Dr. S. G. Foord

ELECTRICAL COMMUNICATION

Technical Journal of the
International Telephone and Telegraph Corporation
and Associate Companies

•

MICROWAVE RADIO LINKS OF BONNEVILLE POWER ADMINISTRATION

MANUFACTURE OF A FIELD TELEPHONE SWITCHBOARD

VERY-HIGH-FREQUENCY OMNIDIRECTIONAL RADIO RANGE ANTENNA

VOICE-FREQUENCY SIGNALING EQUIPMENT

RECEIVERS AND TEST GEAR FOR INSTRUMENT LANDING SYSTEMS

CARRIER CONCENTRATIONS AND FERMI LEVELS IN SEMI-CONDUCTORS

ALIGNMENT OF TUNED MULTIPLE-RESONANT-CIRCUIT FILTERS

Volume 29 **JUNE, 1952** **Number 2**

ELECTRICAL COMMUNICATION

Technical Journal of the
INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION
and Associate Companies

Published Quarterly by the
INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION
67 BROAD STREET, NEW YORK 4, NEW YORK, U.S.A.

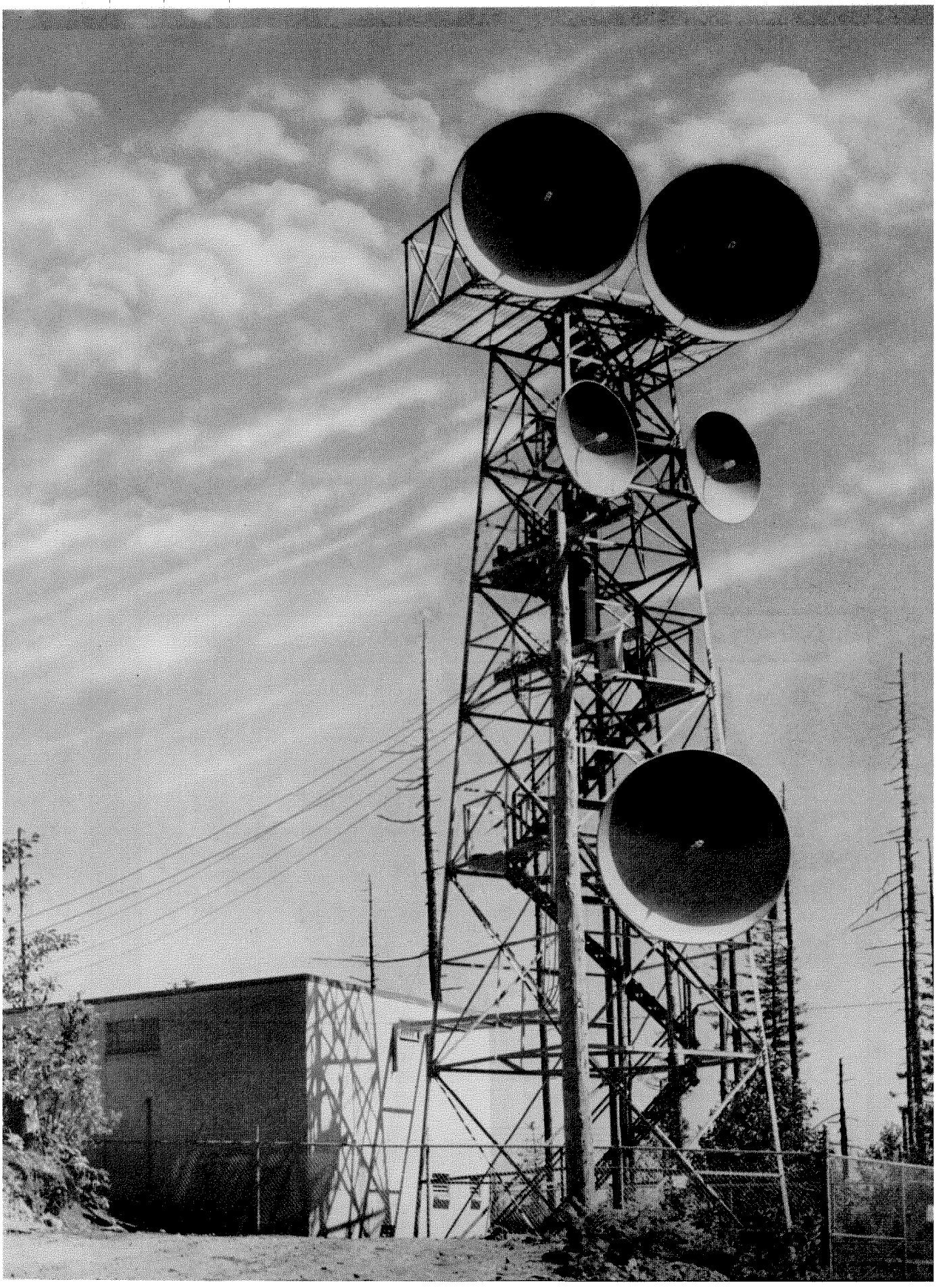
Sosthenes Behn, Chairman William H. Harrison, President
Geoffrey A. Ogilvie, Vice President and Secretary

Subscription, $2.00 per year; single copies, 50 cents

Electrical Communication is indexed in Industrial Arts Index

Volume 29 JUNE, 1952 Number 2

CONTENTS



Microwave antennas on Squak mountain near Seattle, Washington. This relay station is part of the radio system used for supervision and control of the Bonneville Power Administration generating and transmission facilities.

Microwave Radio Links of Bonneville Power Administration*

By SIDNEY METZGER, N. H. GOTTFRIED, and R. W. HUGHES

Federal Telecommunication Laboratories, Incorporated; Nutley, New Jersey

BONNEVILLE Power Administration is responsible for the construction and operation of an extensive electricity supply system based on water power developed by the Columbia River in two of the northwestern states; Washington and Oregon. Mountainous terrain makes the use of microwave radio links among the numerous substations most attractive. Terminal and repeater stations have been installed to provide a communication system linking Seattle, Snohomish, and Covington in Washington State with J. D. Ross substation, which is near Portland, Oregon. A previous paper[1] showed the proposed ultimate system and described a method whereby a fault on a transmission line would automatically be reported and produce a printed record of the time of its occurrence and its distance to within a tenth of a mile from the adjacent radio station.

The microwave system now in operation is shown in Figure 1. It will provide facilities for telephone channels, telemetering, power-line relaying, location of faults on the power lines, very-high-frequency mobile radio tie-in, and any other services normally carried over wire-line or power-line carrier systems.

The original installation provided for 5 channels between J. D. Ross and Snohomish as shown in Figure 2. After a few months of operation, plans were made to expand to the full 23 channels for which the radio equipment was designed. The main backbone of the system consists of the J. D. Ross terminal station; the two repeater stations, Chehalis and Rainier; Olympia, which is now an audio repeating terminal; Squak Mountain; and Snohomish. Seattle and Covington can be considered as branch or spur terminals.

* Presented under the title "Application of a Microwave Radio Link by the Bonneville Power Administration" at the Pacific General Meeting of the American Institute of Electrical Engineers at Portland, Oregon, on August 23, 1951.

[1] R. W. Hughes and Nelson Weintraub, "Fault Location by Pulse-Time Modulation," *Electrical Communication*, v. 28, pp. 90–94; June, 1951: and *Electrical Engineering*, v. 69, pp. 1009–1011; November, 1950.

1. Multiplex Equipment

The multiplex equipment contains means for combining up to 23 independent voice-frequency channels into a time-modulated pulse train and for demodulating the pulse train after transmis-

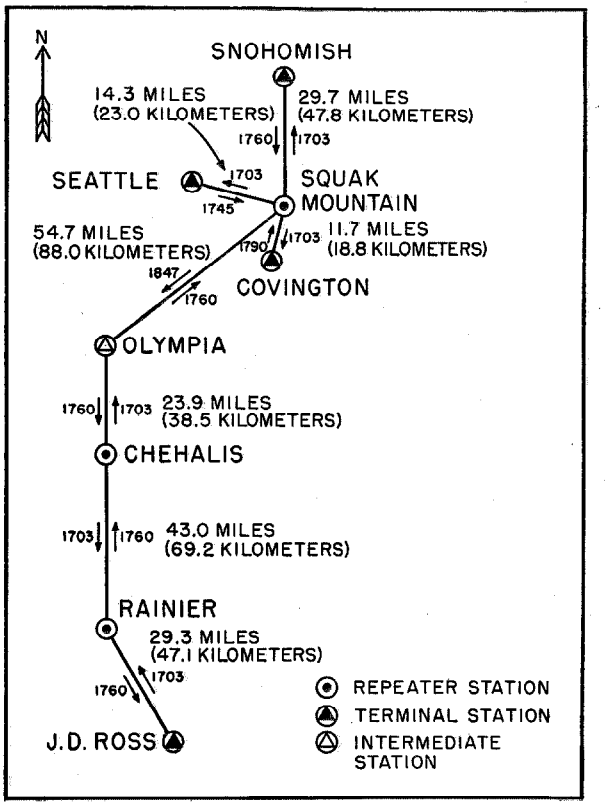


Figure 1—Microwave system now in operation. The distance in miles (kilometers) between adjacent stations is shown together with the transmission frequency in megacycles for the direction indicated.

sion to recover the original audio-frequency signals for delivery to a wire line. The time-modulated pulse train is suitable for keying a radio-frequency carrier. Combination of the several channels into one pulse train is performed in the multiplex modulator equipment. Separation of the several channels and the restoration of the audio signals is performed in the multiplex demodulator equipment. Where one or more

channels are required at repeater points, drop-and-insert equipment may be employed. This equipment is used at the repeater stations, Rainier and Chehalis, for maintenance and alarm transmission.

The multiplex equipment is mounted in cabinets in this particular installation. Figure 3 shows a 23-channel multiplex modulator as used at Squak Mountain. Standard arrangements include a 23-channel multiplex modulator bay, a 23-channel multiplex demodulator bay, or an 8-channel multiplex terminal bay incorporating both modulators and demodulators. Special arrangement of the basic units are constructed to meet particular requirements. For compactness and ease of servicing, the individual channel modulators and demodulators are made on a plug-in unit basis. The multiplex equipment need be equipped with only the number of plug-in units required for the desired number of channels.

Technical characteristics of the over-all system including the modulators, radio-frequency equipment, and demodulators are given in Table 1.

TABLE 1

CHARACTERISTICS OF *FTL-10-B* SYSTEM

Number of Channels	23 per Carrier, Maximum
Channel Bandwidth	300 to 3400 Cycles
Channel Response–Frequency Characteristics	Within +1 and −2 Decibels of the 1000-Cycle Level
Audio Input for 100-Percent Modulation on 4-Wire Basis	−20 Decibels Referred to 1 Milliwatt
Audio Output for 100-Percent Modulation on 4-Wire Basis	+2 Decibels* Referred to 1 Milliwatt
Audio Input and Output Impedance	600 Ohms, Balanced
Signaling-Channel Bandwidth	0 to 10 Pulses Per Second
Harmonic and Spurious Distortion in Voice Channels	<5 Percent
Signal-to-Noise Ratio, Weighted	>60 Decibels
Cross Talk Between Channels	>60 Decibels Below Normal Signal Level
Pulse Circuit Bandwidth	2.8 Megacycles

* Increased to +7 decibels in a later design.

2. *Radio-Frequency Equipment and Antennas*

The radio-frequency equipment is designed for operation both in the government band of 1700–1850 megacycles and in the nongovernment band of 1850–1990 megacycles. For this application, allocations in the first band are used. It consists of the transmitter, receiver, automatic switchover units (when required), and associated power supplies, antennas, radio-frequency transmission lines, and dehydrating equipment. The individual units of the radio-frequency equip-

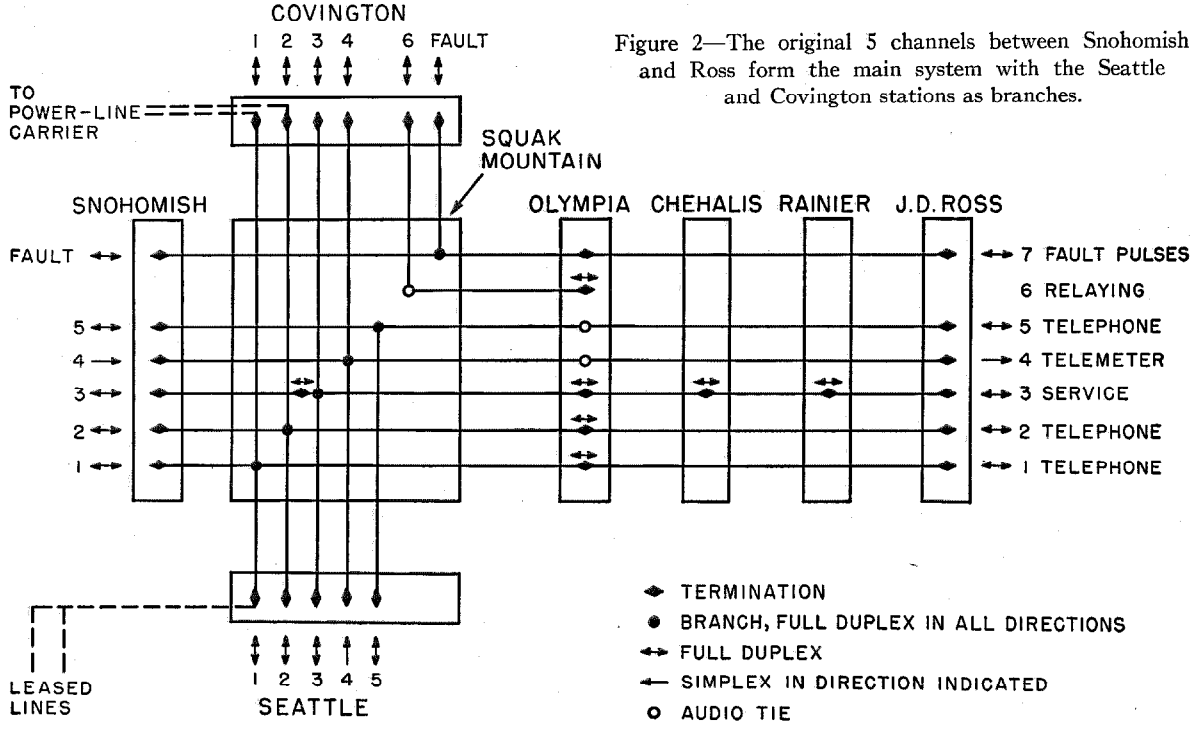


Figure 2—The original 5 channels between Snohomish and Ross form the main system with the Seattle and Covington stations as branches.

ment at terminal and repeater stations are identical but the exact number and type of chassis used depends on the requirements of the particular installation. The equipment is easily serviced because of the open accessible-chassis construction.

The pulse-time-modulated signal, originating in the multiplex modulator equipment, modulates the transmitter oscillator. The pulsed radio-frequency signal is delivered to the transmitting microwave antenna. At the receiving location, the output of the receiving antenna is converted to an intermediate-frequency signal, amplified, and detected in the receiver. The output of the receiver is a time-modulated pulse train suitable for operation of the multiplex demodulator equipment (at a terminal) or of modulating a radio-frequency transmitter (at a repeater).

The transmitter, receiver, power supplies, and switchover equipment are constructed on flat chassis with the various controls and meters used in routine maintenance projecting through the front panel. Figure 4 shows a bay including two transmitters, two receivers, and two power supplies; one set for use and the other for standby.

The transmitter consists of a 2-stage video amplifier and modulator, which amplitude keys a lighthouse-tube cavity oscillator. Built-in metering circuits facilitate performance checking and a monitor is used to measure relative power output and also to provide an alarm indication when pulses fail to appear at the transmitter output. The transmitter has a carrier peak output of 40 watts.

The receiver is a superheterodyne type with a tuned double-cavity preselector, a cavity crystal mixer, lighthouse-tube local oscillator, 30-megacycle intermediate-frequency amplifier, and a video output amplifier. Metering circuits are provided for maintenance and operation.

Identical antenna systems are used in conjunction with both the receiver and transmitter. In later installations, the use of a diplexing filter permits operation of a transmitter and receiver with a single antenna. The antenna itself consists of a half-wave dipole mounted at the focal point of a paraboloidal reflector. The dipole is protected from the elements by a plastic radome. The line feeding the dipole, which also acts as its support, is a rigid $\frac{7}{8}$-inch air-dielectric line with an external stiffener sleeve to prevent excessive vibrations. Both 6-foot and 10-foot reflectors are used in the Bonneville Power Administration link. The frontispiece shows the Squak Mountain tower with its paraboloidal reflectors. Transmitting and receiving antennas are side by side, while the single dish mounted below the two pairs is for a diversity receiver.

The gain of an antenna is a function of the size of the paraboloidal reflector; the 6-foot reflector provides a gain of 29 decibels and the 10-foot reflector a gain of 34 decibels over a point source. The size of the antenna is dictated by the attenuation of the individual transmission path.

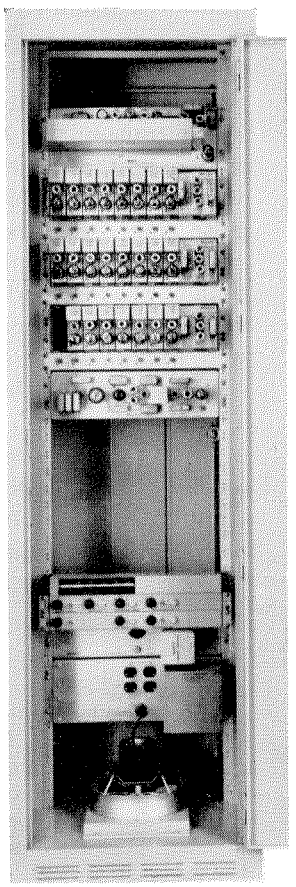


Figure 3—*FTL-10-B* multiplex modulator. The individual-channel plug-in modulator units are near the top of the rack.

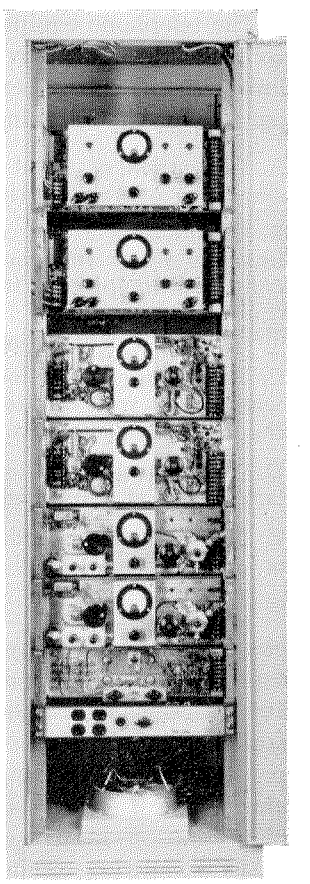


Figure 4—Radio-frequency bay. Operating and standby transmitters, receivers, and power supplies are included.

The antenna assemblies are provided with an azimuth adjustment of approximately ±5 degrees. The initial orientation was based on geodetic maps and a surveyor's transit. Final adjustments were made by orienting the antennas for maximum received signal strength, as indicated by the amplitude of the automatic-gain-control voltages in the receivers.

The feed lines used in conjunction with the antennas are solid-dielectric 50-ohm coaxial cable (*RG-17/U*), where the length of line required at both ends of a link does not exceed 100 feet. Greater lengths introduce excessive attenuation and semirigid or rigid coaxial air-dielectric structures are used for them.

Diversity receivers are used on all hops except the two short ones from Squak Mountain to Seattle and Covington. Diversity reception is simply accomplished by paralleling the outputs of two receivers and paralleling their automatic-volume-control systems.

3. *Telephone Terminating Equipment*

The telephone equipment, shown in Figure 5, connects the multiplex equipment of the radio link with the subscriber circuits. It includes apparatus for the following functions.

A. Conversion of the 4-wire multiplex equipment circuits to 2-wire voice circuits.

B. Operation of 2-wire common-battery telephones from the 4-wire voice and signaling circuits of the multiplex equipment, including selective ringing equipment by means of direct-current dialing on the direct-current path of that particular channel.

C. Power supply, dial tone, busy tone, ringing current, and 24 volts of direct current for operation of the telephone equipment.

D. Test, transfer, and isolation of circuits through jack fields and patch cords.

E. Failure-alarm transmission and remote indication of failures at unattended stations.

The system is provided with dial tone if the channel is idle and busy tone if the channel is in use. Selective ringing equipment permits up to 100 selections on any one channel for telephone stations, other telephone connections such as private branch exchanges or tie lines, failure-alarm indications, or failure-alarm checks. Telephone stations may be connected on a preference basis, permitting talking on a busy channel, or they may be connected to provide privacy on an already-busy channel.

Four-wire terminating sets with compromise networks are provided for conversion of the 4-wire circuits to 2-wire voice circuits. Where the connected lines require more elaborate balancing, an adjustable balancing network has been provided. Fixed pads (plug-in type) are incorporated to adjust levels to standard conditions. Voice connections are provided through networks at all branching points or way stations.

Jack fields permit substitution of multiplex channels whenever necessary to maintain service during technical difficulties. Jacks are also included at other points for convenience in testing and monitoring. All jacks are normalled through and require patch cords only for transfer of channels or during abnormal conditions. Single 3-conductor jacks are used exclusively.

4. *Alarm Transmission and Indication*

Equipment failures that can be made to close an alarm circuit, can actuate the failure-alarm transmitter and the remote indicators. Closure of the alarm circuit causes the alarm transmitter to send a unique series of coded pulses. This pulse train, when received at any trunk circuit, can be made to operate an associated alarm indicator visually signaling the failure.

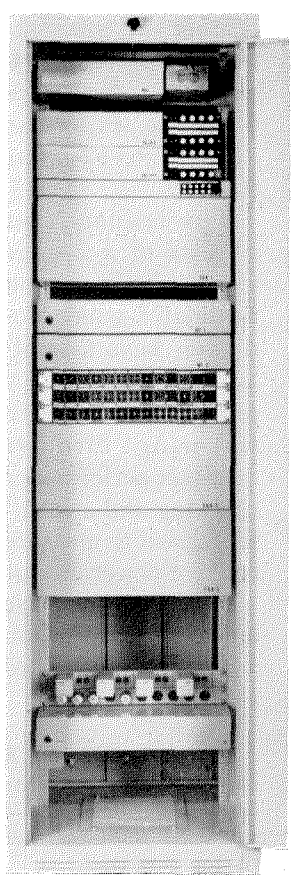


Figure 5—Telephone terminating equipment for the J. D. Ross station. It includes trunk circuits, fault-alarm indicator panels, ringer-circuit terminating networks, and jack fields.

To prevent false codes from providing erroneous information, a checking system is provided which permits: A) interro-

gating any failure-alarm transmitter to determine if a failure-alarm circuit has been closed and an alarm sent, and B) resetting a failure-alarm transmitter and allowing it to retransmit information concerning the particular alarm circuits still closed. The latter operation permits clearing the alarm transmitter at a station when the fault was due to failure of stations elsewhere in the chain.

5. *Automatic Switchover*

Because of the necessity for the maximum possible uninterrupted service, the link has been equipped with complete standby radio-frequency equipment. The transfer of receiving and transmitting antennas to the standby equipment is accomplished by coaxial antenna switches, which are automatically controlled by the switchover chassis. The switchover unit must accommodate a variety of conditions, since the switchover problems at a repeater are different from those at a terminal or at a location using diversity reception. Therefore, a basic design is employed with circuit details varied to meet the requirement of the particular station. In addition to the switchover function, the chassis initiates a secondary alarm to inform the operators of the link that a fault has occurred but that the system is in operation. A primary alarm would also be initiated if complete disruption of communications in one direction occurred.

A pulse-train failure at any point on the main microwave communication link will cause alarms to be displayed at the J. D. Ross and Snohomish terminal stations. The alarm indicates the station of origin and whether the alarm is of primary or secondary nature. Failure information at the intermediate stations is transmitted automatically to one or both of the terminal stations. The two branch stations, Seattle and Covington, are equipped with local alarm indicators only, and do not transmit failure indications to the terminal stations. Outgoing-pulse-train failures at these two branch stations are indicated by automatic alarm-transmitting equipment at the Squak Mountain branching station. Incoming-pulse-train failures at these branch stations, if due to the local receiver or multiplex equipment, will be apparent from other stations by failure to receive revertive ringing tone on the telephone channels.

Failure alarms are transmitted and received by equipment connected to the telephone trunk circuits on service 3. To avoid interference with communication and to reduce the probability of garbled alarms, the trunk circuit is equipped with a lockout feature that prevents transmission of alarms as long as the telephone circuit is busy. As soon as the circuit is idle, the failure-alarm transmitter seizes the circuit and transmits the alarm indication as a series of pulses (similar to dial pulses). Each code is sent twice for greater reliability, and the alarm transmitter then releases the circuit for normal use. If the circuit should be made busy by other equipment during an alarm transmission, the alarm transmitter will stop and wait for the idle condition before continuing its transmission.

6. *Power-Line Fault Locator*

The problem of locating faults on a power line has long plagued the power industry. It is particularly troublesome since approximately 90 percent of all faults are not sustained and hence the location of the fault must be determined on an instantaneous basis. To keep a record of the time and location of all faults, a method[1,2] based on the measurement of the time required for a wave to travel from a fault to a substation appears most practical. The circuit elements of the system are given in Figure 6.

A. A fault occurs between station *A* and station *B*, the two ends of the power line. This starts a steep-wave-front surge propagating down the line in both directions at 0.186 mile per microsecond.

B. Assuming station *B* to be the recording station, the arrival of the surge causes an electronic counter to be triggered.

C. At station *A* the surge is detected, converted to a pulse of convenient size and shape and then transmitted from *A* to *B* via the microwave link.

D. The arrival of this surge-derived pulse at *B* stops the electronic counter, and from the elapsed time the location of the fault can be calculated.

E. The counter is arranged to read distance to the fault directly and by use of a suitable associated printer, a permanent record is made.

The insertion of the power-line-fault-generated surge into the microwave system occurs

[2] R. F. Stevens and T. W. Stringfield, "Transmission Line Fault Locator Using Fault-Generated Surges," *Transactions of the American Institute of Electrical Engineers*, v. 67, pp. 1168–1178; 1948.

in a chassis called the fault-pulse mixer. This fault pulse is inserted into the pulse train, which is momentarily interrupted during the time of insertion. The duration of the interruption is approximately 10 microseconds and no perceptible interruption to voice communication occurs. The normal pulse train plus the fault pulse is transmitted down the link. At the repeaters (Chehalis and Rainier), this combination pulse train is simply passed through in the normal manner. However, at an audio repeater, such as Olympia and Squak Mountain, a fault-pulse

A conventional broadband oscilloscope is used with a synchronizing unit. The function of the synchronizing unit is to allow each channel pulse to be examined on a fast sweep showing only one or two pulses at a time. An additional function of the synchronizing unit is to provide timing pips locked to the pulse train to allow adjustment of channel pulses to the correct time position.

The radio-frequency equipment requires no additional test apparatus since current monitoring meters and frequency-determining wave-

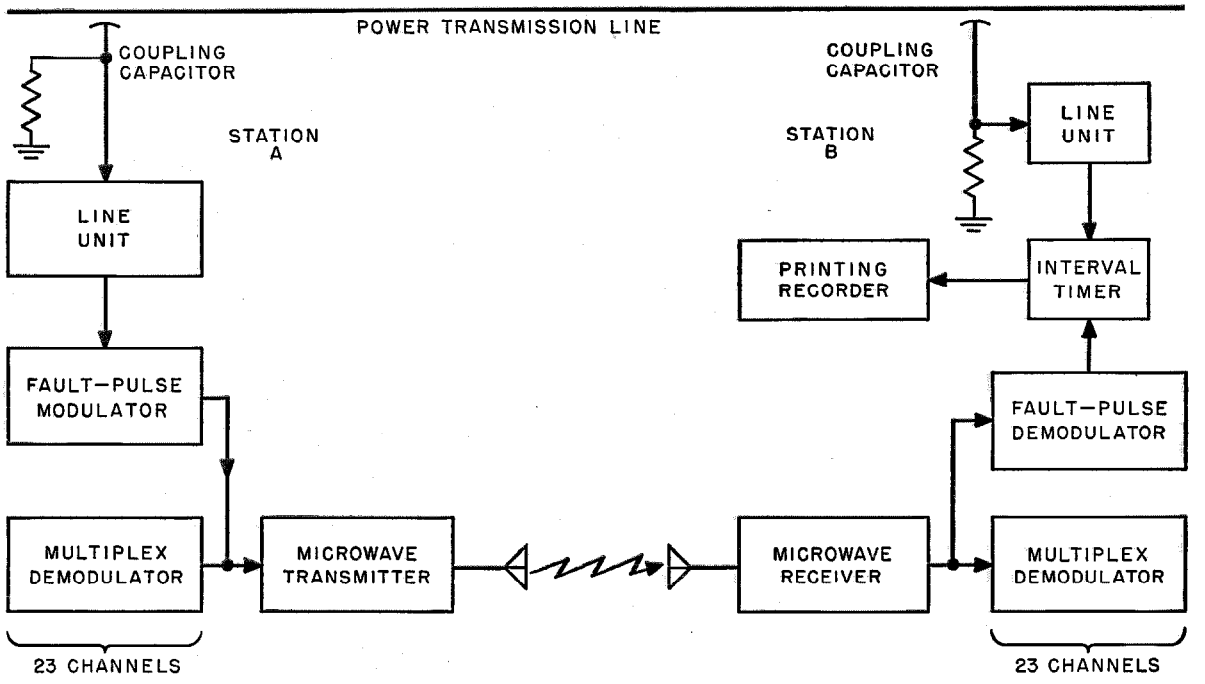


Figure 6—Power-line-fault location using self-generated surges and microwave radio links.

demodulator and a fault-pulse mixer are employed, since it is necessary to detect and reinsert the fault pulse into a new pulse train. The fault-pulse detector recognizes the fault pulse by means of an integrating circuit followed by a differentiating circuit to define sharply the trailing edge of the fault pulse. This information is then sent in suitable form to the fault locator.

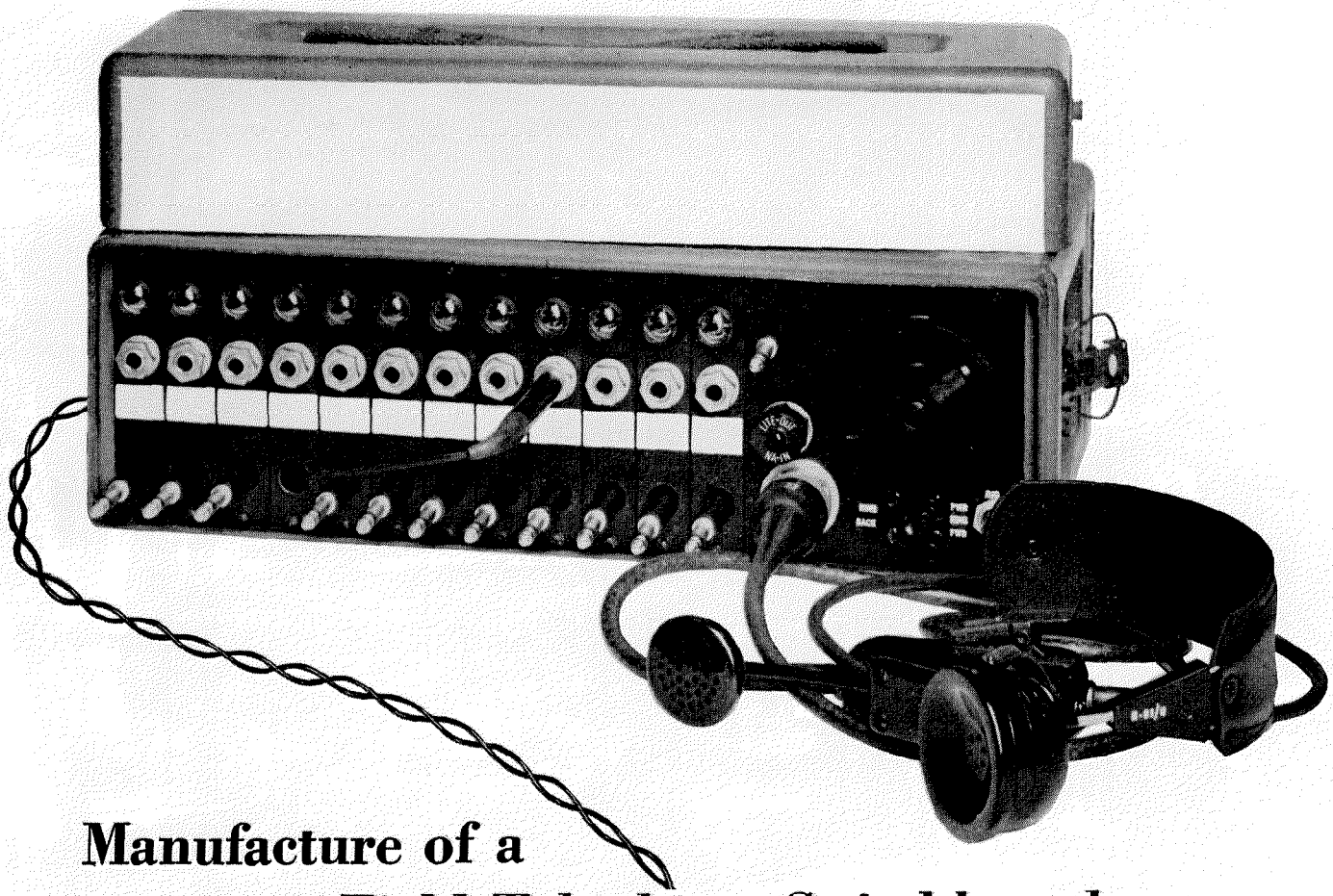
7. Test Equipment

The test equipment required for the microwave radio link consists of a pulse-time-modulation oscilloscope and a vacuum-tube voltmeter.

meters are mounted in the radio-frequency unit.

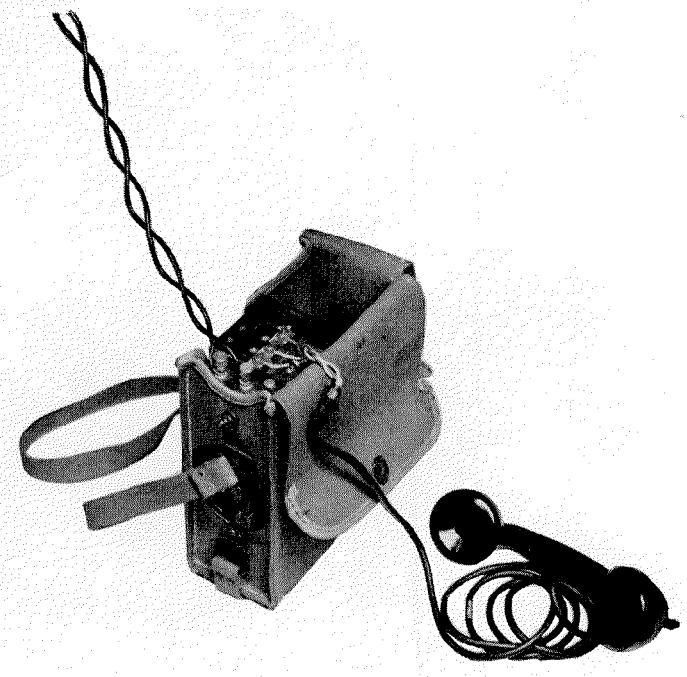
8. Performance

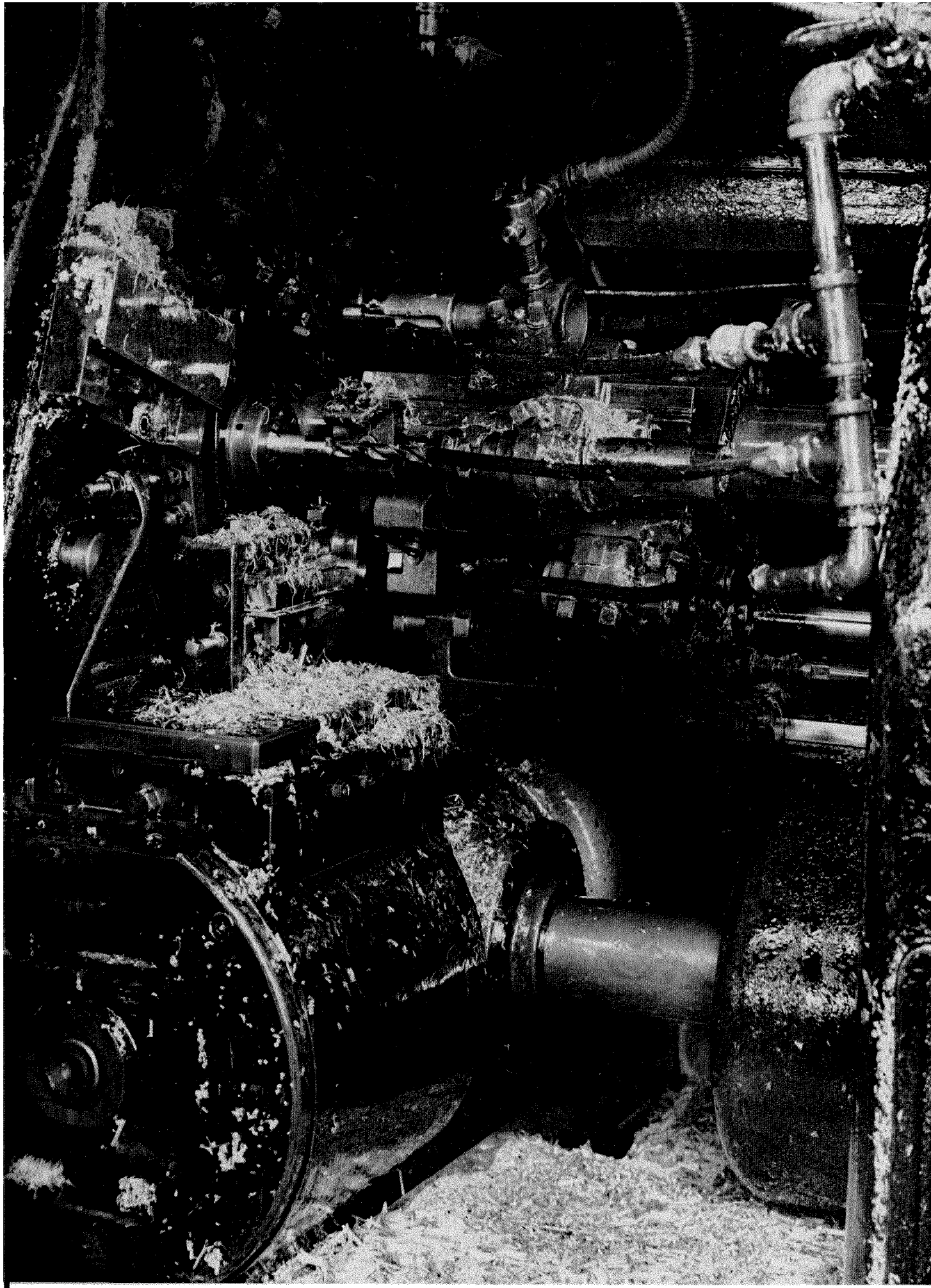
This link was put in service in November, 1950. During the last half of 1951, the ratio of on-the-air circuit time to available circuit time was 99.5 percent. The 0.5-percent outage for all 8 stations included time for maintenance, trouble, and travel incidental to the repair of the trouble. For the first quarter of 1952, this outage figure was reduced to 0.38 percent.



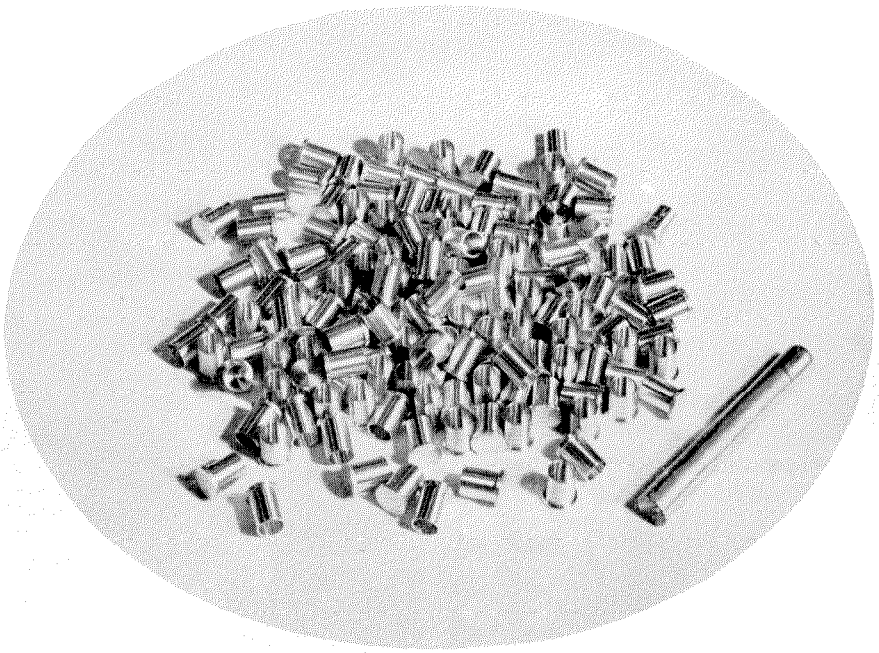
Manufacture of a Field Telephone Switchboard

PRESENTED on the following pages are a series of photographs illustrating the major steps in the manufacture of the *SB-22/PT* field telephone switchboard shown above. This 12-line monocord board, developed by Federal Telephone and Radio Corporation in conjunction with United States Army Signal Corps engineers, is used for switching local-battery telephone lines and voice-frequency telegraph circuits. It will often be used with the familiar type *EE-8* field telephone illustrated at the right. The switchboard has as its main features light weight, small size, and the fact that it is immersion-proof when the cover is fastened. Each line circuit and the operator's circuit are mounted in separate packs that may be unplugged from the case for ready replacement. Up to 3 switchboards may be interconnected to handle 46 lines by replacing the extra operator's packs with line packs.

