STANDARD TECHNICAL TRAINING NOTES

FOR THE

RADIO ENGINEERING TRADE GROUP

(FITTERS)

PART 1A

ELECTRICAL AND RADIO FUNDAMENTALS

(BOOK 3 — SECTIONS 13 to 20)

FOR OFFICIAL USE

STANDARD TECHNICAL TRAINING NOTES

FOR THE

RADIO ENGINEERING TRADE GROUP

(FITTERS)

PART 1A

ELECTRICAL AND RADIO FUNDAMENTALS

(BOOK 3—SECTIONS 13 to 20)

FIRST PROMULGATED BY COMMAND OF THE AIR COUNCIL
By Command of The Defence Council

STANDARD TECHNICAL TRAINING NOTES

FOR THE

RADIO ENGINEERING TRADE GROUP (FITTERS)

FOREWORD

1. These Notes are issued to assist airmen and apprentices under training as Fitters in the Radio Engineering Trade Group. Fitters in this trade group 'require a thorough knowledge of the electrical and radio principles, and the elementary mathematics, appropriate to the theory of the specified equipment' (see A.P. 3282A, Vol. 2). It is with the intention of helping to attain this standard that these Notes are written. They are not intended to form a complete text-book, but are to be used as required in conjunction with lessons and demonstrations given at the radio schools. They may also be used to assist airmen on continuation training at other R.A.F. stations.

2. The Notes, which are based on the syllabuses of training for aircraft apprentices, are sub-divided as follows:—

Part 1A: Electrical and Radio Fundamentals.

This deals with the principles of electricity, electronics and radio at a level suitable for the upper technician ranks and for technician apprentices.

Because of its bulk Part 1A has been split into three separate books: Book 1 covers basic electricity; Book 2, basic electronics; and Book 3, basic radio.

Part 1B: Basic Electricity and Radio.

This deals with the principles of electricity, electronics and radio at a level suitable for the lower technician ranks and for craft apprentices.

Part 2: Communications.

This deals with the applications of the principles covered in Parts 1A and 1B to communication systems and is intended

to be used as required by all fitters in the Radio Engineering Trade Group.

Part 3: Radar.

This deals with the applications of the principles covered in Parts 1A and 1B to radar and is intended to be used as required by all fitters in the Radio Engineering Trade Group.

3. In general, fitters employed on communications equipment will be interested mainly in Part 1A or 1B and Part 2 of these Notes. Similarly, radar fitters will be concerned mainly with Part 1A or 1B and Part 3. However it is difficult to draw a firm dividing line between the knowledge required by fitters engaged in communications and that required by radar fitters. There is considerable overlapping; much of what was once regarded as being exclusively in the province of the radar fitter is now a requirement for the communications fitter also, and *vice versa*. Therefore those under training in the radar trades may find much that is useful in Part 2, whilst those under training in the communications field may find much of interest in Part 3.

4. The Notes deal with the basic theory and the applied principles of electricity, electronics and radio in a general way. They do not cover specific details of equipment in use in the Service. Such details are to be found in the official Air Publication for the equipment and this should always be consulted during the servicing of the equipment.

5. No alteration to these Notes may be made without the authority of official Amendment Lists.

MINISTRY OF DEFENCE

August 1965

LIST OF AIR PUBLICATIONS ASSOCIATED WITH THE TRADE

Principles and Techniques

A.P. 1093   R.A.F. Signal Manual, Part 2 (Radio Communication)

A.P. 1093E   Interservices Radar Manual—Radar Techniques

A.P. 1093F   Radar Circuit Principles, with Aerials and Centimetric Techniques

A.P. 1093G   Radio Circuitry Supplement

A.P. 1093H   Suppressed Aerials

A.P. 1186V   C.V. Register of Electronic Valves

A.P. 2521A   V.H.F. Ground Station Aerial Systems

A.P. 2867   Interservices Standard Graphical Symbols

A.P. 2867A   Interservice Glossary of Terms used in Telecommunications

A.P. 2867B   Interservice Glossary of Terms used in Telecommunications (Radar)

A.P. 2878C   H.F. and M.F. Aerials for Ground Stations

A.P. 2900C   Handbook of Electronic Test Methods and Practices

A.P. 3158C   R.A.F. Technical Services Manual

A.P. 3214 (Series) The Services Textbook of Radio

Equipment

Air Publications applicable to specific radio equipment are listed in:—

A.P. 2463   Index to Radio Publications

INSTRUCTIONAL FILMS

Title	Reference
Current of Electricity	14L/52
Nuts and Bolts	14L/178
Micrometer Calipers	14L/273
Vernier Scale	14L/413
Hammers, Chisels, Punches and Drifts	14L/1605
Files and Filing	14L/1606
Spanners, Screwdrivers and Pliers	14L/1636
Taps, Dies and Reamers	14L/1727
Hacksaws, Shears and Vice Clamps	14L/1728
Locking Devices	14L/1729
Measuring and Marking—Precision Instruments	14L/1730
Transmission Lines—Maintenance of Coaxial Cables	14L/3280
Transmission Lines and Waveguides	14L/3288

INSTRUCTIONAL FILM STRIPS

LIST OF SYMBOLS AND ABBREVIATIONS

TABLE 1

Greek Letters Used in the Text

Letter			Letter		
Small	**Capital**	**Name**	**Small**	**Capital**	**Name**
α	—	Alpha	λ	—	Lambda
β	—	Beta	μ	—	Mu
γ	—	Gamma	π	—	Pi
δ	Δ	Delta	ρ	—	Rho
ε	—	Epsilon	σ	—	Sigma
η	—	Eta	φ	Φ	Phi
θ	—	Theta	ω	Ω	Omega
$\varkappa$	—	Kappa			

TABLE 2

Meaning of Symbols Used in the Text

Letter	Meaning	Letter	Meaning
A	Ampere, Amplification	j	Vector operator $= \sqrt{-1}$
B	Magnetic flux density, Susceptance, Bandwidth	k	Coupling coefficient, Kilo—(prefix), constant
C	Capacitance	l	Length
D	Electric flux density, Distance	m	Modulation factor, Metre, Mass, Milli—(prefix)
E	Electromotive force, Electric field strength	n	Number
F	Farad, Factor, Force	p	Pico—(prefix)
G	Conductance, Giga—(prefix)	q	Instantaneous charge
H	Magnetic field strength, Henry	r	Length (in polar co-ordinates)
I	Electric Current	r_a	Anode slope resistance

(continued overleaf)

J	Joule	s	Second
L	Inductance	t	Time, Temperature
M	Mutual inductance, Mega—(prefix)	u	Velocity
N	Number, Noise factor, Revs. per minute	v	Instantaneous potential difference
		x	Distance, Length
P	Power	y	Length
Q	Quantity or charge of electricity, Coil amplification factor	α	Angle, Number
R	Resistance	β	Number, Feedback factor
S	Magnetic reluctance	γ	Propagation constant
T	Temperature (Absolute), Period, Transit Time, Transformation ratio	δ	Small increment, Loss angle
		ε	Base of natural logs = $2 \cdot 71828$
		η	Efficiency
V	Potential difference, Volt, Volume	θ	Angle
W	Energy or work, Watt	$\varkappa_r$	Dielectric constant
X	Reactance	$\varkappa_0$	Permittivity of free space
Y	Admittance	λ	Wavelength
Z	Impedance	μ	Permeability, Valve amplification factor
a	Area		
c	Velocity of light, Cycle	μ_0	Permeability of free space
d	Distance	μ_r	Relative permeability
e	Instantaneous e.m.f., Electron charge	π	Ratio of circumference to diameter of a circle = $3 \cdot 14159$
f	Frequency	ρ	Specific resistance
g_c	Valve conversion conductance	ϕ	Angle
g_m	Valve mutual conductance	Φ	Magnetic flux
i	Instantaneous current	ω	Angular velocity = $2\pi f$
		Ω	Ohm
		σ	Specific conductance

(continued overleaf)

TABLE 3

Prefixes for Multiples and Sub-multiples

Multiple or sub-multiple	Name	Prefix	Multiple or sub-multiple	Name	Prefix
$1{,}000{,}000{,}000 = 10^9$	Giga-	G	$\dfrac{1}{1{,}000{,}000} = \dfrac{1}{10^6} = 10^{-6}$	Micro-	μ
$1{,}000{,}000 = 10^6$	Mega-	M			
$1{,}000 = 10^3$	Kilo-	k	$\dfrac{1}{10^{12}} = 10^{-12}$	Micro-micro- or Pico	$\mu\mu$ or p
$\dfrac{1}{1{,}000} = \dfrac{1}{10^3} = 10^{-3}$	Milli-	m			

TABLE 4

Abbreviations of Units

Unit	Abbreviation	Unit	Abbreviation
Ampere	A	Gramme	g
Ampere-hour	Ah	Henry	H
Ampere-turn	AT	Joule	J
Cycles per second	c/s	Metre	m
Decibel	db	Ohm	Ω
Degree	° Centigrade = C. Fahrenheit = F.	Second	s or sec.
		Volt	V
Electron-volt	eV	Watt	W
Farad	F	Weber	Wb

CONTENTS OF PART 1A

CONTENTS

Amendment Record Sheet

Foreword

List of Air Publications Associated with the Trade

List of Symbols and Abbreviations

SECTIONS

(A detailed contents list is given at the beginning of each Section and Chapter)

SECTION 13

TRANSMITTER PRINCIPLES

SECTION 13

TRANSMITTER PRINCIPLES

Chapter 1—Generation and Modulation of Radio Frequency Energy

SECTION 13

CHAPTER 1

GENERATION AND MODULATION OF RADIO FREQUENCY ENERGY

GENERATION AND MODULATION OF
RADIO FREQUENCY ENERGY

Introduction

1. A radio transmitter (or 'sender') supplies radio frequency power to an aerial which, in turn, radiates electromagnetic (e.m.) energy into space. By this means, information can be conveyed from one point to a distant point or to a number of distant points without using connecting wires. *Wireless* systems are able to effect communication over hostile or difficult country, and mobile units (e.g. aircraft) are readily controlled by wireless (Fig. 1(a)). In *radar*, the e.m. energy radiated from the aerial does not carry 'intelligence' for communication purposes. With this system (Fig. 1(b)) part of the e.m. energy, on striking a 'target', is reflected and received back at the source where information about the target's height, distance, bearing, etc., may be correlated (see Part 3, Radar).

Relationship Between Frequency and Wavelength

2. If oscillating movements create a disturbance at regular intervals on the surface of a still pond, a series of waves travel away from the source of oscillations with a certain velocity, say u feet per second. If the oscillations occur f times per second, f waves are created every second. Thus, by the time


TRANSMITTER AERIAL (a)

RADAR TRANSMITTER AND RECEIVER (b)
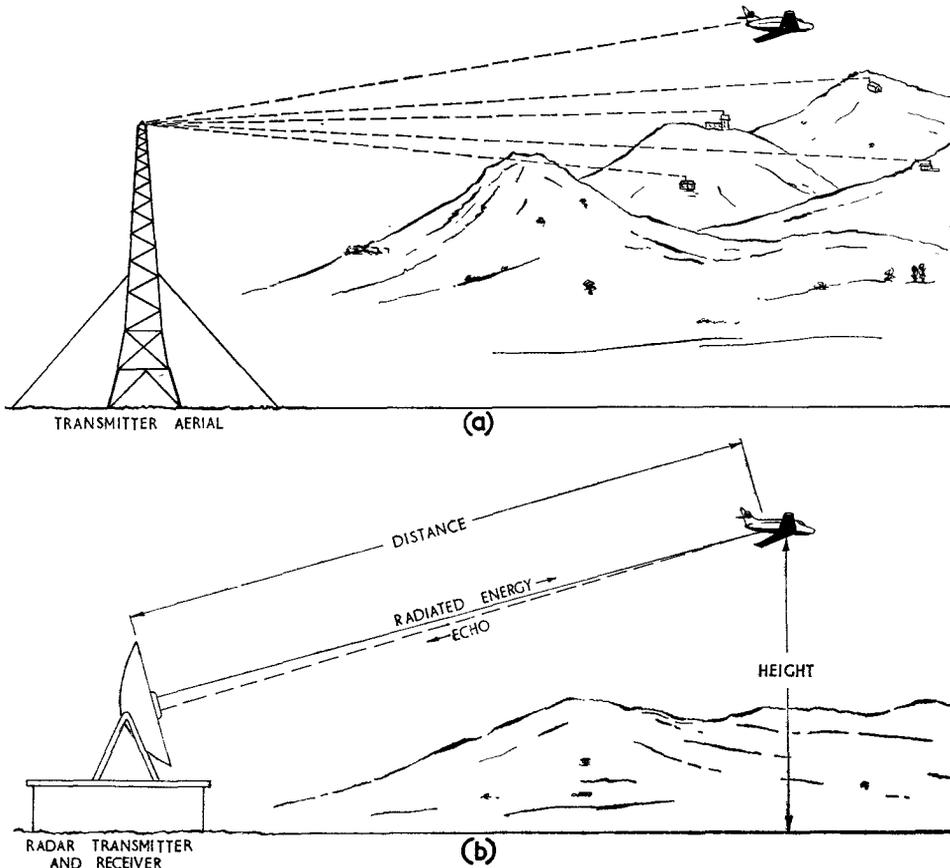
Fig. I. WIRELESS AND RADAR SYSTEMS.

the first wave has travelled u feet, f oscillations will have occurred so that there will be f waves in a distance of u feet (Fig. 2). The *length* of each wave is, therefore u/f feet. If the '*wavelength*' is written as λ, then a simple equation results:—

$$\lambda = \frac{u}{f}.$$

3. The same relationship between wavelength, velocity and frequency exists for radio waves, but the velocity with which radio waves travel out from the aerial is very much higher than that of the waves

Frequencies Necessary for Effective Radiation

4. For adequate radiation of e.m. energy from an aerial it is necessary that the current established in the aerial be at radio frequency. The efficiency of aerials as radiators of e.m. waves depends on a relationship between the size of the aerial and the frequency of the current in the aerial (or the wavelength of the energy radiated). The amount of energy radiated is very small unless the length of the aerial approaches the order of magnitude
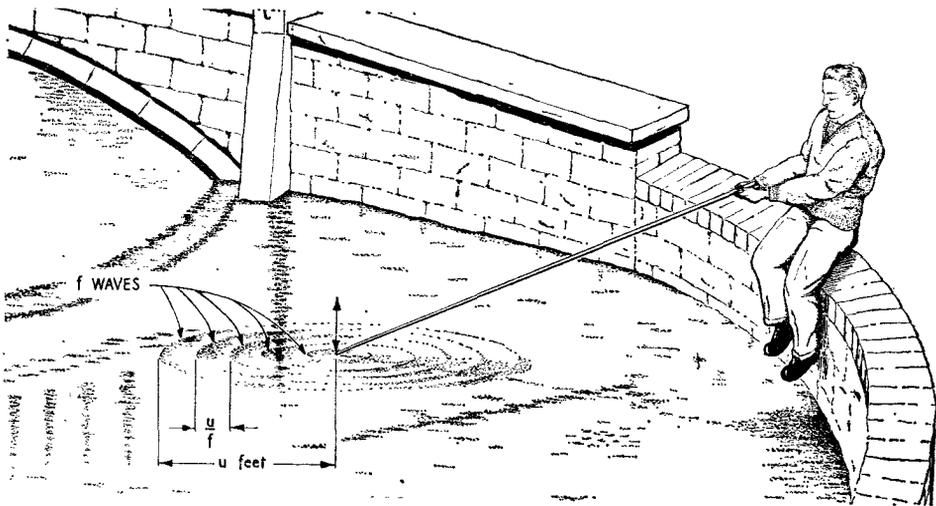
Fig. 2. RELATIONSHIP BETWEEN FREQUENCY AND WAVELENGTH.

in the pond. Radio waves travel at the velocity of light (c), namely 186,000 miles per second or 300,000,000 metres per second. The relationship between wavelength in metres and frequency in cycles per second for a radio wave is therefore:—

$$\lambda = \frac{c}{f}$$

$$\therefore \lambda = \frac{3 \times 10^8}{f} \text{ (metres)}.$$

From this expression if the wavelength is known the corresponding frequency can be calculated, and *vice versa*. The relationship shows, for example, that a *low* frequency corresponds to a *long* wavelength, and a *short* wavelength corresponds to a *high* frequency (see Book 2, Sect. 11, Chap. 1, Table 1).

of a half-wavelength. Thus, at the mains frequency of 50 c/s a half-wavelength aerial approximately 1,750 miles long would be required to give satisfactory radiation. On the other hand, at a frequency of 10 Mc/s a considerable amount of energy would be radiated from a half-wavelength aerial of length 15 metres (49 feet). The *higher* the frequency, the *smaller* is the aerial structure required to give efficient radiation (Fig. 3). For this reason, the lowest frequency used in the Service for wireless communication is of the order of 100 kc/s. The actual frequency used depends on a number of factors including the distance over which communication is required, the atmospheric conditions and the type of transmission (television, broadcasting, point-to-point communication, etc.).
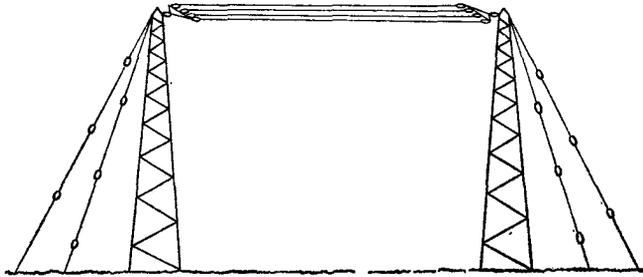
(a)   LARGE AERIAL STRUCTURE FOR LOW FREQUENCIES

(b) SMALL AERIAL STRUCTURE FOR HIGH FREQUENCIES

Fig. 3.   RELATIONSHIP BETWEEN FREQUENCY
AND SIZE OF AERIAL.

Communication Transmitter

5. The characteristics and operation of the basic circuits which go into a typical modern communication transmitter are dealt with in previous Sections. All that remains before the operation of a transmitter can be understood is to combine the necessary circuits in the correct manner. The stages in such a transmitter are shown in block form in Fig. 4 and include:—

(a) *A source of power* for supplying h.t., l.t. and g.b. to the various stages in the transmitter; the power supply may take

(d) A *means for causing the radio frequency current to convey 'intelligence'.* The r.f. component is normally called the '*carrier*' since, in effect, it carries the necessary information.

(e) An *aerial system* that is coupled to the final r.f. power amplifier so that the r.f. currents set up in the aerial give rise to radiation of e.m. waves.

PRODUCTION OF CARRIER

Simple Self-exciting Transmitter

6. A radio transmitter may be constructed by coupling an aerial to the output of any

Fig. 4. STAGES IN A COMMUNICATION TRANSMITTER.

any of the forms described in Sect. 9 (Power Supplies).

(b) An *oscillator* which generates the r.f. oscillation of constant frequency; this may be the same frequency as, or an exact fraction of that finally radiated depending on whether frequency-multipliers are included or not. The oscillator may take any of the forms described in Book 2, Sect. 12 (Valve Oscillators).

(c) A *r.f. power amplifier* section to supply the necessary power to the aerial; this section may include frequency-multipliers. The power amplifier stage may take any of the forms described in Book 2, Sect. 11, Chap. 2 (R.F. Power Amplifiers).

of the oscillators discussed in Book 2, Sect. 12. Fig. 5(a) shows the arrangement in schematic form, and Fig. 5(b) illustrates a simple circuit, where the oscillator is of the series-fed Hartley type. The frequency at the output is determined mainly by the values of L and C, and it is variable between the limits set by the range of the tuning capacitor. The e.m. energy radiated from the aerial is a r.f. signal of constant amplitude and constant frequency (Fig 5 (c)), and is termed a 'Type A_o Wave' or a '*continuous wave*' (C.W.).

7. The self-exciting transmitter can be designed to give a high power output, and it is the system adopted for most radar transmitters (see Part 3). However, for

communication purposes, frequency stability is of prime importance for two reasons:—

(a) The transmitter frequency must remain constant at its assigned value otherwise interference with other transmissions on adjacent channels will result.

(a)

(b)

(c)
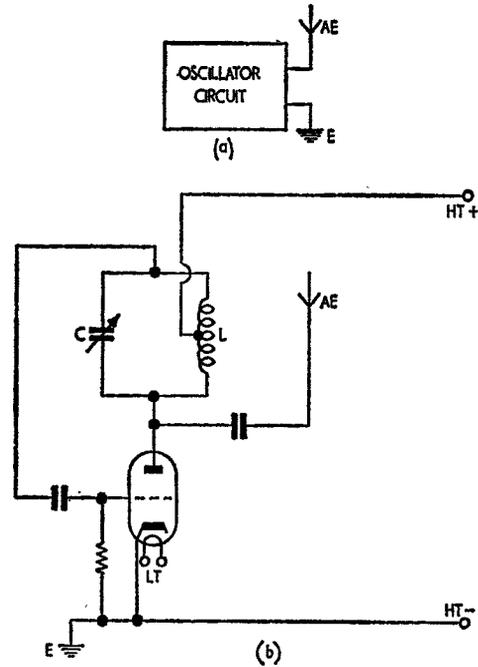
Fig. 5. SIMPLE SELF-EXCITING TRANSMITTER.

(b) If the frequency of the radiated energy is varying, the operator at the receiving end is required continually to re-adjust his receiver.

The frequency stability of a high-power oscillator working directly into the aerial is poor, for the reasons given in Para. 8, and to ensure a stable frequency of transmission, modifications are made to the basic circuit of Fig. 5(b).

8. **Causes of frequency instability.** The causes of frequency instability in oscillators

are described in detail in Book 2, Section 12, Chap. 1, Para. 23. For reference, they are summarized below:—

(a) Instability because of changes in L, C, and R values in the oscillatory circuit caused by:—

(i) temperature variations:

(ii) mechanical vibration:

(iii) variations in the aerial impedance, this being reflected back into the oscillatory circuit; such variations are caused by the motion of the aerial.

(b) Instability because of changes in the maintaining system caused by:—

(i) variations in supply voltages:

(ii) variations in the valve constants.

(c) Instability because of changes in the load to which the oscillator is connected—in this case the aerial.

Master Oscillator—Power Amplifier System

9. The disadvantages of the simple self-exciting transmitter in relation to frequency stability are reduced in the master oscillator—power amplifier (m.o.—p.a.) type of transmitter. This consists essentially of a carefully controlled oscillator, which produces a r.f. signal of very stable frequency working into an amplifier which in turn, supplies power
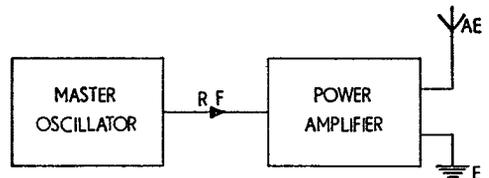
Fig. 6. MASTER OSCILLATOR-POWER AMPLIFIER SYSTEM.

to the aerial. The general layout is shown in Fig. 6. Frequency stability, coupled with a high power output are achieved as follows:—

(a) **Master oscillator.**

(i) The m.o. is run at a low power level to limit variations in temperature; in some instances, the tuned circuit (or crystal) is placed in a temperature-controlled oven.

(ii) The m.o. is rigidly constructed to reduce the effect of mechanical vibration.

(iii) The aerial is no longer directly connected to the oscillator so that movement of the aerial has little effect on the oscillator frequency.

(iv) The supply voltages to the m.o. may be stabilized.

(v) A valve with a high value of r_a is usual in the m.o.

(vi) The load on the oscillator is kept as small and as constant as possible, by coupling the m.o. 'loosely' to the p.a., by using a buffer amplifier between

frequency, steps are taken (including neutralization where necessary) to ensure that the p.a. operates in a stable state.

10. The basic circuit of a m.o.-p.a. type transmitter, incorporating many of the features discussed in Para. 9, is illustrated in Fig. 7.

(*a*) **Master oscillator, V_1.**

(i) The oscillator is a series-fed Hartley.

(ii) The h.t. supply to V_1 is reduced by
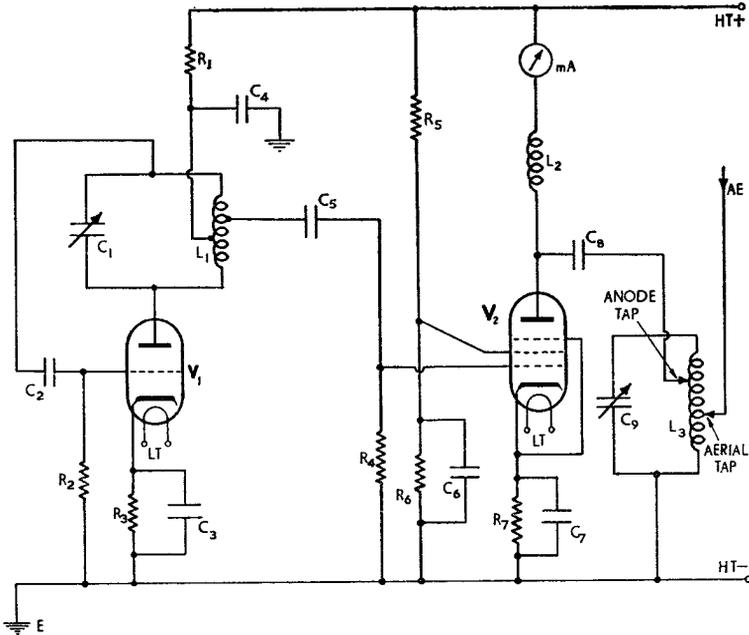
Fig. 7.   CIRCUIT OF BASIC M.O.—P.A. TRANSMITTER.

the m.o. and the p.a. or by using an electron-coupled oscillator.

(*b*) **Power amplifier.**

(i) The p.a. isolates the m.o. from the aerial to prevent frequency instability caused by variations in the aerial.

(ii) The p.a. operates at a high power level and at a high level of efficiency (Class B or Class C bias conditions) and converts the r.f. drive from the m.o. to a sufficiently high power level to feed the aerial without appreciably loading the m.o.

(iii) Since the p.a. is 'driven' from the m.o. and must be prevented from self-oscillating and taking control of the

the p.d. developed across R_1 (de-coupled by C_4) for low power operation of the m.o. and improved frequency stability; in certain circumstances, this supply is stabilized.

(iii) The frequency is determined by L_1 C_1; these may be temperature-compensated or they may be placed in a temperature-controlled oven for improved frequency stability.

(iv) C_2 R_2 give automatic Class C bias for high efficiency and low power dissipation, and safety cathode bias from C_3 R_3 is inserted to safeguard the circuit in the event of cessation of oscillations.

(v) The r.f. signal generated by V_1 is capacitively coupled by C_5 to the p.a. grid. To reduce the loading on the m.o. for improved frequency stability, the m.o. is 'loosely' coupled to the p.a. This condition is obtained by taking the m.o. output from a tapping on L_1 to provide an auto-transformer match.

(b) **Power amplifier V_2.**
(i) The drive from the m.o. is sufficient to run the p.a. into grid current. Thus, the p.a. operates under Class C self-bias conditions (C_5 R_4) to give a high conversion efficiency. Safety cathode bias from C_7 R_7 is inserted to safeguard the circuit if the r.f. drive from the m.o. ceases.

(ii) The tank circuit L_3 C_9 is shunt-fed from L_2 through C_8, and it is tuned to resonance with the r.f. signal from the m.o. This condition is indicated by a 'dip' in the reading of the milliammeter in V_2 anode circuit. At resonance, the oscillatory voltage across the tank circuit is at a maximum and the anode voltage during the conducting portion of the cycle is at a *minimum,* as is the mean anode current (see Book 2, Sect. 11, Chap. 2, Para. 11).

(iii) Maximum power is developed in the tank circuit when its dynamic impedance is matched to the anode slope resistance of the p.a. valve. This is achieved by adjusting the *anode tap* which provides an auto-transformer match.

(iv) To transfer maximum power from the tank circuit to the aerial, the coupling between them is adjusted so that the aerial impedance is matched to the tank circuit resonant impedance. In Fig 7, this is achieved by adjustment of the *aerial tap* which provides an auto-transformer match.

(v) The h.t. supply to the p.a. stage is not reduced so that it is much higher than that applied to the m.o. and a high power output from the p.a. is possible.

(vi) R_5 R_6 C_6 act as a potentiometer chain and r.f. de-coupling network to provide the required d.c. potential at the screen grid.

(vii) A pentode, with its low value of C_{ag} is used to reduce Miller feedback and subsequent tendency to self-oscillation in the p.a. Neutralization is not necessary in this case.

11. **Variations in the basic m.o.-p.a. circuit.** The m.o.-p.a. circuit given in Fig. 7 suffers from certain limitations:—

(a) **Stability of frequency.** Although the m.o. is coupled only loosely to the p.a., the r.f. drive is sufficient to run the p.a. into grid current, thereby providing the necessary Class C self-bias conditions at the p.a. for high efficiency and high power output. However, the p.a. grid power is drawn from the *oscillator* and the load thus placed on the m.o. gives a tendency to frequency instability. This is reduced by inserting a buffer amplifier between the m.o. and the p.a. as shown schematically in Fig. 8. The buffer amplifier supplies the p.a. grid power, but the former is operated under zero grid current conditions
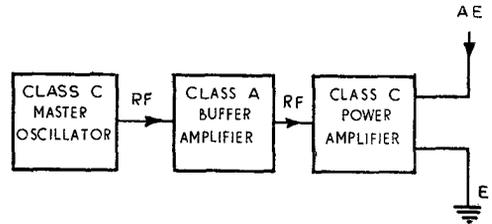
Fig. 8. USE OF BUFFER AMPLIFIER.

(i.e. Class A or Class B_1 bias) so that the power supplied to the buffer from the m.o. is negligible. The result is that the load on the m.o. is small and constant, and frequency stability is improved. Similar results may be obtained by using an electron-coupled oscillator as the m.o.

(b) **Power output.** The power output available with a single p.a. valve may be less than that required for adequate radiation from the aerial. Where a high power output is called for, valves in push-pull or in parallel (or a combination of both) may be used. Fig. 9 shows two triodes connected in Class C push-pull. Because of the wattage dissipation difficulties associated with tetrode and pentode valves at high powers, triodes are preferred in the p.a. stage at power outputs in excess of about 5 kW. Where triodes are used neutralization is essential (Fig. 9).
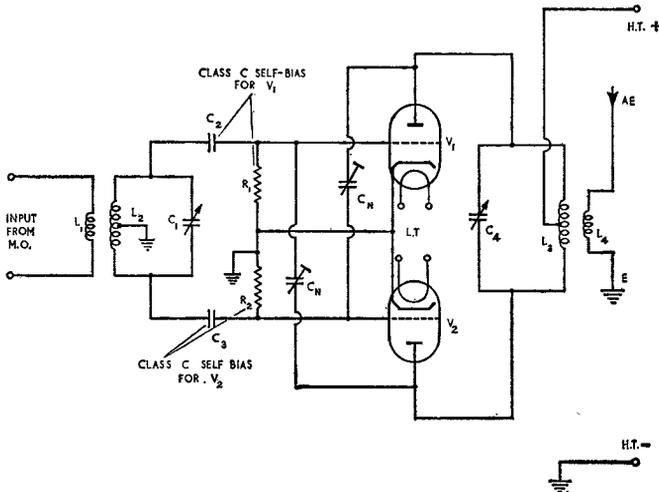
Fig. 9. PUSH-PULL TRIODE POWER AMPLIFIER.

(*c*) **Frequency multiplication.** In many instances, especially at v.h.f., the frequency of the output at the aerial is required to be higher than that at which the oscillator is working. Frequency-multiplier stages are then inserted between the m.o. and the p.a. Frequency stability is even more important in such cases because any deviation in frequency at the m.o. is multiplied to give a greater deviation at the aerial. In such systems it is, therefore, normal to use a crystal-controlled electron-coupled oscillator with its improved frequency stability. The limitation is that the frequency can be altered only by changing the crystal. A block schematic diagram of a typical v.h.f. transmitter is illustrated in Fig. 10.

TRANSMISSION OF INTELLIGENCE

Introduction

12. The preceding paragraphs have shown how a high-power r.f. carrier at the required frequency is obtained at the transmitter aerial. However, for this carrier to be of any practical value for communication purposes it must be caused to carry the desired information. The succeeding paragraphs describe the various methods used to transmit messages.

13. Two methods of transmitting messages by wireless are in general use in the Service; they are telegraphy and telephony.

(*a*) **Telegraphy.** In this system the information is sent in the form of a code in one of the following forms:—

(i) *Morse telegraphy*. Signals are formed in accordance with the morse code.

(ii) *Printing telegraphy*. The received signals are automatically recorded in printed characters.

(iii) *Mosaic telegraphy*. The patterns forming the characters are made up from units transmitted as individual signal elements.

(iv) *Facsimile telegraphy*. Transmission of still pictures, printed matter, etc., producing at the receiver a permanently recorded copy of the original.


Fig. 10. SCHEMATIC DIAGRAM OF TYPICAL V.H.F. TRANSMITTER.

Telegraphy gives accurate copy of the message being relayed but it has the disadvantage that trained operators are normally necessary.

(b) **Telephony.** In this system, the information is sent in the form of speech. Telephony provides immediate personal contact between individuals at each end of the wireless link. For example, the pilot of an aircraft is able to 'talk' directly to the ground controller at an airfield without having to rely on trained operators for the transmission and reception of the messages. The speed of communication is therefore increased. Telephony is, however less secure and less accurate than telegraphy.

Keying a Transmitter

14. The continuous wave radiated from the aerial of a transmitter may be interrupted by 'keying' the transmitter in accordance with a pre-arranged telegraphic code to produce keyed C.W. A transmitter operated in this way is called a 'wireless telegraphy' (W.T.) transmitter. With on-off keying in a W.T. transmitter, the r.f. carrier is switched on and off according to a code which relates the letters and figures to different combinations of time intervals. With the morse code for instance, the transmitter is keyed in such a way that the radiation from the aerial occurs for short and long periods of time to correspond with the dots and dashes of the code. Fig. 11(a) shows a continuous wave and Fig. 11(b) shows the result of keying the transmitter to produce the letter 'L' by morse code.

15. With on-off keying in a W.T. transmitter certain requirements have to be met:—

(a) The r.f. current in the aerial must be arrested *completely* when the key is open (on *'space'*). On *'mark'* (key closed) full radiation is to take place.

(b) The r.f. current must not be interrupted or started *too abruptly* otherwise the square waves so produced (Fig. 11(b)) result in the generation of a large number of harmonics which cause interference in the form of 'key clicks' in neighbouring receivers. This is normally prevented by inserting filters in the keying leads to 'round off' the steep-sided rise and fall of the keying waveform (Fig. 11(c)).

(c) The position in the transmitter at which keying is carried out must be such that:—

(i) keying does not cause frequency instability in the radiated signal;

(ii) the keying system does not involve the direct switching of high voltage or high current circuits, otherwise there is the danger of high voltage to the operator and arcing at the key contacts.
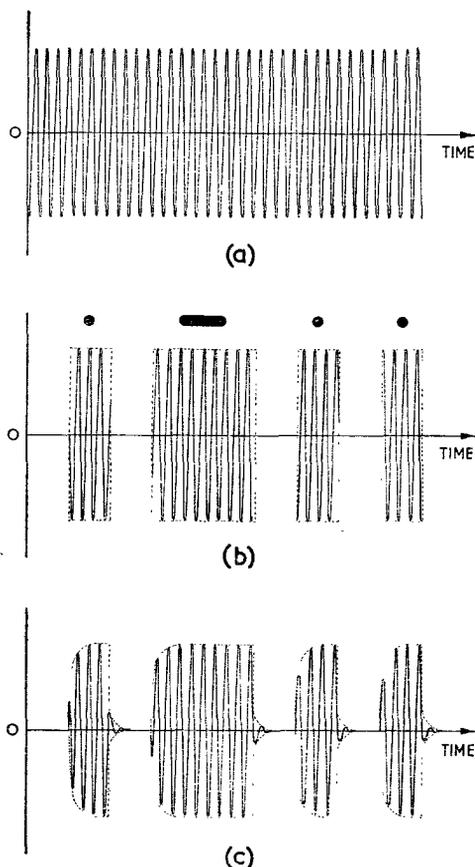
Fig. 11. KEYED C.W.

In practice it is usual for the low-power stages of the transmitter to be keyed, the position chosen being sufficiently far from the m.o. to prevent frequency instability, and yet not so far up the transmitter chain that high-power circuits have to be broken. The usual position is in the frequency multiplier or driver stages.

16. **Methods of keying.** There are many ways of keying a transmitter. One of the simplest on-off methods, known as *'cathode*

keying', is illustrated in Fig. 12. The key, either directly or by means of a contact of a relay operated by the key, interrupts both the anode and the grid circuits. On 'space' (key open) the cathode circuit is broken and the anode current commences to fall. At
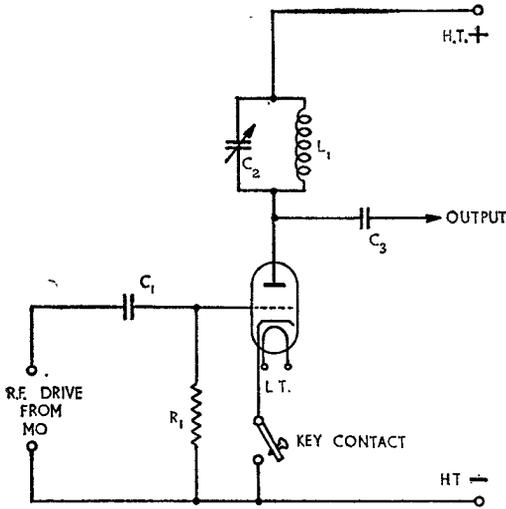
Fig. 12.   CATHODE KEYING.

the same time, the grid return to the cathode is broken so that the negative charge on the grid (built up by the r.f. drive from the m.o.) cannot leak away to the cathode and the bias quickly builds up to cut the valve off. On 'mark' (key closed) these conditions are reversed to give normal operation. This is a fairly rapid method of keying which is useful in low-power transmitters. Other, more efficient methods of keying high-power W.T. transmitters exist and these are discussed in detail in Part 2 (Communications).

Modulation

17. 'Modulation' is the general name given to the process of impressing intelligence upon a r.f. carrier in such a way that the characteristics of the carrier are modified. Modulation may be effected by modifying the *phase* the *frequency* or the *amplitude* of the carrier in accordance with the characteristics of the modulating signal. In telephony, the modulating signal is in the form of speech and a transmitter modulated by speech voltages, is called a 'radio telephony' (R.T.) transmitter. In telegraphy, keyed a.f. tones may be used as the modulating signal. In addition to phase, frequency and amplitude modulation,

a system known as *pulse modulation* may be used. The characteristics of all these systems of modulation are summarized in the succeeding paragraphs.

18. **Phase modulation.** In this system, the audio modulating signal is used to shift the *phase* of the carrier without affecting the carrier amplitude. The *change* in the phase angle above and below that of the unmodulated condition is proportional to the *amplitude* and *sign* of the modulating signal, and the *rate* at which the phase angle alters is proportional to the *frequency* of the modulating signal. Phase modulation has few direct applications in the Service but it is considered briefly in Part 2 (Communications).

19. **Frequency modulation.** In this system, the audio modulating signal is used to shift the *frequency* of the carrier without affecting the carrier amplitude. The *change* in the frequency of the carrier above and below that of the unmodulated condition is proportional to the *amplitude* and *sign* of the modulating signal, and the *rate* at which the carrier frequency varies is proportional to the *frequency* of the modulating signal (Fig. 13). Frequency modulation and frequency-shift-keying (f.s.k.)—a form of frequency modulation—are used extensively in the
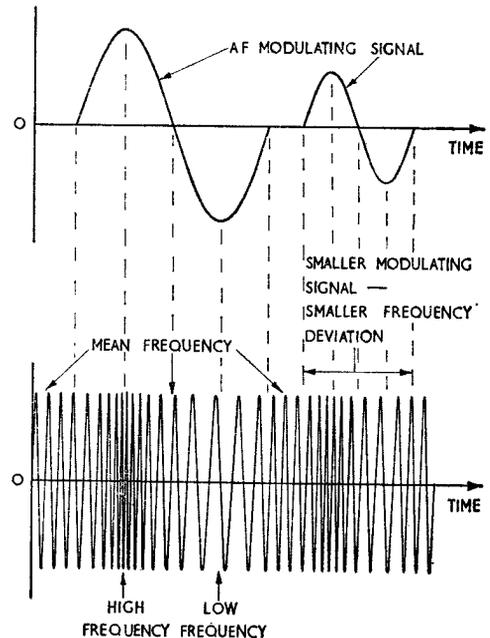

Fig. 13.   FREQUENCY MODULATION.

Service and they are considered in detail in Part 2 (Communications).

20. Amplitude modulation. In this system, the audio modulating signal is used to modify the *amplitude* of the carrier without affecting the carrier frequency. The *change* in the amplitude of the carrier above and below that of the unmodulated condition is proportional to the *amplitude* and *sign* of the modulating signal, and the *rate* at which the amplitude of the carrier varies depends on the *frequency* of the modulating signal (see Fig. 14). Amplitude modulation, together with a
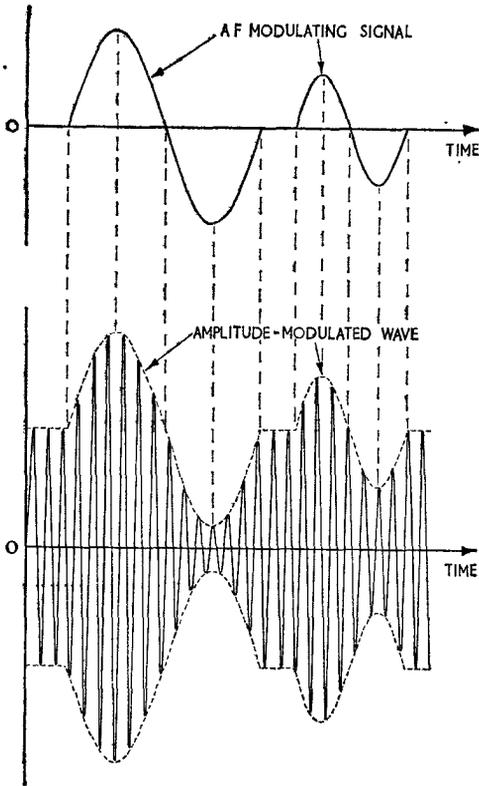

Fig. 14. AMPLITUDE MODULATION.

variation termed '*single sideband modulation*' (s.s.b.) are used extensively in the Service. Since it is so widely used, the outline of an amplitude-modulated transmitter is discussed in Para. 22 and the system is considered in detail in Part 2 (Communications).

21. Pulse modulation. In pulse modulation, the transmitter is switched on and off in such a way that only *very short duration pulses* of energy are radiated from the aerial; the pulses

occur at regular intervals of time but the time interval between the pulses is long compared with the duration of the pulse (Fig. 15). Pulse modulation can be used to convey intelligence. This may be done in at least three ways:—
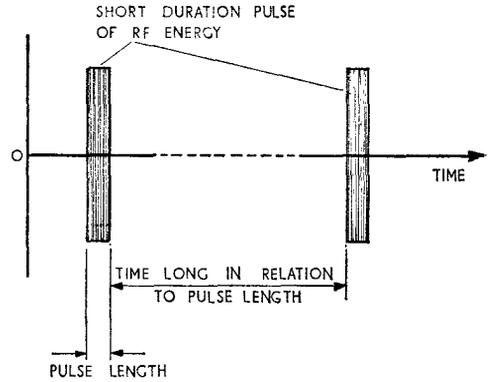
Fig. 15. PULSE MODULATION.

(*a*) **Pulse-amplitude modulation.** The pulses, of uniform length, occur at regular intervals of time but their *amplitude* is varied in accordance with the modulating signal (Fig. 16(b)).

(*b*) **Pulse-length modulation.** The pulses, of constant amplitude, occur at regular intervals of time but their *length* is varied in accordance with the modulating signal (Fig. 16(c)).

(*c*) **Pulse-position modulation.** The amplitude and length of the pulses are kept constant, but the pulse *position* is varied relative to its 'normal' position by an amount depending on the modulating signal; that is, the modulating signal causes the pulse to be delayed or advanced relative to its expected time of arrival (Fig. 16(d)).

These systems of modulation will be further considered in Part 2 (Communications).

Amplitude Modulation

22. The unmodulated C.W. radiation from a transmitter aerial is a radio frequency carrier of constant frequency and constant amplitude, the waveform of which is shown in Fig. 17(a). Fig. 17(b) shows the waveform of a simple audio frequency modulating voltage. When the r.f. carrier voltage and the a.f. modulating voltage are combined in amplitude modulation, the resultant waveform is as shown in Fig. 17(c); that is,

(a) MODULATING SIGNAL

(b) PULSE – AMPLITUDE MODULATION

(c) PULSE – LENGTH MODULATION

(d) PULSE – POSITION MODULATION
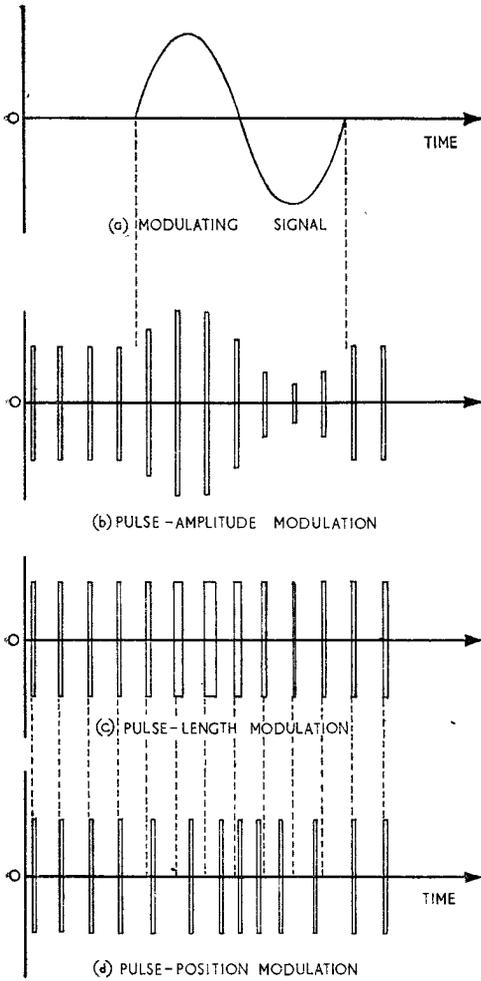
Fig. 16. METHODS OF PULSE MODULATION.

Fig. 17. AMPLITUDE-MODULATED WAVEFORM.

b=a : m=1
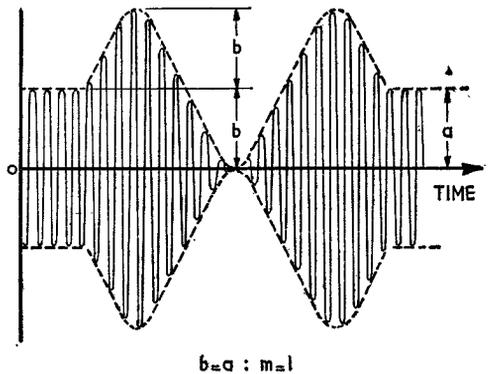
Fig. 18. DEPTH OF MODULATION.

the 'envelope' of the r.f. carrier varies in amplitude in accordance with the modulating signal.

23. **Modulation factor.** In an amplitude-modulated wave, this is the ratio of half the difference of the maximum and minimum amplitude to the mean amplitude of the wave. Thus, for Fig. 17(c), the modulation factor m is:—

$$m = \frac{\frac{1}{2}[(a + b) - (a - b)]}{a}$$

$$\therefore m = \frac{b}{a}.$$

The modulation factor expressed as a percentage is termed the 'modulation percentage' or the '*depth of modulation*'. Thus, if b = a the modulation factor m is unity and the depth of modulation is 100%. This condition is shown in Fig. 18.

24. **Sidebands.** When a r.f. carrier of frequency f_c is amplitude-modulated by an a.f. modulating signal of frequency f_m, the resultant waveform can be shown to consist of *three radio frequency waves* of

constant amplitude (Fig. 19). The three
frequencies are:—

(a) $f_c - f_m$, the lower side frequency;
(b) f_c, the original carrier frequency;
(c) $f_c + f_m$, the upper side frequency.

(a) $f_c - f_m$

(b) f_c

(c) $f_c + f_m$
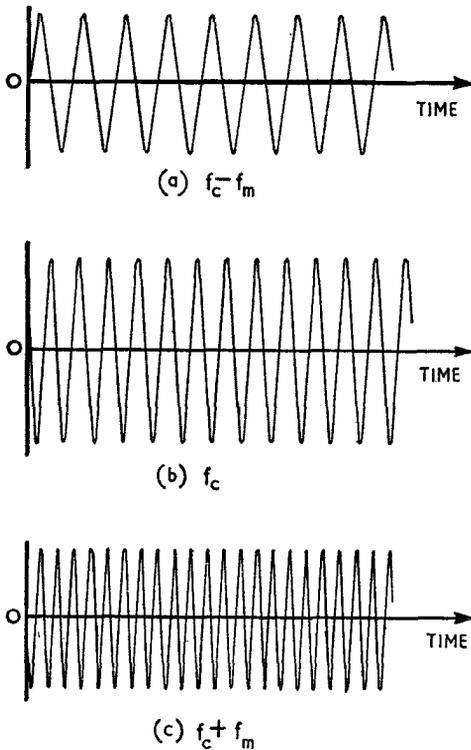
Fig. 19. SIDE FREQUENCIES PRODUCED IN
AMPLITUDE MODULATION.

Amplitude modulation of the carrier of
frequency f_c by speech voltages produces
a *band* of side frequencies above and below
f_c; the band of side frequencies below f_c
is termed the '*lower sideband*' and that above
f_c is termed the '*upper sideband*'. Fig. 20
illustrates the band of frequencies produced
when a r.f. carrier of frequency 1,000 kc/s
is amplitude-modulated by speech voltages
within the frequency range 100 c/s to 4,000
c/s. For the original speech to be faithfully
reproduced at the receiving end, the trans-
mitter circuits and the receiver circuits must
be capable of passing all frequencies from
996 kc/s to 1,004 kc/s; that is, a *bandwidth* of
8 kc/s (twice the highest modulating frequency)
is required.

25. Anode modulation. The next step is to
consider how amplitude modulation may be
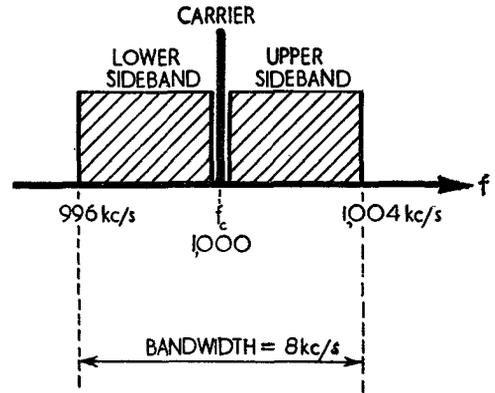
Fig. 20. SIDEBANDS IN AMPLITUDE-MODULATED
WAVE.

effected. In a Class C power amplifier stage
of a transmitter, the amplitude of the r.f.
output is directly proportional to the power
input to the stage from the supply. Thus, to
obtain an amplitude-modulated output, all
that is required is to connect the source of
a.f. modulating voltage in series with the h.t.
voltage supplied to the anode of the p.a.
valve (Fig. 21). By varying the h.t. in accor-
dance with the modulating voltage, the
amplitude of the r.f. carrier rises and falls
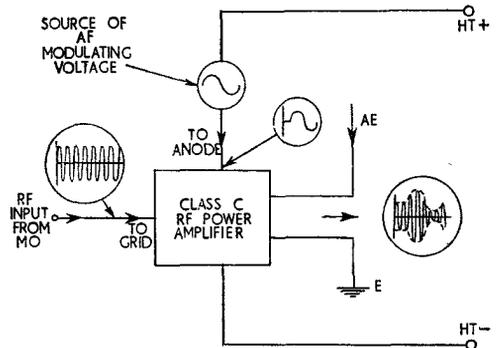
Fig. 21. SCHEMATIC DIAGRAM OF AMPLITUDE
MODULATION SYSTEM.

accordingly. Fig. 22 shows the basic circuit
of a m.o.-p.a. transmitter using this method of
modulation. Neutralization has been omitted
for reasons of simplicity. The oscillator V_1
is a conventional series-fed Hartley that
supplies the r.f. drive to the p.a. valve V_2. The
latter operates as a conventional Class C
amplifier stage whose anode current is in the
form of r.f. pulses at a frequency determined
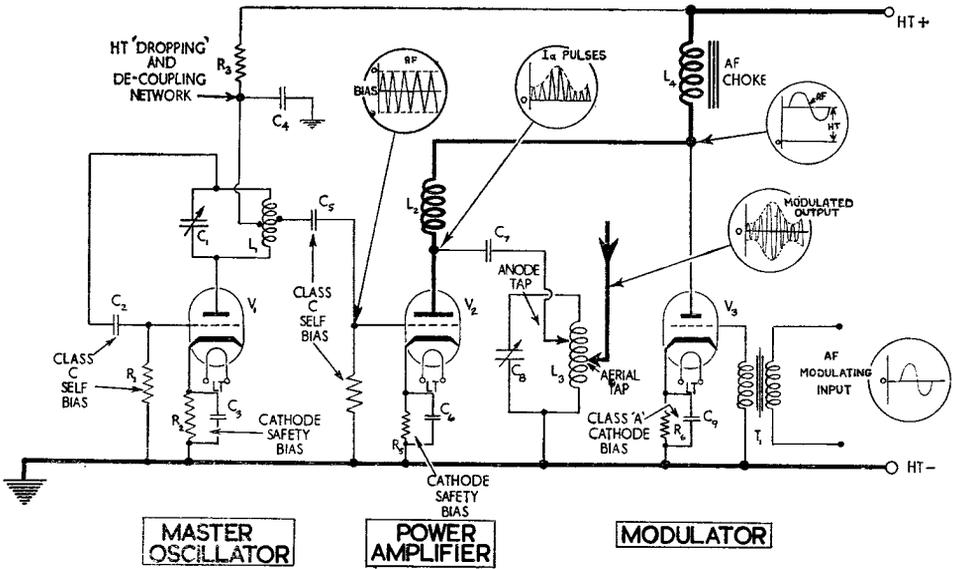
Fig. 22.   SIMPLE ANODE-MODULATED TRANSMITTER.

by the m.o. V_3 is the '*modulator*' stage operating under Class A cathode bias conditions. The a.f. modulating signal input causes an amplified a.f. voltage to appear across the iron-cored choke L_4, and since this voltage is in series with the h.t. applied to the anode of the p.a. V_2, the potential at V_2 anode varies in accordance with the modulating signal. The r.f. pulses of anode current in the p.a. are thus caused to vary in amplitude according to the modulating signal and the resultant forced oscillations in the p.a. tank circuit provide an amplitude-modulated output to the aerial.   The bandwidth of the tank circuit must be sufficiently large to pass all the side frequencies produced as a result of modulation and to provide this the circuit is designed to have a relatively low Q.

Modulation in Radar Transmitters

26.   In radar, the e.m. energy radiated from the transmitter aerial is not intended to convey messages.   However, in order to obtain information about the target's height, bearing, distance and so on, the r.f. carrier must be modulated in some way.   The system generally used is *pulse modulation* as described in Para. 21 and illustrated in Fig. 15.   When the transmitter 'fires', a very short duration pulse of r.f. energy is radiated from the aerial.   The aerial is highly direc-

tional so that information about the bearing of the target may be obtained, and provided that the aerial is beamed towards the target, part of the radiated energy is reflected from the target and received back at the radar aerial during the interval between pulses. In this way information about the distance of the target may be obtained.   In radar, the duration of the pulses (pulse length) and the number of pulses occurring each second (pulse recurrence frequency) are determined by many factors and these are considered in detail in Part 3 (Radar).

Summary

27.   This Chapter has considered briefly how a r.f. carrier is produced at a power level and at a frequency suitable for radiation from an aerial.   The basic systems by which the carrier is caused to convey intelligence have also been introduced.   These ideas will be expanded and the various systems will be examined in detail in the appropriate Sections of Part 2 and Part 3 of these Notes.   However, having discussed basically how a transmitter operates and how messages are sent by wireless it is appropriate to consider next how these messages are received and translated into some form acceptable to the aural or visual senses.   This aspect will be considered in the next Section.

SECTION 14

RECEIVER PRINCIPLES

SECTION 14

RECEIVER PRINCIPLES

SECTION 14

CHAPTER 1

TUNED RADIO FREQUENCY (T.R.F.) RECEIVER

TUNED RADIO FREQUENCY (T.R.F.) RECEIVER

Wireless Communication System

1. A wireless communication system consists of a radio transmitter (or 'sender') and a radio receiver linked by electromagnetic energy radiated from the transmitter aerial. A radio transmitter is an apparatus for the production and modulation of radio frequency energy for the purpose of radio communication, and the necessary stages in such a transmitter are dealt with in Sect. 13. A radio receiver is a device for accepting radio signals and reproducing in suitable form, any modulation present. In this Section it is proposed to consider the basic stages in a radio receiver so that its function and its place in the radio link may be appreciated. Fig. 1 illustrates in block form the fundamental requirements of a wireless communication system.

2. (*a*) **Transmitter.** The basic requirements are:—

(i) A *master oscillator* to generate the r.f. signal at the required frequency.

(ii) A *power amplifier* to raise the power of the r.f. carrier to a sufficiently high level adequately to feed the aerial.

(iii) A *modulator* to impress the required information upon the r.f. carrier.

(iv) An *aerial system* for radiation of the electromagnetic energy.

(*b*) **Receiver.** The following factors have to be considered:—

(i) It is necessary to erect an *aerial system* in which the passing electromagnetic energy radiated from the transmitter induces a signal voltage.

(ii) Induced in any receiver aerial at any
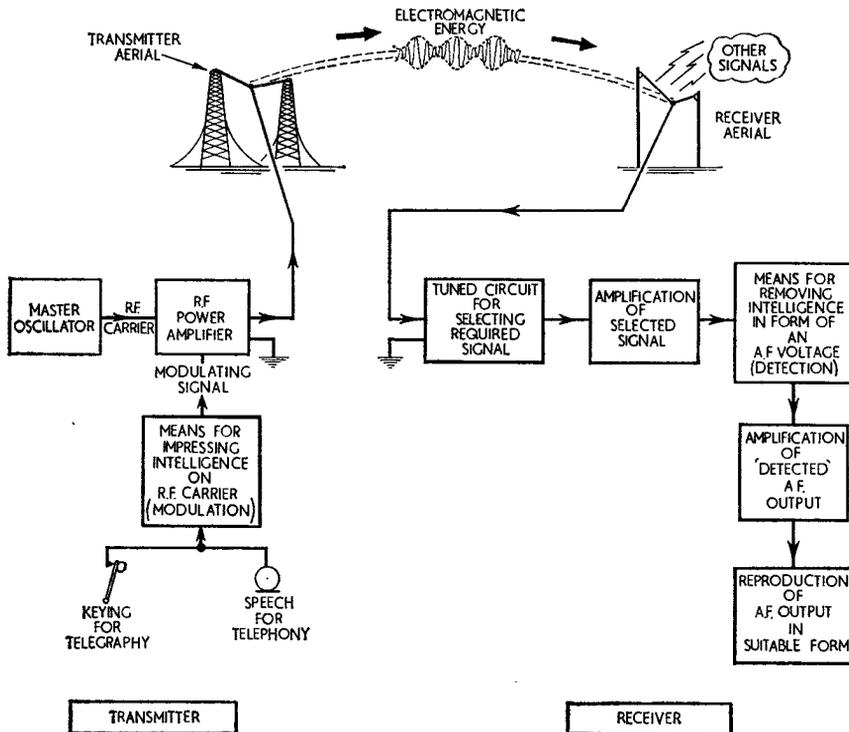
Fig. I. ELEMENTARY WIRELESS COMMUNICATION SYSTEM.

given instant are signal voltages at different frequencies from many sources in addition to the one required. It is therefore necessary to introduce a means of *selecting* the required r.f. signal; this involves the use of a tuned circuit.

(iii) Because of attenuation of the e.m. energy between the transmitter and the receiver, the voltage induced in the receiver aerial by the signal is generally very small (of the order of millivolts or even microvolts). Thus, the next step after selecting the wanted signal is to *amplify* the latter to a level sufficient to operate the next stage.

(iv) Just as the information has to be superimposed on the r.f. carrier at the transmitter, so it must be extracted from the carrier at the receiver. This process is termed '*detection*' (or '*demodulation*'), and the output from the detector should be an a.f. voltage having the same characteristics as the original modulating voltage at the transmitter.

(v) Having selected and amplified the required signal and extracted the intelligence in the form of an a.f. voltage, it is now necessary to *amplify* the a.f. component to a power level sufficient to operate the next stage.

(vi) Finally, it is necessary to include a *reproducer* whose function is to translate the audio output from the receiver into a form acceptable by the aural or visual senses. In most instances, a telephone receiver or a loudspeaker is required.

3. The ability of a communication receiver to perform its job properly is expressed in such terms as:—

(*a*) **Selectivity.** This is the ability of the receiver to discriminate between the desired signal and signals at other frequencies. A receiver with good selectivity will provide an output at the desired frequency only, whereas a receiver with poor selectivity will provide outputs from signals on adjacent frequencies in addition to the one required (see Fig. 2). Selectivity is a property of tuned circuits and it is dependent on the bandwidth of the circuits and on the number of tuned circuits used (see Book 1, Sect. 5, Chap. 3, Para. 18).

(*b*) **Sensitivity.** This is the ability of the receiver to receive weak signals (see Fig. 3). Sensitivity is expressed as the smallest input which will produce a certain output from the receiver at a given *noise* level. It is determined by the amplification of the valves and the magnification of the tuned circuits; that is, it is determined by the *gain* of the stages prior to detection.
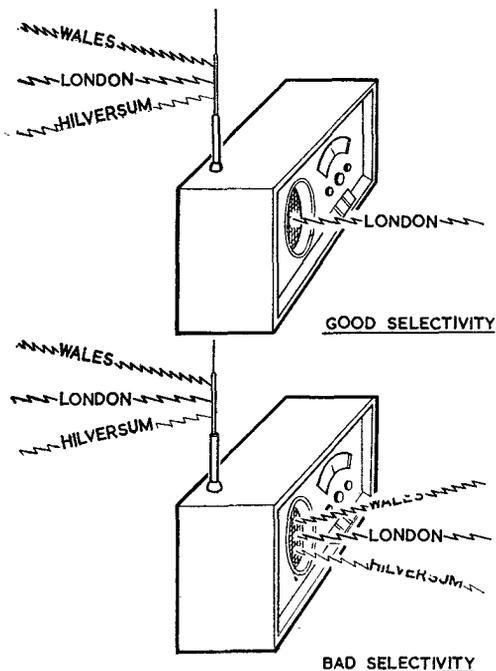
Fig. 2. SELECTIVITY.
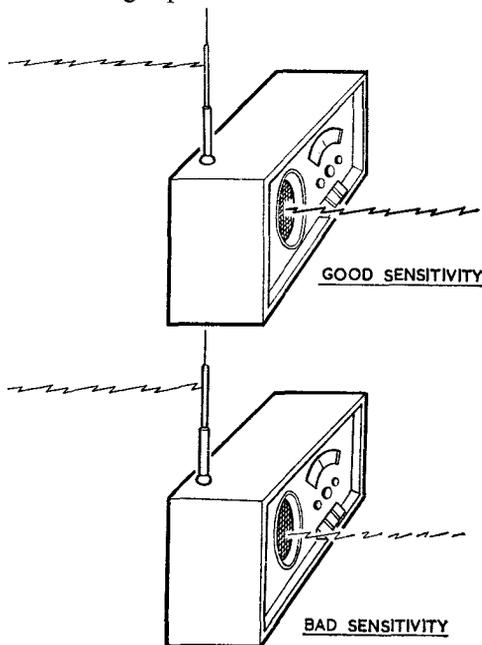
Fig. 3. SENSITIVITY.

(c) **Signal-to-noise ratio.** This is the ratio of the effective output voltage of the desired signal to that of noise, the latter being derived from a large number of sources, some external to the receiver and some internal (see Chap. 2, Para. 26). A
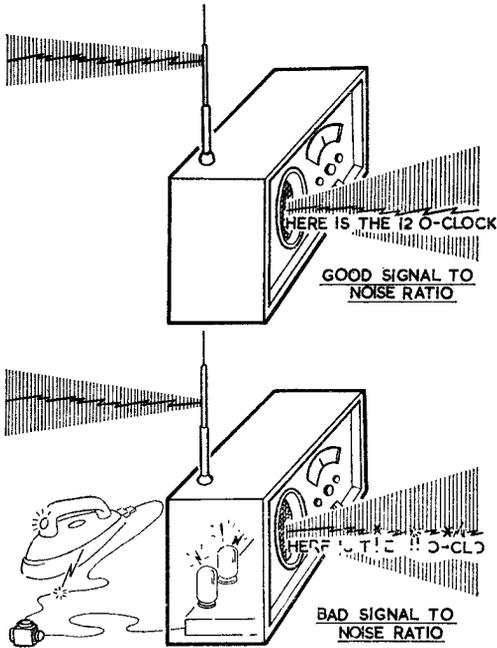
Fig. 4. SIGNAL-TO-NOISE RATIO.

receiver with a good signal-to-noise ratio gives an output relatively free from noise, whereas a receiver with a poor signal-to-noise ratio gives an output that may be completely lost in the noise (see Fig. 4). It is the signal-to-noise ratio that determines the maximum useful sensitivity of a receiver.

(d) **Fidelity.** This is the degree to which the receiver accurately reproduces at its output the essential characteristics of the signal which is impressed upon its input;
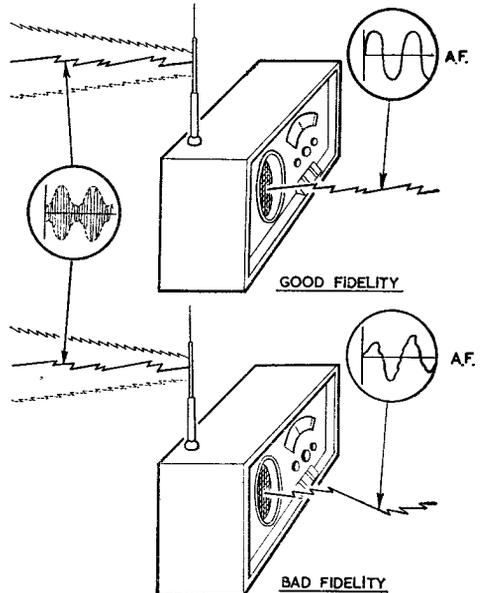
Fig. 5. FIDELITY.

that is, it is a measure of the freedom of the receiver from distortion (see Fig. 5). This depends primarily on the bandwidth and linearity of the stages.

Basic Receiver

4. Detailed examination of Fig. 1 will show that a basic receiver can be broken down into three separate sections;

(a) *One or more stages of r.f. voltage amplification.* This section provides the means for selecting and amplifying the
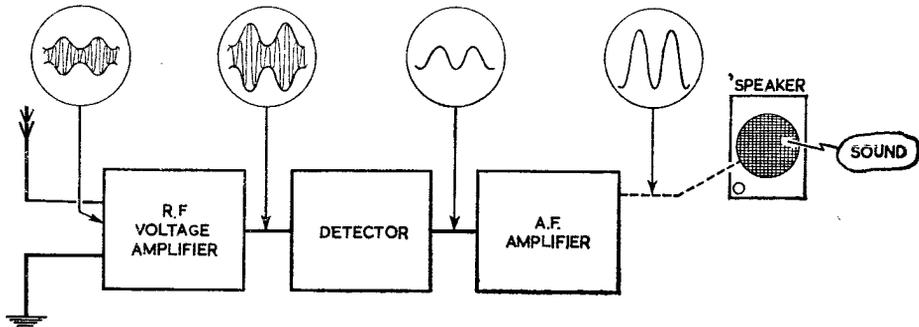
Fig. 6. BASIC TUNED RADIO FREQUENCY (T.R.F.) RECEIVER.

required signal, and determines the selectivity and sensitivity of the basic receiver.

(b) *One stage of detection* which provides the means for extracting the information from the amplified r.f. signal. The signal, after detection, may be a voice or code signal.

(c) *One or more stages of audio frequency amplification.* In communication receivers, the a.f. signal which comes from the detector is amplified by a.f. voltage and power amplifiers to a sufficient strength to operate telephones or a loudspeaker, so that the signal may be heard.
Fig. 6 shows the basic receiver in block form.