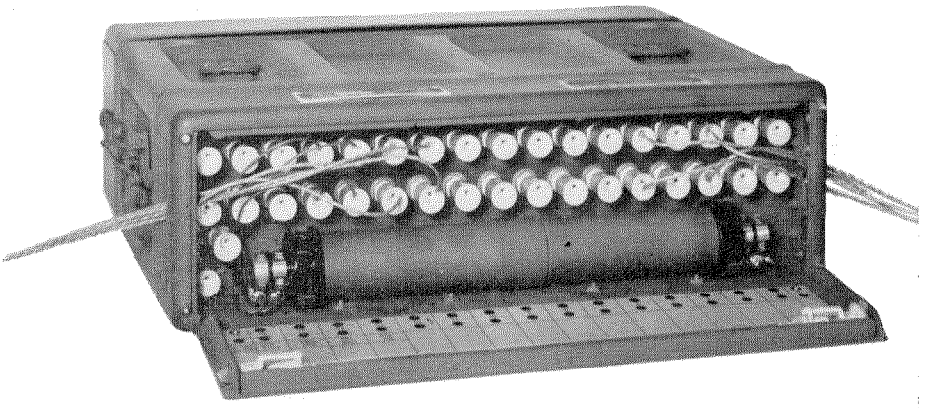


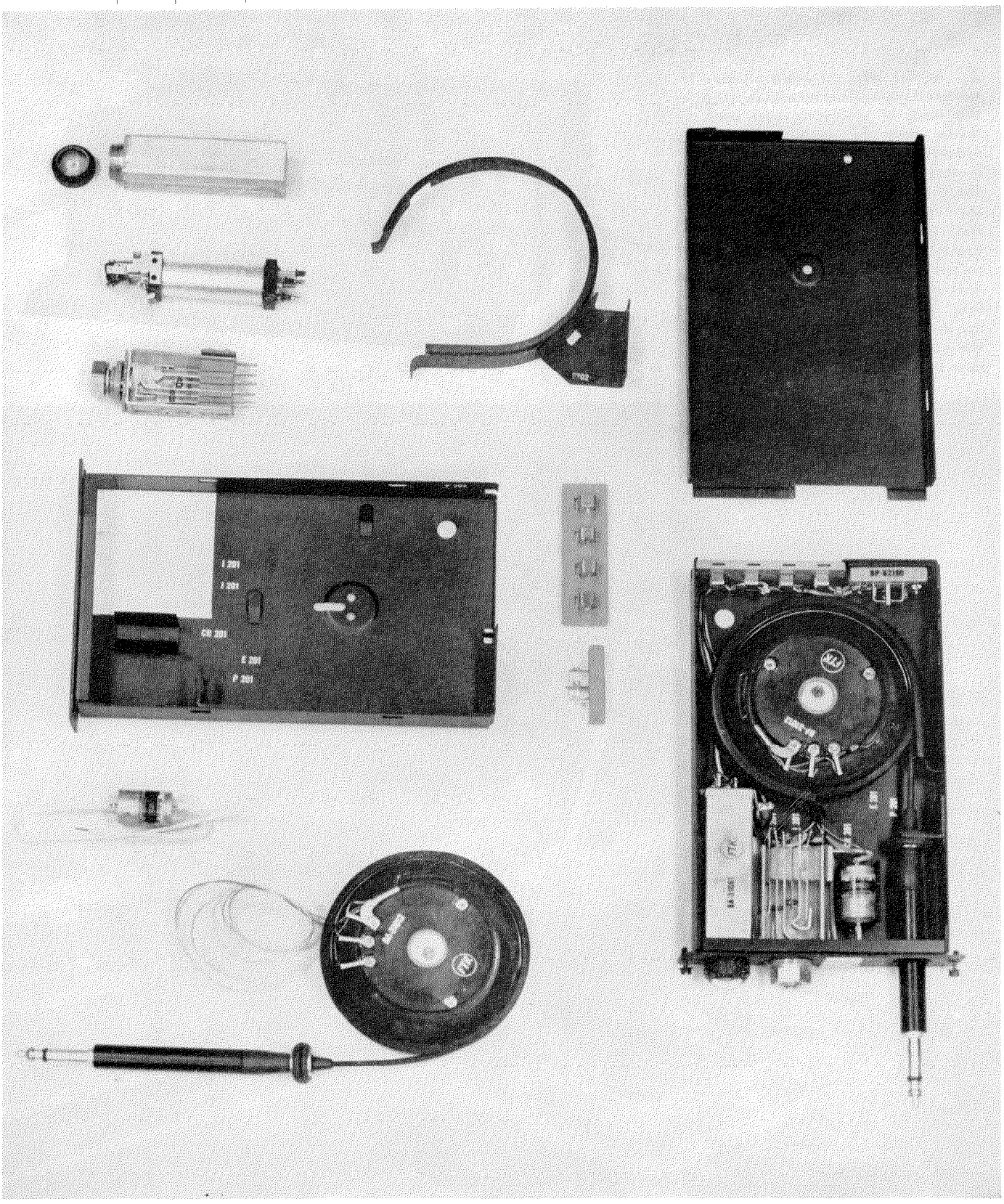


1. The separate pieces that comprise a switchboard number in the thousands. Most of these are made on automatic machines and, for the purpose of illustrating a typical operation of this type, the machining of the parts of a terminal are presented on these two pages. A 6-spindle automatic screw machine (left) is used to cut the cap of the terminal from solid brass rod. Only the head of the screw machine is shown; in operation, oil will flush the chips away. The finished caps are at the right. The body of the terminal is similarly made of brass on a screw machine, and finishing operations on both body and cap are made on machines such as that below. The parts are loaded into the revolving turret, so that two holes may be drilled in the body, and a slot is rough milled and finish milled.



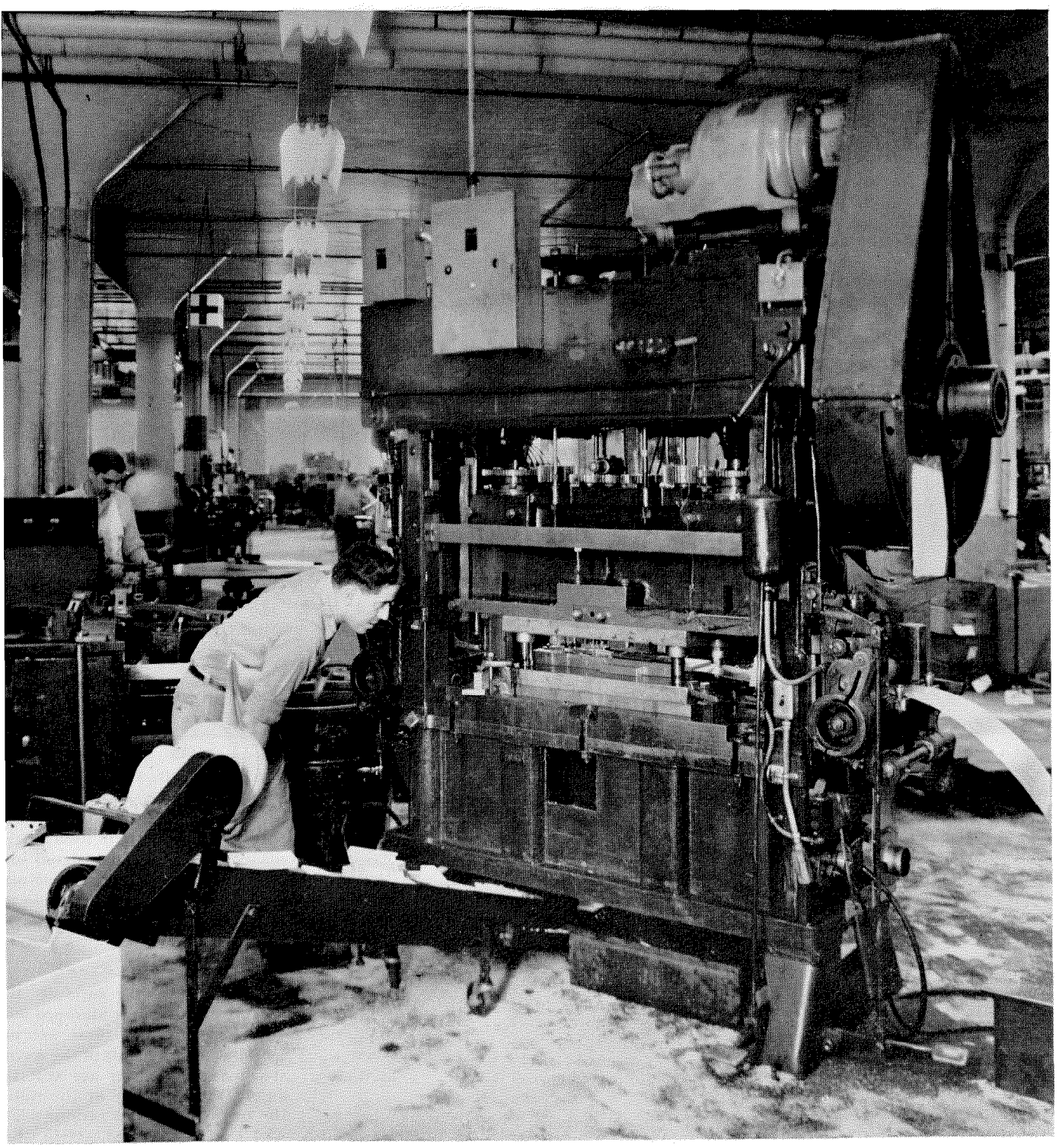
2. At the left, an assembly-line worker fastens the terminals into the back of a switchboard case. Wires may be connected very quickly to these terminals: the cap is depressed with the thumb and the wire is inserted in the hole in the cap. When pressure is released, the terminal grips the wire very firmly (right). The waterproof tube in the back of the case contains batteries for talking and signalling. Below, after the terminals are installed, a cover is put on and the case is checked for watertightness by immersing it for 48 hours.

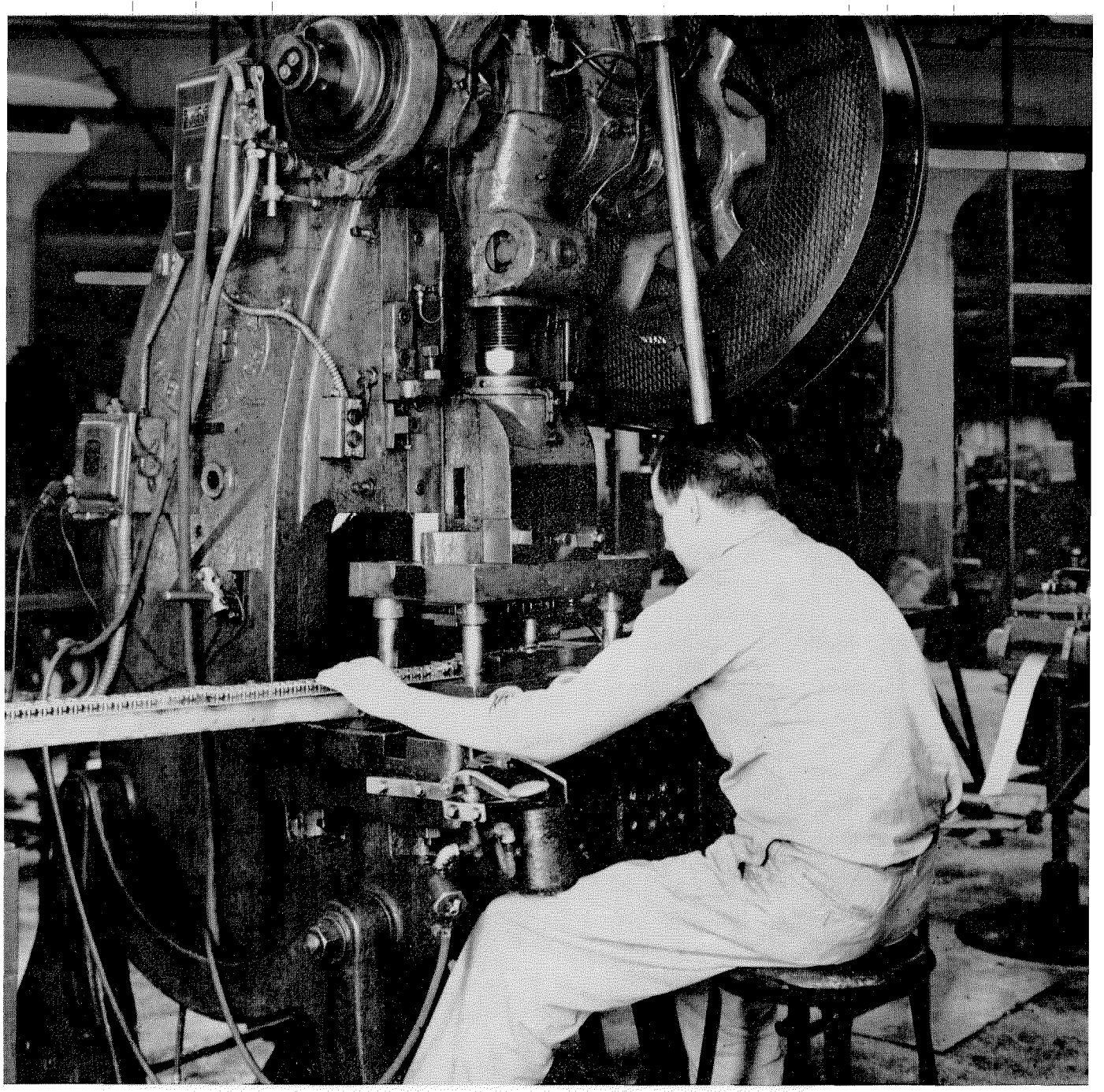


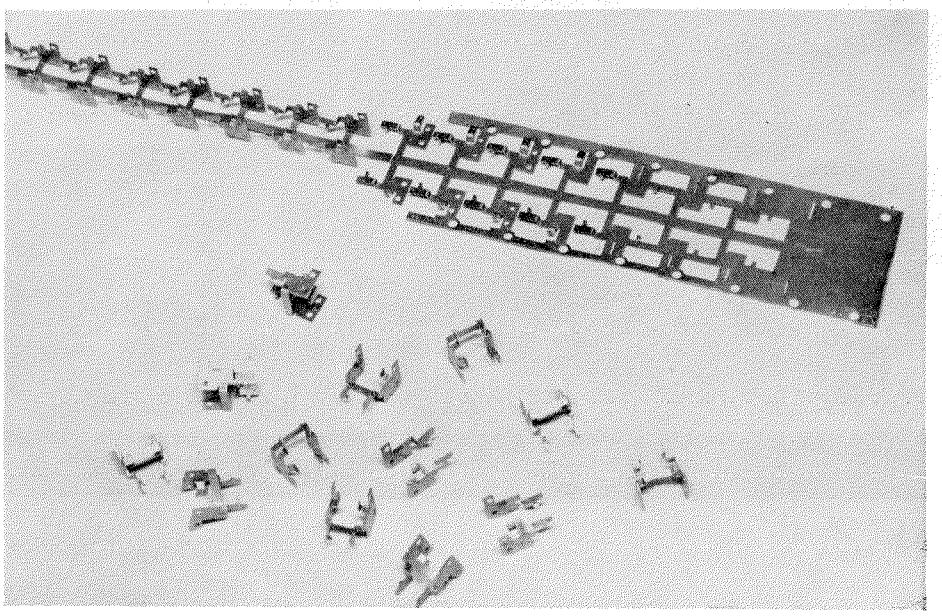


3. The components that make up a line pack are shown here, together with a complete pack. At the upper left is the case for the magnetic signal with the coil and mechanism just below it. Next down is the jack, then come the case for the pack, a selenium rectifier for operation of the signal, and the reel and plug. This reel is a special design in which three spiral beryllium-copper springs inside the body of the reel complete the electrical circuit, thus eliminating the need for sliding contacts. In the center of the picture are the lightning arrester and the contacts that fit into the back of the line pack. When the line pack is inserted in the switchboard case, these gold-filled contact springs engage plugs that are directly connected to the terminals in the back of the case. Removal and replacement of a damaged line pack is thus a simple plug-in operation.

4. Typical of the sheet-metal-forming operations in the manufacture of the *SB-22/PT* is the fabrication of the line-pack case on the press shown below. A wide strip of aluminum is fed into the right side of the press from a large coil. A progressive die in the press first cuts all holes in the cover and then bends up the sides. A case is obtained each time the punch descends. The finished cases fall inside the press to be brought out on a conveyor belt and dropped into the carton at the left. The cases will next be given a protective finish before the apparatus is mounted.

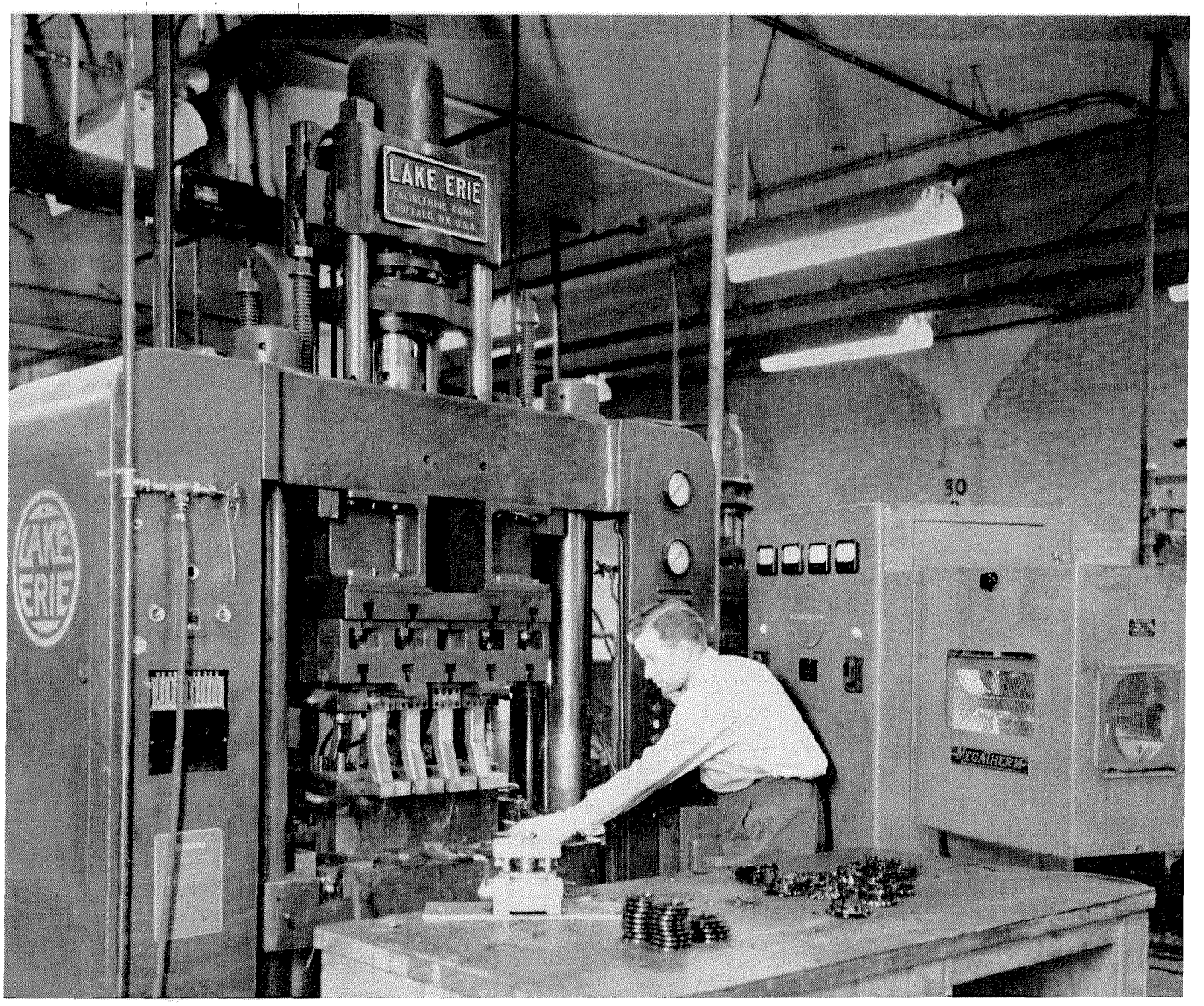


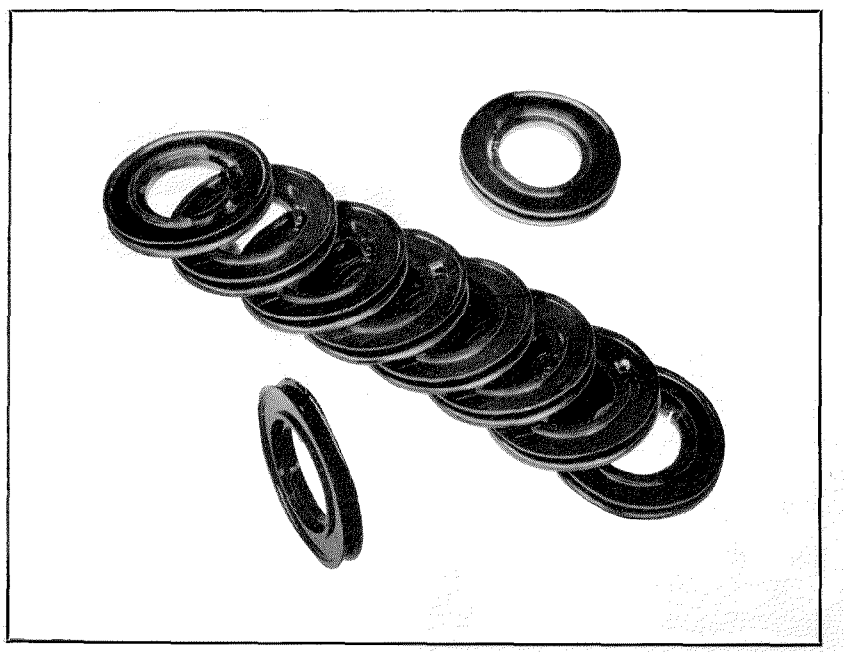




5. At the left, the formation b
a progressive die of a complicate
bracket from a brass strip is indi
cated. This bracket is used in th
magnetically operated signal an
contains several different bends an
a good many holes. The press o
which the die is used is shown abov

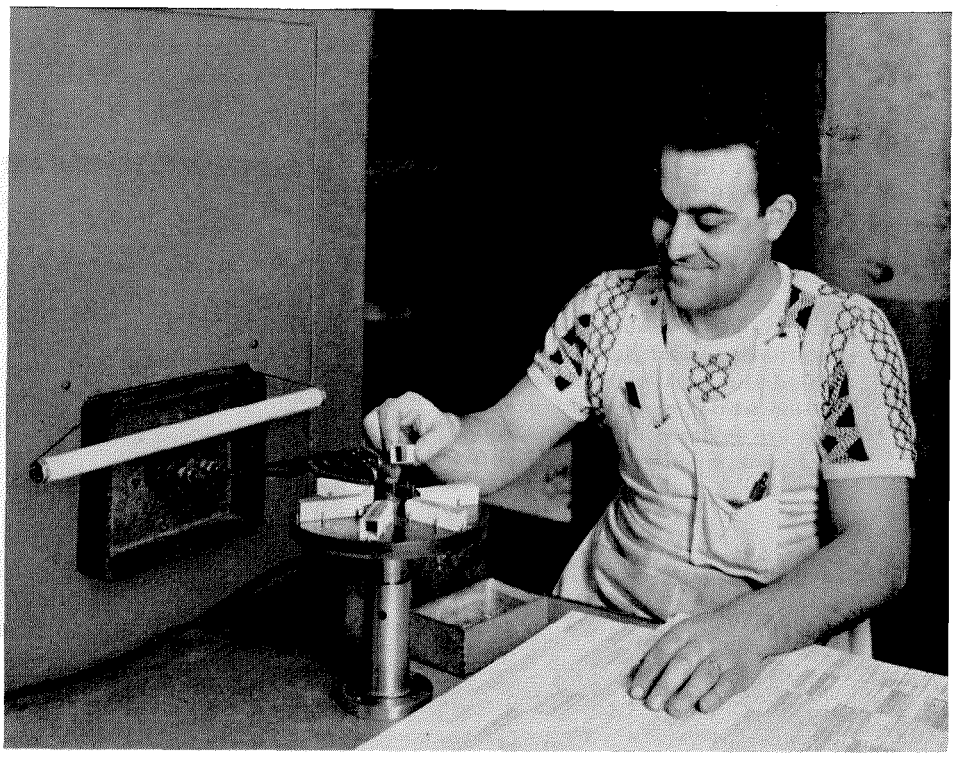
6. No matter how carefully th
parts for the jack are stamped ou
and how carefully these are as
sembled, adjustments of the sprin
spacing and tension are still re
quired. On the page at the righ
the inspector inserts various gag
into the jack to be certain tha
good contact will be made with
standard plug and that the auxiliar
contacts open and close properl





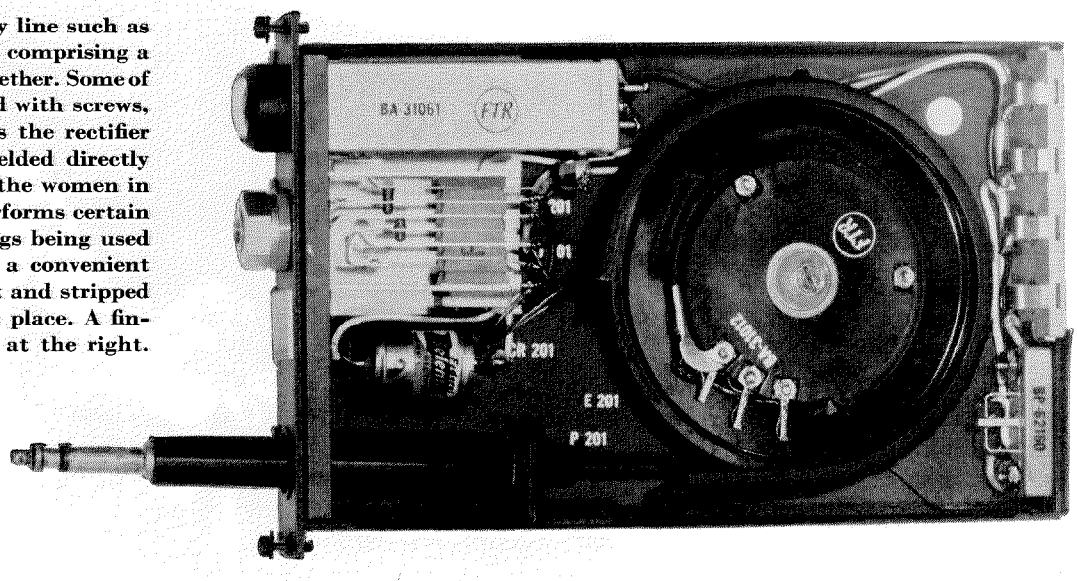
7. When the plugs in the operator's pack and in each line pack are not in use, the 35-inch cord is wound on a reel (left). Four reels at a time are molded from a thermosetting plastic on the press shown above. This is a rather complicated molding, since it contains several holes and the groove for the cord. Briquettes of raw plastic are first preheated in the Megatherm high-frequency heating machine at the right and then are placed in the steam-heated dies on the press. As the press closes, the bars on the sides of the dies force steel slides into the plastic to form the grooves in the reels. The press remains closed for about two minutes while the plastic sets, and the cycle is repeated automatically. When the parts come out of the press, the operator inserts them in a small jig to break off the flashing.

8. The coil for the magnetic signal is wound on its bobbin on the machine above. Three coils are wound simultaneously, each consisting of about 15,000 turns of very fine wire. The function of the drop, which has a fluorescent flag, is to signal the operator when a party is calling and again when the call has been completed.



9. At the right an operator is using a high-frequency induction heater to solder the seam on magnetic-signal cases. Flux and a strip of solder are placed along the seam and the piece is put on the rotating table. As each part is turned under the coil, the solder melts, filling the seam. The heating circuit is monitored by the fluorescent tube, which being in the electromagnetic field, lights when power is applied.

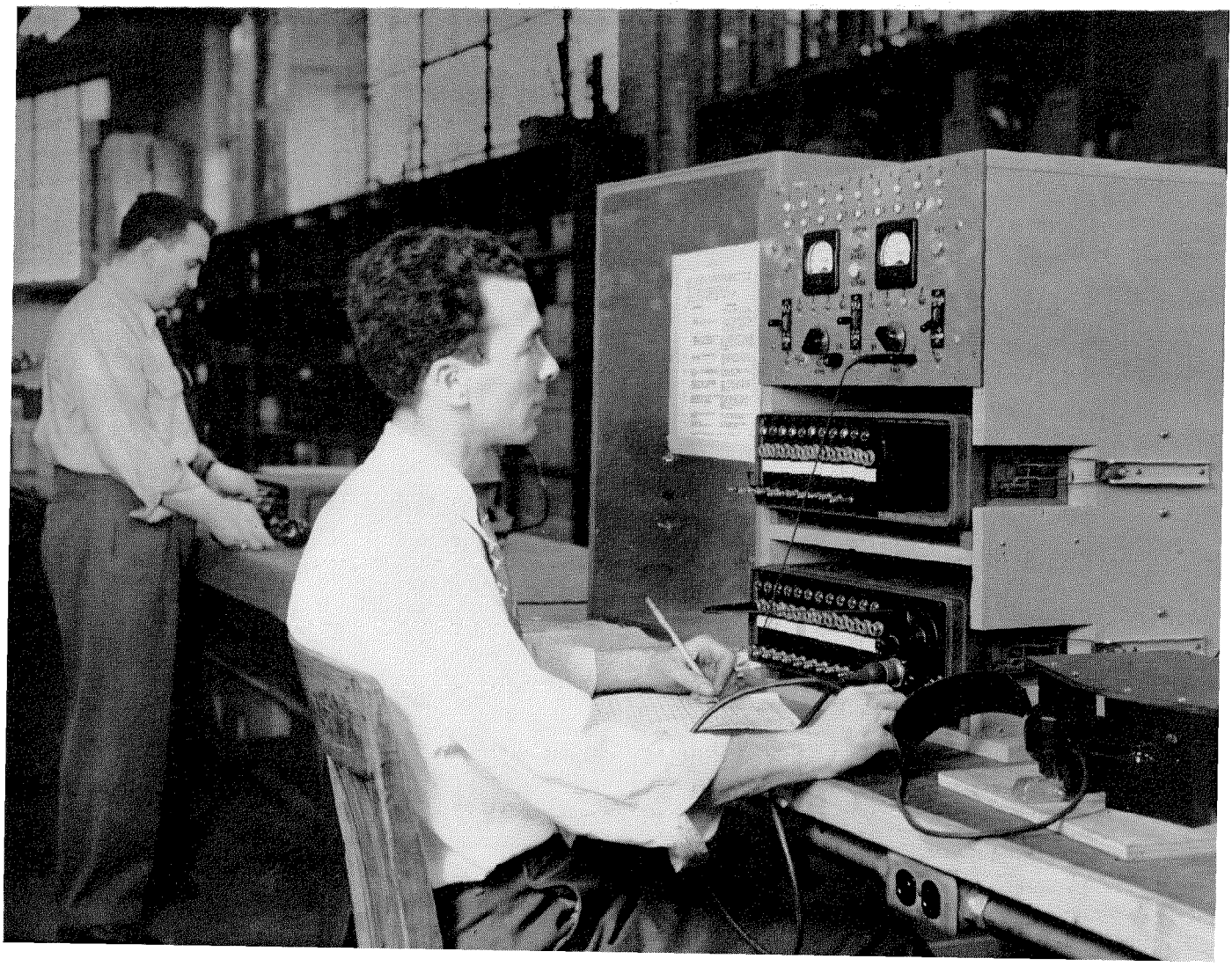
10. On an assembly line such as that above, the parts comprising a line pack are fitted together. Some of the parts are fastened with screws, while others, such as the rectifier bracket, are spot welded directly to the case. Each of the women in the above picture performs certain operations; special jigs being used to hold the pack in a convenient position while precut and stripped wires are soldered in place. A finished pack is shown at the right.



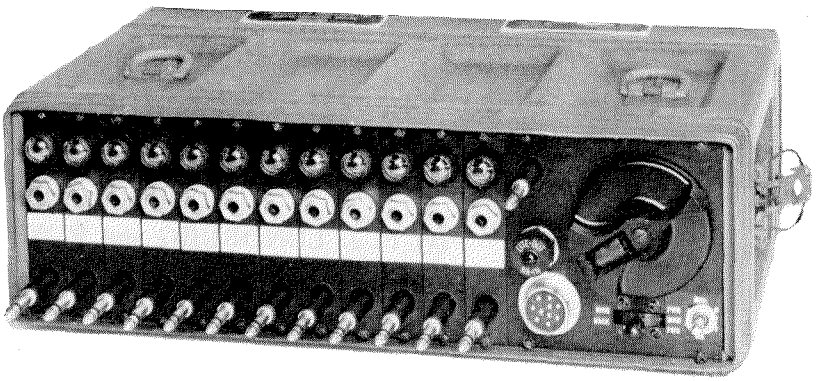
11. An operator's pack is being assembled above. It contains a reel and plug of the kind used in the line packs. In addition, it contains an induction coil, audible night alarm, magneto for ringing extension telephones, and a shielded light for operation in the dark. External ringing current may be connected.

12. Automatic testing apparatu completely checks the operation the finished switchboard. Suc apparatus is shown on this pag At the left is the test of the ope ator's pack. In the picture, th voltage developed by the magnet ringer is being measured. Aft this, by merely depressing a bu ton, the machine will automati ally cycle through a series of oth tests in which each function the pack is checked, the test end ing with high voltage being applie to all circuits to check for ground to the case. Lights on the pan indicate whether there are ar defects.

Below is the test for the comple switchboard. This apparatus ope ates in much the same manner that above, but in addition, th transmission capabilities of th operator's headset are measur by application of its output to network simulating many miles army field wire under wet weath conditions. The upper switchboa case in the test rack is used check spare line packs that accor pany each switchboard to the fiel



13. At the right is a front view of the completed switchboard that is now ready for shipment to the field. The board is placed in a cardboard carton and a moisture- and vaporproof paper barrier is sealed by heat around the carton as shown below. In this photograph, a partial vacuum is developed within the paper bag through the hose to bring the paper tightly against the carton. This package and another containing spare parts are enclosed in one of the overseas shipping crates that are stacked in the background. The filled crate will float.



Cage-Type Very-High-Frequency Phase-Comparison Omnidirectional Radio Range Antenna*

By F. J. LUNDBURG and F. X. BUCHER

Federal Telecommunication Laboratories, Incorporated; Nutley, New Jersey

REVIEWING the history of radio ranges, we see the familiar low-frequency (A-N) type range with its five towers giving way to the very-high-frequency visual–aural (VAR) range with its compact antenna array. The visual–aural range represented a considerable advance in the art with its freedom from low-frequency static, operation in a less congested part of the spectrum, and its visual and aural courses providing definite sector identification. These radio ranges, being limited to four fixed courses, soon yielded to a new system capable of handling more air traffic, the very-high-frequency omnidirectional (VOR) range.

1. *Omnidirectional Range*

The omnidirectional range provides the pilot with an infinity of radial courses. Its principle,[1] as is well known, is that of radiating a rotating 30-cycle-per-second limaçon or variable-phase signal the phase of which is directly proportional to azimuth.

The receiver in the aircraft compares the phase of the detected limaçon with that of a 30-cycle reference frequency that is radiated omnidirectionally and with a phase independent of azimuth throughout the entire 360 degrees. The phase difference is the bearing of the aircraft with respect to the site of the omnidirectional range antenna. The total radiated spectrum consists of a carrier at some assigned frequency in the region between 112 and 118 megacycles, 30-cycle sidebands for the variable-phase signal rotating through space, voice and station-identi-

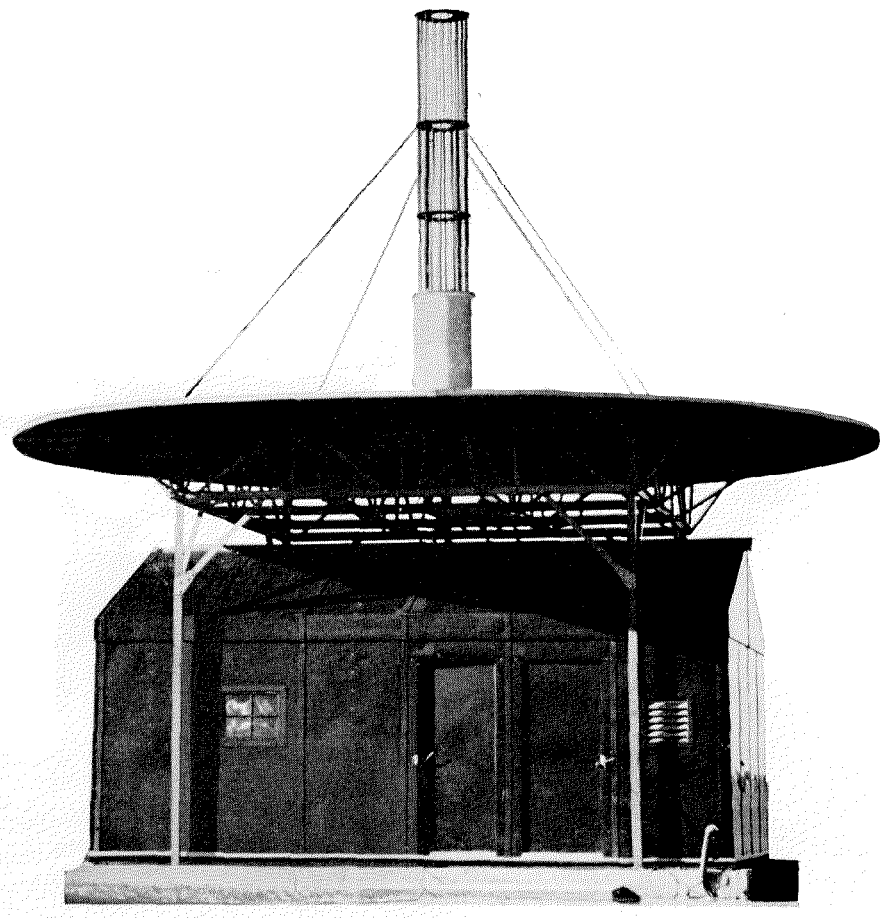


Figure 1—Very-high-frequency omnidirectional range antenna for operation between 112 and 118 megacycles.

* This work was originally sponsored by Rome Air Development Center, Griffiss Air Force Base; Rome, New York. Presented before the Montreal Section, Institute of Radio Engineers, December 5, 1951.

[1] P. C. Sandretto, "Rho-Theta System of Air Navigation," *Electrical Communication*, v. 27, pp. 268–276; December, 1950.

fication sidebands, and the sidebands produced about the carrier by the 9960-cycle subcarriers that are frequency modulated at 30 cycles over a deviation range of ±480 cycles. It is these last-mentioned sidebands that convey the 30-cycle azimuthal phase-independent reference signal. The percentages of modulation of the various parts of the transmitted spectrum are 30 percent for the variable-phase signal, 30 percent for the reference-phase signal, and the remaining 40 percent for the simultaneous voice. Station identification, consisting of a keyed 1020-cycle signal modulated at a 10-percent level, is interrupted when voice is applied.

2. Antenna

One type of antenna system produces the variable-phase signal by means of four antennas arranged at the corners of a square and energized by a rotating goniometer so as to produce two figure-of-eight patterns with mutually perpendicular axes separated 90 degrees in time phase. The reference phase is radiated from a single antenna placed centrally with respect to the four sideband antennas.

Limitations of this type of antenna suggested the desirability of developing a new structure to meet the following requirements.

A. Operation between 112 and 118 megacycles.

B. Simplicity of installation.

C. Use of horizontal polarization and complete absence of vertical polarization.

D. Freedom from octantal errors.

E. Small cone of silence.

It was decided that an antenna using a single rotating element to produce the rotating figure-of-eight pattern was a prime necessity if simplicity of installation and freedom from octantal errors were to be attained. It was realized that when using a rotating dipole severe polarization errors would occur unless steps were taken to suppress the vertical polarization. In the antenna to be described, the polarization purity was made better than that of the five-loop goniometer system.

An antenna that meets the design goal is shown in Figure 1 as set up for field testing. The antenna is a cylindrical cagelike aluminum structure and is divided into three sections:

A. Lower cage.

B. Dipole-rotating unit.

C. Upper cage extension.

The dipole-rotating unit is located coaxially within the lower cage, which is surrounded by a weatherproof housing.

The entire antenna assembly is mounted on a well-bonded circular metallic counterpoise, which is 35 feet in diameter and 15 feet above ground. In the center of the counterpoise, there is a hole over which the antenna is mounted. Bolts pass through holes in a circular plate at the base of the lower cage section to secure the antenna to the counterpoise.

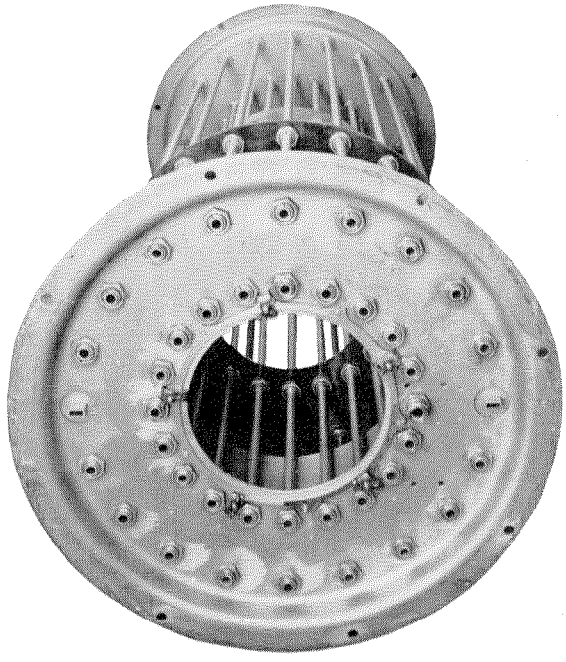


Figure 2—Bottom view of the lower cage showing the recessed ring into which the dipole-rotating unit mounts.

2.1 Installation

The procedure for installing the antenna on the counterpoise is quite simple. The lower cage and its upper extension unit are laid end-to-end on top of the counterpoise and bolted together. Three men can comfortably move the antenna into its upright position over the hole in the

counterpoise. The lower section is then bolted to the counterpoise. The weatherproof housing is next placed over the lower cage section. This housing is made of two half-circular fiber-glass shells that fit snugly around the cage. Four nylon guy ropes with turnbuckles complete the structure. With this arrangement, an antenna has withstood wind velocities up to 100 miles (161 kilometers) per hour.