Such a receiver is termed a '*tuned radio frequency* (t.r.f.) or '*straight*' receiver and the separate stages are considered in detail in succeeding paragraphs.

R.F. VOLTAGE AMPLIFIER

Basic Circuit

5. In a receiver the two steps of selecting and then amplifying the required r.f. signal are combined in a single stage (the r.f. voltage amplifier) or several such stages connected in cascade. The theory of r.f. voltage amplifiers is given in Book 2, Sect. 11, Chap. 1, but the main points to be recalled are summarized below:—

(a) Because of Miller effect, triode valves have a tendency to instability and may produce undesirable oscillations when employed in r.f. amplifier stages. This effect can be reduced by neutralization or by using a grounded-grid triode, and both methods are combined in the *cascode* amplifier to provide a good signal-to-noise ratio at v.h.f. However, at the lower frequencies it is more usual to use tetrodes or pentodes in r.f. voltage amplifiers in order to take advantage of the much lower values for *Cga* in such valves and the higher gain. The valve generally preferred is a variable-mu pentode, in which the gain can be varied by altering the grid bias (see Book 2, Sect. 8, Chap. 3, Para. 24).

(b) The aerial is coupled to the input of the r.f. amplifier by a r.f. transformer (aerial coil), the secondary of which is tuned by a variable capacitor to provide a means for selecting the required signal.

(c) The anode load consists of a r.f. transformer (anode coil), the secondary of which is tuned by a variable capacitor to provide maximum impedance, and hence maximum output voltage at the signal frequency.

The basic circuit of a r.f. voltage amplifier incorporating these features is illustrated in Fig. 7(a). In this circuit, R_1 (de-coupled by C_1) provides Class A cathode bias, and R_2 (de-coupled by C_2) ensures that the screen

(a)

(b)
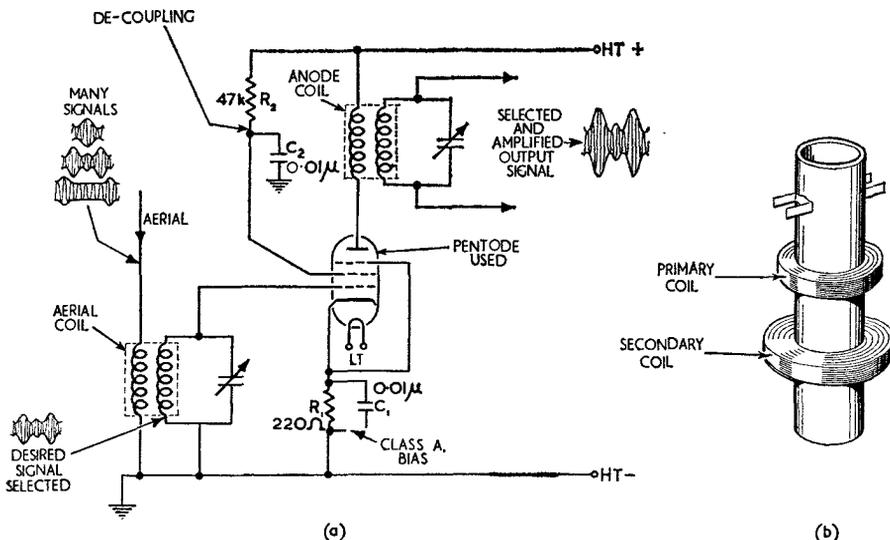
Fig. 7. TYPICAL R.F. VOLTAGE AMPLIFIER.

grid is at the correct d.c. potential for satisfactory operation of the stage. Fig. 7(b) shows the construction of the r.f. transformers; they are generally of the air-core type, although variable powdered cores may also be used.

Multi-stage R.F. Amplifiers

6. The *selectivity* of a receiver depends on the number of tuned circuits used and on the bandwidth of each individual tuned circuit. The *sensitivity* of a receiver depends on the number of tuned voltage amplifiers used and on the gain of each individual stage. Thus, it would appear that the more r.f. amplifiers there are in a receiver, the greater will be its selectivity and sensitivity. This is true only up to a point because, unfortunately, even when pentodes are used in r.f. amplifiers

there is a tendency to instability caused by interaction between the stages, and undesirable oscillations can result. Because of this, and because of the difficulties involved in tuning multi-stage r.f. amplifiers over a range of frequencies, receivers with more than two stages of r.f. amplification are seldom used.

Tuning

7. Fig. 8(a) shows in block form, two stages of r.f. voltage amplifiers connected in cascade. Three tuned circuits are included; the aerial coupling to V_1, the inter-stage coupling between V_1 and V_2, and the anode load of V_2. For maximum selectivity and sensitivity, each tuned circuit is to be resonant at the frequency of the desired signal. In the early days of radio, this was done by adjusting each variable capacitor separately, a considerable
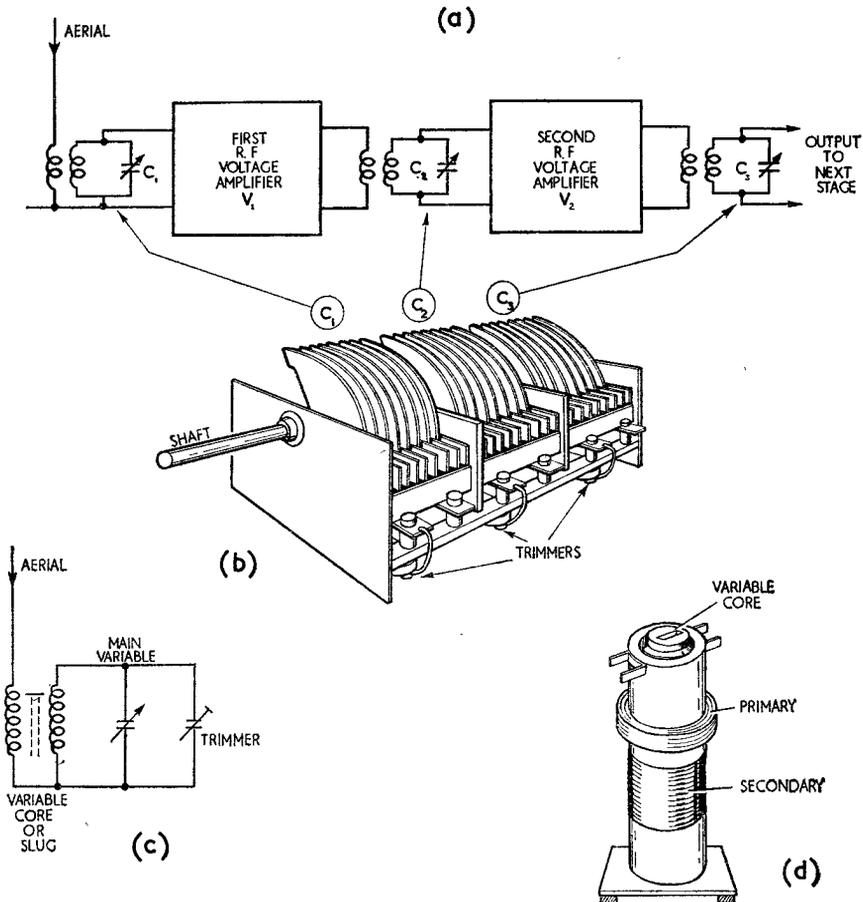
Fig. 8. TUNING R.F. VOLTAGE AMPLIFIER STAGES.

degree of skill on the part of the operator being required to obtain satisfactory tuning. Modern receivers eliminate the need for individual tuning controls by having the variable capacitors of all the tuned circuits mounted on a single shaft, so that a single tuning control alters all the tuned circuits simultaneously. In the block diagram of Fig. 8(a), a 'three-gang' capacitor is required. This is illustrated in Fig. 8(b).

Alignment

8. Because of difficulties associated with manufacture and assembly of components, the individual sections in a ganged capacitor have slightly different capacitances at every setting of the tuning control, and there will therefore be slight differences in the frequencies to which the resonant circuits are tuned. However, for good selectivity and sensitivity, the frequencies to which the three circuits are tuned should coincide exactly at every point within the tuning range. To assist in this:—

(a) *Trimmer capacitors* (Fig. 8(c)) may be inserted in parallel with each section of the three-gang capacitor. Since the total capacitance of two capacitors in parallel is the *sum* of the two capacitances, the trimmer capacitor has the greatest effect when the main variable capacitor is at *minimum* capacitance setting; that is, at the high frequency end of the tuning range (since $f \propto \dfrac{1}{\sqrt{C}}$). If the trimmer capacitor has a maximum capacitance of 20 pF and the minimum capacitance of the main variable capacitor is 50 pF, the total capacitance variation at this setting is 50 pF to 70 pF or $\dfrac{20}{50} \times 100 = 40\%$. If the maximum capacitance of the main variable capacitor is 500 pF, the total capacitance with the trimmer at its maximum value is 520 pF, so that the trimmer is only capable of providing a capacitance variation of $\dfrac{20}{500} \times 100 = 4\%$. The trimmers are, therefore, adjusted at the minimum capacitance (or maximum frequency) setting of the main tuning control, where they have most effect, until the frequencies of all the tuned circuits coincide at this setting.

(b) The r.f. transformers may employ variable-core (or 'slug') tuning, where the slug is moved in and out of the coil to vary

the inductance and hence the frequency of the tuned circuits (see Fig. 8(d)). The slugs may be adjusted at any point in the tuning range, but since the top and bottom of the range can be catered for by other means, it is normal to adjust the slugs at the middle of the tuning range until the frequencies of all the tuned circuits coincide at this setting.

(c) The main variable capacitor vanes are sometimes '*slotted*' (see Book 1, Sect. 4, Chap. 2, Para. 18) so that adjustment of the tuned circuit capacitances can be made at the *maximum* capacitance (or minimum frequency) setting of the main tuning control. This is done by bending the slotted vanes until the frequencies of all the tuned circuits coincide at this setting.
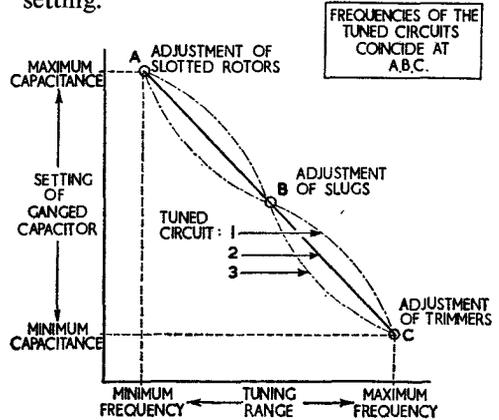
Fig. 9. ALIGNMENT OF R.F. TUNED CIRCUITS.

9. By making appropriate adjustments as given in Para. 8, the three tuned circuits are able to 'track' correctly over the tuning range; that is, the frequency to which any one circuit is tuned will correspond almost exactly to that of the other two tuned circuits over the whole tuning range (see Fig. 9). A receiver in which these adjustments have been made is said to be '*in alignment*', the result being high receiver selectivity and sensitivity.

Band Switching

10. In a receiver employing a single-stage r.f. amplifier, two tuned circuits are used, one in the input to the stage and the other in the output. These circuits will be tuned by a *two-gang* variable capacitor. However, a tuned circuit incorporating a single coil and variable capacitor has a limited frequency range. With the normal type of variable

capacitor, which has a ratio of maximum to minimum capacitance of about 10, the ratio of maximum to minimum frequency is about 3 ($f \propto \dfrac{1}{\sqrt{C}}$); for example, 500 kc/s to 1500 kc/s. If the receiver is to cover a frequency range greater than one coil and capacitor in either of the tuned circuits will allow, it will be necessary to change the tuned circuits. This
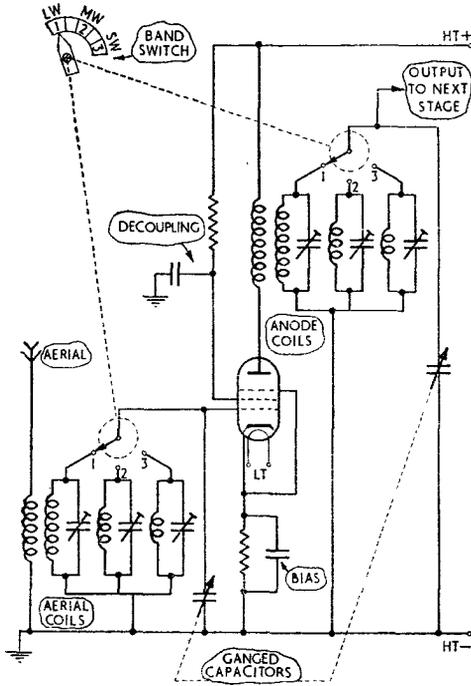
Fig. 10. BAND SWITCHING.

is usually done by switching in different values of inductance. Fig. 10 shows a system which uses several mounted coils, any one of which may be connected by means of a 'band switch' to the appropriate tuning capacitor. In this way, tuning over a very wide range of frequencies is possible. Note that in receivers using band switching, the trimmers for each range are usually mounted on, and parallel with the individual range coils.

Gain Control

11. It is convenient to be able to vary the gain of the r.f. amplifier so that the amplitude of the output voltage can be adjusted according to the strength of the received signal. Thus, for a strong signal it may be necessary to reduce the gain for comfort, and *vice versa* for a weak signal. One of the most common methods of controlling the gain of a r.f.

(a)

(b)
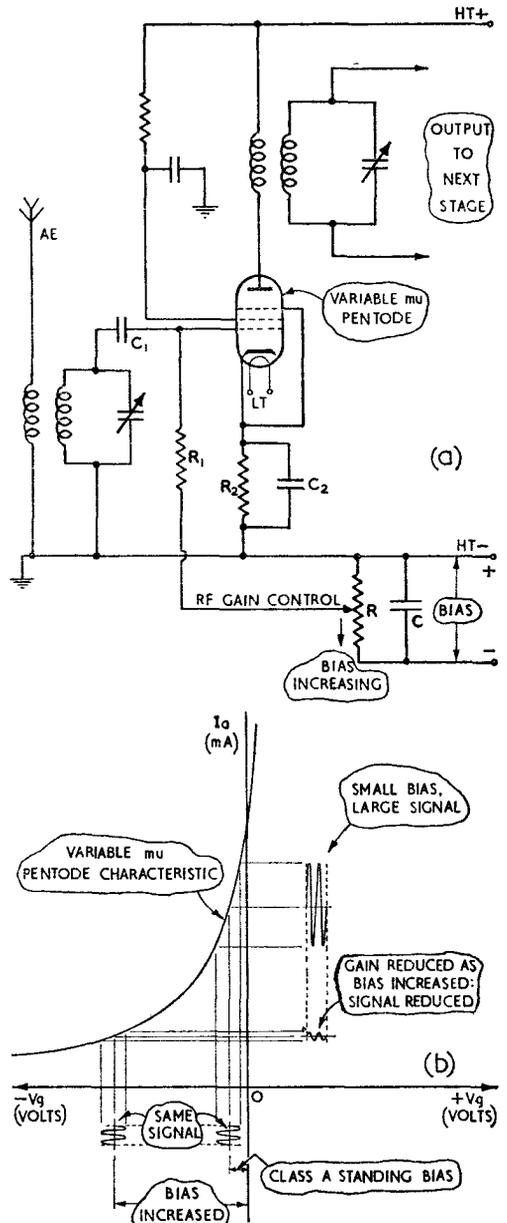
Fig. 11. R.F. GAIN (VOLUME) CONTROL.

voltage amplifier is to use a variable-mu pentode with an adjustable grid bias voltage. One arrangement is illustrated in Fig. 11(a). By increasing the negative bias applied to the grid of the valve, the amplitude of the anode current variations is reduced as shown in Fig. 11(b), and so also is the gain of the stage. $R_2 C_2$ provide the 'standing' or minimum Class A cathode bias conditions; the

negative bias cannot be reduced beyond this value and distortion due to grid current is prevented.

Bandwidth

12. The bandwidth required in the r.f. amplifier stage depends on the '*type*' of signal for which the receiver is designed. For telegraphy signals in morse code, only a very narrow bandwidth of the order of a few hundred cycles per second is required. For reception of amplitude-modulated speech signals, a larger bandwidth of the order of 10 kc/s is needed. For frequency-modulated
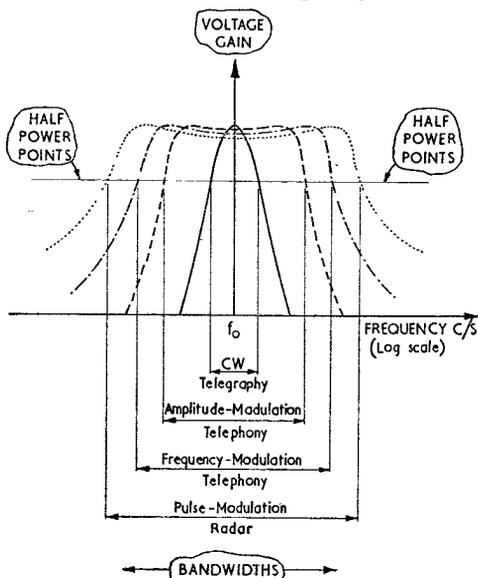
Fig. 12. BANDWITH REQUIREMENTS.

telephony signals, a still larger bandwidth of the order of 200 kc/s is required. For pulse reception in radar, a bandwidth of up to several megacycles per second is called for. These requirements are illustrated in Fig. 12. The bandwidth needs are met in design by adjusting the Q factor of the tuned circuits and by damping (see Book 2, Sect. 11, Chap. 1, Para. 18).

DETECTOR STAGE

Need for Detector

13. The next step in a receiver, after selecting and amplifying the required signal in the r.f. amplifier stage, is to extract from the signal the information that it is carrying. This process is called '*detection*' (or demodulation) and the stage which does this is the detector (or demodulator). The detector is

really the 'heart' of a receiver because without it, no output can be obtained. This may be seen by considering what would happen if the selected and amplified signal from the r.f. amplifier were applied directly to the telephones or the loudspeaker in a communication system. For an amplitude-modulated radio telephony signal, the amplitude of the r.f. carrier is varying in accordance with the information being conveyed (see Sect. 13, Chap. 1, Para. 22); this is shown in Fig. 13(a). If this signal is applied directly to the telephones (Fig. 13(b)) no sound is heard because:—

(*a*) due to its mechanical inertia, the diaphragm of a telephone receiver cannot respond to the rapid changes of the radio frequency signal;

(a)

(b)
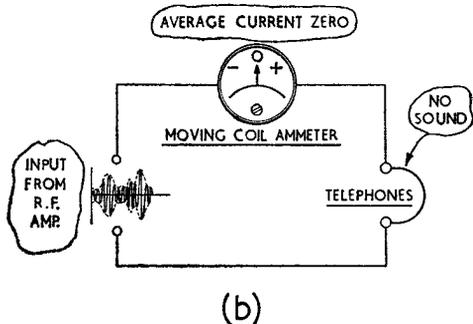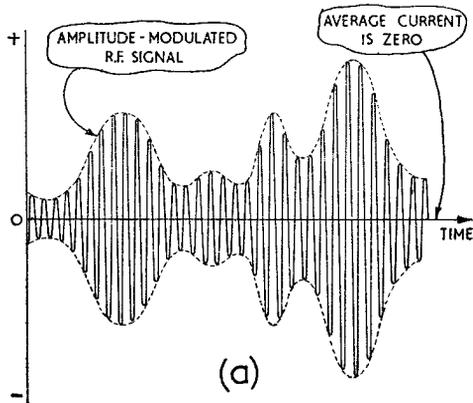
Fig. 13. NEED FOR DETECTION.

(*b*) the *average* current through the telephones is zero since the positive and negative half cycles in each cycle of the r.f. signal are equal in amplitude; hence, since the telephone diaphragm cannot respond to the r.f. variations it will tend to follow the average variation in current and as the latter is zero, the telephone diaphragm does not move.

What is required is some device that will extract from the r.f. signal that component to which the telephones or loudspeaker and the human ear, will respond. This is the work of the detector.

Function of Detector

14. In the ideal detector, the application of an amplitude-modulated radio telephony signal produces an audio output voltage whose frequency and amplitude corresponds to the variations in amplitude of the modulated input signal (Fig. 14). Two processes are involved:—

(*a*) the detector acts as a *rectifier* to remove one half of the input voltage; the *average* current through the detector is then no longer zero but varies in accordance with the variations in amplitude of the applied signal;
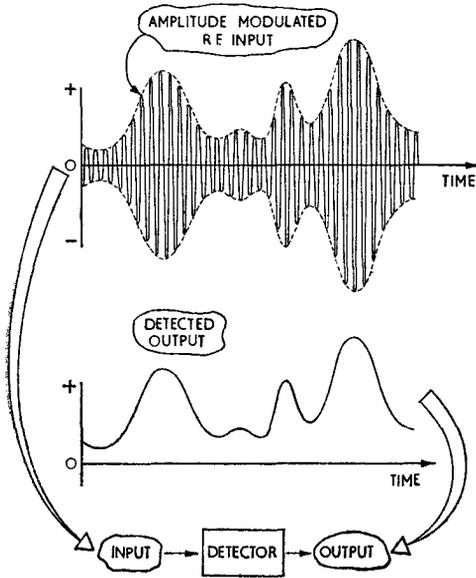
Fig. 14. FUNCTION OF DETECTOR.

(*b*) the unwanted r.f. and d.c. components contained in the rectified output must be eliminated so that the detector output consists only of the wanted audio frequency component; that is, the detector includes *filter circuits*.

Many different types of detector are used in receivers, but all of them employ the two basic processes of rectification and filtering. The types discussed in the following paragraphs are the crystal, the diode, the triode

leaky grid and the triode anode bend detectors.

Crystal Detectors

15. **Introduction.** One of the earliest forms of detection used a crystal diode (not to be
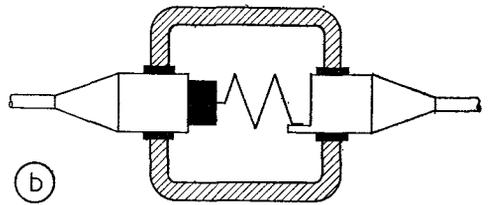
Fig. 15. CRYSTAL DIODE.
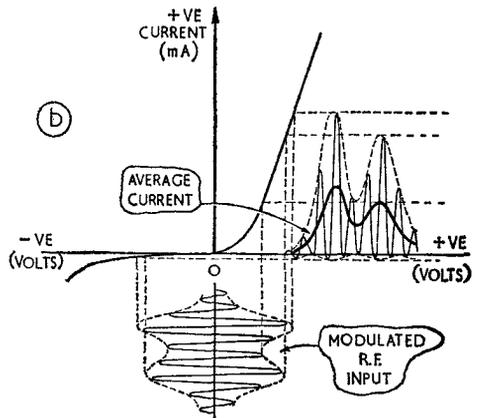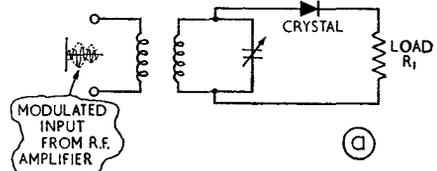
Fig. 16. ACTION OF CRYSTAL DETECTOR.

confused with quartz crystals used for control of oscillator frequency), but because of unreliability of performance crystals were gradually replaced by valve detectors as the latter became available. However, with the advance made in modern semi-conductor research, crystal diodes of the germanium and silicon type are coming back into their own and being increasingly used as reliable detectors in receivers for certain applications. A typical crystal detector is shown in Fig. 15. The crystal diode detector may be either the junction type or the point-contact type depending on the use to which it is to be put (see Book 2, Sect. 8, Chap. 7, Para. 13). Crystal detectors have certain advantages in relation to detectors using valves:—

(a) no heater voltage is required;

(b) they are robust;

(c) they are small in size, and the very small self capacitance of the point-contact type means that it is preferable to valve detectors at the higher frequencies.

16. **Circuit and action.** A simplified circuit of a crystal detector is shown in Fig. 16(a), and Fig. 16(b) illustrates the current-voltage characteristic for a typical crystal. It is seen from Fig. 16(b), that the crystal passes current more easily when the voltage is applied in one direction than it does when the voltage is reversed. Thus if an amplitude-modulated signal is applied, rectification takes place and the average current through the crystal and the load R_1 varies in accordance with the modulation envelope (Fig. 16(b)). This provides the first essential process in detection, namely rectification.

17. **Filter circuit.** The rectified voltage developed across the load contains three components—r.f., d.c., and the required
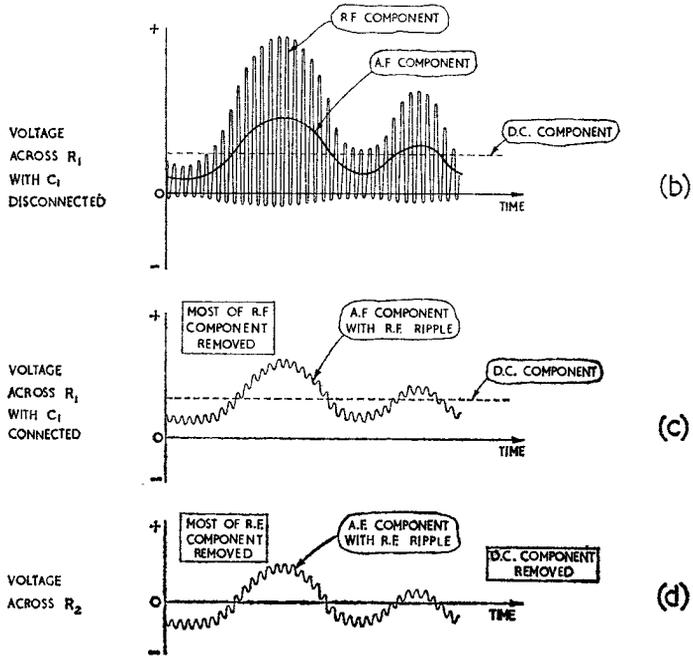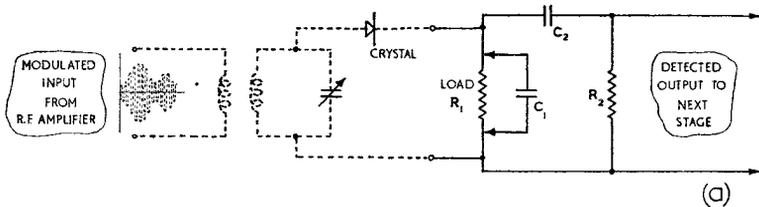
Fig. 17. ACTION OF FILTER CIRCUIT.

audio frequency component. The unwanted r.f. and d.c. components must now be removed so that the detector output consists only of the wanted component. The second process involved in detection (namely, filtering) may be obtained by inserting various types of filter network across the load R_1. One simple and effective type of filter is illustrated in the circuit of Fig. 17(a). Due to the rectifying action of the crystal, the voltage developed across the load R_1, with C_1 disconnected has the waveform shown in Fig. 17(b). The capacitor C_1 is of such a value that it has a low reactance to the r.f. component and a high reactance to the a.f. component. Thus, on inserting C_1 it effectively by-passes most of the r.f. component and the voltage appearing across $R_1 C_1$ then has the waveform shown in Fig. 17(c). This contains a small r.f. ripple and a d.c. component in addition to the wanted component. To remove the d.c. component a blocking capacitor C_2 is inserted. This has a low reactance to the wanted a.f. component and the resultant voltage developed across R_2 is the wanted component with a small r.f. ripple superimposed (Fig. 17(d)). In communication systems, the a.f. output from the detector may be applied directly to the telephones or loudspeaker. However, it is usually necessary to *amplify* the output from the detector to provide adequate power to operate the reproducer.

Diode Detector

18. **Circuit.** This is the most commonly used form of detector for communication receivers. The basic circuit of a diode detector is shown in Fig. 18 and if this is
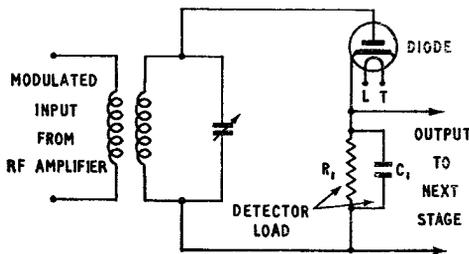
Fig. 18. SIMPLE DIODE DETECTOR CIRCUIT.

compared with the circuit of Fig. 16(a) it is seen that the only fundamental difference between the diode and crystal detectors is

the replacement of the crystal by a diode valve. The operating principles and characteristics of these two detectors are also very similar. The circuit shown in Fig. 18 is a series-fed circuit, where the diode and the load ($R_1 C_1$) are connected *in series* across the input tuned circuit.

19. **Action.** The function of C_1 is explained in simple terms in Para. 17. There is however another, perhaps more satisfactory way of looking at the function of C_1. Fig. 19(a) shows the r.f. modulated input applied to the diode detector from the r.f. voltage amplifier stage. If the detector load ($R_1 C_1$) were short-circuited the current established in the diode would be as shown in Fig. 19(b); that is, rectification takes place. If now, the capacitor C_1 is inserted with R_1 disconnected, the capacitor quickly charges almost to the peak value of the applied voltage as shown in Fig. 19(c), and the current falls rapidly to zero. Finally, if the resistor R_1 is connected across the capacitor C_1, the capacitor charges rapidly on the positive half-cycles of the input and discharges at a slower rate through R_1 on the negative half-cycles of the input when the valve is cut off. The charging time constant is given by $C_1 r_a$ (where r_a is the resistance of the diode). The discharge time constant is given by $C_1 R_1$. The resultant load voltage when C_1 and R_1 are in circuit is shown in Fig. 19(d), and it is seen that this voltage tends to follow the waveform of the envelope.

20. **Filter circuit.** As stated in Para. 17, the rectified voltage developed across $R_1 C_1$ may be analysed into three separate components:—

(a) an a.f. component corresponding to the variations in amplitude of the modulated input;

(b) a r.f. ripple at the carrier frequency;

(c) a d.c. component.

This is shown in Fig. 20. The unwanted r.f. and d.c. components must be eliminated and to do this a suitable filter circuit is connected across $C_1 R_1$ as shown in Fig. 21. The r.f. choke L_1 and the capacitor C_2 form a potential divider across $C_1 R_2$. The values of L_1 and C_2 are such that for the r.f. component the reactance of L_1 is much greater than the reactance of C_2 and the ripple voltage is developed mainly across the choke;
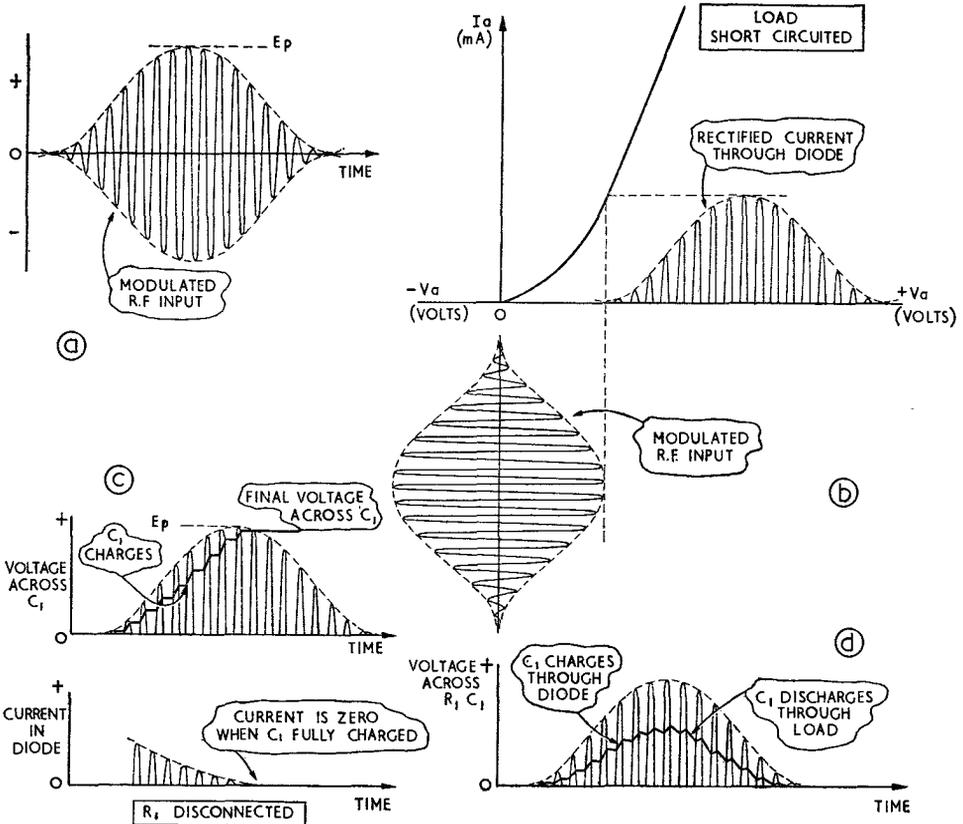
Fig. 19. ACTION OF DIODE DETECTOR.
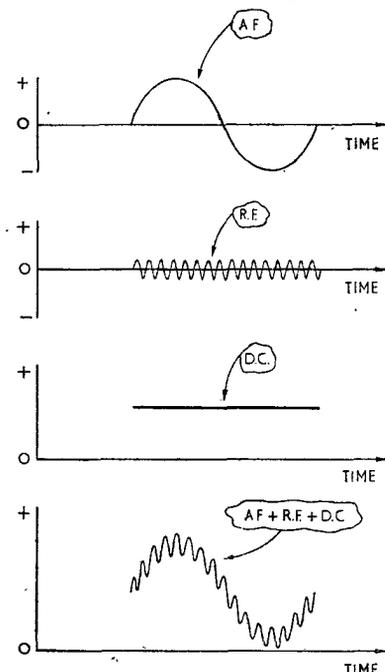
Fig. 20. COMPOSITION OF DETECTED OUTPUT.

for the a.f. component the reactance of L_1 is much less than that of C_2 and the a.f. voltage is developed mainly across C_2. In some circuits, L_1 is replaced by a resistor whose value is much greater than the reactance of C_2 at r.f., and much lower than the reactance of C_2 at a.f. The result is that the voltage developed across C_2 is the wanted a.f. with a d.c. component superimposed, the r.f. ripple being 'dropped' across L_1 (or resistor). The blocking capacitor C_3 prevents the d.c. component being applied to the external circuit, and at the same time it presents a low reactance to the a.f. component which is developed mainly across R_2 and applied to the a.f. amplifier stage.

21. **Distortion.** The diode detector introduces a certain amount of distortion. This arises from two main causes:—

(a) the bottom bend in the diode I_a-V_a characteristic;

(b) an incorrect value for the time constant of the load $C_1 R_1$.
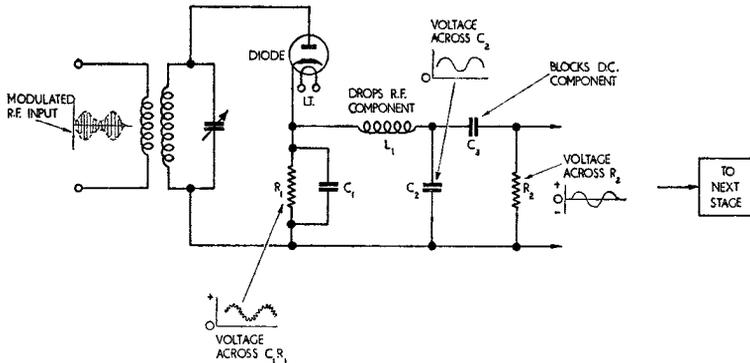
Fig. 21. ACTION OF FILTER CIRCUIT.

22. Effect of bottom bend. The dynamic characteristic of a diode detector is not straight throughout its entire length because the action of the space charge introduces an initial curvature or bottom bend. In Fig. 22(a), the characteristic of an 'ideal' diode is linear throughout and no distortion occurs. In Fig. 22(b), distortion occurs when the modulated input voltage falls below the straight part of the characteristic; the effect is to flatten the shape of the envelope during the troughs of the input signal. Distortion due to the bottom bend may be reduced by ensuring that the signal is sufficiently large and is less than 100% modulated so that only the straight part of the characteristic is in fact used.

23. Effect of load time constant. If the time constant of the load $C_1 R_1$ is such that the load voltage cannot follow the variations in amplitude of the input signal, distortion occurs:—

(a) **Time constant $C_1 R_1$ too long.** The capacitor C_1 charges quickly through the diode, but with a long discharge time

Fig. 22. DISTORTION DUE TO BOTTOM BEND.
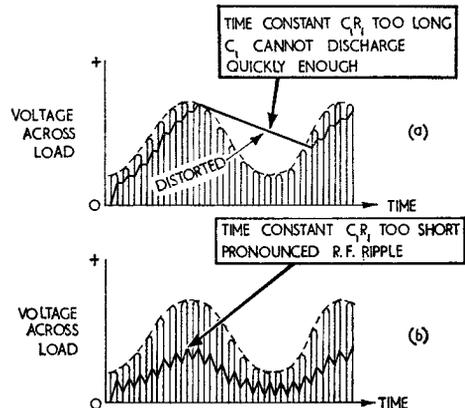
Fig. 23. EFFECT OF LOAD TIME CONSTANT.

constant $(C_1 R_1)$ the capacitor cannot discharge quickly enough in relation to the fall in amplitude of the input signal and the troughs of modulation are distorted (Fig. 23(a)).

(b) **Time constant $C_1 R_1$ too short.** The capacitor charges quickly through the diode, but with a short discharge time constant $(C_1 R_1)$ the capacitor also discharges quickly and a pronounced r.f. ripple results (Fig. 23(b)).

(c) **Time constant $C_1 R_1$ correct.** From (a) and (b) it is seen that the time constant must be a compromise. Thus, the product $C_1 R_1$ seconds should be *smaller* than the time taken for one cycle of the *highest audio modulation frequency* likely to be used, but it must be *greater* than the time taken for one cycle of the *radio frequency* carrier. In addition, R_1 must be large compared with the a.c. resistance r_a of the diode to prevent a large voltage drop across the valve and to reduce the damping effect on the input circuit; that is, for a high efficiency, $R_1 > r_a$. Also, C_1 should offer a low reactance to the r.f. component compared with the reactance offered by the inter-electrode capacitance of the valve. These considerations determine the values for R_1 and C_1.

Example. In a typical broadcast receiver, $C_1 = 100$ pF; $R_1 = 0 \cdot 5$ MΩ.

Then, $C_1 R_1 = 100 \times 10^{-12} \times 0 \cdot 5 \times 10^6$
$\therefore C_1 R_1 = 50 \mu$ **sec.**

The *highest* audio modulation frequency f_m likely to be encountered is 10 kc/s.

Modulation periodic time $t_m = \dfrac{1}{f_m}$

$$= \dfrac{1}{10^4}$$

$\therefore t_m = 100 \mu$ **sec.**

The *lowest* radio frequency carrier f_c likely to be used is about 200 kc/s.

Carrier periodic time $t_c = \dfrac{1}{f_c}$

$$= \dfrac{1}{200 \times 10^3}$$
$\therefore t_c = 5 \mu$ **sec.**

Thus the conditions laid down in Para. 23(c) for the load time constant (namely $t_m > C_1 R_1 > t_c$) have been satisfied.

24. **Diode efficiency.** The ratio of output voltage to input voltage is a measure of the efficiency of a detector. The efficiency will be high when the value of the load resistor R_1 is greater than the a.c. resistance r_a of the diode since then the p.d. across R_1 approaches the magnitude of the applied voltage, and the p.d. dropped across the valve is small. Efficiencies of over 80 per cent. are obtainable in practice.

25. **Input resistance.** A diode detector absorbs power from its input circuit and has the effect of damping the tuned circuit and reducing the selectivity and sensitivity of the preceding (r.f. amplifier) stage. The power absorbed, and hence the amount of damping introduced can be given in terms of the power dissipated in an equivalent resistance R_{eq} connected across the diode input circuit (see Fig. 24). The value of this equivalent resistance is:—

$$R_{eq} = \frac{R_1}{2 \times \text{Efficiency}}.$$

Thus, a high value for the load resistance R_1 reduces the power absorbed and reduces the damping introduced by the diode circuit since:—

$$\text{Power dissipated} = \frac{V^2}{R_{eq}}.$$
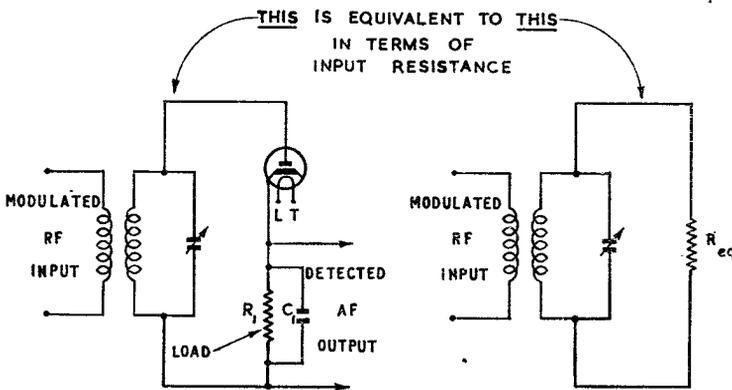
Fig. 24. INPUT RESISTANCE OF DIODE DETECTOR.

By making R_1 as high as possible consistent with other factors mentioned concerning the load time constant, excessive reduction in the selectivity and sensitivity of the preceding stage is prevented.

26. Shunt-fed diode detector. The series-fed diode detector circuit discussed in the preceding paragraphs is extensively used in modern receivers. However for certain applications, the shunt-fed circuit of Fig. 25 is preferred. In this circuit, the diode is *in parallel* with the load R_1 across the input tuned circuit. When the signal causes the
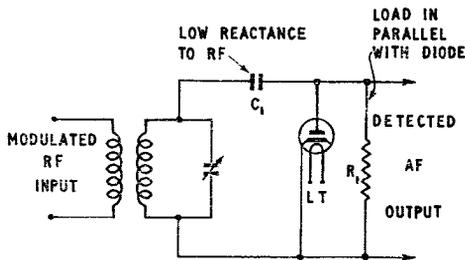

Fig. 25. SHUNT-FED DIODE DETECTOR CIRCUIT.

anode to become positive with respect to its cathode, the valve conducts and charges C_1. On the alternate half-cycle, when the valve is cut off, C_1 discharges through R_1. The action is thus similar to that of the series-fed detector, and for an amplitude-modulated input the voltage across the load follows the variations in amplitude of the applied signal (provided that the time constant $C_1 R_1$ is correct). This circuit has the advantage that there is no d.c. path through the detector from the input tuned circuit; but it has the disadvantage that damping of the preceding stage is greater because, with the valve in parallel with the load, the input resistance is low. This circuit is commonly used to provide the voltage for operating automatic gain control in a receiver.

27. Advantages and limitations of diode detector.

(a) **Advantages.**

(i) The diode detector can handle large input signals without the risk of overloading.

(ii) For large input signals, modulated less than 100%, it acts as a linear detector and introduces very little distortion.

(iii) The d.c. component of load voltage may be used for automatic gain control purposes.

(b) **Limitations.**

(i) The diode detector provides no amplification.

(ii) Power is absorbed from the input circuit, resulting in damping of the previous stage and a reduction in selectivity and sensitivity.

Leaky Grid Detector

28. Introduction. Diode and crystal detectors have the disadvantage that they provide no amplification, and if weak signals are to be received additional amplifier stages prior to the detector have then to be provided. This limitation can be overcome by using a valve which both detects and amplifies. Amplification implies the use of a valve with a control grid, and triode, tetrode or pentode valves may be used to provide the dual function of detection and amplification. The simplest circuit of this type is the triode leaky grid detector.

29. Circuit and action. The simplified circuit of a triode leaky grid detector is illustrated in Fig. 26(a), and the dynamic mutual characteristic for the valve is shown in Fig. 26(b). Fig. 26(a) shows that the grid-cathode circuit is equivalent to that of a diode detector, with the control grid acting as the 'anode' of the equivalent diode. Under no-signal conditions, the grid voltage is zero and the standing anode current is *high* (Fig. 26(b)). On receipt of a constant-amplitude c.w. signal, grid current is established on the positive half-cycles of the input and the capacitor C_1 charges. On the negative half-cycles of the input when grid current ceases, C_1 discharges through R_1 on a time constant of $C_1 R_1$. Under steady signal conditions the charge on C_1 and the voltage across R_1 increase so that the negative bias on the grid increases and the operating point moves further back until a state of equilibrium is reached where just sufficient grid current flows at the peak of the positive half-cycles to replenish the charge lost by C_1 through R_1 on the negative half-cycles. The bias voltage is then steady (Fig. 26(b)). If the amplitude of the signal increases, the bias increases; and if the amplitude of the signal decreases,

so does the bias. Thus, if an amplitude-modulated signal is considered, the charge on C_1 and the voltage across R_1 vary in accordance with the variations in amplitude of the applied signal; that is, the bias voltage varies at audio frequency, and detection has taken place at the grid.


(a)
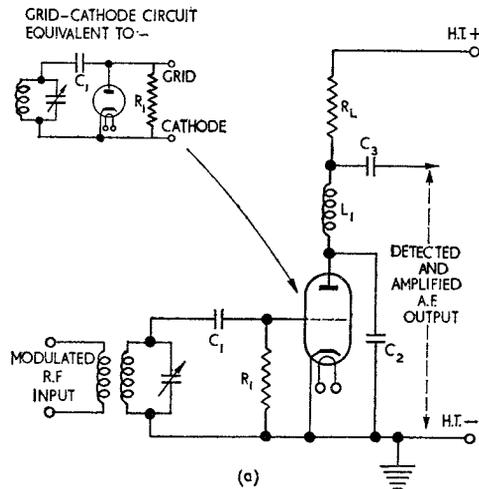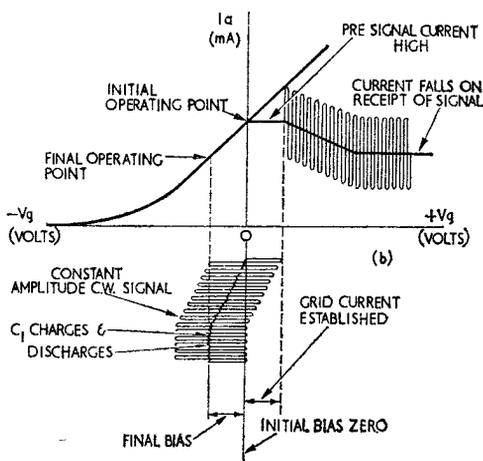
Fig. 26. LEAKY GRID DETECTION.

30. Amplification. Because of the detection which has taken place in the grid-cathode circuit, the voltage applied to the grid of the valve consists essentially of an a.f. voltage with r.f. and d.c. components superimposed and the anode current contains these three components of detection. The r.f. component is developed mainly across the r.f. choke L_1 in Fig. 26(a) and by-passed to earth through a small-value capacitor C_2. The a.f. component is developed across the anode load R_L and capacitively coupled through C_3 to the output. The capacitor C_3 also acts as a d.c. blocking capacitor to the d.c. component of load voltage. Normal triode amplification occurs and the a.f. output voltage is $\dfrac{\mu R_L}{r_a + R_L}$ times larger than the detected a.f. component applied to the grid of the valve.

31. Characteristics.

(a) Before the application of a signal, the bias is zero and the pre-signal value of anode current is high; the grid current is zero. On receipt of a signal the negative bias builds up and the anode current *falls;* the grid current rises from zero to a high value initially and then falls to a small value as the bias builds up.

(b) Leaky grid detection provides high sensitivity and is capable of detecting very weak signals. This is especially so with a large value of grid leak R_1 because the a.f. voltage developed at the grid is then large since $v_g = i_g R_1$. However, with a large value for R_1, distortion occurs for all signals of more than a few volts amplitude; this is shown in Fig. 27(b) where the signal is sufficiently large to cut off the valve for a portion of each cycle.

(c) Some damping of the previous stage occurs in leaky grid detection but it is less than that for diode detection since the grid current established is very small and a large grid leak of the order of 1 MΩ is used. Thus, the reduction in selectivity and sensitivity in the r.f. amplifier stage is small when this method of detection is used.

Anode Bend Detector

32. Introduction. The anode bend detector is another circuit that provides both detection and amplification. A triode or a pentode valve biased nearly to cut-off is used. The bias voltage may be obtained from cathode bias, battery bias or external bias. With the valve biased to the bottom bend, the anode current is nearly zero under no-signal conditions, and if cathode bias is used, a large cathode bias resistor (about 20 kΩ to 50 kΩ) is needed because of this small current.

33. Action. The dynamic mutual characteristic shown in Fig. 28 illustrates the process of anode bend detection. Under no-
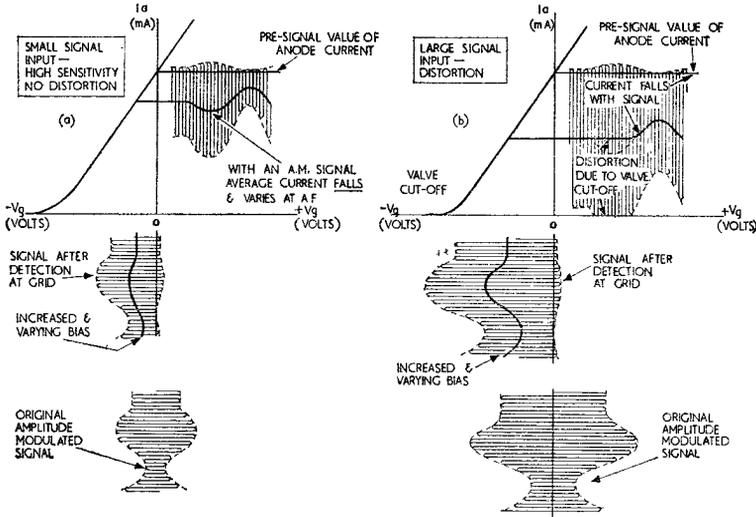
Fig. 27. LARGE SIGNAL DISTORTION WITH LEAKY GRID DETECTION.

signal conditions the anode current is *low* because of the bias. On receipt of an amplitude-modulated signal from the preceding (r.f. amplifier) stage, the detector valve conducts on the positive half-cycles of the input and cuts off on the negative half-
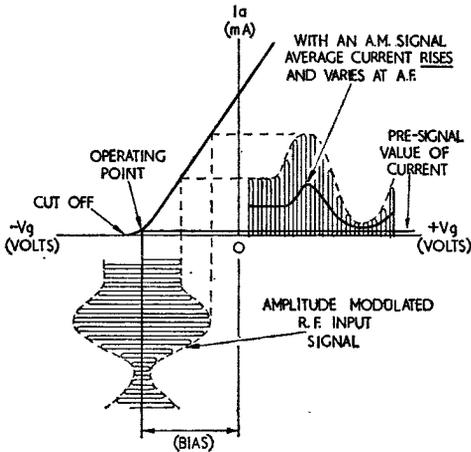
Fig. 28. ACTION OF ANODE BEND DETECTOR.

cycles. The anode current varies in the manner shown, its *average* value varying according to the audio frequency component of the modulated signal. The current also contains the unwanted results of detection, namely the r.f. and d.c. components which must be removed.

34. Circuit. Fig. 29 shows a circuit of a basic anode bend detector. The r.f. component of anode current develops a r.f. voltage mainly across the r.f. choke L_1, and the small value capacitor C_2 by-passes this component to earth. The a.f. component of anode current develops the required a.f. voltage across the anode load R_L, and C_3
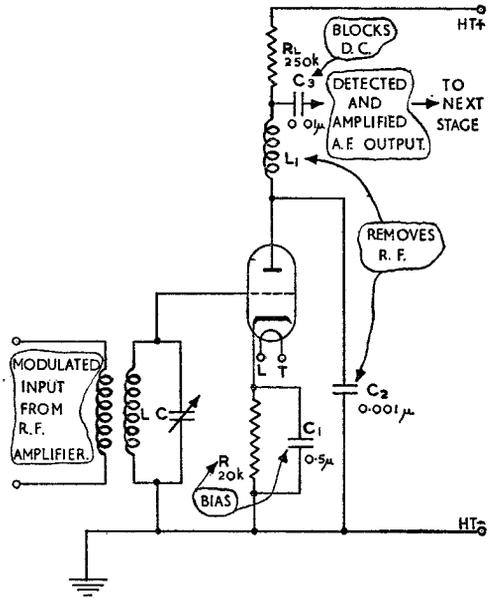
Fig. 29. ANODE BEND DETECTOR CIRCUIT.

acts as a coupling capacitor to the output. Capacitor C_3 is also a block to the d.c. component of load voltage. Normal triode amplification produces a large a.f. output voltage.

35. Characteristics.

(a) In contrast to leaky grid detection, the anode current with anode bend detection is *low* initially and *increases* with the signal.

(b) To prevent the flow of grid current, which results in increased damping of the tuned circuit and reduction in selectivity and sensitivity, the amplitude of the input signal is limited to less than the grid base of the valve.

(c) Since the operating point of the anode bend detector is on the bottom bend of the I_a-V_g curve distortion is appreciable for weak signal inputs. The anode bend detector works best with medium strength signals.

(d) The main use of the anode bend detector is in frequency changer circuits, in certain types of superheterodyne receivers and in valve voltmeters.

Comparison of Detectors

36. Table 1 gives a comparison of the main features of the four types of detector discussed in the previous paragraphs.

OUTPUT STAGE

Introduction

37. The next stage to be considered in the basic t.r.f. receiver is the audio output stage. The required signal has been selected and amplified in the r.f. amplifier stage; the output from the first stage has been detected in the detector stage to provide an audio output voltage; finally, it is necessary to increase the power level of the detected output sufficiently to operate the telephones or loud-speaker (see Fig. 30). The audio amplifier stage provides this function. It is assumed that the incoming signal is an amplitude-modulated telephony signal and the detected output is an a.f. voltage containing the characteristics of the original modulation.

A.F. Power Amplifier

38. Information on the theory of a.f. power amplifiers is given in Book 2, Sect. 10, Chap. 2, but the main points to be recalled are summarized below:—

(a) an a.f. amplifier is, in the ideal case, required to amplify signals equally well at any frequency in the audio band, i.e., from about 20 c/s to about 20 kc/s. However, the range required is usually less than this, 300 c/s to 3400 c/s being adequate for speech signals and 30 c/s to 8000 c/s being adequate for music;

(b) in a receiver, a.f. amplifiers follow the detector stage and are untuned so that they do not add to the selectivity of the receiver;

	Sensitivity	Selectivity	Fidelity	Other Remarks
Crystal	Low	Poor	Good	Capable of handling strong signals; no amplification.
Diode	Low	Poor	Good	Capable of handling strong signals; no amplification.
Leaky Grid	High	Poor	Poor	Easily overloaded; the anode current is high and decreases with the signal; amplifies as well as detects.
Anode Bend	Low on weak signals	Good	Moderate	The anode current is low and increases with the signal; amplifies as well as detects.

TABLE 1. COMPARISON OF FEATURES OF DETECTORS
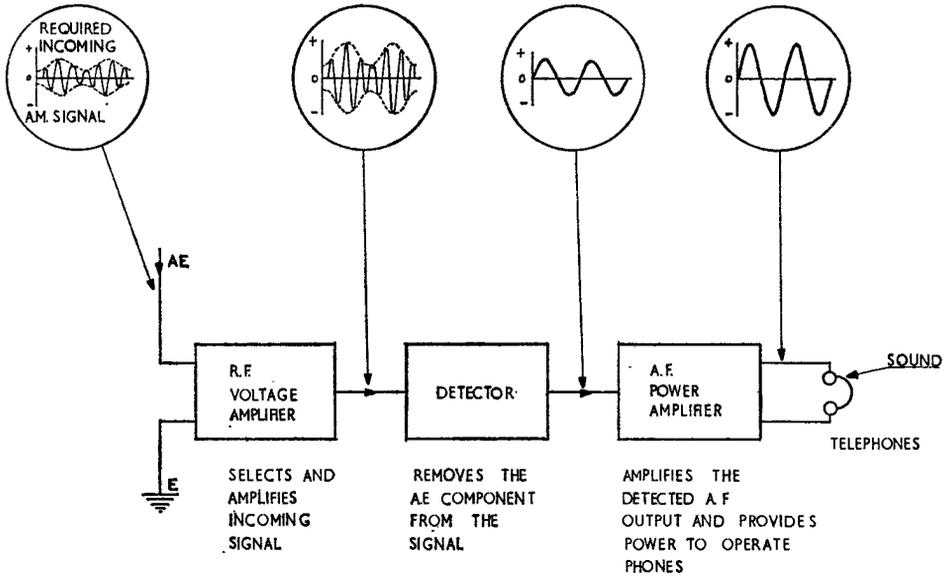
Fig. 30. STAGES IN A T.R.F. RECEIVER.

(c) the telephones are usually coupled into the anode circuit of the output valve by means of a transformer to ensure that no d.c. power is dissipated in the telephones and to give the necessary impedance matching between the telephones and the valve;

(d) a.f. amplifiers are stable and not likely to oscillate; triodes, beam-power tetrodes or power pentodes may be used;

(e) single valves operating under Class A conditions, or valves in Class A, Class AB or Class B push-pull may be used.

Circuit and Action

39. A simplified circuit of a typical a.f. power output stage is shown in Fig. 31. The a.f. output from the detector is coupled via C_1 and R_1 to the control grid of the a.f. amplifier valve V_1; C_1 blocks the d.c. component of the detector output and the a.f. component is developed across R_1. The grid leak R_1 also acts as an a.f. gain (volume) control, the moving arm tapping off the desired amount of signal voltage for application between the grid and cathode of V_1. The resistor R_2 (which must have a sufficiently high wattage rating to cope with the higher current in this stage), acts as a cathode bias

resistor to provide Class A bias conditions; the capacitor C_2 is a by-pass capacitor to ensure a steady d.c. bias voltage and to prevent negative feedback in the circuit. The
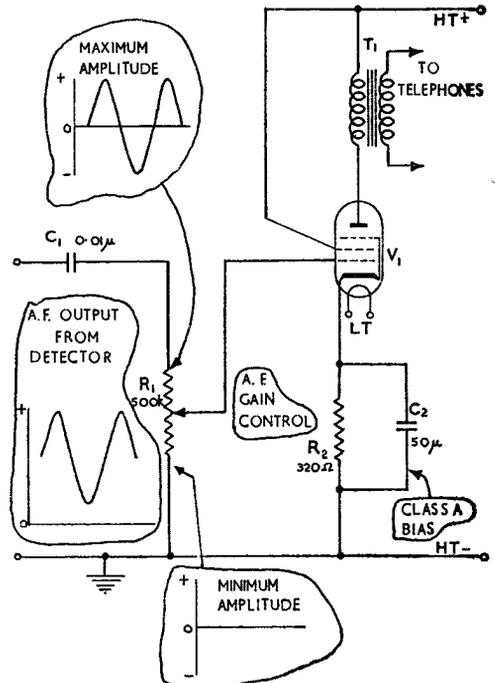
Fig. 31. SIMPLE A.F. POWER OUTPUT STAGE.

primary of the output transformer acts as the anode load and couples the amplifier to the telephones; the turns ratio is arranged to give the correct impedance matching. It is common practice to connect the screen grid of the output valve directly to the h.t. positive line to provide maximum current and a high power output.

COMPLETE T.R.F. RECEIVER

Practical Circuit

40. From the information given in the preceding paragraphs it is possible to construct a simple t.r.f. type receiver. This may be done by combining the circuits shown in Fig. 7 (r.f. amplifier), Fig. 26 (leaky grid detector) and Fig. 31 (a.f. power amplifier). A simple three-stage t.r.f. receiver results, as

between grid and cathode of the detector valve, V_2.

(b) **Leaky grid detector, V_2.** Detection of the amplitude-modulated signal occurs at the grid of V_2 due to the action of C_7 and R_4. The resultant a.f. voltage is developed mainly across R_6 and capacitively coupled through C_{10} to the a.f. amplifier stage; C_{10} also acts as a d.c. blocking capacitor. The r.f. component of voltage at the anode is developed mainly across L_5 and shunted to earth through the small value capacitor C_9 which has a low reactance to r.f. and a high reactance to the a.f. components.

(c) **A.F. amplifier, V_3.** The a.f. output from the detector is developed across the a.f. gain control R_7, and the desired proportion is tapped off for application
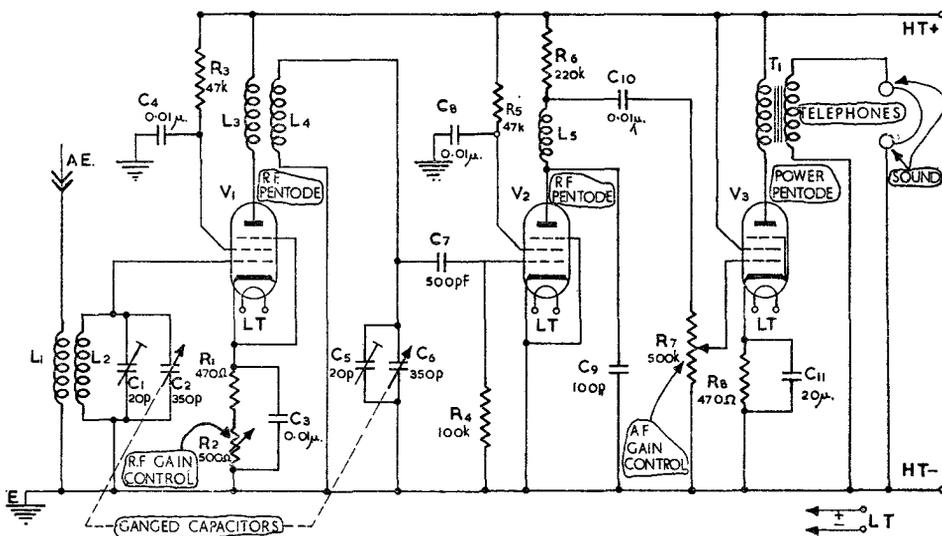
Fig. 32.   SIMPLE THREE-STAGE T.R.F. RECEIVER.

shown in Fig. 32. The main features of this circuit are summarized below:—

(a) **R.F. amplifier, V_1.** This is a conventional circuit as previously described. A variable-mu pentode is used and the cathode bias voltage (from R_1 R_2 C_3) may be varied by the manual r.f. gain control R_2 to give a variation in the volume of the output. The input amplitude-modulated signal is selected in the grid tuned circuit and is amplified in this stage to provide a large r.f. signal voltage at the required frequency across the secondary tuned circuit in the anode, for application

between grid and cathode of the output valve. V_3 is a pentode operating under Class A cathode bias conditions and the a.f. power developed in this circuit is transferred through the transformer for operation of the telephones.

(d) **Power supplies.** The h.t. and l.t. voltages required depend on the valves being used in the circuit. The supply voltages may be obtained from batteries (as in a portable receiver) or a half-wave or full-wave rectifier circuit may be included to convert the a.c. mains supply to the required h.t. and l.t. voltages.

RECEPTION OF C.W.

Introduction

41. The t.r.f. receiver considered so far assumes an amplitude-modulated input signal from which the detector can extract the required a.f. component for operation of the telephones. Information can of course, be conveyed over a wireless link in other ways (see Sect. 13, Chap. 1). For example, the r.f. output from a transmitter can be 'keyed' in accordance with the characters of the morse code to provide a keyed continuous wave or c.w. output.

42. An amplitude-modulated signal applied to a detector produces the required a.f. output (Fig. 33(a)). If however, a c.w. signal

Fig. 33. C.W. SIGNALS AND DETECTION.

is applied to a detector there is no a.f. component in the output and no sound is

heard in the phones. This is because there is no variation in the *amplitude* of the input signal so that the detector output contains r.f. and d.c. components only (Fig. 33(b)). There is in fact, no a.f. component present in the input signal and before the information can be conveyed and the familiar dots and dashes of the morse code heard, something additional must be done to the signal *at the receiver*. To hear a c.w. signal, the '*heterodyne principle*' is applied. This consists of mixing the input c.w. signal with an oscillation produced locally at the receiver. The result of mixing is an amplitude-modulated signal which can then be detected in the normal way so that the dots and dashes of the morse code are heard in the telephones.

Heterodyne Principle

43. If two tuning forks, of slightly different tones, are struck at the same time, a distinct throbbing effect or rhythmic beat can be heard. Similar results may be observed if two adjacent piano keys are struck at the same time, or if two organ pipes of nearly the same frequency are energised simultaneously. The resultant 'beat' in each case has a frequency equal to the *difference* between the frequencies of the two original tones. If the two original tones have frequencies of 264 and 297 cycles per second respectively, the beat frequency will be equal to the difference between them, namely 33 cycles per second.

44. Similarly, when two alternating voltages of slightly different frequencies are combined in a detector, the resultant voltages produced in the output will include a frequency which is equal to the difference between the frequencies of the two original voltages; this is the basis of the heterodyne principle.

45. In the reception of c.w. by the heterodyne principle, one of the original alternating voltages is provided by the incoming signal and the other is produced locally at the receiver by an oscillator; both signals are applied simultaneously to the detector. Unless these two voltages are of exactly the same frequency, they will not rise and fall in time with each other, but will get into step and out of step alternately. The action is
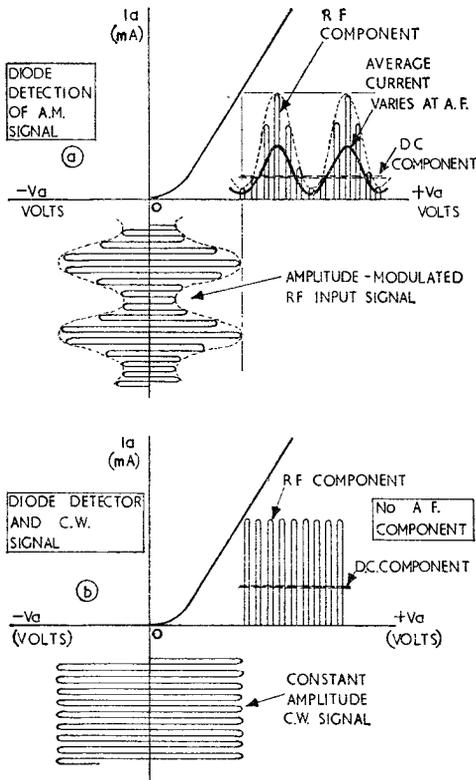
illustrated in Fig. 34. Curve A shows the voltage waveform due to the incoming signal at a frequency, in this case, of 5 kc/s. Curve B shows the voltage waveform produced by the heterodyne oscillator at a frequency of, say, 6 kc/s. Curve C shows the effect of combining these two waveforms at the input to the detector. At times 1, 3 and 5 the two original voltages are exactly out of phase and oppose each other; at times 2 and 4,

Beat Frequency Oscillator

46. The oscillator used at the receiver in the reception of c.w. signals is called a 'beat frequency oscillator' (b.f.o.); it may take any of the forms described in Book 2, Sect. 12, Chap. 1. A typical b.f.o. is illustrated in Fig. 35. This shows a conventional Hartley

Fig. 34. HETERODYNE PRINCIPLE.
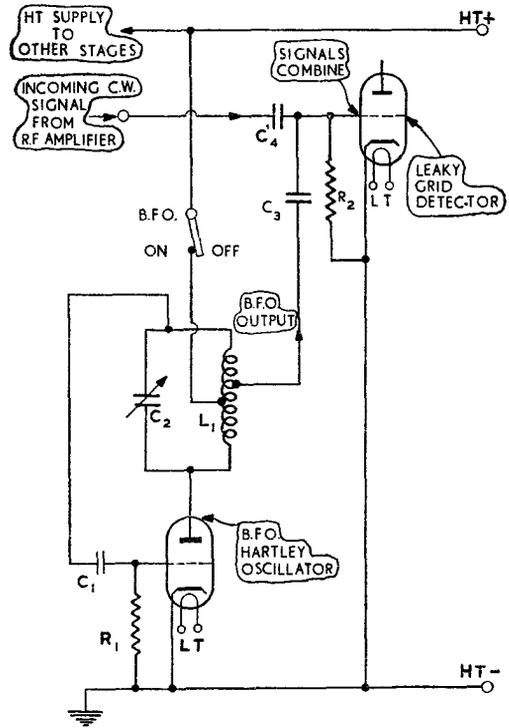
Fig. 35. TYPICAL BEAT FREQUENCY OSCILLATOR CIRCUIT.

they are in phase and additive. The result is that the amplitude of the combined signal varies at a frequency of 1 kc/s, i.e., at the *difference* between the two original frequencies. The same principle applies for any frequencies. Thus, if an incoming c.w. signal at a frequency of 560 kc/s is combined with a locally-produced oscillation at a frequency of 561 kc/s, a difference frequency of 1 kc/s results. The application of the heterodyne principle transforms the incoming c.w. signal into an amplitude-modulated wave which can be detected in the normal manner to produce an a.f. output.

oscillator operating under Class C conditions ($C_1 R_1$) and tuned to the required frequency by C_2. The output from the b.f.o. is coupled by C_3 to the detector circuit where it combines with the incoming c.w. signal to produce the required beat note as explained in the preceding paragraph.