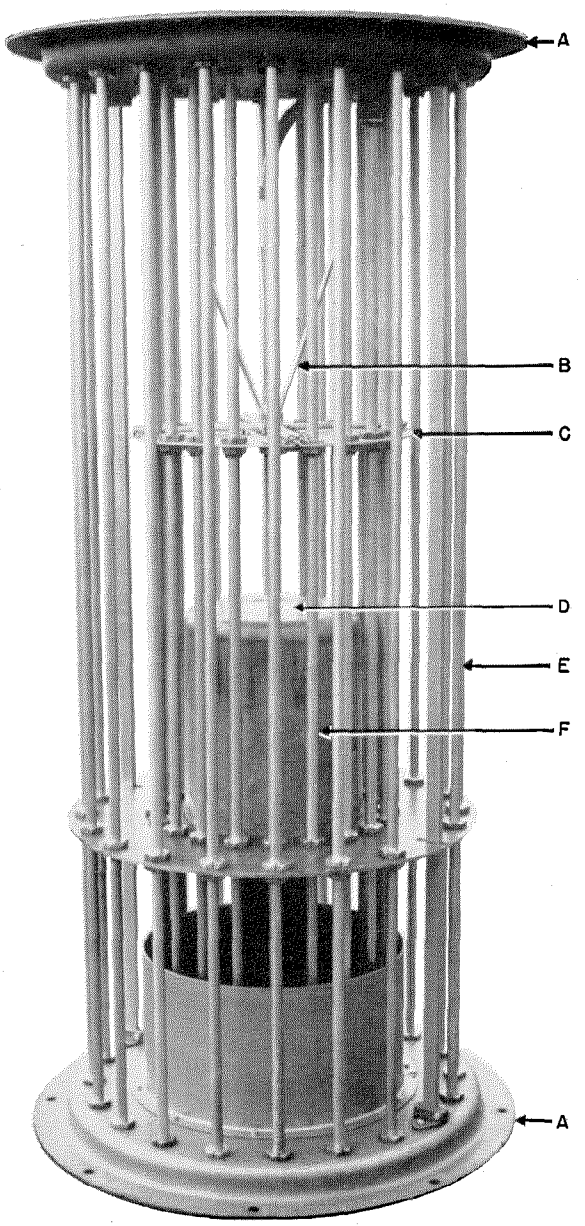


Figure 3—Lower cage without its weatherproof housing.

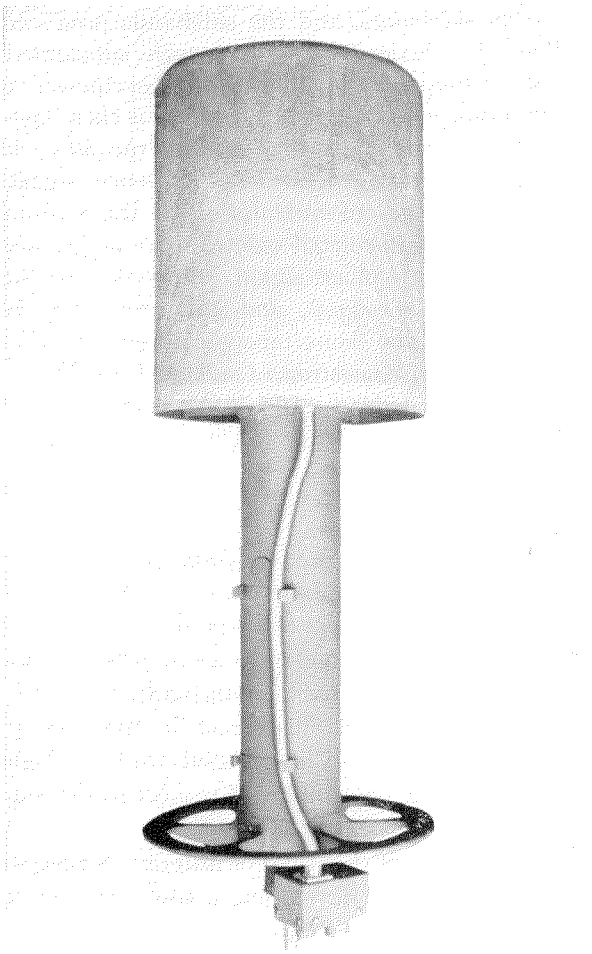


Figure 4—Dipole unit with its fiber-glass protective cover.

Access to the antenna is from underneath the counterpoise. From here the dipole-rotating unit may be inserted easily. Figure 2 shows the bottom of the lower section. A flange on the base of the dipole-rotating unit fits into the recessed ring in the center of the bottom plate and is clamped by 5 wing nuts. When the wing nuts are loosened, the dipole-rotating unit can be turned or may be removed. With this arrangement, it is unnecessary to align the antenna with respect to north at the moment of installation. The course alignment may be made after the installation by turning the dipole-rotating unit in its mouting ring. A circular scale provided for logging its position is particularly useful when replacing a dipole-rotating unit; which process is a matter of minutes only. This allows maintenance work to be performed on the dipole unit without taking the

station off the air for a longer time than is required to interchange two units.

2.2 Lower Cage

Figure 3 shows a close-up of the lower cage with its weatherproof housing removed. This lower cage consists of an outer row of vertical rods *E*, separating two circular end disks *A*. Inside of and coaxial with this outer row of rods is an inner row of rods *F*. Located on this inner row of rods is a third disk *C*, which serves two purposes as will be described later. The dipole-rotating unit *D* may be seen. A circular dielectric plate is used to give mechanical support and to provide a guide for inserting the dipole-rotating unit. *B* are tuning stubs.

2.3 Dipole

Figure 4 shows the dipole unit with its plastic hat, which protects it while it is being inserted in the cage. Figure 5 shows the dipole-rotating unit in the cage with this plastic covering removed to expose the dipole.

The function of this lower cage assembly is to provide two cascaded radial waveguides or transmission lines with discontinuities or obstacles at their junctions to match the dipole to free space. The dipole is only one-tenth of a wavelength long at 115 megacycles [approximately 10 inches (25 centimeters)] and yet when placed in the lower

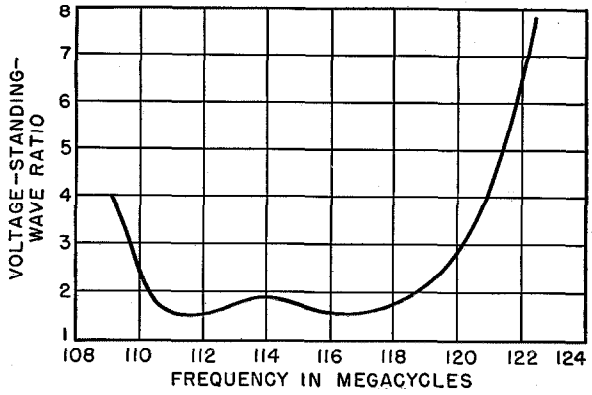


Figure 6—Voltage-standing-wave ratio plotted against frequency of the dipole mounted in the cage.

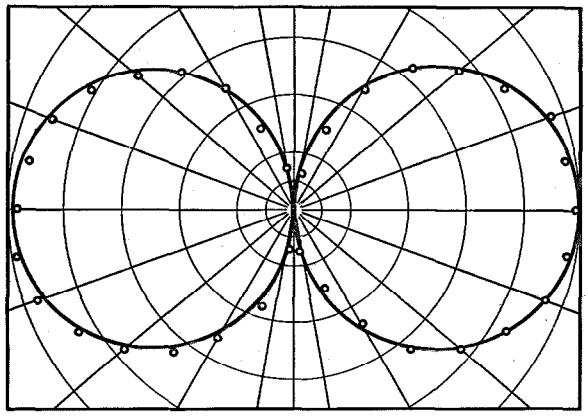


Figure 7—Horizontal radiated field of dipole in cage antenna. The curves are computed and the small circles are measured values.

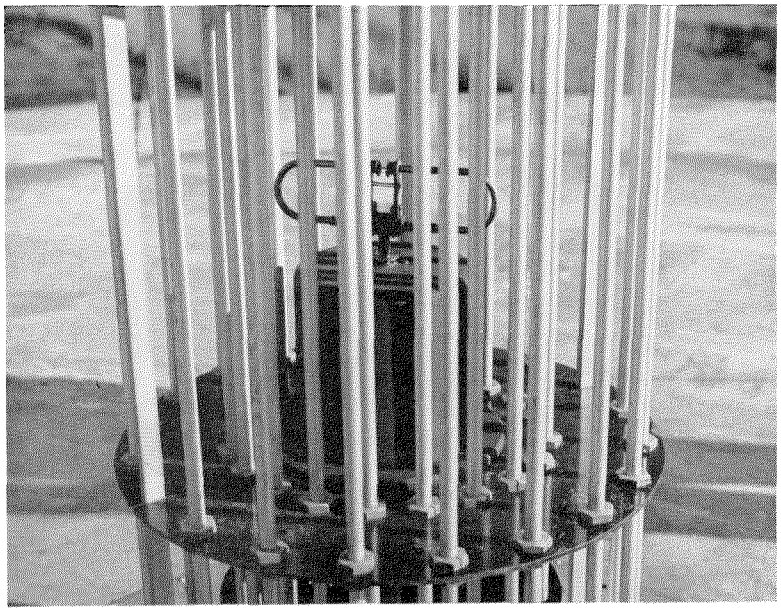


Figure 5—Dipole-rotating unit mounted within the lower cage. The protective cover has been removed.

cage its input impedance is substantially real and 50 ohms; that is, it is substantially the electrical equivalent of a half-wave dipole approximately 4 feet (1.2 meters) long operating at a frequency of 115 megacycles. Thus it constitutes a small efficient dipole which, due to its small physical size, can be turned at 1800 revolutions per minute with a $\frac{1}{16}$th-horsepower synchronous motor located in the dipole-rotating unit.

Figure 6 is a plot of the voltage standing-wave ratio against

frequency for the dipole mounted in the cage and driven from a cable having a 50-ohm characteristic impedance. In the region of 112 to 118 megacycles, the ratio is under 2/1. A folded dipole structure was used only because it provided a convenient way of feeding the dipole.

Since the dipole is small, its directivity pattern in the horizontal plane is a figure-of-eight. Figure 7 shows a plot of the field intensity in the horizontal zero-elevation plane of the dipole in the cage antenna.

A rotating sleeve-type joint with no sliding contacts is located below the motor in the dipole-rotating unit to transfer the radio-frequency energy to the dipole. This rotating joint consists of a capacitive stub formed by the rotating sleeve in series with an inductive stub to make a series-resonant circuit. Low characteristic impedances were used for these stubs to produce a broad-band characteristic for the joint and make tuning adjustments over the 112-to-118-megacycle frequency range unnecessary. The dipole is connected to the top of the rotating sleeve through a 50-ohm cable, which runs through the hollow motor shaft.

2.4 Tone Wheel

Also, on the dipole shaft, as may be seen in Figure 8, is the tone wheel and magnetic pickup that provide the fixed-phase reference signal. This tone wheel generates a spectrum with a center frequency of 9960 cycles frequency modulated with a deviation frequency of ± 480 cycles at a 30-cycle rate. The tone wheel is made from a blank of Swedish iron cut with a changing pitch from tooth to tooth to produce the desired modulation. After assembly, the edge of the tone wheel is ground at a constant radius to insure freedom from any amplitude modulation due to eccentricity. As is well known, this type of reference signal permits a convenient separation in the navigational receiver of the reference phase signal and the amplitude-modulated limaçon or variable-phase signal. The output of the tone wheel pickup is brought out through the base of the dipole-rotating unit, where it connects to a cable from the modulator unit in the transmitter. Since the tone wheel is an integral part of the dipole shaft, there is always a positive fixed relation between the reference- and variable-phase signals. The entire dipole unit consisting of the dipole, tone wheel, motor, and rotating joint is dynamically balanced after assembly.

2.5 Loop Disk

The conducting disk in the inner cage, which was mentioned previously, serves two purposes. First, it acts as an end plate for the inner radial waveguide section, and second it serves to radiate an omnidirectional pattern that, when com-

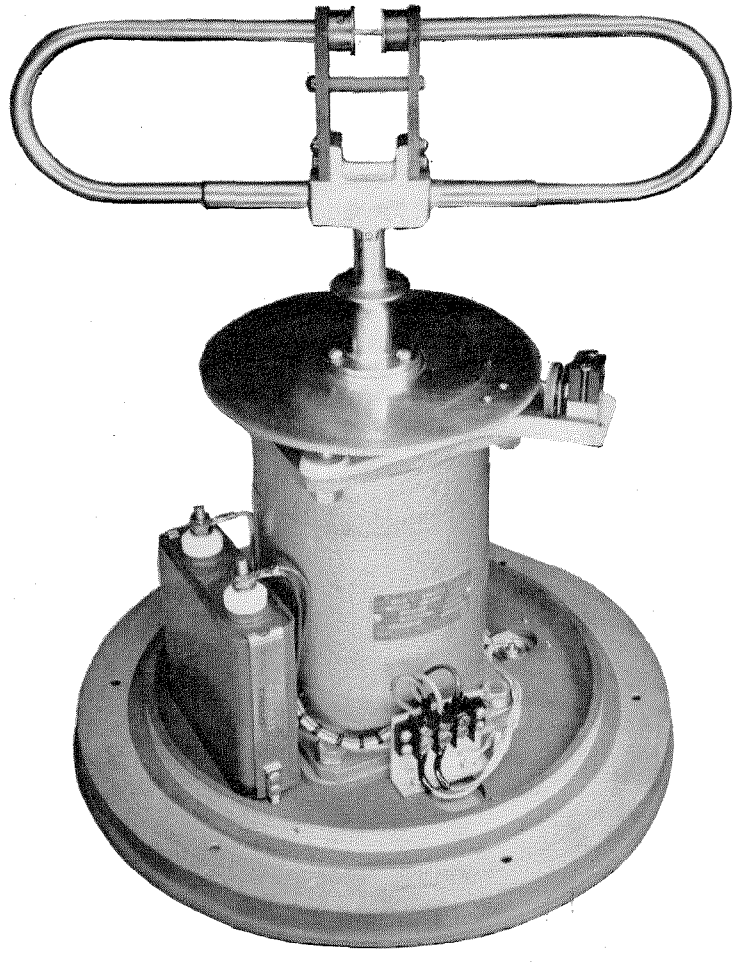


Figure 8—Dipole-rotating unit showing dipole, tone wheel and magnetic pickup, synchronous motor, and the motor-starting capacitor.

bined in the proper radio-frequency phase with the rotating figure-of-eight pattern produced by the rotating dipole, yields the rotating limaçon for the variable-phase signal. To excite the loop disk for the omnidirectional pattern, the disk is provided with 4 orthogonally disposed radial slots. Each of these slots is fed cophaseally to produce the circular loop pattern. Various matching stubs are located on the disk partially to match the loop impedance to a feed line that is brought up to the loop from the bottom of the lower cage section through one of the inner row of rods. A field plot of the loop disk pattern in the horizontal plane at zero elevation angle is given in Figure 9.

As previously mentioned, a coaxial cable running up one of the rods in the inner row supplies the modulated carrier energy to the loop disk. A triple-stub tuner attaches to this coaxial line at the base of the antenna as shown in Figure 10. Also visible in Figure 10 is the base of the dipole-rotating unit with its *RG-9/U* coaxial line and the two cables for the reference signal and the 60-cycle 115-volt motor feed supply. This triple-stub tuner is used to match the *RG-9/U* cable from the transmitter to the loop feed. In operation, a small directional coupler type of standing-wave indicator is used to set the adjustments of the triple-stub tuner. The directional coupler is connected between the triple-stub tuner and the *RG-9/U* feed line and in a few minutes the tuner

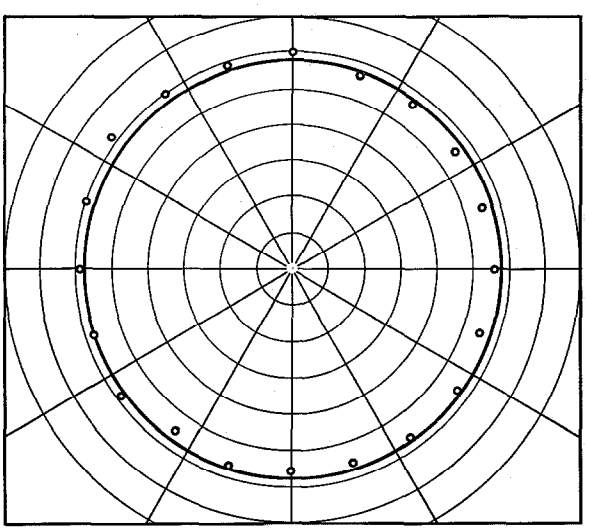


Figure 9—Horizontal directivity pattern of the loop mounted in the cage antenna.

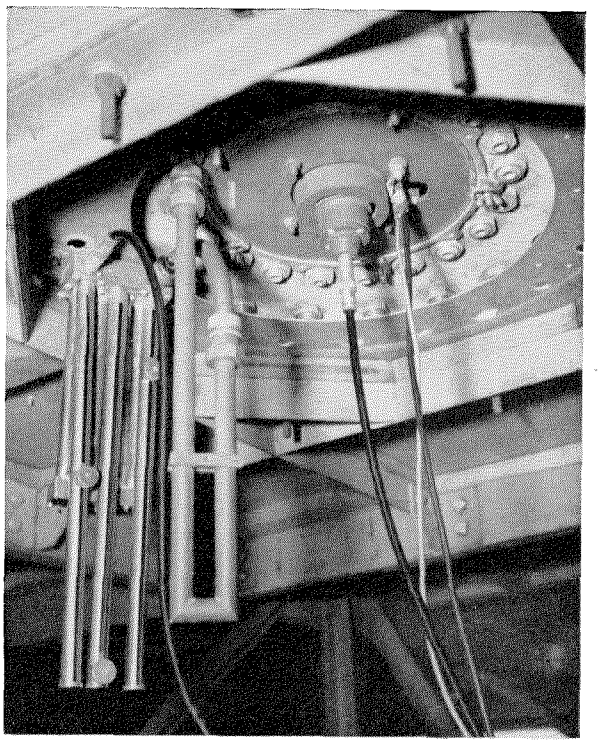


Figure 10—Bottom view of antenna taken from below the counterpoise showing the triple-stub tuner for matching the impedance of the loop disk. The radio-frequency cable from the dipole comes out through the center. The two lines near it are from the tone wheel and the driving motor.

can be adjusted to read zero reflected power, or unity standing-wave ratio. The dipole has broadband characteristics in itself and requires no adjustment over the entire range from 112 to 118 megacycles.

In the first development model tested, it appeared as though the loop disk was not an efficient radiator; however, it was recognized that this shortcoming in the early model should not be attributed to the loop-disk radiator but to the transmission-line harness feeding the loop. This condition was remedied on later models by partially matching the line between the triple-stub tuner and the loop disk with appropriate fixed-length matching stubs on the loop. With this arrangement, the loop has a gain equal to that of a half-wave dipole.

To obtain the proper radio-frequency phase relation between the figure-of-eight pattern and the omnidirectional pattern, an auxiliary unit called a "phaser" is placed in the loop-disk feed line. This phaser unit is located in the transmitter

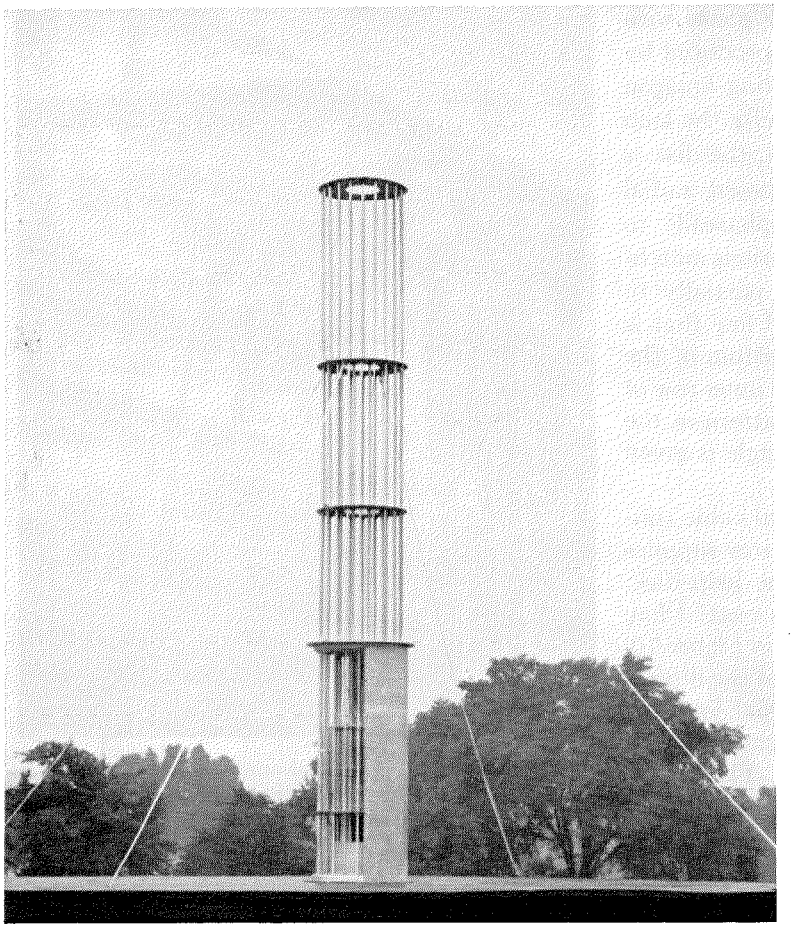


Figure 11—The complete antenna with one of the semi-circular fiber-glass protective housings removed from the lower cage.

house and consists of a trombone-type line stretcher.

2.6 Modulation Eliminator

Another unit essential for stable operation is the "modulation-eliminator" unit. This unit is located in the transmitter house and is bridged across the output of the transmitter feeding the loop disk. The modulation eliminator functions as its name implies. Since the output of the transmitter is modulated up to 70 percent and since the rotating dipole must be supplied with only carrier, the modulation eliminator takes a portion of the modulated carrier from the transmitter output, effectively "clips" off the modulation and delivers only unmodulated carrier to the rotating dipole. This unit assures a constant phase relation between the dipole and the loop disk independent of aging tubes and transmitter tuning.

2.7 Upper Cage

Figure 11 is again the entire antenna showing the upper cage assembly and the lower cage with half the protective housing removed. The function of this upper cage is an extremely important one. A solid conducting cylinder would serve but the cage-like structure, which appears electrically like a solid cylinder, results in reduced weight and windage effects. Without this extension, the amount of vertical polarization radiated from the lower section, when excited by the dipole, would be prohibitive. As is well known, in an aerial navigational system of this sort utilizing horizontal polarization, the addition of any vertical polarization is decidedly detrimental to good operation. Vertical polarization introduces an action referred to as "push." If push exists, then the course observed by the pilot depends on the attitude of the aircraft. A pilot may bank his aircraft so that his instruments indicate that he is on the desired course but on levelling the course appears to shift. He may again bank the

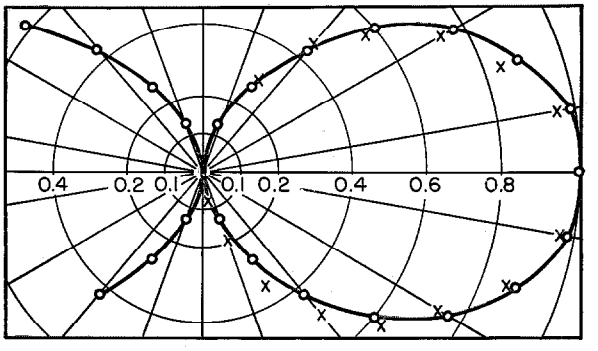


Figure 12—Calculated (circles) and measured (crosses) values of radiation pattern in the vertical plane of the dipole in free space.

aircraft in an attempt to get on course but will find the course has again shifted, and thus his flight will be a series of zig-zags. This push is a result of the vertical polarization that is present

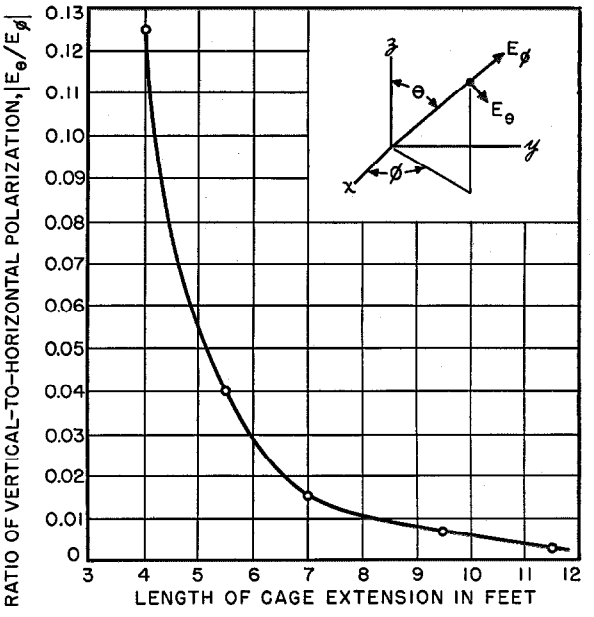


Figure 13—Ratio of vertical polarization to horizontal polorization $|E_\theta/E_\phi|$ as a function of the length of the cage extension.

being a maximum where the horizontal field is a minimum, and vice versa. This effectively produces a figure-of-eight pattern for the vertical polarization that is in space quadrature with the horizontally polarized figure-of-eight pattern. If the omnidirectional-range antenna radiates a pure horizontally polarized field then the course indicated on the cross-pointer instrument in the aircraft will not shift with changes in aircraft attitude.

The upper cage extension makes the cage antenna a source of substantially pure horizontal polarization. Figure 12 is a plot of the calculated and measured vertical-field distribution of the dipole in free space. The null is evidence of the lack of vertical polarization.

Figure 13 shows a plot of actual measured ratio of vertical-to-horizontal polarization versus length of the upper cage extension. If the upper cage were to become infinite in length, this ratio would approach zero. In practice, it is found that an extension of 12 feet yields a satisfactory reduction of vertical polarization. For this length, the vertical polarization is 50 decibels below the horizontal polarization. This ratio of vertical-to-horizontal polarization will yield a calculated amount of push less than half a degree (actually 0.2 degree) on a 45-degree bank of the aircraft.

2.8 Distance-Measuring Equipment

The upper cage extension can also be utilized to house a distance-measuring-equipment (DME) antenna if so desired. With distance-measuring facilities, the omnirange would, of course, provide for a complete rho-theta navigational system.

The vertically polarized distance-measuring antenna operates in the 1000-megacycle region, and to allow it to radiate through the upper cage when placed within it an inner row of rods has been added in the upper section to match the distance-measuring antenna to free space. The field patterns of the distance-measuring antenna are not substantially changed when it is placed within the upper cage.

3. Tests

Figure 14 is a calibration-error curve taken in flight with the aid of a theodolite; the receiver error has been removed. This curve was taken at

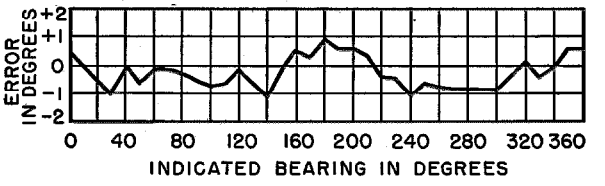


Figure 14—Calibration error measured in flight.

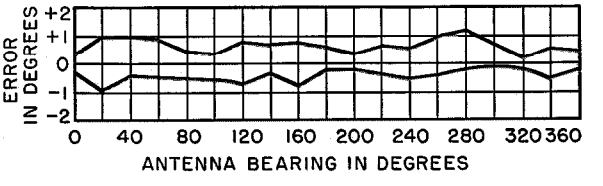


Figure 15—Polarization-error curve.

a radius of 10 miles at an altitude of 2000 feet and shows the departure of indicated azimuth to true azimuth. This azimuth calibration shows a maximum error of ±1 degree.

3.1 Polarization

Figure 15 gives a polarization-error curve of the antenna. This curve was obtained on the ground by using vertical and horizontal receiving dipoles placed a distance of 800 feet away from the omnirange antenna. The course was determined on the horizontal dipole and then the shift in course resulting from paralleling the vertical and horizontal dipoles is the polarization error. The two curves result from reversing the polarity of the vertical dipole by 180 degrees. The polarization error versus azimuth was obtained by rotating the antenna on the counterpoise. This is a severe test as it seems to yield greater errors than appear when the aircraft is banked. The maximum polarization error from this curve is ± 1 degree. Actual flight checks have shown no observable push effects on the airborne cross-pointer instrument during 45-degree banks of the aircraft.

3.2 Cone of Silence

This antenna gives a very narrow cone-of-silence over the station. Above the station, the courses all converge to a point and the direct fields from the omnidirectional and figure-of-eight patterns vanish. If the two patterns do not track substantially, for example if one pattern has a high-angle minor lobe in its vertical pattern, then the course may be destroyed even before the aircraft arrives over the station. The cross-pointer instrument will be erratic and will not settle down until the aircraft has passed over the station and reached a corresponding point on the other side. With this antenna, it is possible with care to fly directly over the antenna and not have the course shift.

While flying over the antenna, however, the TO–FROM indicator in the receiver will change from a TO indication to a FROM indication. In Figure 16, are actual recordings of some flights made over the cone of silence above this antenna. These recordings indicated the existence of a cone with an angle of approximately 6 degrees about the vertical. In this cone, a course does not exist or was too sharp for the pilot to follow. A pilot well experienced in instrument flying can fly over the cone without receiving an indication of a course deviation. This narrow cone with its clearly defined excursion of the TO–FROM meter serves as an excellent marker to tell the pilot that he is over the station.

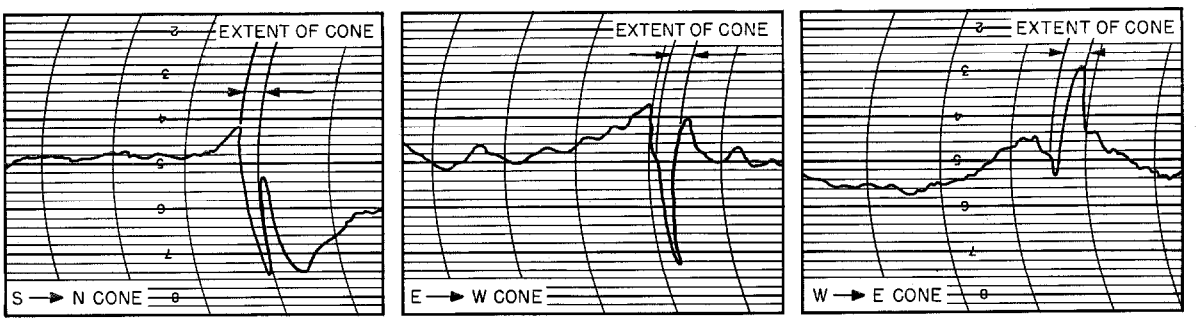


Figure 16—Recordings in flights through the cone of silence directly above the antenna. Flight was at a ground speed of 95 miles (153 kilometers) per hour at an altitude of 3000 feet (914 meters). Each vertical division of the chart is approximately 1.5 degrees.

4. Summary

A cage-type very-high-frequency phase-comparison omnidirectional radio range antenna has been described that has good azimuth accuracy, freedom from vertical polarization, a small cone of silence, and requires a minimum of installation time.

5. Acknowledgement

The authors gratefully acknowledge the support and guidance given by Mr. L. A. deRosa during the development of this antenna.

Thanks are also due Mr. R. J. Fahnstock, who participated in much of the work and flight testing associated with the development.

Voice-Frequency Signaling Equipment

By W. A. BRANDT and JAMES POLYZOU

Federal Telephone and Radio Corporation; Clifton, New Jersey

SIMPLICITY, with concomitant reliability and inexpensiveness, has been stressed in the design of an equipment that will convert the direct-current pulses from a telephone dial into 1600-cycle-per-second pulses suitable for application to the modulator of a carrier-frequency telephone system. False operation of the receiver by interfering signals in the 1600-cycle range is prevented while speech transmission is not deteriorated by elimination of all frequencies in the vicinity of 1600 cycles as has often been the case in previous designs of similar equipment. These advantages are obtained by the use of an ingenious relay switching arrangement that inserts 1600-cycle band-suppression filters in the transmission path only during signaling.

• • •

1. General

The ever-increasing use of automatic switching over carrier-telephone trunk lines emphasizes the role of dialing facilities in these systems. In many of the existing equipments, no provisions have been made to allocate a separate portion of the frequency spectrum for signaling purposes; hence, frequencies within the speech band must be used.

When utilizing a portion of the speech band for signaling purposes, precautionary measures must be taken to prevent the normal speech frequencies from simulating dial or supervisory signals.[1] This danger can be eliminated by the use of expensive multifrequency signaling devices. However, in many cases and particularly in those involving short trunk lines and networks of lesser density, their use is not economically justified.

Single-frequency systems employing band-elimination filters for the removal of the portion of the speech band that is used for signaling purposes are not always acceptable for commercial lines since the reduction in transmission band results in objectionable deterioration of the speech quality.

The signaling equipment to be described is of the single-frequency type, but the objectionable features of band-elimination filters are overcome by the automatic removal of these filters on completion of the necessary signaling operations. The unit has been designed to provide a simple means for transmitting dial pulses and performing the necessary seizures and supervisory functions while retaining immunity to speech interference. It interconnects with standard dialing repeaters designed for carrier or duplex composite dialing operation without any additional equipment. Automatic switching systems that were not designed for this type of service can be adapted for use with it by the addition of an applique consisting of 4 to 7 relays, depending on the switching system. The *904A* signaling equipment shown

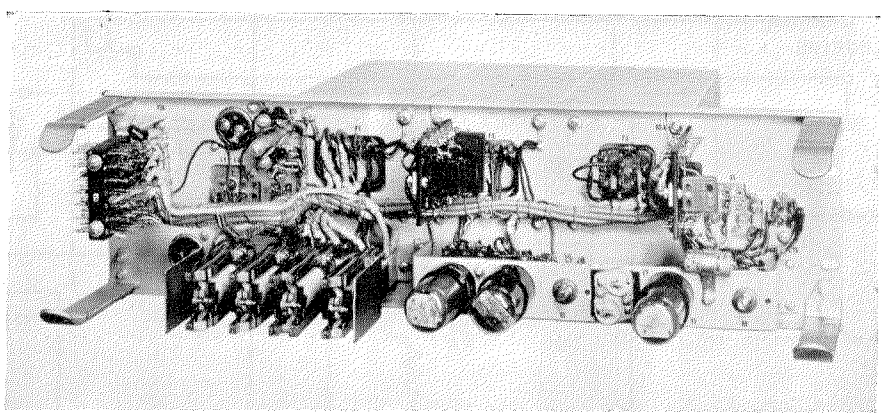


Figure 1—Front view of the *904A* signaling equipment with dust cover removed. The unit mounts on a standard rack.

[1] T. H. Flowers and D. A. Weir, "Influence of Signal Imitation on Reception of Voice-Frequency Signals," *Electrical Communication*, v. 26, pp. 319–337; December, 1949: also *Proceedings of the Institution of Electrical Engineers*, v. 96, Part 3, pp. 223–235; May, 1949.

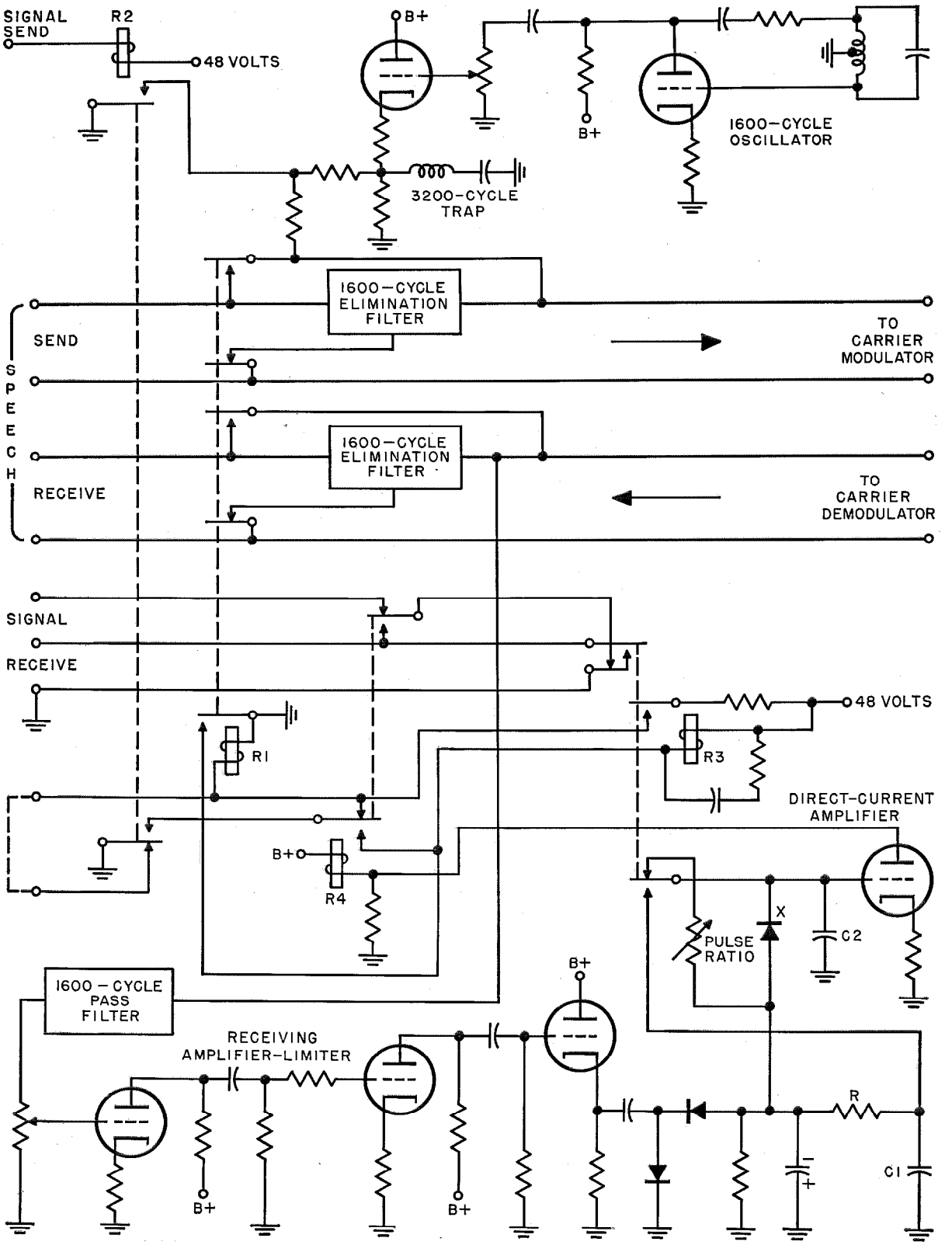


Figure 2—Schematic diagram of the signaling unit in idle condition.

in Figure 1 is designed for 4-wire operation and consists of a transmitter, receiver, and a pair of band-elimination filters, all designed for operation at 1600 cycles, together with the necessary relays to perform the dialing, signaling, and protecting functions.

2. *Circuit Description*

With the circuit in the idle condition as shown in the schematic diagram of Figure 2, the two band-elimination filters are connected before the modulator (or transmitter) and after the demodulator (or receiver) of the carrier or other 4-wire circuit. The 1600-cycle signaling frequency is injected on the carrier-equipment side of the band-elimination filter in the transmitting path, which filter provides protection against signal imitation by incoming speech frequencies. In idle condition, the 1600-cycle signal is transmitted in both directions. The received signals de-energize the relay *R4*.

Seizure of the circuit preliminary to dialing, causes the transmitting relay *R2* to interrupt the 1600-cycle tone in the outgoing direction thereby operating the receiving relay *R4* at the far end of the carrier system. The 1600-cycle signal from the far end still persists, and dial tone, if employed, may be heard from the distant office.

Dialing is now accomplished by relay *R2*, which reapplies the 1600-cycle tone as each dial pulse de-energizes the relay. This pulse train is faithfully reproduced at the contacts of *R4* in the far-terminal equipment, which contacts are connected to the associated automatic-telephone-exchange equipment. On completion of dialing, the calling party may now hear busy signals or ringing signals, and an intercepting operator can speak to the calling party.

When the called station answers, its central-office equipment causes the operation of the far-end relay *R2*, which interrupts the 1600-cycle signal from that terminal. At the near end, relay *R4* repeats the supervisory signal to the associated automatic switching equipment. The absence of 1600-cycle signals from both directions causes the band-elimination filters to be removed at both terminals of the circuit. Simultaneously, a time-delay circuit is inserted in the signal receiving path to preserve speech immunity. Since the band-elimination filters are now removed from the speech circuit, undeteriorated transmission may occur.

When the call has been completed and either end of the circuit is released, 1600-cycle tone is again transmitted from that end to the other terminal. When the second office releases, the equipment is restored to idle condition with 1600-cycle tone transmitted in both directions.

During this idle condition, seizure, and dialing, the circuit is protected against false operation by the two band-elimination filters. Since these filters suppress only a narrow band of frequencies near 1600 cycles, the speech circuit is able to transmit speech and audible dial, busy, or ring-back tones at any time. The receiving circuit, by virtue of its narrow-band-pass filter, is immune to all frequencies except those near 1600 cycles. When the connection is completed, the band-elimination filters are removed from the circuit by relay *R1*, and some other means must be employed to immunize the receiving circuit from speech interference. During this period, both the transmitting relay *R2* and the receiving relay *R4* are energized. When this condition exists, the contacts of the two relays complete a circuit that energizes the slow-acting relay *R3*. The contacts of this relay apply current to another relay *R1* that then removes the two band-elimination filters from the circuit. The contacts of *R1* simultaneously apply a holding voltage to the coil of *R3*. The operation of *R3* also causes a time-delay network *R,C1* to be inserted in the receiving circuit. Receiving relay *R4* under these conditions can be de-energized only by random 1600-cycle signals that persist longer than 0.3 second.

To prevent integration of successive impulses, the delay circuit consists of a capacitor *C1* that charges slowly through the backward resistance of a rectifier *X* but discharges immediately through the forward resistance of *X* once the impulse subsides. If a tone should persist long enough to release the receiving relay *R4*, it will apply a short-circuit across the coil of *R1* causing it to release and momentarily reapply the band-elimination filters. This will immediately remove any interfering impulse from the near end, where it is more likely to exist due to the higher relative level. Relay *R4* will reoperate when the interference is removed, and *R1* will once again remove the filters.

An examination of Figure 2 will reveal that when *R4* released, the supervisory signal was not sent out on the signal-receive lead since *R3* now controls this lead. Although *R1* acted to release *R3*, the release could not be effected since the time constant formed by the resistance and capacitance associated with the coil of *R3* would not permit the relay to de-energize unless *R1* remained inoperative for at least 0.5 second. This is not possible since *R1* will immediately remove the interference that is causing its release by reinserting the band-elimination filters. Under these conditions, the application of a continuous 1600-cycle tone at the near end will only cause a slowly recurrent insertion and removal of the filters without permitting a signal to be sent out on the signal-receive lead.

3. Electronic Circuit

3.1 Transmitting Path

The transmitting path consists of a resistance-stabilized oscillator and a cathode-follower whose output is bridged to the input of the carrier-equipment modulator. The signal level introduced at the modulator is equivalent to 10 decibels below 1 milliwatt in 600 ohms. Lower levels may be obtained by adjustment of the gain control that is provided at the input of the cathode-follower. The bridging loss presented to speech frequencies is approximately 0.25 decibel. When necessary, the output may be increased to 0 decibels by reducing the bridging impedance through a simple strapping arrangement. The bridging loss to speech frequencies under these conditions is approximately 1 decibel.

To avoid the generation of spurious frequencies, the signal keying takes place at the output of the cathode-follower. A 3200-cycle trap circuit is provided at this point to eliminate the first even harmonic of the signal frequency, which is capable of causing interference to some types of carrier equipment.

3.2 Receiving Path

The receiving circuit of the signaling equipment contains a band-pass filter that bridges the output of the carrier-equipment demodulator. The bridging loss is negligible at all frequencies below 1500 cycles and above 1700 cycles, the loss at 1600 cycles being 2.5 decibels. The output of the band-pass filter is applied to a stepped gain control that feeds the signal to a two-stage limiting amplifier. The amplified signal is applied to a cathode-follower, whose output is rectified in a half-wave voltage-doubler to produce the negative voltage indicated. This voltage serves to cut off the direct-current amplifier whose plate current controls relay *R4*. A resistor in parallel with this amplifier permits a small current to flow through the relay coil while the tube is at cut-off. This trickle current causes a delay that tends to shorten the pulses on the relay contacts.

The rectifier–filter circuit, however, includes a capacitor *C2* that is charged through the forward resistance of the rectifier *X* but discharges through the variable pulse-ratio resistor, thereby lengthening the duration of the negative pulse. The pulse shortening of the direct-current amplifier combined with the pulse lengthening of the filter permits the length of the initial pulse to be varied by 10 percent simply by adjusting a variable resistor. The pulse-ratio control is a feature of considerable value in dialing service since the permissible tolerance in pulse distortion that must be maintained to avoid erratic dialing is to a large extent used up within existing dialing systems. The distortion produced by the signaling unit must be very low. The pulse-ratio control will not only correct the local distortion but will also improve the operation of the external direct-current dialing circuits.

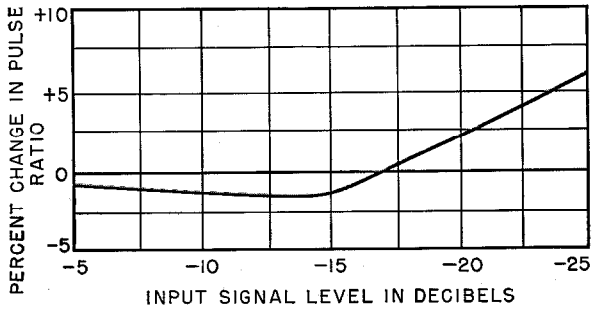


Figure 3—Change in output pulse ratio plotted against the signal input level. Zero decibels refer to 1 milliwatt in 600 ohms.

As indicated in Figure 3, variations in the line loss have little effect on the pulse-ratio output. The amplifying and limiting functions of the signal-receiving circuit preserve the pulse ratio of the transmitting office over the wide range of received signal levels.

In the explanation of the relay circuits, mention was made of a time-delay circuit that is inserted into the signal receiver immediately before the automatic removal of the band-elimination filters. This time delay is formed by removing the pulse-ratio control and substituting a large fixed resistor *R* and capacitor *C1* in the pulse-shaping circuit, thereby delaying the buildup of the negative voltage at the grid of the direct-current amplifier, while still permitting a quick discharge of the capacitor. When the call has been completed, the pulse-ratio-control circuit is restored to its normal condition.

4. Power Supply

Power for the electronic, relay, and filament circuits must be supplied from an external source. Plate power of 130 volts direct current and heater power at 24 volts alternating or direct current, may be obtained from the office batteries provided in carrier and repeater offices, or from any other power supply designed to deliver these voltages, while current for relay operation must be obtained from a 48-volt direct-current source.

Among the features worthy of mention are the use of types *6SN7* and *6SL7* tubes, which may be interchanged with types *5692* and *5691* long-life tubes, the convenient arrangement of the equipment for maintenance purposes, and the accessibility to the wiring and terminals of the large sealed components. Removal of the front dust cover exposes the active components such as tubes, relays, controls, and adjustments. Notches are provided for mounting the equipment on a standard 19-inch (48-centimeter) rack. Vertical space requirements are $5\frac{1}{2}$ inches (14 centimeters). The equipment weight is approximately 21 pounds (9.5 kilograms).

5. Additional Applications

Because of its ability to provide speech-immune signaling channels over any carrier channel or 4-wire facility, the uses of the *904A* signaling equipment are not confined to dialing circuits alone. A number of different optional arrangements are available having slight modifications in the delay circuits to provide for such uses as transmitting railway selector-code impulses at $3\frac{1}{2}$- or 5-cycle rate with 20-cycle or magneto ringing in the opposite direction. The equipment functions in all of these applications in basically the same manner, taking into account the nature of the service. Thus, in operation over a railroad-dispatching circuit that is extended by the use of a carrier channel, the band-elimination filters are normally out of the circuit. When the selector key starts to send a code, the filters are cut in slightly before the actual code starts, and held in until after the entire sequence, including ringing, has been completed. The necessary holding circuits are included as part of the panel, so that no auxiliary adapter circuits are needed when the dispatcher is at the carrier terminal and controls a selector-apparatus case at the other terminal.

Airborne Receivers and Test Gear for Instrument Landing Systems

By F. G. OVERBURY

Standard Telephones and Cables, Limited; London, England

WORK that had been done towards improving the ground side of instrument landing systems and the incorporation of new ideas in the *ILS-2* ground equipment were discussed in a paper[1] in this journal. It is proposed to consider now those problems that are encountered by the designer of airborne installations with particular attention to the development of the *SR.14/15* receivers. Test gear is also mentioned as it plays an equal part with the receivers in obtaining the stability of indication that is an absolute essential to the safe working of the landing aid.

1. Problem

The receivers are the airborne portion of an approach aid that in restricted weather conditions should enable the pilot with complete safety to bring his aircraft to within visual distance of the touch-down point.

When making an instrument approach, information is conveyed to the pilot by means of the cross-pointer indicator shown in Figure 1. The vertical pointer is actuated by the localiser receiver and the horizontal needle by the glide-slope receiver; deflection is such that the pilot is informed in which direction he must turn to correct his approach, rather than of his actual position relative to the descent path. He must, in effect, always correct towards the intersection of the needles, imagining himself to be at the centre of the ring on the indicator. Due to the inertia of an aircraft and the undesirability of making continuous jerky corrections, the pilot does not manage to maintain the needles intersecting over the centre of the circle during the

[1] R. A. Hampshire and B. V. Thompson, "*ILS-2* Instrument Landing Equipment," *Electrical Communication*, v. 27, pp. 112–122; June, 1950.

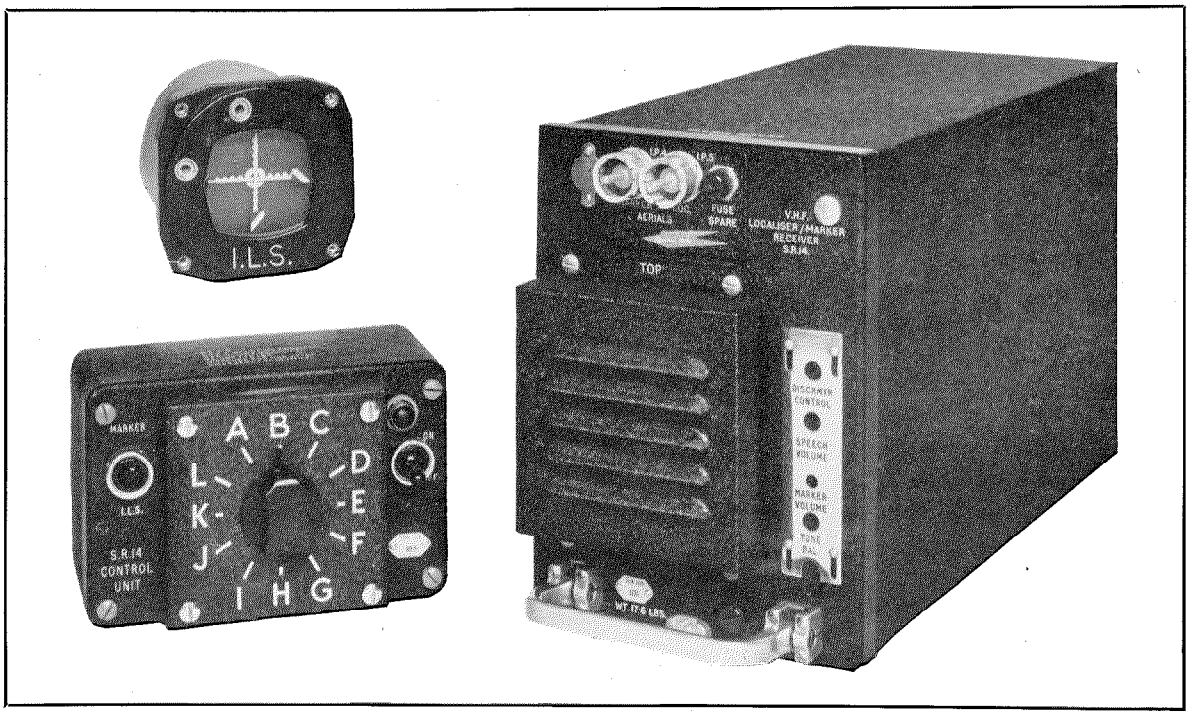


Figure 1—The receivers are in the large case, with the control unit at lower left and cross-pointer indicator above it.

whole approach; if he is to be able to develop a procedure for landing in limited visibility and also to have confidence in this procedure, he must be sure of three things.

A. That when the indicator pointers cross over at the centre of the meter, he has intersected the line of a perfect descent path.

B. That if the indicator shows a finite deflection, he must be confident in the knowledge of how much correction must be applied to the aircraft to avoid violent last-minute manipulation.

C. That if the deflection of the indicator is less than a specified amount (usually full-scale deflection), he is in no danger from ground obstruction. This has a special point with regard to terrain clearance in the case of the glide slope.

From the above points, emerge the two basic requirements of an ideal airborne receiver.

A. When on a predetermined line in space, considered to be the best descent path, the airborne indicator shall always give an "on-course" indication, that is, the needles shall intersect over the centre of the circle.

B. If the aircraft is at a specific angular distance from this descent path, the indicator shall, under all conditions, give the same deflection; thus ensuring that the pilot is aware of the correction he must make and in the case of large deflections of the proximity of obstructions.

It is toward the achievement of these requirements that all development work is directed.

2. *Sources of Error*

The previous article[1] described how each transmitter of the ground system radiates a pair of balloon-shaped lobes and how the intersection of these lobes, corresponding to equal depth of modulation of 90- and 150-cycle-per-second tones, lies along the centre of the course. At any specific angular deviation from the centre, there exists a fixed difference in depth of modulation of the tones.

The function of the airborne receiver is to measure and produce an indicator deflection that is proportional to the difference in the depth of modulation. The deflection is zero on course when this difference is zero and has direction and amplitude determined by which tone is in excess and the angular deviation of the aircraft from course. The steps by which the receiver derives this information from the respective localiser and glide-slope carriers are, firstly, signal amplification, then detection to produce the tones, followed by tone separation in filters, rectification producing direct voltages proportional to the original 90- and 150-cycle modulation depths, and finally the measurement of their difference by feeding them in opposition to the respective cross-pointer movement.

There are two main categories of errors that must be kept small in the design of this airborne measuring instrument; they are centre errors and sensitivity errors, and may be considered separately.

3. *Centre Errors*

Centre errors tend to give an "off-course" indication when equal 90- and 150-cycle modulation is being received. From the pilot's point of view, they represent a bodily shift of the whole approach path. The error is caused by factors that affect the 90-cycle tone in a different way from the 150-cycle tone, and so are due mainly to components in the audio circuits. Harmonic distortion can contribute to unbalance especially with regard to the 90-cycle tone, the 180-cycle second harmonic of which will probably affect the 150-cycle filter. The main sources of trouble are filters and rectifiers, where, of course, each tone is being dealt with separately prior to being compared as direct-current values in the indicator. When it is remembered that a 3·4-decibel ratio in tones gives a full deflection on the glide slope, it will be realised that filter design has to be very carefully approached to accommodate the 2·5-per-cent tolerance allowed by the International Civil Aviation Organization on ground-transmitter tones. Rectifiers must also remain very accurately balanced especially under extremes of temperature and humidity. It is not proposed to discuss these errors any further as the techniques employed for their reduction are conventional instrument and filter methods.

4. *Sensitivity Errors*

The needle deflection, as explained above, is proportional to the difference in amplitude of 90- and 150-cycle tones that are applied to the rectifiers. If this deflection is always to be proportional to the difference in the modulation

depth applied at the aerial, it is necessary that the tone output from the filters should be independent of radio-frequency input, battery voltage, and valve condition.

4.1 Methods of Reducing Sensitivity Errors

An airborne receiver must operate under changes of battery supply from 22 volts to 29 volts and if the output is to be independent of supply, it is necessary for the receiver to carry some reference standard that is independent of external conditions.

4.2 Neon Standard

Neon stabilisers are included in the *SR.14/15*, and the automatic-volume-control delay is derived from them. As the signal at the aerial is increased, the receiver output (which is proportional to deflection sensitivity) is allowed to rise to a level determined by the neon standard, further increase will develop an automatic-volume-control voltage that controls the gain of the receiver. Though a definite level has been established independent of external conditions, it is obvious that a finite increase of signal at the automatic-volume-control detector above the point at which automatic volume control is developed will be necessary to produce the voltage required for controlling the receiver; this would imply a certain increase in deflection sensitivity. There are two methods of reducing the effect.

Some of the control voltage may be applied to a stage following the point at which the automatic volume control is derived, thus reducing the gain of this part of the receiver by an amount that corresponds to the unavoidable increase at the automatic-volume-control detector. Adjustment of the amount of control applied can be used to produce any shape of control characteristics desired. The expression "course softening" is applied to the process whereby the output or deflection sensitivity of the receiver is made to decrease with input signal. Too much control of this type may produce unpredictable results depending on valve conditions and supply voltages.

Alternatively, the automatic volume control may be derived from a very-high-voltage point and a correspondingly high delay voltage used.

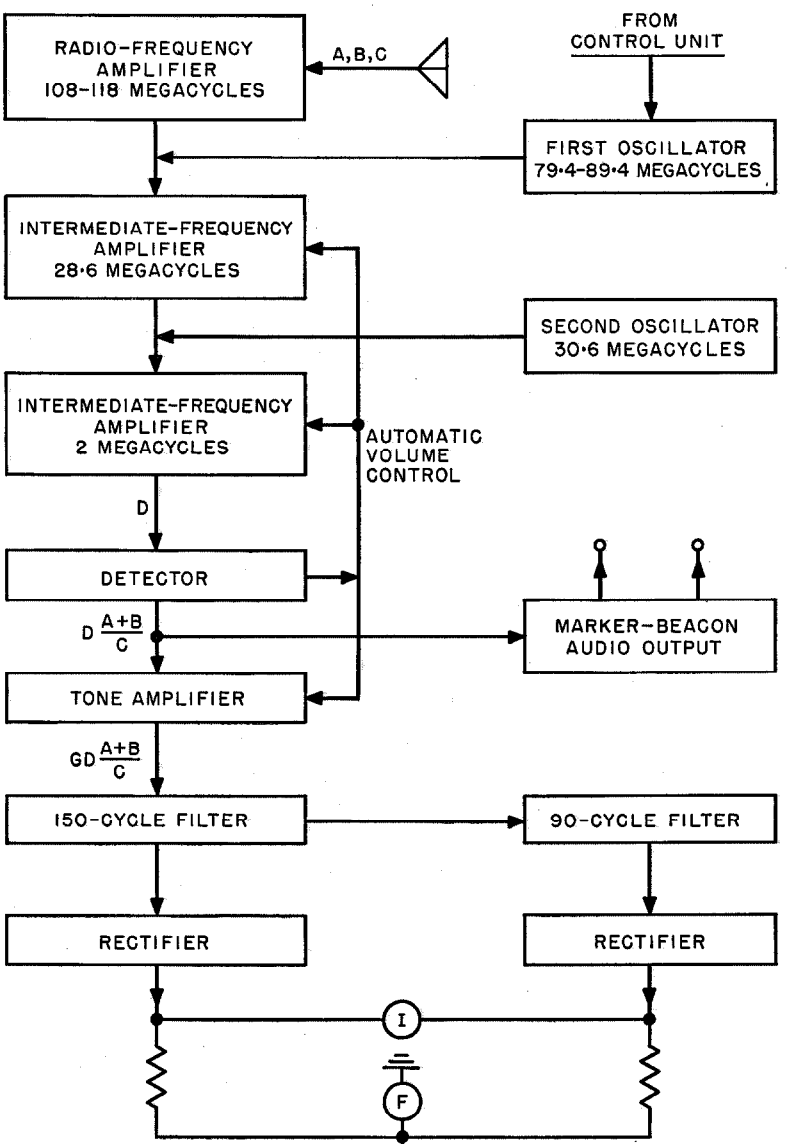


Figure 2—Schematic arrangement of the *SR.14* localiser receiver using carrier-derived automatic volume control.

In this case, the necessary number of volts required to control the receiver may be obtained with a much smaller percentage increase at the controlled point.

On the localiser and glide-slope receivers of the *SR.14/15*, use is made of both these methods. Reference to Figure 2 will show that on the localiser receiver, the automatic volume control is derived at the detector, delay being determined by a neon reference tube. The control voltage is applied to one audio amplifier as well as to the radio-frequency stages, the amount of control is adjusted such that the gain of the audio stage is reduced by about 4 decibels as the input signal is increased between 10 microvolts and 100 millivolts. This corresponds to the unavoidable increase of tone output from the detector between the above limits of input signal. On some receivers, the change of deflection sensitivity is limited to 0·25 decibel over the range; on none does it vary by more than 1 decibel.

In this receiver, the voltage at the detector is referred to the neon reference and stabilised, but changes of gain in subsequent audio stages will affect deflection sensitivity as they are not controlled by the automatic volume control. Such changes are unavoidable due to the effects of battery supply and valve aging.

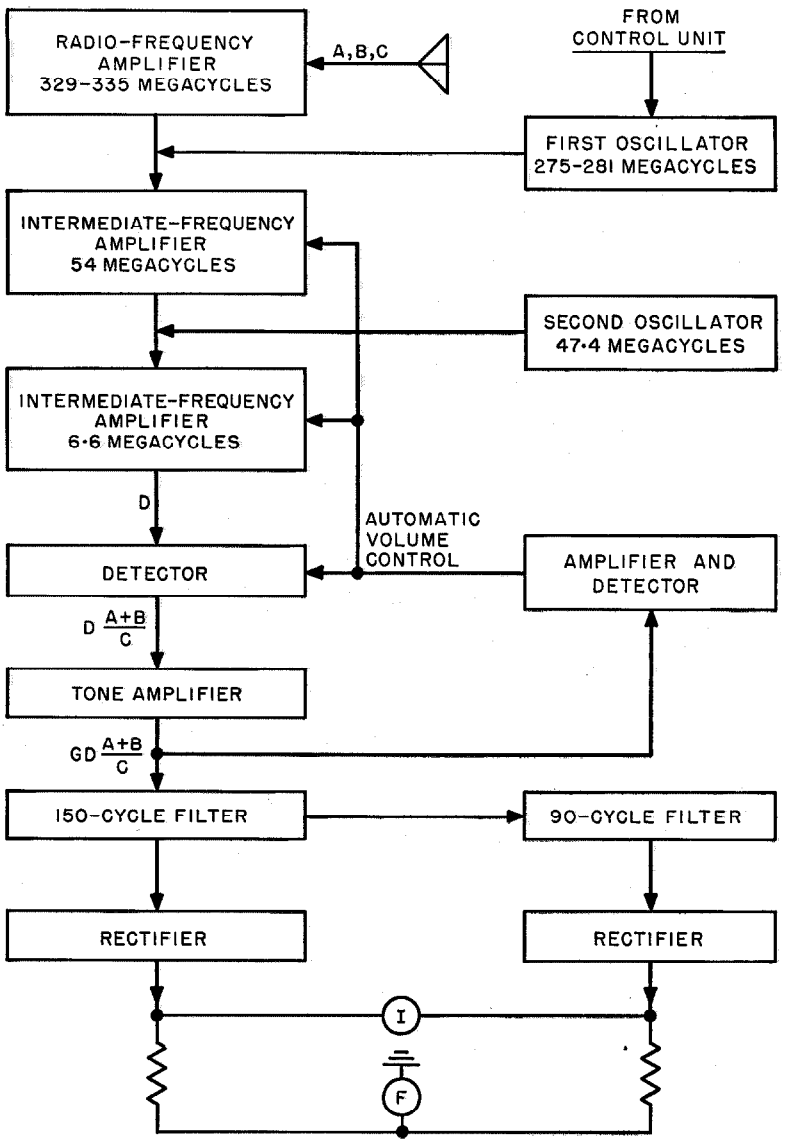


Figure 3—Schematic arrangement of the *SR.15* glide-slope receiver using audio-type automatic volume control.

This effect is reduced to a minimum in the localiser receiver by applying to the audio stages the maximum amount of negative feedback that is consistent with gain requirements. A change of battery supply from 22 to 29 volts causes about 1·8-decibel change in deflection sensitivity. Even so, the total variation due to range of supply and input signals encountered is less than the 3 decibels allowed by the International Civil Aviation Organization.

Further deterioration due to a combination of excessive valve aging and low battery supply will eventually cause the gain to drop to such a point that the alarm flag will show thus indicating the necessity of discontinuing the approach.

On the glide-slope receiver, a different method, which is not applicable to the localiser receiver, is used. This method ensures a greater overall stability and should reduce the number of equipment failures due to loss of audio gain.

Audio-Derived Automatic Volume Control

Figure 3 shows that in the glide-slope receiver the 90- and 150-cycle filters are in series in the anode of the tone amplifier, which is the last valve. At this anode, the complex sum of the two tones is developed and has a peak amplitude between 40 and 50 volts working into a filter impedance of about 7000 ohms. Automatic volume control is developed by the rectification of this voltage. Since the point of rectification is after the last valve in the receiver, gain variations anywhere in the receiver, including the audio stages, will be compensated. Further, since there are 40 volts available at low impedance, it is possible to use voltage-multiplying circuits and a high delay ensuring consequent heavy degeneration of changes in input signal and other parameters. A delay of 120 volts is used on the glide-slope receiver. It is true that there is slightly more increase in receiver sensitivity with input signal (about 1·5 decibels total variation) as it is not possible to apply the post-detector control used in the localiser receiver, but the freedom from random changes, which are especially dangerous, more than compensates for this loss. It is expected that the number of approaches that are discontinued due to receiver deterioration will be much reduced. There are many other implications in the use of audio-derived automatic volume control and these will be discussed at length.

6. *Implication of Audio-Derived Automatic Volume Control*

In analysing the quantity that is measured by two receivers, one using carrier-derived automatic volume control and one using the audio type, reference is made to Figures 2 and 3 and to the following symbols.

A = Field strength at the aerial terminal due to 90- or 150-cycle sidebands* (taken for convenience as the value of the stronger).

B = Field strength at the aerial due to the other sidebands.

C = Field strength at the aerial due to the carrier.

D = Value of carrier voltage at demodulator.

G = Gain of audio stages to either 90 or 150 cycles. (The gain of these stages is adjusted to be the same to both frequencies.)

* More precisely, side frequencies.

For the purpose of establishing the principle, consider a receiver having perfect automatic volume control; that is, one in which control of the radio-frequency gain holds the voltage from which the control is derived to a constant value. The audio voltage produced at the detector due to each tone is the product of modulation depth and carrier voltage, DA/C and DB/C.

Therefore, the components of each tone at the input to the filters will be $DG(A/C)$ and $DG(B/C)$. As the indicator deflection is proportional to the difference between the outputs of the filters, the indicator deflection current will be $kDG\{(A-B)/C\}$.

6.1 Carrier-Derived Automatic Volume Control

In the case where the automatic volume control is derived from the carrier at the detector, D is held at a predetermined and constant value. The indicator current is now proportional to $G\{(A-B)/C\}$ or, *if the gain of the audio stages is assumed to be constant*, it is proportional to $(A-B)/C$, which is the difference in the depth of modulation of the 90- and 150-cycle tones.

6.2 Audio-Derived Automatic Volume Control

If the automatic volume control is derived by the rectification of the combined signal at the anode of the tone amplifier, the input to the two filters is kept constant. In this case

$$GD\frac{A+B}{C}=K.$$

From which

$$D=\frac{KC}{G(A+B)}.$$

But as the deflection current

$$=kDG\frac{A-B}{C},$$

this is proportional to

$$G\frac{A-B}{C}\times\frac{C}{G(A+B)}$$

which is $(A-B)/(A+B)$. This appears to imply that the indication given by a receiver of this

type is entirely different from that of the standard carrier-derived receiver, which simply measures the difference in the depth of modulation of the two waves.

The difference of indication presented by the two receivers is illustrated by

$$\frac{A-B}{A+B}=\frac{A-B}{C}\times\frac{C}{A+B}.$$

$(A-B)/(A+B)$ is the indication given by a receiver of the audio type, $(A-B)/C$ that given by a carrier-derived type, and $C/(A+B)$ is the inverse of overall modulation depth.

The equation shows, therefore, that the indication due to one system differs from that given by the other by a factor that is the overall modulation depth. If the two systems are adjusted to give an equal deflection under specified conditions say at the edge of the glide-slope course, they will always give entirely the same indication provided the overall modulation depth remains constant; if this factor does vary, so will the indication of the systems differ and by a proportional amount. This is obvious as explained by the fact that the carrier level C does not appear in the equation for indicator deflection. Provided there is enough carrier to maintain sidebands, i.e., that the equation for the production of audio at the second detector holds, then indicator deflection is independent of modulation depth. When the automatic volume control is carrier derived, indication is proportional to depth of modulation as the audio produced varies directly with it.

7. *Significance of Indication That is Independent of Modulation Depth*

The International Civil Aviation Organization has specified that the glide-slope modulation depth must fall between 90 and 100 per cent; the indication presented by the *SR.15* glide-slope receiver that will not be conscious of this change may therefore differ by ±5 per cent from that given by a perfect receiver of the carrier-derived type that is indicating difference in depth of modulation. This small departure from the "standard" presentation is easily outweighed by the increased stability achieved.

The system is, however, not at present applicable to the localiser receiver, as the modulation depth of the ground transmitter may be used as a fine adjustment of course width and is permitted to vary between 32 and 44 per cent. To use a receiver of the audio type on the localiser facility would nullify the effect of this adjustment. A further disadvantage is that over 50 degrees off course there is a low-clearance point where the value of $(A-B)/(A+B)$ drops considerably. The quantity $(A-B)/C$ (which is proportional to the deflection given by a receiver with carrier-derived automatic volume control) does not drop to such an extent however, implying a decrease of 90- and 150-cycle sideband ratio accompanied by an increase of modulation depth. A receiver of the audio type would tend to give less deflection and might confuse the pilot into believing that he was approaching the centre of a real course.

The elimination of the effect of modulation depth in the audio type of receiver makes it a ratio-measuring device rather than a meter of the difference of two quantities. This has a considerable advantage in relation to the test equipment used for setting the receiver course sensitivity; it is difficult to establish accurately the known absolute modulation depths of two tones and hence a given difference in depth of modulation. On the other hand, it is comparatively easy to arrange the modulation depths of two tones to be equal and then to establish a known ratio by means of a mathematically calculated attenuator. As the initial setting accuracy of a receiver plays an equal part with its subsequent stability in service in obtaining the required indication of course width, this factor is of no small importance in the operation of the glide-slope facility.

The advantage of the system as a ratio-measuring device may not be confined to its connection with test gear, as the elimination of the carrier as a variable could have advantages when considering ground-equipment stability. In some installations, the radiation intensity of one aerial array affects modulation depth.

8. *Monitor Flags*

Receiver deterioration may cause an apparent broadening of the course until in the limit a completely non-operative receiver will give a permanent "on-course" indication. It is necessary continuously to monitor receiver deflection sensitivity in some way that will give, ideally, a

decisive warning when the sensitivity falls below a pre-determined value into a region of danger.

This is done in both receivers by taking some of the outputs from each rectifier and feeding them in an aiding sense into movements on the cross-pointer indicator. These movements show a red warning flag when the current falls below a predetermined value. The deflection current through the flag movement is therefore proportional to the sum of the outputs from the 90- and 150-cycle filters. Reference to Figure 2 shows that on the localiser receiver this current is proportional to overall modulation depth, whereas in the glide-slope receiver, provided the automatic volume control maintains the input to the filters to a constant value, the flag current does not change. In this type of receiver, the flag current is determined by receiver sensitivity alone and not by modulation depth, hence it is possible to set the flag current to a borderline value such that even a small decrease in sensitivity, and associated course widening, will cause a warning to be given.

On the localiser, however, permissible modulation depths may be between 32 and 44 per cent, and it is necessary to set the flags such that a warning is not given even at the lowest value. If an approach is now made on a ground equipment using 44-per-cent modulation, since the flag current is proportional to modulation depth, the operating current would increase by over 30 per cent, it would now be possible for an increase in course width of over 25 per cent to occur before the flag system even commenced to operate.

Audio-derived automatic volume control thus considerably increases the efficiency of the warning circuits. If the difficulties with regard to its use on the localiser receiver can be overcome, the adoption of this system would reduce the serious weaknesses that can exist in connection with flag alarms and stability in general.

9. SR.14/15

The signal amplification and channel selection mechanism of the *SR.14/15* receivers is not unusual though on the *SR.14* with carrier-derived automatic volume control it is especially necessary to prevent modulation increase or decrease at the mixers. Both receivers are double superheterodynes employing wide-band radio-frequency and first-oscillator stages; there is no mechanical tuning and channel selection is effected by choice of appropriate pairs of localiser and glide-slope crystals in the control box. Crystals are switched into the grids of oscillator valves in the control box and the output is fed by a matched coaxial line into the first-oscillator stages of the appropriate receiver.

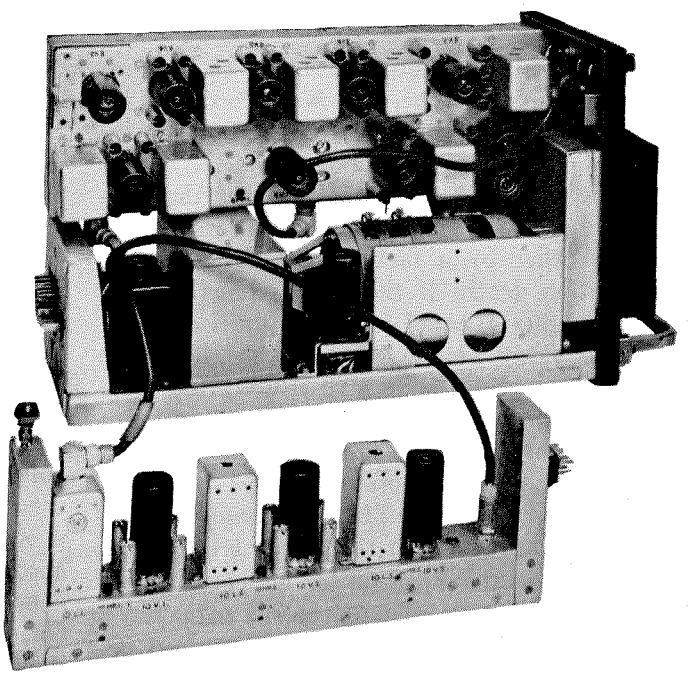


Figure 4—Interior view of the *SR.15* glide-slope receiver with one of the amplifier subassemblies removed from its working position.

These methods ensure simplicity of operation, freedom from mechanical trouble, and reduction of weight and size.

Both localiser and glide-slope receivers weigh approximately 17·5 pounds (8 kilograms) and dimensions are 6 by 8 by 12 inches (15 by 20 by 30 centimetres). The localiser receiver is shown in Figure 1. Figure 4 is an interior view of the glide-slope receiver.

The marker receiver is a two-stage straight receiver followed by an audio-frequency amplifier, a rectifier (which produces a direct current in sympathy with the beacon keying), direct-current amplifier, and relay. The relay is used

to actuate a lamp that will follow the keying of the marker beacon. The audio output containing 400-, 1300-, or 3000-cycle coded signals is fed into the audio stages of the localiser receiver, which is in the same unit as the marker receiver, and is mixed with the localiser identification signals.

10. Test Equipment

10.1 Initial Calibration

If the advantages of greater stability are to be realised, it is necessary to ensure that equipments are calibrated before installation with apparatus that justifies this stability. This requires the use of signal generators that have not hitherto been in use.

The types of very-high-frequency signal generators that have been in general use for checking communication receivers cannot normally be modulated at 90 and 150 cycles without discrimination or distortion. Should even this be possible, most generators are directly modulated oscillators and consequently exhibit considerable frequency modulation especially in the region of 90-per-cent modulation. Such a fault will produce change of apparent modulation depth and possible harmonic distortion when encountering the relatively narrow-band intermediate-frequency stages in the instrument landing receivers. A signal generator, which was under development for general very-high-frequency purposes in the range 95 to 160 megacycles, was slightly modified to make it also suitable for our use. The modulator used in this generator is of the so-called series type, where the cathode of the modulator is directly coupled to the modulating point; this eliminates the use of a modulation transformer and ensures a modulation characteristic that is flat to the lowest required frequencies. A buffer stage is included which, by reducing the coupling of modulating voltage to the oscillator, keeps the frequency modulation to less than 7 kilocycles at 90-per-cent modulation.

A closely balanced 90- and 150-cycle signal must be available for modulating the signal generator. Diodes have to be used to ensure stability of balance and an accurately designed attenuator serves to produce the required ratios of the tones when adjusting receiver course sensitivity.

There is a permissible frequency tolerance of 2 per cent allowed by the International Civil Aviation Organization on the tones of the ground transmitter. This is to allow for mains frequency variation on those transmitters that derive their modulation from a synchronous motor driven off the mains. Better than 1 per cent must be achieved on tone-frequency tolerance in test gear.

To avoid the expense and time involved in the design of a signal generator in the glide-slope frequency range that meets the requirements, a frequency changer is used in conjunction with the localiser signal generator. This unit comprises a 210-megacycle line oscillator, the frequency of which is added to the output of the main generator, to produce signals in the 330-megacycle range. A calibration curve is provided to correlate operating frequency with unit attenuation and enables a correction to be applied to the master-attenuator reading for checking the glide-slope injection signal.

Marker alignment can be carried out with any ordinary very-high-frequency signal generator. As the receiver is a straight set of large bandwidth, frequency modulation is of no major consequence. There is also provision on the glide-slope frequency changer for subtracting the master-signal-generator frequency from the line oscillator to produce 75 megacycles. This would eliminate the necessity for a special marker generator.

10.2 Field Test Set

It is necessary to give an overall check of the airborne installation before each flight. By ensuring that the characteristics of the receiver with regard to balance and deflection sensitivity have not changed appreciably since installation, the danger and expense of last-minute diversion can be much reduced.

The field test equipment, which may be seen in Figure 5, permits accurate checking on one crystal-controlled channel in the localiser and glide-slope frequencies. The audio, filter, and rectifier circuits are carefully checked on this crystal-controlled channel; the operation of the radio-frequency circuits on all frequencies is examined by the functional test. The apparatus is not a laboratory standard but is sufficiently accurate to detect such faults as might endanger the aircraft or prevent the execution of an instrument approach.

The test equipment radiates a 90- and 150-cycle modulated signal from a point external to the aircraft, through the airborne aerial system to the equipment, thus ensuring an overall check. On the crystal-controlled channel, it is possible continuously to adjust the difference in the depth of modulation by means of the deflection control, this quantity being indicated by a centre-zero meter on the front panel. Balance and deflection sensitivity of the airborne receiver are checked against this meter.

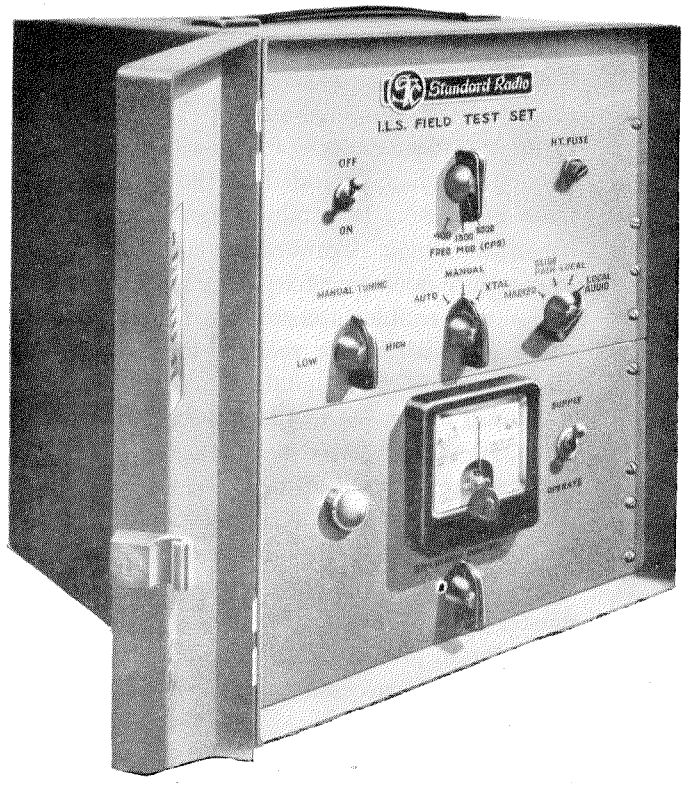


Figure 5—Field test equipment.

The functional check is carried out by sweeping the localiser and glide-slope bands at 150 cycles; the sweep is derived from a reactance valve driven from a synchronised neon saw-tooth oscillator. This causes a deflection to occur on all working channels due to the 150-cycle sweep frequency that will appear at the detector of the receiver.

The variable difference in the depth of modulation is produced from diode-stabilised 90- and 150-cycle oscillators, the outputs of which are differentially controlled by the deflection control on the front panel; the two frequencies are mixed in a modulator-mixer on which there is considerable negative feedback to ensure stability of output. Diodes in the cathodes circuits of each valve of the modulator-mixer produce direct currents proportional to the 90- and 150-cycle components of modulating voltage. These currents are fed in opposition to the centre-reading meter on the front panel, the off-balance current in this meter then indicates the difference of modulating voltages. Neon stabilisers are used to fix the anode voltage of the modulated valves, this ensures that modulation depth bears a fixed relation to the modulating voltage; from the above reasoning, it appears that the meter on the front panel gives a continuous indication of difference in depth of modulation.

The set is calibrated for balance and difference in depth of modulation against a standard receiver and relies on its inherent stability for maintaining calibration. The degree of accuracy to which it maintains the initial setting is adequate in view of the stabilisation of important fixed voltages and the amount of feed-back applied to those circuits that handle the variable voltages.

The equipment is portable and requires less than 90 watts from a 24-volt trolley accumulator. It weighs 26 pounds (12 kilograms) and the dimensions are 11 by $10\frac{1}{2}$ by 13 inches (28 by 27 by 33 centimetres).

11. Acknowledgements

Acknowledgement is made to Mr. W. E. Brunt of the British Overseas Airways Corporation, who was in close touch with us from the early days of the *SR.14/15* development. He gave us considerable information and advice that was of great help.

Mr. J. W. James of Standard Telephones and Cables, Limited, was equally involved with the author in the design of the *SR.14/15*. Without his work, the receiver in its present form could not have been completed.

Carrier Concentrations and Fermi Levels in Semi-Conductors

By J. S. BLAKEMORE

Standard Telecommunication Laboratories, Limited; London, England

IN THIS REVIEW, the statistical theory of the free carrier concentrations, n_e and n_h, and the Fermi level ϵ^* is developed for semi-conductors under a wide range of conditions using Fermi-Dirac statistics. Some parts of the work have been covered by Seitz,[1] Shifrin,[2] and Putley,[3] and this published material has been co-ordinated and tabular and graphical information added to facilitate calculation.

The authors mentioned above have all used different notations, none of which are entirely satisfactory, and no apologies are made for using yet another system of symbols, which retains the most useful features of the previous ones.

In the first few sections, the theory is developed for a simple n-type semi-conductor, leading up to the more complicated cases discussed in Section 6.

1. Nature of Fermi-Dirac Statistics

From general thermodynamic principles, it is clear that if a system contains a large number of particles, such as electrons, then these particles will normally tend to find positions of lowest energy.

Equally clearly, at high temperatures particles will be continually receiving and transmitting energy in a way that will tend to oppose the process of settling towards minimum energy.

In the band theory of solids, first used by Bloch,[4] it is assumed that the outer electrons of each atom form an electron 'gas,' the members of which are free to move through the solid in an electric field due to the periodic array of atomic cores and the other electrons in the gas.*

When the wave function of such an electron is solved, it is found that the wave equation is only satisfied for certain ranges of electron energy. Thus an energy-level diagram for a solid consists of a series of permitted bands separated by forbidden zones.

By application of the Pauli principle, as extended by Dirac, that only two electrons (of opposing spin) can have the same three quantum numbers, it is found that only a limited number of electrons per unit volume can be contained in a given band and that the distribution of available electron levels as a function of energy within each band is fixed.

In a metal, the highest occupied band is only partly full, and by application of an external field, electrons can be accelerated, thus rising to higher, previously unoccupied, levels in the band. This can lead to conduction.

In an insulator, the highest occupied band is completely full, and as now electronic energies can only be exchanged, no net result is obtained on the application of an external electric field.

In a semi-conductor, a number of electrons are removed from the full band or introduced into the empty band above it, or both, by thermal excitation and thus change of electron energy and hence electrical conductivity is made possible. This idea was first suggested by Wilson.[6,7]

The probability that an energy level is occupied by an electron is given by the Fermi-Dirac function

$$f(\epsilon) = \frac{1}{1+\exp\left(\frac{\epsilon-\epsilon^*}{kT}\right)}, \qquad (1)$$

[1] F. Seitz, "Modern Theory of Solids," 1st Edition, McGraw-Hill Book Company, New York, New York; 1940.

[2] K. Shifrin, "On the Theory of Electric Properties of Good Conducting Semi-Conductors," *Journal of Physics* (U.S.S.R.), v. 8, n. 4, pp. 242–252; 1944.