47. Since the frequency of the b.f.o. is normally variable over a certain range, for a given incoming signal it is possible to vary the pitch of the resulting beat note until a satisfactory tone is produced. Table 2 gives the beat note produced as a result of varying the frequency of the b.f.o. either side of the required c.w. signal frequency (500 kc/s).

Frequency of Incoming C.W. Signal (kc/s)	Frequency of B.F.O. (kc/s)	Frequency of Beat Note (kc/s)
	502	2
	501·5	1·5
	501	1
	500·5	0·5
500	500	0 (Dead Space)
	499·5	0·5
	499	1
	498·5	1·5
	498	2

TABLE 2. BEAT FREQUENCIES

These results are illustrated in the graph of Fig. 36, from which it is seen that:—

(a) the *pitch* of the beat note can be varied by the operator;

(b) when the frequency of the b.f.o. coincides with that of the incoming signal, no sound is heard in the telephones; this gives what is known as the *'dead space'*;

(c) the same note can be obtained at two settings of the b.f.o. either side of the dead space.
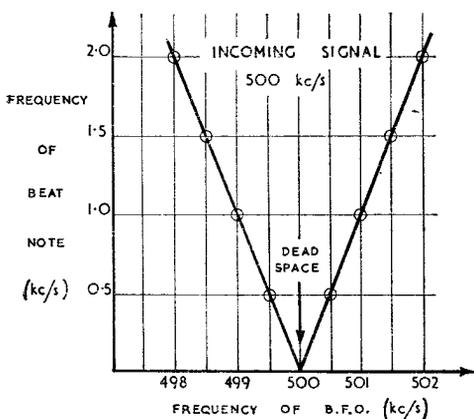
Fig. 36. VARIATION OF BEAT FREQUENCY.

48. The b.f.o. is normally under the control of a switch marked 'R.T.–C.W.'. In the c.w. position, the b.f.o. is switched *on* to give heterodyne reception of the signal. In the r.t. position, the b.f.o. is switched *off*. If the

b.f.o. is left switched on during reception of an amplitude-modulated telephony signal, a continuous whistle is produced as a result of heterodyning between the input signal and the b.f.o. output. This whistle, superimposed on the original modulation, makes reception unintelligible. The b.f.o. must therefore be switched *off*, except when a c.w. signal is being received.

TRANSISTOR T.R.F. RECEIVER

Basic Circuit

49. By making suitable adjustments to the power supply and other circuit details, it is possible to use transistors instead of valves in a t.r.f. receiver. The advantages and limitations of transistors in relation to valves have already been mentioned (see Book 2, Sect. 8, Chap. 7). When transistors are used instead of valves, the same stages are required to carry out the same functions as in a valve receiver. The main advantages in using transistors are miniaturisation, increased reliability and economies in power consumption. The disadvantage is that less power output is possible. Transistor receivers are at present confined to portable and domestic equipment of moderate output and fidelity.

50. Fig. 37 shows the basic circuit of a transistor t.r.f. receiver consisting of a r.f. amplifier stage, a detector stage and an a.f. output stage. Since a b.f.o. is not included in the circuit, reception of amplitude-modulated signals only is possible. The details of the circuit are given below:—

(a) **R.F. amplifier, TR$_1$.** This stage is built round a r.f. transistor, which operates satisfactorily up to 10 Mc/s. The bias network R$_1$, R$_2$ and R$_3$ (de-coupled by C$_3$ and C$_4$) ensures that the junctions are biased in such a way that the emitter is positive with respect to the base (forward bias) and the collector is negative with respect to the base (reverse bias) (see Book 2, Sect.8, Chap. 7, Para. 24). This arrangement ensures that the bias conditions remain constant despite variations in temperature and in current. *D.C. stabilization*, as it is called, is achieved as follows: an increase in collector current (caused by an increase in temperature) means that the emitter current has increased; this causes an increased voltage drop across R$_3$ and the
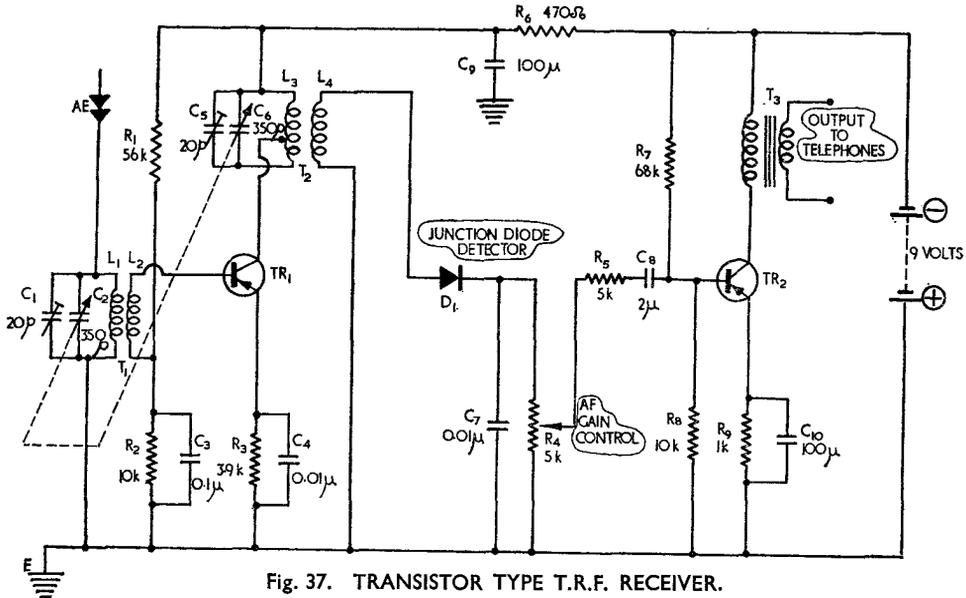
Fig. 37.   TRANSISTOR TYPE T.R.F. RECEIVER.

voltage at the emitter becomes more negative, so reducing the forward bias emitter-to-base; this reduces the base current, which in turn restores the collector current to its original value and stabilization has been achieved.   The rest of the circuit is conventional.   The signal is selected in the input tuned circuit and matched by the step-down r.f. transformer T_1 into the base of TR_1.   The small signal current at the base causes a large variation in collector current so that an amplified signal voltage is developed across the high impedance load in the collector and matched by the step-down r.f. transformer T_2 into the detector circuit.   The impedance match between TR_1 and the collector tuned circuit is by an auto-transformer tap on L_3.

(b) **Detector D_1.**   Rectification is obtained by means of a crystal diode (see Book 2, Sect. 8, Chap. 7), which produces results similar to those obtained from a diode detector valve.   The capacitor C_7 acts as a r.f. by-pass component and the a.f. component is developed across R_4 which operates as an a.f. gain control.   The required a.f. voltage tapped off R_4 is applied through R_5 and C_8 to the base of TR_2; C_8 acts as a d.c. blocking capacitor but in addition, in conjunction with R_5 it provides the correct impedance match between the low input impedance of TR_2 and the impedance of the detector circuit.

(c) **A.F. amplifier, TR_2.**   The transistor used in this stage is a power transistor with a rated dissipation of several milliwatts. The correct operating conditions are provided by the biasing and d.c. stabilization circuit R_7, R_8 and R_9 (de-coupled by C_{10}). This network operates in the same way as the network R_1, R_2 and R_3 in TR_1.   The small a.f. base current resulting from the application of the detector output to TR_2 is amplified by normal transistor action to produce a large a.f. collector current; the latter is applied to the primary of T_3, which acts as an impedance match to transfer the a.f. power output to the telephones. R_6 C_9 are normal de-coupling components to prevent feedback between the output stage and the earlier stages.

LIMITATIONS OF T.R.F. RECEIVER

Sensitivity, Selectivity and Instability

51.   In these days of long-distance wireless communication where a large number of stations are operating on the same waveband with very little separation in frequency, it is essential that a receiver, to operate efficiently has:—

(a) high sensitivity, so that weak signals may be received;

(b) good selectivity, so that the required

signal may be received without interference from other signals on adjacent frequencies. The sensitivity of a receiver depends on the number of tuned amplifiers used and on the gain of each stage; the selectivity of a receiver depends on the number of tuned circuits used and on the bandwidth of each tuned circuit. Thus it would appear that to improve sensitivity and selectivity in a t.r.f. receiver more tuned r.f. amplifier stages would provide the answer. Despite the losses at radio frequencies (due to the shunting effect of stray capacitances, skin effect and so on) this is true, up to a point, provided that the tuned circuits are not required to be tuned *over a range of frequencies*; that is, provided that the circuits are '*fixed*' tuned circuits concerned with one frequency only. Where a large number of r.f. amplifiers in cascade have to

be tuned over a range of frequencies all sorts of difficulties arise—difficulties associated with the ganged capacitors and feedback between the various sections of the capacitors. The overall result is that the whole arrangement becomes unstable with a tendency to self-oscillation, and howls and whistles are produced in the telephones. Thus, in a t.r.f. receiver it is virtually impossible to employ more than about two stages of r.f. amplification without incurring the instability mentioned. The result is that the sensitivity and selectivity obtainable with a t.r.f. receiver are less than those required under modern conditions. The requirements of high sensitivity and good selectivity, coupled with good stability, can only be satisfactorily met in the *superheterodyne receiver*. This type of receiver is discussed in Chapter 2.

SECTION 14

CHAPTER 2

SUPERHETERODYNE RECEIVER

SUPERHETERODYNE RECEIVER

Introduction

1. It is shown in Chapter 1 of this Section that the basic t.r.f. type receiver is incapable of providing the high sensitivity and high selectivity required under modern conditions. Since selectivity and sensitivity depend on the number of tuned voltage amplifiers used, more r.f. amplifiers could be inserted as shown in Fig. 1(c). However, the fact that the r.f. stages must be capable of being tuned over a *range* of frequencies gives rise to certain problems. The need for variable tuning gives an inefficient amplifier because of complications in aligning three or more variable tuned circuits over a range of frequencies. Waveband switching is also complicated and leads to losses at the switch contacts. In addition, a multi-stage r.f. amplifier, tuning over a range of frequencies, becomes generally unstable. This arises because the circuits are 'coupled' via the

main ganged tuning capacitor shaft, and undesirable feedback and interaction between the stages produces howls and whistles in the telephones. The situation becomes even worse if the stages are required to operate at a *high* r.f. because then losses become considerable due to shunting effects of stray capacitance and skin effect.

2. If the receiver were to be tuned to only *one* frequency, the r.f. amplifiers could use pre-tuned circuits which were 'fixed' at that frequency. Correct alignment of the tuned circuits at this fixed frequency would be simple and no waveband switching would be required. No tuning would be necessary, so no ganged capacitor would be needed, and interaction and feedback between the r.f. stages via the common tuning element would be prevented. If at the same time the frequency at which the pre-tuned circuits were required to operate was a *low* radio frequency,
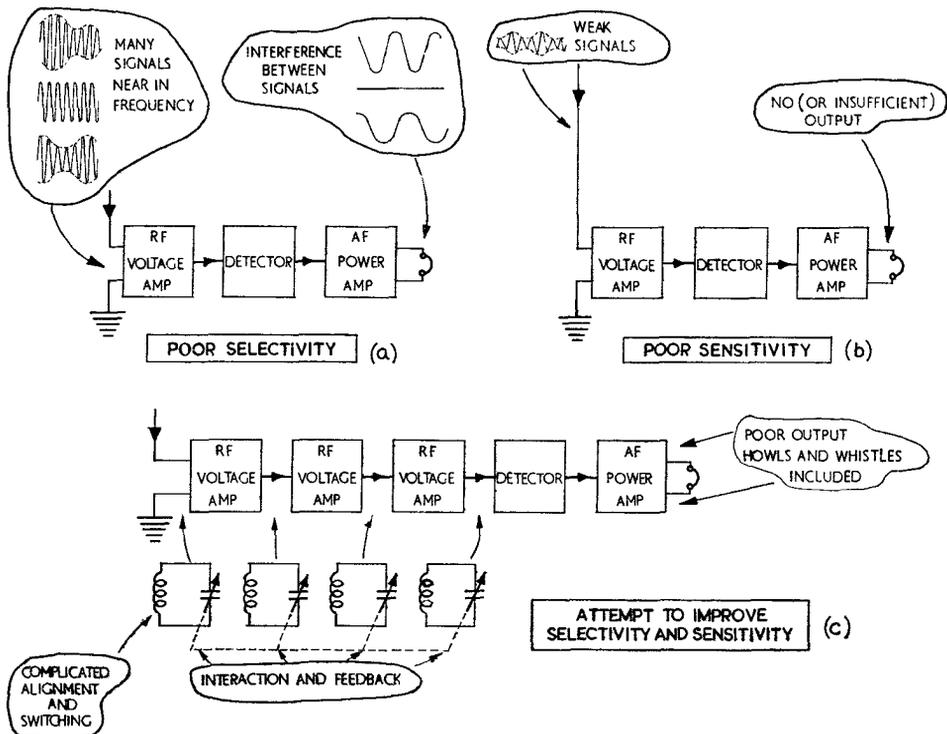

Fig. 1. LIMITATIONS OF T.R.F. RECEIVER.
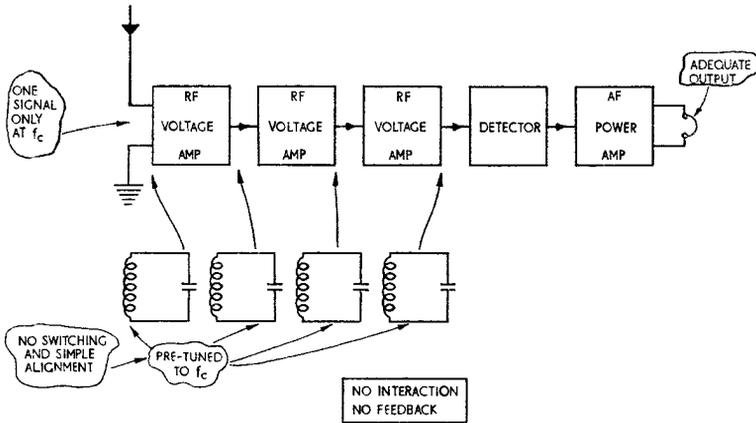
Fig. 2. ADVANTAGES OF FIXED FREQUENCY AMPLIFICATION.

losses caused by stray capacitance and skin effect would be considerably reduced. Thus, since fixed frequency tuned amplifiers are efficient and can be designed for good stability, many more stages of amplification could be inserted to provide the required selectivity and sensitivity *without instability*.

3. It may well be said that a receiver capable of dealing with only one signal at a certain fixed low radio frequency would be of limited use. This is true; but if a way could be found for selecting *any signal* in a given frequency band and then *converting* the signal to the frequency of the pre-tuned amplifier stages, the result would be a practical receiver capable of providing high selectivity and high sensitivity with little tendency to instability.

The system is shown in block form in Fig. 3 and is the basis of the *superheterodyne*

receiver. The required signal is selected and then converted in frequency to that at which the pre-tuned amplifiers operate. This conversion process, as will be seen later, does not change the 'form' of the signal in any way so that after adequate amplification in the pre-tuned stages, it can be detected in the usual way, amplified and applied to the telephones.

Superheterodyne Principle

4. In a superheterodyne receiver, the frequency of the incoming signal is changed to a pre-determined fixed frequency called the *'intermediate frequency'* (*i.f.*). Since the i.f. is normally a low radio frequency in relation to the frequency of the incoming signal all the advantages of amplification at a fixed low radio frequency are obtained.
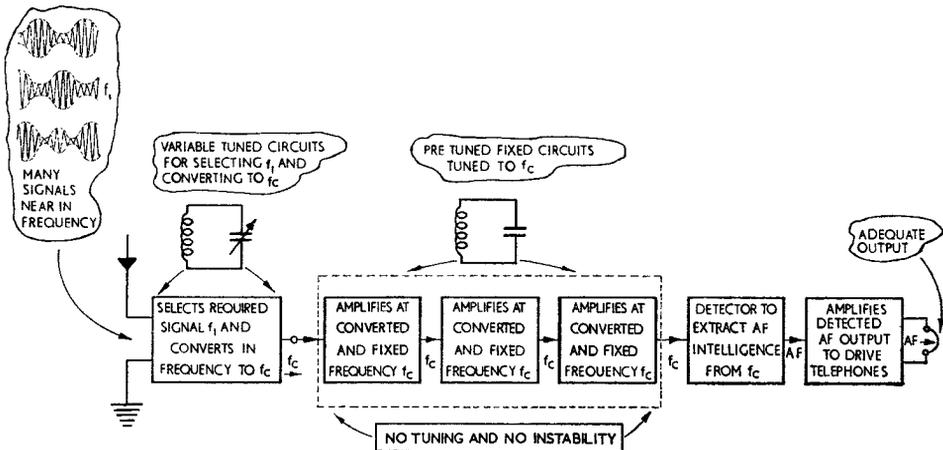
Fig. 3. BASIS OF SUPERHETERODYNE RECEIVER.

The i.f. is obtained as a result of 'mixing' the incoming signal with the output from a 'local oscillator' in a manner similar to that used in the heterodyne principle (see Chap. 1, Para. 43). However, in this case, the difference frequency is above the audible range (supersonic) and the principle is referred to as the supersonic heterodyne principle or more briefly, the superheterodyne principle. This distinguishes it from the heterodyne reception of c.w. where the result of heterodyning produces an audible output.

5. The portion of the superheterodyne receiver which selects the incoming signal and converts it to the i.f. can normally be split up into three separate stages as shown in Fig. 4, namely:—

(ii) provides additional selectivity;

(iii) improves the signal-to-noise ratio;

(iv) reduces radiation from the local oscillator.

(b) **Local Oscillator.** This stage is a variable frequency r.f. oscillator which produces an output at a frequency f_2 for application to the frequency changer. The frequency f_2 is such that it always differs from the selected signal frequency f_1 by an amount equal to the i.f. (f_c). The local oscillator tuned circuit is ganged to the tuned circuits in the r.f. amplifier stage so that operation of the main tuning control alters the frequency f_1 to which the r.f. amplifier is tuned and the frequency f_2 of the local oscillator simultaneously. In
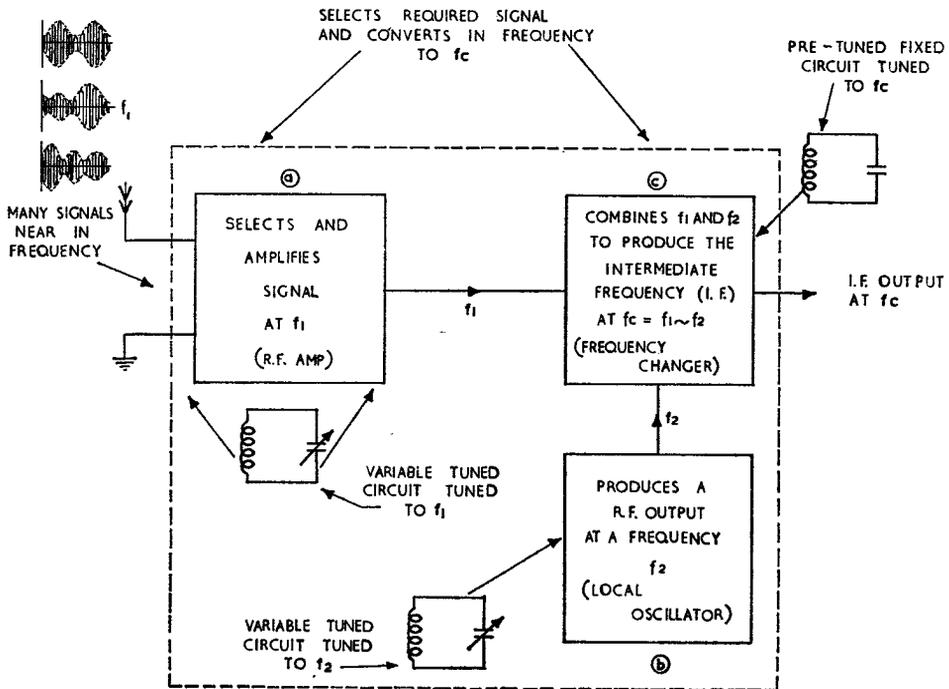
Fig. 4. TUNED STAGES GIVING FREQUENCY CONVERSION IN SUPERHETERODYNE RECEIVER.

(a) **R.F. amplifier.** This stage operates in the usual manner of a r.f. voltage amplifier to select and amplify the incoming signal on any frequency f_1 for application to the frequency changer. The r.f. amplifier:—

(i) provides the frequency changer with a good signal input;

this way, a constant *difference* in frequency equal to f_c is maintained between f_1 and f_2.

(c) **Frequency Changer.** This stage is sometimes referred to as the '*First Detector*' or '*Mixer*' stage. Its purpose is to combine the output from the r.f. amplifier at the signal frequency f_1 and the output from the

local oscillator at a frequency f_2 in such a way that the output from the frequency changer will contain a component that is equal in frequency to the *difference* between f_1 and f_2, namely the intermediate frequency $f_c = f_1 \sim f_2$. Two methods of frequency changing are used—'*additive*' and '*multiplicative*'; these methods will be considered in later paragraphs.

The three stages given in (a), (b) and (c) are the only stages in the superheterodyne receiver that employ variable tuning. Their main function is the efficient changing from the signal frequency to the intermediate frequency. The succeeding tuned amplifiers have only one frequency to deal with—the i.f.—so that the i.f. amplifiers use pre-tuned circuits.

Block Diagram of Superheterodyne Receiver

6. A complete block schematic diagram of a superheterodyne communication receiver is shown in Fig. 5, which also illustrates typical operating frequencies. The function of each stage is explained below:—

(a) **R.F. Amplifier, V_1.** This selects the required signal on any frequency f_1, amplifies it and applies it to the frequency changer V_3 for conversion to the intermediate frequency f_c.

(b) **Local Oscillator, V_2.** This produces a voltage at a frequency f_2 for application to the frequency changer V_3 where it combines with the signal frequency f_1 to provide the intermediate frequency f_c.

(c) **Frequency Changer, V_3.** This combines the outputs from V_1 and V_2 in such a way that the output from the frequency changer is at a frequency equal to the difference between f_1 and f_2, namely the i.f.

(d) **I.F. Amplifier, V_4.** This is a fixed frequency tuned amplifier operating at the i.f. Since this is the only frequency with which the stage is concerned, the associated tuned circuits are pre-set and the whole stage is designed to give high gain with good stability and the required selectivity. It is the i.f. amplifier which determines in the main the degree of sensitivity and selectivity of the receiver, and the number of stages of i.f. amplification used depends on the sensitivity and selectivity required; one Service receiver uses eight i.f. amplifiers.

(e) **Detector, V_5.** This is normally a diode detector and its function is the same as that in a t.r.f. type receiver, namely to extract from the selected and amplified signal the information that it is carrying. It has already been stated that the 'form' of the signal is not changed by converting from the signal frequency to the i.f.; the modulation remains the same and the function of the detector is the same.

(f) **B.F.O., V_6.** This is used only for the reception of c.w. signals, its function being the same as that in the t.r.f. type receiver. The b.f.o. produces a signal at a frequency f_0 which combines at the detector with the i.f. output at a frequency f_c. These two frequencies are such that the output from the detector is an audio output whose frequency is equal to the *difference* between f_c and f_0. Thus, if the i.f. is 465 kc/s and the frequency of the b.f.o. is 464 kc/s, the a.f. output from the detector is 1 kc/s.

(g) **A.F. Output, V_7.** This performs the same function as in a t.r.f. type receiver.
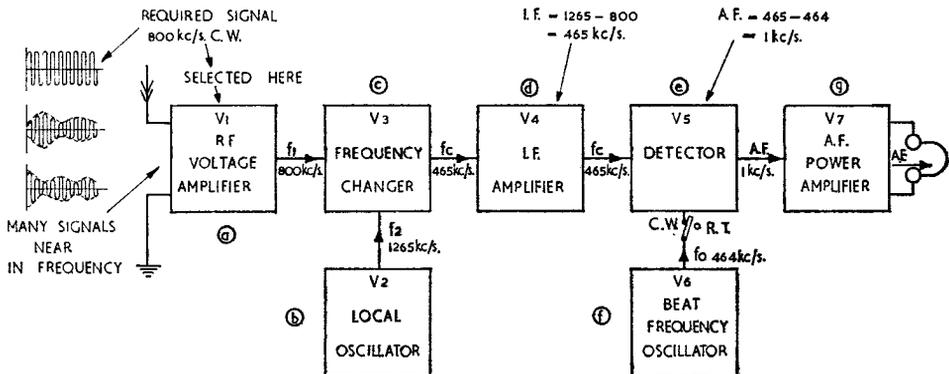

Fig. 5. BLOCK SCHEMATIC DIAGRAM OF SIMPLE SUPERHETERODYNE RECEIVER.

It amplifies the a.f. output from the detector and provides the power necessary to operate the telephones.

Tuning the Superheterodyne Receiver

7. In the superheterodyne receiver any desired incoming signal is converted to a frequency, known as the i.f., that is common for all signals. The i.f. chosen depends on the nature of the equipment and varies considerably from one equipment to another. One common i.f. chosen for domestic receivers is 465 kc/s. All incoming signals are converted to this frequency and all the i.f. amplifiers are pre-tuned to this frequency.

8. To obtain an i.f. of 465 kc/s, it is normal to tune the local oscillator 465 kc/s *above* the signal circuits and to gang the signal and local oscillator tuning capacitors, so that no matter how the tuning is varied, the local oscillator frequency is always 465 kc/s above

difference frequency of $1265 - 800 = 465$ kc/s is obtained; this is the i.f. Similarly by tuning to a signal on 1000 kc/s, the local oscillator is simultaneously and automatically tuned to $1000 + 465 = 1465$ kc/s and the i.f. difference between the two applied voltages is maintained (Fig. 6).

9. There is a good reason for setting the frequency of the local oscillator *above* that of the signal by an amount equal to the i.f. of the receiver. Suppose the receiver i.f. is 465 kc/s. Since the local oscillator frequency must always differ from the signal frequency by 465 kc/s, if it is required to receive the B.B.C. Light Programme on a frequency of 200 kc/s there is only *one possible* setting for the local oscillator, namely $200 + 465 = 665$ kc/s. Thus to receive signals that are at a frequency *lower* than that of the i.f., the local oscillator must be set *above* the signal frequency.
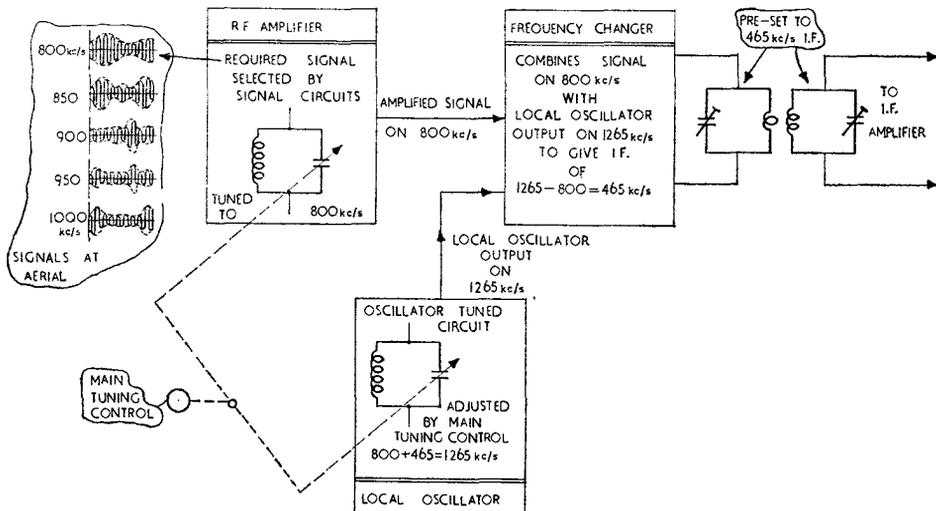
Fig. 6. TUNING THE SUPERHETERODYNE RECEIVER.

the signal frequency. In this way, the difference between the two frequencies is always maintained at the i.f. For example, if it is required to receive a signal on a frequency of 800 kc/s, the main tuning is adjusted so that the signal circuits are tuned to 800 kc/s; since the local oscillator is always 465 kc/s above the signal frequency, the main tuning control will at the same time set the local oscillator frequency to $800 + 465 = 1265$ kc/s. By applying the voltages at 1265 kc/s and 800 kc/s to the frequency changer, the

10. Another reason for setting the oscillator frequency above that of the signal is tied up with the maximum-to-minimum tuning range of a given circuit. Suppose that this receiver (i.f. = 465 kc/s) operates over a tuning range of 1500 kc/s to 500 kc/s on one of its bands. The ratio of maximum to minimum frequencies is $\frac{1500}{500} = 3$, and this is within the range of a normal tuning capacitor.

If the oscillator frequency is set *lower* than the signal frequency by an amount equal to

the i.f., the oscillator tunes between the frequencies of $(1500 - 465) = 1035$ kc/s and $(500 - 465) = 35$ kc/s to produce an i.f. of 465 kc/s in this band. This gives a ratio of $\dfrac{1035}{35} = 29 \cdot 5$, which is impossible to obtain with normal tuning capacitors.

If the oscillator frequency is set *higher* than the signal frequency, the oscillator tunes between the frequencies of $(1500 + 465) = 1965$ kc/s and $(500 + 465) = 965$ kc/s to produce an i.f. of 465 kc/s. This gives a frequency ratio of $\dfrac{1965}{965} = 2 \cdot 04$, which is well within the range of a normal tuning capacitor.

Thus for most purposes, the local oscillator is arranged to tune *above* the signal circuits. There are exceptions to this, mainly at v.h.f., and these exceptions are considered in later paragraphs.

Adjacent Channel Interference

11. This is the name given to interference from a signal that is near (i.e. adjacent) in frequency to the wanted signal. This form of interference is found in all receivers and is *not* peculiar to the superheterodyne receiver. The extent to which the receiver suffers from adjacent channel interference depends on the *selectivity* of the set. Suppose there are two signals, one on a frequency of

1500 kc/s and the other at 1520 kc/s; it is required to select the signal at 1500 kc/s. With a straight t.r.f. type receiver, the signal circuits provide the selectivity and since these are tuned to 1500 kc/s. the percentage separation between the wanted signal at 1500 kc/s and the adjacent channel interference signal at 1520 kc/s is $\dfrac{20}{1500} \times 100 = 1 \cdot 3\%$; the signal circuits find it very difficult to differentiate between two signals with such a small percentage separation and interference from the adjacent channel signal results (Fig. 7(a)).

In a superheterodyne receiver, the i.f. amplifiers provide the selectivity. If an i.f. of 465 kc/s is assumed, the signal circuits will be tuned to 1500 kc/s and the local oscillator will be tuned to 1965 kc/s. The adjacent channel interference signal on 1520 kc/s will also combine at the frequency changer with the local oscillator output to produce a difference frequency of $1965 - 1520 = 445$ kc/s. Since the i.f. circuits are tuned to 465 kc/s, the percentage separation at this frequency between the wanted and unwanted signals is $\dfrac{20}{465} \times 100 = 4 \cdot 3\%$. The interfering signal is now $4 \cdot 3\%$ off resonance as against $1 \cdot 3\%$ off resonance in the straight receiver, so that the response to the interfering signal in the superheterodyne receiver is much less (Fig. 7(b)).
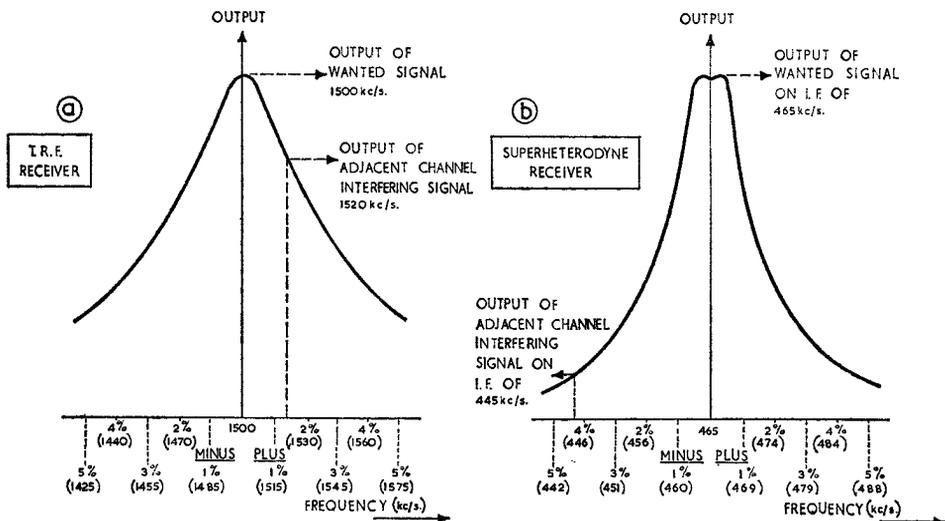

Fig. 7.   ADJACENT CHANNEL INTERFERENCE AND SELECTIVITY IN STRAIGHT AND SUPERHET RECEIVERS.

12. If a lower i.f. had been chosen, the percentage separation between the wanted and unwanted signals at the i.f. would be even greater and the response to adjacent channel interference even less. Thus, the *lower* the i.f., the *less* is the interference from signals on adjacent frequencies.

Second Channel Interference

13. This is the name given to a form of interference peculiar to the superheterodyne receiver. Suppose that the receiver has an unwanted signal). This is illustrated in Fig. 8 where it is shown that the wanted signal on 2500 kc/s combines with the local oscillator output on 2965 kc/s to produce the i.f. of $(2965 - 2500) = 465$ kc/s; and the unwanted signal on 3430 kc/s combines with the local oscillator output on 2965 kc/s to produce the i.f. of $(3430 - 2965) = 465$ kc/s. The unwanted signal is referred to as the '*second channel*' or '*image*' signal. The wanted and image signals are either side of the local oscillator and are separated from
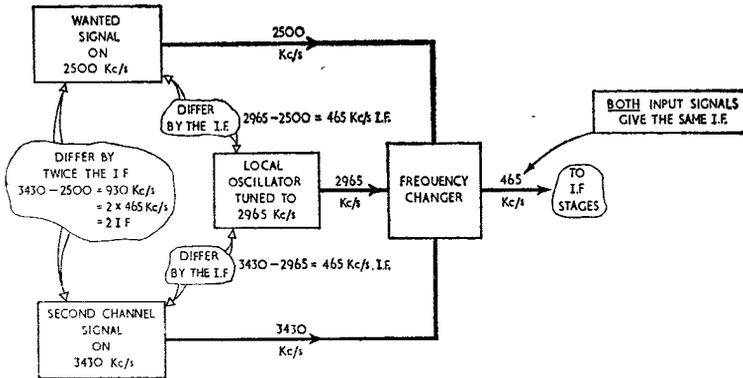
Fig. 8. SECOND CHANNEL (IMAGE) INTERFERENCE.

i.f. of 465 kc/s, and that the oscillator frequency is set higher than the signal frequency. When the receiver is tuned to receive a signal at a frequency of 2500 kc/s the oscillator frequency is $(2500 + 465) = 2965$ kc/s. There are actually *two* signals which can combine at the frequency changer

the local oscillator by an amount equal to the i.f. Thus, the separation between the wanted and second channel signals is equal to *twice* the i.f.

14. In practice there may not be an interfering signal exactly on 3430 kc/s. There
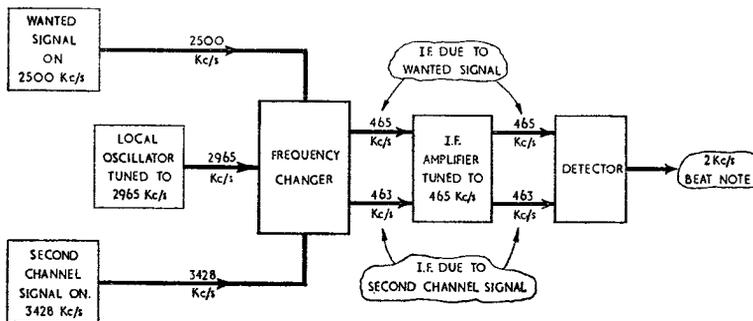
Fig. 9. WHISTLE CAUSED BY SECOND CHANNEL INTERFERENCE.

with the local oscillator output at 2965 kc/s to produce the correct i.f. of 465 kc/s. These are $(2965 - 465) = 2500$ kc/s (the wanted signal) and $(2965 + 465) = 3430$ kc/s (the

may, however, be one slightly off this frequency, say at 3428 kc/s. This unwanted signal will combine at the frequency changer with the local oscillator output on 2965 kc/s

to produce a difference frequency of (3428 − 2965) = 463 kc/s. Simultaneously, the wanted signal on 2500 kc/s will combine with the local oscillator output on 2965 kc/s to produce the correct i.f. of 465 kc/s. Thus, in the i.f. amplifier stages there are two signals—the wanted signal on 465 kc/s and the second channel interfering signal on 463 kc/s. Both signals will pass through the tuned circuits of the i.f. amplifiers and will reach the detector where a beat frequency of (465 − 463) = 2 kc/s will be produced (Fig. 9). This results in an annoying 2 kc/s whistle superimposed on the wanted signal at the receiver output. This heterodyne whistle can be reduced to give the 'dead space' or its pitch may be altered by *detuning* the receiver. This is evident from Table 2 which shows the beat frequencies produced when the receiver is tuned over a small range of frequencies around the wanted signal.

to be a strong local transmitter operating at or near the i.f. of the receiver, this signal may 'break through' to the i.f. stages of the receiver and be superimposed on the i.f. produced by the signal to which the receiver is tuned, resulting in a heterodyne whistle. Another common type of interference is caused by radiation of the local oscillator signals from receivers; for a receiver using an i.f. of 465 kc/s and tuned to the B.B.C. Light Programme on 200 kc/s, the frequency of the local oscillator is 665 kc/s. If this oscillator frequency is radiated from the receiver aerial it will be received by neighbouring sets tuned to the B.B.C. Home Service Programme of 668 kc/s as an unwanted signal, along with the wanted signal. Both signals will pass to the frequency changer and will produce difference frequencies of 465 kc/s (wanted) and 468 kc/s (unwanted); these two frequencies will beat together at the detector

WANTED SIGNAL = 2500 kc/s: I.F. = 465 kc/s: SECOND CHANNEL SIGNAL = 3428 kc/s				
Signal Circuits Tuned to:— (a)	Local Oscillator Tuned to:— (b)	I.F. due to Wanted Signal (c)	I.F. due to Unwanted Signal (d)	Resultant Heterodyne Whistle from (c) and (d) (e)
2496 kc/s	2961 kc/s	461 kc/s	467 kc/s	6 kc/s
2497 ,,	2962 ,,	462 ,,	466 ,,	4 ,,
2498 ,,	2963 ,,	463 ,,	465 ,,	2 ,,
2499 ,,	2964 ,,	464 ,,	464 ,,	0 (Dead space)
2500 ,,	2965 ,,	465 ,,	463 ,,	2 ,,
2501 ,,	2966 ,,	466 ,,	462 ,,	4 ,,
2502 ,,	2967 ,,	467 ,,	461 ,,	6 ,,

TABLE 1. SECOND CHANNEL WHISTLES

15. Since the second channel signal differs from the wanted signal by twice the i.f., the use of a *high i.f.* will give a greater separation between the two signals, and it is then easier for the signal circuits to separate the signals. The value of the r.f. amplifier stage can now be seen. One of its functions is to improve signal frequency selectivity so that the receiver can discriminate against unwanted interfering signals.

Other Types of Interference

16. There are several other ways in which interference takes place in the superheterodyne receiver. For instance, if there happens

and produce an audio whistle of 3 kc/s.

Most interference of the types described is sufficiently reduced by using a r.f. amplifier stage, but if one type of interference is particularly troublesome a *filter* may be necessary to eliminate it.

Choice of I.F.

17. It has been shown that to reduce second channel interference the i.f. should be *high;* to reduce adjacent channel interference the i.f. should be *low*. It is obvious, therefore, that the i.f. selected must be a compromise; it must not be too high, neither must it be too low. The i.f. chosen depends mainly on

the frequency range over which the receiver is required to operate. For receivers operating over the low, medium and high frequency bands a relatively low i.f. (something between 100 kc/s and 600 kc/s) has been found suitable; over these frequency bands, second channel interference can be effectively countered by the selectivity of the signal circuits, the more so if a r.f. amplifier is used.

For receivers operating in the v.h.f. band, second channel interference cannot so easily be eliminated because of the relatively small percentage separation in frequencies. This is one reason for the high i.f. used in v.h.f. receivers (something between about 10 Mc/s and 50 Mc/s i.f.).

18. It is also normal in v.h.f. receivers to set the local oscillator frequency *below* that of the signal by an amount equal to the i.f. This reduces the actual change in oscillator frequency caused by oscillator instability and since v.h.f. receivers usually operate on one of several 'spot' frequencies in a limited frequency range, the ratio of maximum to minimum frequencies is not so important (see Para. 10). Consider a receiver having an i.f. of 500 kc/s receiving a signal on 100 Mc/s. For an instability in the oscillator of $0 \cdot 1\%$, the local oscillator frequency will vary by $\dfrac{0 \cdot 1}{100} \times 100 \cdot 5$ Mc/s = $100 \cdot 5$ kc/s. This represents a variation in the i.f. of $\dfrac{100 \cdot 5}{500} \times 100 = 20 \cdot 1\%$ which would be sufficient completely to detune the i.f. amplifier. However, with a higher i.f. of 10 Mc/s and the oscillator set below the signal frequency of 100 Mc/s, the oscillator will be tuned to 90 Mc/s. For an instability in the oscillator of $0 \cdot 1\%$, the local oscillator frequency will vary by $\dfrac{0 \cdot 1}{100} \times 90$ Mc/s = 90 kc/s, a small change in frequency itself. The variation in the i.f. is only $\dfrac{90}{10,000} \times 100 = 0 \cdot 9\%$, a very small amount of detuning.

Thus, at v.h.f., a high i.f. and the local oscillator set below the signal frequency are both needed to reduce second channel interference and give improved stability.

Ganging and Tracking

19. It is shown in Chapter 1 that with a straight receiver having several tuned r.f.

stages it is difficult to find weak signals unless the tuning capacitors are rotated simultaneously, as the maximum gain is available only when all the circuits are tuned to the same frequency. This led to the introduction of the ganged capacitor. In addition, to make all the circuits tune at the same setting of the variable capacitors it is necessary to provide a small variable trimmer capacitor connected across each tuning coil so that the differences in stray capacitances can be eliminated. It is also necessary for the coils in any range to have the same inductance, as nearly as possible; the inductance adjustment is often made by the use of a 'slug' or inductor trimmer which can be screwed in and out of the coil until the required inductance value is obtained. (See Chap. 1, Para. 8.)

20. With the superheterodyne receiver it is the usual practice to tune the signal frequency and oscillator circuits together. The ganging of the r.f. signal circuits is simple and follows the same practice as that for t.r.f. receivers. The oscillator circuit, however, is more troublesome because of the different frequency range being covered. For example, consider a receiver having an i.f. of 450 kc/s tuning over the range 500 kc/s to 1500 kc/s; the oscillator is required to tune over the range 950 to 1950 kc/s (Fig. 10). The ratio of maximum to minimum frequencies in the signal circuits is $\dfrac{1500}{500} = 3$, and since the frequency of a tuned circuit is inversely proportional to the square root of the capacitance $\left(f \propto \dfrac{1}{\sqrt{C}} \right)$ the ratio of maximum to minimum capacitance required is 9. If the oscillator circuit is tuned by a similar variable capacitor (with a ratio of 9), the oscillator inductance can be made such a value that it will tune with the *minimum* capacitance to give the required oscillator frequency of 1950 kc/s at the *top end* of the tuning range; however, with the variable capacitor at its *maximum* capacitance setting the oscillator will produce an output on 650 kc/s, whereas 950 kc/s is required at the *low end* of the oscillator tuning range. Thus, with the same type of variable capacitor as the signal circuits, it is seen that the oscillator tunes over too great a range; a smaller range of capacitance variation in the oscillator circuit
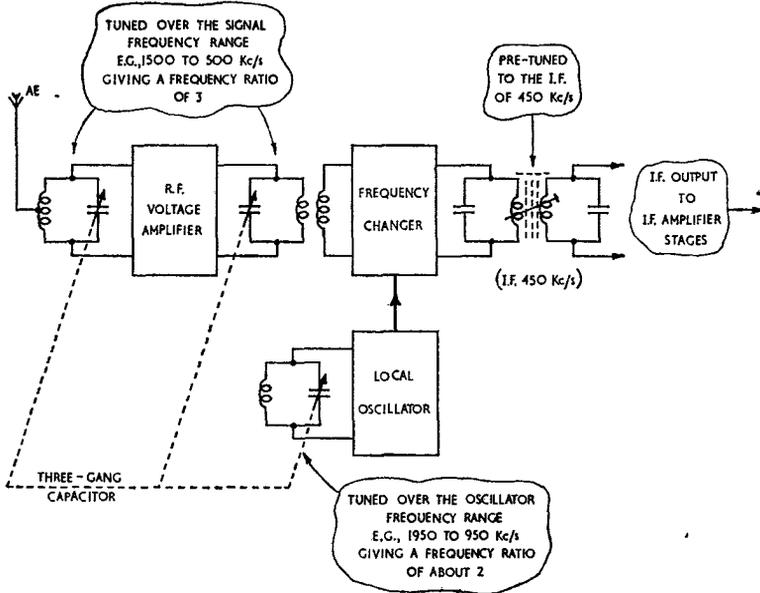
Fig. 10. GANGING OF OSCILLATOR AND SIGNAL TUNED CIRCUITS.

is required.   In some cases the oscillator section of the tuning capacitor has specially shaped vanes to give a smaller capacitance range, but more usually the method described below is used.

21.   The reduction in the ratio of maximum to minimum capacitance across the oscillator coil can be obtained by inserting a variable capacitor, known as a 'padder', in series with the oscillator section of the main tuning capacitor (Fig. 11).   This reduces the effective maximum capacitance to a value sufficient to tune the oscillator inductance to the required 950 kc/s.   Such a capacitance will be fairly large (around 500 pF) and will have little effect on the circuit when the tuning capacitor is at minimum capacitance setting.

Fig. 11. PADDER CAPACITOR.

Thus, at the minimum capacitance setting of the main variable capacitor, the oscillator inductance can be adjusted so that the oscillator produces the frequency of 1950 kc/s; and at the maximum capacitance setting, the padder can be adjusted until the oscillator produces the frequency of 950 kc/s.   This
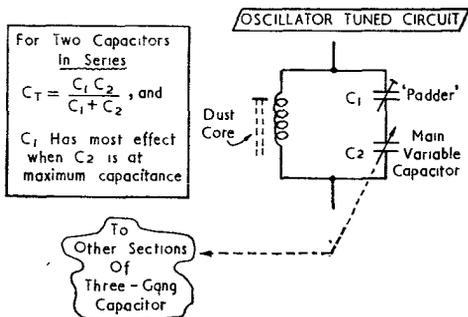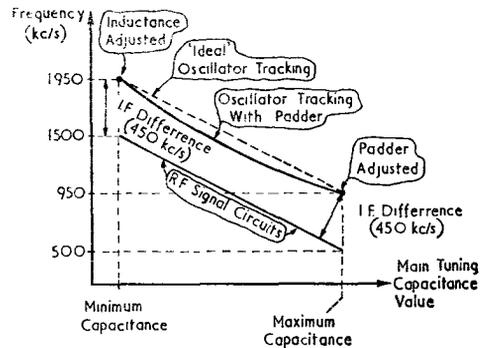


Fig. 12. TWO-POINT TRACKING WITH PADDER.

ensures that the correct frequency difference (equal to the i.f.) is obtained at two points in each tuning range.   There is, however, no guarantee that the 'tracking' will be correct at intermediate points in the range (Fig. 12).

22.   An alternative method is to increase the value of the trimmer in parallel with the oscillator section of the main tuning capacitor.
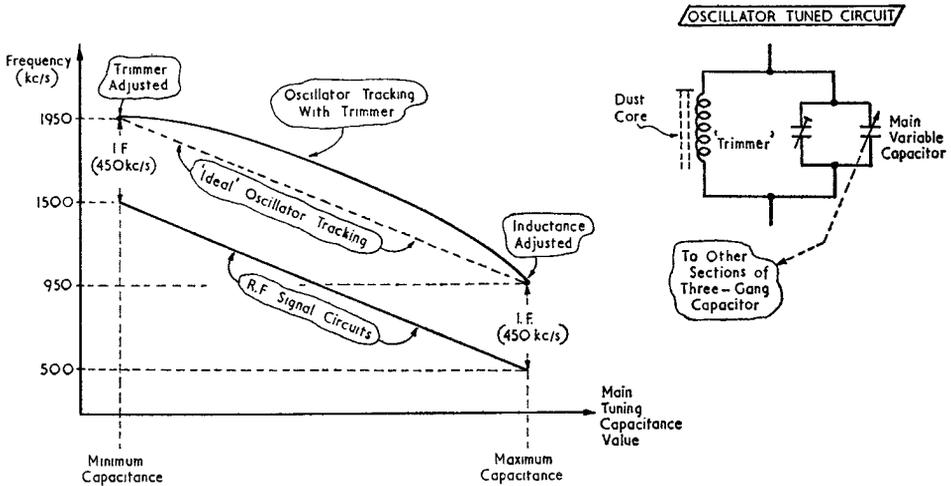
Fig. 13. TWO-POINT TRACKING WITH TRIMMER.

This again restricts the capacitance range so that at the *maximum* capacitance setting of the tuning capacitor the oscillator inductance is adjusted to give an oscillator output on 950 kc/s; and at the *minimum* capacitance setting, the trimmer can be adjusted until the oscillator produces the frequency of 1950 kc/s. This again gives '*two-point tracking*' as shown in Fig. 13.

23. In practice, a combination of padder and trimmer is used to give '*three-point tracking*'. A point is chosen in the *middle* of the range at which the *inductance* value is adjusted to give the correct i.f. difference. The *trimmer* is adjusted at the *top end* of the frequency range, where it has most effect, until the frequency of 1950 kc/s, and therefore the correct i.f. difference, is obtained. The *padder* is adjusted at the *bottom end* of the frequency range, where it has most effect, until the frequency of 950 kc/s, and therefore the correct i.f. difference is obtained. The tracking errors then become very small as shown in Fig. 14. Each frequency range has its own padders and trimmers of course, and these are inserted by the waveband switch.
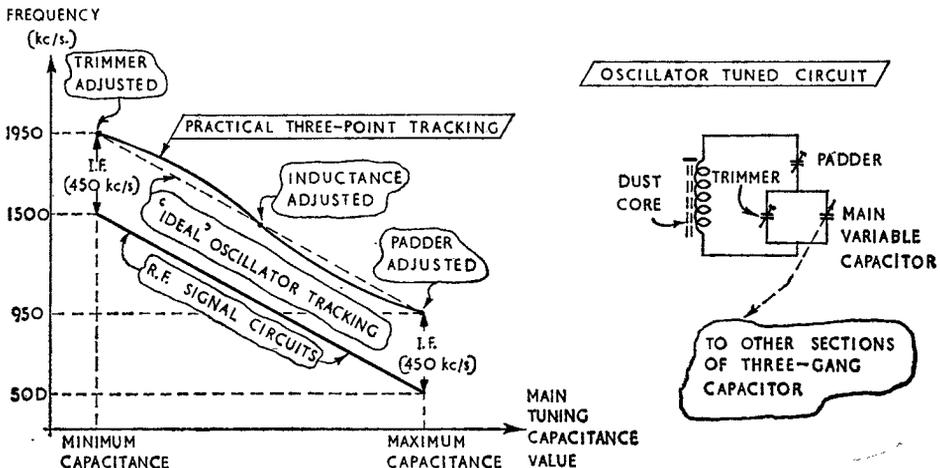

Fig. 14. THREE-POINT TRACKING.

Advantages and Disadvantages of Superheterodyne Receiver

24. Advantages.

(a) By the use of fixed tuning (excepting the signal and oscillator tuning) high efficiency circuits can be obtained, giving high gain and improved selectivity.

(b) Circuit stability is greater with fixed frequency amplifiers operating at a low frequency (the i.f.) than with variable tuned high gain r.f. amplifiers as required in t.r.f. receivers.

(c) The flat top response of the i.f. stages ensures adequate reproduction of the sidebands and improved fidelity.

25. Disadvantages.

(a) Reception can be spoiled by whistles caused by second channel interference, interference from neighbouring receivers and i.f. break-through.

(b) Unless the receiver uses a r.f. amplifier stage, the signal-to-noise ratio may be poor because of the additional noise introduced by the multi-grid frequency changer valve (see Para. 28).

Receiver Noise

26. Noise in a radio receiver is defined as any unwanted electrical disturbance, usually of a random nature, due to any causes. Noise in the output of a receiver generally gives rise to a continuous hissing sound in the telephones. The effect of such noise is to tend to mask or obscure a desired signal, and if the output due to noise is greater than the signal output, the signal is completely submerged in the noise and becomes useless.

27. Noise in the output of a receiver is derived from a large number of sources, some external to the receiver and some internal.

Some external noise, such as that caused by electric motors and generators, could be avoided if precautions were taken to screen such components.

External noise due to what is known generally as '*static*' is unavoidable.

Valve noise is probably the main cause of noise generated inside a receiver. Shot effect, partition noise and induced grid noise are all causes of noise in a valve, and these are dealt with in Book 2, Sect. 8, Chap. 3,

Para. 46. The main point to note is that a valve with many electrodes generates more noise than one with fewer electrodes.

Another cause of noise generated inside a receiver is '*thermal agitation*' of components in the circuit. If a component is heated the movement of electrons between the atoms increases, and since this movement is quite random, a very sensitive voltmeter placed across the component would register a series of readings which vary in amplitude and polarity. These random voltages produced in all components by thermal agitation give rise to noise. The amount of noise generated depends on the temperature of the component, the band of frequencies that the component will pass and the resistance of the component.

28. The usefulness of a receiver for the reception of a given signal depends on how easily that signal can be heard above the noise. A strong signal received under high noise conditions is just as bad as a weak signal received under low noise conditions. The important consideration is the *ratio* of signal output to noise output and the aim is to get as high a signal-to-noise ratio as possible.

Since the receiver itself introduces noise, and noise is amplified equally with the signal, the output signal-to-noise ratio is *always lower* than that at the input, irrespective of the amplification of the receiver. So a high signal-to-noise ratio at the input and a receiver that introduces as little noise as possible are both required.

Because the overall signal-to-noise ratio of a receiver can never be better than that at the first stage, it is important that the first valve in a receiver should introduce as little noise as possible. A multi-grid mixer valve introduces considerably more noise than a pentode. This gives another reason for using a r.f. amplifier stage preceding the mixer in a communications superheterodyne receiver.

Review of Superheterodyne Receiver

29. In the preceding paragraphs, the superheterodyne principle and the basic operation of the superheterodyne receiver have been explained at block schematic level. It is now necessary to 'look inside' the various 'boxes' and find out how the circuits are arranged to give the required conditions. Each stage of the block schematic diagram of Fig. 15 is considered in turn in succeeding paragraphs.
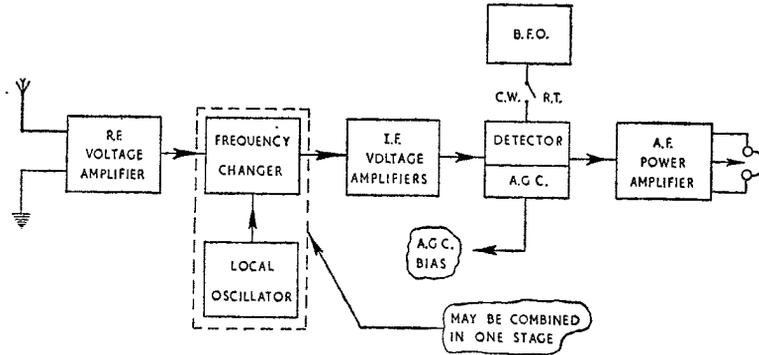
Fig. 15. SUPERHETERODYNE RECEIVER AT " BLOCK " LEVEL.

R.F. Voltage Amplifier

30.   This is a conventional voltage amplifier operating over the required range of signal frequencies. A simplified circuit is shown in Fig. 16.   Only one waveband is illustrated to avoid confusion with waveband switching circuits.   The stage operates under Class A cathode bias conditions ($R_1 C_1$).   The required signal is selected in the grid tuned circuit


Fig. 16.   R.F. VOLTAGE AMPLIFIER.

and applied to the pentode grid for amplification.   The screen potential is reduced to the correct operating level by the voltage drop across R_2 (decoupled by C_2).   The selected and amplified signal is developed across the anode load, the secondary of which is tuned by a second section of the ganged capacitor to the signal frequency.   The signal is then applied for frequency changing.   The trim-

mers are adjusted to ensure correct tracking of the two signal circuits.   The anode is decoupled by $R_3 C_3$.

Not all superheterodyne receivers use a r.f. amplifier stage; in fact, very few domestic receivers use this stage, the signal in this case being selected by the aerial tuned circuit and applied directly to the frequency changer. However, the advantages resulting from the use of a r.f. amplifier stage in reducing second channel and other types of interference and in improving the signal-to-noise ratio are so great that all Service communication receivers employ at least one r.f. amplifier stage.

Local Oscillator

31.   This stage provides the second input to the frequency changer at a frequency differing from the signal frequency by the i.f.   Almost
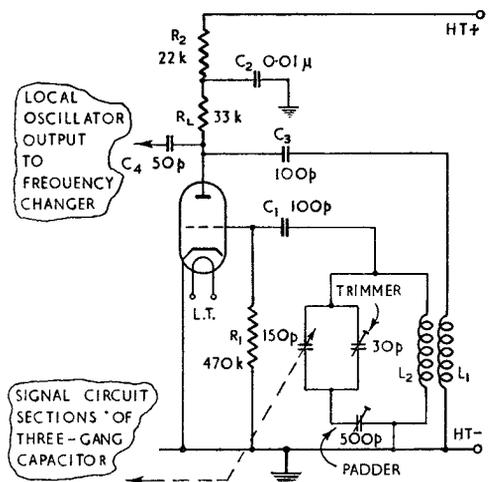

Fig. 17.   LOCAL OSCILLATOR.
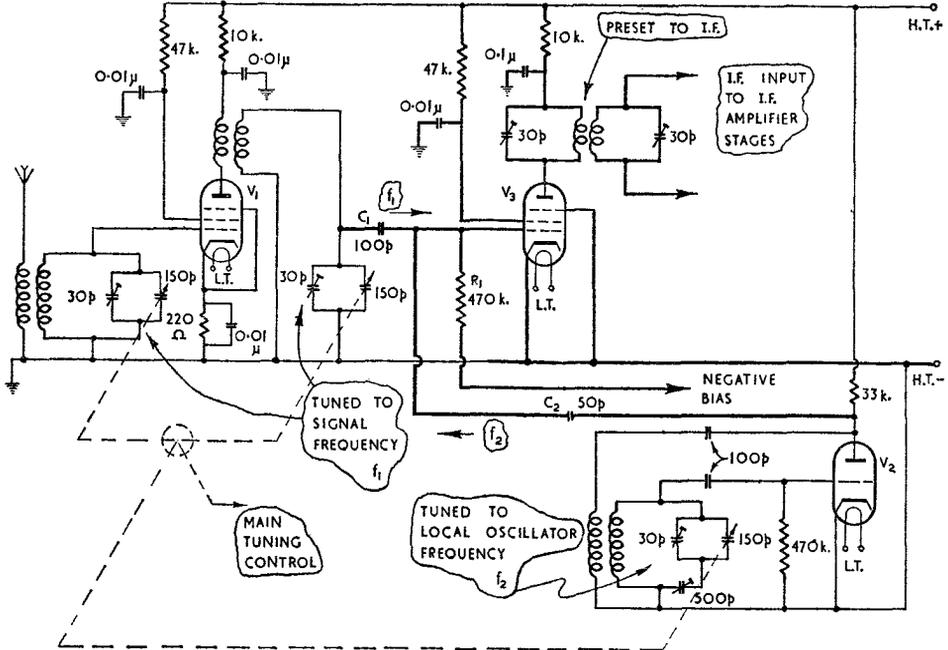
Fig. 18. ADDITIVE FREQUENCY CHANGER.

any of the oscillator circuits discussed in Book 2, Sect. 12 can be used to produce the required r.f. voltage. Fig. 17 shows the simplified circuit of a shunt-fed tuned grid oscillator. The stage operates under grid leak bias from R_1 C_1 and oscillations are maintained by positive feedback from L_1 to L_2 through the mutual inductance. The frequency at which the circuit oscillates is determined by the grid tuned circuit and this is adjusted by the trimmer and padder so that the oscillations differ from the signal frequency by the i.f. at every setting of the tuning capacitor. R_2 C_2 provide anode decoupling and C_3 is the shunt feed capacitor. The oscillations developed across the anode load R_L are coupled via the small value capacitor C_4 to the frequency changer.

Additive Frequency Changer

32. It has been mentioned briefly in Para. 5(c) that two systems of frequency changing are in use in superheterodyne receivers—additive and multiplicative. Additive frequency changing will be considered first. A simplified circuit of an additive type of frequency changer is shown in Fig. 18. V_1 is the r.f. amplifier which selects and amplifies the required signal at a frequency f_1 and applies it to the grid of the frequency changer valve,

V_3 via C_1. V_2 is the local oscillator which provides a voltage at a frequency f_2 differing from the signal f_1 by the i.f.; the local oscillator output at f_2 is also applied to the
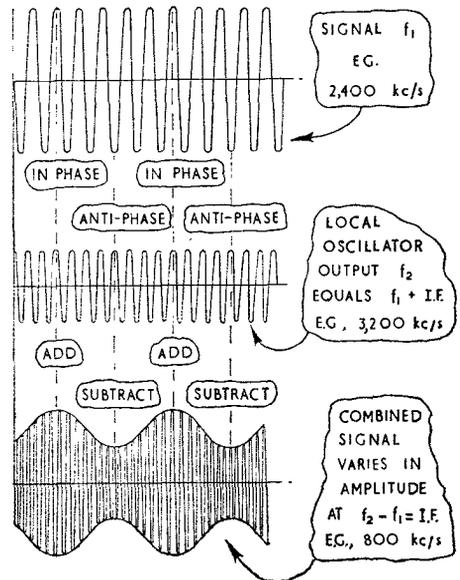

Fig. 19. COMBINING SIGNAL AND OSCILLATOR OUTPUT AT GRID OF ADDITIVE FREQUENCY CHANGER.

grid circuit of the frequency changer valve V_3 via C_2. The two voltages at the grid of V_2 add (hence 'additive') to produce a waveform which varies in amplitude at a frequency equal to the difference between the frequencies f_1 and f_2; that is, the *amplitude* of the combined signal varies at i.f. This is shown in Fig. 19.

33. The additive frequency changer V_3 must be biased so that it operates at the bottom bend of its I_a-V_g characteristic to give non-linear (i.e. detection) conditions. The bias in the circuit of Fig. 18 is obtained from grid leak bias C_1 R_1 plus external bias. With V_3 acting virtually as an anode bend detector
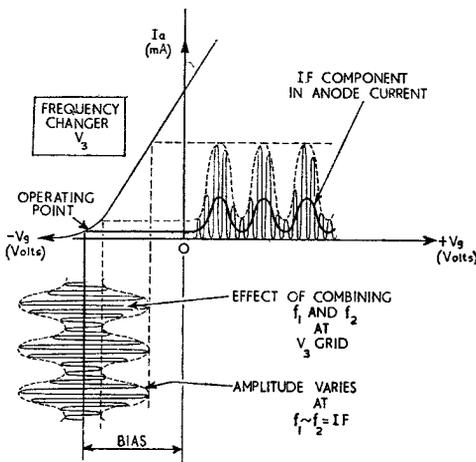

Fig. 20. OPERATION OF ADDITIVE FREQUENCY CHANGER.

the combined signal in the grid circuit causes the anode current to vary in the manner shown in Fig. 20. It is seen that the anode current contains a component at the i.f. This component is selected by the i.f. tuned circuit in the anode of V_3 (see Fig. 18) and applied to the input of the i.f. amplifiers.

34. This method of frequency changing is used mainly at v.h.f. for both communication and radar superhets. The other method of frequency changing (i.e., multiplicative) suffers to a marked degree from '*pulling*' at frequencies higher than about 50 Mc/s. When two circuits operating at two different frequencies, with a small percentage difference between them, are coupled together there is a tendency for one frequency to take over

the control of the other; this effect is known as 'pulling' and is more prevalent at high frequencies.

Pulling is difficult to prevent with multiplicative mixing. With additive mixing it can be prevented by inserting a *buffer* amplifier between the oscillator and the input
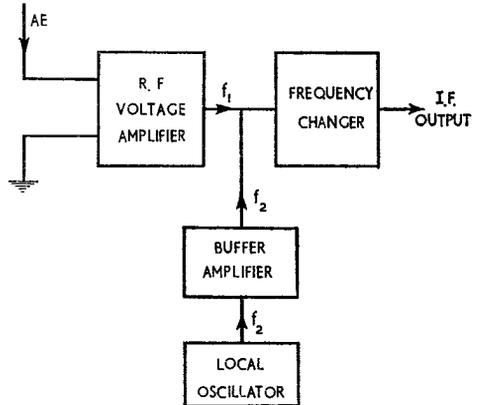

Fig. 21. USE OF BUFFER AMPLIFIER.

to the frequency changer (Fig. 21). Another advantage of a buffer amplifier is that it reduces the coupling between the oscillator and signal circuits and so prevents radiation of the oscillator frequency from the aerial.

Multiplicative Frequency Changer

35. Frequency changing may also be carried out by *multiplying* two signals of different frequencies. The mathematical analysis shows that if this is done, the product contains a frequency which is the *difference* of the two applied frequencies.
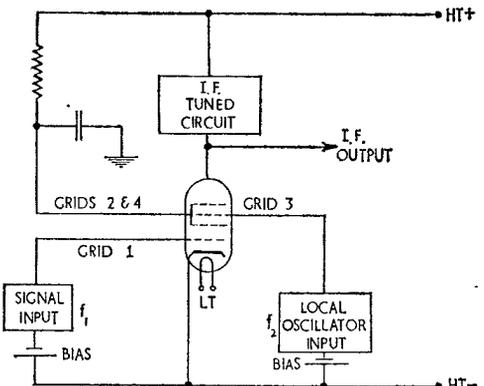

Fig. 22. MULTIPLICATIVE FREQUENCY CHANGER.

A pentode has in the past, been used as a multiplicative frequency changer, with the signal applied to the control grid and the local oscillator output applied to the suppressor grid. In modern circuits, a valve with one more grid than a pentode is used. This is known as a *hexode* valve and the basic circuit for its use as a frequency changer is shown in Fig. 22.

The anode current is controlled by the varying voltages on grid 1 and grid 3. It will be seen that the grids 2 and 4 screen the local oscillator from both the input circuit (signal) and the output circuit (i.f.).

36. The mutual characteristics of such a valve are shown in Fig. 23. Suppose that the signal bias on grid 1 is 3 volts negative, and the oscillator bias on grid 3 is 8 volts negative. Then E represents the static operating point on the mutual characteristics. If the local oscillator generates a r.f. voltage of 8 volts peak, the voltage on grid 3 will vary between zero volts and 16 volts negative and the valve will operate along the path *GEF*.

If at the same time, a signal voltage of 1 volt peak is applied to grid 1, the signal grid will vary between 2 volts negative and 4 volts negative about the signal bias of 3 volts negative.

Since the voltages on grids 1 and 3 are at different frequencies, they will come into step (in phase) and get out of step (anti-phase) at a frequency equal to the difference between the two (the i.f.). When the two voltages are *in phase*, the signal grid rises to 2 volts negative as the oscillator grid rises to zero volts, and falls to 4 volts negative as the oscillator grid falls to 16 volts negative; the operating path is represented on the characteristics by line *CED*. When the two voltages are in *anti-phase*, line *BEA* represents the operating path.