[3] E. H. Putley, "Electrical Conductivity of Germanium," *Proceedings of the Physical Society*, v. A62, pp. 284–292; May, 1949.

[4] F. Bloch, "Über die Quantenmechanik der Elektronen in Kristallgittern," *Zeitschrift für Physik*, v. 52, pp. 555–600; December, 1928.

* The basis of the free-electron theory of solids and of the band approximation is discussed by Hume-Rothery[5], whilst a more mathematical treatment is given by Seitz[1].

[5] W. Hume-Rothery, "Atomic Theory for Students of Metallurgy," Institute of Metals, London, England; 1947.

[6] A. H. Wilson, "The Theory of Electronic Semi-Conductors," *Proceedings of the Royal Society of London*, v. A133, pp. 458–491; October, 1931.

[7] A. H. Wilson, "The Theory of Electronic Semi-Conductors—II," *Proceedings of the Royal Society of London*, v. A134, pp. 277–287; November, 1931.

where ϵ is the energy under consideration and ϵ^* is a reference level known as the Fermi level. The significance of this level will become apparent shortly.

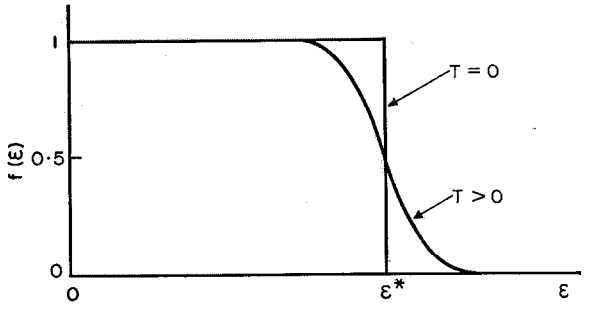


Figure 1—Fermi-Dirac probability function.

The form of this function is shown in Figure 1 for zero temperature and for a temperature T greater than zero.

For energies much greater than ϵ^*, the exponential term in the denominator of (1) is much greater than unity, and this function approximates to the classical Boltzmann probability function

$$f(\epsilon)_{\text{class}} = K.\exp(-\epsilon/kT). \qquad (2)$$

The number of electrons at any energy is governed by the product of the probability with the expression for the density of permissible electron levels at that energy.

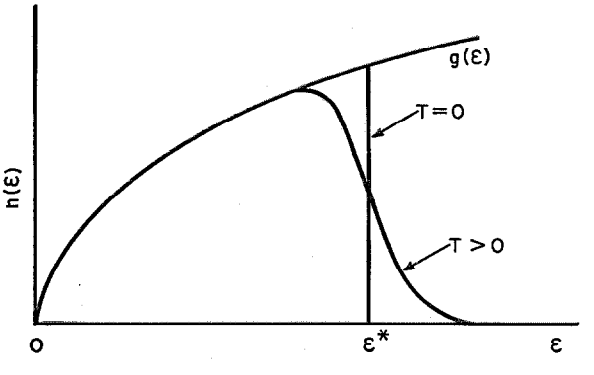


Figure 2—Electron density distribution in a highly degenerate case.

In the forbidden zones, of course, the latter function is zero except at special positions where impurity levels may exist.

Inside a permitted band and fairly near to the bottom, the density of states is given by

$$g(\epsilon) = 4\pi(2m_e/h^2)^{3/2}.(\epsilon-\epsilon_0)^{1/2}, \qquad (3)$$

where m_e is a quantity known as the effective mass of an electron in the band (which may be appreciably different from the actual electronic mass) and ϵ_0 is the energy at the bottom of the band.

Since only differences of potential energy are considered, the origin of energy is purely arbitrary, and it is convenient to choose the bottom of the band, i.e., $\epsilon_0 = 0$.

From (1) and (3), the electron density $n(\epsilon)$ in the band is given by

$$n(\epsilon) = f(\epsilon).g(\epsilon) = 4\pi(2m_e/h^2)^{3/2}.\frac{\epsilon^{1/2}}{1+\exp\left(\frac{\epsilon-\epsilon^*}{kT}\right)}. \qquad (4)$$

In Figure 2, $n(\epsilon)$ is shown as a function of ϵ for a positive value of the Fermi level ϵ^* and for the same temperature conditions as in Figure 1. This is the type of electron distribution dealt with in a metal or a degenerate semi-conductor.

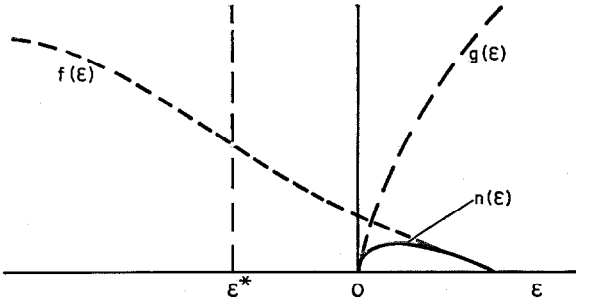


Figure 3—Electron density distribution in a non-degenerate case.

Figure 3 shows the form of $n(\epsilon)$ for a case in which the Fermi level is negative. If ϵ^* is several kT below the bottom of the band, then $f(\epsilon)$ coincides with the classical approximation. This is the case for non-degenerate or 'classical' semiconductors.

The Fermi level adjusts itself automatically as a normalising parameter when temperature changes since it must have a value such that

$$\int_{-\infty}^{\infty} n(\epsilon).d\epsilon = \int_{-\infty}^{\infty} g(\epsilon).f(\epsilon).d\epsilon$$

is equal to the total number of electrons present.

2. *Free-Electron Concentration and Fermi Level in an n-Type Semi-Conductor*

As is customary, the theory is given for an n-type semi-conductor in which the current carriers are free electrons. It can easily be shown, e.g., Lempicki,[8] that an identical analysis applies for a p-type semi-conductor in which the current carriers are 'holes' in an otherwise full band.

The band model used is illustrated in Figure 4. A conduction band, empty at zero temperature, extends upwards from the zero of energy.

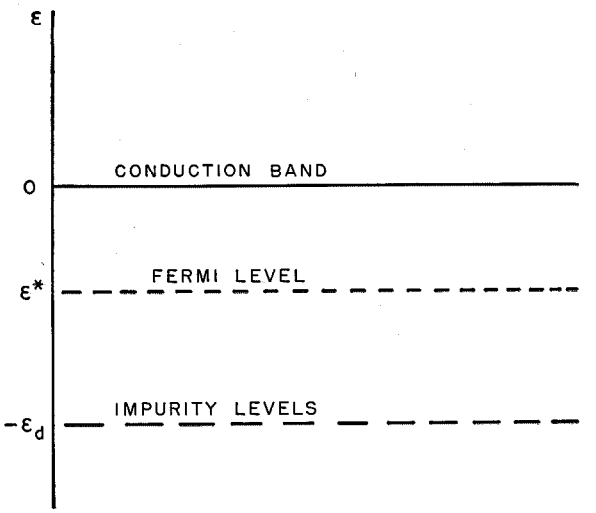


Figure 4—Energy-band structure for an n-type semi-conductor.

A forbidden band extends downwards from this zero, and it is assumed that the highest full band of the material is at a very large negative energy, such that no carriers are produced by direct thermal excitation from this band at normal temperatures. Thus the full bands can be ignored in the analysis of Sections 2 to 5.

At an energy $-\epsilon_d$, N_d donor impurity centres per cubic centimetre are situated, each of which can supply one electron. These impurity levels are localised, so that an electron cannot pass from one impurity to the next without being excited into the conduction band.

The Fermi level ϵ^* may be either in or below the conduction band, depending on the amount of degeneracy, if any.

[8] A. Lempicki, "Electrical Conductivity of Simple p-Type Semiconductors," *Proceedings of the Physical Society*, v. A64, pp. 589–590; June, 1951.

Equations will be derived establishing the value of ϵ^* and the number of free electrons in the conduction band n_e for given values of N_d, ϵ_d, and T.

Now for $\epsilon<0$, $g(\epsilon)=0$ except at $\epsilon=-\epsilon_d$, where there is a high concentration of impurity states. Thus

$$\int_{-\infty}^{0} g(\epsilon).d\epsilon = N_d.$$

Now with a discontinuous function of the type $g(\epsilon)$ and a smooth function of the type $f(\epsilon)$, it can be shown that

$$\int_{-\infty}^{0} g(\epsilon).f(\epsilon).d\epsilon = f(-\epsilon_d)\int_{-\infty}^{0} g(\epsilon).d\epsilon$$

$$\therefore \quad \int_{-\infty}^{0} n(\epsilon).d\epsilon = \frac{N_d}{1+\exp\left(\frac{-\epsilon_d-\epsilon^*}{kT}\right)}. \qquad (5)$$

This expression gives the number of electrons 'bound' in the impurity states for a given N_d, ϵ_d, ϵ^*, and T.

Similarly, the number of electrons in the conduction band is

$$n_e = \int_0^{\infty} g(\epsilon).f(\epsilon).d\epsilon$$

$$= 4\pi(2m_e/h^2)^{3/2}\int_0^{\infty} \frac{\epsilon^{1/2}.d\epsilon}{1+\exp\left(\frac{\epsilon-\epsilon^*}{kT}\right)}. \qquad (6)$$

But the total number of electrons present at any temperature must be N_d. Thus

$$N_d = \frac{N_d}{1+\exp\left(\frac{-\epsilon_d-\epsilon^*}{kT}\right)} + n_e$$

and

$$n_e = \frac{N_d}{1+\exp\left(\frac{\epsilon_d+\epsilon^*}{kT}\right)}. \qquad (7)$$

By solving (6) and (7), n_e and ϵ^* can both be obtained.

In order to do this, it is convenient to change the variables to dimensionless quantities as in the following. Let $\eta=\epsilon/kT$, $\eta_d=\epsilon_d/kT$, $\eta^*=\epsilon^*/kT$, $G=m_e/m$. Let $n_0=4\pi(2mkT/h^2)^{3/2}$ $=5{\cdot}541.10^{15}.T^{3/2}$, centimetres^{-3}. The numerical result is calculated by assuming the values for the fundamental physical constants quoted by

Bearden and Watts.[9] The numerical values of this and similar functions are collected together in the appendix, Section 8.1. Finally, let

$$F_j(\eta^*)=\int_0^\infty \frac{\eta^j\,d\eta}{1+\exp(\eta-\eta^*)}.$$

The functions $F_j(\eta^*)$, where j is an integer or half integer, are collectively known as the Fermi-Dirac integrals. Accurate tables of $F_{1/2}(\eta^*)$ and $(2/3).F_{3/2}(\eta^*)$ have been given by McDougall and Stoner,[10] whilst Wright[11] has tabulated values of $F_1(\eta^*)$ and $F_2(\eta^*)$ for integral values of η^*. In Figure 5, some classical and degenerate approximations to $F_{1/2}(\eta^*)$, which hold for certain ranges of values of η^*, are shown together with $F_{1/2}(\eta^*)$ itself. In Figure 6, graphs are given of $F_{1/2}(\eta^*)$ for small positive and negative values of η^*.

With the substitutions carried out in (6) and (7), the equations reduce to

$$n_e=n_0.G^{3/2}.F_{1/2}(\eta^*) \qquad (6A)$$

and

$$n_e=\frac{N_d}{1+\exp(\eta_d+\eta^*)}. \qquad (7A)$$

These equations will now be solved for non-degenerate and degenerate cases.

3. *Calculation of n_e and ε^* for Classical Semi-Conductors*

These are semi-conductors in which the Fermi-Dirac functions reach their classical limits.

Now, when $\eta^*<0$,

$$F_{1/2}(\eta^*)=\tfrac{1}{2}\pi^{1/2}\exp\eta^*\sum_{s=0}^{\infty}(-1)^s\frac{\exp(s\eta^*)}{(1+s)^{3/2}} \qquad (8)$$

and when $\eta^*<-2$, it is correct to within 5 per cent to take only the first term of this expansion.

Thus, if $\eta^*<-2$,

$$F_{1/2}(\eta^*)=\tfrac{1}{2}\pi^{1/2}.\exp(\eta^*). \qquad (9)$$

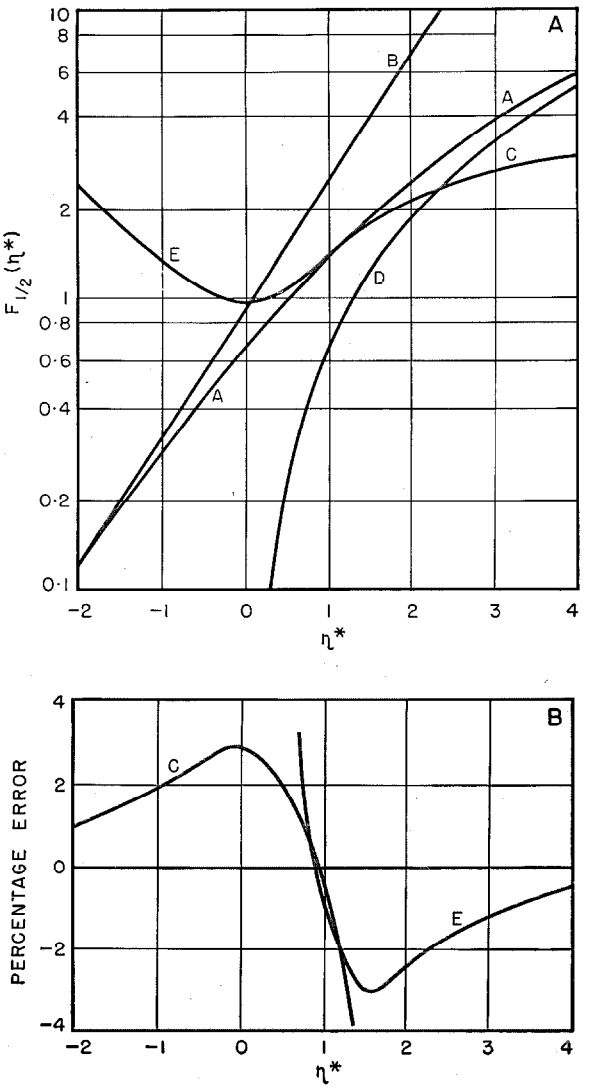


Figure 5—The function $F_{1/2}(\eta^*)$ and its approximations are given in *A* for A) true values, B) classical approximation (9), C) classical approximation (19A), D) degenerate approximation (16), and E) degenerate approximation (18). *B* gives the percentage error in the two approximations C and E.

Hence, from (9) and (6A),

$$n_e=2(2\pi m_e kT/h^2)^{3/2}.\exp\eta^*=N_e.\exp\eta^*. \qquad (10)$$

The symbol N_e, which is useful in classical cases, equals $4{\cdot}831.10^{15}G^{3/2}T^{3/2}$, centimetres^{-3}. When $\eta^*=-2$, n_e has the largest value consistent with the accurate use of simple classical statistics. This value is $6{\cdot}246.10^{14}T^{3/2}$, or $3{\cdot}085.10^{18}$, centimetres^{-3} at room temperature. ($T=290$ degrees Kelvin.)

[9] J. A. Bearden and H. M. Watts, "A Re-Evaluation of the Fundamental Atomic Constants," *Physical Review*, v. 81, pp. 73–81; January, 1951.

[10] J. McDougall and E. C. Stoner, "The Computation of Fermi-Dirac Functions," *Philosophical Transactions of the Royal Society of London*, v. A237, pp. 67–104; February, 1938.

[11] R. W. Wright, "Variation With Temperature of the Electrical Properties of a Degenerate Electronic Semiconductor as exemplified by Cadmium Oxide," *Proceedings of the Physical Society*, v. A64, pp. 350–362; April, 1951.

By elimination of η^* between (7A) and (10), we have

$$n_e^2 = (N_d - n_e) N_e . \exp(-\eta_d). \qquad (11)$$

Now when the Fermi level is several kT above the impurity levels, only a small fraction of the donors are ionised ($n_e \ll N_d$). This is a reserve semi-conductor, for which (11) approximates to

$$n_e = (N_d N_e)^{\frac{1}{2}} . \exp(-\epsilon_d / 2kT) \qquad (11A)$$

and, from (10),

$$\epsilon^* = \frac{-\epsilon_d}{2} + \frac{kT}{2} . \log_e \frac{N_d}{N_e} . \qquad (12)$$

Equations (11A) and (12) are the original equations of Wilson[7] and are frequently assumed in the interpretation of results, sometimes in cases where their use is not justified, such as when a considerable fraction of the impurities are ionised. In such a case, n_e must be found as the positive root of the quadratic equation (11).

A simplification appears again when almost all the impurities are ionised. This is known as an exhaustion semi-conductor, for which

$$n_e = N_d \qquad (13)$$

and

$$\epsilon^* = kT . \log_e (N_d / N_e). \qquad (14)$$

For these to be valid, n_e must not be too great or degeneracy will appear. Moreover ϵ_d must be fairly small for exhaustion to occur at reasonable temperatures.

In Figure 7, the transition from reserve to exhaustion conditions is shown for a semi-conductor with an activation energy $\epsilon_d = 0{\cdot}20$ electron volt and an impurity content $N_d = 10^{17}$ centimetres^{-3}, assuming that free electrons have the normal electronic mass, i.e., $G = 1$. This figure reproduces some results calculated by Huttner et al.[12] The slope of the curve relating n_e to $1/T$ at low temperatures should be approximately equal to $\epsilon_d / 2k$. A small error occurs due to the temperature variation of N_e in (11A). The slope of the linear portion of Figure 7A actually yields $\epsilon_d = 0{\cdot}21$ electron volt, in fair agreement with the true value of $\epsilon_d = 0{\cdot}20$ electron volt.

[12] R. A. Huttner, E. S. Rittner, and F. K. du Pre, "Fermi Levels in Semiconductors," *Philips Research Reports*, v. 5, pp. 188–204; June, 1950.

4. *Calculation of n_e and ϵ^* for Degenerate Semi-Conductors*

4.1 Concept of Degeneracy

In cases of high impurity concentration coupled with low activation energy, the concentration of free electrons in the conduction band may rise well above the figure of 3.10^{18} centimetres^{-3} at room temperature, and appreciable degeneracy will occur in such cases.

In a degenerate system, large changes in the

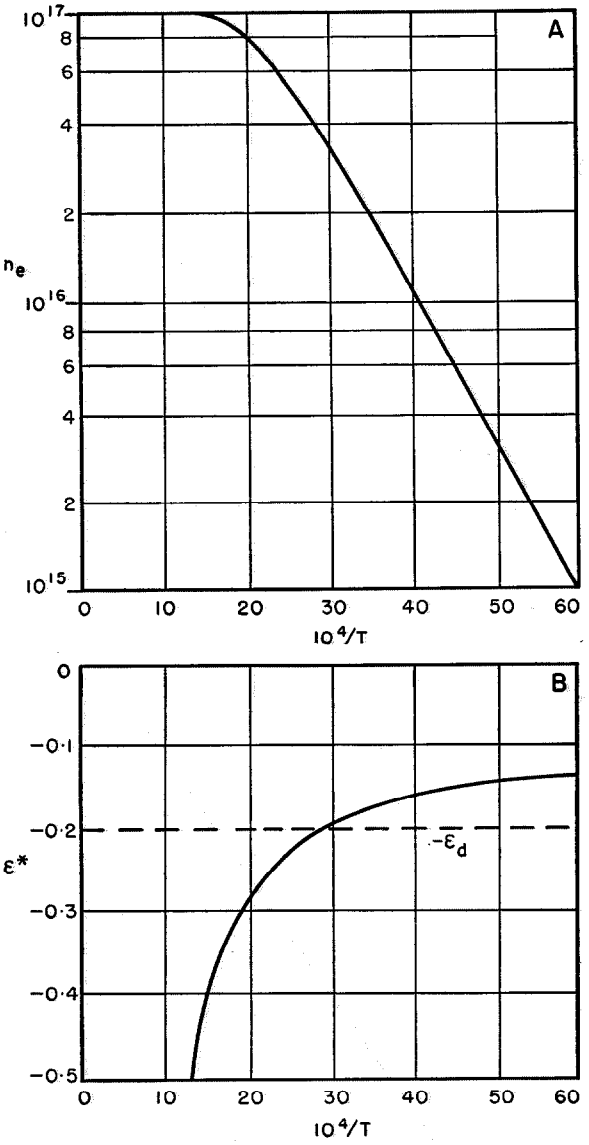
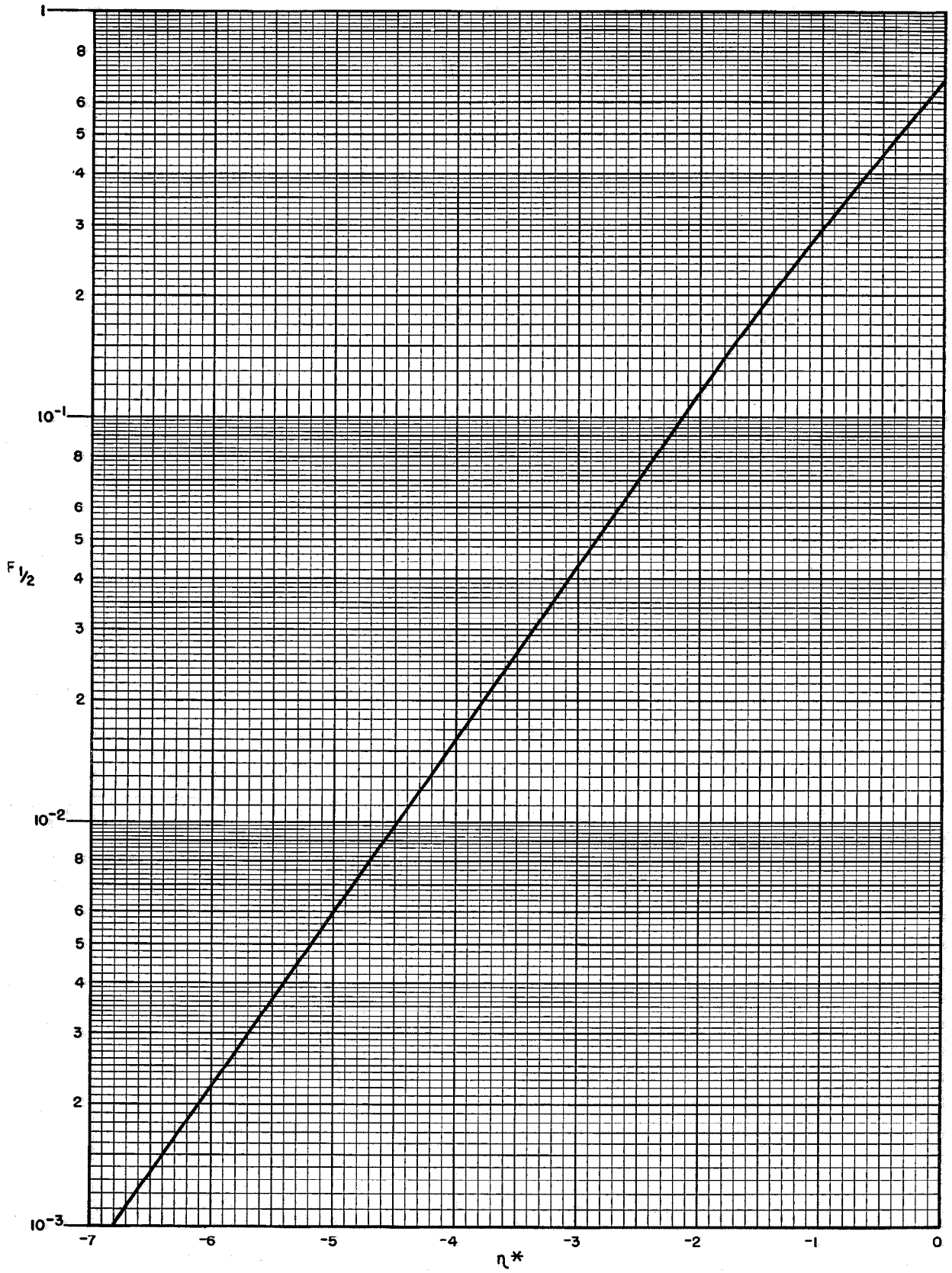


Figure 7—Variation with temperature of the properties of a 'classical' semi-conductor. *A* is for a free-electron concentration and *B* is for the Fermi level. T=degrees Kelvin.


1
8
6
5
4
3
2
10^{-1}
8
6
5
4
3
2
10^{-2}
8
6
5
4
3
2
10^{-3}
$F_{1/2}$
-7
-6
-5
-4
-3
-2
-1
0
η^*

impurity concentration have only a slight effect on the free-electron concentration.

This can easily be understood by considering the state of affairs shown in Figure 2. A large number of non-ionised impurities may be situated at a small energy below the origin. However, as all possible electron states are occupied almost up to the level ϵ^*, an energy $(\epsilon_d+\epsilon^*)$ is required to increase the free-electron concentration any further.

Thus the larger n_e becomes, and in consequence the larger ϵ^* becomes, the more independent n_e becomes of changes in N_d.

Figure 6—$F_{1/2}$ is plotted against small valucs of η^* in the curves on the opposite page and below. When $\eta^*<-6$, $F_{1/2}(\eta^*)=0{\cdot}886$ exp η^* and when $\eta^*>+6$, $F_{1/2}(\eta^*)=0{\cdot}667\,\eta^{*3/2}$.

4.2 Analytical Expressions for $F_{1/2}(\eta^*)$

For cases of strong degeneracy, a simple function can be found for $F_{1/2}(\eta^*)$. Considering Figure 2 again, if $kT\ll\epsilon^*$, i.e., $\eta^*\gg1$, then the area under the curve for $T>0$ is the same as that of the exact half parabola that applies at $T=0$.

Now in (4), we have

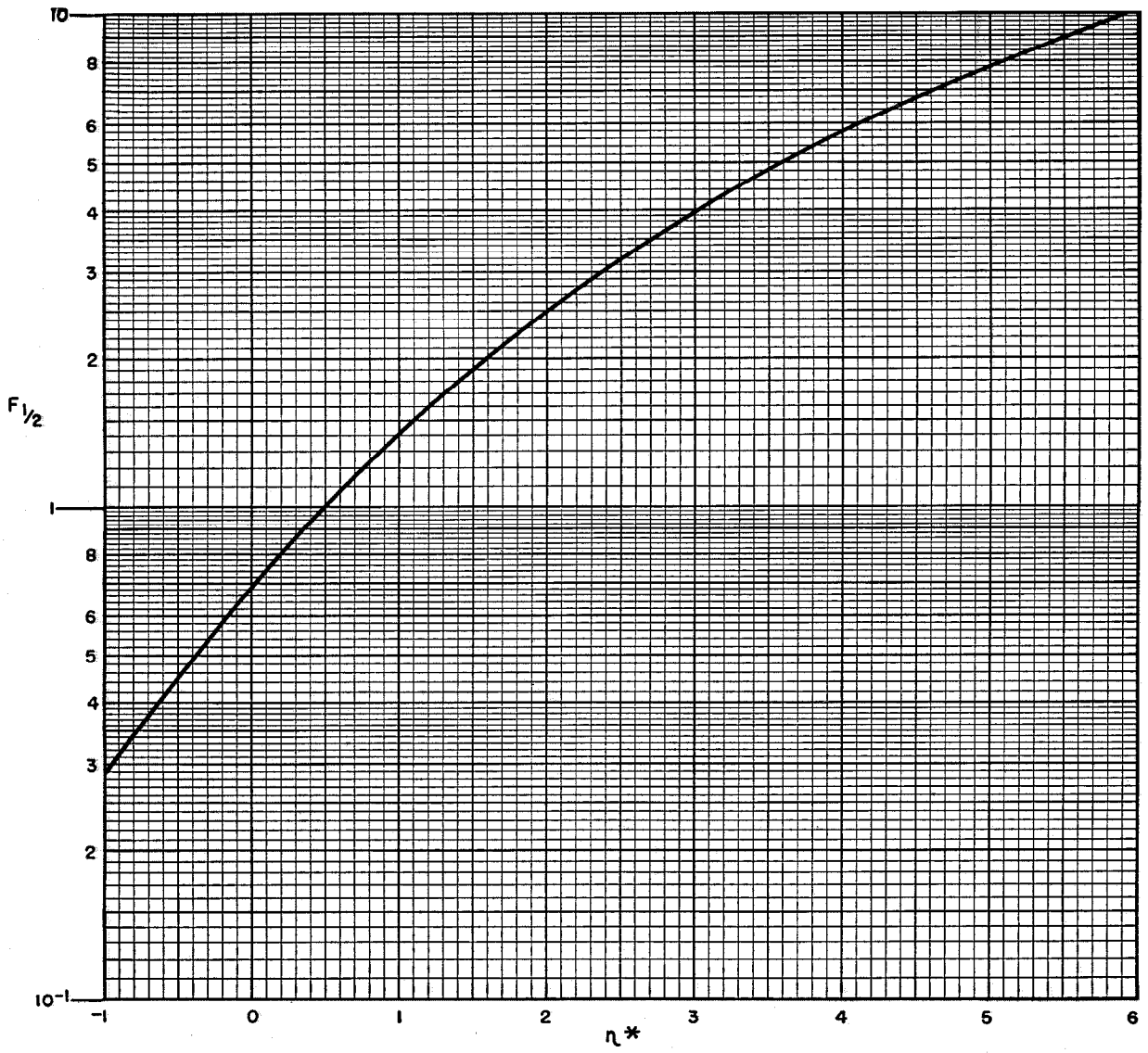
$$n(\epsilon)=4\pi(2m_e/h^2)^{3/2}.\frac{\epsilon^{1/2}}{1+\exp\left(\frac{\epsilon-\epsilon^*}{kT}\right)}.$$

Thus for the parabola at $T=0$

$$\begin{aligned} n(\epsilon)&=4\pi(2m_e/h^2)^{3/2}.\epsilon^{1/2}, \quad \epsilon<\epsilon^* \\ &=0 \qquad\qquad\qquad\quad , \quad \epsilon>\epsilon^* \end{aligned}$$

$$\therefore\quad n_e=\int_0^{\epsilon^*}n(\epsilon).d\epsilon=\frac{8\pi}{3}.\left(\frac{2m_e}{h^2}\right)^{3/2}.\epsilon^{*3/2}$$

$$=(2/3).n_0G^{3/2}.\eta^{*3/2}. \qquad (15)$$



impurity concentration have only a slight effect on the free-electron concentration.

This can easily be understood by considering the state of affairs shown in Figure 2. A large number of non-ionised impurities may be situated at a small energy below the origin. However, as all possible electron states are occupied almost up to the level ϵ^*, an energy $(\epsilon_d+\epsilon^*)$ is required to increase the free-electron concentration any further.

Thus the larger n_e becomes, and in consequence the larger ϵ^* becomes, the more independent n_e becomes of changes in N_d.

Figure 6—$F_{1/2}$ is plotted against small values of η^* in the curves on the opposite page and below. When $\eta^*<-6$, $F_{1/2}(\eta^*)=0{\cdot}886$ exp η^* and when $\eta^*>+6$, $F_{1/2}(\eta^*)=0{\cdot}667\,\eta^{*3/2}$.

4.2 Analytical Expressions for $F_{1/2}(\eta^*)$

For cases of strong degeneracy, a simple function can be found for $F_{1/2}(\eta^*)$. Considering Figure 2 again, if $kT\ll\epsilon^*$, i.e., $\eta^*\gg1$, then the area under the curve for $T>0$ is the same as that of the exact half parabola that applies at $T=0$.

Now in (4), we have

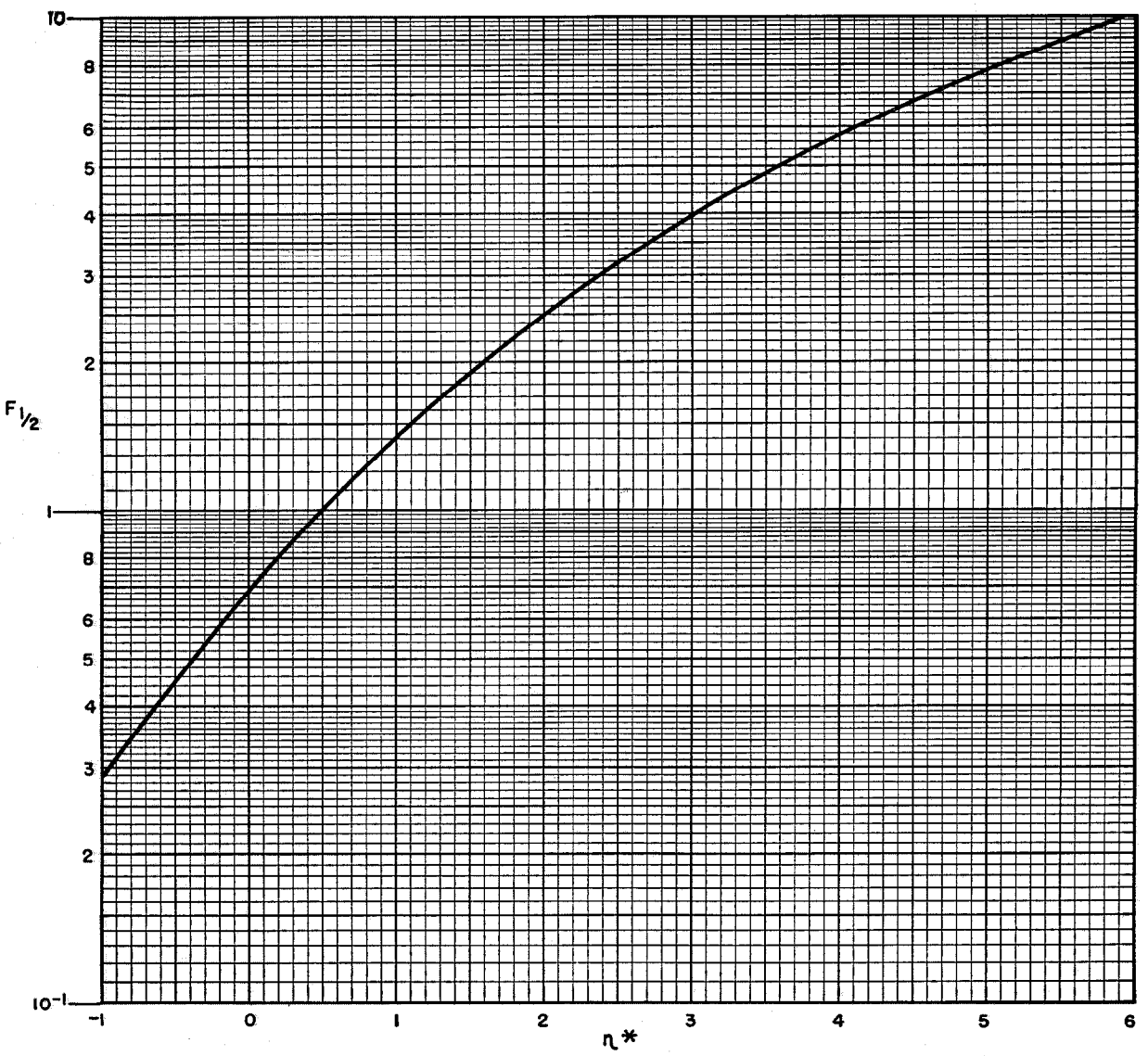
$$n(\epsilon)=4\pi(2m_e/h^2)^{3/2}.\frac{\epsilon^{1/2}}{1+\exp\left(\frac{\epsilon-\epsilon^*}{kT}\right)}.$$

Thus for the parabola at $T=0$

$$\begin{aligned} n(\epsilon)&=4\pi(2m_e/h^2)^{3/2}.\epsilon^{1/2}, \quad \epsilon<\epsilon^* \\ &=0 \qquad\qquad\qquad\quad , \quad \epsilon>\epsilon^* \end{aligned}$$

$$\therefore\quad n_e=\int_0^{\epsilon^*} n(\epsilon).d\epsilon=\frac{8\pi}{3}.\left(\frac{2m_e}{h^2}\right)^{3/2}.\epsilon^{*3/2}$$

$$=(2/3).n_0G^{3/2}.\eta^{*3/2}. \qquad (15)$$



Thus

$$F_{1/2}(\eta^*) = (2/3).\eta^{*3/2}, \quad \eta^* \gg 1. \tag{16}$$

This expansion will obviously fail for small values of η^*, when the transition zone is broad, so that some electron states at the bottom of the band remain unfilled. Curve D in Figure 5A shows that the error is appreciable if $\eta^* < 8$.

A more complete analysis of the area under the $n(\epsilon)$ curve yields an improved expression

$$F_{1/2}(\eta^*) = (2/3)\eta^{*3/2}(1+\pi^2/8\eta^{*2}), \tag{17}$$

which is satisfactory provided that $\eta^* > 1{\cdot}5$. A further improvement can be made by observing that (17) is identical with the first two terms of the expansion

$$\begin{aligned} F_{1/2}(\eta^*) &= (2/3)(\eta^{*2}+\pi^2/6)^{3/4} \\ &= 0{\cdot}667(\eta^{*2}+1{\cdot}645)^{3/4}, \end{aligned} \tag{18}$$

which is accurate to better than 3 per cent for $\eta^* > 0{\cdot}8$, as shown by curve E in Figure 5B. I am indebted to Mr. L. Lewin for pointing out the advantages of (18) over (17).

For semi-conductors that are but slightly degenerate, where $-2 < \eta^* < 1$, modifications of the classical expression

$$F_{1/2}(\eta^*) = \tfrac{1}{2}\pi^{1/2} \exp(\eta^*), \quad \eta^* < -2$$

have been proposed by Ehrenberg[13] and by Landsberg, Mackay, and McRonald.[14] Both of the suggested modifications are of the form

$$F_{1/2}(\eta^*) = \frac{\tfrac{1}{2}\pi^{1/2}}{\phi+\exp(-\eta^*)}, \tag{19}$$

where Ehrenberg suggests $\phi = 0{\cdot}25$, and Landsberg et al., $\phi = 2^{-3/2} = 0{\cdot}354$. By analysing the properties of the family of curves represented by (19), it is found that $F_{1/2}(\eta^*)$ can be represented to within 3 per cent for all values of η^* up to $+1{\cdot}25$ by

$$F_{1/2}(\eta^*) = \frac{\tfrac{1}{2}\pi^{1/2}}{0{\cdot}27+\exp(-\eta^*)}. \tag{19A}$$

This is plotted as curve C of Figure 5A, with the error involved shown in Figure 5B.

By use of (18) and (19A), $F_{1/2}(\eta^*)$ can be represented to within 3 per cent over the entire range of η^*. A convenient point at which to change over from one expression to the other is $\eta^* = +1$; thus we have

$$F_{1/2}(\eta^*) = \frac{0{\cdot}886}{0{\cdot}27+\exp(-\eta^*)}, \quad \eta^* < +1$$

and

$$F_{1/2}(\eta^*) = 0{\cdot}667(\eta^{*2}+1{\cdot}645)^{3/4}, \quad \eta^* > +1.$$

The condition $\eta^* = 1$ can thus be considered as a dividing line between degeneracy and non-degeneracy.

4.3 Degeneracy Temperature

It is customary to define a degeneracy temperature T_{deg} and the corresponding free-electron concentration n_{deg} for which $\eta^* = 1$. These terms are derived on the assumption of the validity of

$$F_{1/2}(\eta^*) = (2/3)\eta^{*3/2}$$

at $\eta^* = 1$. As can be seen from curve D of Figure 5A, this expression is by no means accurate. However, on an approximate basis, this leads to

$$n_{\text{deg}} = (8\pi/3)(2m_e k/h^2)^{3/2}.T_{\text{deg}}^{3/2}.$$

Thus

$$T_{\text{deg}} = \left(\frac{3}{\pi}\right)^{2/3} \frac{h^2}{8m_e k} n_{\text{deg}}^{2/3}. \tag{20}$$

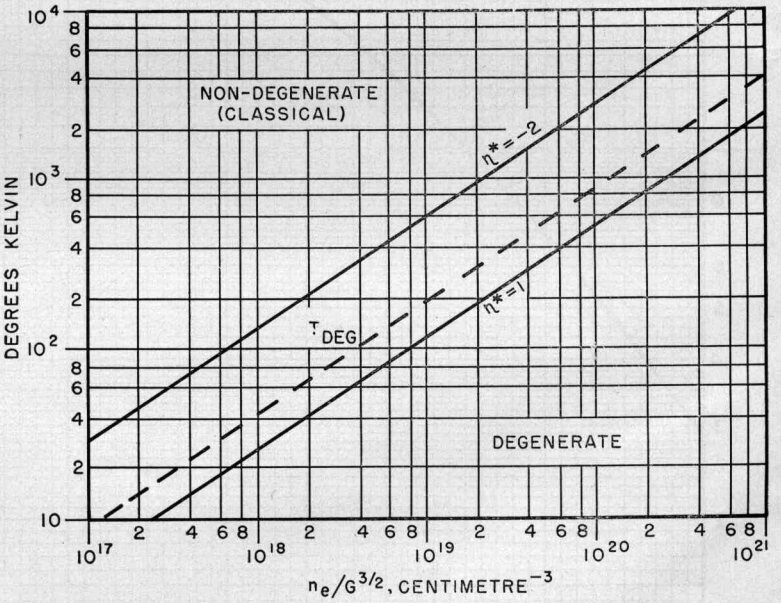


Figure 8—Temperature and conduction-electron concentration for the conditions $\eta^* = -2$ and $\eta^* = +1$ (from (6A)) and for the degeneracy condition (20), which separates classical from degenerate semi-conductors.

[13] W. Ehrenberg, "Electrical Conductivity of Simple Semiconductors," *Proceedings of the Physical Society*, v. A63, pp. 75–76; January, 1950.

[14] P. T. Landsberg, R. Mackay, and A. D. McRonald, "Parameters of Simple Excess Semiconductors," *Proceedings of the Physical Society*, v. A64, pp. 476–480; May, 1951.

This relationship is shown by the dashed curve in Figure 8. It can be seen to differ appreciably from the true line relating temperature and free-electron concentration at $\eta^*=1$. This deviation is due to the inaccuracy of (15) for small values of η^*.

Nevertheless, the concepts of T_{deg} and n_{deg} are useful frequently as rough criteria of the existence of appreciable degeneracy in a semi-conductor.

The solid line in Figure 8 for $\eta^*=1$ should, however, be used as the criterion for the use of (18) or (19A) as approximations to $F_{\frac{1}{2}}(\eta^*)$.

4.4 Evaluation of n_e and η^*

The approximations used for classical semiconductors in deriving n_e and η^*, when N_d, ϵ_d, and T are known, cannot be used in degenerate cases. The results must be obtained by rigorous solution of

$$n_e=n_0G^{\frac{3}{2}}.F_{\frac{1}{2}}(\eta^*)$$

and

$$n_e=\frac{N_d}{1+\exp(\eta_d+\eta^*)}.$$

Thus

$$N_d\{n_0G^{\frac{3}{2}}.F_{\frac{1}{2}}(\eta^*)\}^{-1}=1+\exp(\eta_d+\eta^*). \quad (21)$$

Now

$$N_d(n_0G^{\frac{3}{2}})^{-1}=\frac{N_d}{4\pi(2m_ekT/h^2)^{\frac{3}{2}}}=a.\eta_d^{\frac{3}{2}},$$

where $a=N_dh^3/4\pi(2m_e\epsilon_d)^{\frac{3}{2}}$ is a dimensionless quantity, which depends on the impurity content and activation energy but not on the temperature.

Thus the transcendental equation (21) can be written in the form

$$a.\frac{\eta_d^{\frac{3}{2}}}{F_{\frac{1}{2}}(\eta^*)}=1+\exp(\eta_d+\eta^*), \quad (22)$$

and if either η_d or η^* is known, then the other may be found by solution of $y_1=y_2$, where

$$y_1=a.\frac{\eta_d^{\frac{3}{2}}}{F_{\frac{1}{2}}(\eta^*)} \quad (23)$$

and

$$y_2=1+\exp(\eta_d+\eta^*). \quad (24)$$

The solution of this equation can be carried out either analytically or graphically.

Examples of the graphical solution of (23) and (24) are given in Figures 9A and 9B for a semi-conductor in which $N_d=10^{20}$ centimetres^{-3} and

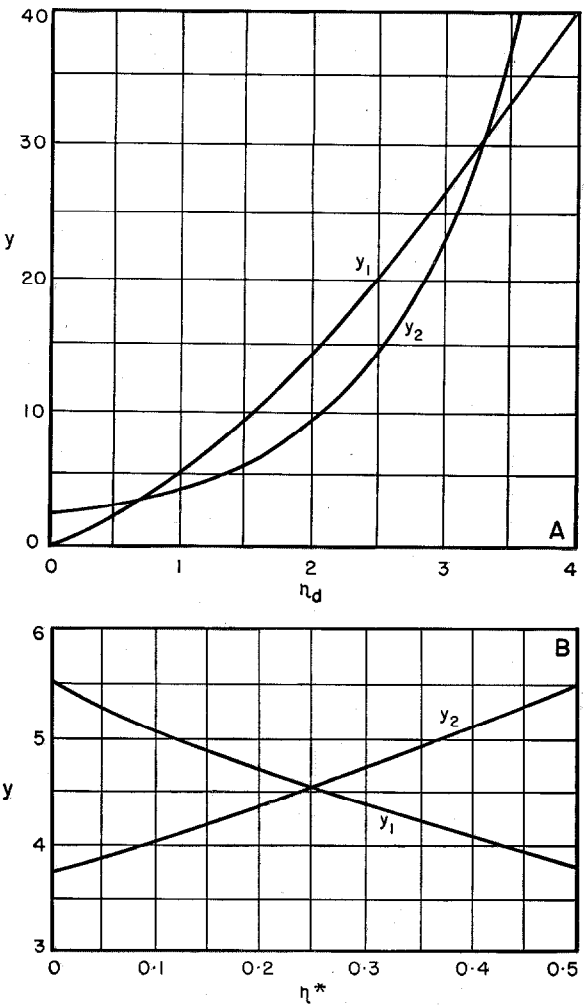


Figure 9—Graphical solutions of (22): A for when η^* is known and B for when η_d is known.

$\epsilon_d=0{\cdot}025$ electron volt; so that if G is assumed to be unity, then $a=3{\cdot}714$.

The example in Figure 9A illustrates the use of the method when the temperatures are required corresponding to a given value of η^*, in this case 0·10. The roots are two values of η_d that with the known value of ϵ_d yield temperatures of 89 and 382 degrees Kelvin as the solutions.

In Figure 9B, the converse process is shown, when a temperature in this case of 290 degrees Kelvin is initially assumed, so that $\eta_d=1{\cdot}00$.

The solution in this case gives a value for η^* of 0·253, which enables n_e to be calculated for this temperature as $2{\cdot}23.10^{19}$ centimetres^{-3}.

The values of η^* and n_e as functions of temperature in Figure 10 were obtained by repeating

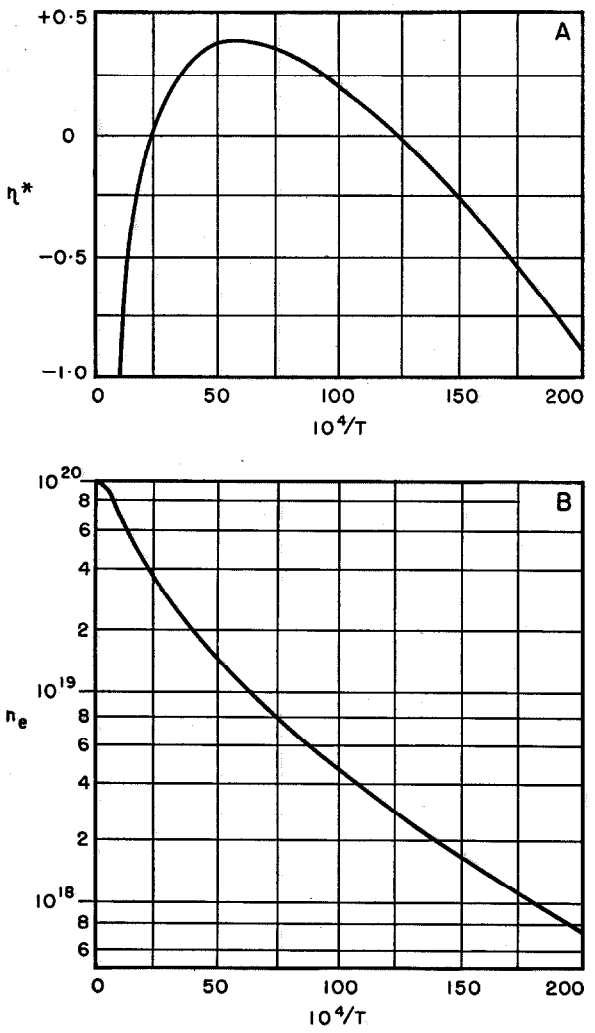


Figure 10—Temperature variation of the properties of a partially degenerate semi-conductor. *A* is for a reduced Fermi level η^* and *B* is for free-electron concentration. T = degrees Kelvin.

this procedure at a number of temperatures. The shape of the curve for n_e should be contrasted with that for a classical semi-conductor given in Figure 7A. Whereas in the latter case, a reliable result for the activation energy could be derived from the slope, an attempt to do so with the results in Figure 10B would lead to serious errors.

4.5 Maximum Value of η^*

It was noted in referring to Figure 9A that there were two roots or two temperatures corresponding with a given value of η^*. As η^* rises, the two roots approach until a limiting case occurs when the two curves in Figure 9A touch at only one point. There is no real solution of (22) for a higher value of η^*.

The form of the behaviour of η^* as a function of temperature is shown in Figure 10A. In the limiting case for Figure 9A, (23) and (24) have equal slopes at the point of contact.

$$(3/2).a\eta_d^{1/2}/F_{1/2}(\eta^*)=\exp(\eta_d+\eta^*),$$

but

$$a.\eta_d^{3/2}/F_{1/2}(\eta^*)=1+\exp(\eta_d+\eta^*),$$

thus

$$(2/3).\eta_d.\exp(\eta_d+\eta^*)=1+\exp(\eta_d+\eta^*).$$

Hence

$$\eta^*_{\max}=-\log_e\{(2\epsilon_d/3kT_{\max})-1\}-\epsilon_d/kT_{\max}. \quad (25)$$

This maximum can be either positive or negative. As N_d increases, both $\eta^*_{\max}$ and $T_{\max}$ increase, whilst an increase in ϵ_d leads to a rise in $T_{\max}$ but a fall in $\eta^*_{\max}$.

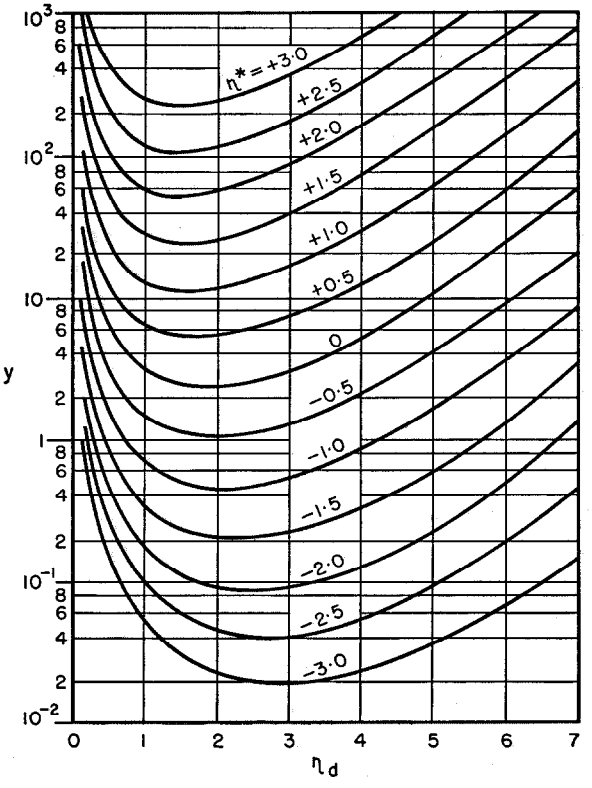


Figure 11—Graphical solutions of (27) for small positive and negative values of η^*. This figure with different symbolism is from Lehovec and Kedesdy[15].

5. *Approximate Evaluation of n_e and ϵ^**

It will be obvious that the method described in Section 4.4 for the evaluation of n_e and ϵ^* (or η^*) applies not only in degenerate cases but is valid for any value of η^*.

However, as the method is very slow and tedious in use, it is not generally employed when other methods are possible. In the classical range, for instance, use of (10) and (11) is preferable, especially if the expressions for reserve or exhaustion semi-conductors apply.

In many more cases, only an approximate solution is required and two recent papers[14,15] give graphical information that is useful in the approximate estimation of n_e and η^* from N_d, ϵ_d, and T, or vice versa. When more accurate results are desired, these approximate solutions greatly reduce the amount of calculation required.

Lehovec and Kedesdy[15] base their curves on (22), except that the temperature-independent parameter, which is characteristic of the semi-

[15] K. Lehovec and H. Kedesdy, "Graphical Determination of the Fermi Level in a Simple Semiconductor," *Journal of Applied Physics*, v. 22, pp. 65–67; January, 1951.

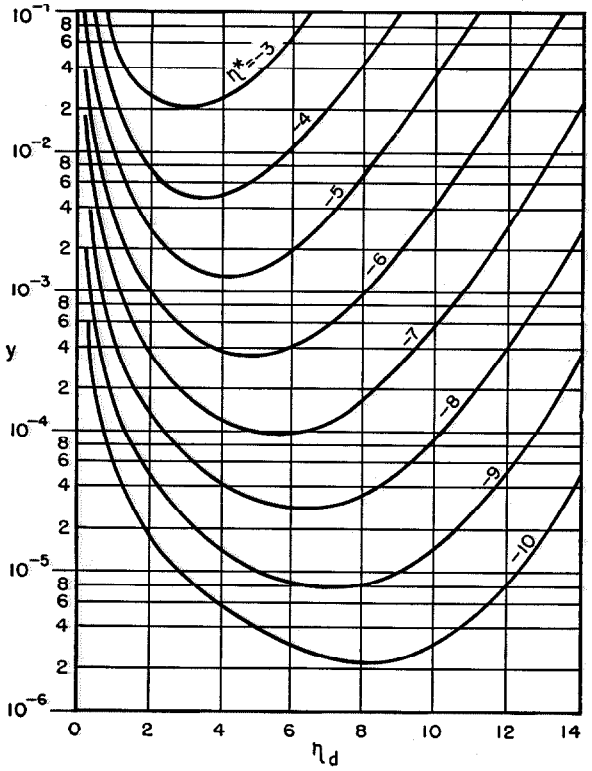


Figure 12—Graphical solutions of (27) for negative values of η^*. From Lehovec and Kedesdy[15].

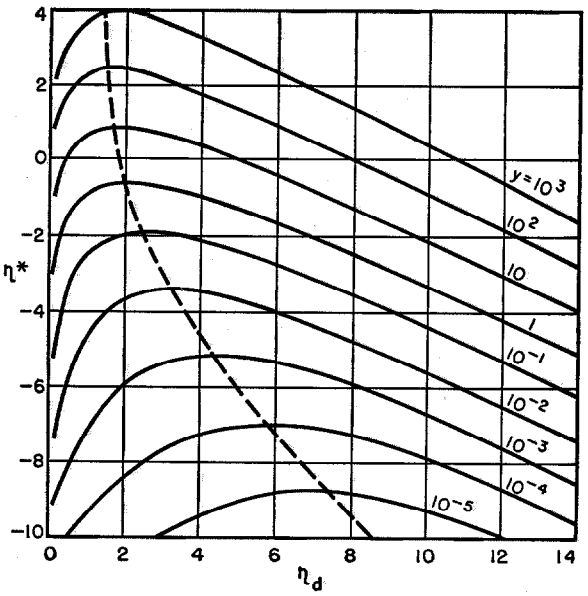


Figure 13—Variation of η^* with η_d for given values of y (when N_d and ϵ_d are both known). From Lehovec and Kedesdy[15].

conductor, differs by a factor of $2/\pi^{1/2}$ from that used in Section 4. By making the substitution

$$y=(2/\pi^{1/2}).a$$
$$=N_d h^3/2(2\pi m_e \epsilon_d)^{3/2}, \qquad (26)$$

then (22) can be rewritten as

$$y=2\pi^{-1/2}.\{1+\exp(\eta_d+\eta^*)\}.F_{1/2}(\eta^*).\eta_d^{-3/2}. \qquad (27)$$

Three diagrams from the Lehovec and Kedesdy paper are reproduced in Figures 11 to 13. In the first two of these, y is plotted as a function of η_d for set values of η^*, whilst in the third, graphs are plotted of η^* as a function of η_d for particular values of y.

Interpolation between the curves will give results that are sufficiently accurate to satisfy many requirements.

When it is required to calculate η^* and n_e, the other parameters being known, the use of these diagrams is perfectly straightforward. An important use, however, is the calculation of N_d and ϵ_d from experimental observations of n_e as a function of temperature. If the observations include the range for which η^* has its maximum, the process is simple, for the magnitude of this maximum together with the aid of Figure 11 or 12 establishes the values of η_d and y, and a

knowledge of T_{max} enables N_d and ϵ_d to be evaluated separately. (The values of η^* are of course obtained from the values of n_e at known temperatures.)

If the experimental results do not cover the range in which η^* has a maximum, the values of η^* are required for at least two temperatures in order to obtain N_d and ϵ_d, whilst results at several other temperatures are useful for substitution in the relevant equations for confirmation.

Figure 14 is a slightly more detailed form of a figure given in the Landsberg paper.[14] This relates $n_e/n_0.G^{3/2}$ with $N_d/n_0.G^{3/2}$ for constant values of η_d and illustrates the variation of the degree of ionisation of a semi-conductor with impurity content at constant temperature.

Thus the curve for $\eta_d=1$ could correspond with room temperature conditions for a semi-conductor with an activation energy of 0·025 electron volt, the abscissa reading $N_d/n_0G^{3/2}=1$ corresponding with $N_d=2{\cdot}69.10^{19}$ centimetres^{-3} for normal carrier mass.

6. *More General Semi-Conductor Band Model*

6.1 General Considerations

In the preceding sections, it has been assumed that the valence-band electrons have been so far below the Fermi level that none are excited into the conduction band at normal temperatures. This is not always the case, and in this section a more complete band model is assumed, as shown in Figure 15. As before, the origin of energy is assumed to be the bottom of the conduction band; and the valence band is now assumed to extend downwards from the level $\epsilon=-\epsilon_i$. The energy ϵ_i is usually referred to as the intrinsic gap. Known values of ϵ_i vary from 0·1 electron volt in grey tin to several electron volts in highly insulating materials.

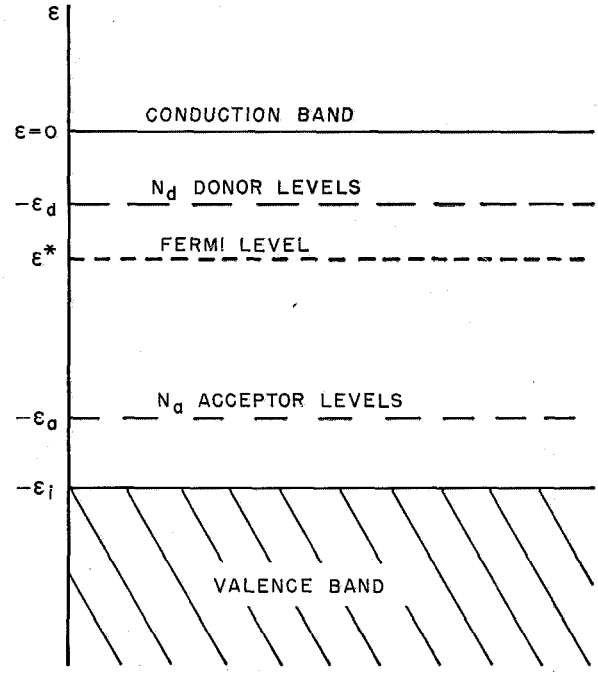


Figure 15—A more complete band model for a semi-conductor.

Two sets of impurities are considered in Figure 15, N_d donor centres at $\epsilon=-\epsilon_d$, each of which can contribute one electron; and N_a acceptors at $\epsilon=-\epsilon_a$, each of

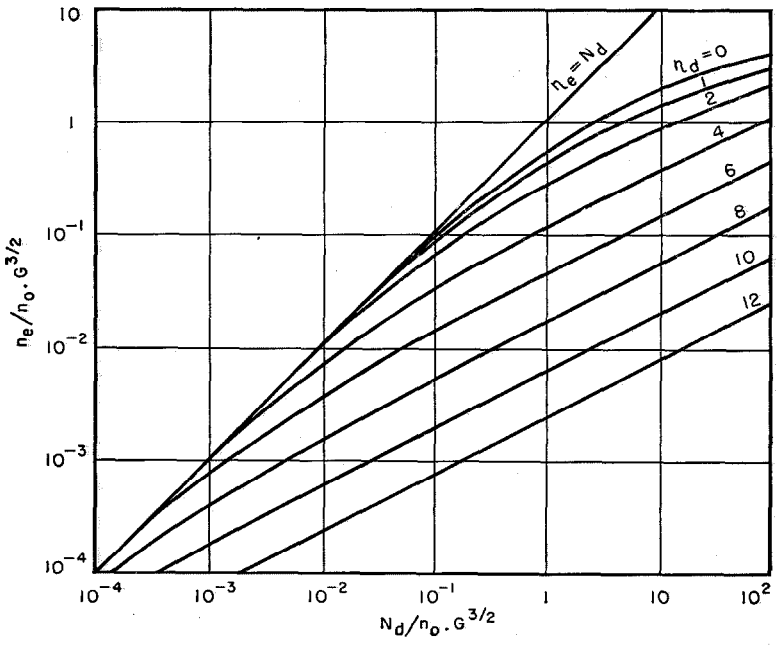


Figure 14—Variation of the degree of ionisation with change of impurity content.

which by accepting an electron can, in effect, donate one hole.

If at $T=0$ degrees Kelvin, the valence-bond band contains N electrons, then the total number of electrons in the system must be $(N+N_d)$. Thus at any temperature

$$\int_{-\infty}^{\infty} n(\epsilon).d\epsilon = \int_{-\infty}^{\infty} f(\epsilon).g(\epsilon).d\epsilon = (N+N_d). \quad (28)$$

The distribution of electrons and holes can be considered in four sections of the energy scale.

6.1.1 Conduction Band, $\epsilon>0$

Let the effective mass of a free electron be m_e, which may be different from that of a hole m_h. Then the free-electron concentration is given by

$$n_e = 4\pi(2m_e kT/h^2)^{3/2} \int_0^{\infty} \frac{\eta^{1/2}.d\eta}{1+\exp(\eta-\eta^*)} = n_0(m_e/m)^{3/2}.F_{1/2}(\eta^*) \quad (29)$$

exactly as in Section 2.

6.1.2 Donor Levels, $\epsilon=-\epsilon_d$

The number of electrons bound in the donor levels is, as in Section 2, given by the product of the density of donor levels with the probability $f(-\epsilon_d)$. This product is

$$(N_d)_e = \frac{N_d}{1+\exp(-\eta_d-\eta^*)} \quad (30)$$

and the number of unoccupied donor levels, i.e., the number of holes in the donor levels, is thus given by

$$(N_d)_h = N_d-(N_d)_e = \frac{N_d}{1+\exp(\eta_d+\eta^*)}. \quad (31)$$

6.1.3 Acceptor Levels, $\epsilon=-\epsilon_a$

Similarly, the number of *electrons* in the acceptor levels is given by

$$(N_a)_e = \frac{N_a}{1+\exp(-\eta_a-\eta^*)} \quad (32)$$

and the number of *holes* in these states (i.e., the number of non-ionised acceptors) is

$$(N_a)_h = N_a-(N_a)_e = \frac{N_a}{1+\exp(\eta_a+\eta^*)}. \quad (33)$$

6.1.4 Valence Band, $\epsilon<-\epsilon_i$

In this band, it is required to know how many electrons have been removed (i.e., the density of free holes n_h); rather than the number of electrons left behind, which must be $(N-n_h)$.

The probability that an electron state contains a hole is

$$f_p(\epsilon) = 1-f(\epsilon).$$

Thus

$$f_p(\epsilon) = \frac{1}{1+\exp\left(\frac{\epsilon^*-\epsilon}{kT}\right)}, \quad (34)$$

whilst the density of states increases parabolically as energy *decreases* in the upper part of the valence-bond band

$$g(\epsilon) = 4\pi(2m_h/h^2)^{3/2}.(-\epsilon-\epsilon_i)^{1/2}, \quad \epsilon<-\epsilon_i.$$

Thus the number of free holes per cubic centimetre n_h is given by

$$n_h = \int_{-\infty}^{\epsilon=-\epsilon_i} g(\epsilon).f_p(\epsilon).d\epsilon = 4\pi\left(\frac{2m_h}{h^2}\right)^{3/2} \int_{-\infty}^{\epsilon=-\epsilon_i} \frac{(-\epsilon-\epsilon_i)^{1/2}\,d\epsilon}{1+\exp\left(\frac{\epsilon^*-\epsilon}{kT}\right)}.$$

The form of the integral is more obvious if the variable is changed to $\lambda=-(\epsilon+\epsilon_i)/kT$. Then the expression for n_h can be rewritten as

$$n_h = 4\pi\left(\frac{2m_h kT}{h^2}\right)^{3/2} \int_0^{\infty} \frac{\lambda^{1/2}.d\lambda}{1+\exp(\lambda+\eta_i+\eta^*)}$$

$$\therefore \quad n_h = n_0(m_h/m)^{3/2}.F_{1/2}(-\eta^*-\eta_i). \quad (35)$$

6.1.5 Electron and Hole Densities in Thermal Equilibrium

By use of the symbols defined in (29) to (35), the integral on the left-hand side of (28) can be expressed as

$$n_e+(N_d)_e+(N_a)_e+(N-n_h) = (N+N_d) \quad (36)$$

$$\therefore \quad n_e+(N_d)_e-n_h-(N_a)_h = N_d-N_a. \quad (37)$$

This important equation states the point that the excess of available electrons for conduction over the number of holes available for conduction must equal the excess of donors over acceptors.

Equation (36) can be also written in the form

$$n_e+(N_a)_e = n_h+(N_d)_h, \quad (38)$$

i.e., the excess of free electrons over free holes is equal to the excess of bound holes over bound electrons. When substitutions are made from

(29), (31), (32), and (35), this becomes,

$$n_0\left(\frac{m_e}{m}\right)^{3/2}.F_{1/2}(\eta^*)+\frac{N_a}{1+\exp(-\eta_a-\eta^*)}$$
$$=n_0\left(\frac{m_h}{m}\right)^{3/2}.F_{1/2}(-\eta^*-\eta_i)+\frac{N_d}{1+\exp(\eta_d+\eta^*)}. \quad (39)$$

This equation expresses the thermal equilibrium condition of a semi-conductor containing donors and acceptors, provided that the impurities are not distributed over ranges of energy. The form of the equation in various cases is considered below.

6.2 Intrinsic Semi-Conductors

In the intrinsic case, there are no impurities, and free carriers are produced by direct excitation of electrons from the valence band to the conduction band. Thus two terms disappear from (38) leaving

$$n_e=n_h=n_i \quad (40)$$

(the subscript i denoting the intrinsic condition), so that (39) can be expressed as

$$n_i=n_0\left(\frac{m_e}{m}\right)^{3/2}.F_{1/2}(\eta^*)$$
$$=n_0\left(\frac{m_h}{m}\right)^{3/2}.F_{1/2}(-\eta^*-\eta_i), \quad (41)$$

i.e.,

$$F_{1/2}(\eta^*)=\left(\frac{m_h}{m_e}\right)^{3/2}.F_{1/2}(-\eta^*-\eta_i). \quad (42)$$

The ratio m_h/m_e is unlikely to be greatly different from unity, thus the two Fermi integrals must be roughly equal.

Thus

$$\eta^*\simeq-\eta_i/2,$$

i.e., the Fermi level is near the centre of the intrinsic gap, being at the exact centre only if $m_e=m_h$.

Since in all known semi-conductors the intrinsic gap is more than $4kT$, then the classical approximations

$$F_{1/2}(\eta^*)=\tfrac{1}{2}\pi^{1/2}\exp\eta^*$$

and

$$F_{1/2}(-\eta^*-\eta_i)=\tfrac{1}{2}\pi^{1/2}\exp(-\eta^*-\eta_i)$$

can be used. When these substitutions are made in (42), then

$$\exp(2\eta^*)=(m_h/m_e)^{3/2}.\exp(-\eta_i)$$
$$\therefore\quad \epsilon^*=\frac{-\epsilon_i}{2}+\left(\frac{3kT}{4}\right).\log_e\left(\frac{m_h}{m_e}\right) \quad (43)$$

and

$$n_i=\tfrac{1}{2}\pi^{1/2}n_0\left(\frac{m_em_h}{m^2}\right)^{3/4}.\exp\left(\frac{-\epsilon_i}{2kT}\right) \quad (44)$$

$$=(N_eN_h)^{1/2}\exp\left(\frac{-\epsilon_i}{2kT}\right), \quad (44A)$$

where

$$N_e=2(2\pi m_ekT/h^2)^{3/2}$$

and

$$N_h=2(2\pi m_hkT/h^2)^{3/2}$$

are equivalent forms for the conduction and valence bands respectively of the symbol N_e used in Section 3 for classical n-type semi-conductors. These symbols can be considered as representing 'effective numbers of states' at $\epsilon=0$ and $\epsilon=-\epsilon_i$, which can replace the distributed ranges of levels in the conduction and valence bands when the classical limit of the Fermi distribution is reached.

It will be shown later that when impurities are introduced, changing both n_e and n_h, the relationship

$$n_en_h=n_i^2=(N_eN_h)\exp(-\epsilon_i/kT) \quad (45)$$

is still valid unless the Fermi level approaches within $2kT$ of either band, when the classical approximations for the Fermi integrals are no longer reliable.

It follows from (44) and (45) that

$$\frac{n_en_h}{T^3}=4\left(\frac{2\pi mk}{h^2}\right)^3.\left(\frac{m_em_h}{m^2}\right)^{3/2}.\exp\left(\frac{-\epsilon_i}{kT}\right) \quad (46)$$
$$=2{\cdot}33.10^{31}\left(\frac{m_em_h}{m^2}\right)^{3/2}.\exp\left(\frac{-\epsilon_i}{kT}\right).$$

Thus a semi-logarithmic plot of n_en_h/T^3 against $1/T$ for either an intrinsic or an extrinsic (impurity) semi-conductor should have a slope of $-\epsilon_i/k$ and an intercept corresponding with the coefficient just quoted.

In practice, this procedure usually leads to a value of ϵ_i, which agrees with the results of independent experiments, but the intercept may differ from the above value by an amount that is too large to be accountable by deviations of carrier masses from m.

Pearson and Bardeen[16] found for specimens of silicon a coefficient 34 times larger than N_eN_h/T^3. They explain this on the assumption that the width of the intrinsic gap depends on temperature.

[16] G. L. Pearson and J. Bardeen, "Electrical Properties of Pure Silicon and Silicon Alloys Containing Boron and Phosphorus," *Physical Review*, v. 75, pp. 865–883; March, 1949.

Assume that $\epsilon_i = \alpha - \beta T$, then

$$\frac{n_e n_h}{T^3} = \left\{ N_e N_h . T^{-3} . \exp\left(\frac{\beta}{k}\right) \right\} \exp\left(\frac{-\alpha}{kT}\right). \quad (47)$$

Experimental measurements of β for silicon by other methods show that the factor $\exp(\beta/k)$ is of the correct order of magnitude to explain the results of Pearson and Bardeen.

The temperature shift of ϵ_i is of the order of 5.10^{-4} electron volt per degree centigrade for both germanium and silicon and is believed to be considerably greater for lead sulphide, lead selenide, and lead telluride.

6.3 Semi-Conductors with *n*-Type Impurities

In this case, $N_a = 0$, and thus

$$n_e = \frac{N_d}{1+\exp(\eta^*+\eta_d)} + n_0\left(\frac{m_e}{m}\right)^{3/2} . F_{1/2}(-\eta^*-\eta_i). \quad (48)$$

When $(-\eta^*-\eta_i) \ll \eta^*$, the second term on the right-hand side of (48), which is n_h, becomes negligibly small and the equation is identical in form with (7A), derived in Section 3.

Thus the analysis of Sections 2 to 4 applies when the Fermi level is well above the centre of the intrinsic gap.

This is the case when N_d is not too small and $\epsilon_d \ll \epsilon_i$. If these restrictions are not valid and the Fermi level is not far above the centre of the intrinsic gap, then (6A) and (48) must be solved.

This would be awkward if either $F_{1/2}(\eta^*)$ or $F_{1/2}(-\eta^*-\eta_i)$ were in the semi-degenerate region. Fortunately, with one exception, this is not the case and complete analysis of (6A) and (48) is required only when

$$F_{1/2}(\eta^*) = \tfrac{1}{2}\pi^{1/2} \exp(\eta^*)$$

and

$$F_{1/2}(-\eta^*-\eta_i) = \tfrac{1}{2}\pi^{1/2} \exp(-\eta^*-\eta_i).$$

The exception is grey tin, for which $\epsilon_i = 0{\cdot}1$ electron volt. Thus at temperatures slightly below the transition point of 13 degrees centigrade, ϵ_i is only a little more than $4kT$, and if ϵ^* moves more than a small fraction of kT from the mid-point of the intrinsic gap, then the classical approximation (9) cannot be accurately valid for both electrons and holes.

For all other known semi-conductors, ϵ_i is sufficiently large to ensure that either classical approximations can be made in evaluating n_e and n_h or that one of the two sets of carriers is present in a negligibly small quantity. Thus $n_e = N_e \exp(\eta^*)$ and $n_h = N_h \exp(-\eta^*-\eta_i)$. Hence $n_e n_h = (N_e N_h) \exp(-\eta_i)$, the right-hand side of which is identical with that of (45). This demonstrates the equality

$$n_e n_h = n_i^2,$$

which was quoted earlier.

When, for an *n*-type semi-conductor, n_h is not negligible by comparison with n_e, equations (6A) and (48) must be equated

$$n_e = N_e \exp(\eta^*) = \frac{N_d}{1+\exp(\eta^*+\eta_d)} + N_h \exp(-\eta^*-\eta_i). \quad (49)$$

The general solution of this cubic equation is complicated, but two limiting cases are more amenable to treatment.

6.3.1 Reserve Semi-Conductors

When the Fermi level is more than $2kT$ above the donors, by neglecting unity compared with $\exp(\eta^*+\eta_d)$, (49) becomes

$$n_e = N_e \exp(\eta^*) = N_d \exp(-\eta^*-\eta_d) + N_h \exp(-\eta^*-\eta_i) \quad (50)$$

$$\therefore \quad n_e^2 = N_d N_e \exp(-\eta_d) + N_e N_h \exp(-\eta_i)$$

and

$$n_e = \{N_d N_e \exp(-\eta_d) + n_i^2\}^{1/2}. \quad (51)$$

This equation is trivial unless the impurity activation energy is at least half the intrinsic gap. For if $\epsilon_d < \epsilon_i/2$, and ϵ^* is at least $2kT$ above $-\epsilon_d$, then n_e will be more than 7 times as great as n_i, and n_i^2 in (51) will be negligibly small. In such a case, (51) reduces to the form of (11A), derived in Section 3 for a simple classical *n*-type semi-conductor.

When $\epsilon_i/2 < \epsilon_d < \epsilon_i$, the contributions of the donors and the valence band towards the conduction-electron concentration in the upper band may be comparable. The temperature at which the contributions are equal depends on N_d and ϵ_d, for when this is the case, $n_e = 2n_h$ and the two terms on the right-hand side of (50) are equal.

$$N_d \exp(-\eta^*-\eta_d) = N_h \exp(-\eta^*-\eta_i),$$

thus

$$\exp(\eta_i - \eta_d) = N_h/N_d$$

and the temperature at which this condition is satisfied is

$$T = \frac{\epsilon_i - \epsilon_d}{k . \log_e(N_h/N_d)}. \quad (52)$$

At temperatures greater than this, though the degree of ionisation of the donors increases, n_h increases more rapidly and the behaviour of the semi-conductor approaches that of the pure intrinsic material.

As ϵ_d increases towards ϵ_i, the transition temperature decreases, whilst an increased value of N_d has the opposite effect.

Donor centres with activation energies of more than $\epsilon_i/2$ are believed to occur in a number of semi-conductors, due to interstitial atoms. Such semi-conductors should show a transition towards intrinsic conduction at high temperatures governed by (51). However, quantitative work on most of these materials, which are usually oxides, sulphides, etc., is complicated by the fact that the measured activation energy usually depends on the impurity content and the previous thermal history. Much of the uncertainty in these materials is due to the presence of intergranular contact resistance.

Interstitial atoms of germanium in the covalent germanium lattice are also believed to act as donor centres with activation energies of about 0·60 electron volt, compared with an intrinsic gap of $\epsilon_i = 0{\cdot}75$ electron volt. Thus these donors are well below the midpoint between the bands, and the same behaviour should be found here. However, as far as is known, interstitial donors in germanium are only found in conjunction with vacant lattice sites (which act as acceptors), when the germanium is subjected to nucleon bombardment or thermal shock. The statistics of semi-conductors containing defect donors and acceptors are discussed in Section 6.5.2.

6.3.2 Exhaustion Semi-Conductors

With donor activation energies less than half the intrinsic gap, the contribution of the valence band towards n_e can only be comparable with that of the donors if the Fermi level lies near to or below the donor levels; for the influence of the full band can be ignored if ϵ^* is more than $2kT$ higher than the midpoint of the bands.

If the Fermi level is assumed to be at least $2kT$ below the donors, these are almost entirely ionised and (49) may be written in the form,

$$n_e = N_e \exp(\eta^*) = N_d + N_h \exp(-\eta^* - \eta_i) \quad (53)$$
$$= N_d + n_h$$

now

$$n_e n_h = n_i^2$$

thus

$$n_e = n_i^2/n_h$$
$$= n_i^2/(n_e - N_d).$$

This quadratic equation in n_e has the solution

$$n_e = \tfrac{1}{2}N_d + (\tfrac{1}{4}N_d^2 + n_i^2)^{\frac{1}{2}}. \quad (54)$$

This result may appear at first sight rather surprising. It is not always appreciated that when a few donors are added to intrinsic material, although all the donors are ionised, only half the electrons so produced go into the conduction band, the other half serving to reduce the number of free holes in the valence band.

As N_d increases, n_h falls towards zero and n_e approaches N_d asymptotically in the way shown

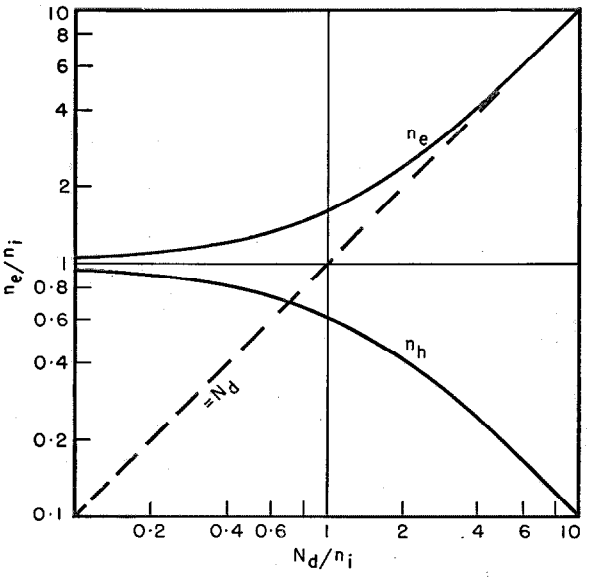


Figure 16—Variation of free-electron and free-hole densities with donor content, when n_i is not negligibly small.

in Figure 16 for the range of donor content in which n_e and n_i are comparable in magnitude. The contributions of the donors and the valence band towards the free-electron concentration are equal if

$$n_e = N_d + n_h = 2n_h.$$

This condition is satisfied if $N_d = 2^{-\frac{1}{2}}n_i$.

The behaviour given by (54) is appropriate to high-resistivity n-type germanium at room temperature, for the chemical n-type impurities in this semi-conductor are all ionised from quite a small temperature upwards. At room temperature, the Fermi level is not very far above $-\epsilon_i/2$ and germanium with a resistivity of about 5 ohm-

centimetres at this temperature behaves intrinsically above about 360 degrees Kelvin (90 degrees centigrade).

6.3.3 Transition from Extrinsic to Intrinsic Conditions

The various expressions derived in the previous sections for the free-electron concentration n_e under different sets of conditions can be summarised as follows.

A. In *all* cases

$$n_e = n_0 G^{3/2} . F_{1/2}(\eta^*). \qquad (6A)$$

B. When $\epsilon_d > \epsilon_i/2$, the conditions are always classical at all temperatures and for any impurity content.

a. At low temperatures

$$n_e = (N_e N_d)^{1/2} \exp(-\eta_d/2). \qquad (11A)$$

b. As temperature rises towards the intrinsic region, this must be generalised to

$$n_e = \{N_d N_e \exp(-\eta_d) + n_i^2\}^{1/2}. \qquad (51)$$

c. At still higher temperatures, the impurities have little effect and

$$n_e = n_i = (N_e N_h)^{1/2} \exp(-\eta_i/2). \qquad (44A)$$

C. When $\epsilon_d < \epsilon_i/2$, but $N_d < (N_e/10)\{1+\exp(\eta_d)/10\}$, so that the conditions are classical at all temperatures, then:

a. At very low temperatures, very few donors are ionised and

$$n_e = (N_d N_e)^{1/2} \exp(-\eta_d/2). \qquad (11A)$$

b. At slightly higher temperatures, when an appreciable fraction of the donors is ionised, n_e is found as the positive root of the quadratic equation

$$n_e^2 = (N_d - n_e) N_e \exp(-\eta_d). \qquad (11)$$

c. At still higher temperatures, ionisation of the donors becomes complete and

$$n_e = N_d. \qquad (13)$$

d. As temperature rises, the Fermi level approaches the energy $-\epsilon_i/2$, and the effect of the valence band becomes appreciable, thus

$$n_e = \tfrac{1}{2} N_d + (\tfrac{1}{4} N_d^2 + n_i^2)^{1/2}. \qquad (54)$$

It should be noted that if ϵ_d is only a little smaller than $\epsilon_i/2$, then the full band may become important before ionisation of the donors is complete. It is obvious that in this case (52) can be expressed in the form

$$n_e = \frac{N_d}{1+(n_e/N_e)\exp(\eta_d)} + \frac{n_i^2}{n_e}. \qquad (55)$$

e. At a sufficiently high temperature, the material becomes intrinsic and

$$n_e = n_i = (N_e N_h)^{1/2} \exp(-\eta_i/2). \qquad (44A)$$

D. When $\epsilon_d < \epsilon_i/2$, but $N_d > (N_e/10)\{1+\exp(\eta_d)/10\}$ for a part of the temperature range, then for this part n_e must be found as the solution of

$$n_e = n_0 G^{3/2} . F_{1/2}(\eta^*) \qquad (6A)$$

and

$$n_e = \frac{N_d}{1+\exp(\eta_d+\eta^*)}. \qquad (7A)$$

Outside this part of the temperature range, the expressions given in sub-section *C* are valid.

The expressions given in sub-sections *C* and *D* can be identified with various parts of the curves in Figure 17, which shows the variation of n_e with $1/T$ for temperatures from 33 degrees Kelvin upwards in a semi-conductor with a band structure rather similar to that of germanium.

In the calculations for this diagram, it was assumed that both electrons and holes have the normal electronic mass. An intrinsic gap of 0·75 electron volt and a donor activation energy of 0·014 electron volt were used, so that in (27) $y = 10^{-19} N_d$.

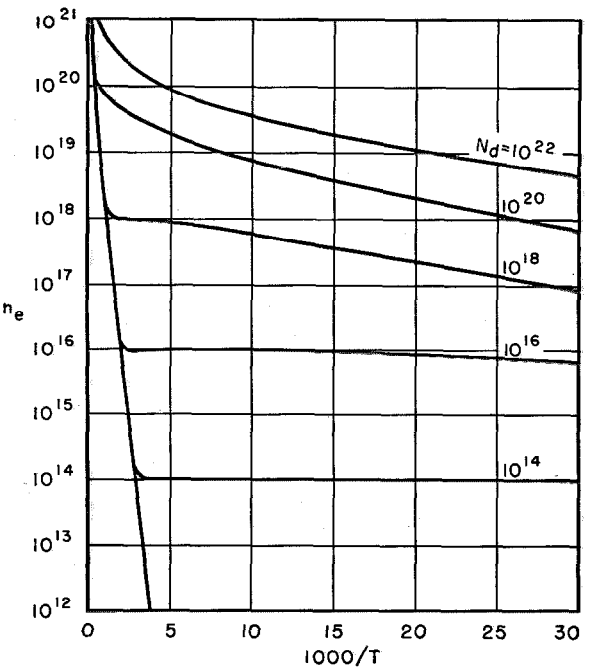


Figure 17—Variation of free-electron concentration with temperature for a semi-conductor with a band structure similar to that of germanium. T=degrees Kelvin and n_e is in centimetres^{-3}.