The mean anode current when the valve is operating over the path *C E D* is greater than the mean anode current when the valve is operating over the path *B E A*; and since the two inputs work into phase and out of phase with each other at a frequency equal to the *difference* of the two (the i.f.), the anode current contains a beat frequency component equal to the i.f. which is extracted by an i.f. tuned circuit in the anode.
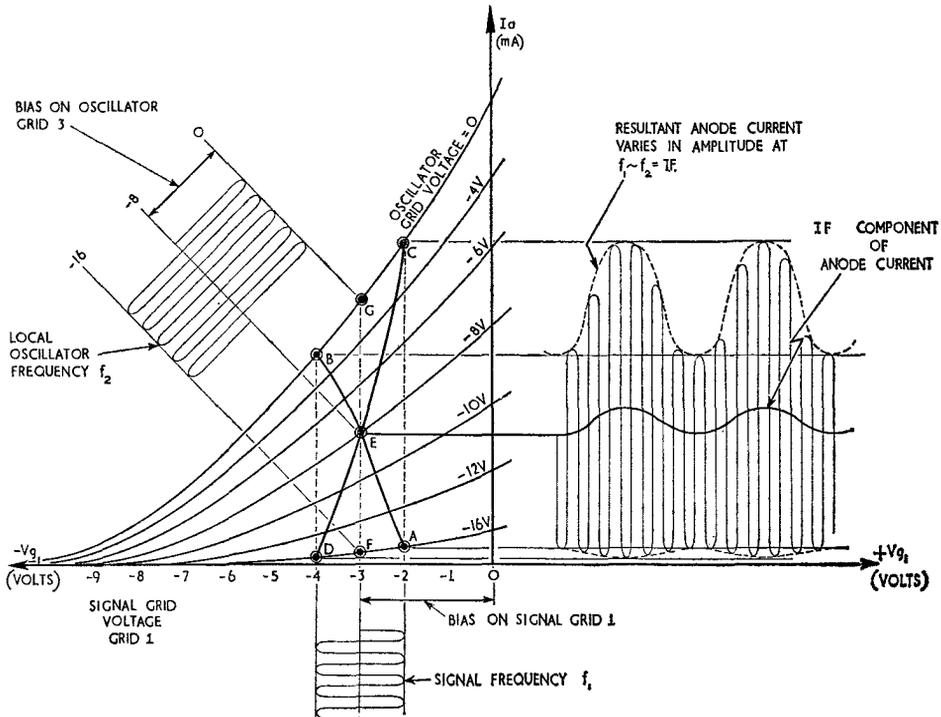

Fig. 23. ACTION OF MULTIPLICATIVE FREQUENCY CHANGER.

37. Triode-hexode. In modern practice, the hexode mixer valve is combined with a triode oscillator in one envelope. A circuit using such a combination is shown in Fig. 24.

V_1 is the r.f. voltage amplifier which selects and amplifies the required signal and applies it to the control grid of the hexode section of the frequency changer valve V_2. The triode section of the frequency changer is arranged as a shunt-fed tuned anode oscillator; it operates under Class C self-bias and its anode circuit is tuned by one section of a three-gang capacitor to a frequency that is above the signal frequency by the i.f.; the padder and trimmer ensure correct tracking with the signal circuits. The oscillation at a frequency
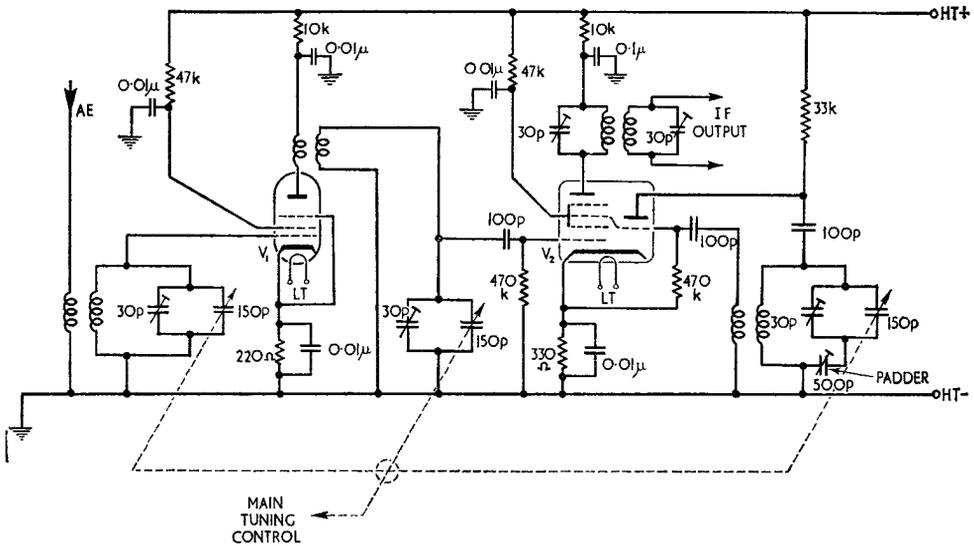

Fig. 24. TRIODE-HEXODE FREQUENCY CHANGER.
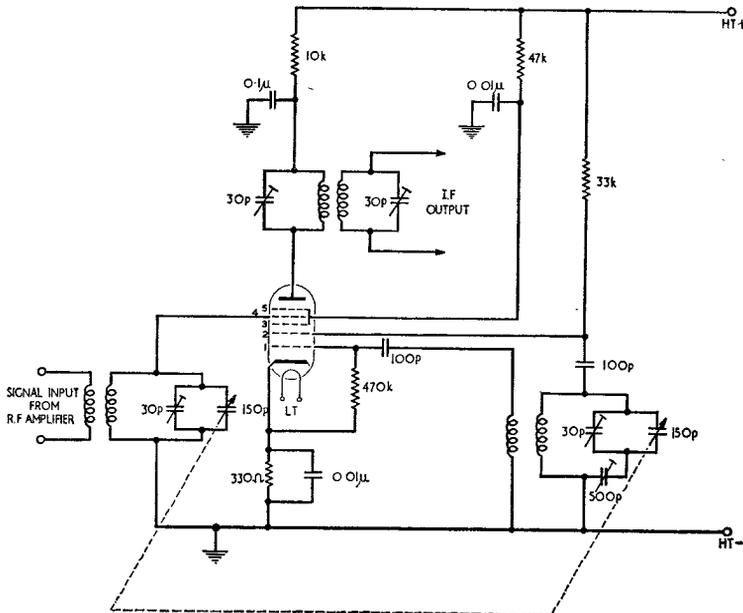
Fig. 25. PENTAGRID FREQUENCY CHANGER.

of (signal + i.f.) is applied from the control grid of the triode section to the injector grid (grid 3) of the hexode section for mixing with the signal.

The hexode section of V_2 operates under Class A cathode bias. The signal from the r.f. amplifier V_1 is applied to the control grid (grid 1) of the hexode and combines with the local oscillation at the injector grid (grid 3) in multiplicative frequency changing to produce a component at the i.f. in the hexode anode current. The hexode anode load consists of a band-pass pre-tuned i.f. transformer which extracts the i.f. component and applies it to the i.f. amplifiers. All stages in the circuit of Fig. 24 are de-coupled in the usual manner at anodes and screen grids.

38. Other multi-grid frequency changers. Other multi-grid valves are used as multiplicative frequency changers. A *heptode* (or *pentagrid*) is one such valve. This is a valve with five grids and it may be arranged in a circuit such as that shown in Fig. 25.

The cathode and the first two grids are connected in a shunt-fed tuned anode oscillator circuit, with the second grid acting as the 'anode'. The signal input from the r.f. amplifier is applied to grid 4. Thus, the electron emission from the cathode is subjected to r.f. variations at two grids—grid 1 and grid 4; the resultant multiplicative mixing produces the i.f. component which is selected in the anode load and applied to the i.f. amplifiers.

Conversion Conductance
39. To measure the efficiency of a frequency changer valve, a term similar to 'mutual conductance' in an amplifying valve is used. This term is '*conversion conductance*' and is the ratio of the i.f. component of anode current to the signal input voltage. It is normally expressed in milliamperes or microamperes per volt.

Comparison of Frequency Changers
40. The additive and multiplicative frequency changers differ as follows:—

(*a*) In additive mixing both signals are applied to the *same* grid (the control grid); in multiplicative mixing the signal input and the local oscillator input are applied to *separate* grids.

(*b*) The additive frequency changer operates under *non-linear* (bottom bend) conditions to provide detection; the mixing portion of a multiplicative frequency changer (as distinct from the oscillator portion) operates under *linear* (Class A) conditions.

These factors are illustrated in Fig. 26.

The multiplicative frequency changer has a higher conversion conductance than the additive type, but because of the tendency to pulling at the higher frequencies, the multiplicative type is limited in use to frequencies up to about 50 Mc/s. Above this frequency, additive types are more common.

I.F. Amplifiers
41. The purpose of the i.f. amplifier or amplifiers is to amplify the i.f. output from the frequency changer to a sufficiently high level to provide adequate detection. The i.f. amplifier stages determine the selectivity and sensitivity of the receiver as a whole.

42. It was shown in Section 13 that a modulated signal contains 'sidebands' and to avoid distortion of the signal, the sidebands must pass through the receiver intact. The bandwidth required depends on the 'type' of signal being received (see Chap. 1, Para. 12).

To give the required bandwidth in i.f amplifiers, band-pass coupled circuits and sometimes staggered tuned stages are used (see Book 2, Sect. 11, Chap. 1, Para. 15). The band-pass requirements of the r.f. signal circuits are not nearly so stringent. This is


Fig. 26. COMPARISON OF ADDITIVE AND MULTIPLICATIVE FREQUENCY CHANGERS.

because the bandwidth as defined in these Notes is a function of the midband frequency. Thus, at the signal frequency the percentage separation between the upper and lower side-bands is smaller than the percentage separation at the i.f., and the signal circuit can easily pass the required band of frequencies. At the i.f. however, the bandwidth requirements must be met by using band-pass circuits or staggered tuning as explained above.

43. A typical two-stage i.f. amplifier circuit is shown in Fig. 27. The i.f. output from the frequency changer is applied to the grid of the first i.f. amplifier; it is then amplified through the two stages and applied to the detector.

44. The band-pass tuned circuits of i.f. amplifiers are made up as transformer units mounted in metal screening cans. Coupling between the two circuits of an i.f. transformer may be inductive as in Fig. 28(a) or it may be capacitive as in Fig. 28(b).

There are two methods of ensuring that the i.f. transformers are tuned to the correct frequency:—

(a) the capacitor across each coil may be a small mica trimmer capacitor, adjustment of which alters the frequency (Fig. 28(a));

(b) the transformer may have fixed capacitors and use variable 'slugs' in the inductors to adjust the frequency (Fig. 28(b)).


Fig. 27. TWO-STAGE I.F. AMPLIFIER.

These stages operate as conventional tuned voltage amplifiers working under Class A cathode bias conditions. However, since they are fixed frequency tuned amplifiers, they can be designed to give high selectivity together with good stability and the required band-pass characteristics. In some receivers, the i.f. tuned circuits are arranged to give variable selectivity (and variable bandwidth) in accordance with the type of signal being received and the conditions existing. The bandwidth is altered by varying the *coupling* between primary and secondary.

A typical i.f. transformer using the first method of alignment is illustrated in Fig. 28(c).

Detector

45. The detector in a superheterodyne receiver is often referred to as the '*second detector*' to distinguish it from the frequency changer which is sometimes termed the 'first detector'. The purpose of the second detector is the same as that of the detector in a t.r.f. receiver; namely, to extract the a.f. component of the modulated i.f. input from the

Fig. 28. I.F. TRANSFORMERS.

final i.f. amplifier, and to apply the detected output to the a.f. stages of the receiver.

46. A diode detector is always used in a superheterodyne receiver because of its good linearity (i.e., freedom from distortion) and because the signal amplitude at this stage in a superheterodyne receiver is sufficiently large to allow a diode to operate efficiently.

The action of the diode detector and filter appears in Chap. 1, Para. 18 of this Section. The circuit shown in Fig. 29 is a series diode detector where the voltage developed across R_3 is the wanted component which is applied to the a.f. stage for amplification.

In practice C_3 and R_3 are usually the grid capacitor and grid leak of the next stage. The

diode connected across the last i.f. tuned circuit tends to damp this circuit which will therefore be less selective than the preceding i.f. tuned circuits.

Reception of C.W.

47. A beat frequency oscillator (b.f.o.) is required for the reception of c.w. signals. Its function is the same as that in a t.r.f. receiver. The b.f.o. produces a signal at a frequency f_0 which combines at the second detector with the i.f. output at a frequency f_c. These two frequencies are such that the output from the detector is an audible beat note whose frequency is equal to the *difference* between f_0 and f_c.
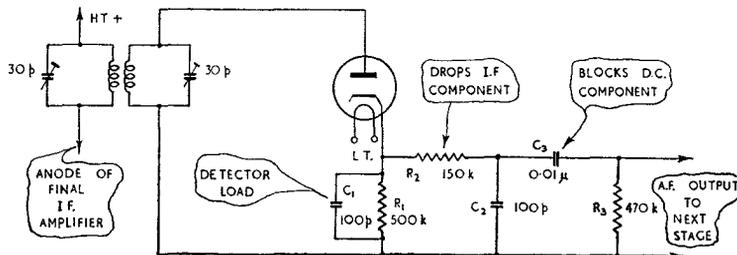
Fig. 29. CIRCUIT OF DIODE DETECTOR.
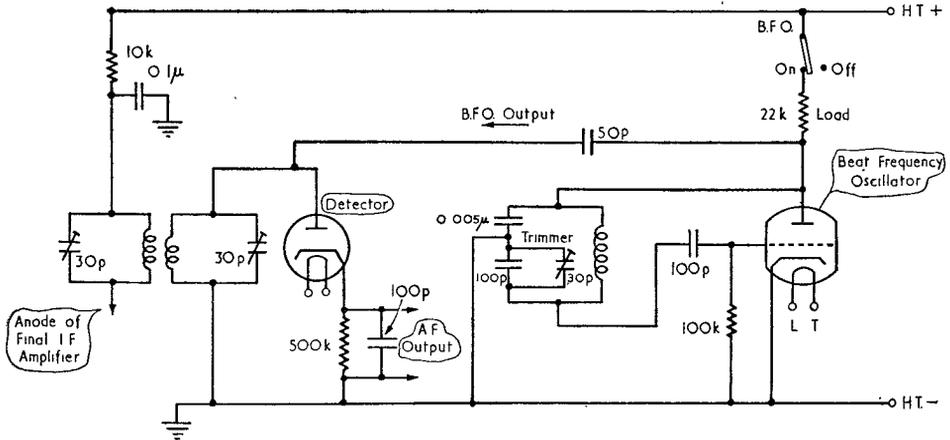
Fig. 30. B.F.O. FOR RECEPTION OF C.W.

48.   The circuit of a typical b.f.o. is shown in Fig. 30.   The circuit is that of a shunt-fed Colpitts oscillator operating under Class C grid leak bias.   It is pre-tuned to a frequency that differs from the i.f. by the required a.f., and this frequency may be altered slightly by the trimmer capacitor in the tuned circuit so that the beat note may be adjusted as necessary.   The output from the b.f.o. is capacitively coupled to the diode detector where it " heterodynes " with the i.f. output to produce the required beat note.   Since the b.f.o. is required only for the reception of c.w. signals, an ON-OFF switch is provided.

49.   The waveforms shown in Fig. 31 may be helpful in understanding the operation of the superheterodyne receiver when it is being used to receive a c.w. signal.

Waveform (a) represents the required input signal to the r.f. amplifier V_1 at a frequency of 700 kc/s; the time axis indicates a period of $\frac{1}{10,000}$ second so that, in this time, 70 cycles will be shown for a frequency of 700 kc/s.

Waveform (b) shows the same signal after it has been amplified in V_1.

Waveform (c) shows the output from the local oscillator V_2 at a frequency of 1050 kc/s; in the time of $\frac{1}{10,000}$ second, 105 cycles appear.

Waveform (d) shows the effect of combining waveforms (b) and (c) at the grid of the additive frequency changer V_3; the com-

bined signal is seen to vary in amplitude at the difference between (b) and (c), i.e., varying in amplitude at $1050 - 700 = 350$ kc/s.   For the period shown, 35 cycles appear.

Waveform (e) represents the output from the frequency changer at a frequency of 350 kc/s (35 cycles shown) after it has been amplified in the i.f. amplifier V_4; since this signal contains no a.f. component, detection would produce no output and a b.f.o. is necessary.

The b.f.o. output at a frequency of 340 kc/s (34 cycles shown) is illustrated by waveform (f).

Waveform (g) shows the effect of combining waveforms (e) and (f) at the input to the detector; the amplitude of the resultant signal varies in frequency at the difference between the i.f. and the b.f.o. inputs, i.e., at $350 - 340 = 10$ kc/s.   For a period of $\frac{1}{10,000}$ second one cycle appears.

Waveform (h) shows the detected output at a frequency of 10 kc/s (1 cycle shown); this can be amplified in a.f. amplifier stages before being applied to the telephones.

In practice, the b.f.o. would be more likely to be set to a frequency of about 349 kc/s so that the a.f. difference produced by combining with the i.f. of 350 kc/s would be about 1 kc/s; 10 kc/s was chosen merely for the sake of illustration and for convenience on a time axis of $\frac{1}{10,000}$ second.
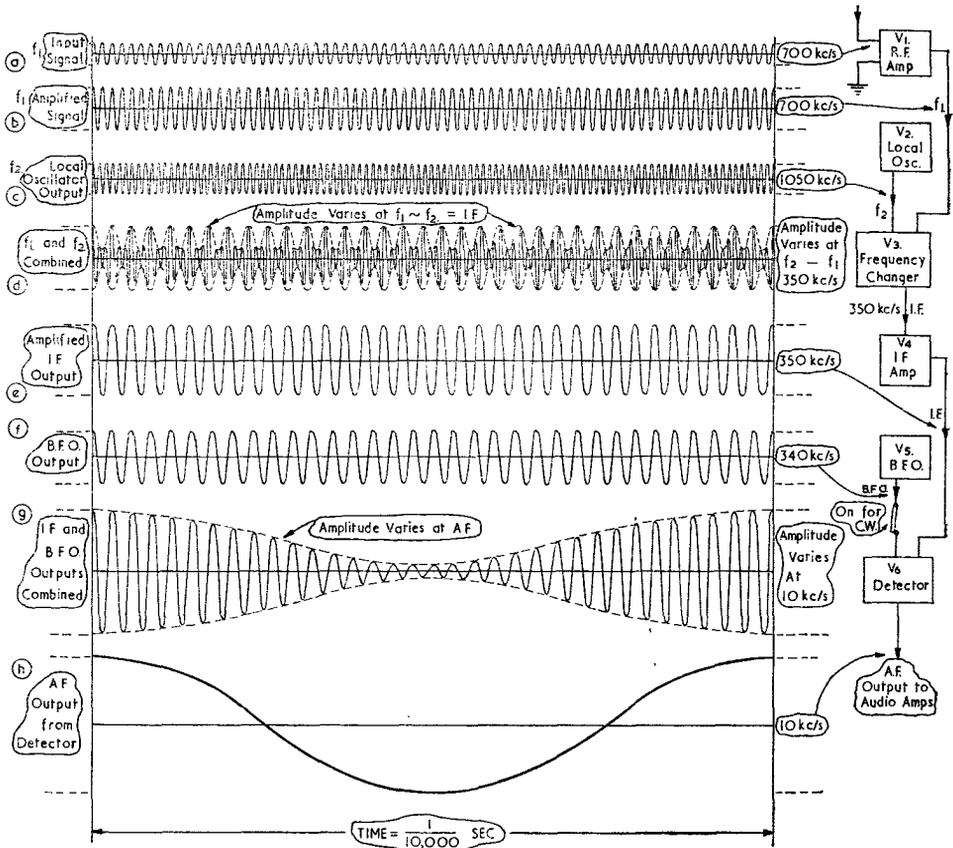
Fig. 31. WAVEFORMS ILLUSTRATING RECEPTION OF C.W. IN A SUPERHETERODYNE RECEIVER.

Reception of Amplitude-modulated Signal

50. For an amplitude-modulated telephony signal, the intelligence is already present in the signal and the b.f.o. must now be switched OFF. If this is not done, a continuous whistle superimposed on the speech will result at the output. The waveforms shown in Fig. 32 may be helpful in understanding the operation of the superheterodyne receiver when it is being used to receive an amplitude-modulated signal.

Waveform (a) represents the required signal at a frequency of 700 kc/s after it has been selected and amplified in the r.f. amplifier stage V_1. The 700 kc/s signal is amplitude-modulated at a frequency of 10 kc/s so that for the period of time indicated, 70 cycles of signal and 1 cycle of modulation appear.

Waveform (b) shows the output from the local oscillator V_2 at a frequency of 1050 kc/s; in the time indicated, 105 cycles appear.

Waveform (c) shows the effect of combining waveforms (a) and (b) at the input to the frequency changer V_4; the combined signal varies in amplitude at the difference between (a) and (b), i.e., it varies in amplitude at $1500 - 700 = 350$ kc/s, and the signal still contains the original modulation at 10 kc/s. For the period shown, 35 cycles of i.f. and 1 cycle of modulation appear.

Waveform (d) represents the output from the frequency changer at a frequency of 350 kc/s after it has been amplified in the i.f. amplifier V_4; the i.f. at 350 kc/s (35 cycles shown) still contains the original modulation (1 cycle shown) so that normal detection can take place. In fact, the only difference between waveforms (a) and (d) is the substitution of the i.f. for the r.f.

Waveform (e) shows the result of detection; an a.f. output at the original modulation frequency of 10 kc/s (1 cycle shown) is obtained. This can be amplified in a.f. amplifier stages before being applied to the telephones.
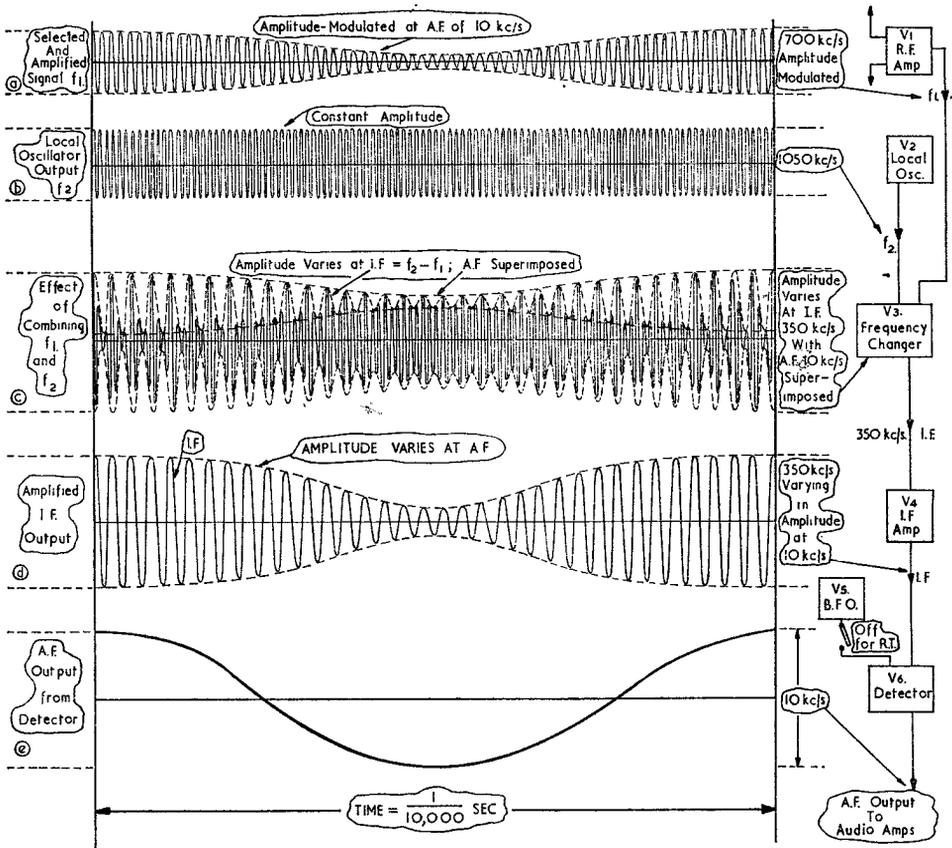
Fig. 32. WAVEFORMS ILLUSTRATING RECEPTION OF AN AMPLITUDE-MODULATED SIGNAL IN A SUPERHETERODYNE RECEIVER.

Output Stage

51. This performs the same function as in a t.r.f. type receiver. It amplifies the output from the detector and provides the power necessary to operate the telephones. A simple circuit is shown in Fig. 33. The a.f. output from the detector is coupled via C_1 and R_1 (which acts as an a.f. gain control) to the control grid of the power output valve. The stage operates under Class A cathode bias conditions. The primary of the output transformer acts as the anode load and couples the amplifier to the telephones; the turns ratio is arranged to give the correct impedance matching.

In some receivers the a.f. output from the detector is not sufficiently large to drive the output stage adequately. In such cases an a.f. *voltage* amplifier is interposed between the detector and the power output valve.

Automatic Gain Control (A.G.C.)

52. The sensitivity of the modern receiver is such that the reception of signals over great distances is common. The intelligibility of these signals is usually limited by '*fading*'. Fading consists of a series of unpredictable fluctuations in the strength of a signal from a distant transmitter. It arises from a number of different causes and will be considered in Section 17. The effect of fading is such that the output of the receiver may at one moment be excessively loud while at the next moment it may fade to the point of inaudibility.

In an attempt to prevent this the operator could continually adjust the manual gain control in such a way as to try to keep the
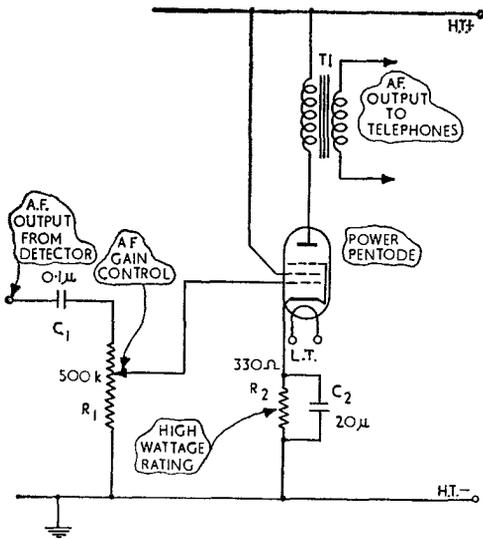
Fig. 33. POWER OUTPUT STAGE.

modulation depth has no effect. The general scheme of control is shown in Fig. 34.

The signal passes through the amplifying stages to the detector and a.g.c. device. Here a control voltage is developed which is proportional to the carrier strength. This control voltage consists of a negative bias which is fed back to the amplifying stages and reduces the gain of each stage, again proportionately to the carrier strength.

55. The valves to be controlled (r.f. amplifier, frequency changer and i.f. amplifiers) must have *variable-mu* characteristics so that the mutual or conversion conductance of the valve is reduced as the negative bias on the control grid is increased. This is shown in Fig. 35 where an increase in the negative bias reduces the gain of the stage.

Simple A.G.C.

56. Fig. 36 illustrates a simple a.g.c. circuit.
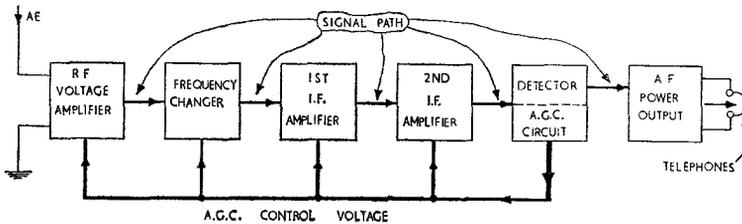
Fig. 34. SYSTEM OF AUTOMATIC GAIN CONTROL (A.G.C.).

output constant despite variations in the signal strength. A much better way is by the addition of a circuit which will accomplish this task automatically—an automatic gain control (a.g.c.) circuit.

53. In a receiver fitted with a.g.c., the sensitivity or gain of the receiver is made to vary *inversely* with the strength of the received signal. With simple a.g.c. the gain is maximum with no signal input. For a strong signal the gain of the receiver is reduced, and as the signal weakens the gain returns towards maximum. The result is that the output of the receiver remains fairly constant despite variations in signal strength. Besides minimizing fading, a.g.c. reduces the necessity for continual re-adjustment of the manual gain control when the receiver is being tuned from one station to another.

54. The operation of this system is controlled by the level of the *carrier* component of the signal which is being reproduced; the
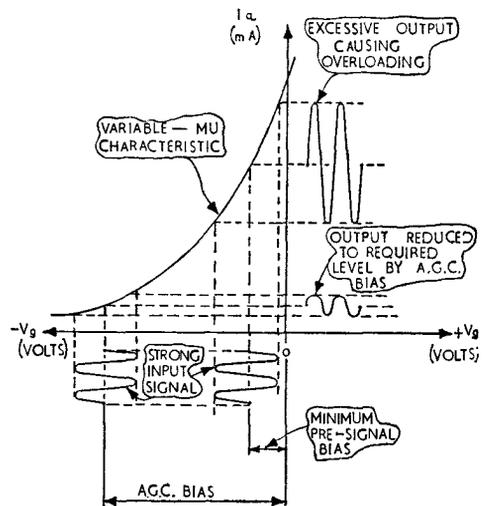
Fig. 35. EFFECT OF A.G.C. BIAS ON VARIABLE-MU VALVES.

The diode in this case has a dual function; it operates as the second detector to provide an a.f. output to the audio stage and it also provides the control voltage for the controlled valves.

In Fig. 36, the resistor R_1 is the diode load, the filter network being omitted for clarity. When the anode becomes positive with
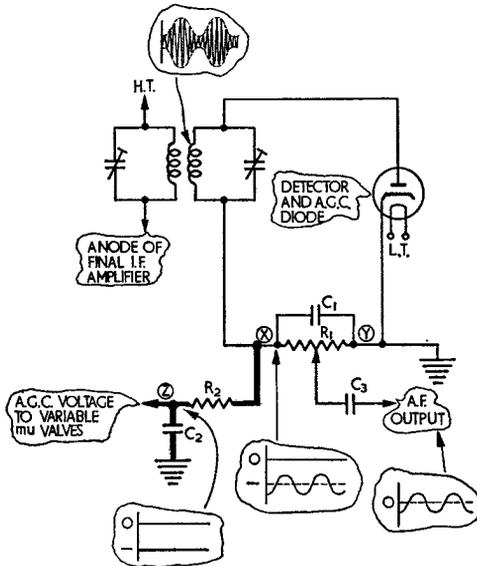

Fig. 36.   SIMPLE A.G.C. CIRCUIT.

respect to the cathode on receipt of a signal, the diode conducts and electrons flow through R_1 from X to Y. Thus X becomes *negative* with respect to Y (earth) by an amount depending on the strength of the received carrier. The voltage across R_1 actually contains the three normal components of detection; namely, the i.f. ripple, the d.c. component and the a.f. component. The i.f. component can be filtered off; the a.f. component can be tapped off R_1 and applied via C_3 to the a.f. stages; and the d.c. component can be applied via R_2 as the a.g.c. voltage.

57.   It is shown in Fig. 36 that the waveform at X consists of an a.f. component super-imposed on a negative d.c. voltage. The a.g.c. filter, consisting of R_2 C_2, acts as a low-pass filter and is so proportioned that to the a.f. component C_2 offers a low reactance; most of the a.f. voltage is therefore developed across R_2. For the d.c. component on the

other hand, C_2 offers an infinite impedance so that the d.c. voltage is developed across C_2 and the voltage at Z is negative with respect to earth by an amount depending on the strength of the received carrier. It is this negative voltage that is applied through the a.g.c. line to the grids of the variable-mu valves in the preceding stages.

58.   The filter is necessary to ensure that the a.f. variations superimposed on the negative d.c. potential are not applied to the grids of the variable-mu valves. If this were allowed to happen the gain of the controlled valves would vary at an audio rate and undesirable feedback effects would result.

On the other hand, the slow variations in the negative d.c. voltage caused by variations in the strength of the received carrier through fading, must be passed down the line. The time constant of the filter must therefore be adjusted accordingly. Typical values are $R_2 = 1$ MΩ, $C_2 = 0 \cdot 1$ μF, giving a time constant of $0 \cdot 1$ second. Since C_2 becomes fully charged in a time of $5\ C_2\ R_2$, these values will cope comfortably with two or three variations every second caused by fading. The time constant is too long however, for a.f. variations to pass down the a.g.c. line.

59.   From the foregoing it is seen that when a signal is being received, a negative voltage proportional to the strength of the received carrier is produced. This voltage is applied down the line to the grids of the variable-mu valves in the r.f. amplifier, frequency changer and i.f. amplifier stages. If the signal increases in strength, the negative voltage produced increases; this results in an increase in negative bias to the variable-mu valves and a consequent reduction in gain. If the signal strength falls, the a.g.c. bias voltage falls and the gain of the controlled stages rises. In this way, an output whose volume is reasonably constant is obtained despite variations in the strength of the received signal.

Delayed A.G.C.

60.   The simple system described in the preceding paragraphs develops an a.g.c. voltage from *any signal* which will produce a detected output. This is a disadvantage with very weak signals where it is required that the *maximum* gain of the valves should be available; such weak signals, therefore, should

be prevented from developing an a.g.c. voltage.

A practical method of effecting this condition is to arrange that no control bias is generated by signals with a strength less than that necessary to produce the 'normal' output. This necessitates the use of a separate valve to produce the a.g.c. voltage. This valve is given a bias voltage to prevent current flowing until the input reaches a pre-determined level. If the same valve were used as the diode detector, no output would be obtained for weak signals below the pre-determined level. Thus, the detector is a *separate* unbiased diode.

61. A circuit providing this facility is shown in Fig. 37. The final i.f. tuned circuit is connected directly to the detector, the a.f. output being tapped off the load resistor R_1 and passed to the a.f. stage via C_2.

The signal is also fed to the a.g.c. diode via C_3, which has a value of the order of 100 pF. The a.g.c. diode is connected as a shunt detector (see Chap. 1, Para. 27 of this Section) and the control voltage is developed across the load resistor R_2.

The delay facility is obtained by biasing the a.g.c. diode via the resistor R_2. This valve therefore does not conduct until the peak signal voltage from the i.f. amplifier exceeds the negative bias voltage on the anode. As soon as the signal is sufficiently strong to
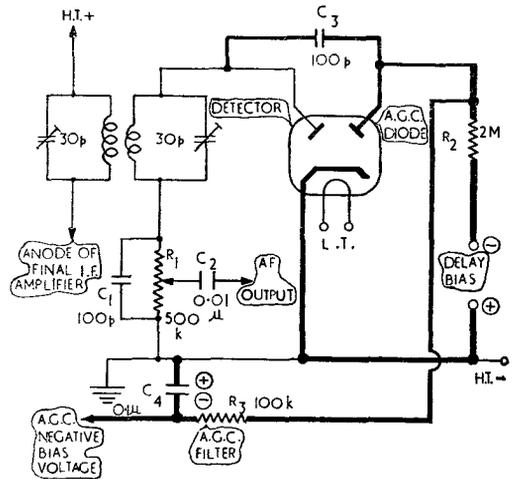
Fig. 37. DELAYED A.G.C. CIRCUIT.

overcome the delay bias, the diode conducts and a voltage is developed across R_2. A connection is taken from the 'negative end' of R_2 to the a.g.c. line via the filter R_2 C_4. Across C_4 a steady d.c. voltage proportional to the strength of the received carrier is then developed. That side of C_4 which is negative with respect to earth supplies the bias for the variable-mu valves.

Double-diode Triode

62. In most modern superheterodyne receivers, the detector, the a.g.c. diode and the first a.f. amplifier valves are combined in a
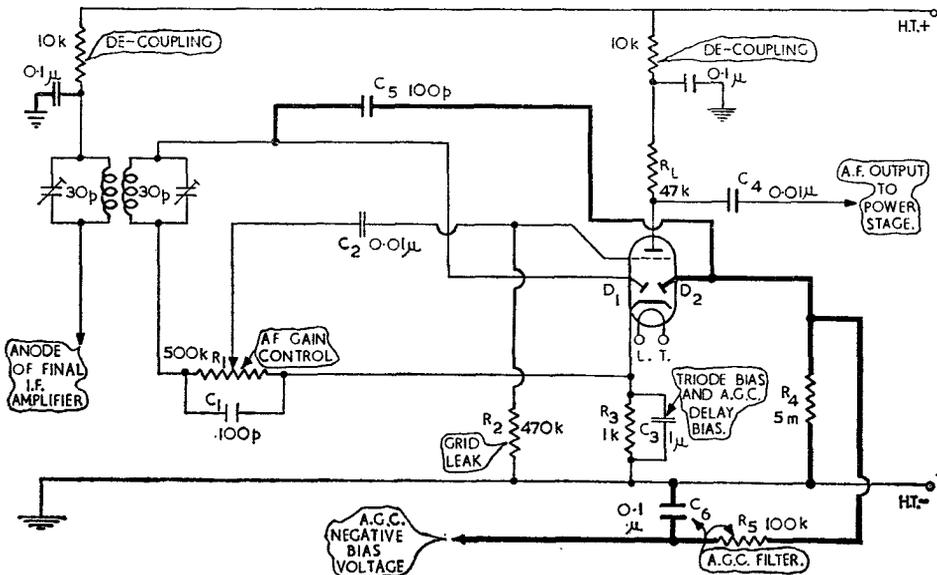
Fig. 38. CIRCUIT USING DOUBLE-DIODE-TRIODE VALVE.

double-diode-triode. This results in a considerable saving in space. A typical circuit arrangement is shown in Fig. 38.

63. The final i.f. tuned circuit is connected directly to the detector D_1, the a.f. output being tapped off the load resistor R_1 and passed to the grid of the first a.f. amplifier (the triode portion of the valve) via the d.c. blocking capacitor C_2 and the grid leak R_2.

The triode portion operates as a conventional Class A amplifier, cathode bias being provided by R_3 C_3. The resultant amplified a.f. voltage developed across the anode load R_L is coupled via C_4 to the output stage.

The bias developed across R_3 C_3 by the triode current also provides the delay bias for the a.g.c. diode D_2. This is clear from the fact that the anode of D_2 is returned to earth via R_4 while its cathode is positive to earth by the bias voltage developed across R_3. D_2 anode is thus negative with respect to its cathode.

The diode detector D_1 on the other hand is returned to the cathode, not to earth, otherwise there would be a 'delay' on the detector also.

The signal from the final i.f. tuned circuit is fed to the diode D_2 via the coupling capacitor C_5 (of the order of 100 pF), and when the signal is sufficiently strong to overcome the delay bias, D_2 conducts and develops a voltage across R_4. That end of R_4 which is negative with respect to earth is connected to the a.g.c. line via the filter R_5 C_6.

64. The graph given in Fig. 39 compares the gain characteristic of receivers incorpor-

ating no a.g.c., simple a.g.c. and delayed a.g.c. The X-axis represents the input to the receiver and the Y-axis represents the voltage available at the detector.
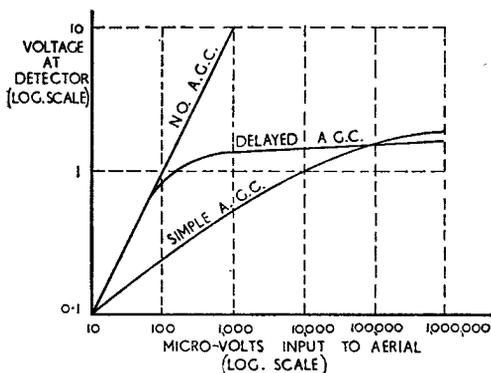
With no a.g.c. it is seen that the output rises with the input and on strong signals the receiver will be overloaded. With simple a.g.c., the control bias is available at once and the receiver gain on weak signals is considerably reduced. With delayed a.g.c. the receiver operates with full gain on weak signals (to about 100 μV input in this case) and then when a.g.c. becomes operative an almost level characteristic is obtained.

Amplified A.G.C.

65. In some cases the voltage produced by a delayed a.g.c. circuit may be insufficient to prevent overloading on strong signals. A partial improvement can be obtained by feeding the a.g.c. circuit from the *anode* of the final i.f. amplifier; this voltage is greater than that developed across the secondary and the a.g.c. voltage is thus greater for the same signal input. For the same reason the final i.f. amplifier is often allowed to operate at full gain, no control bias or only a reduced bias being applied to this stage. In other receivers, a separate i.f. amplifier is used to feed the a.g.c. circuit; again, this amplifier is not controlled by the a.g.c. bias so that the full increase of signal is allowed to develop the control voltage.

However, with any of these systems, the action of a.g.c. automatically cuts down the source of the control voltage itself, and it is impossible to obtain a truly level characteristic by these methods. Various forms of 'amplified' a.g.c. circuits have therefore been devised in which the a.g.c. voltage is applied to a d.c. amplifier which develops an adequate control bias.

66. With delayed and amplified a.g.c. systems the receiver is operating at maximum sensitivity on weak signals since no control bias is then produced. Thus, during alterations to the tuning of high gain receivers provided with this form of a.g.c. and during 'no-signal' periods, the background noise rises to a high level. This is sometimes reduced by arranging to '*mute*' the receiver until a signal is received; in effect, muting consists of disconnecting the telephones from the receiver during no-signal conditions.

Fig. 39. RECEIVER GAIN CHARACTERISTICS WITH AND WITHOUT A.G.C.

Applying the A.G.C. Voltage

67. Methods of applying the control voltage to the variable-mu valves are illustrated in Fig. 40.

Fig. 40(a) represents a portion of an i.f. amplifier. The a.g.c. voltage is applied via a de-coupling resistor R_1 to the 'earthy' end of the transformer secondary and thence to the grid of the variable-mu valve. A 'standing' or minimum value of bias must be provided

In all these circuits R_1 C_1 represent the a.g.c. filter to each valve. In every case, C_1 charges to the d.c. potential of the control bias on a time constant of R_1 C_1. This must be such that the charge on C_1 can follow the variations in bias due to fading but is sufficiently long to prevent the charge on C_1 following the variations due to modulation in the signal. C_1 has a typical value of $0 \cdot 1$ μF and R_1 between 100 kΩ and 1 MΩ.
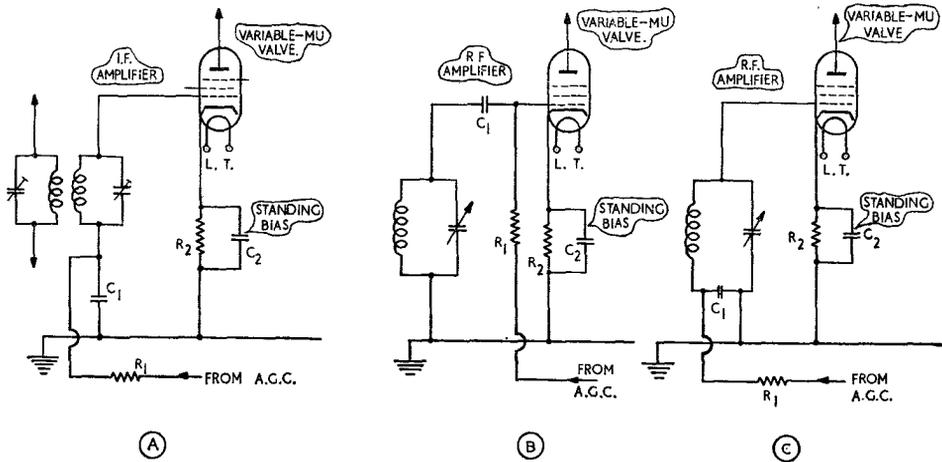
Fig. 40. METHODS OF APPLYING THE A.G.C. BIAS VOLTAGE.

until an a.g.c. voltage is developed; this is provided by the cathode bias resistor R_2 de-coupled by C_2 and its value is such as to give the required maximum gain without distortion.

In tuned r.f. stages, the variable capacitor in the tuned circuit has normally one set of plates earthed. Thus the circuit of Fig. 40(a) cannot be used. Either the circuit of Fig. 40(b) or the circuit of Fig. 40(c) may be used.

Fig. 40(b) illustrates shunt feed. It is not a popular method of applying the a.g.c. voltage to the valve because R_1 is virtually in parallel with the tuned circuit and introduces damping losses.

Fig. 40(c) shows series feed. The a.g.c. voltage is applied to the 'earthy' end of the tuning inductor, and the tuned circuit is completed via C_1. One set of plates in the tuning capacitor can now be earthed. If C_1 has too small a value its presence in the tuned circuit will upset tracking as it effectively reduces the value of the variable capacitor. A typical value for C_1 is $0 \cdot 1$ μF.

Tuning Indicators

68. As well as compensating for fading, a.g.c. also tends to compensate for any drop in output caused by detuning the receiver. Since the output does not alter appreciably as the station is tuned in, it is often difficult to adjust the tuning accurately. To avoid this, *tuning indicators* are often fitted on receivers employing a.g.c. Since the control bias will be a maximum when the set is accurately tuned to a station, a simple meter could be used to indicate either the maximum a.g.c. voltage or the minimum current in the controlled valves. Another method is to use a 'magic eye' tuning indicator.

69. **Magic eye.** The magic eye tuning indicator is a multi-unit valve consisting of a triode amplifier section and a visual indicator section. The symbol for such a valve and its general construction are shown in Fig. 41.

The triode section is conventional. The cathode is extended into the indicator section
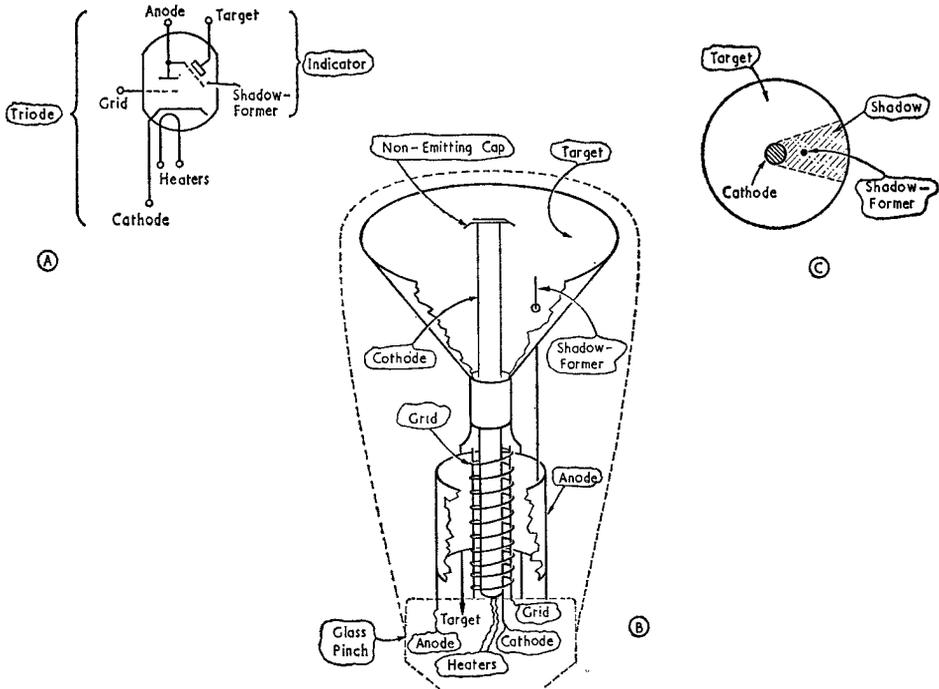
Fig. 41. 'MAGIC-EYE' TUNING INDICATOR.

which consists of a conical 'target' with a fluorescent coating and a thin deflector rod or 'shadow-former' which protrudes through a hole in the target and is connected to the triode anode. The valve is positioned in the receiver in such a way that only the cone is visible as shown in Fig. 41(c).

70. In the indicator section, electrons are emitted equally in all directions by the cathode, and this causes uniform illumination on the fluorescent coating of the target. If the potential of the shadow-former is the same as that of the electric field in the target-cathode space at the position in which the shadow former is placed, then it has negligible effect and no shadow is formed.

If, however, the potential of the shadow-former is reduced below that of the target-cathode electric field at that point, the portion of the fluorescent coating behind the shadow-former no longer receives electrons and a 'shadow' is formed. In practice, the angle of the shadow can be varied from 0° to about 100° by variation of the shadow-former potential.

If the indicator is adjusted to give maxi-mum shadow for no signal being received, then as soon as a signal is received the angular width of the shadow is reduced. The smallest shadow indicates maximum signal and accurate tuning.

71. A typical circuit arrangement for a magic eye tuning indicator is shown in Fig. 42. The target anode is connected directly to the
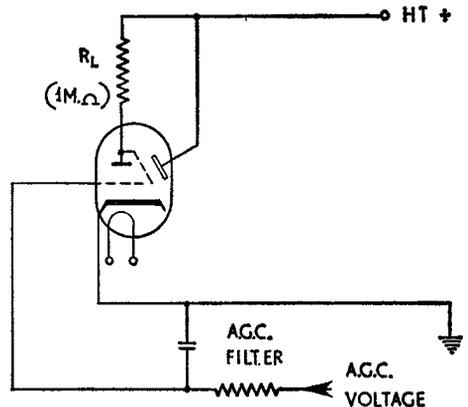
Fig. 42. CIRCUIT FOR 'MAGIC-EYE' TUNING INDICATOR.

h.t. positive line. The shadow former and triode anode are connected to h.t. positive via the anode load R_L (typically about 1 MΩ). The input applied between grid and cathode of the triode is a d.c. voltage derived from the signal and is usually the a.g.c. bias voltage.

72. In the absence of a signal, no a.g.c. voltage is developed. Thus, the grid and cathode potentials are the same, the current through the valve is high and there is a large voltage drop across R_L. The potential of the shadow-former is then less than that of the target-cathode electric field at the position of the shadow-former and a shadow is formed.

As a signal is tuned in, the a.g.c. negative bias increases and the triode current falls. The decrease in voltage drop across R_L causes the potential of the shadow-former to rise, and as it approaches the potential of the electric field at that point, the shadow narrows. The receiver is correctly tuned when the shadow is a minimum.

Complete Superheterodyne Receiver

73. A block diagram of a typical communication superheterodyne receiver is shown in Fig. 43. This includes all the stages discussed in the preceding paragraphs.

A complete circuit diagram is shown in Fig. 44 (pull-out leaf); waveband switching has been omitted for the sake of clarity.

signal grid of the frequency changer V_2. Standing bias is provided by $R_4 C_5$ and a reduced level of a.g.c. bias is applied through R_1 (de-coupled by C_1) and L_2 to the grid of V_1. The reduced a.g.c. is necessary because of the short grid base of the CV1091 used in this stage. The anode is de-coupled by $R_5 C_6$ and the screen by the network $R_2 R_3 C_4$; this potential divider tends to keep the screen voltage constant for any variation in the a.g.c. voltage.

(b) V_2, **Frequency Changer.** The triode portion of this valve operates as a shunt-fed tuned anode oscillator generating a voltage that is above the signal frequency by the i.f. It is biased to Class C by $R_{11} C_{13}$. The anode load is R_{12}, and $R_{13} C_{18}$ provide de-coupling.

The standing bias for the hexode portion is provided by R_9 (de-coupled by C_{11}) and full a.g.c. bias is also applied through R_6 (de-coupled by C_7) and L_4 to the signal grid of V_2. Normal multiplicative mixing takes place in the hexode and the i.f. component is selected by I.F.T$_1$. Anode de-coupling is by $R_{10} C_{12}$ and the screen is supplied from the potential divider $R_7 R_8$ (de-coupled by C_{10}).

(c) V_3, **First I.F. Amplifier.** This is a conventional i.f. amplifier. The standing Class A bias is provided by R_{18} (de-coupled by C_{21}) and full a.g.c. bias is applied to the grid through R_{14} (de-coupled by C_{19}) and

Fig. 43. BLOCK SCHEMATIC DIAGRAM OF COMMUNICATIONS SUPERHETERODYNE RECEIVER.

74. The stages in the circuit of Fig. 44 are as follows:—

(a) V_1, **R.F. Amplifier.** This is a conventional r.f. voltage amplifier circuit, the required signal being selected in the grid tuned circuit, amplified and applied to the

the secondary of I.F.T$_1$. The i.f. signal developed across I.F.T$_1$ is amplified by V_3 and appears across I.F.T$_2$. The anode de-coupling is $R_{19} C_{22}$ and the screen is supplied from $R_{16} R_{17}$ (de-coupled by C_{20}).

(d) **V$_4$, Second I.F. Amplifier.** This stage is similar to V$_3$. The standing bias is from R$_{23}$ (de-coupled by C$_{25}$) but a reduced a.g.c. bias is applied to V$_4$ grid through R$_{20}$ (de-coupled by C$_{23}$) and the secondary of I.F.T$_2$. This reduced bias ensures a higher gain at this stage so that the input to the detector and the source voltage for the a.g.c. diode are adequate; to further assist in this, the a.g.c. diode is fed from the *anode* of V$_4$ via C$_{36}$ instead of from the secondary of I.F.T$_3$. The anode is de-coupled by R$_{24}$ C$_{26}$ and the screen by R$_{21}$ R$_{22}$ C$_{24}$. The i.f. signal is further amplified in this stage and developed across I.F.T$_3$. The selectivity of the receiver is determined by the bandwidth of the circuits I.F.T$_1$, I.F.T$_2$ and I.F.T$_3$.

(e) **V$_5$, Detector, A.G.C. and A.F. amplifier.** This valve is a double-diode triode. One diode (the left-hand diode) provides detection, the other diode is for a.g.c. and the triode operates as an a.f. voltage amplifier.

The secondary of I.F.T$_3$ is connected to the detector. On receipt of an i.f. signal, the diode conducts on the positive half-cycles and develops an a.f. voltage across the detector load R$_{26}$ C$_{29}$. The i.f. component is by-passed to earth by the filter C$_{27}$ R$_{25}$ C$_{28}$. The d.c. component of detection is blocked by C$_{30}$ and the a.f. component is tapped off R$_{26}$ (acting as an a.f. gain control) and applied via C$_{30}$ to the grid of V$_5$. R$_{27}$ is the grid leak. The detector circuit is returned to the cathode of V$_5$ and not to earth, otherwise the bias voltage developed across R$_{30}$ would give a 'delay' in detection.

The a.f. component is amplified through the triode portion of V$_5$ which is cathode biased by R$_{30}$ (de-coupled by C$_{33}$) to Class A conditions. An amplified a.f. voltage is developed across the load R$_{31}$ and coupled by C$_{35}$ to the grid of V$_7$. R$_{32}$ C$_{34}$ provide anode de-coupling.

The a.g.c. diode is fed from the anode of V$_4$ through C$_{36}$. Since the diode anode is returned to *earth* through the load R$_{33}$ R$_{34}$, the cathode bias provided by R$_{30}$ also provides a delay bias for the a.g.c. diode. When the signal exceeds this delay, the diode conducts on the positive half-cycles and develops a control voltage across R$_{33}$ R$_{34}$ in series. The full available negative a.g.c. bias is applied via the low-pass filter R$_{29}$ C$_{32}$ to the grids of V$_2$, V$_3$ and V$_8$; the reduced negative a.g.c. bias (developed across R$_{34}$ only) is applied via the low-pass filter R$_{28}$ C$_{31}$ to the grids of V$_1$ and V$_4$.

(f) **V$_6$, B.F.O.** This is switched ON only for the reception of c.w. signals. In the OFF position (for r.t. reception) the h.t. positive line to V$_6$ anode is broken at the switch. When receiving c.w. signals, the b.f.o. operates as a conventional shunt-fed Colpitts oscillator working under Class C self-bias (R$_{35}$ C$_{41}$) to produce a voltage across the anode load R$_{36}$. The frequency of this voltage differs from the i.f. by an a.f., and the b.f.o. output is coupled via C$_{37}$ to the diode detector where it combines with the i.f. to give a heterodyne beat note. The pitch of this beat note can be altered either by mistuning the receiver so that the i.f. produced at the frequency changer is incorrect or more correctly by adjusting the trimmer C$_{40}$ in the b.f.o. Anode de-coupling is provided by R$_{37}$ C$_{42}$.

(g) **V$_7$, Output.** This is an a.f. power pentode and is operated with its screen at h.t. positive potential. The stage operates under Class A cathode bias (R$_{39}$ C$_{43}$) and the output from V$_5$ is coupled via C$_{35}$ to the grid leak R$_{38}$ for application to V$_7$ grid-cathode. The resultant a.f. current through the primary of T$_1$ controls the amount of power drawn from the supply. The turns ratio of T$_1$ ensures correct matching to the telephones for optimum power output. No de-coupling is necessary at the output stage.

(h) **V$_8$, Tuning Indicator.** The operation of this valve is explained in Para. 66. Its grid is connected to the full a.g.c. line so that when the receiver is correctly tuned to a signal, the a.g.c. bias developed by that signal is a maximum and the current through the triode portion of the magic eye is a minimum. The potential of the shadow-former then approaches that of the target anode and the shadow angle is a minimum for that particular signal.

Note:—The power supplies for this receiver have not been shown. These differ widely. For mains operated sets, a rectifier would be incorporated: for portable sets, batteries are used; for air-borne and mobile sets, rotary machines provide the power supplies. All these methods of power supply are dealt with in Book 2, Sect. 9.

Transistorized Superheterodyne Receivers

75. Superheterodyne receivers incorporating transistors instead of valves are becoming increasingly popular, especially in circumstances where economies in power supply are important and miniaturization is a requirement. The transistorized t.r.f. type receiver is dealt with in Chap. 1, Para. 49 of this Section. A circuit of a transistorized superheterodyne receiver is shown in Fig. 45 (pull-out leaf).

This receiver is suitable only for reception in the long-wave and medium-wave bands: band-switching has been omitted for reasons of clarity. Since a b.f.o. is not included in the circuit, no reception of c.w. signals is possible.

76. The circuit consists of the following stages: r.f. amplifier, frequency changer, two i.f. amplifiers, detector and a.g.c. diode, a.f. amplifier and output stage. All transistor stages in the receiver derive their bias from a d.c. stabilization circuit, details of which are given in Chap. 1, Para. 50(a) of this Section. The receiver power supplies consist of a 6 volts battery. Details of the circuit are given below:—

(a) **TR$_1$, R.F. Amplifier.** This is a conventional tuned voltage amplifier biased by the d.c. stabilization circuit, R$_1$ R$_2$ R$_4$ and the associated de-coupling capacitors C$_3$ C$_7$. A.g.c. bias is also applied to the base via R$_2$ and L$_2$.

The signal is selected in the input tuned circuit and matched into the base of TR$_1$ by transformer action. The small signal current at the base causes a large variation in collector current and an amplified signal voltage is developed across the high impedance load in the collector and matched into the base of the frequency changer TR$_2$. The impedance match between TR$_1$ and its collector tuned circuit is by an auto-transformer tap on L$_3$. All other collector tuned circuits in the receiver are similarly matched.

The input tuned circuit and the collector tuned circuit are tuned to the required signal frequency by two sections of the three-gang capacitor; trimmers are inserted for tracking. The collector is de-coupled by R$_3$ C$_6$.

(b) **TR$_2$, Frequency Changer.** This stage is arranged as a self-oscillating additive type frequency changer.

The oscillator is a Meissner type where energy is fed back from the collector (L$_6$) to the emitter (L$_5$) via the tuned circuit inductance L$_7$, to maintain oscillations. The frequency of the oscillations is controlled by the tuned circuit associated with L$_7$. This circuit is tuned by one section of the three-gang capacitor. A padder (C$_{12}$) and a trimmer (C$_{13}$) are included to give three-point tracking.

The signal from TR$_1$ is coupled through C$_8$ to the base of TR$_2$ where it mixes with the local oscillation to produce a component at the i.f. in the collector current. The i.f. component is selected by I.F.T$_1$ and passed for amplification to TR$_3$.

Bias is obtained from the d.c. stabilization circuit R$_5$ R$_6$ R$_7$ and the associated de-coupling capacitors C$_9$ C$_{10}$. The collector is de-coupled by R$_8$ C$_{11}$.

(c) **TR$_3$ and TR$_4$, I.F. Amplifiers.** Both stages are conventional. TR$_3$ is biased by the d.c. stabilization circuit R$_9$ R$_{10}$ R$_{11}$ and the associated de-coupling capacitors C$_{15}$ C$_{16}$; a.g.c. bias is also applied to TR$_3$ base via R$_{10}$ and the secondary of I.F.T$_1$.

TR$_4$ is biased by the d.c. stabilization circuit R$_{12}$ R$_{13}$ R$_{14}$ and the associated de-coupling capacitors C$_{17}$ C$_{18}$; no a.g.c. bias is applied to TR$_4$ and this stage operates at maximum gain to provide a large output into D$_1$.

The i.f. input from the frequency changer is amplified by TR$_3$ and TR$_4$ in cascade and applied to the detector D$_1$.

A capacitor network C$_{19}$ C$_{20}$ is connected between the secondary coils of I.F.T$_1$, I.F.T$_2$ and I.F.T$_3$. These are neutralizing components and provide negative feedback between the output and input of each i.f. stage to prevent undesirable oscillations in the i.f. amplifier. This is necessary because of the fairly high capacitances existing between the collector-base and base-emitter electrodes in a transistor.

(d) **D$_1$, Detector and A.G.C.** This consists of a germanium diode circuit arranged in a manner similar to that of a series-fed diode detector circuit. The circuit gives simple a.g.c. as well as providing detection.

The i.f. output developed across the secondary of I.F.T$_3$ is applied to D$_1$ and the diode conducts on the positive half-cycles of input. The i.f. ripple component of detection is by-passed to earth through

Valve block labels:

- V_1 R.F. AMPLIFIER
- V_2 FREQUENCY CHANGER
- V_3 AND V_4 I.F. AMPLIFIERS
- V_5 DETECTOR A.G.C. AND A.F. AMPLIFIER
- V_6 B.F.O.
- V_7 POWER OUTPUT

AERIAL

I.F.T. 1   I.F.T. 2   I.F.T. 3

B.F.O. ON OFF

PHONES

H.T. +

H.T. –

FULL A.G.C. LINE   FULL A.G.C. LINE

REDUCED A.G.C. LINE   REDUCED A.G.C. LINE

L.T.
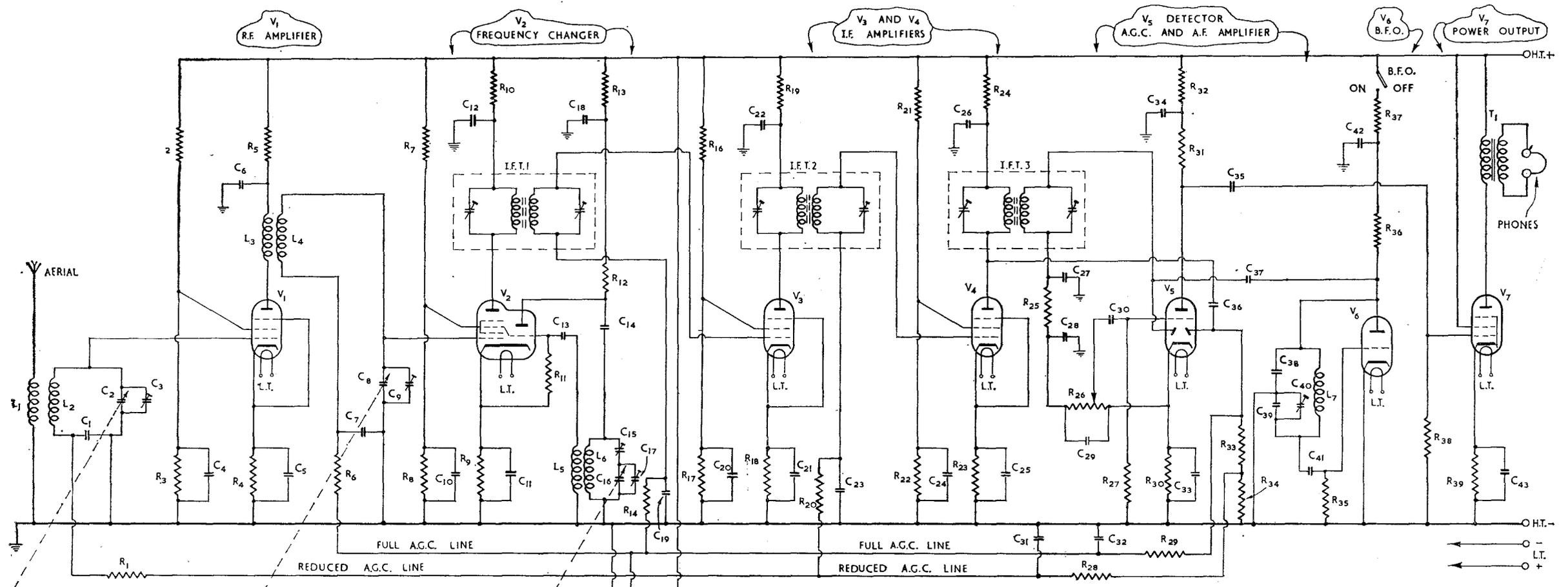
VALVE TYPES

V_1 — R.F. AMPLIFIER — CV 1091
V_2 — FREQUENCY CHANGER — CV 1347
V_3 and V_4 — I.F. AMPLIFIERS — CV 1053
V_5 — DETECTOR A.G.C. AND A.F. AMPLIFIER — CV 587
V_6 — B.F.O. — CV 583
V_7 — OUTPUT — CV 1911
V_8 — TUNING INDICATOR — CV 1103

CAPACITOR VALUES

C_1 = 0·01 μ	C_{16} = VARIABLE 150 p	C_{31} = 0·1 μ
C_2 = VARIABLE 150 p	C_{17} = TRIMMER 30 p	C_{32} = 0·1 μ
C_3 = TRIMMER 30 p	C_{18} = 0·01 μ	C_{33} = 1 μ
C_4 = 0·01 μ	C_{19} = 0·02 μ	C_{34} = 1 μ
C_5 = 0·01 μ	C_{20} = 0·1 μ	C_{35} = 0·01 μ
C_6 = 0·01 μ	C_{21} = 0·1 μ	C_{36} = 100 p
C_7 = 0·01 μ	C_{22} = 0·1 μ	C_{37} = 50 p
C_8 = VARIABLE 150 p	C_{23} = 0·02 μ	C_{38} = 0·005 μ
C_9 = TRIMMER 30 p	C_{24} = 0·1 μ	C_{39} = 100 p
C_{10} = 0·01 μ	C_{25} = 0·1 μ	C_{40} = TRIMMER 30 p
C_{11} = 0·01 μ	C_{26} = 0·1 μ	C_{41} = 100 p
C_{12} = 0·1 μ	C_{27} = 100 p	C_{42} = 0·1 μ
C_{13} = 100 p	C_{28} = 100 p	C_{43} = 20 μ
C_{14} = 100 p	C_{29} = 100 p	
C_{15} = PADDER 500 p	C_{30} = 0·01 μ	

RESISTOR VALUES

R_1 = 47 k	R_{16} = 27 k	R_{31} = 47 k
R_2 = 27 k	R_{17} = 22 k	R_{32} = 22 k
R_3 = 22 k	R_{18} = 330 Ω	R_{33} = 4·7 M
R_4 = 220 Ω	R_{19} = 10 k	R_{34} = 2·2 M
R_5 = 10 k	R_{20} = 47 k	R_{35} = 100 k
R_6 = 47 k	R_{21} = 27 k	R_{36} = 22 k
R_7 = 27 k	R_{22} = 22 k	R_{37} = 47 k
R_8 = 22 k	R_{23} = 330 Ω	R_{38} = 470 k
R_9 = 330 Ω	R_{24} = 10 k	R_{39} = 330 Ω
R_{10} = 10 k	R_{25} = 150 k	
R_{11} = 470 k	R_{26} = 500 k	
R_{12} = 33 k	R_{27} = 470 k	
R_{13} = 22 k	R_{28} = 100 k	
R_{14} = 47 k	R_{29} = 100 k	
R_{15} = 1 M	R_{30} = 1 k	

Fig. 44. BASIC CIRCUIT OF SUPERHETERODYNE RECEIVER.

C_{23}, the a.f. and d.c. components appear across R_{15} (acting as an a.f. gain control) and the d.c. component is blocked by the coupling capacitor C_{24}. The a.f. output is thus applied through C_{24} to the base of TR_5.

R_{15} also acts as the a.g.c. load and the 'positive end' of R_{16} is connected to the a.g.c. line via the filter R_{15} C_{22}. Note that for p-n-p type transistors, the potential applied by the a.g.c. line to the base is a *positive* bias. The stronger the input signal the more positive this becomes and the greater is the barrier to the flow of holes from emitter to base; the amplification falls accordingly. A.g.c. bias in this receiver is applied to TR_1 and TR_3.

(*e*) **TR_5, A.F. Amplifier.** The detected a.f. output developed across R_{16} is capacitively coupled through C_{24} to the base of TR_5. The high value for C_{24} ($4\mu F$) is necessary because of the low input impedance of the transistor. Bias for TR_5 is obtained from the d.c. stabilization circuit R_{17} R_{18} R_{19} R_{20} and the associated de-coupling capacitors C_{25} C_{26}.

A negative feedback voltage, derived from the secondary of the output transformer via R_{23} C_{27} is developed across R_{20} in the emitter of TR_5. Negative feedback reduces any distortion present and the values of R_{23} C_{27} are such that feedback *increases* with frequency thereby compensating for the rise in the gain-frequency response characteristic.

The a.f. base current of TR_5 is amplified and the resultant amplified a.f. current through T_1 primary develops the required a.f. voltage across the secondary for driving the push-pull stage TR_6-TR_7. The correct impedance match is obtained by adjusting the turns ratio of T_1.

(*f*) **TR_6-TR_7, Power Output.** The two transistors in the output stage have their bases biased by R_{21} R_{22} across the battery supply. Bias is applied to the emitters by the voltage developed across R_{24}. The bias resistors in this stage are not decoupled and the resultant negative feedback reduces the distortion and improves the performance of the stage.

For efficient operation of the push-pull stage the two transistors should be 'matched'; for this reason they are normally supplied in pairs.

The a.f. drive from TR_5 is applied to the bases of TR_6 and TR_7 from the centre-tapped secondary of T_1. Conventional push-pull action results in a considerable amount of d.c. power from the supply being converted to give the required a.f. power output to the telephones or loud-speaker. The turns ratio of T_2 is adjusted to give the required impedance match.

Summary

77. This Section has dealt with receiver principles in a general way. Chapter 1 considered the t.r.f. type receiver and Chapter 2 discussed the superheterodyne receiver. Techniques and problems peculiar to either communication receivers or radar receivers will be considered in Parts 2 and 3 of these Notes.

However, having talked about transmitters in Section 13 and receivers in this Section, it is logical to carry on now in order to see:—

(*a*) how the transmitters and receivers are connected to their aerials; Section 15 (Transmission Lines) considers this:

(*b*) the basic theory behind the operation of aerials; this is dealt with in Section 16 (Aerials):

(*c*) the various ways in which the energy radiated by the transmitter aerial reaches the receiver aerial; this is discussed in Section 17 (Propagation).
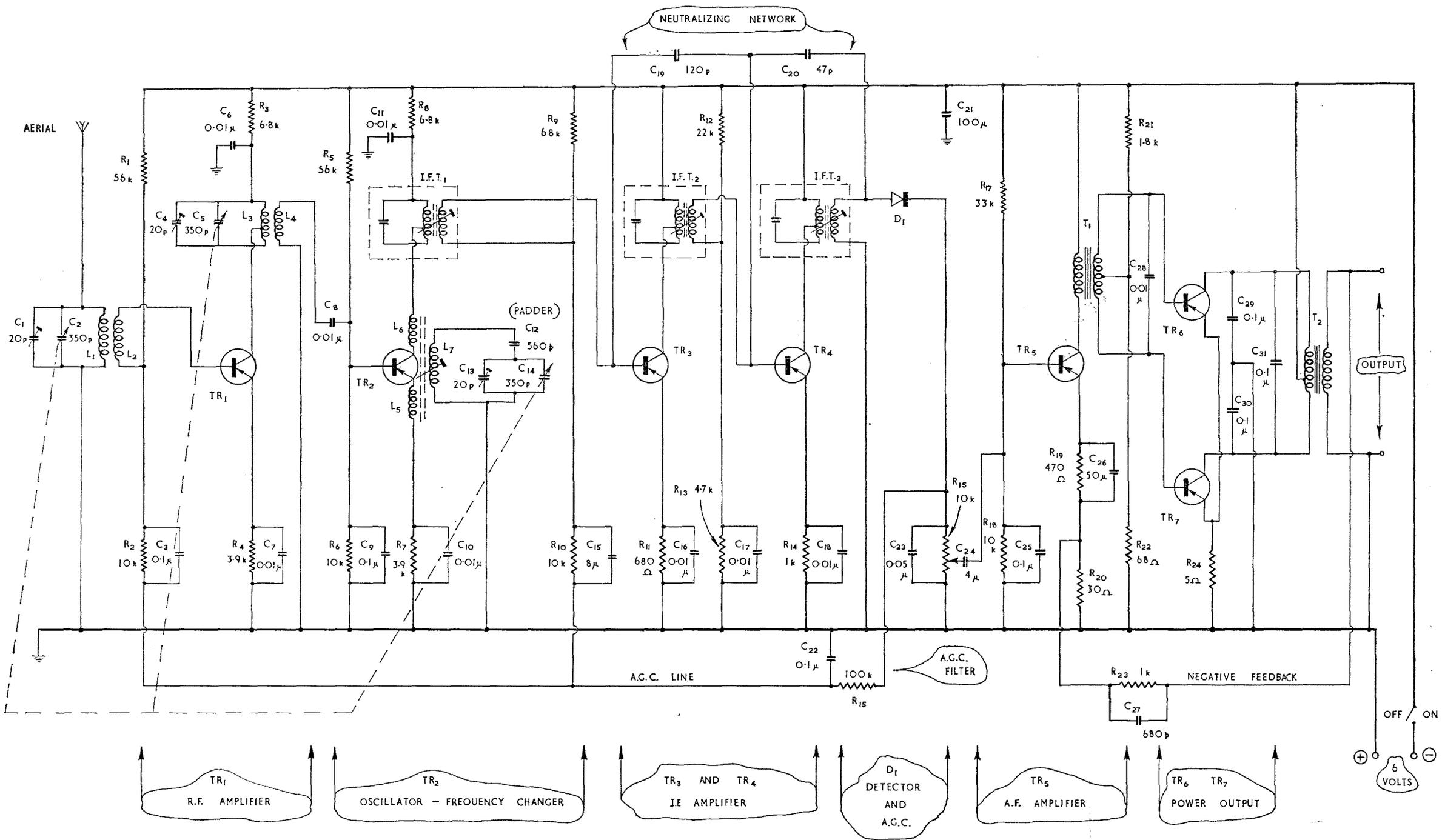
Fig. 45. TRANSISTORIZED SUPERHETERODYNE RECEIVER.

SECTION 15

FILTERS AND TRANSMISSION LINES

SECTION 15

FILTERS AND TRANSMISSION LINES

SECTION 15

CHAPTER 1

SIMPLE FILTER CIRCUITS

SIMPLE FILTER CIRCUITS

Introduction

1. The word 'filter' has been in everyday use for many years; in whatever context it is used, it describes some device which has the property of discrimination, i.e., it will allow virtually unobstructed passage to one thing whilst restricting or completely stopping the flow of others.

PARTICLES IN CONTAMINATED OIL
CANNOT PASS THROUGH FILTER

CONTAMINATED OIL   CLEAN OIL

WIRE MESH FILTER
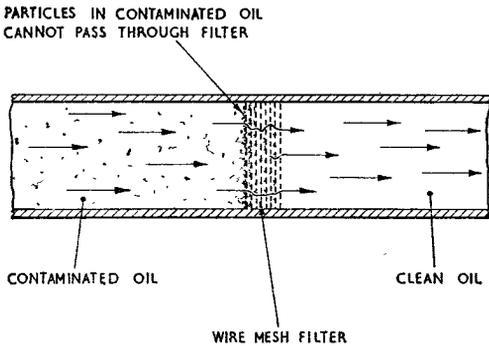
Fig. I. OIL FILTER.

A familiar example is the oil filter in an internal combustion engine. Oil which circulates through the engine becomes contaminated by particles caused by combustion. When the contaminated engine oil is passed through the filter, as shown in Fig. 1, the particles are held by the filter, while the oil itself is allowed almost unimpeded progress. The filter thus discriminates against the particles suspended in the engine oil.

2. Many other types of filter exist besides that just mentioned, and filters can be used in conjunction with almost every physical quantity. The atmosphere for example, possesses the property of discriminating *against* light at the red, or low frequency end of the visible spectrum; this explains why, when the sky is unobscured by clouds, most of the sunlight reaching the earth is at the upper, or blue end of the visible spectrum, and the sky appears blue.

3. Generally, the purpose of an electrical filter is to discriminate between alternating currents of different frequencies, and severely to attenuate one component whilst offering almost zero impedance to others. The complexity of the filter depends on the difference in frequency of the two currents; a small difference in frequency requires a complex filter to achieve reasonable discrimination, whilst discrimination between widely differing frequencies can be achieved with a relatively simple filter.

Simple RC Filter

4. A familiar example of an electrical filter is the RC network used to couple two amplifier stages; such a network is shown in Fig. 2(a). The input to the filter consists of a
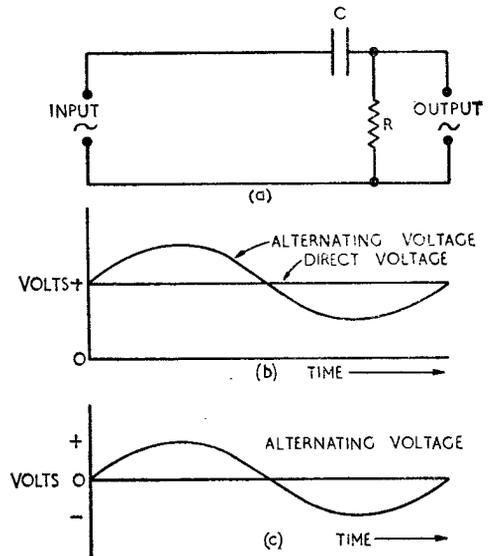
Fig. 2. SIMPLE R.C. FILTER.

direct voltage upon which is superimposed an alternating voltage as shown in Fig. 2(b). The direct voltage has a frequency of zero and the capacitive reactance to this voltage is infinite; thus the direct voltage is completely blocked by the capacitor. The capacitive reactance to the alternating voltage is relatively small compared with the value of R, and most of the alternating voltage is thus developed across R and appears at the output terminals, as shown in Fig. 2(c).

5. Since the reactance of a capacitor is inversely proportional to frequency, the circuit shown in Fig. 2, can also be used to discriminate between two alternating voltages. Fig. 3 shows a graph of output voltage
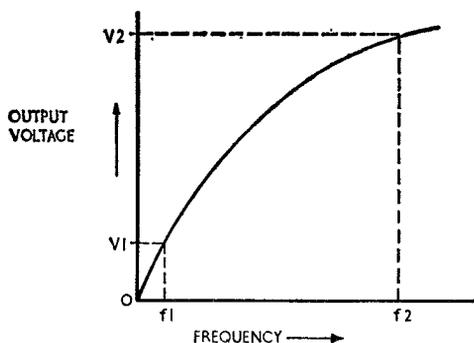
Fig. 3. SIMPLE FILTER CHARACTERISTIC.

plotted against frequency for the circuit of Fig. 2, and also shows output voltages V_1 and V_2 which are derived from input voltages at frequencies f_1 and f_2 respectively. Since the two frequencies are widely separated, V_1 is much less than V_2 due to the discriminating action of the CR network; that is, the reactance of C is much greater and the output much less at frequency f_1 than at frequency f_2.

This type of filter functions reasonably well when audio and radio frequency voltages are both present in a circuit, and only radio frequency voltages are required. For example, in the simplified modulation circuit shown in Fig. 4, the anode current of V_1 has both a.f. and r.f. components, and
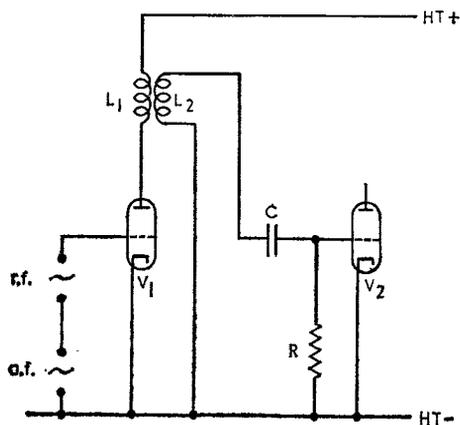
Fig. 4.   CIRCUIT USING SIMPLE R.C. FILTER.

some a.f. voltage is developed across L_2, the secondary of the r.f. transformer. To prevent a.f. voltages being applied to the control grid of V_2, a filter consisting of C and R is included. Since the reactance of C is inversely proportional to frequency, most of the a.f. voltage is dropped across C, while most of the r.f. voltage is developed across R, and is thus applied between grid and cathode of V_2.

6. The RC filter shown in Fig. 4 discriminates against low frequencies, but a rearrangement of the circuit will reverse its characteristics. Fig. 5(a) shows a circuit in which the component in series with the
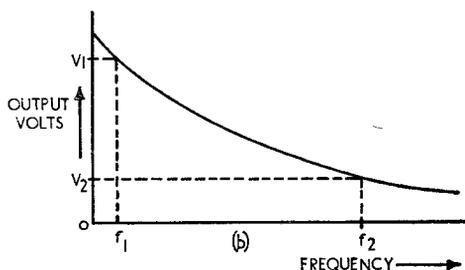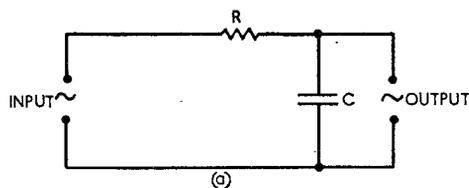
Fig. 5.   SIMPLE R.C. FILTER (C IN PARALLEL).

output is R, and the component in parallel with the output is C. The output voltage depends on the relationship between the resistance of R and the reactance of C; as the reactance of C is infinite when the input frequency is zero and is still very high at audio frequencies, the greater part of audio frequency input voltages are developed across the capacitor and thus appear at the output terminals. Input voltages at r.f. develop little voltage across C and the r.f. output voltage is thus very small. Fig. 5(b) shows a graph of output voltage plotted against frequency, and it can be seen that for two widely separated frequencies f_1 and f_2, the lower frequency f_1 causes an output voltage much greater than that caused by the higher frequency f_2.

7.   A typical example of the use of this type of filter is shown in Fig. 6(a) which is a circuit of a diode detector.   In the process of detection a considerable amount of unwanted r.f. voltage at the signal frequency
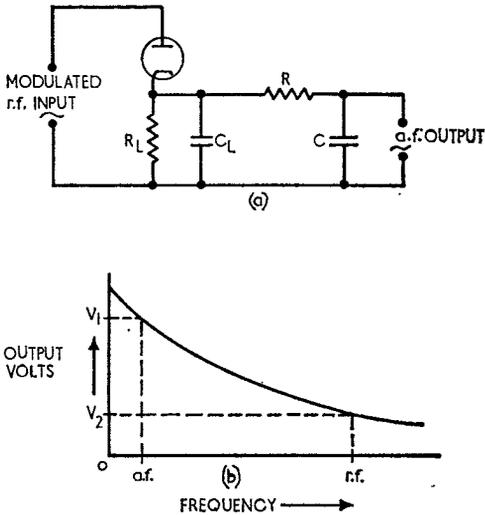
Fig. 6.   CIRCUIT USING SIMPLE R.C. FILTER (C IN PARALLEL).

is developed across the diode load components $C_L R_L$, and if this were applied to the subsequent a.f. amplifier stage, unwanted feedback could result.   However, as shown in Fig. 6(b), the a.f. and r.f. voltages are widely separated, and the relationship between the resistance of R and the reactance of C is such that the voltage V_2 which is derived from the r.f. is negligible with respect to the required voltage V_1 which is derived from the a.f.   The input to the next stage thus consists almost entirely of a.f. voltages.

A similar filter is often used in power rectifier circuits; in this case however the higher frequency component discriminated against is the a.f. ripple and the output is virtually a pure d.c. voltage.

Disadvantage of Simple RC Filter

8.   The filters dealt with so far are extensively used, but their application is strictly limited.   Their main disadvantage is shown in Fig. 7, which is a graph similar to that shown in Fig. 6.   In this case however, the input consists of two voltages at frequencies which are *not* widely separated.   The two frequencies f_1 and f_2 cause output voltages

V_1 and V_2, respectively, and because of the characteristic of the filter V_1 and V_2 are almost identical; that is, the filter cannot adequately discriminate between adjacent frequencies.   Even if the characteristic is varied by altering the capacitor value, the change in the filter discrimination properties is negligible, as can be seen from the dotted characteristic in Fig. 7.
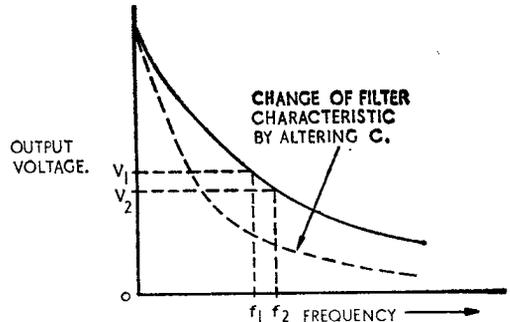
Fig. 7.   DISADVANTAGE OF SIMPLE FILTER.

Ideal Filter

9.   Before dealing with filters in greater detail it is important to know the characteristics of an *ideal* filter in order to appreciate the requirements that practical filters should satisfy.   An ideal filter possesses at least one frequency band in which attenuation is zero (no suppression) and another frequency band in which attenuation is infinite (infinite suppression).   The change from infinite to zero attenuation occurs suddenly, and the frequency at which the change takes place is called the *cut-off frequency.*

10.   Filters may be required to suppress high or low frequencies, while in other applications they may be required to admit or suppress a specific *band* of frequencies. Whatever the type of filter, there is an ideal response with which that of the practical filter can be compared.   The ideal and practical characteristics will never be exactly the same, but the efficiency of the practical filter is related to the degree of similarity between them.

11.   An example of an ideal filter characteristic is shown in Fig. 8.   This ideal filter is required to pass frequencies *above* a specific value, and to suppress those frequencies *below* that value.   The characteristic is a graph of attenuation plotted against frequency, and it can be seen that the graph

consists of a straight vertical line originating at the cut-off frequency. Voltages at frequencies to the left of the vertical line are subjected to infinite attenuation while those
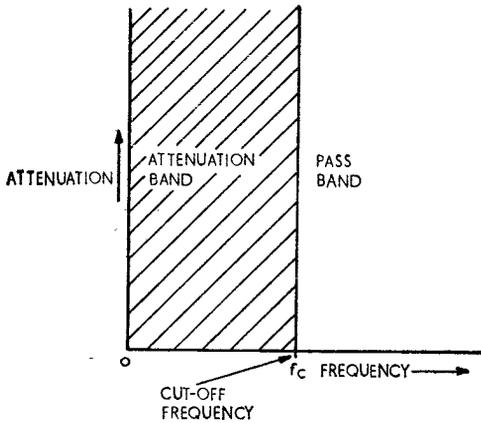
Fig. 8. CHARACTERISTIC OF IDEAL FILTER.

to the right of the vertical line are not attenuated. From zero cycles per second to the cut-off frequency f_c is termed the *attenuation band*, and from f_c to infinity is termed the *pass band*. The output from such
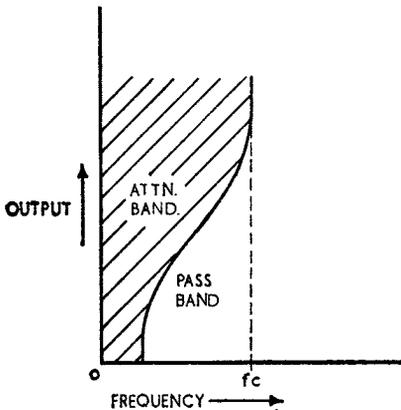
Fig. 9. PRACTICAL FILTER CHARACTERISTIC.

a filter would be zero for input voltages at frequencies within the attenuation band; this state of affairs never exists in practice, but well-designed filters approach the ideal fairly closely, as can be seen from inspection of the graph shown in Fig. 9.

Practical Filters

12. Two basic facts emerge from the discussion on the ideal filter. Firstly, a filter must be frequency sensitive to a high degree in order to ensure a sudden change in attenuation. Secondly, the attenuation for voltages which are at frequencies within the pass band must be as near to zero as possible. These facts indicate that a practical filter must be constructed from *reactive* components, since inductors and capacitors are frequency sensitive and inductors possess small resistance. Thus most filters use only inductors and capacitors, the number and position of the components depending on the required filter characteristics.

13. Practical filters may be divided into two main types, both of which however, perform the same functions. The more complex type is called a balanced filter, and is used when each arm of the filter is required to be at some fixed r.m.s. voltage with respect to reference (usually earth). The simpler and more commonly used type is called an unbalanced filter; in this type, one arm of the filter is at reference potential. Also the reactances which normally comprise the filter may be arranged in one of two ways. The first is the T network, where the input reactance is in series with the supply. The second is the π network, where the first reactance shunts the supply. The application of these types is governed by the output and input impedances of the circuits to which the filter is connected.

14. The diagrams shown in Fig. 10 are really filter sections. A filter may consist of one or more of the sections shown, but the filter characteristics (i.e. the relation between attenuation and frequency) depends almost entirely on the types of reactances used. For example, in a filter designed to pass high frequencies without attenuation it would be useless to use inductors as the series reactances since at high frequencies more voltage would be developed across the inductors than across the output terminals. If this is remembered it is easy to deduce the filter type by inspection of the component positions.

High-pass Filter

15. A high-pass filter is so designed that alternating currents below the cut-off frequency are severely attenuated while currents

BALANCED T FILTER   Ⓐ T FILTER   UNBALANCED T FILTER

GENERAL FILTER SYMBOL

BALANCED π FILTER   UNBALANCED π FILTER
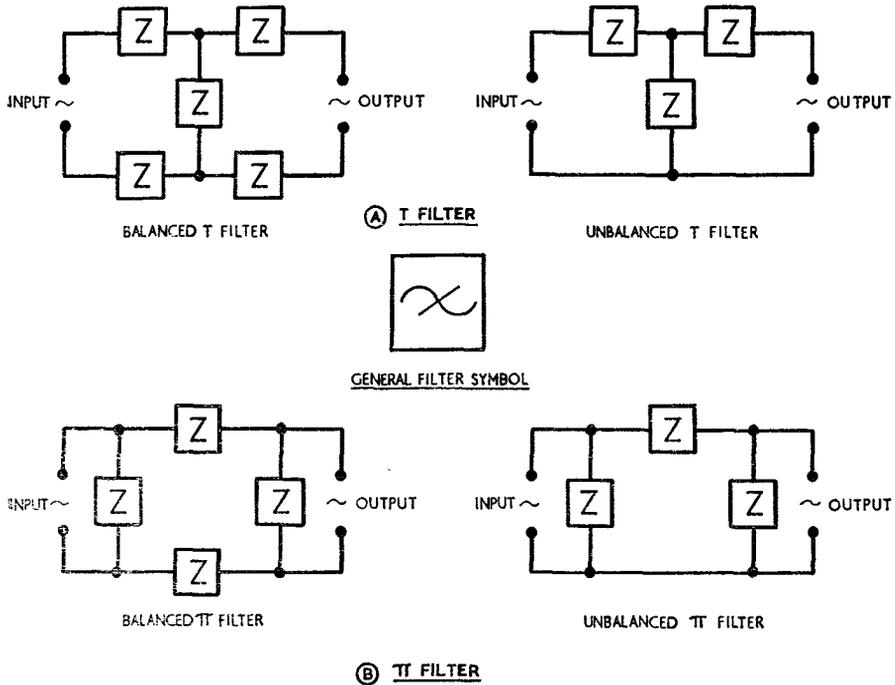
Ⓑ π FILTER

Fig. 10. TYPES OF FILTER.

Ⓐ T SECTION   Ⓑ π SECTION
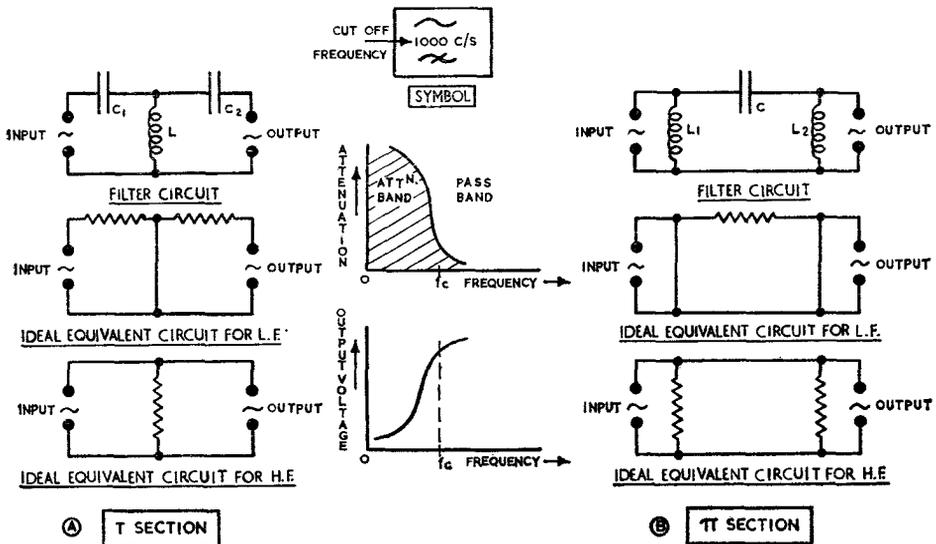
Fig. 11. HIGH-PASS FILTER.

above the cut-off frequency are passed with almost zero attenuation; that is it 'passes' high frequencies but does not admit low frequencies. The position of cut-off in the frequency spectrum is a matter of design, and depends on component values and the relation between inductive and capacitive reactances.

16. The networks shown in Fig. 11 are typical T section and π section high-pass filters. In the T section filter the capacitors C_1 and C_2 are equal in value, while in the π section filter the inductors L_1 and L_2 have equal values. In each case the series reactance is a capacitor; therefore if the input consists of a low frequency voltage most of the voltage will be developed across the series capacitors and little will appear across the output terminals. In Fig. 11 this is indicated by the presence of series resistors which represent reactances of very high value. If the frequency of the input voltage is increased a frequency is reached where resonance takes place and this causes the characteristic to drop sharply, indicating that the output voltage has increased and that the filter attenuation has suddenly decreased; this frequency is the cut-off frequency f_c. If the frequency of the input voltage is further increased, the reactance of the shunt inductor increases while the reactance of the series capacitors decreases; thus the voltage at the output terminals remains almost constant at a high value. This is indicated in Fig. 11 by shunt resistors which represent a high value of shunt reactance, and an absence of series reactance.

17. The characteristic of the high-pass filter shows firstly that it can be used to discriminate between voltages at frequencies which lie relatively near to, but on either side of the cut-off frequency, and secondly that its efficiency in the cut-off region depends on the steepness of the characteristic.

Low-pass Filter

18. In many radio and radar applications the need arises for circuits which pass currents at frequencies up to a certain value, and attenuate all others. One important example of this can be seen in line communication systems where, if high frequency currents were allowed to exist, unwanted feedback would occur and the communication system would be rendered unserviceable by virtue of self-oscillation. The circuit commonly used to attenuate high frequency currents while passing currents at low frequencies is the *low-pass* filter.

19. Since the pass-band of a low-pass filter must be from zero cycles per second to cut-off frequency the series components must be reactors whose reactance at low frequencies is very small; the series components are therefore inductors, and the shunt components are capacitors.

20. The circuits shown in Fig. 12 are circuits of typical T section and π section low-pass filters. As in Fig. 11, the value of cut-off frequency shown in the filter symbol is only an example; the cut-off frequency for any filter depends on the purpose for which it is required. If a low frequency voltage is applied to the terminals of either the T section or the π section, since the inductor reactance is very small at low frequencies most of the applied voltage is developed across the high reactance of the capacitors and therefore appears at the output terminals. As the frequency of the input voltage is increased, a frequency is reached where, as in the high-pass filter, the phenomenon of resonance causes the attenuation characteristic to rise sharply and the output voltage falls. As the frequency of the input voltage is further increased the reactance of the series inductors increases while the reactance of the shunt capacitor decreases, and the voltage at the output terminals remains almost constant at a low value.

21. The frequency at which resonance occurs is the cut-off frequency. Above cut-off the output voltage is very small, while below cut-off almost all the applied voltage appears at the output terminals. Even for single section filters such as those shown in Fig. 12 the characteristic is fairly steep at the cut-off frequency and the circuit can therefore be used to discriminate between voltages at frequencies which lie adjacent to, but on either side of the cut-off frequency.

Band Filters

22. The high-pass and low-pass filters already discussed are used in circumstances where one pass-band and one attenuation band only are required. They are generally quite efficient, but they are useless if the need arises for a filter which must be sensitive to a *band* of frequencies. Filters which satisfy
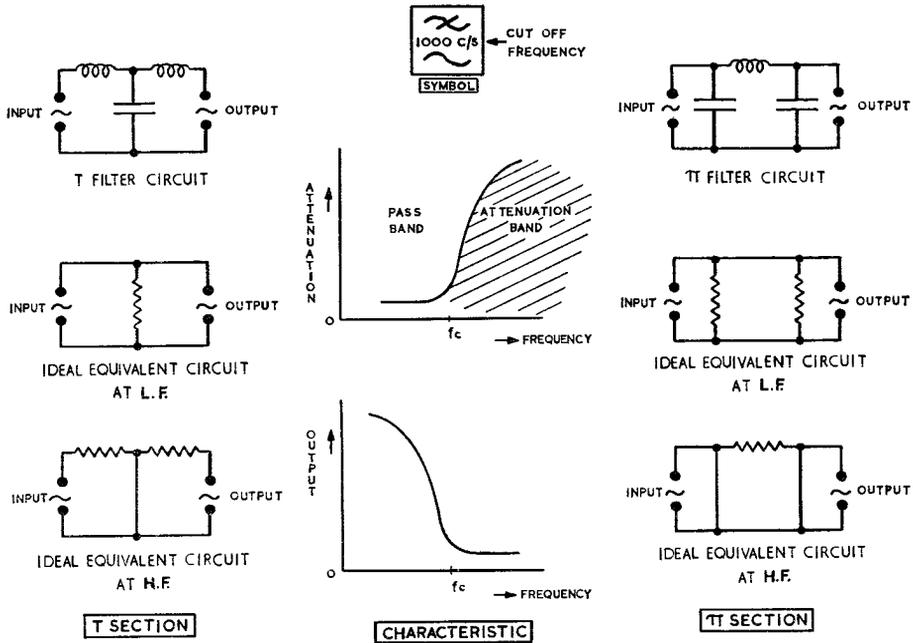
Fig. 12. LOW-PASS FILTER.

this need are called *band filters* and are extensively used in radio and radar systems. There are two types, the band-pass filter and the band-stop filter.

Band-pass Filter

23. A band-pass filter is a filter designed to pass alternating currents throughout a specific band of frequencies and to attenuate currents on either side of the band. It has two cut-off frequencies, the difference between them being equal to the width of the band of frequencies required to be passed.

For example, a superheterodyne receiver used for reception of radio telephony possesses intermediate frequency circuits which will pass a band of frequencies. This is because the amplitude modulated input signal consists of a voltage component at carrier frequency, plus voltages which differ from the carrier frequency by an amount equal to the modulation frequency. This is shown in Fig. 13 where f· is the carrier frequency and f_m the modulation frequency; it can be seen that if all components of the input signal are to be amplified the intermediate frequency circuits must pass a band of frequencies which extends from $(f-f_m)$ to $(f+f_m)$. This

can be achieved only by means of a band-pass filter which should possess a characteristic approaching the ideal shown in Fig. 13.
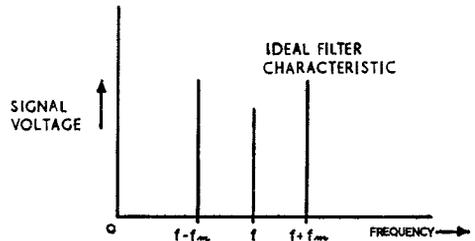
Fig. 13. BAND-PASS FILTER CHARACTERISTIC.

24. This type of filter is illustrated in Fig. 14. It consists of two parallel tuned circuits which are tuned to a common frequency and coupled by mutual inductance. Since inductive coupling is used the circuits are in effect coupled by an inductor common to both circuits, and it can be seen from the equivalent circuit that this type of filter is really a π network. As both primary and secondary are parallel tuned circuits, any input at low frequencies (frequencies below 2000 c/s in this case) is virtually short circuited by the low reactance of the primary and
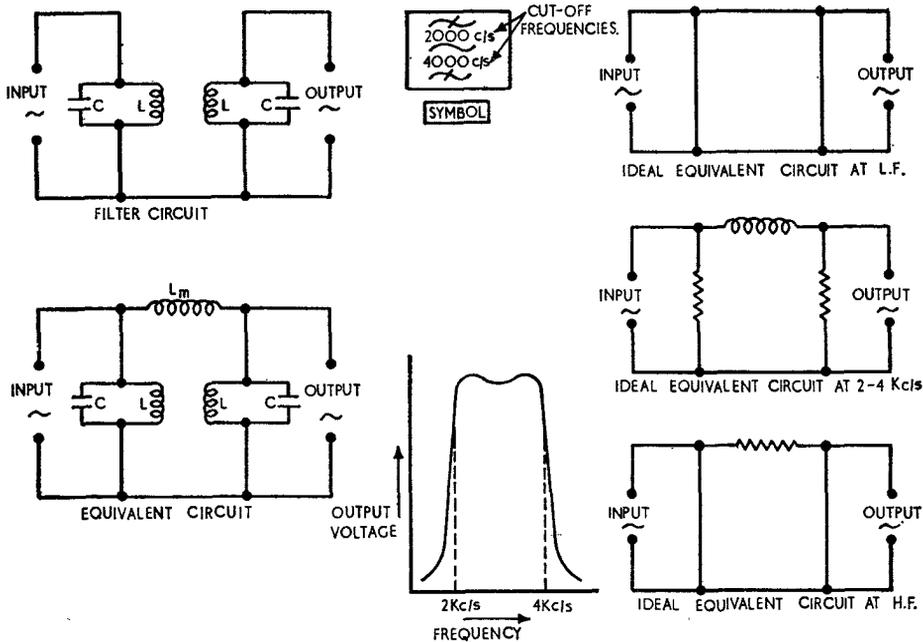
Fig. 14. BAND-PASS FILTER.

secondary inductors. As the input frequency is increased, the filter, since it consists of two coupled parallel tuned circuits, exhibits the effects of resonance over a band of frequencies, the width of which depends on the value of the coupling factor. Over this frequency band both circuits tend to act as high value resistors, and almost all the voltage appears at the output terminals. The filter attenuation is thus greatly decreased over the frequency band at which resonance effects occur, and the frequencies at which resonance effects suddenly disappear are the cut-off frequencies. Above the upper cut-off frequency (4000 c/s in the case illustrated in Fig. 14) the reactance of the circuit capacitors is small, virtually shunting both input and output as did the circuit inductors at low frequencies.

This type of band-pass filter is extensively used in the IF stages of a superheterodyne receiver since it meets the requirements mentioned in para. 23.

Band-stop Filter

25. In many radio and radar systems the need exists for a filter which will attenuate voltages over a band of frequencies and pass voltages outside the band. The *band-stop* filter meets this requirement. Like the band-pass filter it possesses two cut-off frequencies. It consists of tuned circuits connected in such a manner that throughout the frequency band in which resonance effects occur, the output voltage is very small.

26. An example of the use of a band-stop filter is in the case of a transmitter which is required to radiate only signals which are the sum and difference frequencies arising from modulation (the sidebands). This type of transmission occurs in suppressed carrier
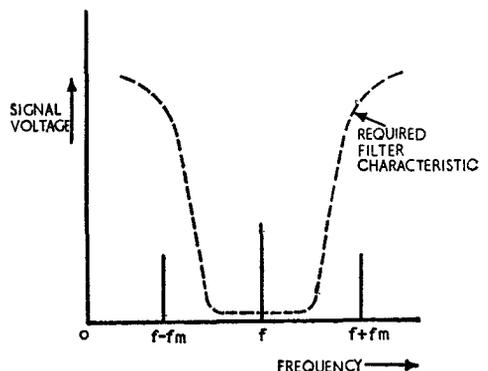
Fig. 15. BAND-STOP FILTER CHARACTERISTIC.
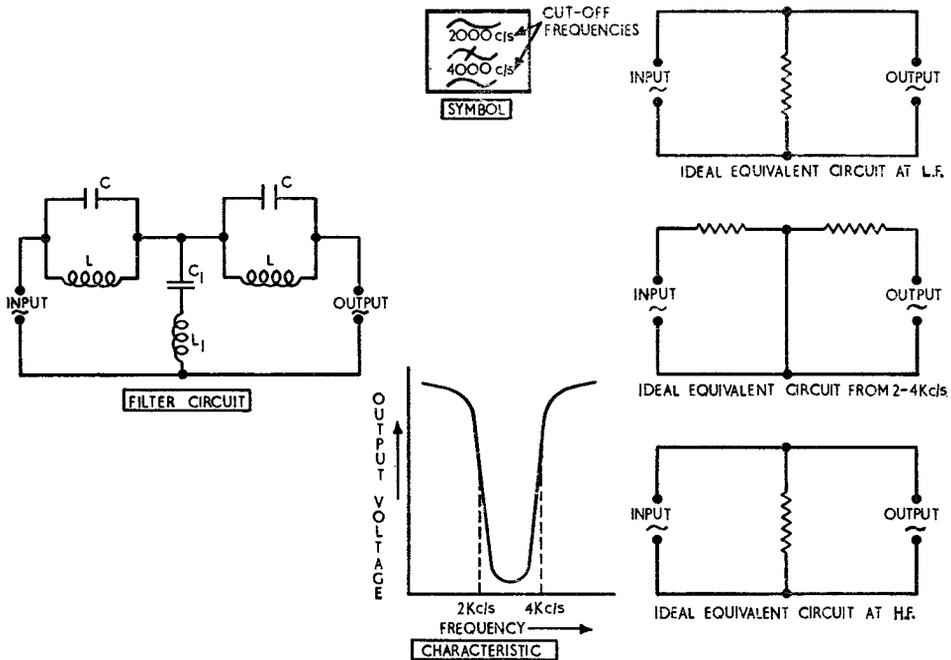
Fig. 16. BAND-STOP FILTER.

systems. In this case voltages at frequencies within a specific band must be attenuated, while voltages which are at frequencies just outside the band should remain unaltered. The required characteristic for this type of filter is shown in Fig. 15 where the carrier frequency is shown as f and the sum and difference frequencies are shown as (f + f_m) and (f − f_m) where f_m is the modulation frequency. As stated in para. 22, the only satisfactory filter for this purpose is one which is sensitive to a band of frequencies, i.e. a filter which possesses *two* cut-off frequencies.

27. As attenuation and resonance must coincide, the series filter components consist of parallel tuned circuits and the shunt components form a series tuned circuit. A typical band-stop filter, together with its equivalent circuits and its characteristic is shown in Fig. 16. The tuned circuits are tuned to a common frequency, but in order to provide attenuation over a band of frequencies, the *ratio* of L_1 to C_1 differs from that of L to C.

Crystal Filters

28. Since a quartz crystal exhibits the properties of a tuned circuit, and has an elec-

trical equivalent circuit as shown in Fig. 17, it can be used in place of the conventional inductors and capacitors so far discussed. The magnification factor (Q) of a quartz crystal is many times greater than that of a conventional tuned circuit and so the voltage/frequency response curve is much steeper. This means that the crystal can discriminate between two closely spaced frequencies much more effectively than can the LC circuit.
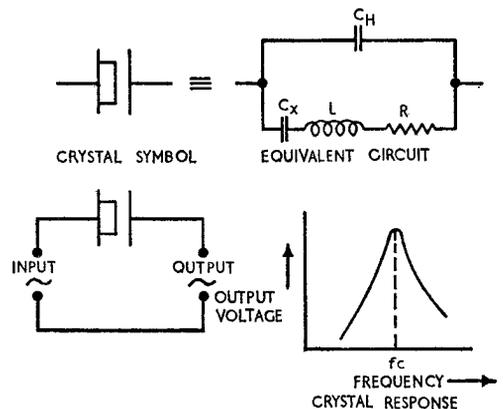
Fig. 17. CRYSTAL FILTER.

When the crystal is used as a filter component the cut-off frequency is clearly defined and the filter characteristic approaches that of the ideal filter.

29. A single crystal filter has a useful application in a telegraphy receiver where it may be desired to separate a signal on one frequency from signals on adjacent frequencies. The circuit of Fig. 18 shows this
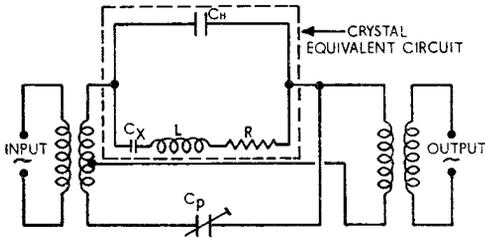
Fig. 18. CRYSTAL FILTER WITH PHASING CONTROL.

application. A limited control over the frequencies passed by the filter can be obtained by a variable phasing capacitor (Cp) placed in the filter circuit as shown.

Crystal Band-pass Filter

30. When it is required to pass a band of frequencies, crystals may again be employed. The advantage of a crystal band-pass filter over the LC band-pass filter is that for fewer components they approach nearer the ideal filter.
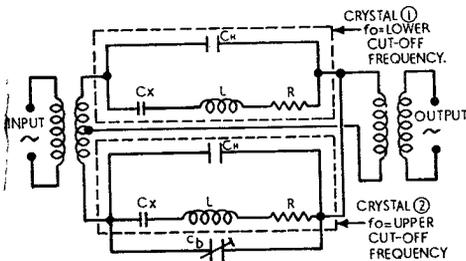
Fig. 19. BAND-PASS CRYSTAL FILTER.

The double crystal filter shown in Fig. 19 has two cut-off frequencies, the pass band being between these two. In order to achieve this band pass, the resonant frequencies of the crystals are such that one crystal resonates at the lower cut-off frequency and the other crystal resonates at the higher cut-off frequency.

31. Again, a limited control over the width of the band of frequencies to be passed is obtained by capacitor C_b placed in parallel with one of the crystals. In practical filters of this type, the maximum obtainable bandwidth is about 1% of the mid-band frequency; a typical double crystal filter having a mid-band frequency of 500 kc/s has a bandwidth of 5000 c/s.

Multi-Section Filters

32. The most important aspect of any type of filter is the steepness of its attenuation/frequency characteristic at the cut-off frequency. In an ideal filter the characteristic is a vertical line, but in practical filters cut-off does occupy a part of the frequency spectrum.

With single section high-pass and low-pass filters, cut off is not sharp enough, with the result that high values of input voltages at frequencies adjacent to cut-off frequency, cause an appreciable output voltage.

33. In the same way that the impedance/frequency curve for a tuned circuit system can be 'sharpened' by adding more tuned circuits so can the attenuation/frequency curve for a filter be made steeper by adding more filter sections. This is shown in Fig. 20
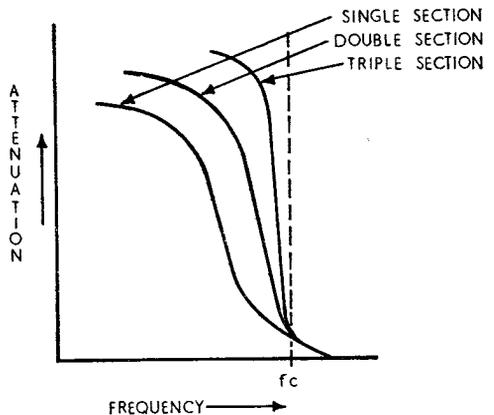
Fig. 20. ATTENTUATION/FREQUENCY CHARACTERISTIC FOR SINGLE, DOUBLE AND TRIPLE SECTION FILTERS.

where the attenuation/frequency characteristics of single, double and triple section high-pass filters are compared, all of which have the same nominal cut-off frequency.
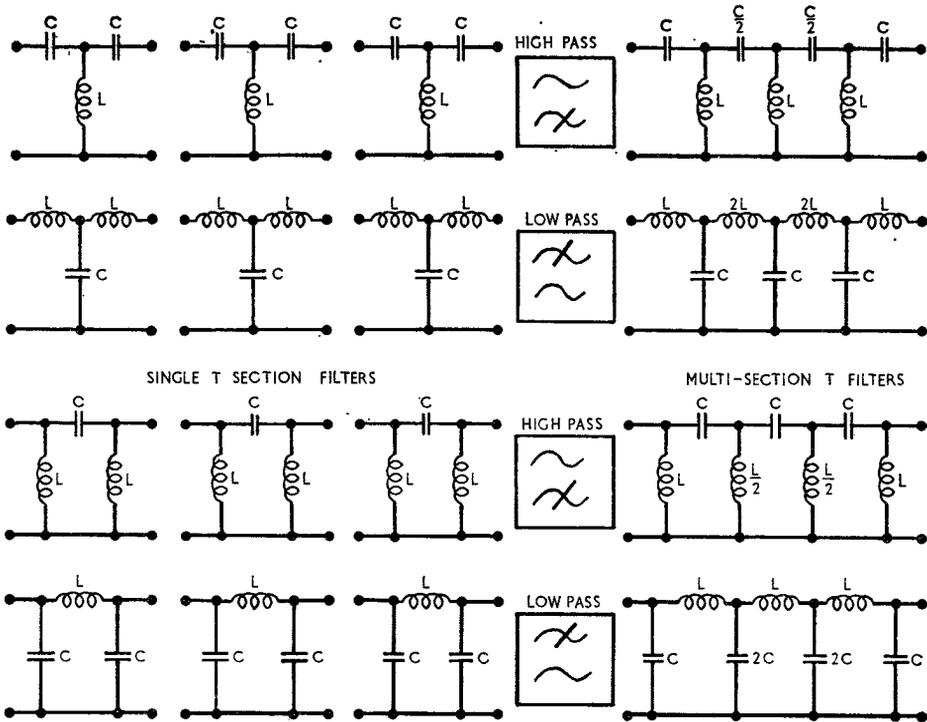
SINGLE T SECTION FILTERS   MULTI-SECTION T FILTERS

SINGLE π SECTION FILTERS   MULTI-SECTION π FILTERS

Fig. 21. CONSTRUCTION OF MULTI-SECTION FILTERS.

LOW-PASS	HIGH-PASS	BAND-PASS	BAND-STOP
INDUCTORS IN SERIES ARMS	CAPACITORS IN SERIES ARMS.	REJECTOR CIRCUITS IN PARALLEL ARMS	REJECTOR CIRCUITS IN SERIES ARMS

Fig. 22. TYPICAL FILTER SECTIONS.
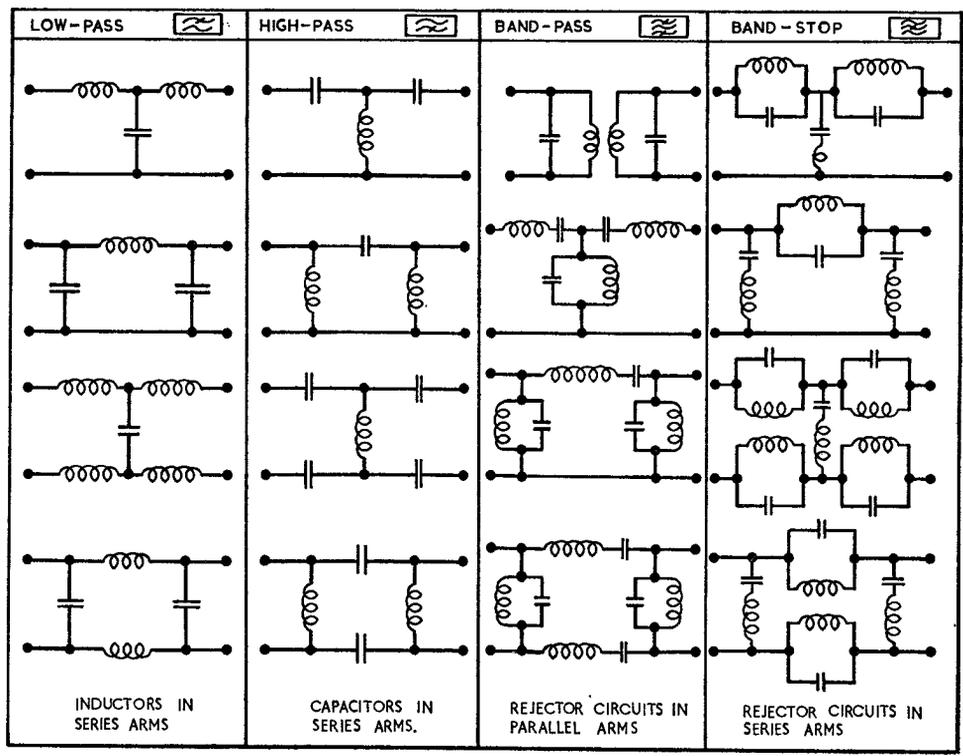
Construction of Multi-Section Filters

34.   In a single section T filter the series components are equal in value, and in a single section π type filter the parallel components are also equal in value.   A multi-section filter can thus be constructed by joining a number of sections in series and, to avoid component multiplicity, suitably altering the value of the appropriate components.   Thus in the high-pass multi-section T filter shown in Fig. 21, all capacitors except the first and last should have a value of $\dfrac{C}{2}$, where C is the capacitor value in the single section; similarly in high-pass multi-section π filters all inductors except the first and last should have a value of $\dfrac{L}{2}$ where L is the inductor value in the single section.

35.   The filters dealt with are a representative cross-section of the filters used in radio communication, radar, telephone and telegraph systems.   Many types of filters exist, but they are all based on those discussed in this Chapter, and may be derived from one of the single section filters shown in Fig. 22.

SECTION 15

CHAPTER 2

THE INFINITE TRANSMISSION LINE

THE INFINITE TRANSMISSION LINE

Introduction

1. The title of this chapter refers to a line which is infinitely long, and it may be wondered why time should be spent in reading about a device that is not practicable. The reason however becomes clear when it is realized that in a great number of cases, the attributes and characteristics of a *practical* transmission line are required to be the same as those of a line of infinite length; the transmission line is of value as an efficient means of conveying energy only if it *behaves* as an infinitely long line.

2. Before discussing the infinite line, it is desirable to understand what is meant by the term 'transmission line'. A transmission line, or 'feeder', is any system of conductors by means of which electrical energy can be transferred from one point to another with negligible loss. Some examples of the practical use of transmission lines are illustrated in Fig. 1, and it will be seen that as the lines must be fairly long, a large energy loss would render the systems useless.

In radio systems, the majority of transmission lines are used to transfer r.f. energy either from a transmitter to an aerial, or from an aerial to a receiver. The loss in each case must be as small as possible. In the first case, any loss of energy is wasteful as it involves a needless use of power at the transmitter, while in the second case a loss of r.f. energy could mean a serious loss of signals at the receiver input. Thus in any transmission line system, steps must be taken to ensure that any energy losses incurred are negligible in comparison with the magnitude of the energy being transferred.

Theoretical Construction of a Transmission Line

3. In the radio engineering context, a transmission line consists of two conductors suitably insulated from each other and supported throughout the line length. One type of r.f. transmission line, known as an '*open wire feeder*' consists of two conductors which are supported by insulators placed at intervals along the line as shown in Fig. 2.

TO TELEVISION RECEIVER

TO AERIAL

TO TRANSMITTER

TYPICAL TRANSMITTING STATION FEEDER
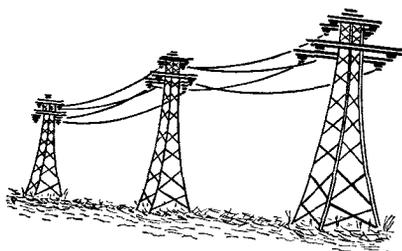
GRID TRANSMISSION SYSTEM.

Fig. 1. USES OF TRANSMISSION LINES.

CABLES. INSULATORS.

SUPPORT.
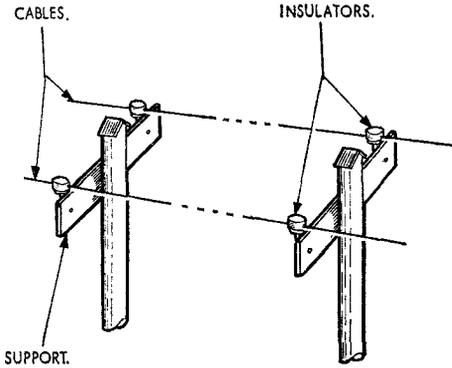
Fig. 2. BASIC TRANSMISSION LINE.

4. Any transmission line must possess a certain amount of *resistance*, although in practice this is very small. Also, since no insulator is perfect, there is a certain leakage current between the conductors, but in any practical transmission line, insulator resistance is very high, and *conductance* is therefore small.

In addition to resistance and conductance, each conductor has *inductance*, and *capacitance* exists between the two wires. The value of the inductance depends on the diameter of the wire and that of the capacitance on the distance the wires are apart. At radio frequencies, inductive reactance of the line is much greater than the line resist-

ance, which can therefore be neglected in comparison; and the shunt capacitive reactance between the lines is much less than the conductance, which can therefore be neglected in comparison.

Thus, at radio frequencies, a transmission line can be thought of as a series of inductors and capacitors distributed along its length as shown in Fig. 3.
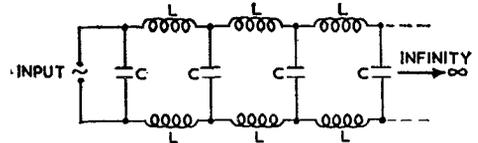
Fig. 3. TRANSMISSION LINE REACTANCES.

Characteristics of an Infinite Line

5. One of the main characteristics of any transmission line is its input impedance; that is, the impedance that the line presents to the source of supply. This is important because maximum transfer of energy from supply source to load is achieved only under properly matched conditions. It might seem that, irrespective of the impedance at the remote end of the line, the input impedance would vary with the length of the line, but this is true only to a limited extent. This can be seen by considering a line consisting of resistors arranged as shown in Fig. 4.
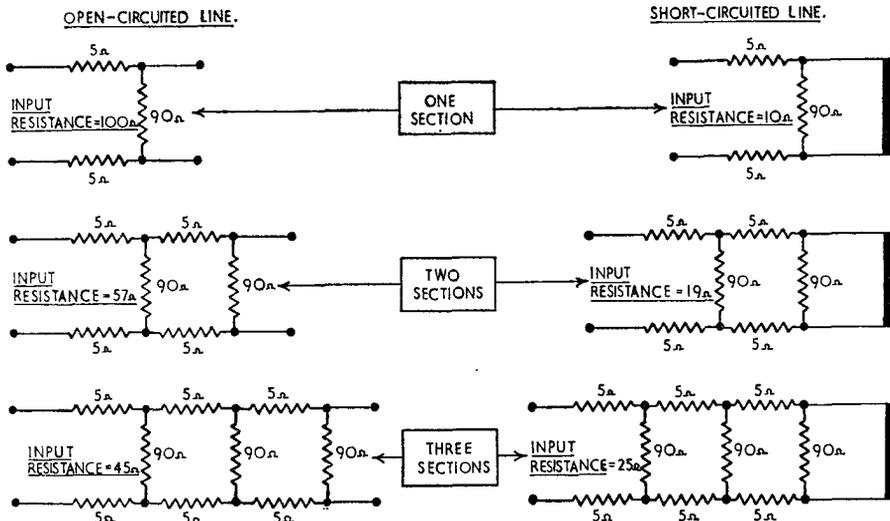
Fig. 4. LINE INPUT RESISTANCE.

6. In Fig. 4, two types of line termination are shown. The terminating resistance of the left-hand line sections is infinity (an open circuit), while that of the right hand sections is zero (a short circuit). The input resistance is shown in each case, and while the input resistance of the open line *decreases* with an increase in the number of line sections, the input resistance for the short circuited line *increases*. The increase and decrease in each case are by no means linear; in fact, the input resistance for both lines tends towards the same value (30 ohms in this case). In Fig. 5, graphs are shown of
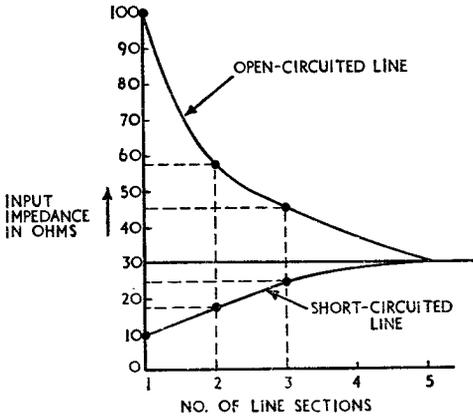
Fig. 5. VARIATION OF INPUT IMPEDANCE WITH LENGTH OF LINE.

line input resistance plotted against the number of line sections. As the length of the line is increased (more sections added) the difference in input resistance between the open and short-circuited lines becomes negligible, and with an infinitely long line, the input resistance would remain constant at a definite value. The actual value of input resistance would depend on the component values in each line section.

7. By replacing the resistors of Fig. 4 with series inductors and shunt capacitors and applying a r.f. input, similar conclusions about the value of the input impedance would be reached. Namely, that if a transmission line is infinitely long, the input impedance remains constant. Also the value of input impedance depends on the component values in each line section; that is, on the line inductance and capacitance per unit length (per foot, metre, yard, and so on).

Furthermore, if the line is infinitely long, energy travels down the line indefinitely and power is continuously absorbed from the supply. Since power cannot be dissipated by reactance but only by resistance, the infinitely long line is acting as a *resistance* of value equal to the input impedance (see Fig. 6). This means that the current and voltage travelling down the line must be *in phase*.
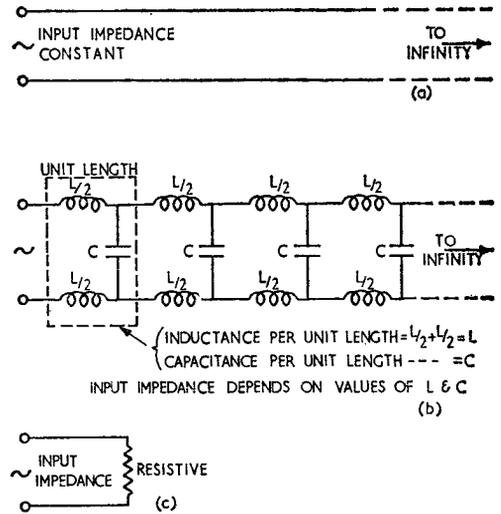
Fig. 6. INPUT IMPEDANCE.

Value of Input Impedance

8. It will be seen from the foregoing that the value of input impedance when the line is infinitely long depends on the values of line inductance and capacitance per unit length of line. As shown in Fig. 7, the input
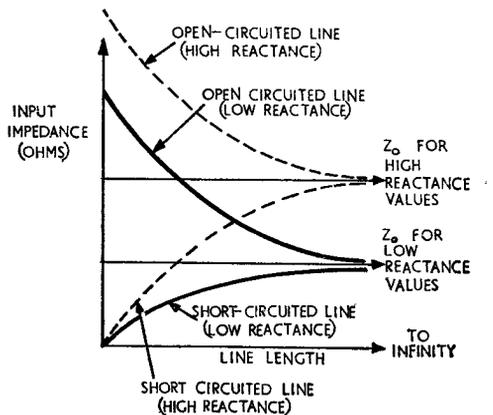
Fig. 7. VALUE OF INPUT IMPEDANCE.

impedance for a line where reactance values are high is higher than that for a low reactance line. However, irrespective of its actual value, the input impedance of an infinite line determines the value of line current and therefore determines the *ratio* of applied voltage to input current. Since a line consists primarily of distributed inductance and capacitance, some voltage must be dropped across each element of inductance and some current must be shunted through each element of capacitance. However, although both line voltage and line current decrease with line length in this way, since the line is uniform, the ratio of line voltage to line current remains constant; that is, *the ratio is the same at any point along the line* (Fig. 8). This ratio of line voltage to line current is a most important transmission line feature and is called the line *characteristic impedance*; the symbol is Z_0.
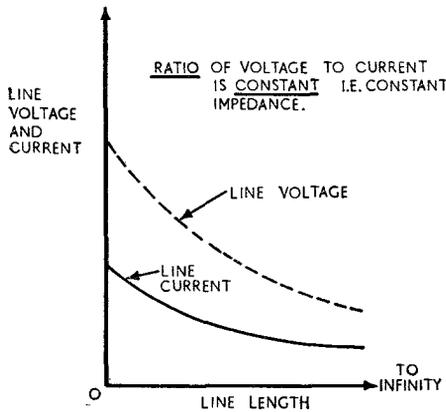
Fig. 8. VARIATION OF VOLTAGE AND CURRENT WITH LINE LENGTH.

The ratio of voltage to current at the line input is Z_0 and this is also the value of the input impedance of an infinite line. Thus, for an infinite line, Z_0 is resistive as explained in para. 7 in connection with input impedance.

Characteristic Impedance

9. It is stated earlier that the value of input impedance (and hence of the characteristic impedance Z_0) of an infinite line depends on the values of line inductance and capacitance per unit length. A high value of inductance increases the series reactance and increases the value of Z_0, while a high capacitance value decreases the shunt reactance and thus decreases the value of Z_0. It should be noted that Z_0 is practically independent of frequency, because as the frequency increases, inductive reactance increases while capacitive reactance decreases; Z_0 thus remains substantially constant with frequency.

10. At radio frequencies, when resistance and conductance can be neglected in comparison with inductive reactance and capacitive reactance respectively, the value of Z_0 is given by $Z_0 = \sqrt{\dfrac{L}{C}}$, where L and C are inductance and capacitance per unit length of line as shown in Fig. 9.
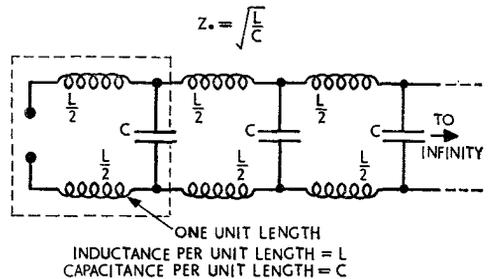
Fig. 9. RELATION BETWEEN INDUCTANCE AND CAPACITANCE PER UNIT LENGTH AND Z_0.

11. If the inductive reactance per unit length is 10 ohms, and the capacitive reactance is 90 ohms, then:—

$$X_L = \omega L = 10 \text{ ohms}$$
$$\therefore L = \frac{10}{\omega} \text{ henrys}$$
$$X_C = \frac{1}{\omega C} = 90 \text{ ohms}$$
$$\therefore C = \frac{1}{\omega 90} \text{ farads}$$
$$\text{and } Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{10}{\omega} \times \frac{\omega 90}{1}}$$
$$= \sqrt{900}$$
$$\therefore Z_0 = 30 \text{ ohms.}$$

The value of 30 ohms is merely an example. In practice, the value of characteristic impedance depends on the type of line, and values for transmission lines used in radio systems vary between about 45 ohms and 600 ohms.

Reflection of Energy

12. The purpose of a transmission line is to transfer energy from a source to some

type of load. This process should, of course be unidirectional; that is no energy should be returned from the load to the source and all the energy reaching the load should be dissipated in the load. If a line is of infinite length no energy will ever reach the remote end, and therefore no energy will be returned. This may seem a rather pointless statement, but when it is realised that if a practical transmission line can be made to behave as an infinitely long line, the practical transmission line will then possess the very desirable characteristic of being able to transfer e.m. energy from one point to another without causing any return of energy; that is, no energy will ever be reflected back from the load to the source.

Features of an Infinite Line

13. To conclude this chapter the main features of an infinite transmission line will now be stated. An infinite transmission line possesses an input impedance which remains constant; this input impedance is termed the characteristic impedance (symbol Z_0) and is resistive, i.e. voltage and current are in phase. When a source of voltage is applied, Z_0 gives the ratio of voltage to current at any point throughout the length of the line. The value of Z_0 depends on the ratio of inductance to capacitance per unit length of line, and varies little with frequency.

14. If a line is infinitely long, or possesses the characteristic of an infinitely long line, no energy can return from the load to the source; that is, no reflection can occur. The energy source is thus perfectly matched to the load.

15. Although an infinitely long transmission line is a physical impossibility the characteristics of such a line can be made to exist in a practical line. Chapter 3 shows how this is done, and shows how the features of an infinite line apply to any practical transmission line.

SECTION 15

CHAPTER 3

THE FINITE TRANSMISSION LINE

THE FINITE TRANSMISSION LINE

Introduction

1. The previous chapter dealt with a theoretical transmission line of infinite length, and discussed the important characteristics of such a line. This chapter deals with the characteristics of a *finite* transmission line such as that used in radio systems to carry both large and small amounts of r.f. energy.

2. One of the main uses of a finite transmission line is to convey r.f. energy from source to load. For example a transmitter (the source) situated some distance from its aerial (the load) must be connected to it by means of a length of transmission line. In order that maximum energy be transferred from transmitter to an aerial, the losses in the transmission line must be kept to a minimum. This entails correct matching of the internal impedance of the transmitter to the transmission line, and at the load end, correct matching of the transmission line impedance to that of the load (Maximum Power Theorem).

   If these conditions are satisfied, no re-flection of energy will occur and losses in the transmission system will be at a minimum.

3. It is equally important that these conditions be obtained in the case of a receiver remotely sited from its aerial. In this case the voltages induced in the aerial (the source) may be in the order of microvolts and to ensure that maximum available energy be presented to the receiver input (the load) transmission line losses must be kept to a minimum by correct matching.

4. If the impedance of the finite transmission line is known and is constant throughout its length, the matching problem can be solved. The *infinite* line has a known, constant impedance, the Z_0 equal to $\sqrt{\dfrac{L}{C}}$ ohms and no energy is reflected back along it. Thus if these characteristics can be reproduced in the finite line, the matching problem is solved.

Termination of a Finite Transmission Line

5. In the infinite line shown in Fig. 1(a), the input impedance is $\sqrt{\dfrac{L}{C}}$, where L is inductance and C capacitance per unit length of line.
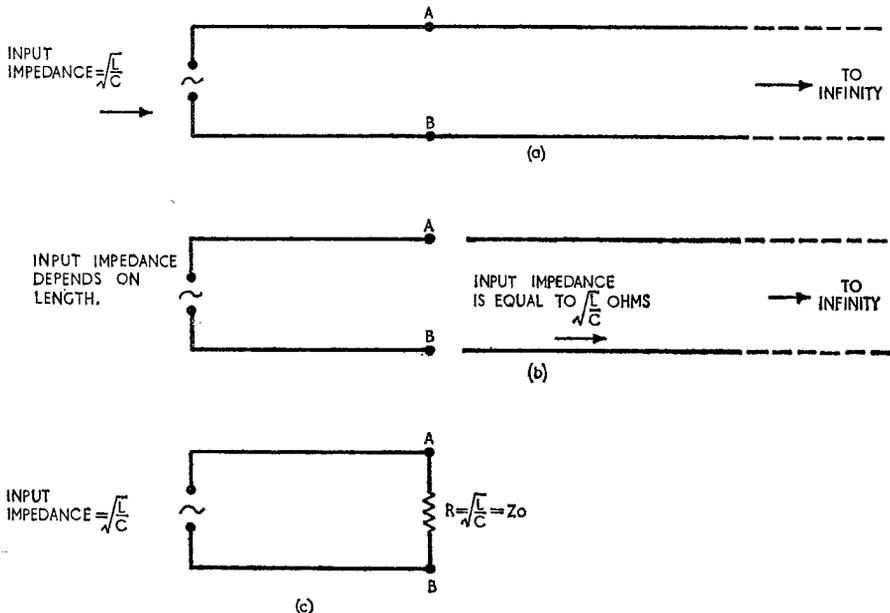
Fig. 1. TERMINATION OF FINITE TRANSMISSION LINE.

If this line is now broken at points AB, as shown in Fig. 1(b), the short length of line on the left will possess an input impedance depending on its length, while the length on the right, since it is of infinite length, must possess an input impedance of value $\sqrt{\dfrac{L}{C}}$. This means that the short length of line in Fig. 1(a) was terminated by a resistive impedance of value $\sqrt{\dfrac{L}{C}}$; that is, the short length of line was terminated at the points AB by an impedance equal to the line characteristic impedance Z_0. If now the short line in Fig. 1(b) is terminated by a resistor the value of which is equal to $\sqrt{\dfrac{L}{C}}$, the short length of line will behave in the same way as a line of infinite length. It will possess an input impedance equal in value to the characteristic impedance of the line and, since its characteristics are the same as those of an infinite line, no energy will be reflected back from the load to the source.