The model chosen differs from germanium in a number of minor ways.

A. The intrinsic gap of germanium is slightly temperature dependent.

B. The mass of a free hole in germanium is greater than that of a free electron. This slightly modifies the intrinsic curve.

C. The donor activation energy in germanium generally changes with donor concentration, and for N_d very large, the activation energy is reduced effectively to zero, the behaviour becoming that of a semi-metal. (See Section 6.6.) Thus the curves for the highest donor concentrations in Figure 17 have a much larger temperature dependence than is found in germanium itself.

Nevertheless, the results of Figure 17 can be taken as broadly representing the behaviour of germanium with a wide range of impurity contents.

The degree of ionisation is complete at room temperature for donor concentrations up to 10^{18} centimetres^{-3}, whilst with the high-purity material required for most germanium devices, the donors are almost entirely ionised down to liquid hydrogen temperatures.

6.4 Semi-Conductors with *p*-Type Impurities

All the expressions developed for *n*-type semi-conductors are directly applicable to *p*-type conconductors, for the density of states in the valence band increases parabolically as *electron* energy decreases. Thus the density of states in this band increases as *hole* energy *increases*, since electron and hole energies are obviously opposite in sign.

If the arbitrary zero of energy for a *p*-type semi-conductor is assumed to be at the apex of the valence band and if hole energies are considered, then the expressions for *n*- and *p*-type semi-conductors are identical in form.

The similarity of processes in *n* and *p* types of semi-conductors is of great assistance, but the analogy should be treated with caution since some processes in one type of conductor have no counterpart in the other.

Thus, if a semi-conductor absorbs an incident photon, the energy could be used to expel an electron into free space surrounding the material. However, since a hole is an absence of a particle, the corresponding process of hole emission can not exist. On the other hand, holes can be created in the valence band of a semi-conductor by removal of electrons to an adjacent electrode, and it is legitimate to regard this as a process of hole injection from the electrode to the semi-conductor.

The terms electron injection and extraction should only be used when a transfer occurs between the conduction band and an electrode.

6.5 Semi-Conductors Containing Both Donors and Acceptors

All the previous analysis has been carried out on the assumption that impurities of only one sign were present. In any practical case, this is an impossible condition. Any *n*-type semi-conductor inevitably contains at least a few acceptor impurities, and similarly any *p*-type semi-conductor must be expected to have some donors.

Though these impurities of anomalous sign may have a negligible influence on the properties at normal temperatures, the low-temperature properties can be considerably affected by them.

For example, consider an *n*-type classical semi-conductor. At low temperatures under reserve conditions, the Fermi level is determined by

$$\epsilon^* = \frac{-\epsilon_d}{2} + \frac{kT}{2} \log_e \frac{N_d}{N_e}$$

and as T tends towards zero, the Fermi level approaches the level $-\epsilon_d/2$.

However, if a small number $N_a < N_d$ acceptors are now considered to be situated near the valence band, there will be a strong tendency for electrons to be trapped in the acceptor levels, leaving only $(N_d - N_a)$ electrons in the donor levels and conduction band. Even at the lowest temperatures, the donor levels remain incompletely filled. But the Fermi probability of occupation

$$f(\epsilon) = \frac{1}{1+\exp\left(\frac{\epsilon-\epsilon^*}{kT}\right)}$$

has a value intermediate between zero and unity only for energies within a few kT of ϵ^*, hence as T tends to zero, the Fermi level must approach that of the donor levels themselves.

A number of sets of conditions for semi-conductors with both donors and acceptors are discussed by Huttner et al.,[12] and the application of

the theory to germanium with chemical impurities is discussed by Shockley.[17] Of the two cases to be discussed, the first is similar to that of Shockley, whilst the second is of interest also in connection with germanium but under different circumstances.

6.5.1 Chemical Donors and Acceptors in a Semi-Conductor Similar to Germanium

Consider a semi-conductor containing $N_d=(d+x)$ donors just below the conduction band and $N_a=x$ acceptors just above the valence band. Then all the acceptors are permanently ionised, and d electrons are available for excitation into the conduction band in the impurity range.

Hence, (39) becomes

$$n_e=n_0G^{3/2}.F_{1/2}(\eta^*)=\frac{d+x}{1+\exp(\eta_d+\eta^*)}-x.$$

Assuming that classical statistics apply, then

$$n_e=N_e\exp(\eta^*)=\frac{d+x}{1+(n_e/N_e)\exp(\eta_d)}-x \qquad (56)$$

$$\therefore\ n_e^2+n_e\{x+N_e\exp(-\eta_d)\}=dN_e\exp(-\eta_d). \qquad (57)$$

Consider the effect on n_e of increasing the donor and acceptor concentrations by equal amounts at constant temperature. By differentiating n_e with respect to x, we have from (57)

$$\left(\frac{\partial n_e}{\partial x}\right)_T=\frac{-n_e}{x+2n_e+N_e\exp(-\eta_d)}.$$

This must always be negative, and the effect of adding impurities of both signs is thus one of reducing the free-carrier density. Since the saturation value at high temperatures is $n_e=d$, independent of x, then n_e must drop off more sharply as temperature falls in the presence of impurities of both signs.

At low temperatures, when $n_e\ll x$, (56) reduces to

$$x=\frac{d+x}{1+(n_e/N_e)\exp(\eta_d)},$$

i.e.,

$$n_e=(dN_e/x)\exp(-\epsilon_d/kT), \qquad (58)$$

which changes twice as rapidly with temperature as

$$n_e=(dN_e)^{1/2}\exp(-\epsilon_d/2kT)$$

for a semi-conductor with donors alone in the reserve region.

This behaviour is demonstrated by the curves in Figure 18 for a semi-conductor with a donor

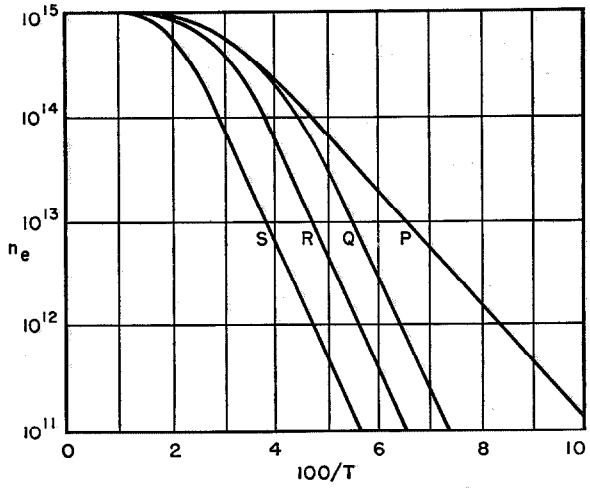


Figure 18—Variation of free-electron concentration with temperature for a semi-conductor with $10^{15}+x$ donors and x acceptors per cubic centimetre. Curves P, Q, R, and S are for values of x of 0, 10^{14}, 10^{15}, and 10^{16} centimetres^{-3} respectively. T = degrees Kelvin and n_e is in centimetres^{-3}.

activation energy $\epsilon_d=0{\cdot}020$ electron volt. An initial donor concentration of 10^{15} centimetres^{-3} is assumed with additions of $x=10^{14}$, 10^{15}, 10^{16} donors and acceptors.

From (58) and (10), the Fermi level at low temperatures when both donors and acceptors are present is determined by

$$\exp(\eta^*)=n_e/N_e=(d/x)\exp(-\eta_d)$$
$$\therefore\quad \epsilon^*=-\epsilon_d+kT\log_e(d/x). \qquad (59)$$

This linear variation of ϵ^* with T at low temperatures is shown by the curves in Figure 19. For curve R, where $d=x=10^{15}$ centimetres^{-3}, the Fermi level is invariant with respect to T for the range in which (59) is valid, up to about 25 degrees Kelvin. As temperature rises, the Fermi level approaches a common value for all values of x.

6.5.2 Effect of Lattice Defects on a Semi-Conductor Containing Chemical Impurities

Any crystalline solid contains a number of lattice defects in thermal equilibrium (Frenkel,[18]

[17] W. Shockley, "Electrons and Holes in Semiconductors," 1st Edition, D. van Nostrand Company, New York, New York; 1950.

[18] J. Frenkel, "Über die Wärmebewegung in festen und flüssigen Körpern," *Zeitschrift für Physik*, v. 35, pp. 652–669; February, 1926.

Wagner and Schottky[19]), which can act as donors and acceptors, and certain treatments can result in concentrations of these defects far larger than the thermal equilibrium values.[20] This situation can be produced by cooling the solid rapidly from a high temperature, for the equilibrium density of defects increases rapidly as temperature rises, and on quenching large numbers of lattice imperfections can be 'frozen in.'

Bombardment with high-energy nucleons can also produce lattice dislocations of similar form, which 'heal out' at normal temperatures after only a very long time.

In semi-conductors with ionic lattices, usually only the positive metallic ions are found out of position, interstitial ions giving rise to donor levels, whilst vacant positive-ion lattice sites act as acceptors.

Semi-conductors with atomic lattices, such as germanium and silicon, are also affected by lattice defects, with atoms in interstitial positions forming donors, and vacant sites, acceptors. However, the activation energies of these defects are very much larger than those of chemical impurities, which leads to interesting results.

Consider a semi-conductor with N_d chemical donors per cubic centimetre at a small activation energy ϵ_d below the conduction band.

If equal numbers X of defect donors and acceptors are now created, ϵ_D, ϵ_A, at energies well below the conduction band, at first the interstitial donors will be ionised by the equal number of acceptors and will tend to capture electrons from the conduction band governed by

$$n_e=\frac{N_d}{1+\exp(\eta_d+\eta^*)}+\frac{X}{1+\exp(\eta_D+\eta^*)}+\frac{X}{1+\exp(-\eta_A-\eta^*)}+n_h. \quad (60)$$

If N_d is small, so that the donor levels are exhausted, and X is smaller still, then (60) may be approximated by

$$n_e=N_e\exp(\eta^*)=N_d-X. \quad (61)$$

Hence

$$\epsilon^*=-kT\log_e\{N_e/(N_d-X)\} \quad (62)$$

and both n_e and the Fermi level fall as X increases.

What happens when X becomes very large depends on the values of ϵ_D and ϵ_A. In silicon, it is believed that ϵ_D is slightly less than half the intrinsic gap, whilst ϵ_A is a little more than $\epsilon_i/2$. As X increases, the Fermi level tends towards $-(\epsilon_D+\epsilon_A)/2$, and since this is almost $-\epsilon_i/2$, the behaviour approaches that of intrinsic silicon. If silicon with p-type chemical impurities had been the starting material, the fall in carrier concentration towards the intrinsic value would have followed the same course.

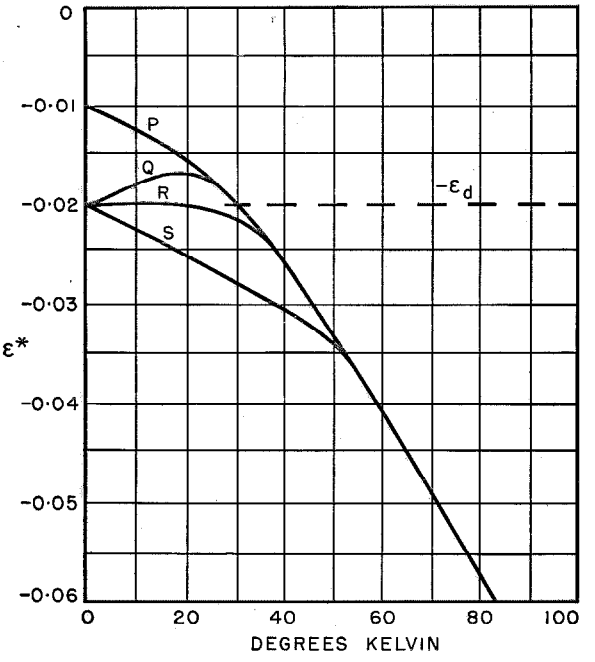


Figure 19—Variation of Fermi level with temperature for a semi-conductor containing $10^{15}+x$ donors and x acceptors per cubic centimetre. Curves P, Q, R, and S are for values of $x=0$, 10^{14}, 10^{15}, and 10^{16} centimetres^{-3} respectively.

In germanium, the situation is rather different. On creating pairs of lattice defects, the free-hole concentration in impure p-type germanium decreases, but the asymptotic value for very large values of X is considerably greater than the intrinsic value. Highly pure p-type germanium becomes more conducting as X increases, whilst germanium that is initially n-type becomes p-type for large values[21] of X.

[19] C. Wagner and W. Schottky, "Theorie der geordneten Mischphasen," *Zeitschrift für physikalische Chemie*, v. B11, pp. 163–210; December, 1930.

[20] B. R. A. Nijboer, "On the Theory of Electronic Semiconductors," *Proceedings of the Physical Society*, v. 51, pp. 575–584; July, 1939.

[21] "Semi-Conducting Materials," Butterworths, London; 1951: See K. Lark-Horovits, "Nucleon-Bombarded Semi-Conductors," pp. 47–69.

This can be explained on the assumption that both ϵ_D and ϵ_A are greater than $\epsilon_i/2$, and as the limiting position of ϵ^* is midway between $-\epsilon_D$ and $-\epsilon_A$, this must be nearer the valence band than the conduction band. Thus

$$n_h = n_i \exp(-\eta^* - \eta_i/2) \tag{63}$$

is greater than

$$n_e = n_i \exp(\eta^* + \eta_i/2). \tag{64}$$

James and Lehman[22] have made some calculations on the changes in the Fermi level for a semi-conductor model similar to germanium at 293 degrees Kelvin, on the assumption that chemical impurities have very small activation energies, free electrons and holes both have the normal electronic mass, and $\epsilon_D = 0{\cdot}60$ electron volt and $\epsilon_A = 0{\cdot}70$ electron volt compared with $\epsilon_i = 0{\cdot}75$ electron volt.

6.6 Semi-Metals

The group of materials described as semi-metals represent a modification of the simple band theory of semi-conductors. Whilst the term semi-metal is sometimes loosely applied to semi-conductors in which impurities have small activation energies, so that the free-carrier concentration is relatively independent of temperature, this use of the word is unfortunate. In the true semi-metal, the impurity-activation energy is completely absent, and the impurity levels are continuous with the normal lattice levels of the valence or conduction bands.

A detailed discussion of the way this situation can arise would lie outside the proper scope of this paper. It is sufficient to note that interaction between impurity centres in highly impure materials could cause a coalescence of the impurity levels with the normal band levels. This situation is believed to occur in silicon[16] and cadmium oxide[11] amongst other materials.

Consider an n-type case of a semi-metal. As before, n- and p-type cases are mirror images.

A fundamental difference exists between a semi-conductor with a very small activation energy and a semi-metal, in that the impurity levels in the former are bound states in which electrons are not free, whilst in a semi-metal these states are no longer bound ones. Electrons in these states can pass freely into non-localised states *at the same energy*, thus the impurity levels serve merely to increase the density of free states at the bottom of the conduction band.

Whereas for a semi-conductor, the number of free electrons was given by

$$n_e = n_0 G^{3/2} . F_{1/2}(\eta^*)$$

with $N_d/\{1+\exp(-\eta_d-\eta^*)\}$ electrons bound in the impurity states, in a semi-metal both these sets of carriers are free and

$$n_e = N_d = \frac{N_d}{1+\exp(-\eta^*)} + n_0 G^{3/2} . F_{1/2}(\eta^*).$$

Hence

$$n_e = N_d = n_0 G^{3/2} . F_{1/2}(\eta^*)\{1+\exp(\eta^*)\}. \tag{65}$$

The value of η^* corresponding with a given value of n_e is considerably larger in a degenerate case when (6A) is incorrectly used instead of the appropriate expression (65). This can possibly account for the very low values of free carrier mass ($\simeq 0{\cdot}09\ m$) reported for cadmium oxide by Wright.[11]

7. *Conclusion*

An attempt has been made to review the present state of this part of semi-conductor theory rather than to extend the analysis in any particular direction. It is hoped that a number of results of the usual theory that are not always appreciated have been clarified.

In dealing with actual materials, one is interested not only in the carrier concentrations and the Fermi level, but also in the observable quantities that depend on them, such as conductivity, thermo-electric power, Hall effect, etc. Determination of n_e from experimental results must always depend on the interpretation of these effects. When lattice scattering alone is important, a recent paper by Wright[23] relates many of these effects to the reduced Fermi level.

---

[22] "Semi-Conducting Materials," Butterworths, London; 1951: See H. M. James and G. W. Lehman, "Fermi Levels in Bombarded Semi-Conductors," pp. 74–77.

[23] R. W. Wright, "The Effect of Mean Free Path of Electrons on the Electrical Properties of Non-Metals," *Proceedings of the Physical Society*, v. A64, pp. 984–999; November, 1951.

8. *Appendixes*

8.1 Numerical Values of Some Functions

The numerical values given in Tables 1 and 2 have been calculated using the values for the fundamental physical constants quoted by Bearden and Watts.[9]

TABLE 1

$$n_0 = 4\pi(2mk/h^2)^{3/2}T^{3/2} = 5{\cdot}451 \,.\, 10^{15}\, T^{3/2}, \text{ centimetres}^{-3}$$

$$n_0 \text{ at } T = 290 \text{ degrees Kelvin} = 2{\cdot}692 \,.\, 10^{19}, \text{ centimetres}^{-3}$$

$$N_e = 2(2\pi m_e kT/h^2)^{3/2} = 4{\cdot}831 \,.\, 10^{15}\, G^{3/2}T^{3/2}, \text{ centimetres}^{-3}$$

$$N_e \text{ at } T = 290 \text{ degrees Kelvin} = 2{\cdot}386 \,.\, 10^{19}\, G^{3/2}, \text{ centimetres}^{-3}$$

$$(n_e)_{\eta^*=0} = n_0 F_{1/2}(\eta^*)_{\eta^*=0} \,.\, G^{3/2} = 3{\cdot}696 \,.\, 10^{15}\, G^{3/2}T^{3/2}, \text{ centimetres}^{-3}$$

$$(n_e)_{\eta^*=0} \text{ at } T = 290 \text{ degrees Kelvin} = 1{\cdot}825 \,.\, 10^{19}\, G^{3/2}, \text{ centimetres}^{-3}$$

$$(n_e)_{\eta^*=-2} = n_0 F_{1/2}(\eta^*)_{\eta^*=-2} G^{3/2} = 6{\cdot}246 \,.\, 10^{14}\, G^{3/2}T^{3/2}, \text{ centimetres}^{-3}$$

$$(n_e)_{\eta^*=-2} \text{ at } T = 290 \text{ degrees Kelvin} = 3{\cdot}085 \,.\, 10^{18} G^{3/2}, \text{ centimetres}^{-3}$$

$$n_{\text{deg}} = (8\pi/3)(2m_e k/h^2)^{3/2} \,.\, T_{\text{deg}}^{3/2} = 3{\cdot}634 \,.\, 10^{15}\, G^{3/2}T_{\text{deg}}^{3/2}, \text{ centimetres}^{-3}$$

$$T_{\text{deg}} = (3/\pi)^{2/3}(h^2/8m_e k) \,.\, n_{\text{deg}}^{2/3} = 4{\cdot}231 \,.\, 10^{-11}\, G^{-1} n_{\text{deg}}^{2/3}, \text{ degrees Kelvin}$$

$$a = N_d h^3/4\pi(2m_e\epsilon_d)^{3/2} = 1{\cdot}468 \,.\, 10^{-22}\, N_d/(G\epsilon_d)^{3/2} \text{ (for } \epsilon_d \text{ in electron volts)}$$

$$y = 2\pi^{-1/2} \,.\, a = 1{\cdot}653 \,.\, 10^{-22}\, N_d/(G\epsilon_d)^{3/2} \text{ (for } \epsilon_d \text{ in electron volts)}$$

$$\eta = \epsilon/kT = 11{,}600\ \epsilon/T \text{ (for } \epsilon \text{ in electron volts)}$$

When

$$n = \text{constant} \exp(-\epsilon/2kT),$$

then

$$\epsilon = -3{\cdot}97 \,.\, 10^{-4} \,.\, \frac{d(\log_{10} n)}{d(1/T)}, \text{ electron volts.}$$

TABLE 2

n_0 AND N_e BETWEEN 0 AND 500 DEGREES KELVIN

Degrees Kelvin	n_0	N_e ($G=1$)	Degrees Kelvin	n_0	N_e ($G=1$)
0	0·00	0·00	250	2·15 . 10^{19}	1·91 . 10^{19}
10	1·72 . 10^{17}	1·53 . 10^{17}	260	2·28	2·02
20	4·88	4·22	270	2·42	2·14
30	8·96	7·94	280	2·55	2·26
40	1·38 . 10^{18}	1·22 . 10^{18}	290	2·69	2·39
50	1·93	1·71	300	2·83	2·51
60	2·53	2·25	310	2·98	2·64
70	3·19	2·83	320	3·12	2·77
80	3·90	3·46	330	3·27	2·90
90	4·65	4·13	340	3·42	3·03
100	5·45	4·83	350	3·56	3·16
110	6·30	5·59	360	3·72	3·30
120	7·17	6·36	370	3·88	3·44
130	8·09	7·10	380	4·04	3·58
140	9·02	8·00	390	4·20	3·72
150	1·00 . 10^{19}	8·88	400	4·35	3·86
160	1·10	9·78	410	4·52	4·01
170	1·21	1 07 . 10^{19}	420	4·69	4·16
180	1·32	1·17	430	4·86	4·31
190	1·43	1·27	440	5·02	4·46
200	1·55	1·37	450	5·20	4·61
210	1·66	1·47	460	5·37	4·77
220	1·78	1·57	470	5·54	4·92
230	1·90	1·69	480	5·72	5·08
240	2·02	1·80	490	5·90	5·24
250	2·15	1·91	500	6·09	5·40

8.2 Symbols

a = parameter depending on impurity content and activation energy
d = excess of chemical donors over acceptors
e = electronic charge
$f(\epsilon)$ = Fermi-Dirac probability function
$f_p(\epsilon) = 1 - f(\epsilon)$
$F_j(\eta^*)$ = Fermi-Dirac integral of the jth degree
G = ratio of conduction electronic mass to true electronic mass
$g(\epsilon)$ = density of electronic states at energy ϵ
h = Planck's constant
k = Boltzmann's constant
K = Normalising constant for classical statistics
m = actual electronic mass
m_e = mass of a free electron
m_h = mass of a free hole
n_e = concentration of free electrons
n_h = concentration of free holes

n_i = intrinsic concentration of electrons and holes

n_0 = an equivalent number of electron states

$n(\epsilon)$ = density of electrons at energy ϵ

n_{deg} = concentration of free carriers at the degeneracy point

N = total number of states in the valence band

N_a = concentration of chemical acceptors

N_A = concentration of lattice-defect acceptors

N_d = concentration of chemical donors

N_D = concentration of lattice-defect donors

N_e = classical equivalent number of states in conduction band

N_h = classical equivalent number of states in valence band

$(N_d)_e$ = number of non-ionised donors

$(N_d)_h$ = number of ionised donors

$(N_a)_e$ = number of ionised acceptors

$(N_a)_h$ = number of non-ionised acceptors

T = absolute temperature in degrees Kelvin

T_{deg} = degeneracy temperature

T_{max} = temperature for which η^* is a maximum

x = number of chemical donor-acceptor pairs

X = number of lattice-defect donor-acceptor pairs

y = parameter depending on impurity content and activation energy

α = intrinsic gap at the absolute zero

β = temperature coefficient of intrinsic gap

ϵ = energy

ϵ_0 = energy at bottom of the conduction band (usually $\epsilon_0 = 0$)

ϵ^* = energy at Fermi level

$-\epsilon_a$ = energy at chemical-acceptor levels

$-\epsilon_A$ = energy at defect-acceptor levels

$-\epsilon_d$ = energy at chemical-donor levels

$-\epsilon_D$ = energy at defect-donor levels

$-\epsilon_i$ = energy at the top of the valence band

η = reduced energy ($= \epsilon/kT$)

η^* = reduced Fermi level

$-\eta_a$ = reduced energy at chemical-acceptor levels

$-\eta_A$ = reduced energy at defect-acceptor levels

$-\eta_d$ = reduced energy at chemical-donor levels

$-\eta_D$ = reduced energy at defect-donor levels

$-\eta_i$ = reduced energy at the top of the valence band

η^*_{max} = maximum value of reduced Fermi level

$-\lambda$ = reduced energy with respect to top of valence band

ϕ = constant used in a modification of the classical approximation to the Fermi integral $F_{1/2}(\eta^*)$.

Alignment and Adjustment of Synchronously Tuned Multiple-Resonant-Circuit Filters*

By MILTON DISHAL

Federal Telecommunication Laboratories, Incorporated; Nulley, New Jersey

A SIMPLE METHOD of "tuning up" a multiple-resonant-circuit filter quickly and exactly is demonstrated. The method may be summarized as follows: Very loosely couple a detector to the first resonator of the filter; then, proceeding in consecutive order, tune all odd-numbered resonators for maximum detector output, and all even-numbered resonators for minimum detector output (always making sure that the resonator immediately following the one to be resonated is completely detuned).

Also considered is the correct adjustment of the two other types of constants in a filter. Filter constants can always be reduced to only three fundamental types: f_0, d_r $(1/Q_r)$, and $K_{r(r+1)}$. This is true whether a lumped-element 100-kilocycle filter or a distributed-element 5000-megacycle unit is being considered. d_r is adjusted by considering the rth resonator as a single-tuned circuit (all other resonators completely detuned) and setting the bandwidth between the 3-decibel-down points to the required value. $K_{r(r+1)}$ is adjusted by considering the rth and $(r+1)$th adjacent resonators as a double-tuned circuit (all other resonators completely detuned) and setting the bandwidth between the resulting response peaks to the required value.

Finally, all the required values for K and Q are given for an n-resonant-circuit filter that will produce the response $(V_p/V)^2=1+(\Delta f/\Delta f_{3db})^{2n}$.

• • •

1. Introduction

This paper attempts to answer two questions: "Exactly how can one 'tune up' a synchronously tuned multiple-resonant-circuit filter and be sure the tuning is correct?" and "Exactly how can one make sure that the mechanical design is actually supplying the required circuit constants?"

It should be noted that, for brevity, the paper will refer only to band-pass filters; the reader should realize that the discussion also applies similarly to the alignment and adjustment of low-pass, high-pass, and band-rejection filters when analogous frequency-variables and circuit constants are used.

The physical embodiment of a constant-K or equivalent type of filter, i.e., a filter having n complex frequency roots and no finite frequencies of infinite attenuation (all zeros at infinity), must exactly supply the numerical values of three kinds of quantities:[1] A) resonant frequency f_0, B) coefficients of coupling between adjacent resonators $K_{r(r+1)}$, and C) resonator decrements d_r $(Q_r=1/d_r)$. It may be helpful to note that the above is true whether the elements of the unit are lumped, quasi-lumped, or distributed, so long as the percentage bandwidth is less than approximately 10 percent for the latter two cases, as is most always true.

It should be realized that in the literature concerning filters a number of seemingly different types of constants have been used to describe the same or exactly equivalent networks. For example, classical filter theory, which gives only approximate design data, usually produces the required values for L, C, M, and R; late in the 1930's, a number of papers described circuits by the so-called "ladder-network coefficients" for each arm; and at present many papers speak of the doubly or singly loaded Q of each resonator. In every case, the many different types of constants are all equivalent, but it has been the experience of the author that the constants f_0, $K_{r(r+1)}$, and Q_r are the "best" to use, particularly when dealing with *dissipative* filters.

* Reprinted from *Proceedings of the IRE*, v. 39, pp. 1448–1455; November, 1951. Section 4 has been modified substantially in this reprint. This paper was written in connection with work done on a contract for the Department of Air Force—Air Materiel Command, Wright-Paterson Air Force Base.

[1] M. Dishal, "Design of Dissipative Band-Pass Filters Producing Desired Exact Amplitude-Frequency Characteristics," *Proceedings of the IRE*, v. 37, pp. 1050–1069; September, 1949: and *Electrical Communication*, v. 27, pp. 56–81; March, 1950.

For practical reasons usually involving mechanical tolerances, most selective-circuit designs incorporate a trimming adjustment for setting the resonant frequency of each resonator. After the filter is mechanically finished, the unit is aligned, i.e., all resonant frequencies are somehow correctly adjusted. Section 3 describes a method of alignment for multiple-resonant-circuit filters that is precise, requires no sweep-frequency generator, and can be performed by unskilled personnel.

The coefficient of coupling between adjacent resonators is usually not made variable as this adjustment requires a person "skilled in the art"; each K is carefully set by the designer as part of the mechanical design, which must be sufficiently stable to maintain it at the required value. Section 4 describes an easy method for experimentally adjusting each coefficient of coupling to the exact desired value.

For the sake of completeness, a few comments are made about measuring Q, the third of these constants, in Section 5.

Section 6 presents some useful design data on what values of K and Q are required in a multiple-resonator-circuit filter.

2. *Symbols*

n = total number of resonators used in a filter.

r = resonator number in filter chain. The resonator at the input end is number *1*.

f_0 = resonant frequency of each resonator; this must include all coupling reactances. See Section 3.3.

$K_{r(r+1)}$ = coefficient of coupling between the rth and $(r+1)$th adjacent resonators. This may be defined fundamentally as the fractional bandwidth between the resulting response peaks that exist when each *pair* of adjacent resonators is considered separately (and the resonator Q's are infinite).

d_r = decrement of the resonator. This may be defined fundamentally as the fractional bandwidth between the 3-decibel-down points when each resonator is considered separately.

Δf_{3db} = total bandwidth between 3-decibel-down response points.

$V_{1,r}$ = voltage across *first* resonator when all following resonators up to the rth resonator have been correctly tuned.

$p_{AB} = K_{AB}(Q_A Q_B)^{1/2}$ = fraction of "critical coupling" in a double-tuned circuit made up of resonators A and B.

$(\Delta f_p)_A$ = total bandwidth between response peaks in resonator A, to which a generator is coupled, when A is coupled *only* to an adjacent resonator B.

$F_p = (\Delta f_p / f_0)$ = fractional bandwidth between response peaks.

$t = Q_A/Q_B$.

Δf_β = total bandwidth between response points that are V_p/V_β down from the peak response.

V_p = voltage output at peak of response curve.

V_β = voltage output at point of response curve where the bandwidth is Δf_β.

3. *Alignment of Multiple-Resonator Filters*

3.1 General Principle

This paper will refer mainly to small-percentage-bandwidth node networks. The reader should realize that in accordance with the principle of duality and with the following substitutions of words: mesh for node, current for voltage, voltage for current, open circuit for short circuit, and the like, the alignment procedure applies similarly to mesh networks. The procedure applies also to the large-percentage-bandwidth constant-K configuration discussed in Section 3.3.

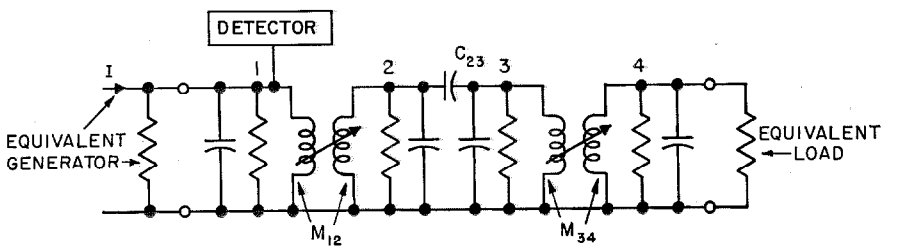


Figure 1—Quadruple-tuned filter used to demonstrate alignment procedure of Section 3. It should be realized that the alignment procedure applies to all the different types of synchronously tuned constant-K and coupled-resonant-circuit filters.

The fundamental principle proposed in this section is that alignment can best be done by *completely assembling the filter and then concentrating on the amplitude phenomena occurring in the first resonator of the filter chain at the desired resonant frequency.* In Section 3.3, it is shown that if all the resonators are first completely detuned and if they are resonated in numerical order, calling the input resonator *1*, then all odd-numbered resonators place an open circuit (high resistance) and all even-numbered resonators place a short circuit (low resistance) across the input terminals when correctly tuned.

3.2 Alignment Procedure

The alignment procedure will be described using the quadruple-tuned node-type band-pass filter shown in Figure 1 as an example. Figure

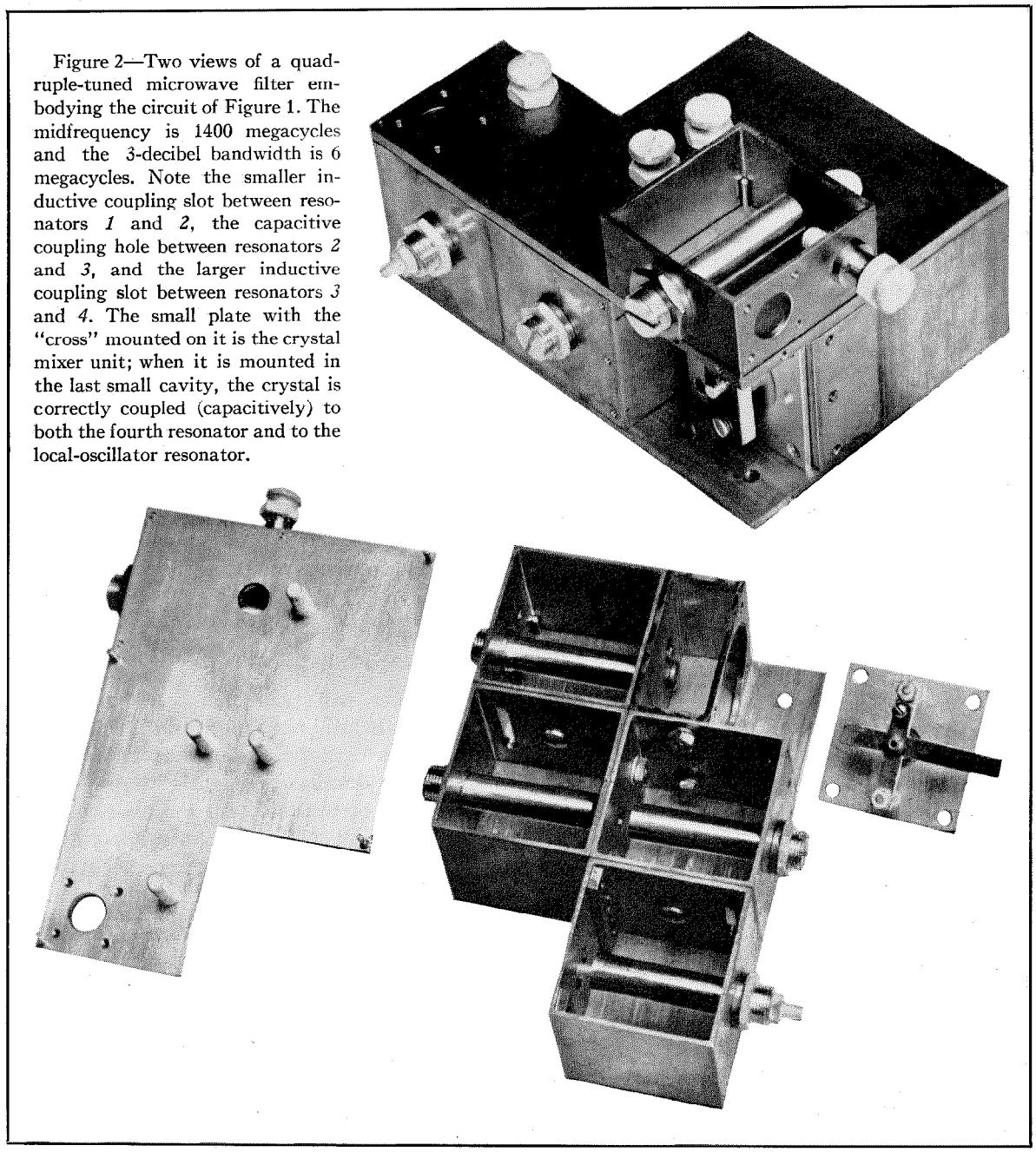


Figure 2—Two views of a quadruple-tuned microwave filter embodying the circuit of Figure 1. The midfrequency is 1400 megacycles and the 3-decibel bandwidth is 6 megacycles. Note the smaller inductive coupling slot between resonators *1* and *2*, the capacitive coupling hole between resonators *2* and *3*, and the larger inductive coupling slot between resonators *3* and *4*. The small plate with the "cross" mounted on it is the crystal mixer unit; when it is mounted in the last small cavity, the crystal is correctly coupled (capacitively) to both the fourth resonator and to the local-oscillator resonator.

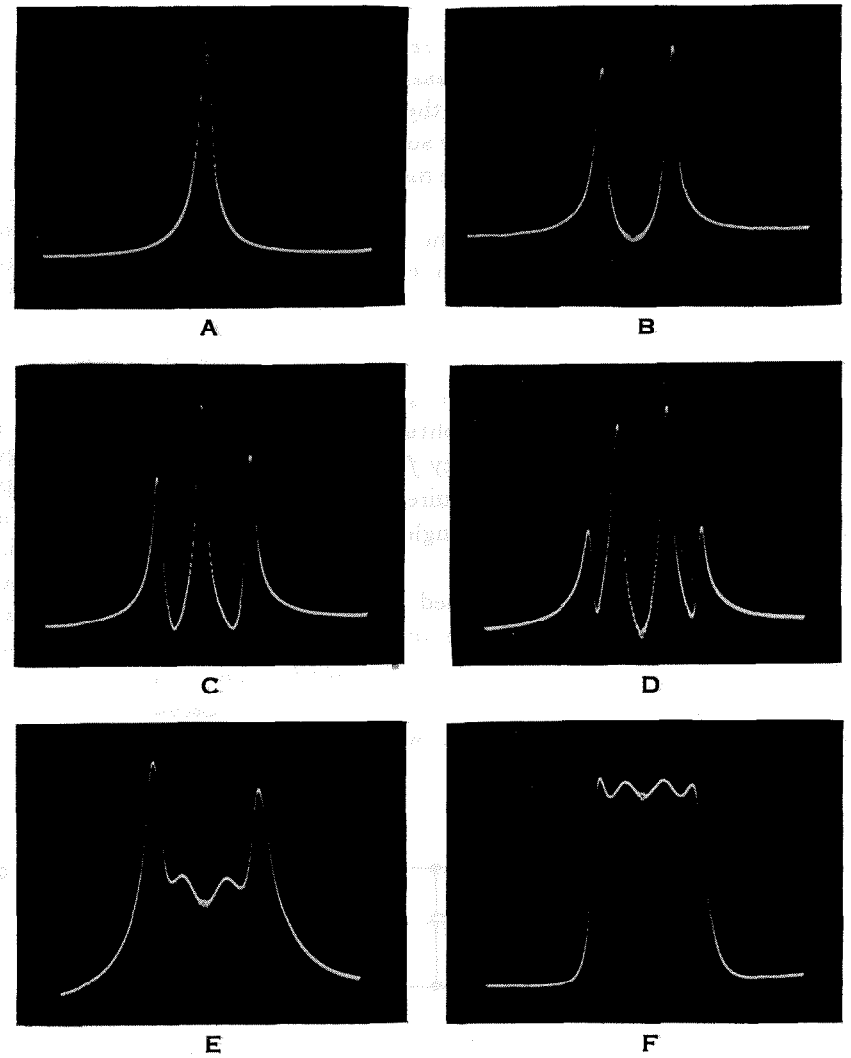


Figure 3—Oscillograms of the amplitude–frequency phenomenon occurring in resonator *1* as the alignment steps of Section 3 are performed.

A—Resonator *2* is short-circuited or detuned and the input resonator *1* (odd numbered) is adjusted for maximum marker signal of f_0.

B—Resonator *3* is detuned and the second resonator (even-numbered) is tuned for minimum amplitude of f_0.

Oscillograms *C* and *D* show the continuation of the procedure of tuning odd-numbered resonators to produce maximum and even-numbered resonators to produce minimum values of f_0 response in resonator *1*. It can be seen that as the rth resonator is tuned, there will be r peaks and $r-1$ valleys produced in resonator *1*; this is a simple restatement of Foster's reactance theorem.

Oscillogram *E* shows the voltage across resonator *1* as, with correct applied loading, the last resonator (even-numbered) of Figure 1 is tuned for minimum output at f_0.

Oscillogram *F* shows the resulting Chebishev transfer response shape (no tuning adjustments were retouched).

2 shows a practical physical embodiment of the circuit of Figure 1.

The procedure is applicable to all coupled-resonant-circuit filters, whether they be low-frequency constant-K configurations, medium-frequency coupled circuits, microwave quarter-wave-coupled waveguide filters, or the like.

A. Connect the generator to the first resonator of the filter and the load to the last resonator of the filter in exactly the same manner as they will be connected in actual use.

B. Couple a nonresonant detector directly and very loosely[2] to either the electric (voltage) or magnetic (current) field of the *first* resonator of the filter chain.

C. Completely detune[3] all resonators.

D. Set the generator frequency to the desired mid-frequency of the filter.

E. Tune resonator *1* for *maximum output* indication on the detector. Lock the tuning adjustment.

F. Tune resonator *2* for *minimum output* indication on detector. Lock the tuning adjustment.

G. Tune resonator *3* for maximum output and lock the tuning adjustment.

H. Tune resonator *4* for minimum output and lock the tuning adjustment. The alignment of the filter shown in Figure 1 is now complete.

If it is impracticable completely to detune all

[2] A nonresonant detector (or generator) may be said to be "very loosely" coupled when it lowers the unloaded Q of the resonator by less than 5 percent (say).

[3] A resonator is sufficiently detuned when its resonant frequency is at least 10 pass-band-widths away from the pass-band midfrequency.

the resonators in a node network, a *single* device may be used to short-circuit *the resonator immediately following the one being tuned* since this will remove the effect of all the following resonators. It is important to make sure that this short circuit is fully effective at the measurement frequency.

Figure 3 clearly demonstrates the amplitude–frequency phenomena that occur in each step of the alignment procedure. A sweep-frequency generator was used, and attention is called to the resonant-frequency marker. It should be clearly realized that since the alignment adjustments depend exclusively on the amplitude of the response at the resonant frequency f_0, a sweep-frequency generator is *not* required and all adjustments can be made with a single-frequency input f_0.

These oscillograms were obtained in aligning the quadruple-tuned filter shown in Figure 1. It was designed to produce a Chebishev transfer shape[1] having a ½-decibel peak-to-valley ratio when loaded on one side only; i.e., it was fed by a high-impedance generator.

tuning adjustments should be set midway between two points of equal output.