Need for Resistive Termination

6. The short transmission line mentioned in para. 4 is of course, of finite length, and energy will soon travel from the source to the load. It has been stated that providing the line is terminated by a resistor of value equal to the characteristic impedance of the line $\left(\sqrt{\dfrac{L}{C}}\right)$, no reflection will occur. It must be remembered that the line termination must be resistive; a reactance, whatever its value, cannot dissipate power, and if the line is terminated by a reactance, energy will inevitably be reflected back from the termination to the source. Transmission lines are seldom terminated by resistors; usually they are terminated by a practical load such as an aerial or a receiver. Whatever the actual load however, it must be resistive, i.e., it must be such that load voltage and current are in phase so that power is dissipated, and the resistance value must equal the line characteristic impedance so that *all* the power is dissipated.

Transmission Line Terminated in Z_0

7. If a transmission line of characteristic impedance Z_0 is terminated by a *resistive* load equal in value to Z_0, all the energy

propagated down the line is dissipated by the load, and there is no reflection back up the line.

Suppose this transmission line is being used to connect a transmitter to its aerial. If the transmitter output stage is correctly matched into the line, the line is correctly terminated at *both* ends and this is the condition for maximum power transfer from the transmitter, through the line, to the aerial. Losses are negligible.

The same considerations apply to transmission lines used to connect an aerial to a receiver. Provided that the line is correctly terminated at both ends, maximum power is transferred from the generator (the aerial) to the load (the receiver).

Production of Standing Waves— Terminating Resistance Equal to Zero

8. In order to show the effects which result from mismatching a transmission line, two extreme cases will be discussed. The first case is where the transmission line is termin-
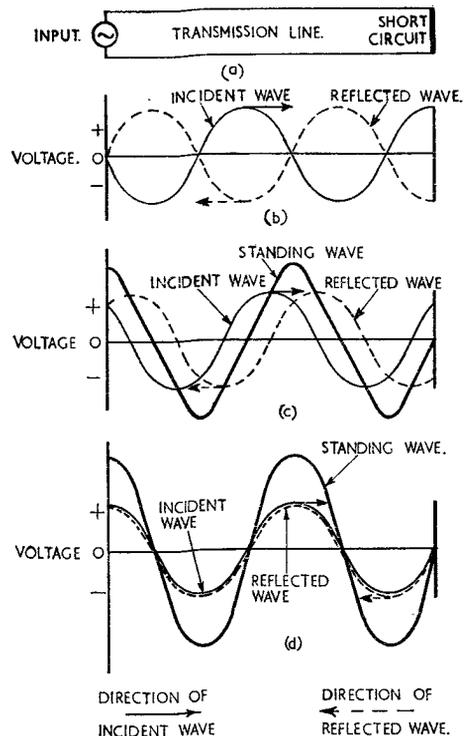
Fig. 2. STANDING WAVE ON A SHORT-CIRCUITED TRANSMISSION LINE.

ated by a short circuit; that is, when the value of the terminating resistance is zero as in Fig. 2(a).

Since the termination resistance value is zero, and the voltage across a short circuit is zero, the voltage at the end of the line must always be zero.

However, since the input voltage varies sinusoidally, the variation in voltage along the line is also sinusoidal. As the voltage at the termination must be zero, another voltage, equal and opposite to the forward travelling or incident voltage wave, must exist at the line termination, as shown in Fig. 2(b). The equal and opposite voltage is the *reflected* voltage. In Fig. 2(b), the two voltages are equal and opposite throughout the line, but this is not always the case.

9. Fig. 2(c) shows the situation one-eighth of a cycle later. The incident voltage wave has progressed forward by an eighth of a cycle, and the reflected voltage wave has advanced towards the source by an eighth of a cycle; the resultant voltage at the termination is still zero, as it must be across a short circuit, the reflected voltage being equal and opposite to the incident voltage at this point. At some points along the line the waves cancel, but at other points they add, causing a resultant voltage to exist which has an amplitude equal to the sum of the incident and reflected voltages. The resultant voltage produced by the combination of the incident and reflected voltages is referred to as a 'standing wave' of voltage.

10. Fig. 2(d) shows the state of affairs yet another eighth of a cycle later; the incident and reflected voltage waves are in phase, and the standing wave of voltage is thus at maximum amplitude.

11. The *position* of the resultant voltage does not change, although both incident and the reflected waves move along the line; hence the term 'standing wave' for the resultant voltage. Although the position of the standing wave is fixed, it rises and falls sinusoidally about this fixed position according to the frequency of the input. Note that a standing wave exists only when incident and reflected waves are both present, and for a short circuited line the amplitude of the voltage standing wave at the termination is zero.

Terminating Resistance Equal to Infinity

12. This is the second case of extreme mismatching, and it occurs when the value of the termination resistance is infinity; that is, when the transmission line is terminated by an open circuit, as in Fig. 3(a).

As in the previous case, graphs can be drawn to illustrate both incident and reflected waves, and the resultant standing wave. Similar conclusions result, the only point of difference being the position of the voltage standing wave on the line. For the short-circuited line, the voltage standing wave has zero amplitude at the termination; for the open-circuited line, the voltage standing wave has maximum amplitude at the termination.
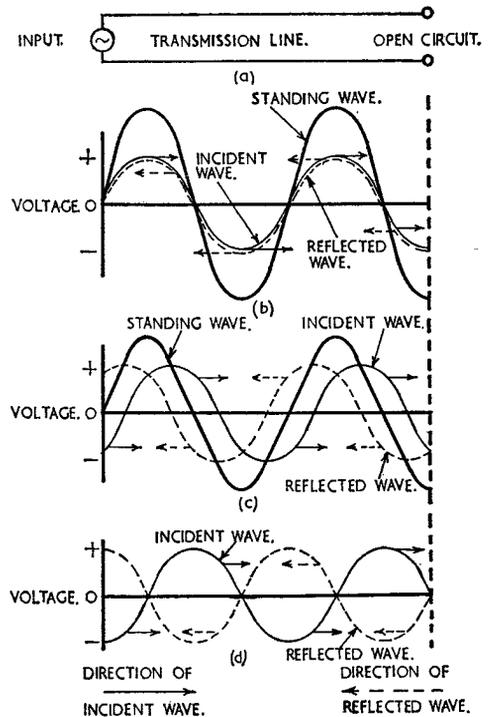
Fig. 3. VOLTAGE STANDING WAVE ON AN OPEN-CIRCUITED TRANSMISSION LINE.

13. The incident voltage wave progresses down the line from the input until it reaches the termination. Since no power can be absorbed in an open circuit, complete reflection occurs at the termination and a reflected wave, equal in amplitude to the incident wave, travels back up the line to-

wards the source. In the short circuited line it was noted that the reflected voltage at the termination was equal and of opposite sign to the incident voltage to give a resultant of zero at the termination. In the open-circuited line, the reflected voltage at the termination is equal to the incident voltage but is of the *same* sign (i.e. reflected in phase with the incident voltage). This is shown in Fig. 3(b) where over the time depicted the reflected and incident voltages are in phase all the way along the line and combine to produce a voltage standing wave of maximum amplitude.

14. Fig. 3(c) shows the situation one-eighth of a cycle later. The incident voltage wave has progressed forward by an eighth of a cycle, and the reflected voltage wave has advanced towards the source by an eighth of a cycle, the reflected and incident voltages at the termination being equal and of the same sign. The resultant standing wave of voltage, obtained by combining the incident and reflected voltage waves, has a smaller amplitude than that shown in Fig. 3(b).

15. Fig. 3(d) shows the state of affairs yet another eighth of a cycle later. The incident and reflected voltages are now such that their resultant is zero all along the line.

16. As in the short-circuited case, the resultant standing wave of voltage does not move along the line although, of course, it varies sinusoidally at the frequency of the input voltage. Its maximum amplitude during one cycle of the input voltage indicates the amount by which the value of the line termination differs from Z_0.

Effect of Terminating Resistance not equal to Z_0

17. So far three possible terminations to the finite transmission line have been considered. (i) The correctly terminated line, where the terminating resistance is equal to the Z_0 of the line. All the energy sent down the line is dissipated by the resistance, and none is reflected. (ii) The line terminated in a short circuit and (iii) the line terminated in an open circuit. In both these latter cases no energy is dissipated in the termination, all the energy sent down the line being reflected.

18. A practical line designed to convey energy from transmitter to aerial would be terminated by the aerial which would be unlikely to be purely resistive *and* equal in value to the Z_0 of the line. If a transmission line is terminated in an impedance that is *reactive* or not equal in value to the Z_0 of the line, or both, the line is said to be 'mismatched'. In this case only part of the r.f. energy travelling down the line is dissipated in the load, and reflection occurs.

19. The amount of reflection, and hence the magnitude of the resultant standing wave, depends on the degree of mismatch at the termination. In the short-circuited and open-circuited line the degree of mismatch is absolute; all the incident energy is reflected and the magnitude of the standing wave is a maximum. However, the closer the load impedance is in value to the Z_0 of the line, the greater is the energy absorbed at the termination, the smaller the amount of reflected energy, and the less the magnitude of the standing wave on the line.

The magnitude of the standing wave is therefore an indication of how close the load impedance is in value to the Z_0 of the line, and since the magnitude of the standing wave can be measured by means of a valve voltmeter or other suitable instrument, it is usually easy to determine whether or not a transmission line is correctly terminated.

20. Fig. 4 illustrates the effect of a terminating load of incorrect value. In Fig. 4(a) energy is fed to the input of a transmission line and progresses towards the load. The value of the load resistor is *less* than that of the line Z_0, and all the forward energy cannot be dissipated; some is therefore reflected back towards the source.

In Fig. 4(b) the load resistor has a value *greater* than the line Z_0 and energy is again reflected back towards the source.

Fig. 4(c) shows the situation when the value of the load resistor equals Z_0; all the forward energy is dissipated in the load.

The only condition which permits all the forward energy to be dissipated in the load, is when the load is purely resistive and equal in value to the characteristic impedance of the line.

21. It is worth noting at this point that when the terminating impedance is *less* than Z_0, conditions approaching that of the short-circuited transmission line exist; the reflected voltage at the termination is of
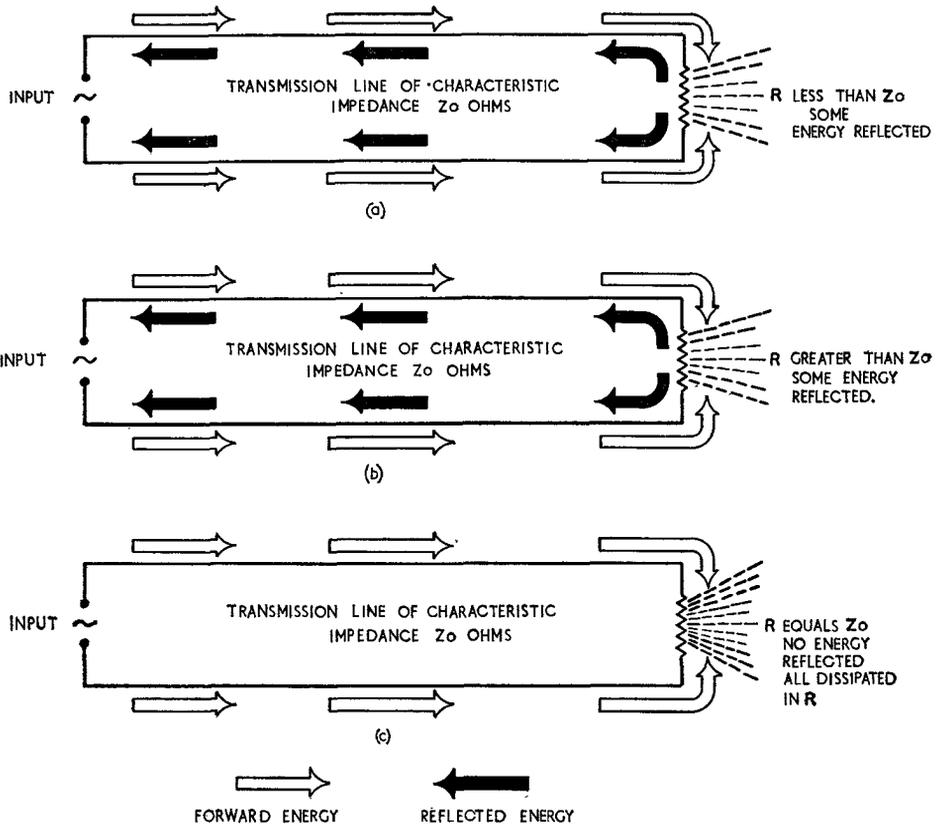
Fig. 4. ENERGY REFLECTION ON A TRANSMISSION LINE.

opposite sign to the incident voltage, and the amplitude of the standing wave of voltage at the termination is a minimum.

Similarly, when the terminating impedance is *greater* than Z_0, conditions approaching that of the open-circuited transmission line exist; the reflected voltage at the termination is of the same sign as the incident voltage, and the amplitude of the standing wave of voltage at the termination is a *maximum*.

Standing Wave Ratio (SWR)

22. When a transmission line is connected between a transmitter and an aerial, an alternating voltage and an alternating current are propagated down the line from the transmitter. The r.m.s. value of voltage (or current) at any point along an unscreened line can be measured by connecting a suitable instrument to a loop of wire and placing the loop close to one of the wires in the feeder.

The voltage induced in the loop by the alternating magnetic (or electric) field produces a reading in the meter (Fig. 5(a)).

When the line is correctly terminated by a resistive impedance equal to the Z_0 of the line, no reflection occurs and no standing waves are produced. Only the incident wave is present and the meter gives the same reading at all points along the line (Fig. 5(b)).

23. The result of a mismatch, on the other hand, is standing waves and the r.m.s. value of the voltage (or current) will vary from point to point, rising and falling between a maximum and a minimum value as the meter is moved along the line.

In the short-circuited transmission line, complete reflection occurs and the reflected wave is equal in amplitude to the incident wave, so that the standing wave on the line has a maximum amplitude of twice the value of the incident voltage (or current) and a
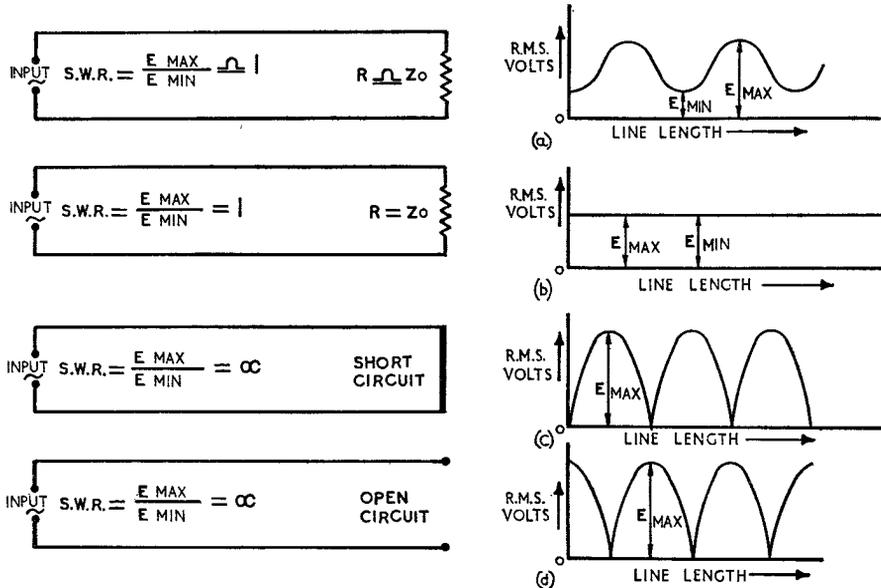
Fig. 5. EFFECT OF TERMINATION ON STANDING WAVE RATIO.

minimum amplitude of zero (Fig. 5(c)). This is also the case for the open circuited transmission line (Fig. 5(d)).

For any termination between the two extreme cases other than a termination of Z_0, some energy is absorbed and some is reflected so that the amplitude of the reflected voltage is less than that of the incident voltage. The r.m.s. value of the resultant standing wave of voltage therefore never attains a maximum amplitude of twice the value of the incident voltage or a minimum amplitude of zero (Fig. 5(a)).

24. The *range* of variation between maximum and minimum values gives an indication of the degree of mismatch and is usually expressed in terms of the '*standing wave ratio*' (SWR). This ratio is given by E_{max}/E_{min} (or I_{max}/I_{min}), where E_{max} denotes the maximum r.m.s. voltage indicated by the meter and E_{min} denotes the minimum r.m.s. voltage.

When the line is correctly terminated in a resistive impedance equal to Z_0, there is no standing wave, E_{max} equals E_{min} and the SWR is unity; this is the case in Fig. 5(b). With a slight mismatch, the reflected wave is small; E_{max} is then only slightly greater than E_{min} (Fig. 5(a)) and the SWR is slightly greater than one. As the degree of mismatch

increases, the SWR gets greater and for the open-circuited or short-circuited transmission line the SWR is $\frac{E_{max}}{0}$ = Infinity (Figs. 5(c) and 5(d)).

25. It is seen that correct termination of a transmission line is indicated by a SWR of unity. In practice, this may be difficult to obtain, but the maximum permissible SWR for an installation is given in the appropriate Air Publication for the equipment and the line must be examined and adjusted where necessary to give the required conditions.

Velocity of Propagation

26. The time taken for r.f. energy to travel from the beginning to the end of a transmission line is very small; in fact its velocity is comparable, in most cases, with the velocity of propagation of e.m. energy in space. However, the velocity of propagation in a transmission line depends to some extent on the physical line constants. As mentioned previously, any transmission line consists of distributed inductance and capacitance; if a voltage is applied to the input of a transmission line, as shown in Fig. 6, the voltage across the first section depends on the charge on C_1. Since the charging current of C_1 is opposed by inductive reactance, the time
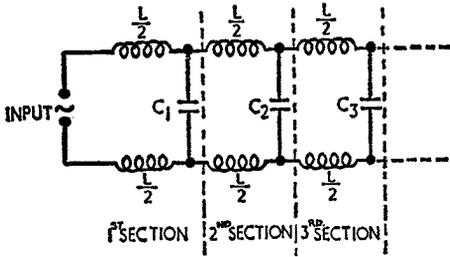
Fig. 6. TRANSMISSION LINE REACTANCE.

taken for C_1 to charge depends on the value of the inductance and capacitance of the first section. Since the capacitance of the second section is charged from the voltage which exists across the first section, the energy travels along the line at a velocity which is proportional to the line inductance and the capacitance per unit length. In fact, the time taken for r.f. energy to propagate along a transmission line is $\sqrt{LC}$ seconds, where L and C are inductance and capacitance per unit length; the velocity of propagation is thus $\dfrac{1}{\sqrt{LC}}$ unit lengths per second.

Propagation Coefficient

27. One of the most important aspects of a transmission line is the relationship between the input and output voltages, i.e., the source voltage and the load voltage. Since the source voltage is always alternating, the relationship between the input and output voltages must include phase as well as voltage difference. For this reason, the *propagation coefficient*, as it is called, is divided into two parts; the first part, called the *attenuation coefficient*, gives the amount of voltage (or current) decrease throughout the line length; the second part, called the *phase-change coefficient*, gives the phase difference between the source voltage and the voltage at any other part of the line.

Attenuation Coefficient

28. In a perfect transmission line the voltage at the remote end would be equal in amplitude to the source voltage. A perfect transmission line of course, is an impossibility and some voltage and current attenuation must occur due to the fact that a practical line possesses series resistance and conductance between the conductors.

Due to these and other factors, the line voltage decreases with length in the manner shown in Fig. 7(a). As the line voltage is alternating, its amplitude at any point on the line can be indicated by a series of rotating vectors which diminish in amplitude, as shown in Fig. 7(b). The attenuation coefficient of any transmission line is the ratio of two voltage vectors at two points on the line. For example, the attenuation coefficient of a given length of line may be given by ratio of E_1 to E_2, which denotes the amplitude of the
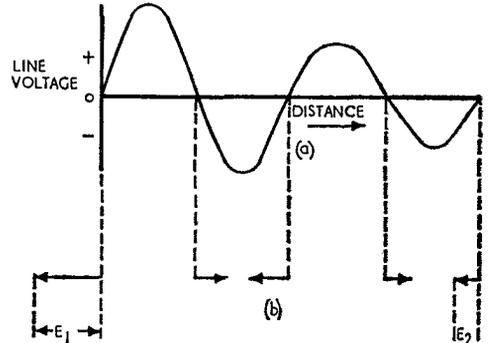
Fig. 7. MEASUREMENT OF ATTENUATION COEFFICIENT.

voltage at the beginning and end of that length of line. The attenuation coefficient of a transmission line is usually expressed as so many nepers per mile. The neper, as pointed out in Sect. 6, Chap. 3, is a logarithmic unit giving the gain or loss of a system. Using the neper to calculate the loss in a transmission line:—

Attenuation coefficient =

$\log_e \dfrac{E_1}{E_2}$ nepers per mile, where E_1 is the amplitude of the voltage at the beginning of a mile of line, and E_2 is that at the end.

The neper can be converted to decibels by multiplying the answer by $8 \cdot 686$.

29. Since the presence of a standing wave would complicate the vector system, the attenuation coefficient of a transmission line is valid only when the line is correctly terminated; that is, when it is terminated by a resistive impedance equal to the Z_0 of the line.

Phase-change Coefficient

30. Since the voltage on a transmission line is alternating, a phase difference can exist

between the voltage at the source and the voltage at any given point along the line; this phase difference is defined by the phase-change coefficient, which is a measure of the amount by which the voltage (or current) is shifted in phase as it passes down the line. In one measured wavelength of line, the voltages at the beginning and end should be in phase. They may not be however, due to the reactive components in the line. The amount they are out of phase is a measure of the phase-change coefficient. The phase-change coefficient is generally expressed in radians; a typical figure is 0·25 radians per mile.

Types of Transmission Lines

31. The four examples of transmission lines shown in Fig. 8 represent the main types of line used in radio systems. Each type has a particular application, and the type in use in any radio system depends on factors such as the required characteristic impedance, the power to be carried and the frequency of the r.f. energy being propagated.

32. **Open wire feeder.** The open wire feeder (Fig. 8(a)) consists of two parallel wires spaced a small fraction of a wavelength apart; it is usually run about ten feet above the ground and supported on insulators at intervals of 70 to 100 feet. It is generally used when a relatively high value of characteristic impedance and the ability to handle large amounts of power are required.

Since $Z_0 = \sqrt{\dfrac{L}{C}}$, an increase in spacing decreases C and increases Z_0, and the diameter of the wires determine the value of L. In practice, the Z_0 of an open wire feeder lies between 150 and 1000 ohms, common figures being 300 and 600 ohms.

The open wire feeder has certain advantages over the other types of transmission line shown in Fig. 8, but it also has certain limitations:—

(a) **Advantages**
(i) High transmitter power outputs can be handled without danger of breakdown of the air dielectric.
(ii) Standing waves can be easily measured and maintenance of the line is relatively simple.
(iii) The open wire feeder is a *balanced* line; that is, the impedance between wire and earth is the same for each wire; this is an important property when considering matching the line to the load (see Chapter 4).

(b) **Limitations**
(i) It is bulky and rigid and can be used only on static installations.
(ii) To limit radiation losses, the spacing between the wires is a small fraction of a wavelength. At very high frequencies the spacing becomes so small that if high powers are being handled there is a danger of 'flashover' between the wires.
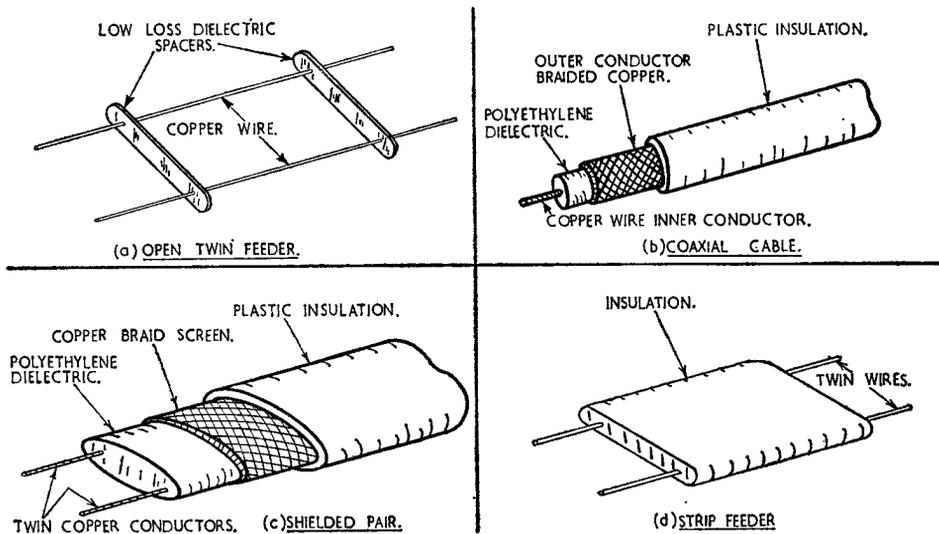
Fig. 8. TYPES OF TRANSMISSION LINES.

The practical upper frequency limit for open wire feeders on high power installations is about 100 Mc/s.

(iii) It must be kept clear of the ground and walls.

33. Coaxial feeder. This is a transmission line in which one conductor is in tubular form and completely surrounds and screens the inner conductor (Fig. 8(b)). The inner conductor is a wire running along the axis of the outer tube. It is held in the correct position relative to the outer conductor by the use of insulating washers spaced along the line at frequency intervals, or by completely filling the space between the conductors with a low-loss dielectric (e.g. polyethylene). The outer conductor may be a metal tube or it may be constructed of metal braiding to give flexibility. It is protected by an outer covering of some plastic material.

The characteristic impedance of a coaxial feeder depends on the radius of each conductor, the spacing between them and on the dielectric constant of the separating medium; in practice it generally lies between 40 and 100 ohms, common figures being 45 and 75 ohms.

In relation to the open wire feeder, the coaxial feeder has certain advantages and certain limitations:—

(*a*) **Advantages**
(i) The coaxial line is a screened cable and losses due to radiation are negligible. It can therefore be used up to frequencies of the order of 1000 Mc/s; above this, waveguides are used.
(ii) It is flexible and compact and can be buried in the ground.

(*b*) **Limitations**
(i) The power handling capabilities are less than those of an open wire feeder.
(ii) The coaxial feeder is an *unbalanced* line; that is, the outer conductor is earthed to provide a screen and the inner conductor is isolated from earth; this introduces additional problems when matching the line to the load (see Chapter 4).

34. Twin wire feeder. The twin wire feeder (or shielded pair) consists of two parallel conductors mounted in a flexible cable and suitably insulated from each other by polyethylene (Fig. 8(c)). A metal braiding surrounds the polyethylene and acts solely as a screen, the whole cable being protected by an outer covering of some plastic material. In this way, the advantages of the coaxial feeder are combined with one advantage of the open wire feeder. The twin wire feeder is thus a screened and flexible line and also a balanced line. Its characteristic impedance is of the same order as that of a coaxial cable (40 to 100 ohms approximately), but since its power handling capacity is relatively small for a given size and weight, it is used only for special applications.

35. Strip feeder. This consists of two covered conductors mounted at the sides of a strip of insulating material (Fig. 8(d)). This type of line is extremely flexible, but has a relatively small power carrying capacity and for this reason is usually used only for receiving systems.

Summary

36. In this chapter, the finite length of transmission line has been discussed from the very important aspect of conveying r.f. energy from source to load with mimimum losses. It has been shown that to achieve this, the line must be terminated with a pure resistance equal to the Z_0 of the line. If this is not achieved, standing waves are set up along the line and the standing wave ratio indicates the degree of mismatch.

37. In discussing this, the effects of a short circuit and open circuit on a finite length of line, were examined. These effects are very important and have numerous practical applications in wireless and radar systems; these applications are discussed in detail in Chapter 4.

SECTION 15

CHAPTER 4

TRANSMISSION LINE TECHNIQUES

TRANSMISSION LINE TECHNIQUES

Introduction

1. If it were suggested that the two conductors of a high power r.f. transmission line could be held in position by metal spacers, as shown in Fig. 1, the suggestion would, on the face of it, seem ridiculous. Yet the use of metal spacers, which can
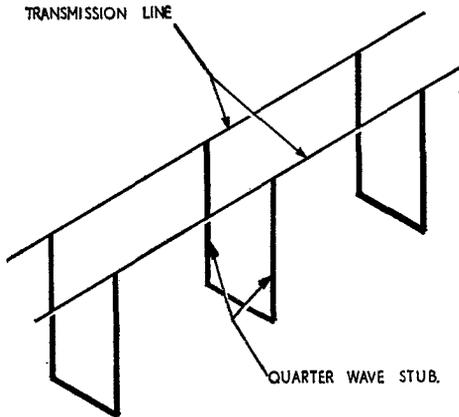
Fig. 1. TRANSMISSION LINE SUPPORTED BY QUARTER-WAVE STUBS.

paradoxically be termed 'metallic insulators', is merely one application of transmission line techniques. The metal spacers are really short lengths of transmission lines, so constructed that they act as extremely high value resistors, and they can therefore be used as mechanical supports and spacers for a high power transmission line without causing undue power loss.

Many other uses exist for short lengths of transmission lines, and the purpose of this chapter is to show how transmission line techniques enable short transmission lines to be used in radio systems.

2. Most of the phenomena exhibited by short transmission lines can be explained by considering the distribution of current and voltage along the line. The electrical behaviour of any component is determined by the phase relationship between the current established in the component and the voltage across it. If current and voltage are in

phase the component behaves as a resistor, while if current and voltage have a phase difference of 90° the component behaves as a reactor. A transmission line behaves in the same way as any other component and its influence on the circuit to which it is connected depends on the phase difference between line voltage and line current at the point of connection.

Short Length of Open-Circuited Line

3. When a transmission line is terminated by an open circuit, no energy is absorbed in the termination. Complete reflection occurs and a standing wave of voltage is produced on the line as shown in Fig. 2.
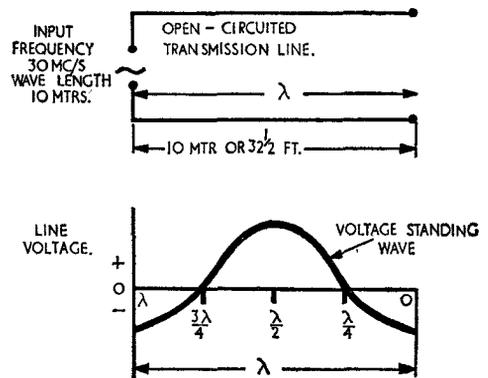
Fig. 2. VOLTAGE STANDING WAVE ON AN OPEN-CIRCUITED TRANSMISSION LINE ONE WAVELENGTH LONG.

The frequency of the input voltage in this case is such that one complete cycle of voltage variation occurs within the length of the transmission line; that is the line is one wavelength (λ) long. Note that a voltage maximum occurs at the termination, and a voltage minimum one quarter of a wavelength $\left(\frac{\lambda}{4}\right)$ back from the termination.

4. It is usual in radio engineering, to refer to distances along the transmission line in terms of wavelengths of the energy being sent down the line, and it is necessary to understand what is meant by this.

INPUT

10 METRES = 32.5 FEET.

(a)

FREQUENCY OF INPUT 40 MC/S. WAVELENGTH 7·5 METRES.

VOLTAGE

$\frac{5\lambda}{4}$

λ $\frac{3\lambda}{4}$ $\frac{\lambda}{2}$ $\frac{\lambda}{4}$

ELECTRICAL LENGTH OF LINE = 1·25 WAVELENGTHS

LENGTH OF LINE

(b)

FREQUENCY OF INPUT 24 MC/S. WAVELENGTH 12·5 METRES

VOLTAGE

$\frac{3\lambda}{4}$ $\frac{\lambda}{2}$ $\frac{\lambda}{4}$

ELECTRICAL LENGTH OF LINE = 0·75 WAVELENGTHS

LENGTH OF LINE

(c)
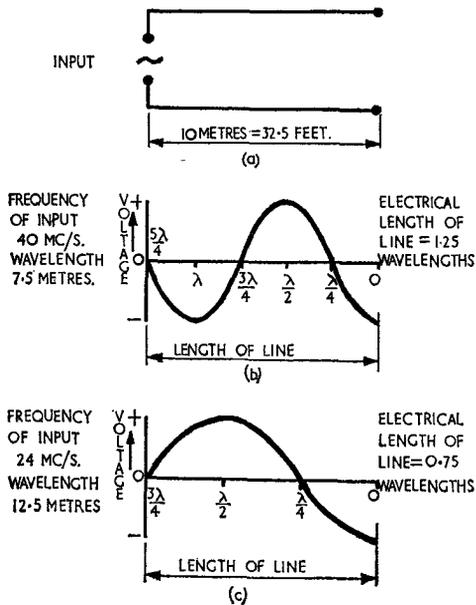
Fig. 3.   ELECTRICAL LENGTH OF LINE.

Suppose a transmitter, operating on a frequency of 30 Mc/s, corresponding to a wavelength of 10 metres, is connected to a short length of open circuited transmission line. If the line is exactly $32\frac{1}{2}$ feet long (10 metres) it is acting as a one wavelength section of transmission line, and the voltage standing wave has the distribution shown in Fig. 2. It does so only at this frequency, however. If the transmitter frequency is *increased* to 40 Mc/s corresponding to a wavelength of 7·5 metres ($24\frac{1}{2}$ feet), the $32\frac{1}{2}$ feet length of line is now *greater* than one wavelength of the transmitted energy. This is shown in Fig. 3(b). Similarly, if the frequency of the transmitter is *decreased* to 24 Mc/s the wavelength increases to 12·5 metres (40·6 feet) and the $32\frac{1}{2}$ feet length of line is now *less* than one wavelength (Fig. 3 (c)).

Thus it is usually more convenient to define the length of a section of transmission line in terms of its 'electrical' length; that is, to state that at a given frequency, a section of line is one wavelength long, a half wavelength long, a quarter wavelength long, greater or less than a quarter wavelength long, and so on. The electrical length varies with frequency as described.

5. Although current has not yet been mentioned, a standing wave of current is produced in the same way as a standing

wave of voltage. The incident current travels along the line from the source and at the open end (since no energy can be dissipated), complete reflection occurs and a reflected wave of current travels back along the line to the source. The combination of incident and reflected waves produces a standing wave of current. At the open end, the standing wave of current is zero (no current can flow through an open circuit), while at a distance $\frac{\lambda}{4}$ back from the termination, the current has a maximum value as shown in Fig. 4.

INPUT

OPEN CIRCUITED TRANSMISSION LINE

λ

CURRENT

λ $\frac{3\lambda}{4}$ $\frac{\lambda}{2}$ $\frac{\lambda}{4}$
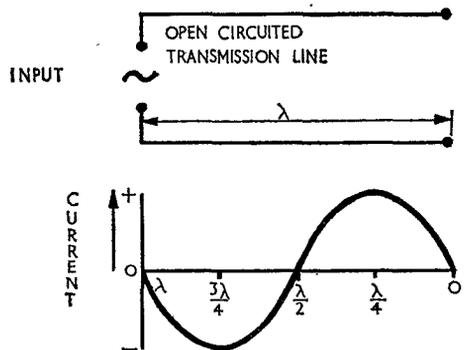
Fig. 4.   CURRENT STANDING WAVE ON AN OPEN-CIRCUITED TRANSMISSION LINE.

6. If the voltage standing wave for an open circuited line is now superimposed on the current standing wave, as in Fig. 5(b) the important characteristics of the line can be deduced. It can be seen that the voltage and current have a phase difference of 90°, and since impedance is given by the ratio of voltage to current ($Z = \frac{V}{I}$), the impedance varies along the line. At the input end, V is a maximum and I is zero, and thus, at this point, the impedance is infinite, or in practice, very high, such as is obtained from a *parallel* tuned circuit at resonance. The line therefore appears to the source at input 1 as a very high impedance at this frequency.

7. If the input is now moved nearer the termination, to input 2 (Fig. 5(a)), the ratio of V to I at this point is zero, and so the impedance presented to the source is zero, or in practice, very low; such an impedance is associated with a series tuned circuit at resonance. This is shown in Fig. 5(c) at the
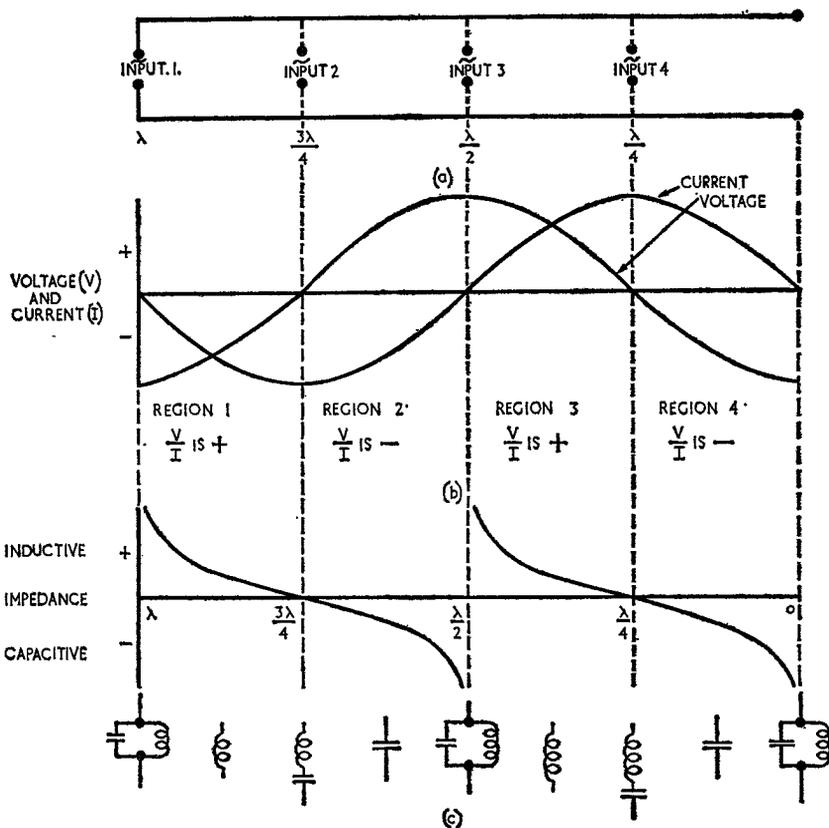
Fig. 5.   OPEN-CIRCUITED TRANSMISSION LINE IMPEDANCE.

$\frac{3\lambda}{4}$ point on the line.   In region 1, between inputs 1 and 2, both I and V are negative, and so the ratio $\frac{V}{I}$ is positive, indicating an *inductive* reactance.   The magnitude of this reactance varies as shown in region 1 of Fig. 5(c).

If the source is now moved to input 3, $\frac{\lambda}{2}$ from the open circuited termination, the conditions are again, voltage maximum and current zero, giving another point of high impedance corresponding to a parallel tuned circuit at resonance.   In region 2, V is positive and I negative, giving a negative reactance; this indicates a *capacitance* reactance which varies in magnitude as shown in region 2 of Fig. 5(c).

In regions 3 and 4 the above sequence of impedance variation is repeated, being inductive in region 3, falling in magnitude to zero (or in practice a very low value) $\frac{\lambda}{4}$ from the termination, and rising through capacitive reactance to a very high value at the open circuited termination (region 4).

Thus the impedance presented to the source of energy, depends on the *electrical* distance the source is from the termination and varies as shown in the impedance/ electrical distance graph of Fig. 5(c).

Short Length of Short-Circuited Line

8.   When a transmission line is terminated in a short circuit as in Fig. 6(a), total energy reflection from the termination occurs, and standing waves of voltage and current are

set up along the line. At the termination, the standing wave of voltage is zero, since no voltage can exist across a short circuit, while the current standing wave is maximum. Moving back from the termination, the voltage standing wave increases in amplitude and the current standing wave falls, until, $\frac{\lambda}{4}$ from the short circuit, the current is zero and the voltage maximum. As with the open circuited line, this pattern is repeated down the line, the voltage and current having a phase difference of 90°. However, relative to the open circuited line, the standing waves on the short circuited line have been displaced by $\frac{\lambda}{4}$, and at the source, the ratio of voltage to current (impedance) is zero.

9. By superimposing the voltage and current standing waves of Fig. 6(b) and (c), Fig. 7(a) is obtained, the electrical length again being one wavelength of the energy being propagated. Fig. 7(b) shows the ratio of V to I (impedance) throughout this length of line.
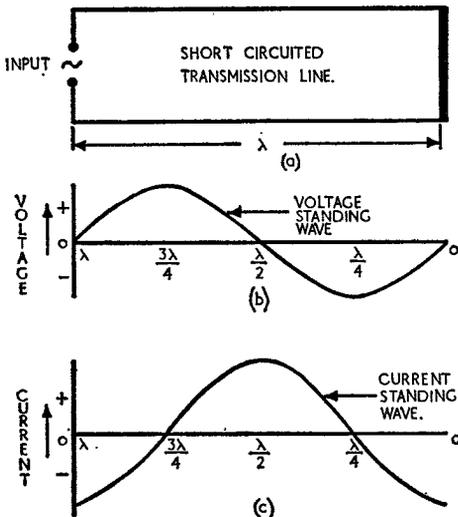
Fig. 6. VOLTAGE AND CURRENT STANDING WAVES ON A SHORT-CIRCUITED TRANSMISSION LINE ONE WAVELENGTH LONG.

At the termination, $\frac{\lambda}{2}$ back from the termination, and at the source (one wavelength back from the termination) voltage is zero while current is maximum; the impedance at these points is thus zero, or in

practice, a very low value associated with a *series* tuned circuit at resonance.

At distances $\frac{\lambda}{4}$ and $\frac{3\lambda}{4}$ from the termination, voltage is a maximum while current is zero; the impedance at these points is thus very high as is obtained from a parallel tuned circuit at resonance.

In regions 1 and 3, since the phase difference between V and I is 90°, and since the impedance has a negative sign, the line behaves as a capacitor. In regions 2 and 4, the impedance has a positive sign and the line behaves as an inductor. The magnitude of the inductive and capacitive reactances throughout the length of line, are shown in Fig. 7(b).

Thus the impedance the source 'sees' depends on the electrical length of the line and the type of termination. Comparison of Fig. 7 with Fig. 5 shows that the impedance points along the line are displaced by a quarter wavelength.

Comparison of Open and Short-Circuited Lines

10. Short lengths of open-circuited and short-circuited lines behave electrically in similar ways. In both cases the impedance varies throughout the length, from a low or a high value, through inductive or capacitive reactance, to a high or low value, and this is repeated over succeeding half wavelengths. The electrical length that gives these values and types of impedances depends on the type of termination, and Fig. 8 compares the characteristics of open-circuited and short-circuited lines of various lengths.

Practical Uses of Short Lengths of Line

11. It will now be appreciated that a short length of line, terminated in an open circuit or short circuit, is a most versatile and useful component. It can act as an inductor, capacitor, rejector circuit or acceptor circuit, depending on its electrical length and the type of termination.

It has already been stated that if energy is not to be reflected, a transmission line must be terminated by a resistance impedance equal in value to the Z_0 of the line. In many cases aerials are partially reactive, and since the reactive component cannot dissipate energy, reflection from the aerial back down the transmission line to the transmitter,

would occur. However, by means of a short transmission line, aerial reactance can be cancelled, leaving only the resistive component, and no energy reflection occurs.

This is only one example of the many uses of a short section of transmission line, and will be dealt with more fully when 'matching stubs' are discussed. Other uses are given in the succeeding paragraphs.

Quarter-Wave Stubs

12. At the beginning of this Chapter, it was stated that a transmission line could be supported and spaced by means of metal spacers, known as quarter-wave stubs (sometimes 'metallic insulators'). A quarter wave stub is a short-circuited transmission line quarter of a wavelength long as shown in Fig. 9. R.F. energy, moving down the main transmission line, enters the $\frac{\lambda}{4}$ short-circuited stub and standing waves are set up on the stub. This is shown in Fig. 9(c).

At the end remote from the short circuit, where the stub joins the main transmission line, the voltage is a maximum and the current zero. Thus the main transmission line sees a very high impedance, and since this is in parallel with the Z_0 of the line, the stub has no electrical effect on the r.f. energy being propagated down the main transmission line. Therefore very little further r.f. energy enters the stub. Further, since the voltage at the short-circuited end of the stub is zero, this point can be earthed without affecting the stub action.

This is a very efficient method of supporting and spacing a twin wire feeder; much more efficient than conventional insulation spacers. However, if the frequency of the transmitter changes, the stub is no longer $\frac{\lambda}{4}$ long and
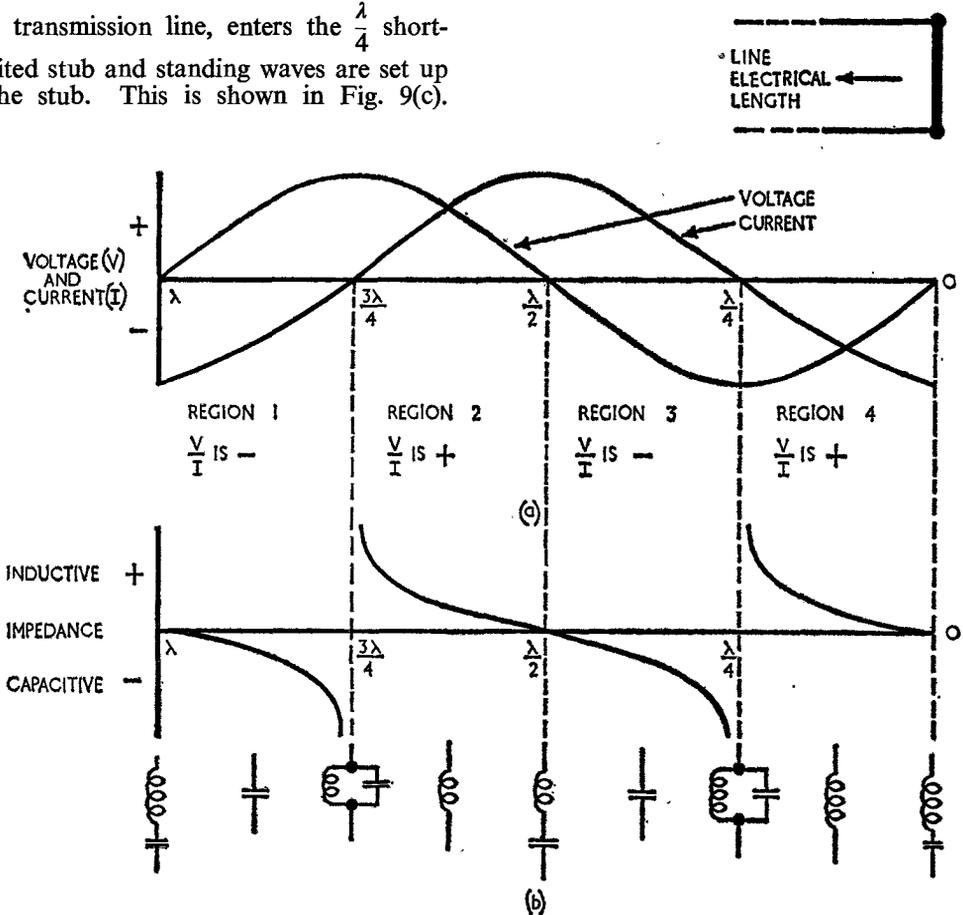
Fig. 7. SHORT-CIRCUITED TRANSMISSION LINE IMPEDANCE.

ELECTRICAL LENGTH OF LINE	OPEN-CIRCUITED TERMINATION	SHORT-CIRCUITED TERMINATION
LESS THAN $\frac{\lambda}{4}$	CAPACITIVE REACTANCE	INDUCTIVE REACTANCE
$\frac{\lambda}{4}$	LOW IMPEDANCE ACCEPTOR CIRCUIT	HIGH IMPEDANCE REJECTOR CIRCUIT
BETWEEN $\frac{\lambda}{4}$ AND $\frac{\lambda}{2}$	INDUCTIVE REACTANCE	CAPACITIVE REACTANCE
$\frac{\lambda}{2}$	HIGH IMPEDANCE REJECTOR CIRCUIT	LOW IMPEDANCE ACCEPTOR CIRCUIT
BETWEEN $\frac{\lambda}{2}$ AND $3\frac{\lambda}{4}$	CAPACITIVE REACTANCE	INDUCTIVE REACTANCE
$3\frac{\lambda}{4}$	LOW IMPEDANCE ACCEPTOR CIRCUIT	HIGH IMPEDANCE REJECTOR CIRCUIT
BETWEEN $3\frac{\lambda}{4}$ AND λ	INDUCTIVE REACTANCE	CAPACITIVE REACTANCE
λ	HIGH IMPEDANCE REJECTOR CIRCUIT	LOW IMPEDANCE ACCEPTOR CIRCUIT
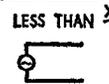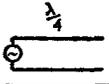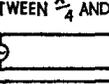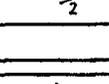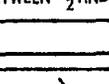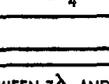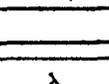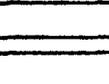
Fig. 8. BEHAVIOUR OF SHORT LENGTHS OF LINE.

a reactance, either inductive or capacitive, is placed across the main transmission line. This causes energy to be reflected and standing waves to be set upon the main transmission line with a consequent increase in losses.

Quarter-Wave Matching Transformer

13. When the load is a pure resistance but is not equal in value to the characteristic impedance Z_0 of the line, standing waves are set up on the line and a quarter wave section of line acting as a transformer can be inserted between the line and the load to achieve matching. A quarter wave matching transformer is useful only when the load is purely *resistive*.

At low frequencies the input impedance of a source can be matched into its load by means of a conventional transformer as shown in Fig. 10(a); the turns ratio of the transformer is adjusted according to the values of Z_0 and R. At r.f. the same principle is employed, but the transformer takes the form of two parallel conductors a specific distance (D) apart, and mounted between the transmission line and the aerial as shown in Fig. 10(b). To achieve the required match between the Z_0 of the main transmission line and the input impedance of the aerial the impedance of the $\frac{\lambda}{4}$ transformer must be such that $Z_c = \sqrt{Z_0 \times Z_1}$. This is accomplished by varying the diameter of the conductors or, more commonly, by varying their distance D apart, such that the standing wave ratio on the main transmission line is as close to unity as is possible. EXAMPLE: It is required to match a 320 ohms open wire feeder into a dipole aerial of input impedance 80 ohms (resistive). The

cancelled out, the effective impedance at that point would be purely resistive and equal to the Z_0 of the line. The length of line between the transmitter and that point would then be correctly terminated, and no standing waves would exist on it. This is done in practice by mounting a reactive stub at that point.

MAIN TRANSMISSION LINE
TO AERIAL
FROM TRANSMITTER
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
QUARTER WAVE STUBS
(a)

$\frac{\lambda}{4}$
INPUT
QUARTER WAVE STUB
(b)

V. OR I.
$\frac{\lambda}{4}$
VOLTAGE
CURRENT
(c)

Fig. 9. QUARTER-WAVE STUB OR METALLIC INSULATOR.

LINE IMPEDANCE $=Z_o$
LOAD RESISTANCE$=R$
LOW FREQUENCY MATCHING
MATCHING TRANSFORMER
(a)

$\frac{\lambda}{4}$
AERIAL
TRANSMISSION LINE IMPEDANCE $=Z_o$
Z_c
D
AERIAL INPUT IMPEDANCE$=Z_i$
HIGH FREQUENCY MATCHING

MAXIMUM ENERGY TRANSFERRED TO AERIAL WHEN $Z_c = \sqrt{Z_o Z_i}$
(b)

Fig. 10. QUARTER-WAVE MATCHING TRANSFORMER.

impedance of the required $\frac{\lambda}{4}$ transformer is given by:—

$$Z_c = \sqrt{Z_0 \times Z_1} \text{ ohms}$$
$$Z_c = \sqrt{320 \times 80} \text{ ohms}$$
$$= \sqrt{25600} \text{ ohms}$$
$$= 160 \text{ ohms.}$$

Thus the distance D would be adjusted accordingly.

Reactive Matching Stubs

14.   A line terminated in a reactive impedance, or in a resistive impedance not equal to the Z_0 of the line, gives rise to standing waves. The impedance along the line then varies from point to point. There are, however, always two points in any half wavelength of line, where the impedance is equivalent to a pure resistance, equal in value to the Z_0 of the line and shunted by a reactance (see Fig. 11(a)). If the reactive component of such a point (Y or Z) could be

When mounting such a stub, the things that have to be decided are:—

(a) The distance d_1 at which the stub is to be inserted.

(b) Whether the reactance at this point is inductive or capacitive.

(c) The exact length d_2 of the stub.

15.   These factors can be found by trial and error, using the SWR on the main line as a guide. The stub should be mounted as near the aerial as possible, so that the length of line between the stub and aerial on which standing waves will still exist, is short. With the SW indicator, the point of maximum standing current nearest to the aerial is found (point X) Fig. 11(a). The impedance between X and Y is resistive with a capacitive component, and so an inductive stub, less than $\frac{\lambda}{4}$ and short circuited, could
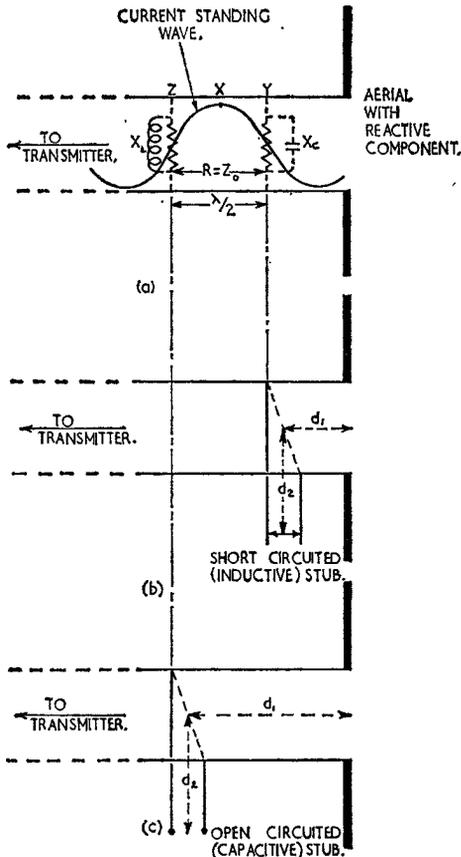
Fig. II.   TRANSMISSION LINE MATCHING
POINTS.

be inserted between these points.   Between X and Z the impedance is resistive with an inductive component, and an open circuited stub, less than $\frac{\lambda}{4}$ could be mounted between these points.   In this way the reactive component can be cancelled out by connecting a reactive component of *opposite sign* in parallel with the line.

The exact distance of the stub from the load (d_1) and the length of the stub (d_2), can be determined by trial and error, until the SWR on the transmitter side of the stub is as near as possible to unity.   A short circuited stub (Fig. 11(b)) is preferable to an open circuited stub, since it is mechanically more robust, and its effective length is easier to adjust.   Should, however, the point of maximum current be too near the aerial to use an inductive stub, an open circuited stub (Fig. 11(c)) is used.

16.   If the impedance of the aerial is widely different from the Z_0 of the transmission line, the standing waves on the stub and on the short section of line between stub and aerial, may be very large, and the losses high. Under these circumstances it is necessary to introduce a $\frac{\lambda}{4}$ matching transformer between the aerial and the line, before inserting the stub, as shown in Fig. 12.   A good practical example of this application is discussed in Section 16.
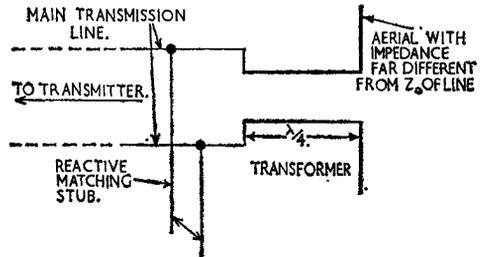
Fig. 12.   USE OF A $\frac{\lambda}{4}$ TRANSFORMER AND
REACTIVE STUB.

Balanced to Unbalanced Matching

17.   When discussing the types of transmission line (Chapter 3, Paras. 32–34) it was stated that an open-wire feeder was a 'balanced' line, both conductors having the same impedance to earth.   A coaxial feeder on the other hand, has the outer conductor earthed, and is therefore an 'unbalanced' line.   If an open-wire balanced line is connected to an aerial that is not balanced about earth, then the line itself becomes unbalanced and unequal currents flow in each wire.   The same applies if the transmitter output stage is not symmetrically balanced about earth.   If an open-wire feeder is not balanced it will radiate and losses increase.   To prevent this an open-wire feeder must be terminated at the load end by an aerial, such as a centre-fed $\frac{\lambda}{2}$ dipole that is balanced about earth, and at the transmitter end, by a circuit that is symmetrical about earth.   This is illustrated in Fig. 13(a).

Similarly an unbalanced coaxial feeder must be terminated at the load end by an unbalanced aerial, such as a $\frac{\lambda}{4}$ Marconi, and the transmitter output stage must have one end earthed (Fig. 13(b)).
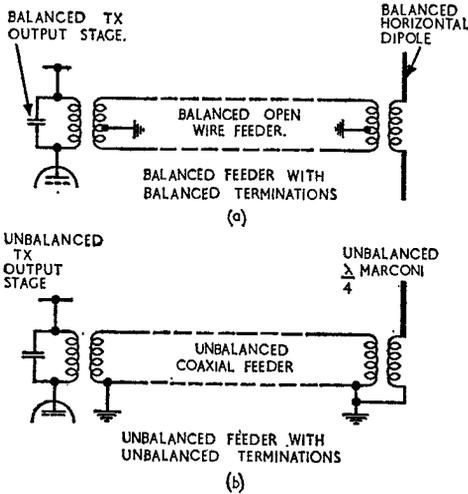
Fig. 13. BALANCED/UNBALANCED FEEDER
WITH BALANCED/UNBALANCED
TERMINATIONS.

18. When it is necessary to feed an unbalanced load, such as a $\frac{\lambda}{4}$ Marconi aerial with a balanced open wire feeder, or a balanced $\frac{\lambda}{2}$ dipole with an unbalanced coaxial feeder, a balanced to unbalanced matching device must be used; typical of these are the $\frac{\lambda}{2}$ phasing loop, and the balun (or Bazooka).

$\frac{\lambda}{2}$ Phasing Loop

19. A balanced open-wire line can be connected to an unbalanced load such as a $\frac{\lambda}{4}$ aerial by using a phasing loop as shown in Fig. 14. The phasing loop consists of a half wavelength of line folded back on itself to prevent radiation, and connected as shown. The voltage standing wave at points on a wire half a wavelength apart, are in antiphase. Thus if the voltage at point A is + V volts with respect to earth, that at B is − V volts with respect to earth and AB presents a balanced load to the line. Connecting the aerial at point A does not appreciably upset this balance. The effective earth is a point midway between A and B.

In addition to balancing the load about earth, the phasing loop also provides a

matching device between line and load. Since the voltage at A with respect to earth is + V volts and that at B is − V volts, then the voltage across AB is 2V volts. But the power supplied to the unbalanced aerial
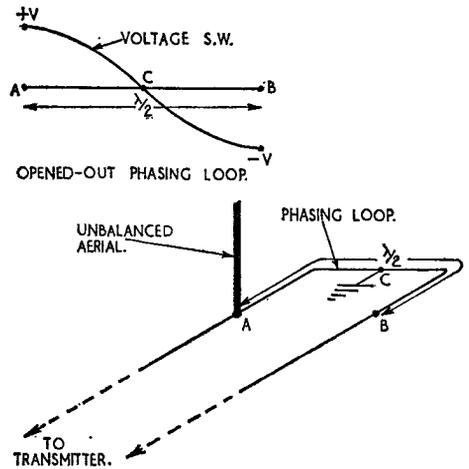
Fig. 14. USE OF PHASING LOOP IN OPEN
WIRE LINE.

of impedance Z_1, equals the power supplied by the balanced feeder of impedance Z_0. Therefore:—

$$\frac{V^2}{Z_1} = \frac{(2V)^2}{Z_0}$$
$$\text{and } Z_0 = 4Z_1$$

Thus a $\frac{\lambda}{2}$ phasing loop gives a 4:1 impedance match as well as balancing about earth. It could be used, for example, to match and balance a 320 ohms open-wire feeder to a 80 ohms unbalanced aerial.

20. The $\frac{\lambda}{2}$ phasing loop can also be used to match and balance a coaxial feeder to a balanced centre-fed $\frac{\lambda}{2}$ dipole as shown in Fig. 15. The action of the phasing loop is similar to that already described for the balanced feeder supplying the unbalanced aerial. The aerial in this case however, is balanced and the line unbalanced; the phasing loop enables the aerial to 'see' a balanced line across AB. Balanced to unbalanced matching is thus achieved. The impedance transformation is also reversed, being 4:1 step up from feeder to aerial.

In practice a $\frac{\lambda}{4}$ matching transformer and a reactive matching stub would be required to reduce the SWR on the line to acceptable limits. These would be inserted as previously explained.
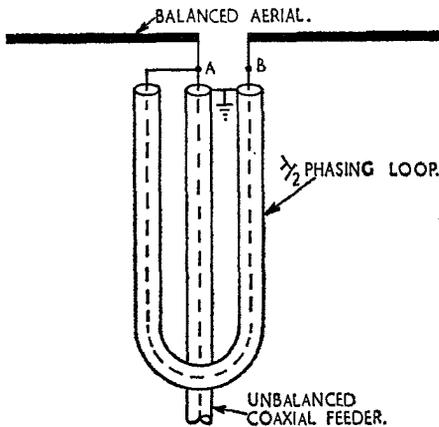
Fig. 15. COAXIAL PHASING LOOP.

The Balun

21. At higher frequencies a device known as a balun is employed to provide balanced to unbalanced matching between coaxial feeder and balanced aerial. This is shown in Fig. 16.

The balun consists of a metal cylinder, surrounding the last quarter wavelength of line, and connected to the outer conductor of the coaxial cable $\frac{\lambda}{4}$ from the termination
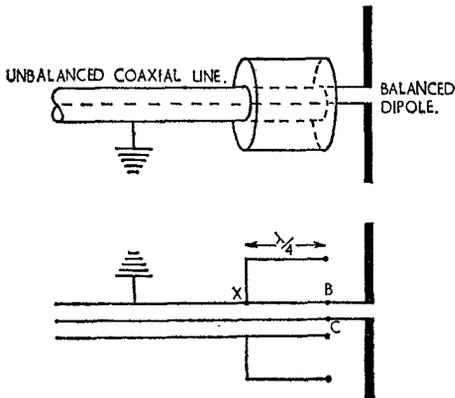
(at point X). The outer surface of the last $\frac{\lambda}{4}$ of the outer conductor and the inner surface of the balun, and the piece which joins the balun to the line, together form a short-circuited quarter-wave coaxial stub. Thus the short-circuit at X (Fig. 16) is reflected over quarter of a wavelength to give a high impedance at the open end of the balun. The impedance between the outer conductor of the coaxial cable at B and earth (point X) is therefore very high so that the outer conductor of the coaxial cable at B is effectively isolated from earth. The inner conductor C of the coaxial cable is also isolated from earth, so that B and C have practically the same impedance to earth. Thus the aerial may be connected to the line at points B and C without disturbing the balanced condition of the load. Unlike the phasing loop, the balun gives no impedance transformation.

It should be noted that matching stubs and $\frac{\lambda}{4}$ matching transformers will still be required on the lines using phasing loops and baluns; the primary function of the latter is to provide balanced to unbalanced matching.

Lecher Bars

22. Another important application of a short section of transmission line is the lecher bar. As has already been noted, a short circuited section of line, presents

Fig. 17. LECHER BAR.

Fig. 16. THE BALUN.
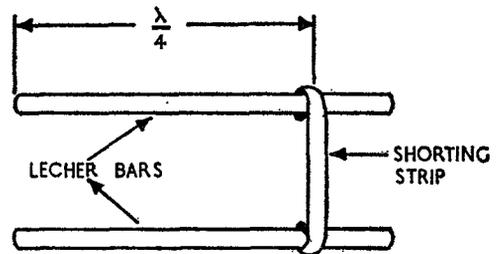
a very high impedance, equivalent to a parallel tuned circuit at resonance, quarter of a wavelength from the short circuited end. Thus at uhf, when the physical length of such a line will be short, it can be used in place of the conventional tuned circuit in a valve oscillator.

In practice lecher bars may take the form of hollow, silver-plated tubes with a movable shorting bar, which varies the electrical length, and thus the frequency at which the lecher bar will resonate (Fig. 17). Since the resistance of the bars will be small due to their short length and silver plating, a magnification factor (Q) of the order of 10,000 is easily obtained. Oscillator circuits using lecher bars are frequently employed in wireless and radar equipments working in the uhf band.

Summary

23. This section has dealt with the theory of transmission lines and some of the uses to which transmission lines, and short sections of transmission lines, may be put. Many further applications of the latter will be encountered in wireless and radar equipments, but with a sound understanding of the basic principles explained in this section, no difficulty should be experienced in understanding these applications.

SECTION 16

AERIALS

SECTION 16

AERIALS

SECTION 16

CHAPTER 1

BASIC PRINCIPLES OF RADIATION

BASIC PRINCIPLES OF RADIATION

Introduction

1. Previous Sections of these Notes have dealt with the production of electromagnetic energy in the transmitter and the reception of this energy from the input to the receiver. Section 15 considered the transfer of the energy from the source to the load, and the transmitter aerial, which forms the load, was mentioned. In this Section the aerial itself will be considered in greater detail, basic ideas on how radiation from an aerial occurs will be given, and the properties and characteristics of some simple aerials will be discussed. Finally a selection of some of the more common aerial systems in use will be reviewed.

2. The aerial is an essential part of a communications system. No matter how efficient the transmitter and receiver may be, without an equally efficient aerial system, maximum range and quality of reception cannot be expected.

3. An aerial may be defined as a device for the efficient transmission and reception of e.m. energy. In this Section aerials will be considered in the main as radiating elements, but the properties of a transmitting aerial apply equally well to a receiving aerial. In fact many installations use, in conjunction with necessary switching devices, a common aerial for transmission and reception.

The Basic Half-wave Dipole

4. The simplest form of aerial and one often used in practice, is a straight wire conductor with a generator of alternating current placed at the centre (Fig. 1). If the aerial is transformer coupled to the generator the e.m.f. induced in the secondary circuit acts in series with the aerial, and so the generator is still effectively in series with the aerial. The aerial wire has inductance throughout its length and capacitance is indicated by the dotted lines at X and Y in Fig. 1.

5. The generator voltage will now be followed throughout one complete cycle and the charge distribution on the aerial over the same period will be noted.

At time t_1 in Fig. 1(a), when the generator voltage is maximum, the aerial capacitor is fully charged and no current is flowing in the wire. As the voltage falls the capacitor discharges and current builds up to a maximum at t_2 (Fig. 1(b)), when the generator voltage is zero. This current re-distributes the electrons in the wire and a quarter of a cycle later, at time t_3 end X will be positively charged and end Y negatively charged as shown in Fig. 1(c). At this instant the voltage is maximum and the current zero. A further quarter of a cycle later (t_4) the generator voltage has fallen to zero and the current is again maximum but in the opposite direction. This is shown in Fig. 1(d).
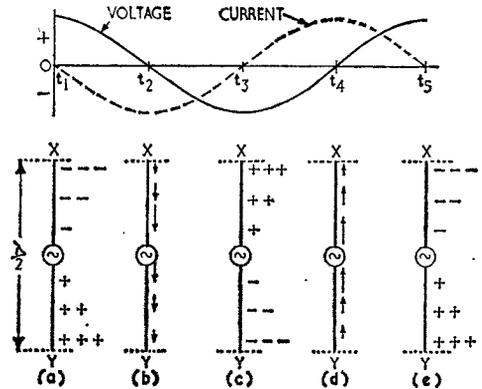
Fig. I. PRINCIPLE OF THE HALF-WAVE DIPOLE.

There is a natural tendency for the electrons at Y to move back towards X and as the generator voltage is changing polarity, this movement of electrons is assisted by the generator and end Y becomes positive and end X negative. This is shown in Fig. 1(e) at t_5. For this change of electron distribution to occur between time t_3 and t_5 the electrons must leave end Y and arrive at end X half a cycle later. If the length of the wire from X to Y is half a wavelength at the frequency of the exciting voltage, the electrons will take half a period to move from one end of the wire to the other end. The charging and discharging of the aerial capacitor will then be synchronised with the frequency of the applied voltage. The wire will be resonant at the

generator frequency and is an open oscillatory circuit. Note that throughout the cycle the current and voltage are 90° out of phase.

This is the principle of the half wave dipole and is dealt with in more detail in Chapter 2.

Electrical Length of Aerial

6. The electrons in the conductor XY will have a velocity less than that in free space because of the resistance and reactance of the wire. Thus to enable the wire to be resonant at the generator frequency its *physical* length must be less than half a wavelength of the exciting energy. If a correction factor is introduced, the *electrical* length of the wire can be made equal to half a wavelength.

For half wave dipoles the physical length should be 0·48λ.

Elementary Concept of Radiation

7. A moving electron produces a moving electric field, and since a moving electron is in fact a current, a magnetic field is also produced. In a resistive circuit these two fields are in phase, but at right angles to each other. With an alternating current the electrons are continually changing their velocity, going from zero to maximum speed in one direction, back to zero and to maximum speed in the opposite direction.

8. Thus the electrons surging up and down the wire of the half wave dipole cause electric lines of force as shown in Fig. 2(a). The direction of this E field is, by convention,
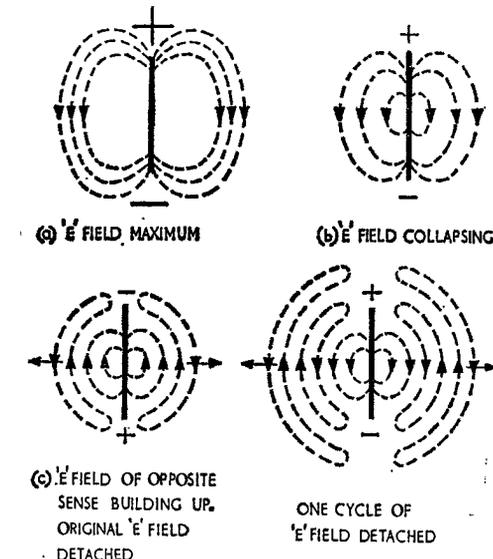
from positive to negative. Since the potential across the aerial is alternating the E field will fall from maximum intensity in one sense, through zero to maximum intensity in the opposite sense. However, the E field is spread some distance away from the wire and so its collapse will lag on the current which is causing it. Thus when the current changes direction and a new E field of opposite sense begins to build up, the original E field will still be present. This is illustrated in Fig. 2(c); the first E field is forced outwards in the form of a closed loop because the two fields repel each other. The detached E field moves away from the aerial in the direction shown, with the speed of light (3 × 10⁸ metres per second).

9. A changing or moving E field causes a magnetic field and so the energy radiated from the aerial is in the form of electric (E) and magnetic (H) fields *in phase* with each other but in planes at right angles to each other (in space quadrature).

10. The diagrams of Fig. 2 are sectional in the vertical plane and show the E field only. Radiation takes place in three dimensions and it would be more correct to visualise an expanding sphere with the aerial at the centre; as the radius of the sphere increases other concentric spheres are produced at regular intervals. This is illustrated in Fig. 3.
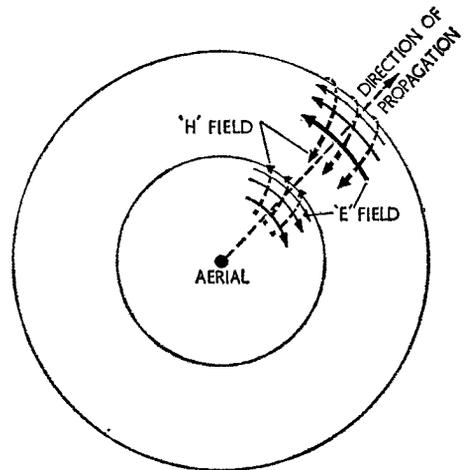
Fig. 3. CONCEPT OF THE RADIATED FIELD.

(a) 'E' FIELD MAXIMUM   (b) 'E' FIELD COLLAPSING

(c) 'E' FIELD OF OPPOSITE SENSE BUILDING UP. ORIGINAL 'E' FIELD DETACHED

ONE CYCLE OF 'E' FIELD DETACHED

Fig. 2. ELECTRO-MAGNETIC RADIATION.

Nature of the Radiated Energy

11. The energy leaves the aerial in the form of electric and magnetic fields at right angles to each other and at right angles, or transverse

to the direction of propagation. Because of this, electro-magnetic waves are called *transverse* waves; this distinguishes them from sound waves which are longitudinal. The relationship between the senses of the E and H fields and the direction of propagation is shown in Fig. 4(a). This relationship can be remembered by considering the E field to be the handle of a corkscrew. If the E field is rotated towards the H field then the direction of propagation is given by the direction in which the tip of the corkscrew moves.

If the direction of propagation is reversed either the E field or the H field is reversed, but not both. Usually it is the E field which reverses in direction, or rather changes its phase by 180° at a reflecting surface, the H field being unaltered. (Fig. 4(b)).

wavelength. As the transmitter is continually exciting the aerial successive cycles are produced and the wave front at A moves further away from the aerial with the speed of light, followed by successive cycles. Fig. 5(b) and (c) illustrate this, Fig. 5(b) being a vertical section and 5(c) a schematic representation of the E and H fields.
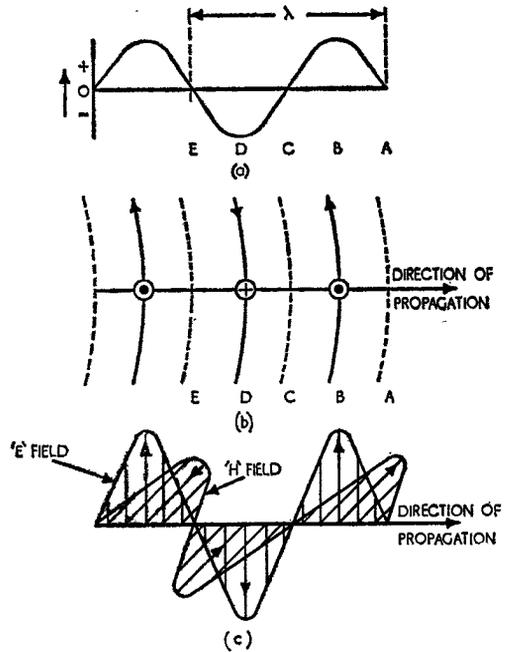
Fig. 5. COMPOSITION OF THE RADIATED FIELD.

Fig. 4. COMPOSITION OF THE PLANE WAVE.

12. **The plane wave.** The E and H lines as drawn in Figs. 2 and 3 are curved when they leave the aerial and form the surface of a sphere. This curvature may be quite pronounced close to the aerial but at distances away from the source the curvature is less noticeable and the wave can be considered as occupying a single plane.

13. The voltage and current on the aerial vary sinusoidally and so the radiated fields will vary in the same way. At point B in Fig. 5 the E field is maximum in the sense shown by the arrow. From B to C the field intensities decrease to zero and rise again from C to D to a maximum in the opposite sense to that at B, falling again to zero at E and so on. The wave has completed one cycle of oscillation between A and E and the distance AE is one

Plane of Polarisation

14. When the electric lines of force are in the vertical plane the wave is said to be *vertically polarised*. When the E lines are in the horizontal plane the wave is *horizontally polarised*. Thus the plane containing the E field and the direction of propagation is called the *plane of polarisation*.

In the case of a half wave dipole aerial since the E lines are parallel to the aerial, the aerial is parallel to the plane of polarisation. This is not true for all types of aerial however, and it must be remembered that it is the plane of the E field which defines the polarisation of the wave. If a vertical transmitter aerial radiates a vertically polarised wave then the aerial designed to receive the wave must also be vertical, i.e. the transmitter and receiver aerials must be parallel.

15. **Circular and elliptical polarisation.** In a plane polarised wave the plane through the E field and the direction of propagation is constant as shown in Fig. 6(a). If the direction of the E field (and of the H field) rotates as the wave progresses, and the amplitude of the fields remains constant, the tip of the E field vector will trace out a circular path (Fig. 6(b)). This wave is said to be *circularly polarised.*

If the plane of polarisation is rotating and the amplitude of the E vectors varies sinusoidally the tip of the E vector will trace an ellipse as shown in Fig. 6(c). The wave is said to be *elliptically polarised.*

Plane polarisation is usually employed for communications but circular and elliptical polarisation is used in some special equipments.

PLANE POLARIZATION.
E FIELD CONSTANT.
IN DIRECTION

CIRCULAR POLARISATION.
E FIELD CONSTANT. IN
AMPLITUDE BUT ROTATING
IN DIRECTION.

ELLIPTICAL POLARISATION.
E FIELD ROTATING IN DIRECTION.
AND VARYING SINUSOIDALLY IN
AMPLITUDE.
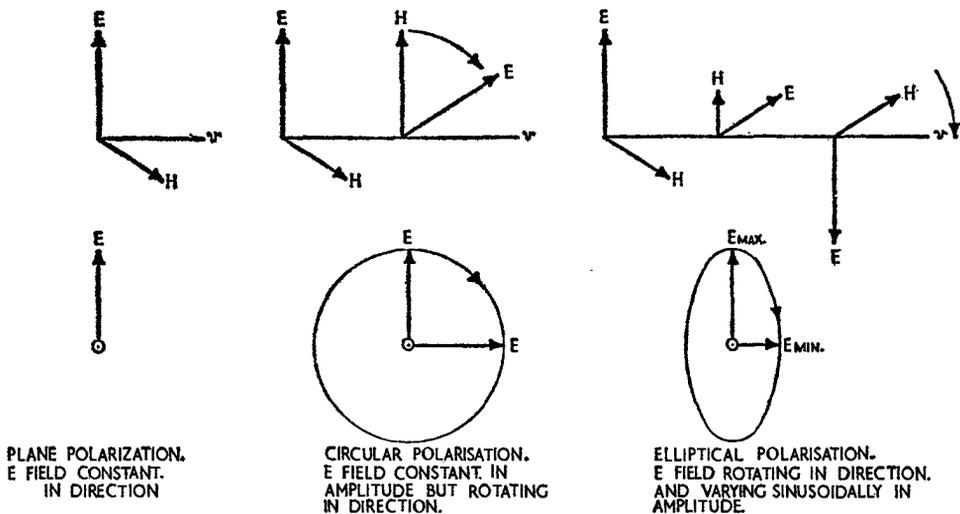
Fig. 6. CIRCULAR AND ELLIPTICAL POLARISATION.

SECTION 16

CHAPTER 2

SIMPLE RESONANT AERIALS

SIMPLE RESONANT AERIALS

Introduction

1. In dealing with the elementary principles of radiation in Chapter 1, the simple half-wave dipole was used to illustrate these principles. The half-wave dipole is a *resonant* or *standing wave* aerial. *Non-resonant* or *travelling wave* aerials are also used, and will be considered in Chapter 4; this chapter will consider some simple resonant aerials and their important features. The three simple resonant aerials which will be dealt with are the half-wave dipole, the Marconi quarter-wave aerial and the slot aerial, all in common use in ground and air radio installations.

Standing Wave or Resonant Aerials

2. An aerial of this type is basically a series resonant circuit in which electrons surge backwards and forwards along the wire which forms the aerial. The electrons can be considered as effectively forming one large charge of electricity which is continuously being accelerated and decelerated by the exciting voltage of the generator. The result is that radiation occurs; maximum radiation takes place perpendicular to the aerial and zero from the ends.

The Half-wave Dipole—Production of Standing Waves

3. A half-wave dipole consists of a wire half a wavelength long at the frequency considered, with an alternating generator at its centre. It is shown in Fig. 1. When terminal A is negative, a travelling wave of current, accompanied by a travelling wave of voltage moves from A towards X. When the current reaches X it suffers a reversal of phase and is reflected back towards the generator. When it arrives at the generator after having travelled half a wavelength, it will be in phase with the incident wave of current. The voltage wave travels the same distance but is reflected without a change of phase and so arrives back at the generator in anti-phase to the incident voltage. Similar conditions apply for waves travelling from B to Y.
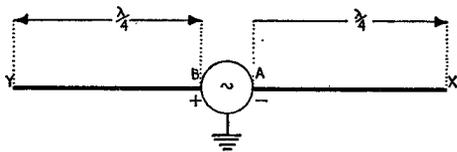
4. The incident and reflected waves travelling along the aerial combine at each instant throughout the cycle to form resultant voltage and current standing waves (Section 15 Chapter 3). Fig. 2 illustrates the build-up and collapse of the standing waves at ⅛ of a cycle intervals over half a period. The current standing wave is a maximum at the centre, forming a current *antinode* and zero at the ends forming current *nodes* while the voltage standing wave has antinodes at the ends and a node at the centre.
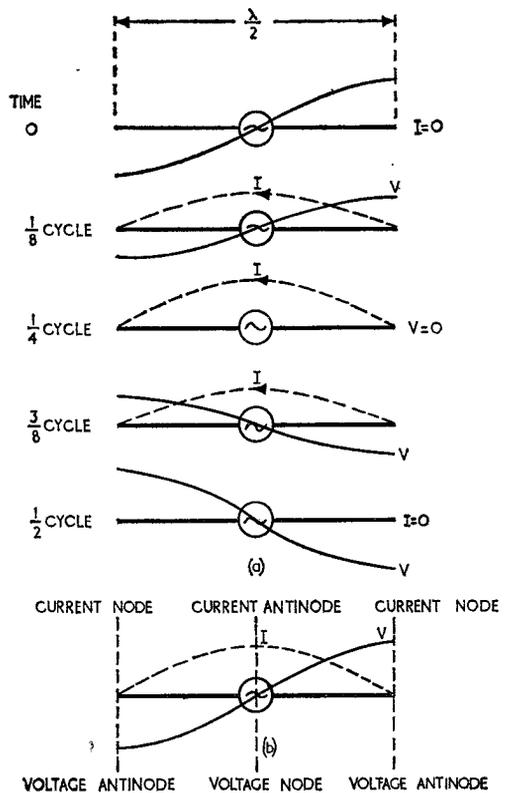
Fig. 2. STANDING WAVES ON A HALF-WAVE DIPOLE.

Fig. 1. PRINCIPLE OF THE HALF-WAVE DIPOLE.

L

It has already been mentioned that a resonant aerial is an open series resonant circuit. The voltage at the ends of the aerial is equivalent to the voltage across the capacitor, and is many times the applied voltage. Were there no losses in the aerial, its Q would be infinity and there would be voltage at the ends but zero voltage at the centre. In practice the aerial does suffer losses because it radiates energy, and a relatively small driving voltage is required to maintain standing waves, in the same way that a small voltage is required to overcome the resistance losses in a series resonant circuit.

5. As shown in Fig. 2(a) the voltage and current standing waves on the aerial are 90° out of phase in time, i.e. when voltage is a maximum, current is zero and quarter of a cycle later, when current is a maximum voltage is a minimum. However, when depicting standing waves on an aerial, graphs of the r.m.s. values of voltage and current are shown as illustrated in Fig. 2(b). These waves can readily be measured by using a suitable meter.

Aerial Resistance
6. The purpose of an aerial is to dissipate energy to the surrounding space in the form of electromagnetic radiation. Thus the aerial converts energy from one form to another in the same way that a resistor converts electrical energy into heat energy when a current flows through it. Therefore an aerial can be considered as a fictitious resistor which dissipates energy in the form of e.m. radiation. This effective resistance is known as the *radiation resistance* of the aerial.

7. The r.m.s. current supplied by the generator driving a half-wave dipole, and necessary to overcome the radiation losses, constitutes energy dissipated in the radiation resistance of the aerial. In the case of a half-wave dipole the radiation resistance is approximately constant at 73 ohms.

8. If the aerial was perfect, all the energy supplied to it would be converted into e.m. radiation. However, losses occur and some of the supplied energy is wasted. This constitutes a *loss resistance* which added to the radiation resistance, increases the total aerial resistance. The sources of these losses are summarised as follows.

   (a) **Dielectric losses.** These are due to the energy dissipated in the aerial insulators and nearby objects which are affected by the radiated fields. High quality low-loss insulators have been developed, which reduce these losses to a minimum but to be fully effective the insulators must be kept clean.

   (b) **Brush discharge losses.** These are caused by a discharge due to ionisation of the air near the aerial. Brush discharge is most apparent from high voltage points on the aerial and can be reduced by rounding the ends of the aerial.

   (c) **Copper losses.** The inherent ohmic resistance of the aerial wire converts electrical energy into heat energy, causing a loss of supplied energy. These losses increase with frequency, due to skin effect.

   (d) **Eddy current losses.** Any near-by metal bodies will have currents induced in them by the aerial energy. These currents produce heat, the energy for which has come from the aerial. Thus to reduce these losses, the aerial should be mounted as far away as possible from metal objects.

Aerial Efficiency
9. The efficiency (η) of an aerial is:—

$$\eta = \frac{\text{Power radiated}}{\text{Power supplied}} = \frac{I^2(R_R)}{I^2(R_R + R_{loss})}$$

Where R_R is the radiation resistance of the aerial and R_{loss} the total loss resistance. This formula represents the fraction of the total power supplied to the aerial which is converted into radiated power. For high efficiency, loss resistance must be small compared with radiation resistance. For practical aerials the efficiency can vary from 15% at low frequencies to 90% and above at higher frequencies.

Aerial Impedance
10. It can be seen from Fig. 2(b) that the ratio of voltage to current varies along the length of the aerial. At the centre the current is large and voltage is zero, ideally, and in practice very small, while at the ends the voltage is large and the current zero. At intermediate points the ratio varies between these two extremes. The ratio of voltage to current is impedance and so the impedance of the dipole varies throughout its length, being minimum at the centre and maximum at the ends.

   Thus the impedance that the aerial presents to the generator depends on the point where the generator is connected to the aerial. This

is termed the *input impedance* of the aerial. If the generator is connected at the centre of the aerial it is feeding into a point of low impedance and is said to be *current fed* or *centre fed*. If connected at a point of maximum voltage, the aerial is said to be *voltage fed* or *end fed*. Both these methods of feeding are illustrated in Fig. 3.
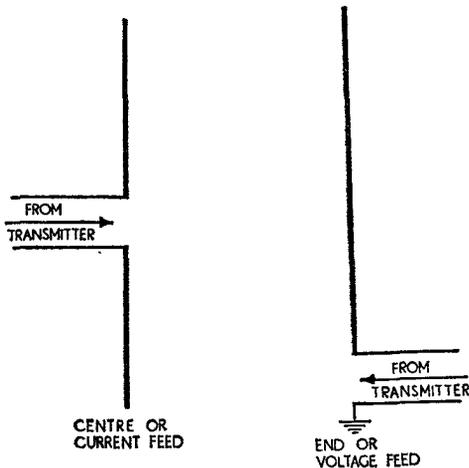
Fig. 3. METHODS OF FEEDING A DIPOLE.

11. The input impedance of a current fed half-wave dipole is 73 ohms while for a voltage fed half-wave dipole it is approximately 2,500 ohms. Thus the input impedance and radiation resistance of a half-wave dipole are of the same value only when the aerial is centre fed.

Aerial Matching

12. In order to achieve maximum transfer of power, the internal resistance of the generator must be matched to the resistance of the aerial at the point of feed. If this is done the generator "sees" a resistance equal in value to its own internal resistance, and the travelling wave reflected from the end of the aerial is absorbed without reflection, by the generator. Maximum power transfer takes place under these conditions.

13. If a transmission line is used to convey the energy from transmitter to aerial the internal resistance of the transmitter must be matched to the characteristic impedance of the line; at the aerial end the characteristic impedance of the line must be matched to the input resistance of the aerial. This often means employing a matching transformer. (See Section 15, Chapter 4).

The Delta Match

14. One method of achieving a correct match between an open wire transmission line of characteristic impedance 600 ohms and the impedance of a dipole is illustrated in Fig. 4. Since the impedance of the dipole increases towards the ends, two points X and Y can be found where the aerial impedance matches that of the twin feeder at the point of connection. As the voltage between X and Y is practically zero, the centre of the aerial can be made continuous. This method of matching is often used and is known as the *delta match*.
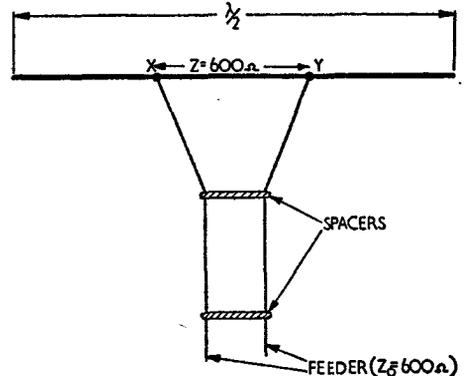
Fig. 4. THE DELTA MATCH.

Because the transmission line is opened out to form the delta, radiation occurs from the delta. This upsets the radiation pattern and reduces the aerial efficiency.

The Folded Dipole

15. In cases where it is necessary to use a transmission line of high characteristic impedance (of the order of 300 ohms) to feed a centre fed half-wave dipole, matching is made easier by using a *folded dipole*. The important feature of this aerial is its high input impedance which is approximately four times that of a simple dipole, i.e. $4 \times 73 = 292$ ohms Thus it presents a suitable load to a 300 ohm feeder.

The current and voltage distribution on an

aerial one wavelength long is shown in Fig. 5(a). When the portion BC is folded over the portion AB the current in these two halves is in the same direction and the current standing waves will be in phase and reinforce each other (Fig. 5(b)).

(a)

(b)
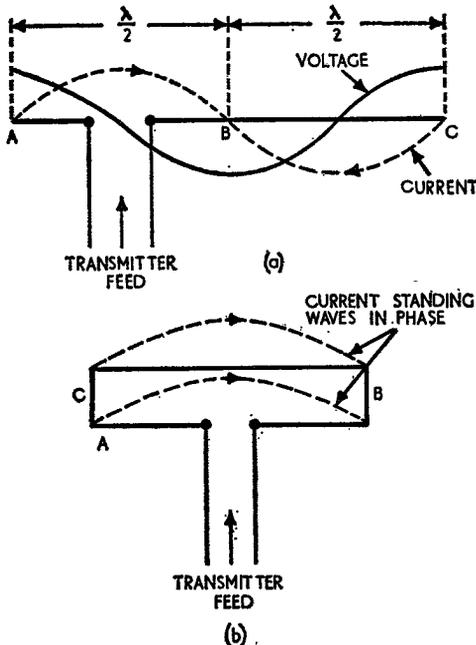
Fig. 5. THE FOLDED DIPOLE.

16. So far resonant aerials half a wavelength long have been discussed. It is possible to have resonant aerials of greater length i.e. λ or $\dfrac{3\lambda}{2}$ or even longer. However, for simplicity, ease of construction and ease of matching, a half-wavelength aerial is by far the most common.

Polar Diagrams

17. A vertical aerial in space does not radiate fields of equal intensity in all directions. It cannot, for example, radiate any energy from the ends. Thus, if a field strength meter was placed directly above or below a vertical aerial it would read zero. If the meter, still in the vertical plane, were moved around the aerial and readings taken at regular intervals, a diagram could be constructed from these readings. An indication of the strength of the field radiated in all directions in the vertical plane could thus be obtained. The same could be done for the horizontal plane and another diagram drawn.

These diagrams are called *polar diagrams* and are very useful in determining the radiation characteristics of an aerial.

A polar diagram does not show the limits of reception in various directions. This depends, among other things, upon the strength of the transmitter feeding the aerial, and the sensitivity of the receiver.

Polar Diagrams of a Half-wave Dipole Suspended in Free Space

18. The polar diagrams of a half-wave dipole removed from any reflecting surface such as the earth, are shown in Fig. 6. Fig. 6(a) shows the horizontal and vertical polar diagrams for a *vertical* half-wave dipole in free space, and Fig. 6(b) shows the same for a *horizontal* dipole. A composite picture of the radiation pattern for a vertical half-wave
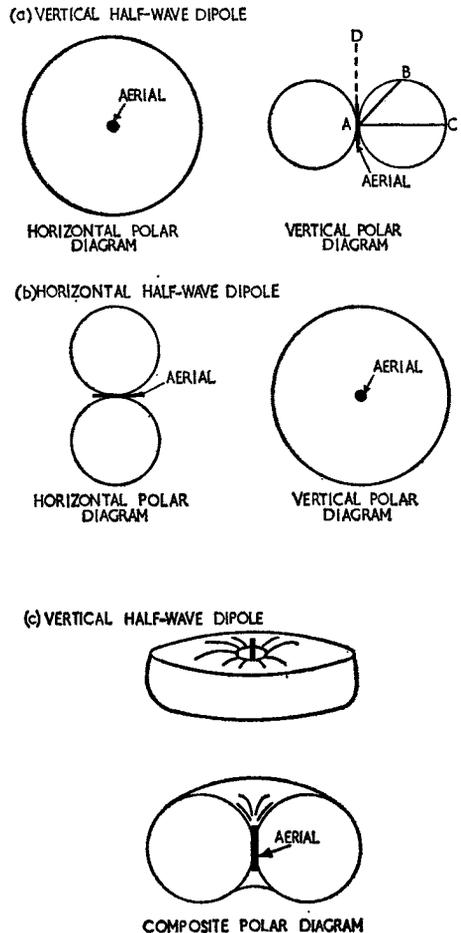
Fig. 6. POLAR DIAGRAMS FOR HALF-WAVE DIPOLES SUSPENDED IN FREE SPACE.

dipole is given in Fig. 6(c). Notice that if a line is drawn on the diagram in any direction, from the aerial to the curve (AB in Fig. 6(a)), its length indicates the relative strength of the wave radiated in that direction. Thus maximum radiation occurs along AC (90° to the aerial) and zero along AD (from the ends of the aerial). Equal all-round radiation occurs in the horizontal plane for a vertical half-wave dipole, and in the vertical plane for a horizontal half-wave dipole.

The Marconi Quarter-wave Aerial

19. At h.f., v.h.f. and u.h.f., the half-wave dipole is widely used. At low frequencies however its physical length makes it difficult to mount and support. For example, a half-wave dipole required to operate on a frequency of 500 kc/s would have to be over 300 metres long. To overcome this problem the quarter-wave Marconi aerial was developed.

Since the voltage at the centre of a half-wave dipole is nearly zero, this point can be earthed through the generator without affecting the voltage and current distribution on the aerial. The aerial, now quarter of a wave-length long, can be regarded as a modified half-wave dipole with the wire forming one half and the earth, acting as a reflector or mirror, forming the other half. This is illustrated in Fig. 7(a). The image aerial can be considered as radiating half the total power. The electric lines of force near such a radiator are shown in Fig. 7(b) and since they meet the conducting surface at right angles, the conducting surface produces no change in the field pattern.

Fig. 7. THE MARCONI QUARTER-WAVE AERIAL.

20. The conductivity of the earth surrounding a Marconi aerial must be high, and if, as in the case of dry sandy soil, this is not so,
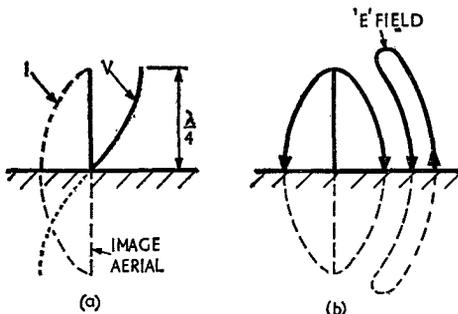
then a metal mesh round the base of the aerial, called a *counterpoise* is often found necessary, to give satisfactory termination of the electric lines of force.

21. One of the main differences between the Marconi quarter-wave aerial and the half-wave dipole, apart from the physical lengths, is the method of feeding. A centre fed dipole is a *balanced* load and must therefore be fed by a balanced transmission line such as an openwire feeder. The Marconi quarter-wave aerial on the other hand presents an *unbalanced* load, and should therefore be fed by an unbalanced feeder such as coaxial cable.

A further point of difference is the value of the radiation resistance which for a quarter-wave Marconi aerial is approximately half that of a half-wave dipole (i.e., about 36 ohms).

The polar diagrams for a quarter-wave Marconi aerial are shown in Fig. 8. They show a similar radiation pattern to that of the half-wave dipole, save that in the vertical plane the earth bisects the figure of eight.
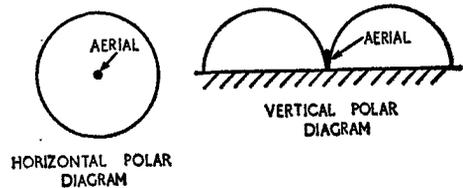
Fig. 8. POLAR DIAGRAMS FOR A VERTICAL MARCONI QUARTER-WAVE AERIAL.

Aerial Tuning

22. An aerial of fixed physical length presents to the generator an impedance which varies as the frequency of the generator voltage varies. This follows from the fact that a standing wave aerial acts as a series tuned circuit which is resonant at one frequency. This frequency is shown as f_0 in Fig. 9 and at this frequency the aerial presents a pure resistive load to the generator equal to the radiation resistance of the aerial. If the frequency is increased without altering the physical length, then the aerial becomes electrically too long and acquires an inductive reactance. If the frequency is decreased the aerial becomes electrically short and presents a capacitive load to the generator.

23. In many installations it is required to operate an aerial over a band of frequencies,

and to attempt to bring the aerial into tune by varying its physical length would be impracticable. If it is wished to operate an aerial of length l_0 (Fig. 9) at a frequency
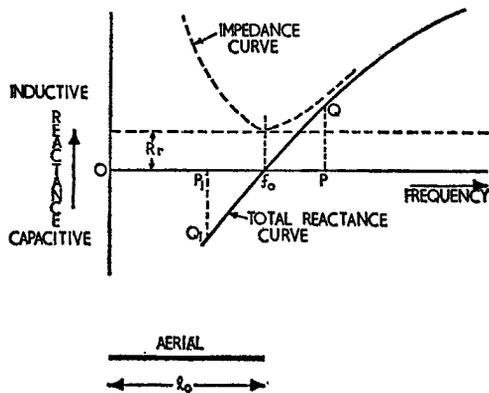
Fig. 9. VARIATION OF AERIAL INPUT IMPEDANCE WITH FREQUENCY.

represented by OP, the inductive reactance of the aerial, of value PQ, can be cancelled by inserting in the aerial a *capacitor* of value PQ. The aerial reactance then becomes zero and the aerial will act as a pure resistance equal to its radiation resistance, i.e. the aerial will be in tune. If the required operating frequency is represented by OP_1, the aerial reactance P_1Q_1 is capacitive and an *inductance* of the same value must be inserted to tune the aerial. Because of the excessive length of low and medium frequency aerials, they are rarely constructed of resonant physical length, but are made electrically resonant by inductively loading the aerial.

Thus an aerial of *fixed length* may be tuned to a particular frequency by using a loading inductor or shorting capacitor; in this way the value of L and C of the aerial is altered and it is made electrically resonant at a frequency $f_0 = \dfrac{1}{2\pi\sqrt{L.C}}$ cycles per second, where L is the total inductance and C the total capacitance of the aerial.

Reflection of Electro-Magnetic Waves

24. The Marconi quarter-wave aerial depends on the reflecting properties of the earth for efficient radiation. The fact that the ground can act as a mirror to some e.m. waves has an important effect on the polar diagrams of any aerial mounted near the surface of the earth.

Before considering the radiated field patterns of such aerials it is necessary to know more about the effect that the earth has on the radiated e.m. wave. The behaviour of a radio wave on striking a surface depends upon (a) the nature of the surface (b) the polarisation of the wave (c) the angle at which the ray strikes the surface and (d) the frequency of the wave.

The nature of the surface is of considerable importance since its effect on an e.m. wave depends on whether the surface acts as a conductor or as a dielectric to the wave. The types of reflecting surfaces on the earth vary widely, from sea to dry deserts and these cannot be clearly defined as either a dielectric or a conducting surface. Furthermore, the electrical properties of the surface vary with the frequency of the wave striking it. For example water behaves as a conductor to low frequency waves and as a dielectric to waves of high frequency.

A wave striking a surface which acts as a dielectric is partly reflected and partly refracted. Thus the reflected ray is of smaller amplitude than the incident ray. The amount of refraction depends on the refractive index of the surface, which in turn, depends on the dielectric constant k_r of the surface. The ratio of the amplitude of the reflected ray to the amplitude of the incident ray, is called the *reflection coefficient*.

Reflection from a Perfect Conductor

25. (a) **Horizontally polarised wave.** When a horizontally polarised wave strikes a perfect conducting surface all the energy is reflected. Thus in Fig. 10(a), E_2 the reflected ray is equal in amplitude to E_1 the incident ray and therefore the reflection coefficient $\dfrac{E_2}{E_1}$ is unity. The wave suffers 180° phase change on reflection and moves away from the surface at an angle of reflection, measured to the normal, equal to the angle of incidence.

(b) **Vertically polarised wave.** When a wave of vertical polarisation strikes a perfect conducting surface, there is again total reflection, the reflected ray being equal in amplitude to the incident ray and so the reflection coefficient is unity. However, there is *no change of phase* on reflection and the reflected ray moves away from the conducting surface at an angle of reflection

equal to the angle of incidence and with the same phase as the incident ray. (Fig. 10(b).)

$$\text{REFLECTION COEFFICIENT} = \frac{E_2}{E_1} = 1$$

(a) HORIZONTAL POLARISATION

(b) VERTICAL POLARISATION

Fig. 10. REFLECTION FROM A PERFECT CONDUCTOR.

Reflection from a Dielectric

26. (a) **Horizontally polarised wave.** When a horizontally polarised wave strikes a dielectric surface some of the energy is refracted and some is reflected. Thus there is a loss of energy at the surface and the reflection coefficient is therefore *less* than unity. For a wave travelling from air and striking a dielectric surface, the reflection coefficient is equal to $\dfrac{\sqrt{k_r} - 1}{\sqrt{k_r} + 1}$, where k_r is the dielectric constant of the surface. For water $k_r = 81$ and for soil $k_r = 10$. In practice most of the energy is reflected with a phase change of 180°.

(b) **Vertically polarised wave.** The behaviour of a vertically polarised wave when it strikes a dielectric surface depends upon the angle of incidence θ. When θ is small most of the wave is reflected with no change of phase, E_2 being much greater than E_3 (Fig. 11(a)). As the angle of incidence increases the reflection coefficient decreases until at a certain angle to the horizontal known as the *Brewster angle* (Φ_B in Fig. 11(b)), reflection, for all practical purposes

is zero and refraction is a maximum. Note that the angle the refracted ray makes with the path that would be taken by the reflected ray if there were one, is 90°.

(a) ϕ GREATER THAN BREWSTER ANGLE.

(b) BREWSTER ANGLE

(c) ϕ LESS THAN BREWSTER ANGLE
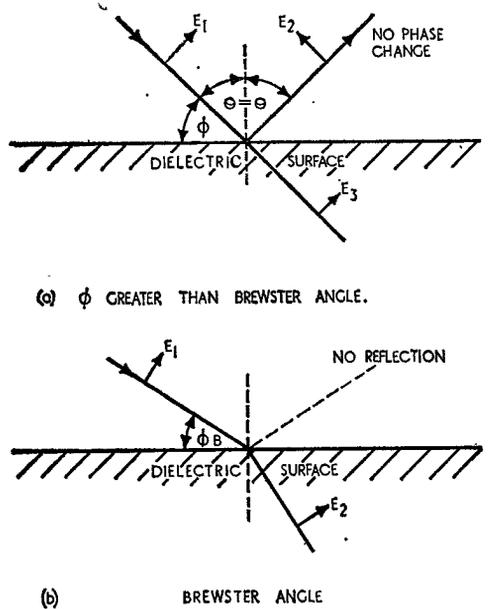
Fig. 11. VERTICAL POLARISATION—REFLECTION FROM A DIELECTRIC.

For angles of incidence where $90° - \theta$ is less than the Brewster angle (Fig. 11(c)), reflection again occurs but with a phase reversal of 180°. The reflection coefficient increases to a maximum of unity when the angle of incidence is 90°.

Reflections from Land and Sea

27. At low frequencies land and sea behave as conductors but at high frequencies they behave to the radio wave as dielectrics. The frequency at which the change occurs is not

sharply defined and is known as the *critical frequency*. For sea the critical frequency is about 900 Mc/s (33 cms) and for land about 2 Mc/s (150m). Thus land acts as a dielectric for radar frequencies while sea acts as a conductor for metre waves and a dielectric for microwaves.

The Brewster angle for dielectrics depends upon the dielectric constant which varies with frequency. For sea the Brewster angle is approximately 6·5° and for land approximately 17°.

Polar Diagrams for Practical Aerials

28. The polar diagrams shown in Fig. 6 apply only to a dipole mounted away from any reflecting surface. When an aerial is mounted on or near the earth the resultant polar diagrams are considerably changed. The field strength at a point remote from the transmitter aerial is the vector sum of the direct and reflected rays (Fig. 12). If these two rays arrive at the receiver aerial in phase, the resultant field strength will be maximum, while if they arrive out of phase it will be less than maximum, the actual value depending on the relative strengths of the two waves and their phase difference. The phase difference will depend upon the difference in path lengths and any phase change that occurs on reflection i.e. the phase difference will depend on the polarisation.
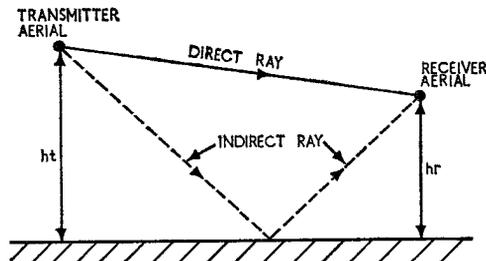
Fig. 12.—DIRECT AND INDIRECT RAYS.

29. **Polar diagrams for a vertical dipole.** The vertical polar diagram for a vertical half-wave dipole will be governed by its height above the earth. Fig. 13 shows the polar diagrams for aerials mounted at heights of $\lambda/_4$, $\lambda/_2$ and λ, the height being measured from the ground to the centre of the dipole. Radiation along the ground occurs at any height. The polar diagrams shown assume the earth to be a perfect conductor: the effect of a poor conducting surface is to deflect the ground lobe upwards by a small angle.
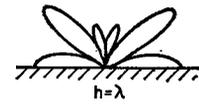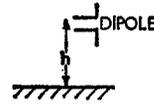
Fig. 13. VERTICAL POLAR DIAGRAMS FOR A VERTICAL HALF-WAVE DIPOLE.

30. **Polar diagrams for a horizontal dipole.** These are shown in Fig. 14. There is no lobe parallel to the ground and vertical radiation occurs at odd quarter wavelengths above the ground.
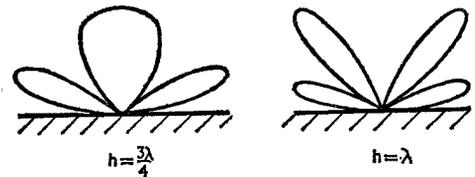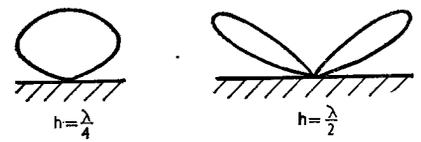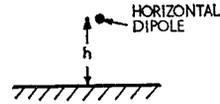
Fig. 14. VERTICAL POLAR DIAGRAMS FOR A HORIZONTAL HALF-WAVE DIPOLE.

Aerial Bandwidth

31. At h.f. and above it is often necessary to operate one aerial of fixed length over a band of frequencies. If the aerial is of correct length at the mid-band frequency it will become inductive at higher frequencies and

will not present a correct termination to the transmission line. Standing waves will be set up along the transmission line and losses will increase. For frequencies below the mid-band frequency the aerial becomes capacitive and mismatch again occurs.

32. If the band of frequencies over which the aerial must work is not too wide, the matching problem can be solved by reducing the Q factor of the aerial i.e. by reducing the L/C ratio and thus widening the aerial bandwidth. This can be accomplished by connecting several wires in parallel (Fig. 15(a))
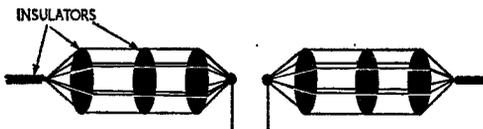
(a) MULTI-WIRE CAGE DIPOLE
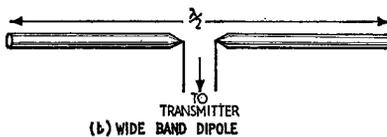
(b) WIDE BAND DIPOLE
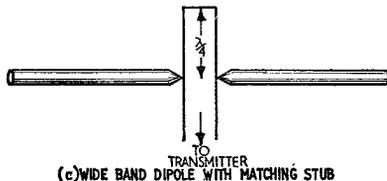
(c) WIDE BAND DIPOLE WITH MATCHING STUB

Fig. 15. AERIAL BANDWIDTH.

or by increasing the diameter of the aerial. Thus the reactive component of the input impedance for a given change in frequency, is less. By using a thick rod or a tube in place of a thin wire (Fig. 15(b)), the s.w.r. on the main transmission line is brought closer to unity and the aerial will operate satisfactorily over the required band.

33. The bandwidth of an aerial can be further increased by the device illustrated in (Fig. 15(c)). The aerial is shunted by a short-circuited section of transmission line, quarter of a wavelength long at the mid-band frequency. For an increase in frequency the aerial becomes inductive, but the stub, instead of presenting an open circuit at the aerial terminals, becomes capacitive and tends to cancel the inductive reactance of the

aerial. The opposite happens when the frequency is below the mid-band frequency; the aerial becomes capacitive and the stub inductive. Thus the aerial and the line remain sensibly matched over the required frequency band.

34. When the aerial is required to operate over a very wide band of frequencies, it may be necessary to introduce a $\lambda/4$ matching transformer, as well as a reactive stub over part of the frequency range, in order to keep the s.w.r. within acceptable limits.

A practical arrangement for use on a v.h.f. airborne transmitter/receiver is shown in Fig. 16. The equipment covers the frequency band 100–156 Mc/s in two ranges, 100–125 Mc/s and 125–156 Mc/s. For the lower frequency range a short-circuited inductive stub is sufficient to match the aerial to its 45 ohm coaxial cable (Fig. 16(a)).

Over the range 125–156 Mc/s the aerial impedance changes, with its resistive component increasing and its reactive component changing sign. To counteract this a $\lambda/4$ matching transformer is inserted between the reactive stub and the aerial, as shown in Fig. 16(b). This $\lambda/4$ transformer effectively reduces the resistive component and reverses the sign of the aerial reactance so that the short-circuited stub is still effective in counteracting the reactance on the line, and maintaining the s.w.r. within the specified limits.

Slot Aerials

35. Another basic type of resonant aerial is the slot aerial. A slot, such as the one illustrated in Fig. 17(a) will radiate in a manner similar to a dipole provided the surface area of metal surrounding the slot is large in terms of the wavelength being radiated. Because of this proviso the use of slot aerials is confined to frequencies above 200 Mc/s. Slot aerials are used on airborne wireless and radar installations and in certain ground radar equipments. The polar diagram for a simple half-wave slot aerial is similar to that for a half-wave dipole.

One of the main differences in the characteristics of a dipole aerial and a slot aerial is the reversal of the plane of polarisation. A vertical slot will radiate a horizontally polarised wave and a horizontal slot, a vertically polarised wave. This is because the standing wave distribution along the slot
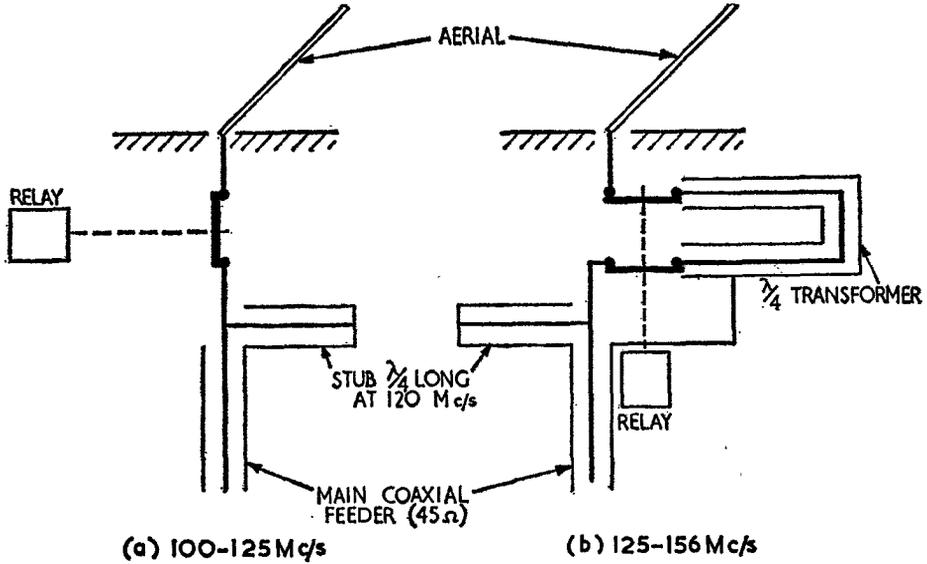
Fig. 16. WIDE BAND AERIAL MATCHING.

is different as shown in Fig. 17(b). As might be expected, the current standing wave is a maximum at the short-circuited ends and a minimum in the centre, whilst the voltage standing wave has nodes at the ends and an antinode at the centre. This means that the input impedance of a slot aerial varies in the opposite manner to that of a dipole, being maximum (about 485 ohms) at the centre and minimum at the ends.
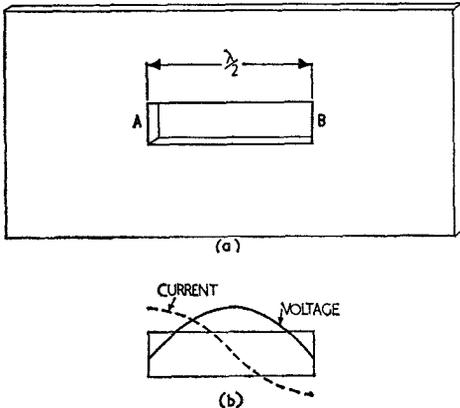
Fig. 17. THE SLOT AERIAL.

Methods of Feeding Slot Aerials

36. A slot aerial may be fed in the same way as a dipole, i.e. voltage fed or current fed. Fig. 18(a) shows voltage feed. The impedance of the transmission line can be matched to the
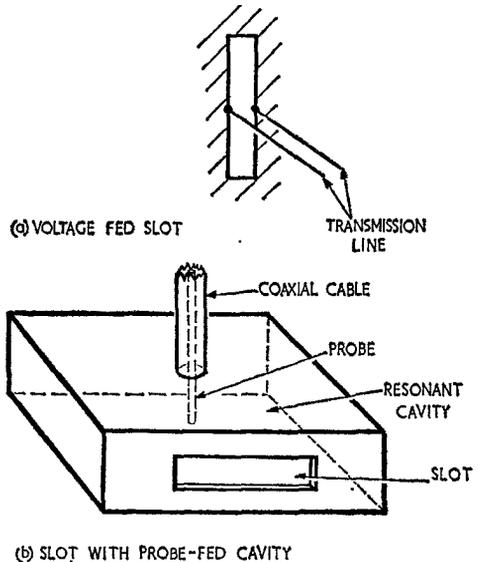
Fig. 18. METHODS OF FEEDING A SLOT AERIAL.

input impedance of the slot by moving the feed point to one end of the slot, since the slot acts as its own auto-transformer.

In certain applications, where the slot aerial is used as a radiator in one direction only, the slot is backed by a resonant cavity, as shown in Fig. 18(b). The cavity size is such as to provide resonance over the required frequency band. Energy is fed into the cavity via a voltage probe projecting through the roof of the cavity.

SECTION 16

CHAPTER 3

DIRECTIONAL AERIAL ARRAYS

DIRECTIONAL AERIAL ARRAYS

Introduction

1. The simple aerials so far considered do not possess directional properties. For example, the field radiated from a single vertical aerial is of equal amplitude in all directions in the horizontal plane. An aerial system designed to radiate most of its energy in one direction would have many uses. The operational range of the equipment could be increased or the transmitter power could be reduced. Further, such a system, used as a receiver aerial, would indicate the direction in which a transmitter lies.

Directional properties can be obtained by using combinations of aerials and systems such as these are called *aerial arrays*. The common television receiving aerial is one example, and many service radio installations employ directional aerial arrays.

Parasitic Elements

2. If a metal rod is placed in the field radiated from a driven dipole it will have currents induced in it, and it will radiate. The rod obtains its energy from the driven dipole and so is called a *parasitic element* (Fig. 1). The phase of the energy radiated by the parasite will depend upon its length, lp and its distance (S) from the dipole. The length of the parasite determines the phase difference between the current and induced voltage. The distance of the parasite from the driver determines the phase of the voltage induced in the parasite relative to the field radiated by the driver.
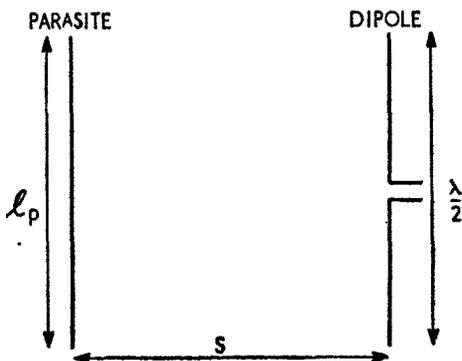
Thus the parasitic element and the driven element radiate e.m. energy. These fields will reinforce each other in one direction, while they will weaken each other in the opposite direction. In practice it is found, by trial and error, that for radiation to be increased in one direction the parasite has to be reactive. This is achieved simply by making the parasite longer or shorter than half a wavelength.

3. Best results are obtained when an inductive parasite longer than the dipole, is placed quarter of a wavelength behind the dipole. This results in the horizontal polar diagram shown in Fig. 2(a). This parasitic element is called a *reflector*.

Similar results are obtained if a parasite shorter than the dipole is placed 0.15λ in front of the dipole. The parasite is now called a *director* and the combination produces a horizontal polar diagram as shown in Fig. 2(b).

The aerial array shown in Fig. 2(a) is often used as a roof-top television receiver aerial.

The Yagi Array

4. A combination of reflector and director used with a driven dipole, gives greater forward radiation than either used separately. This arrangement is shown in Fig. 3, and is known as a *Yagi array*. The spacing and lengths of the elements are usually found by trial and error; typical figures are given in Fig. 3(a). The horizontal polar diagram is shown in Fig. 3(b).

5. The directivity can be increased if the number of directors used is increased. The further the director is from the dipole the greater is the capacitive reactance required to obtain correct phasing of the parasitic current. The lengths of the directors therefore taper off as shown in Fig. 4(a). The maximum practical number of directors which can be used is four or five.

Input Impedance of Yagi Array

6. Because of the proximity of the parasitic elements and the driven element of a Yagi array, there is considerable inductive coupling between them. This reduces the input

Fig. 1. THE PARASITIC ELEMENT.

REFLECTOR

DIRECTOR
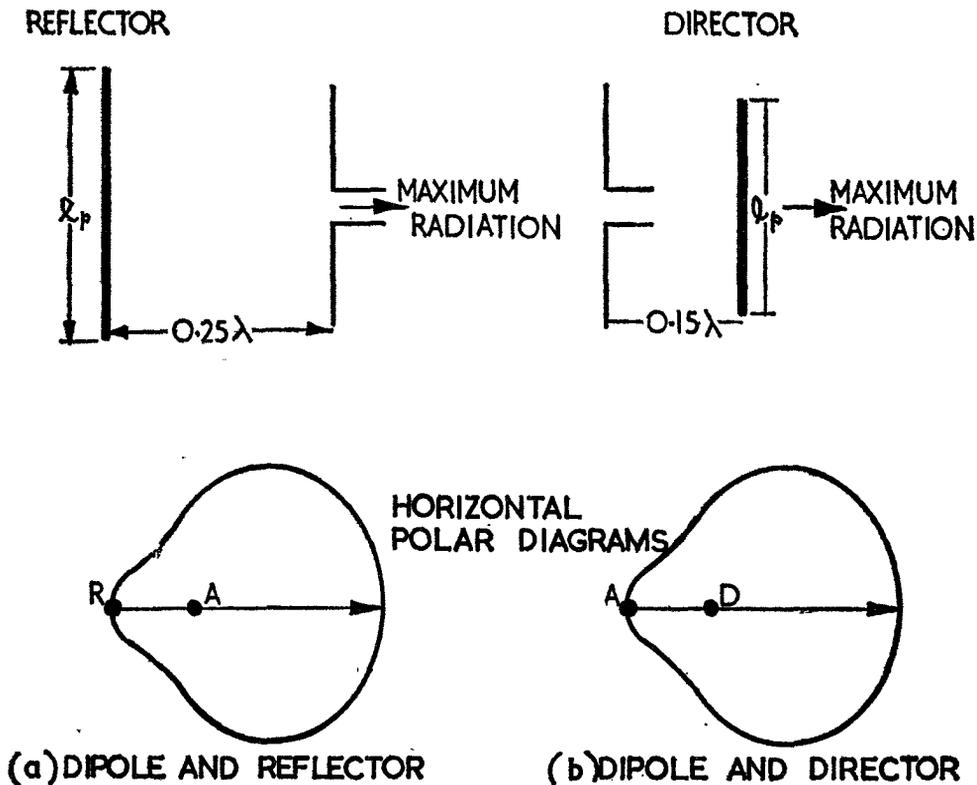
Fig. 2.   DIPOLE WITH PARASITES.

(a) DIPOLE AND REFLECTOR

(b) DIPOLE AND DIRECTOR
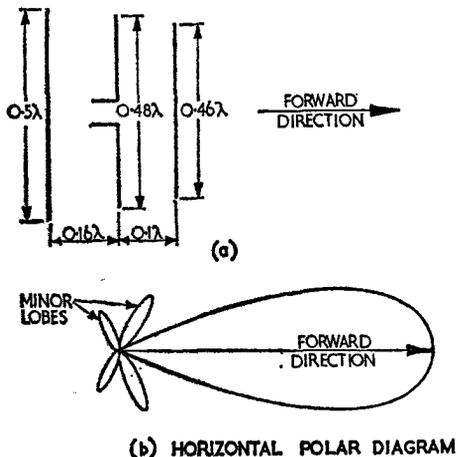
Fig. 3.   THREE ELEMENT YAGI ARRAY.
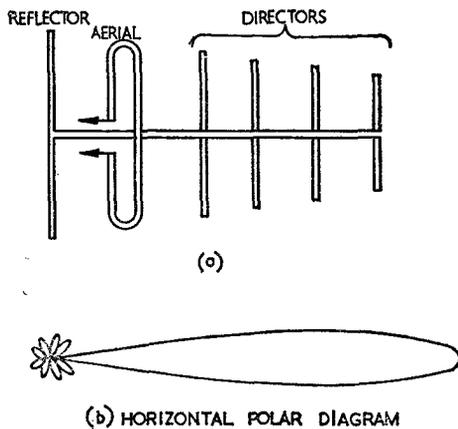
Fig. 4.   SIX ELEMENT YAGI ARRAY.

impedance of the dipole considerably and its value can fall as low as 20 ohms. Matching the dipole to the transmission line becomes difficult at this low value. To facilitate matching, a folded dipole, which has an input impedance four times that of a simple dipole, is used.

Forward Gain
7. An indication of the effectiveness of a directional array can be obtained by comparing the field strength radiated from a reference aerial (either an omni-directional aerial or a simple half-wave dipole), with the field strength radiated in the forward direction by the array under test, both aerials being fed with the same current. The ratio:

$$\frac{\text{max. field strength radiated from given aerial}}{\text{field strength radiated from reference aerial}}$$

is called the *forward gain* of the array under test. It is measured in decibels.

Back to Front Ratio
8. Another method of stating the effectiveness of an aerial array such as a Yagi, is by giving the ratio of the energy radiated in a backward direction to that radiated in a forward direction. This is termed the *back to front ratio* of the array and is also measured in decibels.

Beam Width
9. A convenient and often-used measure of the directivity of an aerial array is the *beam width*. This is the angle measured on the polar diagram between points where the radiated power has fallen to half of its maximum value. Polar diagrams usually show the distribution of field strength, not power, but power is proportional to the square of the field strength, and so on a field strength polar diagram the 0·707 points correspond to the half power points. This is shown in Fig. 5. A typical beam width for a 6 element Yagi array is 40°.
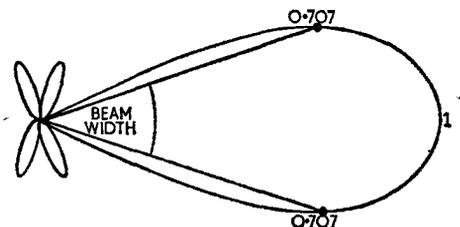
Fig. 5. BEAM WIDTH.

Multiple Driven Arrays
10. So far one driven aerial used in conjunction with parasitic elements has been considered. Directional arrays of two or more driven elements are often used, particularly at h.f. when the lengths and spacings of the elements are such that the arrangement is of practical size.

Driven arrays are used more for transmission than reception, but although a receiver aerial cannot be driven in the same way as a transmitter aerial, the polar diagrams will be the same for both transmitter and receiver aerials.

Linear Broadside Array
11. Consider two half-wave vertical dipoles (A and B in Fig. 6), spaced half a wavelength apart and fed with equal amplitude in-phase currents.
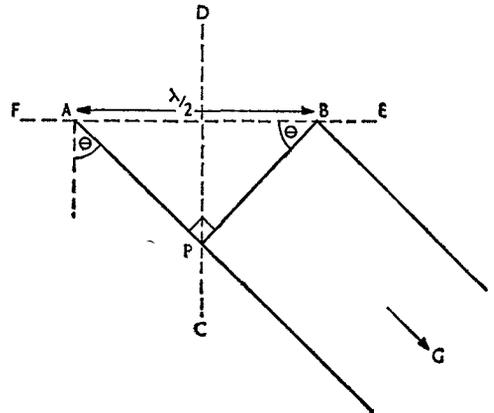
Fig. 6. TWO ELEMENT BROADSIDE ARRAY.

When the energy radiated from A in the direction AE arrives at B it will be 180° out of phase with the energy then being radiated from aerial B, since the wave path AB is half a wavelength, or 180°. Thus, since the two energies are of equal amplitude, they will cancel, and no radiation will take place in the direction AE. Similarly, there will be no radiation in the direction BF.

Along the line DC, which is the perpendicular bisector of AB, the wave paths of the energy from A and B are the same length. Therefore since the aerials are fed in phase the radiated energy from the two aerials will be in phase and add anywhere along the line DC. Thus the energy radiated at 90° to the array will be maximum.
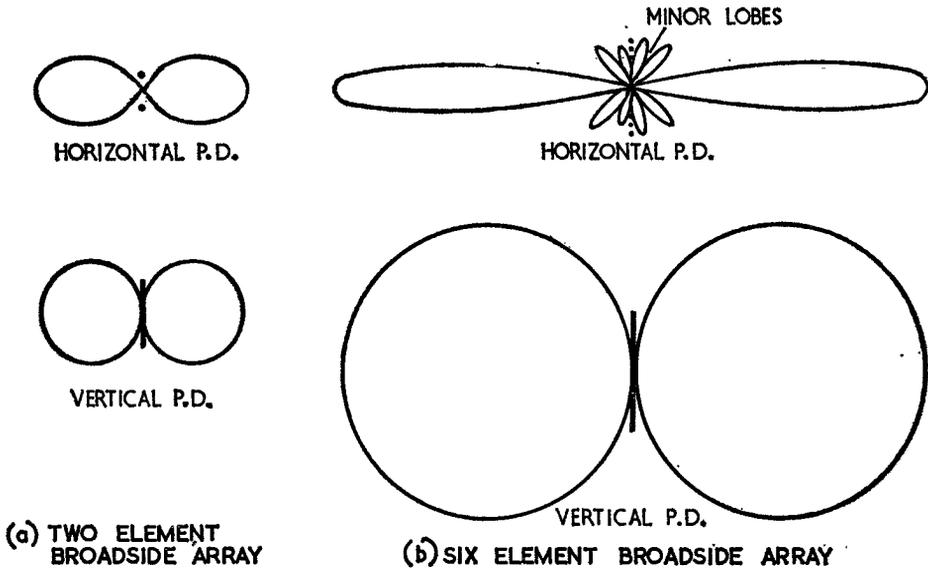
Fig. 7. POLAR DIAGRAMS FOR TWO-ELEMENT AND SIX-ELEMENT BROADSIDE ARRAYS.

12. At a distant point G (Fig. 6), the resultant field strength will depend upon the power radiated from each aerial and upon the angle θ. This type of array is called a *linear broadside array*. The horizontal and vertical polar diagrams for a two-element array are shown in Fig. 7(a).

13. If the number of elements forming the array is increased, the beam width narrows and the gain of the array increases. In general, for a broadside array, by increasing the size of the array in one plane the directivity in the plane at right angles to the array is increased. The horizontal and vertical polar diagrams for a six-element broadside array are shown in Fig. 7(b).

Broadside Array with Reflectors

14. An obvious saving in transmitter power could be obtained if the broadside array were made to radiate in one direction only. This can be done by placing parasitic elements one quarter of a wavelength behind the driven elements. These parasites will then act as reflectors augmenting radiation in the forward direction and reducing it in the backward direction. The horizontal polar diagram of a six-element broadside array with reflectors is shown in Fig. 8.

At high frequencies, where the wavelength is short, a wire mesh or solid metal reflector can be used in place of the metal rods. For large aerials however, solid metal reflectors are avoided because of their wind resistance.
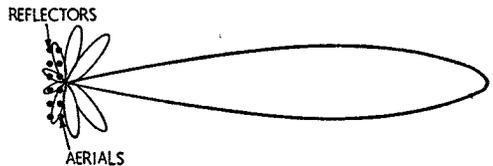
Fig. 8. HORIZONTAL POLAR DIAGRAM FOR BROADSIDE ARRAY WITH REFLECTORS.

The End Fire Array

15. Another type of directional array is that illustrated in Fig. 9(a), and called the end fire array. The driven elements are spaced a quarter wavelength apart and are fed with equal amplitude currents, the phase of the current in each aerial being such that it lags by 90° the current in the aerial on its left.

The energy radiated from aerial A in the forward direction, will thus be in phase with the energy then being radiated from aerial B when it arrives at B after having moved through a quarter wavelength. Similarly by the time the combined radiation from A and B arrives at C, $\lambda/4$ or 90° later, the energy being radiated from C will also be in phase with that from A and B in the forward direction; and so on through the array. In

the backward direction however, energy from B will be 180° out of phase with energy then being radiated from A in the backward direction, and the two will cancel. Thus because of the phasing and the spacing of the aerials, radiation in the backward direction is cancelled while that in the forward direction adds.

(a)

(b) HORIZONTAL POLAR DIAGRAM
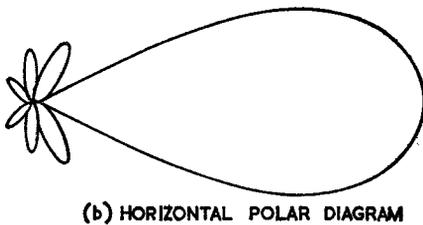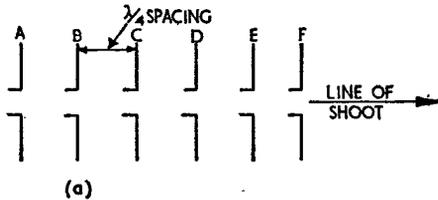
Fig. 9. END FIRE ARRAY.

16. This arrangement therefore gives a horizontal polar diagram as shown in Fig. 9(b); the direction of maximum radiation is in the line of the array and is called the line of shoot. For an eight-element array the beam width is approximately 60° and for a sixteen-element array it is 41°.

The difference between this array and the Yagi array is that in the Yagi array only one element is driven, the remainder being parasites.

Stacked Arrays

17. So far arrays which give directivity in the horizontal plane have been discussed. It is often desirable to have an array which possesses directional properties in both the horizontal and vertical planes. This can be achieved by stacking a number of broadside arrays one above the other as shown in Fig. 10. The array illustrated is the pine tree array employing horizontal aerials. It gives a narrow beam in both the horizontal and vertical planes.

The method of feeding the array is impor-

tant since each aerial must be fed in phase with the other aerials. The aerials are therefore spaced half a wavelength above each other and the feeder line is crossed over as shown. Also, the feed to each stack must be in phase and so the length of feeder between the stacks and the feed point must be equal.
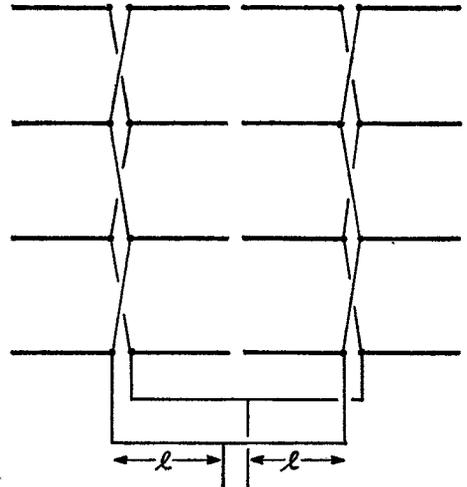
Fig. 10. PINE-TREE ARRAY.

The dipoles are end fed so the input impedance to each is high but since all the aerial impedances are in parallel the impedance presented to the feeder line is conveniently low to make a match possible.

Summary

18. The different types of driven aerial arrays dealt with in this chapter are summarised in Fig. 11. The broadside array gives horizontal directivity at 90° to the line of the array; the gain can be increased by increasing l, i.e. by adding more aerials. Radiation can be concentrated in one direction by using undriven reflectors, as shown.

In the end fire array maximum gain is in the direction of the array and again, the beam width can be narrowed and the forward gain increased by adding more driven elements.

The stacked array gives directivity in the vertical plane, the aerials being mounted one above the other. In this case, to increase the gain in the vertical plane, h must be increased.

ARRAY   CHARACTERISTICS

LINEAR BROADSIDE ARRAY (PLAN)

HORIZONTAL P.D.

AERIALS FED IN PHASE
AND SPACED λ/2

LINEAR BROADSIDE ARRAY
WITH REFLECTORS (PLAN)

HORIZONTAL P.D.

AERIALS FED IN PHASE
REFLECTORS SPACED λ/4
BEHIND AERIALS

END FIRE ARRAY (PLAN)

HORIZONTAL P.D.

AERIALS FED 90° PHASE
LAGGING AND SPACED λ/4

STACKED ARRAY (SIDE VIEW)   VERTICAL P.D.

AERIALS FED IN PHASE
AND SPACED λ/2 VERTICALLY

Fig. 11. CHARACTERISTICS OF DRIVEN ARRAYS.

19.   The theory of the aerial arrays discussed in this chapter can be applied to many of the more complicated arrays in current use. Each array is designed for its particular purpose but the basic principles used in the design are common.

SECTION 16

CHAPTER 4

TRAVELLING WAVE AERIALS

TRAVELLING WAVE AERIALS

Introduction

1. The aerials so far considered in these Notes have been standing wave aerials. In a standing wave or resonant aerial, a wave sent along the wire to an unterminated end is reflected back to the source and the combination of incident and reflected waves forms a standing wave on the aerial.

   If the wire forming the aerial is terminated in a resistor equal in value to the characteristic impedance of the wire, then the incident energy arriving at the termination would be absorbed by the resistor. There would be no reflection and the only waves on the wire would be the incident waves travelling towards the termination. These waves are called *travelling waves* and give the aerial its name.

2. Aerials of this type will be discussed in this chapter. They are widely used and have certain advantages over standing wave aerials. They are simple to construct and service, and possess good directional properties. As they are non-resonant, they can be used over a wide band of frequencies without being retuned. This is of considerable advantage when the system employs reflections from the ionosphere and is required to operate on several widely spaced frequencies during day and night.