3.3 Simple Theory of Alignment Procedure

Perhaps the simplest way of showing that the alignment procedure is correct is to consider the large-percentage-bandwidth constant-K filter chain shown in Figure 4A, to which all small-percentage-bandwidth coupled-resonant-circuit filters are exactly equivalent no matter what type of coupling is used between adjacent resonant circuits; and then to consider as a further example the small-percentage-bandwidth node circuit shown in Figure 4B.

The reasoning applicable to the constant-K configuration of Figure 4A requires previous knowledge of two simple facts.

First, complete detuning of all the resonators means that all the *series resonators* are effectively *open-circuited* and all the shunt resonators are effectively *short-circuited.*

Second, when correctly aligned, the resonant frequency f_0 of each separate resonator is identical.

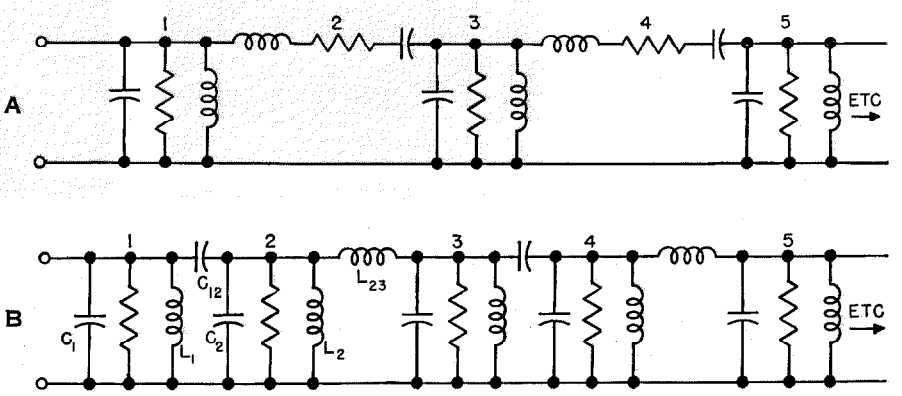


Figure 4—*A* and *B* are, respectively, large- and small-percentage-bandwidth networks referred to in the explanations given in Section 3.3. It should be clearly realized that Figure 3B shows alternate C and L couplings purely as an example; the explanation is true no matter what type of reactive coupling is used.

Efficient filters with low internal losses, i.e., those using resonators having unloaded Q's very much greater than the fractional midfrequency ($f_0/\Delta f_{3db}$), produce deep and broad minimums when the even-numbered resonators are properly tuned, as may be seen in Figure 3B. Therefore, it is important to use a large-amplitude signal input and high detector gain so that the middle of the minimum can be tuned accurately to the midfrequency. If the maximum generator input and detector gain still produce a broad null, the

Thus, when resonator *1* is tuned to f_0 (with resonator *2* open-circuited), maximum voltage will appear across the high parallel-resonant resistance of resonator *1*. When resonator *2* is tuned to f_0 (with resonator *3* short-circuited), the low series-resonant resistance of *2* will shunt the terminals of *1*, thus dropping the voltage across resonator *1* to a minimum. On tuning *3* to f_0 (with resonator *4* open-circuited), the high parallel-resonant resistance of *3* will remove the low series-resonant resistance of *2* from across

the terminals of *1* so that the voltage across *1* will again rise to a maximum. Thus, starting at the front end of the filter, all odd-numbered resonators will produce a maximum voltage and all even-numbered resonators will produce a minimum voltage across resonator *1*.

The reasoning applicable to the small-percentage-bandwidth node network of Figure 4B requires previous knowledge of three simple facts.

First, complete detuning of a resonator means that the node involved is effectively short-circuited.

Second, when correctly aligned, the resonant frequency of each node is identical and the elements that resonate a node consist of every susceptance that touches the node, e.g., node *2* of Figure 4B is resonated by adjusting C_2 (or L_2) to resonate with parallel combination of C_{12}, C_2, L_2, and L_{23}.

Third, if a group of reactances parallel resonate together, then any one of the reactances also series resonates with all the others, e.g., C_{12} series resonates with the parallel resultant of C_2, L_2, and L_{23}.

Thus, where node *1* is tuned to f_0 (with node *2* short-circuited), the high parallel-resonant resistance of C_1, L_1, and C_{12} will produce a voltage maximum at f_0. When node *2* is tuned to f_0 (with node *3* short-circuited), C_{12} will series resonate with the parallel resultant of C_2, L_2, and L_{23}, thus placing a short circuit across node *1* and producing a voltage minimum across node *1*. The process repeats as alignment proceeds, producing maximums for alignment of odd-numbered and minimums for even-numbered resonators.

It will be shown in Section 4.4 that if we know the Q's of each resonator being used, then the ratios of maxima and minima occurring in resonator *1* can be used to set or check all the $(n-1)$ coefficients of coupling in a filter.

4. *Exact Adjustment of Coupling Between Resonators*

4.1 General Principle

Before this section can be applied to the mechanical design and adjustment of a filter, it is, of course, necessary to determine by some synthesis procedure just what adjacent-resonant-circuit coefficients of coupling the mechanical embodiment must produce. As mentioned in the Introduction, no matter what type of constants are used to describe the synthesized network, they can always be transformed into f_0, $K_{r(r+1)}$, and Q_r.

The fundamental procedure being proposed in this section is to consider every pair of adjacent resonators as a double-tuned, i.e., two-pole, circuit (with all the other resonators completely detuned), and so be able to use the exactly known relation between the circuit constants and the resulting amplitude–frequency characteristic of a double-tuned circuit.

In practice, two cases face the designer. In the first case, he has available resonators with unloaded Q's that are very much greater than the fractional midfrequency $(f_0/\Delta f_{3db})$ being used; whereas in the second case he has available resonators with unloaded Q's that are only 2 or 3 (say) times the fractional midfrequency being used.

An example of the first case is the filter shown in Figure 2; here the fractional midfrequency being used is 1400/6 or 234 and the unloaded Q's of the resonators being used are approximately 2500. Another example would be a 40-megacycle-wide, intermediate-frequency interstage filter with a geometric mean midfrequency of 20 megacycles to be made with resonators having an unloaded Q of 100. (It should be realized that for this design the large-percentage-bandwidth configuration of Figure 4A would be used).

An example of the second case is an intermediate-frequency-amplifier interstage filter with a required midfrequency of 2 megacycles and a required 3-decibel bandwidth of 50 kilocycles using unloaded resonator Q's of (say) 90. Here $(f_0/\Delta f_{3db})$ is 40 giving a ratio of unloaded Q to fractional midfrequency of only 2.25 to 1.

Section 4.2 gives a recommended procedure for adjusting a coefficient of coupling when the unloaded Q of the resonators is much greater than $(f_0/\Delta f_{3db})$ and Section 4.3 gives a recommended procedure when Q unloaded is approximately equal to $(f_0/\Delta f_{3db})$.

4.2 Adjustment Procedure When Q Unloaded is $\gg (f_0/\Delta f_{3db})$

In a double-tuned circuit *with Q_A and Q_B equal to infinity*, the fractional bandwidth $(\Delta f_p/f_0)_A$

between primary response peaks is exactly equal to the coefficient of coupling between resonators A and B. (*In fact, this may be used as the basic definition of the constant that is commonly called the coefficient of coupling.*) This direct relationship makes this phenomenon an excellent one to use as the basis of an experimental procedure for adjusting coefficients of coupling to a desired value, when the unloaded Q's of the resonators are very much greater than the desired coefficient of coupling. When the resonators do not have infinite Q, the above equality is not exactly true, and Figure 5 together with the described procedure supplies a way of finding the exact coefficient of coupling between adjacent resonators.

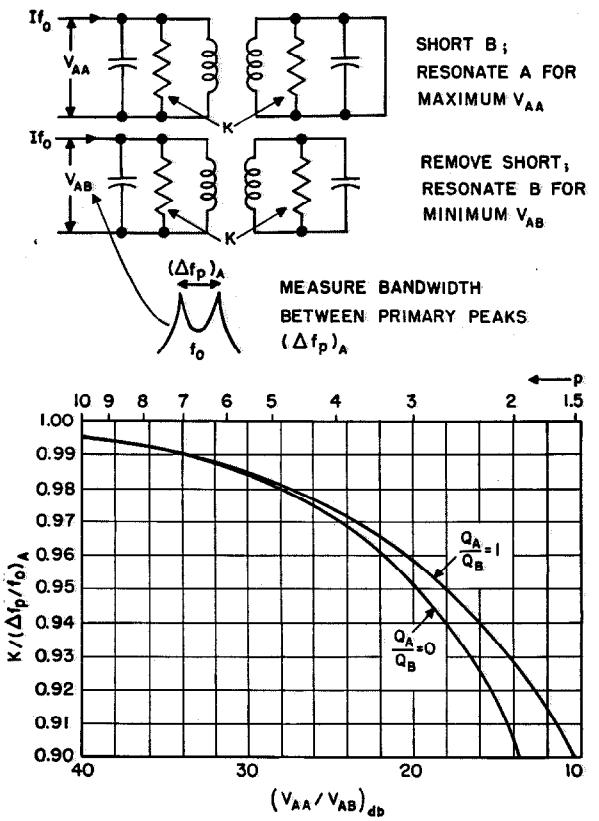


Figure 5—Method of obtaining exact coefficient of coupling K between two resonators when the K required is much greater than the unloaded decrements. See (1) and (2).

The procedure for measuring adjacent-resonator coupling, which is applicable to all coupled-resonant-circuit filters whether they be called low-frequency constant-K configurations, medium-frequency coupled circuits, or microwave quarter-wave coupled-waveguide filters, is as follows:

A. Designate as A and B the two adjacent resonators between which the coefficient of coupling is to be adjusted. If A and B have different Q's, A should be the lower-Q resonator.

B. Couple a nonresonant signal generator directly and very loosely to either the electric (voltage) or magnetic (current) field of resonator A.

C. Couple a nonresonant detector directly and very loosely to either the electric (voltage) or magnetic (current) field of resonator A.

D. Completely detune all the resonators in the filter chain.

E. Tune resonator A for maximum output from the detector. Record the signal-generator input and detector output.

F. Tune resonator B for minimum output from the detector (as in alignment procedure, Section 3). Increase the signal-generator input to produce the same output obtained in step E.

G. The ratio of the signal-generator input in step F to that in step E is the abscissa of the graph of Figure 5. From the ordinate of this graph, read the ratio of the coefficient of coupling K between resonators A and B, to the percentage bandwidth $(\Delta f_p/f_0)_A$ between the response peaks that are now present across resonator A.

H. Carefully measure the bandwidth $(\Delta f_p)_A$ between the response peaks of resonator A.

J. The exact coefficient of coupling is equal to the fractional bandwidth between these peaks times the ordinate obtained in step G.

Examination of the ordinate values of Figure 5 shows that even with a $(V_{A,A}/V_{A,B})$ ratio as small as 12 decibels, the existing coefficient of coupling is only 8.5 percent less than the percentage bandwidth between the primary peaks; therefore, in many cases it may be permissible simply to measure the bandwidth $(\Delta f_p)_A$ between primary response peaks and make the approximation that the coefficient of coupling is equal to about 0.96 times the measured fractional bandwidth.

Figure 7 shows the frequency–amplitude relation on the primary side when the above procedure is used.

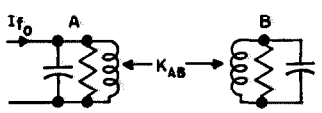
4.3 Adjustment Procedure When Q Unloaded $\doteq (f_0/\Delta f_{3db})$

For this second case, it will be found that the coefficients of coupling required in the network to produce "desirable" response shapes are not much greater than the unloaded decrements. When this is true, the resulting peak-to-valley ratio in the primary will not be very great, when the previous procedure of Section 4.2 is used, and therefore the primary peaks will not be sharply defined. When this is the case, the coefficient of coupling can be found in terms of the measured Q of the resonators being used. The procedure is as follows:

Steps *A* through *F* are identical to those given in the previous procedure of Section 4.2.

G. The ratio of the signal-generator input in step *F* to that in step *E* is the abscissa of the graph of Figure 6. From the ordinate of this graph, read the ratio of the coefficient of coupling K_{AB} to the geometric means of the decrements of resonators *A* and *B*.

H. Carefully measure the Q of each resonator *A* and *B* by measuring the 3-decibel (6-decibel, etc.) bandwidth as outlined in Section 5.



D. SHORT B

E. TUNE A FOR MAXIMUM V_{AA}

F. REMOVE SHORT, TUNE B FOR MINIMUM V_{AB}

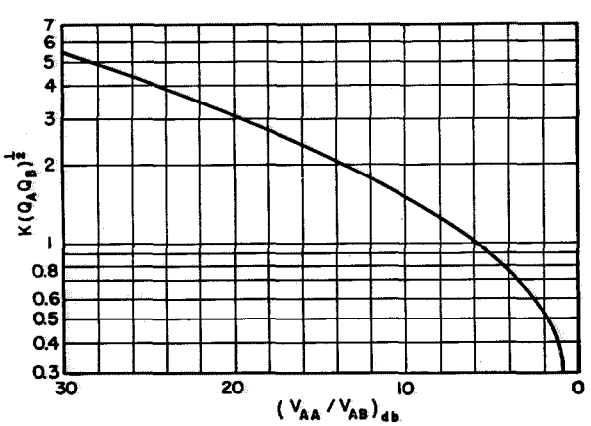


Figure 6—Method of obtaining exact coefficient of coupling K between two resonators when the K required is not much greater than the unloaded decrements. See (1).

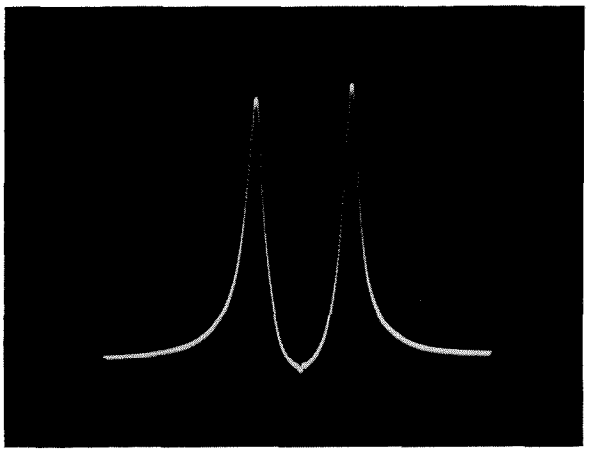


Figure 7—Response peaks occurring in the input resonator *A* of a properly resonated pair of adjacent resonators with all other resonators completely detuned. By measuring the frequency bandwidth between the peaks, the exact coefficient of coupling may be obtained. (See Figure 5.)

J. The exact coefficient of coupling is equal to the geometric mean of the *A* and *B* decrements times the ordinate value obtained in step *G*.

4.4 Quantitative Theory of Measuring Coefficient of Coupling

All the amplitude–frequency characteristics of a double-tuned circuit are most conveniently related to each other by means of two variables, the Q ratio of primary to secondary Q_A/Q_B and the $K_{AB}(Q_AQ_B)^{1/2}$ product. For the preparation of Figures 5 and 6, we need the relations between the above variables and two quantities: A) drop in primary voltage when the secondary is correctly resonated $V_{A,B}/V_{A,A}$ and B) bandwidth between response peaks in the primary side $(\Delta f_p)_A$.

Straightforward analysis of a correctly resonated double-tuned circuit results in the following equations for these quantities in which the Q ratio will be denoted by $t=Q_A/Q_B$ and the $K_{AB}(Q_AQ_B)^{1/2}$ product by p.

$$\left(\frac{V_{A,A}}{V_{A,B}}\right)=1+p^2. \tag{1}$$

$$\left(\frac{\Delta f_p}{f_0}\right)_A=K\left\{\left[1+2(1+t)\frac{1}{p^2}\right]^{1/2}-t\frac{1}{p^2}\right\}^{1/2}. \tag{2}$$

$$t=\frac{Q_A}{Q_B};\quad p=K_{AB}(Q_AQ_B)^{1/2}. \tag{3}$$

Equations (1) and (2) are used straightforwardly to obtain the graphs of Figures 5 and 6.

4.5 Checking Coefficients of Coupling

During construction, or after a filter has been constructed, it is often desirable to be able to set or to double check each coefficient of coupling without going through the procedure of converting each pair of adjacent resonators into a double-tuned circuit as is required by Sections 4.2 and 4.3.

As shown below, this can be accomplished by measurements made entirely at the input resonator. There are, in practice, two cases that have to be considered: In the first case, the unloaded Q's of the resonators being used are very much greater than the fractional midfrequency $(f_0/\Delta f_{3db})$ being used; i.e., the unloaded individual Q's are essentially infinite. In the second case, the unloaded Q's of the resonators are, say, only 4, or 5, or fewer times $(f_0/\Delta f_{3db})$.

For the first case above, the K's can be approximately measured, in consecutive order, by measuring the bandwidth between the various response peaks appearing in resonator *1*, as each of the following resonators is resonated in consecutive order (see Figure 3). If the unloaded Q's truly approach infinity, i.e. are, say, 100 times $(f_0/\Delta f_{3db})$, then this procedure gives exact answers.

It will be remembered from Figure 3 that there will be r response peaks occurring in the input resonator when the rth resonator is correctly tuned.

To calculate the peak bandwidths that should be measured when the rth resonator is tuned, straightforward analysis shows that we must solve the polynomial (4) for its roots

$$F_p{}^r-(\sum K^2)F_p{}^{(r-2)}+(\sum K^2K^2)F_p{}^{(r-4)} -(\sum K^2K^2K^2)F_p{}^{(r-6)}\ldots=0. \quad (4)$$

The polynomial stops at the first- or zero-power term; i.e., no negative exponents are considered.

The coefficient of the $F_p{}^{(r-2)}$ term is the sum of all the products of the coefficients of coupling squared taken one at a time; the coefficient of the $F_p{}^{(r-4)}$ term is the sum of all the K^2 products taken two at a time, *but in any pair a subscript number must not appear more than once;* the coefficient of the $F_p{}^{(r-6)}$ term is the sum of all the K^2 products taken three at a time, *but in any triplet a subscript number must not appear more than once;* and so forth.

As an example, as the first to fifth resonators are tuned consecutively, the following five polynomials must be solved consecutively to calculate the fractional bandwidth $(\Delta f_p/f_0)$ that should occur between response peaks.

$$F_p=0 \quad (4A)$$

$$F_p{}^2-K_{12}{}^2=0 \quad (4B)$$

$$F_p{}^3-(K_{12}{}^2+K_{23}{}^2)F_p=0 \quad (4C)$$

$$F_p{}^4-(K_{12}{}^2+K_{23}{}^2+K_{34}{}^2)F_p{}^2+(K_{12}K_{34})=0 \quad (4D)$$

$$F_p{}^5-(K_{12}{}^2+K_{23}{}^2+K_{34}{}^2+K_{45}{}^2)F_p{}^3 +(K_{12}{}^2K_{34}{}^2+K_{12}{}^2K_{45}{}^2+K_{23}{}^2K_{45}{}^2)F_p=0. \quad (4E)$$

These first five polynomials require simple linear- and quadratic-equation solutions, and, because the coefficients are known numerical values, the Graeffe root-squaring process can be used to solve accurately any of the polynomials.

For the second case described above, the coefficients of coupling should be set or measured in consecutive order as follows: Accurately measure the Q of each resonator in the filter, then proceed step by step through the alignment procedure of Section 3.2, accurately measuring (and recording) the magnitudes of the alternate maxima and minima produced. Straightforward analysis shows that the ratio of the detector output obtained when resonator *1* is alone resonated to that obtained when resonator r is resonated is given by the continued fraction of (5).

$$\left(\frac{V_{1,1}}{V_{1,r}}\right)=\left[1+\cfrac{p_{12}{}^2}{1+\cfrac{p_{23}{}^2}{1+\cdots\cfrac{}{1+p^2{}_{(r-1)r}}}}\right] \quad (5)$$

It is important to realize that Q_1 is the Q of resonator number *1* with both the generator and detector coupled to it.

Since we know the desired value for each K and have measured each Q, we know the value of each $p_{12}{}^2=K_{12}{}^2Q_1Q_2$, and so on in (5); and the measured value of the voltage ratios $(V_{1,1}/V_{1,r})$ should, of course, equal those calculated from (5).

5. *Adjustment of Resonator Decrement 1/Q*

It has been the author's experience that any method of measuring Q that removes the resonator from the exact position it occupies in the filter chain is potentially inaccurate.

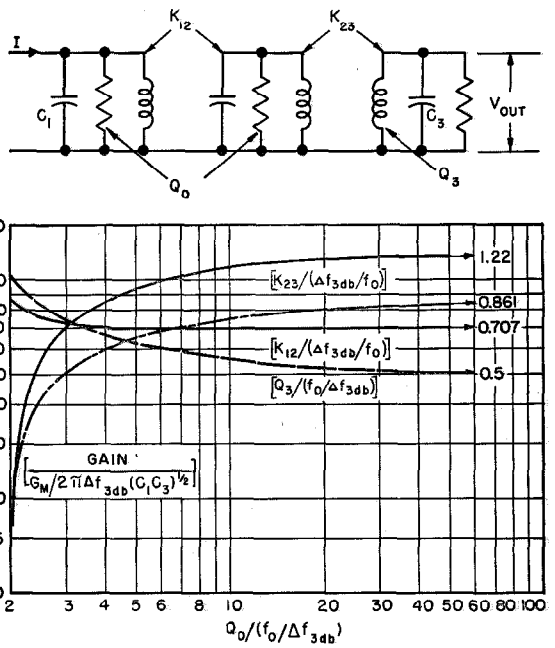


Figure 8—Exact design for a finite-Q triple-tuned node circuit producing a Butterworth response shape when driven by an infinite-resistance generator.

It has also been noted that measurements involving an amplitude-modulated oscillator can lead to erroneous results particularly in the ultra-high-frequency and microwave regions because of spurious frequency modulation. If an amplitude-modulated carrier is being used, an obvious check for appreciable spurious frequency modulation is to use an oscilloscope to examine the envelope of the waveform being measured and a narrow-band receiver to examine the frequency content of the supposedly purely amplitude-modulated carrier.

The most trustworthy method of making accurate unloaded- or loaded-Q measurements on a resonator that is part of a filter chain seems to be as follows:

A. Completely assemble the filter.

B. Completely detune all resonators except the one to be measured. Obviously, complete detuning of the resonator on each side of the one being measured should be satisfactory.

C. A nonresonant signal generator is coupled directly and very loosely to either the electric or magnetic field of the resonator.

D. A nonresonant detector is coupled very loosely and preferably to the field opposite to that being used for the generator; i.e., make sure that there is negligible direct coupling between generator and detector.

E. Using an unmodulated wave or an amplitude-modulated wave checked for negligible frequency modulation from the signal generator, measure the frequency difference Δf_β between the points that are V_p/V_β down from the peak response; the resonator Q is given by

$$Q=(f_0/\Delta f_\beta)[(V_p/V_\beta)^2-1]^{1/2}. \qquad (6)$$

Obviously, when high Q's are to be measured, the apparatus must be capable of measuring very-small-percentage bandwidths. This may be accomplished by "beating down" the measurement frequency with a very stable local oscillator and a mixer and by making the measurements at the resulting difference frequency. By this means, the accuracy of the measuring apparatus is increased by the ratio of the original to the

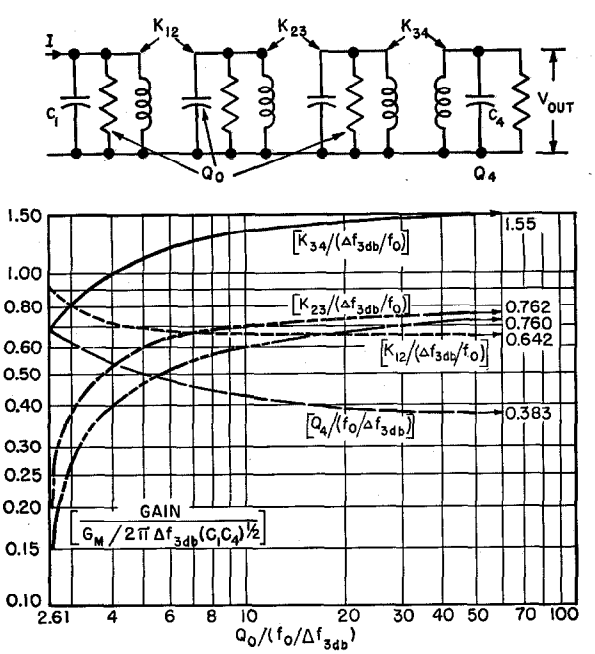


Figure 9—Exact design for a finite-Q quadruple-tuned node circuit producing a Butterworth response when driven by an infinite-resistance generator.

difference frequencies. The "cost" of this simplification is, of course, the necessity of using a very stable local oscillator.

6. *K's and Q's to Produce Response Shape* $V_p/V=[1+(\Delta f/\Delta f_{3db})^{2n}]^{1/2}$

A straightforward synthesis procedure[1,4,5] shows that for the transfer response shape given just above, and if infinite-Q resonators are used, i.e., in practice, resonators whose unloaded Q's are greater than $10/[\sin(90°/n)]$ times the fractional midfrequency $(f_0/\Delta f_{3db})$, it is possible to write very concisely the exact values of K and Q required for any number n of resonators for two important practical cases. The rate of cutoff obtained with the transfer response shape given in the title of this section is exactly $6n$ decibels per octave and the size of an octave is any bandwidth greater than $3^{1/m}$ times the 3-decibel bandwidth.

For the case where *one end of the network can have only a pure reactance placed across it* (by either a reactive generator, e.g., the plate of a pentode tube, or by a reactive load), the required Q and K values are given exactly by (7A) and (7B).

$$\frac{Q_1}{(f_0/\Delta f_{3db})}=\sin\frac{90°}{n}, \quad Q_{2\to n}=\infty \tag{7A}$$

$$\frac{K_{r(r+1)}}{(\Delta f_{3db}/f_0)}=\frac{\cos(r\,90°/n)}{\{[\sin(2r-1)(90°/n)][\sin(2r+1)(90°/n)]\}^{1/2}}, \tag{7B}$$

[4] E. L. Norton, United States Patent 1,788,538; January, 1931.

[5] W. R. Bennett, United States Patent 1,849,656; March, 1932.

where r is made equal to 1, 2, 3, and so forth, up to $(n-1)$; and n is the total number of resonators used. For the above design equations, *the end resonator that is loaded is called resonator 1.*

For the case where *both ends of the network must have resistances placed across them* (e.g., a 50-ohm generator and 50-ohm load are being used), then the required Q and K values are given exactly by (8A) and (8B).

$$\frac{Q_1(=Q_n)}{(f_0/\Delta f_{3db})}=2\sin\frac{90°}{n}, \quad Q_{2\to(n-1)}=\infty \tag{8A}$$

$$\frac{K_{r(r+1)}}{(\Delta f_{3db}/f_0)}=\frac{0.5}{\{[\sin(2r-1)(90°/n)][\sin(2r+1)(90°/n)]\}^{1/2}}, \tag{8B}$$

where r is set equal to 1, 2, 3, and so forth, up to $(n-1)$; and n is the total number of resonators being used. Since the resulting circuit is symmetrical, it makes no difference which end resonator is called resonator *1.*

For the unfortunately practical case where the unloaded Q of the resonators being used is not infinite, it does not seem to be possible to obtain elegantly simple design equations like (7) and (8). Figures 8 and 9 give the K and Q values required to produce exactly the transfer response shape $(V_p/V)^2=1+(\Delta f/\Delta f_{3db})^{2n}$ for triple- and quadruple-resonator filters, respectively, for the reactive-generator case. The abscissa of these graphs is the ratio of the unloaded Q (Q_0) of the resonators being used to the fractional midfrequency $(f_0/\Delta f_{3db})$.

In Memoriam

JACOB SUTER JAMMER

JACOB SUTER JAMMER, a director and vice president of International Standard Electric Corporation, was associated with the International System since its formation in 1924. His many years of loyal service in the fields of manufacturing and sales have contributed immeasurably to the growth of the Corporation. With an unusually broad grasp of System activities, he readily undertook many difficult and varied assignments, which he fulfilled capably and with imagination. He possessed a keen mind, a lively sense of humor, and a friendly nature combined with a gift of showmanship.

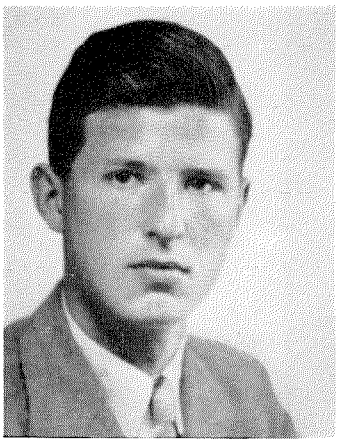
Mr. Jammer was born in Cumberland, Maryland, in 1898. He was graduated in 1918 from Johns Hopkins University with a degree in electrical engineering. After graduation, he joined the research department of the Western Electric Company in New York. In 1924, he was transferred to the International Western Electric Company and was stationed in Australia to supervise the installation of the first long-distance carrier telephone system between Sydney and Melbourne. He further assisted the Australian administration in the planning and installation of their long-distance telephone network. During this period, the International Western Electric Company was acquired by the International System, and Mr. Jammer was given a long series of increasingly responsible positions in both the United States and abroad.

In 1929 he was appointed European general sales manager responsible for all sales and commercial activities of the International System in Europe. In 1931, as a member of the general staff in London, he had charge of developing the printing-telegraph business of the company on a world-wide basis. He was made regional vice president of the eastern and central European area in 1939, and in 1940 was transferred to the New York headquarters, where he directed various sales and manufacturing activities of the Corporation.

At the time of his death, Mr. Jammer was also a vice president and director of the International Standard Trading Corporation, where he was instrumental in bringing to the American market many of the electronic developments of the associate companies abroad.

Mr. Jammer died on March 16, 1952, at the Community Hospital in Montclair, New Jersey, after a short illness.

Contributors to This Issue



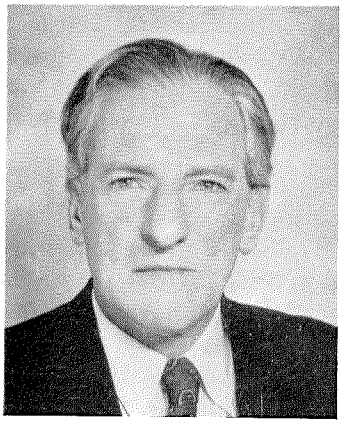
J. S. BLAKEMORE

J. S. BLAKEMORE was born at London on May 25, 1927. He studied physics at Queen Mary College of London University, receiving the B.Sc. degree in 1948 and a Ph.D. degree after doing postgraduate research in semi-conductors.

Dr. Blakemore joined Standard Telecommunication Laboratories late in 1950 and has been working on problems relating to the use of semi-conductors in rectifiers and transistors.

• • •

WALTER A. BRANDT was born in Austria and received the degree of doctor of technical sciences in 1922 from the Technical University at Vienna.



WALTER A. BRANDT

On graduation, he joined the staff of the Austrian Telephone Manufacturing Company, which company became associated with the International System in 1930. Dr. Brandt was transferred to the United Telephone and Telegraph Works in Vienna until 1938 when he went to the Bell Telephone Manufacturing Company in Antwerp, Belgium, for two years.

From 1941 to 1943, he was engaged in development work for the United States Army and the Columbia Broadcasting System in New York City. He joined the staff of Federal Telephone and Radio Corporation in 1943, working on carrier systems and automatic switching methods.

On October 16, 1951, Dr. Brandt died after a long illness. He was 59 years old.

• • •

FRANK X. BUCHER was born at Miamisburg, Ohio, on January 14, 1919. He received the B.S. degree in electrical engineering from the University of Dayton in 1941 and the M.S. degree from Stevens Institute of Technology in 1950.

During the war, he was in the electronics development laboratory of the National Cash Register Company. In 1946, he joined the engineering staff of Federal Telecommunication Laboratories.

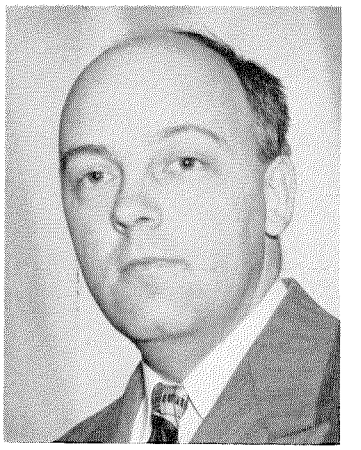
Mr. Bucher is a Senior Member of the Institute of Radio Engineers.

• • •

MILTON DISHAL was born on March 20, 1918, at Philadelphia, Pennsylvania. Temple University conferred on him the B.S. degree in 1939 and the M.A. degree in physics in 1941. He served as a teaching fellow in physics during the postgraduate years.

In 1941, he entered the employ of Federal Telecommunication Laboratories, where he is now a department head in charge of a group engaged in the development of radio receivers having special characteristics.

Mr. Dishal is a Senior Member of the Institute of Radio Engineers. He has



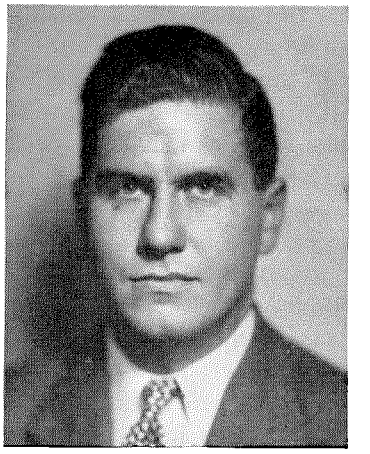
FRANK X. BUCHER

been active in the work of the Radio Technical Committee on Aeronautics. He is also serving as a part-time instructor at Polytechnic Institute of Brooklyn.

• • •

NOAH J. GOTTFRIED was born at New York, New York, on February 28, 1916. He graduated from RCA Institute in 1939.

On graduation, he joined the United States Army Signal Corps laboratories at Fort Hancock, New Jersey. In 1940, he started four years of service with the Civil Aeronautics Administration at La Guardia Field (New York), where he was in charge of maintenance of facilities for aerial navigation.



MILTON DISHAL

NOAH J. GOTTFRIED

In 1944, he became associated with Federal Telecommunication Laboratories, working on pulse-time-modulation systems. He has supervised the installation of various commercial links in North America, starting with the first in 1948 in Canada. He is now a project engineer in charge of a research and development group in the special projects laboratory.

Mr. Gottfried is a Member of the Institute of Radio Engineers.

• • •

ROBERT W. HUGHES was born on March 21, 1922, at New York, New York. He received the B.S. degree in electrical engineering from Cornell University in 1943.

ROBERT W. HUGHES

He served in the United States Army Signal Corps for the next three years and introduced the first Army pulse-time-modulation equipment in the European and Pacific theaters. From 1946 to 1948, he worked on radar plotting boards for Electronic Associates.

Since 1948, Mr. Hughes has been with Federal Telecommunication Laboratories working on pulse-time modulation and is now a project engineer.

Mr. Hughes is a Member of the Institute of Radio Engineers.

• • •

FRANK J. LUNDBURG was born at Ashton, Idaho, on December 28, 1916. He received a B.S. degree in electrical engineering from Purdue University in 1940 and an M.S. degree from Stevens Institute of Technology in 1946.

He was associated with the Columbia Broadcasting System from 1940 to 1942, serving during the last year in the high-frequency broadcasting division. Since 1942, he has been with Federal Telecommunication Laboratories.

Mr. Lundburg is a Senior Member of the Institute of Radio Engineers.

• • •

SIDNEY METZGER was born on February 1, 1917, at New York, New York. In 1937, he received the B.S. degree in electrical engineering from New York University and the M.E.E. degree from Polytechnic Institute of Brooklyn in 1950.

He was a member of the staff of the United States Army Signal Corps laboratories at Fort Monmouth, New Jersey, from 1939 to 1945. He then joined Federal Telecommunication Laboratories and is now head of the microwave communications division.

Mr. Metzger is a Senior Member of the Institute of Radio Engineers and a member of Tau Beta Pi and of Iota Alpha.

• • •

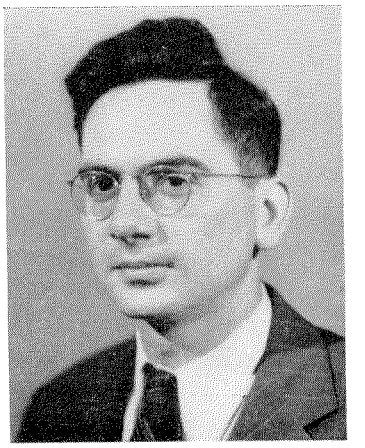
FRANK J. LUNDBURG

FRANCIS GILES OVERBURY was born at Nottingham on October 27, 1923. He received his technical training at the University College of South West England at Exeter.

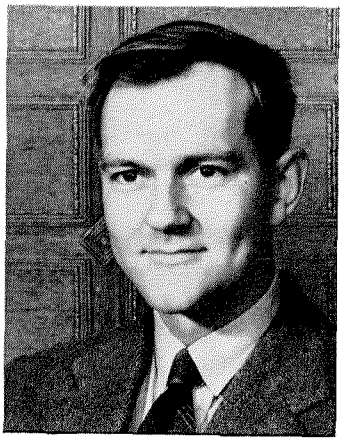
He joined the Royal Air Force in 1943, serving at the Royal Aircraft Establishment on the development of automatic radio compasses.

In 1947, Mr. Overbury came to Standard Telephones and Cables, Limited, where he has been concerned with the development of airborne instrument landing equipment and its associated test gear.

• • •



SIDNEY METZGER



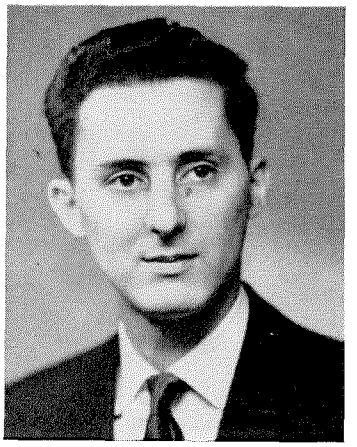
FRANCIS GILES OVERBURY

JAMES POLYZOU was born at Farrell, Pennsylvania, on January 20, 1923. He received the B.S. degree in 1947 from Carnegie Institute of Technology. During the second world war, he served as an electronics instructor in the United States Navy.

Since 1947, he has been with Federal Telephone and Radio Corporation, working on development and design of carrier telephone and telegraph equipment.

Mr. Polyzou is an Associate of the Institute of Radio Engineers.

• • •



JAMES POLYZOU

INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION

Associate Manufacturing and Sales Companies

United States of America

Capehart-Farnsworth Corporation, Fort Wayne, Indiana
Flora Cabinet Company, Inc., Flora, Indiana
Thomasville Furniture Corporation, Thomasville, North Carolina
The Coolerator Company, Duluth, Minnesota
Federal Telephone and Radio Corporation, Clifton, New Jersey
International Standard Electric Corporation, New York, New York
International Standard Trading Corporation, New York, New York
Kellog Switchboard and Supply Company, Chicago, Illinois

Great Britain and Dominions

Standard Telephones and Cables, Limited, London, England
Creed and Company, Limited, Croydon, England
International Marine Radio Company Limited, Croydon, England
Kolster-Brandes Limited, Sidcup, England
Standard Telephones and Cables Pty. Limited, Sydney, Australia
Silovac Electrical Products Pty. Limited, Sydney, Australia
Austral Standard Cables Pty. Limited, Melbourne, Australia
New Zealand Electric Totalisators Limited, Wellington, New Zealand
Federal Electric Manufacturing Company, Ltd., Montreal, Canada

South America

Compañía Standard Electric Argentina, Sociedad Anónima, Industrial y Comercial, Buenos Aires, Argentina
Standard Electrica, S.A., Rio de Janeiro, Brazil
Compañía Standard Electric, S.A.C., Santiago, Chile

Europe

Vereinigte Telephon- und Telegraphenfabriks Aktiengesellschaft Czeija, Nissl & Co., Vienna, Austria
Bell Telephone Manufacturing Company, Antwerp, Belgium
Standard Electric Aktieselskab, Copenhagen, Denmark
Compagnie Générale de Constructions Téléphoniques, Paris, France
Le Matériel Téléphonique, Paris, France
Les Téléimprimeurs, Paris, France
C. Lorenz, A.G. Stuttgart, Germany
Mix & Genest Aktiengesellschaft and Subsidiaries, Stuttgart, Germany
G. Schaub Apparatebau G.m.b.H., Pforzheim, Germany
Süddeutsche Apparatefabrik Gesellschaft m.b.H., Nuremberg, Germany
Fabbrica Apparecchiature per Comunicazioni Elettriche, Milan, Italy
Nederlandsche Standard Electric Maatschappij N.V., The Hague, Netherlands
Standard Telefon og Kabelfabrik A/S, Oslo, Norway
Standard Electrica, S.A.R.L., Lisbon, Portugal
Compañía Radio Aérea Marítima Española, Madrid, Spain
Standard Eléctrica, S.A., Madrid, Spain
Aktiebolaget Standard Radiofabrik, Stockholm, Sweden
Standard Telephone et Radio S.A., Zurich, Switzerland

Telephone Operating Systems

Companhia Telefônica Nacional, Rio de Janeiro, Brazil
Compañía de Teléfonos de Chile, Santiago, Chile
Cuban American Telephone and Telegraph Company, Havana, Cuba
Cuban Telephone Company, Havana, Cuba
Compañía Peruana de Teléfonos Limitada, Lima, Peru
Porto Rico Telephone Company, San Juan, Puerto Rico

Radiotelephone and Radiotelegraph Operating Companies

Compañía Internacional de Radio, Buenos Aires, Argentina
Compañía Internacional de Radio Boliviana, La Paz, Bolivia
Companhia Radio Internacional do Brasil, Rio de Janeiro, Brazil
Compañía Internacional de Radio, S.A., Santiago, Chile
Radio Corporation of Cuba, Havana, Cuba
Radio Corporation of Porto Rico, San Juan, Puerto Rico

Cable and Radiotelegraph Operating Companies

(Controlled by American Cable & Radio Corporation, New York, New York)

The Commercial Cable Company, New York, New York[1]
Mackay Radio and Telegraph Company, New York, New York[2]
All America Cables and Radio, Inc., New York, New York[3]
Sociedad Anónima Radio Argentina, Buenos Aires, Argentina[4]

[1] Cable service. [2] International and marine radiotelegraph services.
[3] Cable and radiotelegraph services. [4] Radiotelegraph service.

Laboratories

Federal Telecommunication Laboratories, Inc., Nutley, New Jersey
International Telecommunication Laboratories, Inc., New York, New York
Laboratoire Central de Télécommunications, Paris, France
Standard Telecommunication Laboratories, Limited, London, England