Principle of the Travelling Wave Aerial

3. A simple travelling wave aerial consists of a horizontal wire terminated with a resistance equal to the Z_0 of the wire (Fig. 1). Energy
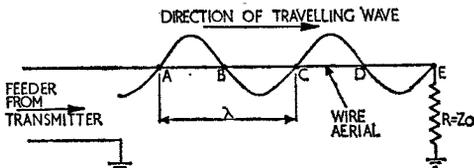
Fig. I. SIMPLE TRAVELLING WAVE AERIAL.

from the transmitter causes travelling waves to move towards the resistor. These travelling waves radiate e.m. energy as they progress. The radiation pattern is shown in Fig. 2, the complete polar diagram being

a solid of revolution about the axis of the wire. The angle that the two major lobes make with the axis of the wire (angle θ in Fig. 2(a)), and the beam width of the lobes depend upon the length of the wire measured in wavelengths. The longer the wire the smaller the angle θ and the narrower the beam width. Fig. 2(a) shows the polar diagram for an aerial 2λ long; this gives a value for θ of 35°. Fig. 2(b) shows the polar diagram for an aerial of length 4λ. The beam width is narrower than that of Fig. 2(a) and the angle of the major lobe is 24°.
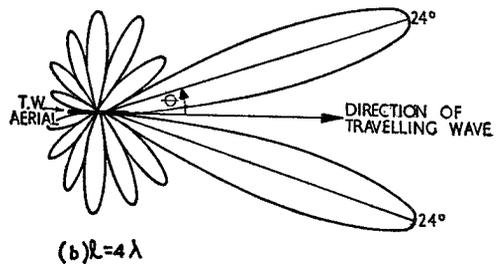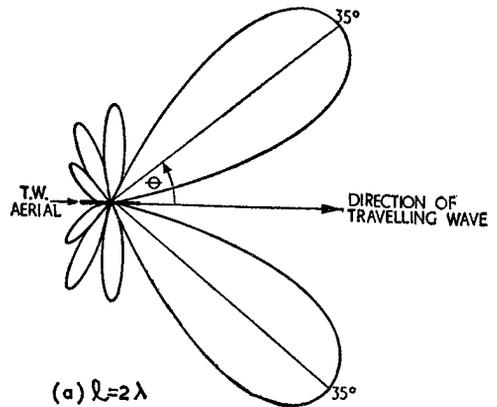
Fig. 2. POLAR DIAGRAMS FOR TRAVELLING WAVE AERIALS.

The Beverage Aerial

4. This is a simple travelling wave aerial used mainly for reception of long wave signals. It consists of a wire several wavelengths long mounted about 30 feet above the

earth and terminated to earth at both ends through resistances equal to the Z_0 of the wire (Fig. 3). The wave front of a vertically polarised wave is tilted slightly, due to diffraction and so a voltage will be induced in the wire. This will cause a travelling wave to move along the wire towards the receiver and a voltage will be developed across the receiver input.
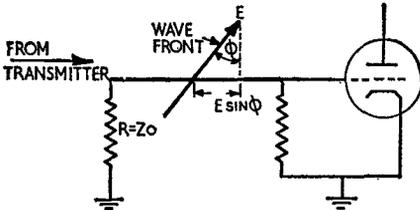
Fig. 3. THE BEVERAGE AERIAL.

5. The polar diagram for this aerial is similar to those shown in Fig. 2. The forward gain can be increased by connecting an array of these aerials in parallel. The Beverage aerial is simple to construct and maintain and gives satisfactory reception over a wide band of frequencies in the v.l.f. and l.f. bands.

The Inverted-V Aerial

6. This travelling wave aerial is similar to the Beverage aerial but is used on higher frequencies. As shown in Fig. 4, it requires only one supporting mast. When used for transmitting it is fed by an unbalanced feeder at point A, and point C is connected to earth through a resistor R, of such a value that no standing waves are set up on the aerial. The direction of maximum radiation is towards the terminating resistor R, as shown by the arrow in Fig. 4.
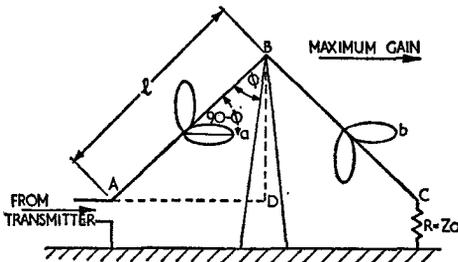
Fig. 4. THE INVERTED-V AERIAL.

7. When excited by the transmitter the legs AB and BC will cause major lobes to form as shown. For maximum radiation in the direction shown by the arrow, the angle that

the major lobes make with the wire must equal $90° - \phi°$ where $-\phi$ is the angle of tilt; further, the energy in each lobe must be in phase at a distant point. To satisfy these requirements the lengths of the legs measured in wavelengths, and the angle ϕ must be carefully chosen.

The main advantages of the inverted-V aerial lie in its simplicity of construction, its narrow beam width and its broad bandwidth.

The Rhombic Aerial

8. This is a non-resonant aerial widely used for transmitting and receiving at h.f. It consists of four wires forming a diamond or rhombus shape in the horizontal plane, as shown in Fig. 5(a). It has a very wide bandwidth, with a high gain and fairly constant input impedance to frequencies within this band. The direction of maximum gain is along the major axis of the rhombus towards the terminating resistor. The value of this resistor is such that travelling waves are set up in the four legs AB, BC, CD and DA, when they are excited.
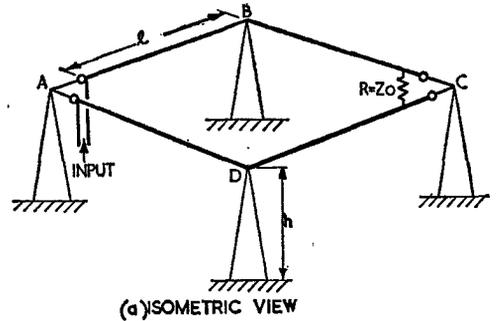
(a) ISOMETRIC VIEW
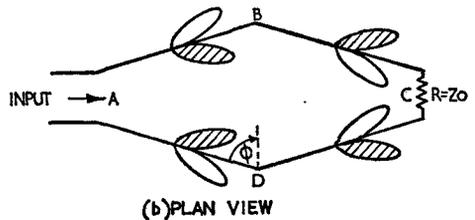
(b) PLAN VIEW
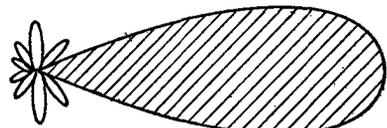
(c) POLAR DIAGRAM

Fig. 5. THE RHOMBIC AERIAL.

Each wire produces a main hollow cone of radiation as shown in the plan view of the aerial in Fig. 5(b). The object of the design is to make the four lobes shaded in Fig. 5(b) combine in phase and direction, to form a concentrated forward lobe of radiation as shown in Fig. 5(c). To obtain this, the factors which have to be considered and which are interdependent are the leg length l, measured in wavelengths, the tilt angle ϕ, and the height of the aerial above the ground, h. As with the inverted-V aerial l and ϕ

characteristic impedance of the wires, some of the transmitter power is dissipated as heat in the terminating resistor. This loss amounts to between 30% and 50% of the input power. In the case of the rhombic aerial the loss may be reduced to a minimum by constructing the legs of several spaced wires in parallel as shown in Fig. 7. This reduces the Z_0 and so gives a lower power loss in the terminating resistor for a given aerial current. Losses also occur due to the large number of side lobes of appreciable amplitude. Because of

(a) VERTICAL POLAR DIAGRAMS

(b) HORIZONTAL POLAR DIAGRAMS
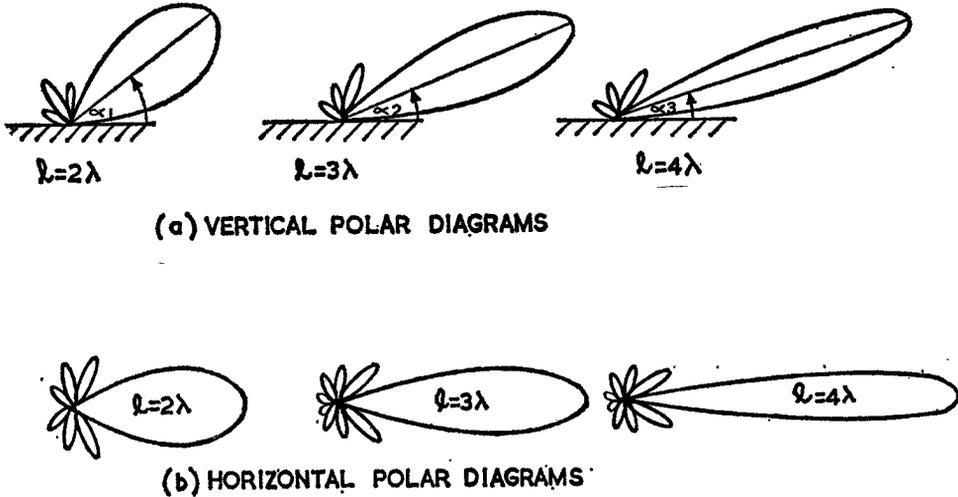
Fig. 6. POLAR DIAGRAMS OF RHOMBIC AERIALS.

control the gain of the main beam. The factor h, in conjunction with l, controls the angle of elevation of the main beam.

9. Horizontal and vertical polar diagrams for rhombic aerials with different leg lengths are shown in Fig. 6. Note that as the leg length increases, the beam width narrows and the angle of elevation ($\propto$) decreases.

As with all travelling wave aerials terminated with a resistance equal to the

these side lobes the aerial can cause interference with other stations working on the same frequency when transmitting, and can be subjected to interference when receiving.

The input impedance of a multi-wire rhombic aerial is approximately 600 ohms, and so is suitable for matching to a 600 ohm open wire transmission line. The approximate power gains over a half-wave dipole, for various leg lengths are shown in Table 1.

Side Length (wavelengths)	Power Gain (db)
2	6·5
3	9·0
4	12·0
6	18·0
8	23·0

Table 1.
MULTI-WIRE RHOMBIC
POWER GAINS
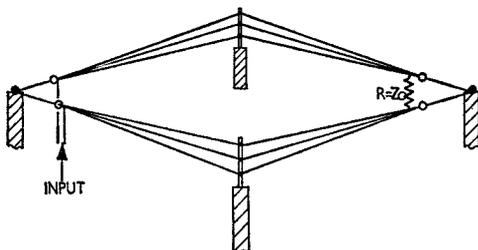
Fig. 7. MULTI-WIRE RHOMBIC.

These figures allow for the loss in the terminating resistor.

10. The advantages of travelling wave aerials far outweigh the disadvantages. They are widely used for long distance point to point communication where several changes of frequency throughout a period of 24 hours may be necessary. They work well over a wide frequency range; a typical frequency coverage is a $2 \cdot 5 : 1$ change in frequency without need for adjustment.

Their main disadvantage is the large ground area needed for the installation.

SECTION 16

CHAPTER 5

PRACTICAL AERIALS

PRACTICAL AERIALS

Introduction

1. Previous chapters in this Section have considered basic aerials and aerial arrays from a theoretical aspect and certain types of aerial installations have been used to illustrate the theory. The service technician will encounter a wide variety of aerial installations, the designs of which are based on the theoretical considerations already discussed. A small selection of these aerial installations will be considered in this Chapter. Since the type of aerial installation used for a particular purpose largely depends on the frequency employed, they have been classified according to the frequency and wavelength bands of Table 1. Representative aerials used in each of these bands and problems associated with their design will be considered in this Chapter.

Band	Frequency Range	Wave-length Range (Metres)	Typical Services
Very low freq. V.L.F.	3–30 kc/s	10^5–10^4	World-wide telegraphy.
Low freq. L.F.	30–300 kc/s	10^4–10^3	Navigation, broadcasting, long distance communication.
Medium freq. M.F.	0·3–3 Mc/s	1000–100	Broadcasting.
High freq. H.F.	3–30 Mc/s	100–10	Beamed communication services.
Very High freq. V.H.F.	30–300 Mc/s	10–1	Air to ground communications. Radio relay telephony. Television.
Ultra High freq. U.H.F.	300–3,000 Mc/s	1–0·1	Air to ground communications, navigational and landing aids, radar.
Super High freq. S.H.F.	3,000 Mc/s and above	0·1 and below	Radar. Radio and television relay links.

TABLE 1

Low Frequency Aerials

2. The major problem encountered in the design of aerials for operation in the v.l.f. and l.f. bands is the physical size of the elements involved. For example if it is required to have a quarter-wave aerial operating on 100 kc/s, the height of the aerial would have to be approximately 2,500 feet. The structural problems involved make such an aerial impracticable, and in general heights are restricted to about 300 feet. This means that aerials operating at low frequencies are only a fraction of a wavelength high.

Directional arrays would require massive structures and so omnidirectional systems are generally used at these frequencies. Owing to the heights involved vertical polarisation must be employed, and this means using end fed vertical aerials. Since the current at the base of such an aerial used for transmitting is high, losses occurring in the ground surrounding the aerial are high, especially if the ground is of poor conductivity. This means that the efficiency of the aerial will be low; the efficiency can be improved by burying conductors in the ground surrounding the aerial so as to form a radial mesh extending for a distance from the base of the aerial equal to the height of the aerial. Alternatively a counterpoise as mentioned in Chapter 2, para. 19, may be employed.

3. The aerial efficiency can be further increased by making the aerial with wire of a low resistance. Stranded copper or phosphor bronze wire is often used; the heavier the gauge of the wire used, the smaller is the ohmic resistance and the more efficient the aerial.

4. One method of increasing the radiation from a vertical end-fed aerial without increasing the mast height is illustrated in Fig. 1. The current distribution for such an aerial of height h is shown in Fig. 1(a). By connecting a horizontal wire to the top of the aerial, the current distribution in the vertical (and most important) portion is as shown in Fig. 1(b); the current at the top of the aerial is no longer zero, and so the effective height of the aerial is increased, thus increasing radiation from the aerial.

This type of aerial is often used at v.l.f. and l.f., and is called an inverted L aerial. A variation, known as the T aerial is shown in Fig. 1(c).
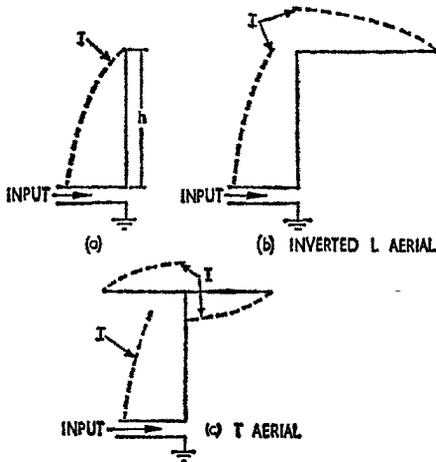
Fig. I. V.L.F. AND L.F. AERIALS.

ference between the ground wave and sky wave will be minimum, and there is less risk of fading at points remote from the transmitter. Towards the upper end of the band aerials half a wavelength long can be used. For example at 1 Mc/s (300m) a tower 450 feet high would be required, which is a practical construction. The current at the base of a vertical half-wave dipole is small and thus the ground losses close to the aerial are low.

6. Fig. 2(a) shows a typical half-wave vertical m.f. aerial. It consists of a parallel mast half a wavelength high supported by wire guys. The guys are broken up into small sections less than quarter of a wavelength long and each section is separated by insulators. Thus the guys are non-resonant and do not affect the radiation pattern. The radial earth wires at the base should be at least as long as the mast is high to provide a good ground system. The standing wave distribution and the polar diagram of such an aerial are shown in Fig. 2(b). The radiation
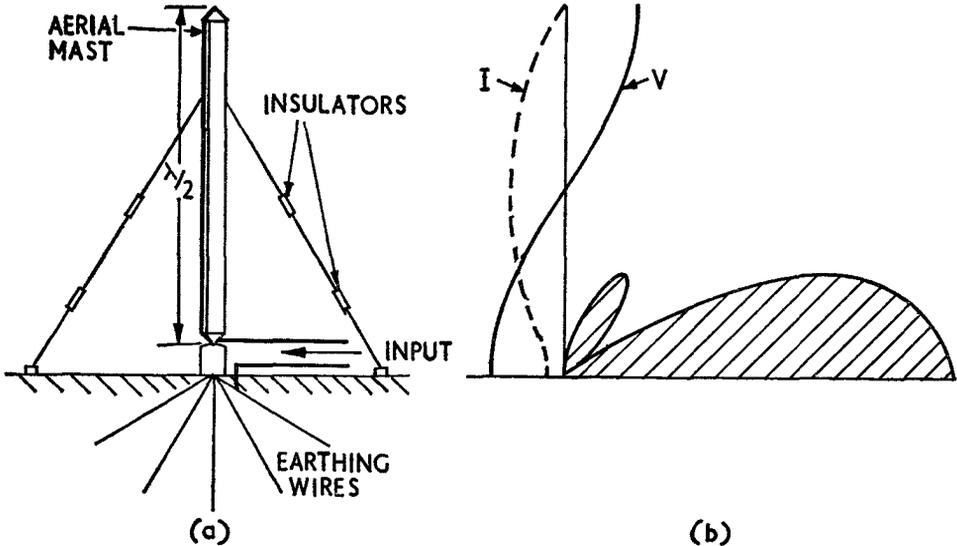
Fig. 2. HALF-WAVE VERTICAL M.F. AERIAL.

Medium Frequency Aerials

5. Aerials in this category are designed to cover the frequency band 300 kc/s to 3 Mc/s; they are usually employed on broadcast systems and are omnidirectional. Vertical polarisation is required and maximum gain along the ground with minimum radiation in the vertical plane is necessary. Thus inter-

resistance is high compared to the total loss resistance and so a high efficiency is obtained.

7. The height of the aerial described in para. 6 can be reduced if a capacitance 'hat' is used. The hat consists of a ring with radial spokes as shown in Fig. 3. If the diameter of the ring is 5 times that of the mast, the

physical height of the aerial can be reduced by 5 to 10% without affecting the radiation pattern to any marked degree.
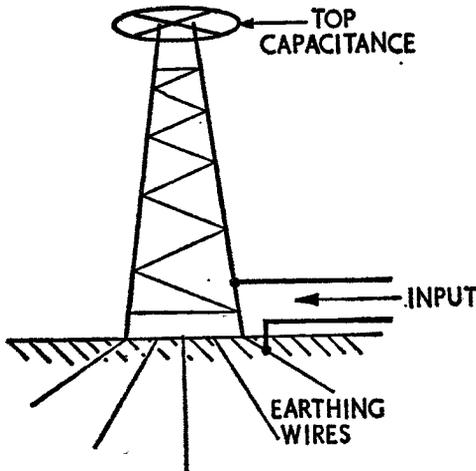
Fig. 3. CAPACITANCE HAT AERIAL.

8. The outrigger folded unipole aerial illustrated in Fig. 4 is one other type of m.f. aerial. It consists of a central mast from which are suspended wires quarter of a wavelength long which form the radiating elements. It is end fed and an efficient earthing system must be constructed to reduce ground losses. It is
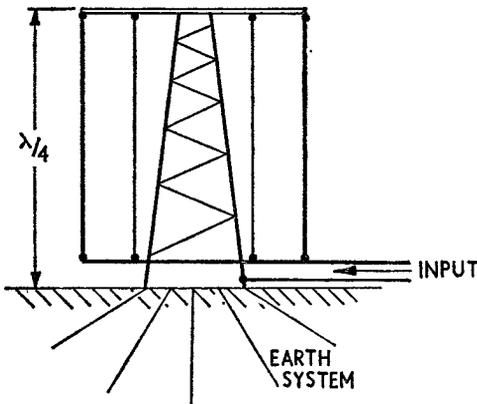
Fig. 4. OUTRIGGER FOLDED UNIPOLE.

an omnidirectional aerial using the ground wave for propagation. Two such aerials spaced quarter of a wavelength apart and driven by the same transmitter would have directional properties.

9. Aerials for use in the l.f. and m.f. bands are expensive to construct. When designing these aerials it is usual to construct small scale models mounted on wooden masts of corresponding scale, and driven at a proportionally higher frequency. With this comparatively inexpensive model, field strength readings can be taken in all planes around the aerial and the design adjusted to meet requirements.

High Frequency Aerials

10. Communication in the band of frequencies between 3 and 30 Mc/s generally employs sky wave propagation. At the lower end of the band the simplest form of aerial used is the horizontal half-wave dipole shown in Fig. 5.
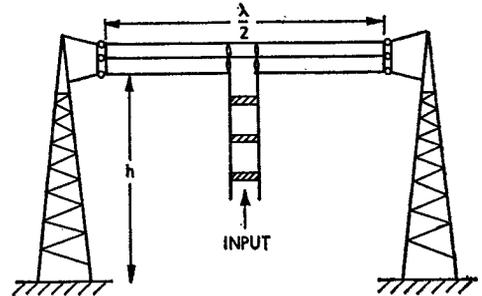
Fig. 5. HORIZONTAL CENTRE FED HALF-WAVE DIPOLE.

The aerial is rigged between two support masts and uses the earth as a reflector. Heights of between $\lambda/8$ and $\lambda/4$ are used, when a polar diagram similar to that shown in Chapter 2, Fig. 4 is produced. Maximum gain is obtained at high angles of elevation and no ground wave is present. This aerial is thus suitable for ionospheric communication using horizontally polarised waves (see Section 17). A folded dipole is used in order to provide a high input impedance to the open wire feeder. Alternatively the delta match may be employed.

The end fed horizontal half-wave dipole, known as the Zeppelin aerial is shown in Fig. 6. This aerial has the same characteristics as the centre fed dipole already discussed.

11. For all-round radiation using the ground wave, a vertical end-fed dipole can be used at the lower end of the band. In the middle of the band, two or more half-wave dipoles mounted one above the other give
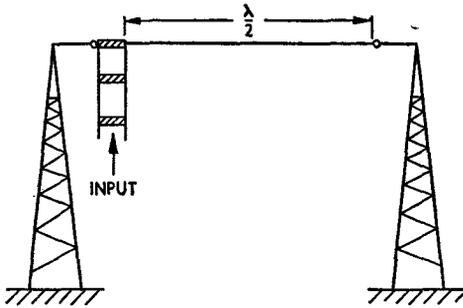
Fig. 6.   THE ZEPPELIN AERIAL.

extra gain in the horizontal plane. This type of aerial, called a Franklin aerial, is illustrated in Fig. 7. In Fig. 7(a) radiation from alternate half waves is suppressed by inserting coils which have distributed inductance and capacitance equivalent to half a wavelength. A better method is shown in Fig. 7(b). By folding alternate half wavelengths of wire as shown, fields radiated by the upper and lower folded lengths cancel and alternate half wavelengths are suppressed.
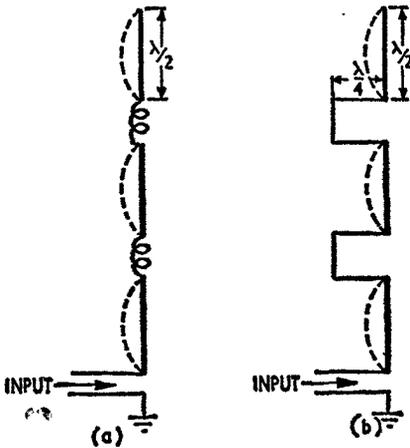
Fig. 7.   THE FRANKLIN AERIAL.

12.   In the h.f. band increased ranges can be obtained by using directional arrays with the main beam directed at an angle to the ionosphere such as to produce optimum results at the receiver (see Section 17). Examples of standing wave and travelling wave directional arrays have been given in Chapters 3 and 4.

13.   **Ferrite rod aerials.** The development of ferrite materials possessing high permeability

and high resistivity (i.e. negligible eddy current losses) has resulted in considerable improvement in many electrical components.   One application of these materials is the ferrite rod aerial used on many modern receivers.

When a rod of a suitable ferrite material is placed in a magnetic field, the field around the rod is distorted as shown in Fig. 8.   The
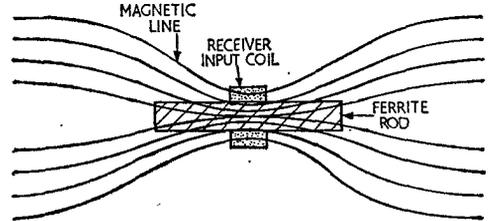
Fig. 8.   FERRITE ROD AERIAL.

magnetic field from a large area is concentrated in the rod, and so the effective pick-up area of the rod, is increased.   Maximum flux density occurs at the centre of the rod.

If the aerial input coil of a receiver is wound on the centre of the ferrite rod maximum e.m.f. will be induced in the coil.   Thus the rod and coil form a small but efficient aerial, suitable for portable receivers.   The receiver input circuit can be tuned by moving the coil along the ferrite rod.   For maximum reception the axis of the rod must point in the direction of the transmitter.

Aircraft Aerials for Use in the M.F. and H.F. Bands

14.   With present day high-speed aircraft the aerodynamic properties of the aircraft have to be considered in the design of aircraft aerials.   For the m.f. band a simple wire aerial such as an inverted L or T rigged between the top of the tail fin and a small mast near the cockpit is normally used. Because of the short distance between the aerial and aircraft skin the polar diagram is not well defined.   When the aerial is used for transmitting, corona discharge from high voltage points can occur; when used for reception precipitation static interference, caused by electric charges from dust or rain discharging from pointed parts of the aircraft structure, is likely to be present.

15.   On higher speed aircraft suppressed aerials are commonly used.   These aerials are built into the structure of the aircraft so as to

preserve the streamlining and reduce drag caused by projections. Suppressed aerials fall into three basic classes, (a) concealed aerials, (b) insulated aerials, (c) aerials which use the aircraft skin as the radiating element. Radar scanners and slotted waveguide aerials mounted inside the aircraft and radiating through a radome, are examples of concealed aerials (a) in Fig. 9. An insulated aerial is illustrated at (b) in Fig. 9. A short section of the wing tip is insulated from the rest of the aircraft and when tuned with a variable inductor the tip can be made to resonate at the operating frequency. An example of the last class of aerial mentioned is the aerial shown at (c) in Fig. 9. An inductance situated in a cavity in the trailing edge of the wing is tuned to resonance with a capacitor. Coupling then occurs between the inductance and the aircraft structure and the aircraft surfaces radiate.
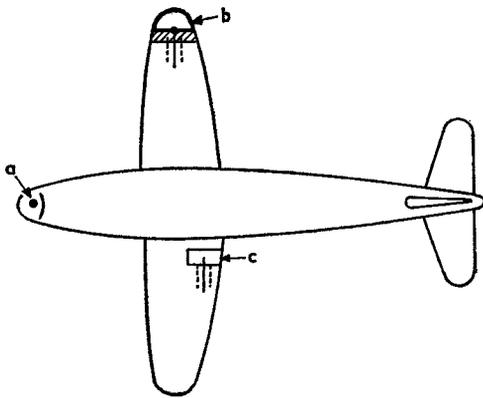
Fig. 9. AIRCRAFT SUPPRESSED AERIALS.

. A variation of this principle is used in the notch aerial. A notch in the skin of the aircraft is lined with copper foil and so forms a resonant cavity which directly excites the required portion of the aircraft skin.

Polar diagrams of suppressed aerials are difficult to predict and are ragged. Their shape depends on the size of the aircraft and the position of the aerial. In general the efficiency of suppressed aerials is low compared with conventional aerials.

V.H.F. and U.H.F. Aerials

16. Aerials for use in the v.h.f. and u.h.f. bands are generally made of hollow aluminium or copper tubing and as the diameter of the elements can be increased, so the band-width is increased. Since the wavelength is relatively small, aerials which are large in terms of wavelength and often possessing elaborate reflecting systems, can be used. In addition, such aerials mounted on fairly low masts are several wavelengths above the ground and their polar diagrams approach closely those of an aerial in free space. The basic theory still applies but impedance matching requires careful consideration. Co-axial feeders are preferred, and as pointed out in Chapter 4 of Section 15 this arrangement requires a balanced to unbalanced matching device. Transmission is almost solely by means of the space wave (Section 17) and therefore the range depends on the heights of the transmitter and receiver aerials. Horizontal or vertical polarisation can be employed.

17. **The Quarter-Wave Ground Plane Aerial.** As shown in Fig. 10 this aerial consists of a $\lambda/4$ radiating element with four radial rods joined to the outer conductor of the coaxial feeder and forming an artificial earth. The
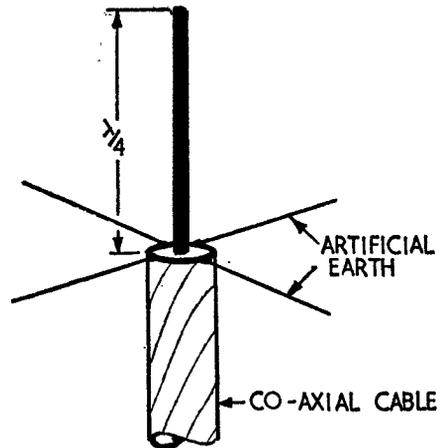
Fig. 10. THE QUARTER-WAVE GROUND PLANE AERIAL.

length of these rods is between $\lambda/4$ and $\lambda/8$, and sometimes a solid or semi-solid circular disc is used. The aerial gives all round radiation in the horizontal plane and low angle radiation in the vertical plane. If the diameter of the artificial earth is between one and two wavelengths the angle of elevation is increased.

18. **The Turnstile Aerial.** An aerial giving all round horizontally polarised radiation in

the horizontal plane suitable for broadcast transmissions, is the turnstile aerial shown in Fig. 11(a). Several tiers of crossed dipoles are mounted on a central mast. Fig. 11(b) shows the plan view of two crossed dipoles. Energy fed to dipole A is 90° out of phase with that fed to dipole B. In a direction θ degrees to A, the field due to A at a distant point P is proportional to sin θ and that due to B is proportional to cos θ. As the two fields are 90° out of phase with each other, the resultant

The beamwidth in the vertical plane can be made less by stacking the crossed aerials in tiers. This increases horizontal radiation and decreases vertical radiation.

19. **Biconical and discone aerials.** Aerials for use in the v.h.f. and u.h.f. bands are required to have a wide bandwidth, often of the order of 4:1 ratio. This requirement is obtained by designing conical-shaped aerials. The biconical aerial shown in Fig. 12(a) is a commonly used example of such aerials. It

(a)

(b)

(c)
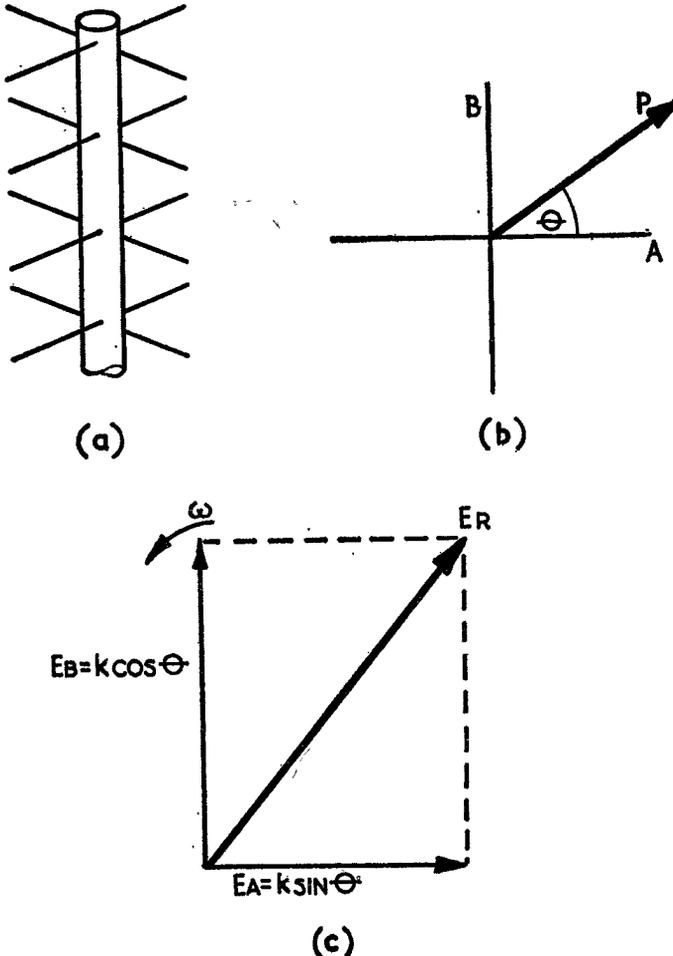
Fig. 11. THE TURNSTILE AERIAL.

field at P is the vector sum of two vectors at 90° to each other and of amplitude proportional to sin θ and cos θ. This is shown in Fig. 11(c). As sin θ decreases cos θ will increase and the resultant will remain constant i.e. the field strength in the horizontal plane around the aerial is constant.

will operate efficiently over the band for which the cone slant length (r) is between one quarter and one wavelength long. The cones are usually made in the form of a wire cage to reduce weight and wind resistance.

Another wide band conical aerial is the discone aerial shown in Fig. 12(b). It consists

of a disc D and a cone C and is fed by coaxial cable as shown. Its performance is similar to that of the biconical aerial.

(a) BICONICAL AERIAL
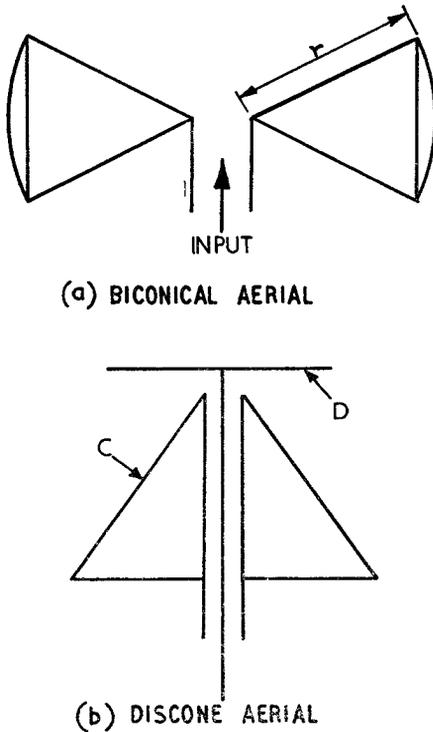
(b) DISCONE AERIAL

Fig. 12. WIDE–BAND U.H.F. AERIALS.

The discone and biconical aerials are used on u.h.f. ground installations employed on air-to-ground communications.

20. **The helical aerial.** This type of aerial provides a circularly polarised beam of radiation. It consists of a helix made from thick copper wire or tubing and is fed between one end and a ground plane as shown in Fig. 13(a). The aerial functions on the travelling wave principle and produces a polar diagram with maximum radiation in the direction of the axis of the helix as shown in Fig. 13(b). To obtain this polar diagram the diameter of the helix and the spacing between the turns must be carefully chosen.

21. **Aircraft slot aerials.** The slot aerial described in Chapter 2 is often used at v.h.f. and u.h.f. as an aircraft suppressed aerial.

A slot is cut in a cavity mounted in the aircraft structure, and when the cavity is excited at its resonant frequency the slot will radiate.

(a) HELICAL AERIAL

(b) POLAR DIAGRAM
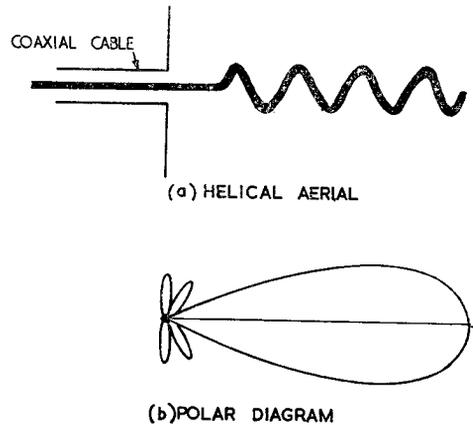
Fig. 13. THE HELICAL AERIAL.

S.H.F. or Microwave Aerials

22. In chapters 3 and 4 it was shown that to produce a narrow beam the array had to be large in comparison with the wavelength. Thus at low frequencies large cumbersome constructions are required for narrow-beam arrays. At centimetric wavelengths, however, a beam width of 2° can be produced with a comparatively small array. Such an array can be rotated at high speeds in any required plane, and is used widely in radar scanning equipments.

In radar, pencil-shaped beams and fan-shaped beams are often required. To produce such beams use is made of the reflecting properties of metallic sheets and the refracting properties of dielectrics.

Centimetric aerials are dealt with in detail in Part 3 of these Notes. Some of the more common types are shown in Fig. 14.

Summary

23. This section has provided a brief, overall picture of electro-magnetic radiation, and the devices employed to achieve it. The main points covered are summarised as follows.

Radiation of e.m. energy occurs from a single conductor if it is carrying a current of a sufficiently high frequency. The energy leaves the conductor with the speed of light in the form of electric and magnetic fields at right angles to each other and to the direction of propagation. The plane of the electric

RADIATING SLOTS

INPUT→

(a) WAVEGUIDE HORN WITH
PARABOLIC REFLECTOR

(b) SLOTTED WAVEGUIDE

WAVEGUIDE

DIELECTRIC

WAVEGUIDE

DIELECTRIC

(c) DIELECTRIC ROD

(d) DIELECTRIC HORN
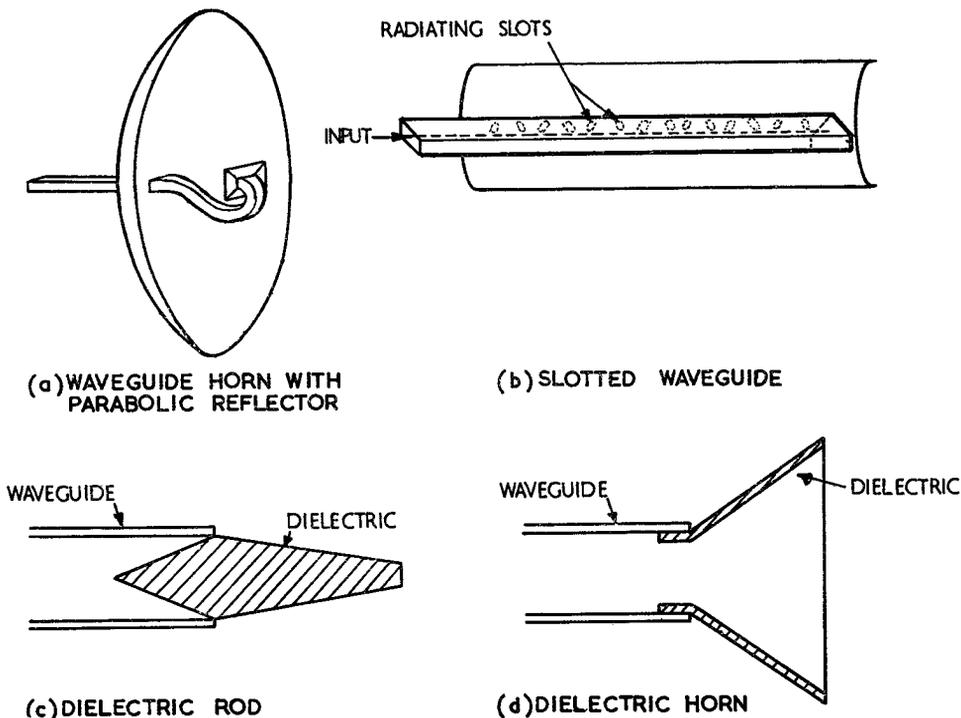
Fig. 14. CENTIMETRIC AERIALS.

field is taken as the plane of polarisation. The two fields are inter-dependent and in phase. At a distance from the aerial they form a plane wave, the strength of the fields being inversely proportional to distance from the source. They will cause currents to flow in a suitably designed receiver aerial.

The simplest form of standing wave radiator is the dipole which can be likened to an open oscillatory circuit. As the dipole dissipates energy in the form of e.m. radiation it can be said to have a resistance, which is called the radiation resistance of the aerial. The aerial also possesses a loss resistance, and the ratio of the radiation resistance to the total resistance is termed the aerial efficiency. The pattern that the radiated energy forms in space is called the polar diagram of the aerial. The Marconi quarter-wave aerial uses the reflecting properties of the earth to effect a reduction in the required length of the aerial. The effect the ground has depends on the condition of the soil and the frequency of the wave, and governs the shape of the polar diagram. The bandwidth of a standing wave

radiator can be increased by increasing its effective diameter.

The simple dipole can acquire directional properties by using it in conjunction with a reflector spaced behind the radiator and directors spaced in front. The beamwidth, the forward gain and the back-to-front ratio are means of stating the effectiveness of a directional array. Multiple driven aerials forming directional arrays include the broadside and end fire arrays.

Because of their simplicity of construction and wide bandwidth, travelling wave aerials are widely used at l.f., m.f. and h.f. They possess directional properties and the beam width decreases as their length, in terms of wavelength, increases.

This final chapter will have given the reader an idea of the wide variety of shapes and sizes that aerial arrays can take, depending on use and frequency of operation. Each aerial is designed for a specific purpose and new shapes can be expected as new needs arise. The basic theory involved in their design will, however, remain the same.

SECTION 17

PROPAGATION

SECTION 17

PROPAGATION

SECTION 17

CHAPTER 1

ELEMENTARY PROPAGATION THEORY

ELEMENTARY PROPAGATION THEORY

Introduction

1. Previous sections of this publication have dealt with the production of radio energy, its transfer from the source to the transmitting aerial, radiation from the aerial, and devices to ensure its efficient reception at a receiving station some distance away. This Section deals briefly with what can happen to the e.m. energy between the transmitter and receiver aerials, the different paths it can take and the factors which can affect it during its passage. This knowledge of the *propagation* of radio waves is necessary in order to understand the reasons for certain circuits in radio equipments, and why some frequencies are used in preference to others. It will also provide an appreciation of the performance limitations of different equipments, and will explain some phenomena which might otherwise be attributed to partial failure of the equipment.
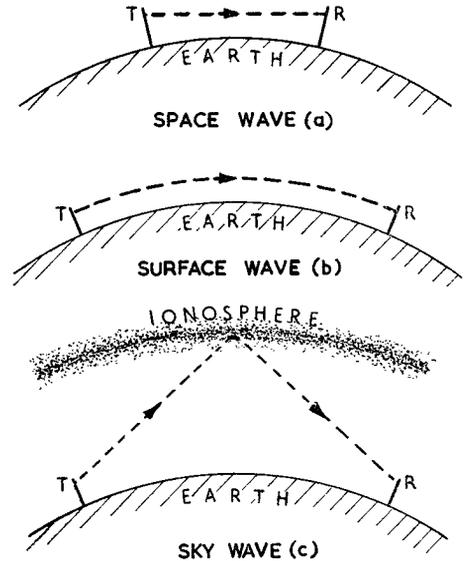
Types of Wave Paths

2. The energy radiated from a transmitter aerial travels outward into space at the velocity of light (3×10^8 metres/sec.). This energy can be considered to be made up of an infinite number of rays, similar to rays of light, all moving away from the aerial in straight lines. If the transmitter and receiver aerials are in line of sight, i.e. if they are situated above the earth's surface such that a direct line can be drawn between the two, the radio energy will take a direct path and voltages will be induced in the receiver aerial (Fig. 1(a)). This is called the *space wave*. If this were the only path the energy could

Fig. I. TYPES OF RAY PATH.

take, reception would not be possible at a receiver aerial situated below the horizon.

However, while e.m. waves in free space travel in straight lines, if they encounter an obstacle they are bent in the same way that light and sound waves are bent, and so a wave radiated close to the surface of the earth is bent by the earth and follows the curvature of the earth as shown in Fig. 1(b). This is called the *surface wave* and by using this path an increased range can be achieved. A combination of space and surface waves is termed the *ground wave*.

Frequency	Wavelength	Main Wave Path Employed
Below 300 kc/s	Longer than 1000m	Ground wave
300 kc/s to 3 Mc/s	1000m to 100m	Ground wave for short ranges. Sky wave for longer ranges
3 Mc/s to 30 Mc/s	100m to 10m	Sky wave
Above 30 Mc/s	Shorter than 10m	Space wave within line-of-sight

TABLE 1

The third possible path for the radio energy to take is the direct path from the aerial into the atmosphere as shown in Fig. 1(c). This is called the *sky wave*, and it would be lost in outer space if it were not for a reflecting surface in the earth's atmosphere, known as the ionosphere. This bends waves below certain frequencies back towards the earth's surface, and so under certain conditions, which will be dealt with more deeply in later paragraphs, reception due to this sky wave is possible at receiver stations below the horizon.

3. The main wave path taken by the radiated energy depends on frequency. Table 1 classifies these paths in relation to the given frequency bands, but it should be remembered that these frequency divisions are approximate, and no hard and fast dividing line can be drawn.

It is now proposed to discuss in greater detail the factors affecting propagation in these three wave paths, with particular reference to the frequency of the propagated energy.

The Ground Wave

4. As has already been stated, the ground wave can consist of the surface and space waves. The surface wave tends to follow the curvature of the earth, due to diffraction, in the same way that a sound wave is bent round an obstacle such as the corner of a building. With the surface wave, the obstacle is the earth and as the bottom of each wave front is in contact with the earth it is bent such that each wave front is tilted towards the one in front. This is shown in Fig. 2. The lower the radio frequency the
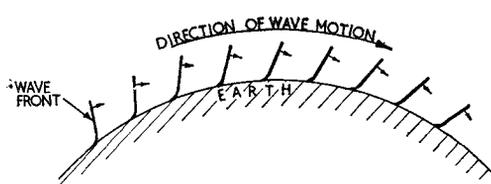
Fig. 2. BENDING OF A RAY BY DIFFRACTION.

greater is the bending, and at low frequencies the wave follows the curvature of the earth for considerable distances and an increased range results. As frequency increases, diffraction decreases and the range of the surface wave is less. For frequencies below

300 kc/s most of the energy is propagated by this surface wave, and regular reception is possible at all times of day and night up to ranges of several thousands of miles, dependent on radiated power.

Attenuation

5. There are two main causes for the reduction in field strength of a radio wave. The first, applicable to any ray path, is due to the natural spreading of the wave over large areas as it moves away from the aerial. This is called spatial attenuation and is proportional to the distance the wave is from the point of origin. The field strength of a radiated e.m. wave varies inversely as the distance from the transmitter, i.e. $E \propto \dfrac{1}{d}$.

The second cause of attenuation of the strength of the surface wave is termed absorption attenuation. As the wave travels over the surface of the earth, currents are induced in the earth and energy is absorbed from the wave. Thus the wave is attenuated as it progresses. The degree of attenuation is inversely proportional to the conductivity of the surface over which the wave travels. On the earth's surface water has the greatest conductivity and dry land the least. Maximum ranges are therefore obtained with waves propagated over the sea. The range decreases with propagation over wet land and becomes a minimum over dry desert. Furthermore the attenuation increases as the frequency of the wave being propagated increases, since the currents induced tend to flow in a thinner layer of earth (skin effect) and so the earth's conductivity to these currents decreases and losses increase.

6. Thus in addition to spatial attenuation, the two factors affecting the range of the surface wave are frequency, and type of terrain. The graphs of Fig. 3 show that the range progressively decreases as the frequency increases, and also that it is reduced over dry land compared to sea.

7. **Polarisation.** It will be remembered from Section 16 that the e.m. wave consists of an electric and magnetic field at right angles to each other and to the direction of propagation. The plane of polarisation is defined as the plane of the E field and a vertically polarised wave is radiated from a vertical dipole aerial. Since the conducting surface of the earth would tend to short-circuit the

Fig. 3. RANGE VARIATION WITH FREQUENCY
OVER DRY LAND AND SEA.

E field of a horizontally polarised wave, the surface wave would have a very short range. Therefore propagation by surface waves at low frequencies normally employs vertically polarised waves, and horizontal polarisation is confined to frequencies above 3 Mc/s.

While the state of polarisation remains constant in the surface wave, it may be altered

when passing through the upper regions of the atmosphere. This will be referred to in later paragraphs.

The Ionosphere

8. The atmosphere surrounding the earth is subjected to a bombardment of ultra-violet and cosmic rays emitted by the sun. The energy from these rays is absorbed by the atoms of gas in the atmosphere, and some of the orbiting electrons are displaced, forming positive ions and free electrons. Thus the gases are ionised and, due to the increased numbers of free electrons present, the conductivity is greatly increased compared with that of the normal gas. As the solar radiation penetrates deeper into the atmosphere so it becomes weaker, and ionisation decreases; furthermore, since pressure of the gases increases nearer the earth, the electrons recombine with the positive ions to re-form neutral atoms. In the upper regions however, where radiation is more intense and atoms more widely spaced, ionisation persists. The part of the atmosphere where ionisation occurs is called the *ionosphere*. It extends from approx-

Fig. 4. IONISATION LAYERS.

imately 30 miles above the earth's surface to at least 250 miles. In this region there are several ionised layers.

Ionisation Layers

9. There are three main layers of ionisation in the ionosphere, called the D, E and F layers in ascending order. They vary in degree of ionisation and in effective height, according to the time of day or night, and season of the year. At times the F layer can be divided into two layers, F_1 and F_2. Because of the limited rate of recombination in the higher regions, a certain degree of ionisation persists throughout the night, despite cessation of solar radiation.

An indication of the heights of these layers for winter and summer, day and night is given in Fig. 4 but it should be noted that the heights vary and are only approximate. During daytime the ionosphere consists of the D, E, F_1 and F_2 layers, whilst at night the D layer disappears and the F_1 and F_2 layers combine to form a single F layer. The degree of ionisation and hence the conductivity of the E layer decreases at night. The layer is then referred to as the residual E layer.

Propagation in the Ionosphere

10. **Refraction.** Before considering the effects that the ionised layers have on radio waves, it is necessary to recall a basic law of light. Fig. 5 shows a ray of light travelling from one medium (air) to a more dense
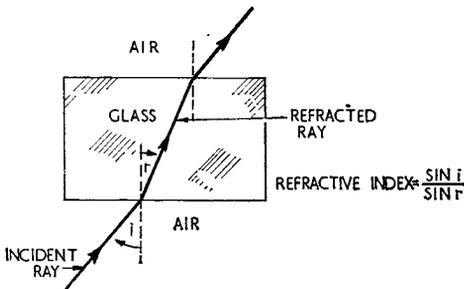
Fig. 5. REFRACTION OF LIGHT.

medium (glass). The ray of light is bent or refracted towards the normal to the horizontal glass surface. The angle i is termed the angle of incidence and r the angle of refraction. The ratio of sin i to sin r is the *refractive index* of the glass. On emerging from the glass and entering the less dense air the ray is again refracted, this time *away* from the normal to the glass surface. The angle

of refraction r depends upon (a) the angle of incidence i (b) the refractive indices of the two media, and (c) the frequency of the light.

11. Radio waves behave in a manner similar to that of light rays. An ionised layer has a lower refractive index than that of air; thus when a radio wave enters the layer it is refracted *away* from the normal, i.e. towards the earth. The angle of refraction will depend upon the angle of incidence, the frequency of the wave and the refractive index of the layer. The degree of ionisation, or the electron density of the layer throughout its depth, is not constant however, but varies with height, being greatest at the middle of the layer. The refractive index decreases as the electron density increases. Therefore, the refractive index of the ionised layer *decreases* towards the middle of the layer.

The refractive index n of an ionised layer for a wave of frequency f is given by:—

$$n = \sqrt{\left(1 - \frac{81N}{f^2}\right)},$$

where N is the number of electrons per cubic metre and f is measured in c/s. From this formula it can be seen that n decreases with an increase in electron density or a decrease in frequency.

It was stated in para. 10 that $n = \dfrac{\sin i}{\sin r}$. Thus since n decreases towards the middle of the layer, the angle of refraction r of a radio wave entering the layer at an angle of incidence i and of fixed frequency, will increase according to the expression $\sin r = \dfrac{\sin i}{n}$. When the right hand side of this equation equals unity i.e. when sin i = n, the angle r will equal 90° (Fig. 6). The ray will now have
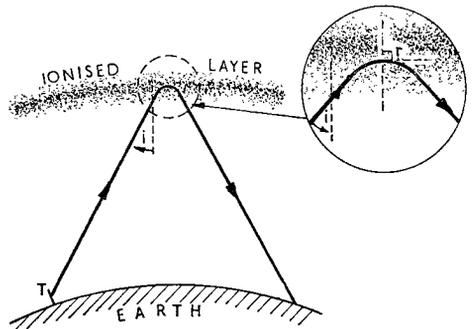
Fig. 6. REFRACTION OF A RADIO WAVE BY AN IONISED LAYER.

reached its highest point and will start its downward journey to the earth. Thus if the layer is deep enough, reflection of the wave occurs.

12. Critical Angle. If the angle of incidence with the vertical is increased (Fig. 7 ray a) the ray will not penetrate so far into the layer before being returned and will strike the earth at a greater distance from the transmitter. On the other hand, if the angle of incidence is reduced (Fig. 7 ray b) the ray penetrates further into the layer before being returned to earth, the ground distance covered being less. For a given frequency and given state of ionisation, an angle of incidence will be reached, below which the ray will not be reflected but will escape through the layer. This angle is called the *critical angle* (Fig. 7 ray c). An escape ray is shown in Fig. 7 ray d.

13. Critical Frequency. The refractive index of the ionised layer increases with an increase in frequency. Thus for a given angle of incidence a wave of higher frequency will penetrate deeper into the layer before being returned to earth. When the angle of incidence of a ray is reduced to zero, i.e.,

take place at vertical incidence. However, if the angle of incidence is increased from zero, then the frequency can be increased above the critical frequency and reflection back to earth still be achieved. For a given angle of incidence a required range can be obtained (Fig. 8). The maximum frequency which will be reflected at this angle of incidence and thus enable communication at the required range to be established, is called the *maximum usable frequency* (m.u.f.). If the frequency is increased above this the wave completely penetrates the layer and reflection does not occur.

Fig. 8 shows the ray path taken at the m.u.f. for a required range, and paths taken by a higher and lower frequency. Since the intensity of ionisation of the layer varies continuously, then if the m.u.f. were used there would be a risk of total penetration of the ionised layer at some time during the required reception period and communication would be intermittent. This can be overcome by choosing a frequency lower than the m.u.f. so that despite local variations in the refractive index of the layer, reliable communication can be ensured. Losses in the ionosphere increase with a decrease in frequency however, and if too low a frequency were chosen losses would be high. Generally a frequency of about 85% of the m.u.f. is chosen. This is called the *optimum working frequency* (o.w.f.).

No regular sky-wave signals can be expected from frequencies above 30 Mc/s. Even for an angle of incidence such that the ray just glances the layer, a wave of higher frequency than this passes through the

Fig. 7. EFFECT OF DIFFERENT ANGLES OF INCIDENCE (FREQUENCY CONSTANT).

when it enters the layer at right angles, and the frequency is increased there will be a frequency beyond which the wave will not be returned to earth, but will completely penetrate the layer. This frequency is called the *critical frequency*.

Maximum Usable Frequency

14. The critical frequency is thus the maximum frequency at which reflection can

Fig. 8. MAXIMUM USABLE FREQUENCY.

ionosphere without sufficient refraction to return it to earth.

15. **Skip-Distance.** At the m.u.f. the return wave from the ionised layer is refracted just enough to strike the earth at the receiver aerial. For greater angles of incidence the wave will strike the earth beyond the receiver aerial but for smaller angles of incidence the wave will penetrate the layer and will not be returned to earth. Thus every frequency has a shortest distance at which it will return to earth, for a given layer height. This is called the *skip-distance* (Fig. 9) and increases as the frequency increases. Within this distance from the transmitter no sky wave reception is possible, and if reception is required, the frequency must be lowered. The skip distance for a given m.u.f. also varies with the height of the ionised layer and degree of ionisation, and therefore with the time of day, night and season of the year.
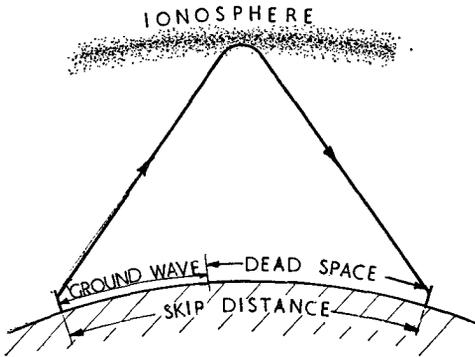
Fig. 9. SKIP DISTANCE AND DEAD SPACE.

Ground wave reception is still possible over part of this distance, but the ground wave suffers greater attenuation than the sky wave and after a certain distance will die out. Between this point and the receiver, no reception from either ground or sky wave is possible. This distance is called the *dead space* (Fig. 9). For high frequencies and long hops the dead space is approximately equal to the skip distance since the range of the ground wave is short. Hence, sometimes the term skip distance is confused with that of dead space. Note that while an increase in the power output of the transmitter might increase the range of the ground wave, it will make no difference to the point of the first sky wave return. This latter is dependent on frequency and degree of ionisation only.

Regular Ionospheric Variations

16. Because of the continuous changes in the earth's position relative to the sun, the ionisation density of the layers is subject to daily and seasonal changes. This causes variations in the m.u.f. for a required range, from hour to hour during the day and night and during summer and winter. Fig. 10 shows a set of curves relating m.u.f., local time and skip distance (d) for the month of March 1960. Each curve represents a different skip distance and is plotted for the frequency which gives this distance at different times of the day.

For example, using the chart of Fig. 10, the skip distance for a frequency of 6 Mc/s at 12 noon is 200 miles, while if the frequency is increased to 23 Mc/s, the skip distance would rise to 1200 miles. Thus for point to point working at a range of 1200 miles, the frequency chosen must be such that the skip distance is *less* than 1200 miles. The best frequency to use around noon would be 19 Mc/s (85% of the m.u.f. of 23 Mc/s) and overnight a frequency of 8–9 Mc/s would give a skip distance less than the required 1200 miles thus ensuring that the receiving station is within the reception distance. These frequencies are called the day and night frequencies, the latter always being lower than the former. By using two or more frequencies over the 24 hours, the skip distance is maintained sensibly constant and uninterrupted communication is maintained.
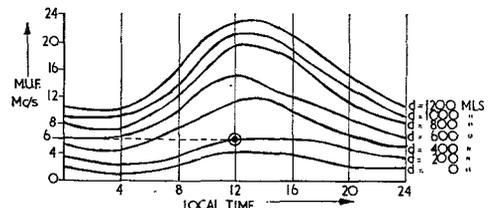
Fig. 10. M.U.F. VARIATION WITH TIME AND SKIP DISTANCE OVER 24 HOURS DURING MARCH, 1960.

Multiple Hop Propagation

17. For communication over distances greater than 2500 miles, propagation takes place by successive reflections from the ionosphere and the earth's surface. This form of propagation is termed multiple hop propagation and very long ranges can be obtained by using waves travelling in this manner (Fig. 11). It should be noted that

with multiple hop and single hop propagation, the time, as indicated by the sun's position, at the points of reflection in the ionosphere is different from that at the transmitter, and thus the refractive index of the layer at these points is different. This factor has to be considered when choosing a suitable frequency.
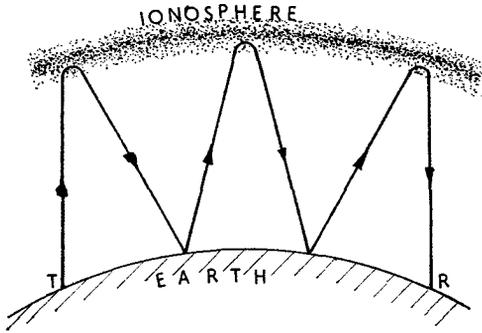
Fig. 11. MULTIPLE HOP PROPAGATION.

Fading

18. If the resultant voltage induced in the receiver aerial varies in amplitude the output of the receiver will vary. This fluctuation in receiver output is called fading. The different causes and forms of fading will now be considered.

In addition to the fairly regular variation in the intensity of ionisation already discussed, the degree of ionisation is subject to continuous slight variations. The signal arriving at the receiver aerial is made up of a large number of rays, each of which has travelled through the ionosphere by slightly different paths. As these ray paths will vary with the degree of ionisation, they will suffer slight variations, and so the phase of the rays arriving at the receiver aerial will vary in a random manner. The amplitude of their resultant thus varies continuously, giving rise to what is known as *interference fading*.

Another factor contributing to the fading of the sky wave is the random change in the plane of polarisation of the wave as it passes through the ionosphere. As the amplitude of the voltage induced in the receiver aerial depends on the plane of polarisation of the wave, the signal strength at the receiver will vary from instant to instant. Automatic gain control circuits are included in the receiver in order to reduce this form of fading and in

certain systems diversity reception is employed.

19. **Selective Fading.** If the wave being propagated is amplitude modulated, then it contains a variety of frequencies in the sidebands. Since the path length in the ionosphere varies with frequency, phasing and attenuation of the component frequencies of the signal will vary. This results in distortion of the intelligence contained in the signal.

Irregular Ionospheric Disturbances

20. **Sunspots.** Since the ionisation of the earth's upper atmosphere is caused mainly by ultra-violet radiation from the sun, and since the latter increases with the sunspot activity, the number of sunspots affects the propagation of radio waves. It has been noticed over the years, that in addition to daily and seasonal variations, the intensity of ionisation is subject to long period variations, one cycle of variation from minimum to maximum intensity taking approximately eleven years. The variation is by no means uniform throughout the sunspot cycle, but in general the m.u.f. increases as the sunspot activity increases.

21. **Solar Flares.** Fairly frequently there occur eruptions of the surface of the sun which are known as *solar flares*. These result in an increase in ultra-violet radiation, which produces intense ionisation of the D layer. The flares themselves last for one or two minutes only, but the ionisation effects continue for an hour or more. The increase in electron density in the D layer causes increased absorption of the wave entering the layer, and complete and sudden fadeout of all sky wave reception. Communication returns after one or two hours, more gradually than it ceased. There is as yet no method of predicting the onset of these fadeouts.

22. **Ionospheric Storms.** These constitute the main form of disturbance to sky wave communication. While their effect is not so intense as that produced by solar flares, the storms last for much longer periods. During ionospheric storms, radiation from the sun is in the form of minute ionised particles beamed towards the earth, as well as the normal ultra-violet radiation. This radiation commences at the same time as solar flares, but the particles travel at a much slower

speed than the ultra-violet waves, and therefore arrive at the earth's upper atmosphere about thirty hours after the flare has been observed. On entering the ionosphere the particles are affected by the earth's magnetic field and are carried towards the magnetic poles. The F layer rises in height and the ionisation intensity decreases, allowing waves which would normally be refracted back to earth to penetrate the layer. At the same time absorption of the waves by the lower ionised layers increases. Ionospheric storms cause subnormal sky wave communication for periods of two or more days, and can be roughly forecast by observing the solar flares at the sun's centre.

23. Magnetic Storms. These are closely associated with ionospheric storms, and take the form of violent fluctuations in the earth's magnetic field, caused by ionised particles from the sun forming large currents in the ionosphere. Since the particles are carried towards the magnetic poles these areas are most severely affected, and cause polar aurorae.

24. Meteor Ionisation. Meteors are small particles of matter entering the earth's atmosphere from outer space at very high speeds, and frequently in showers. They produce a local volume of increased ionisation in their track, which persists for short periods of up to 20 seconds, at heights of approximately 75 miles. Thus they cause fluctuations in the ionisation of the layers, and some techniques employ the meteor trails as reflecting surfaces to increase the range of high frequency transmission. Sudden ionospheric disturbances due to solar flares, ionospheric storms, magnetic storms and meteor ionisation are often referred to by the abbreviation s.i.d.

Tropospheric Propagation

25. The region of the earth's atmosphere immediately above the earth's surface is called the troposphere. This region extends to approximately 36,000 feet, and its upper limit is termed the tropopause. In this region normal meteorological conditions are found with currents of air forming winds, and temperatures decreasing with height. Propagation of e.m. waves in the troposphere will now be considered.

26. Regular sky wave reception is possible only at frequencies below approximately 30 Mc/s. Except under special conditions, a wave of higher frequency than this passes through the ionosphere and is lost. Therefore reception of signals at v.h.f. and above is almost entirely dependent on the space wave.

The space wave consists of a direct ray and a ray reflected from the surface of the earth, as shown in Fig. 12(a). At high frequencies the aerial is small enough to enable it to be raised to heights of at least one wave length above the surface of the earth, and at this height, the surface wave can be neglected in comparison to the space wave. Thus the space wave provides the only effective means of propagation at v.h.f. and above. Owing to slight variations in the density of the troposphere with height, the space wave suffers a small amount of refraction, and so the range is increased above line of sight by a factor of about $\frac{4}{3}$.

27. Direct and reflected ray paths. Since the wave arriving at the receiver aerial consists of the direct and reflected rays (Fig. 12(a)) the voltage induced in the aerial will be the vector sum of the components of these two rays. Therefore the relative phases of the e.m. energy at the receiver aerial, from these two rays, governs the strength of the received signal; if they arrive in phase, reception will be maximum; if they arrive 180° out of phase, it will be a minimum. The phase difference between these two rays will depend on the difference in path lengths and the change in phase on reflection at the earth's surface.

The phase change occuring on reflection depends on whether the wave is horizontally

(a) DIRECT & REFLECTED RAY PATH

(b) BREWSTER ANGLE
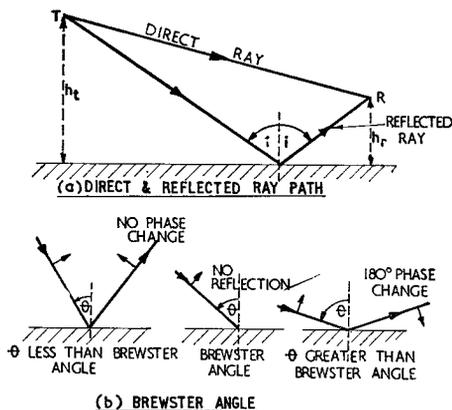
Fig. 12. REFLECTION OF A VERTICALLY POLARISED WAVE.

or vertically polarised. If horizontally polarised, a phase change of 180° occurs, no matter what the angle of incidence may be. If the wave is vertically polarised however, the phase change depends on the angle of incidence and the nature of the reflecting surface. At one critical angle of incidence, called the Brewster angle, no reflection occurs (see Sect. 16, Chap. 2, Para. 24). If the angle of incidence is such that the angle made to the normal is less than the Brewster angle, reflection occurs with no phase change, whilst if greater than the Brewster angle, 180° phase change occurs. (Fig. 12(b).)

Therefore, the factors which govern the signal strength at the receiver aerial are:—

(*a*) Heights of transmitter and receiver aerials above ground.

(*b*) Distance apart of the transmitter and receiver aerials.

(*c*) The wavelength (or frequency) of the radiated energy.

(*d*) The polarisation of the radiated energy.

(*e*) The transmitter power.

(*f*) The nature of the reflecting surface.

Scatter Propagation

28. **Ionospheric Scatter.** A form of propagation employed at v.h.f. to extend the range of the space wave beyond the optical horizon is known as *scatter propagation.* The E layer is in a continuous state of turbulence and this gives rise to localised clouds of ionisation. These clouds or 'blobs' of ionised gas have a lower refractive index than the surrounding gas. If the wavelength of the energy being propagated is comparable with these irregularities, then while most of the energy will penetrate the ionised layer, some will scatter in a narrow cone surrounding the forward direction of propagation of the incident wave (Fig. 13). In order to receive a scattered signal at a point beyond the horizon, high transmitter power and highly directional transmitter and receiver aerials must be employed. Best results are obtained when the scatter angle is small, of the order of 2°. Maximum range (about 1200 miles) occurs when the transmitted beam is directed at the horizon. The minimum range is about 600 miles, and the useful frequency range for this *ionospheric scatter*

propagation is between 30 and 60 Mc/s. Communication using this form of propagation is not affected by s.i.d.s. and may be maintained when h.f. stations are blanked out. Because of the random nature of the scattering the received signal is subject to fading, but the effect of this is reduced by using a.g.c. in the receiver and diversity reception, i.e. using several receiver aerials spaced about 100 feet apart and feeding separate receivers.
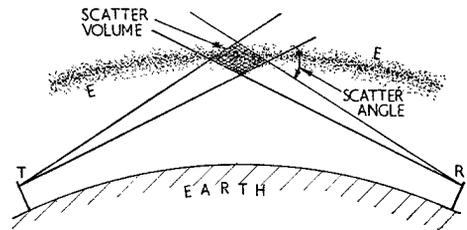
Fig. 13. IONOSPHERIC SCATTER PROPAGATION.

One system at present employing ionospheric scatter techniques is the Janet system. This uses ionised meteor trails to reflect the waves, and ranges of 600 to 1250 miles are obtained. Communication is reliable and subject to little interference. The system is most suitable for telegraphy and other narrow band transmissions.

29. **Tropospheric Scatter.** With frequencies of the order of 300 to 2000 Mc/s, localised variations of refractive index caused by meteorological conditions enable scatter propagation to be achieved in the tropopause. The height of the scatter volume is about 40,000 feet and therefore the range is less than that obtained with ionospheric scatter, being from 75 to 400 miles. Again high power transmitters and directional transmitter and receiver aerials directed at the horizon are necessary. Bandwidths of from 3 to 4 Mc/s can be transmitted up to 200 miles, but the useful bandwidth decreases considerably as range increases. Diversity reception is employed.

Whilst scatter propagation installations have a high initial cost compared with h.f. systems, their main advantage lies in the reliable long range communication they provide which is unaffected by variations in the ionosphere caused by s.i.d.s.

Super Refraction

30. Under certain abnormal climatic conditions, temperature can increase and humidity decrease with height. When this occurs the refractive index may decrease with height, and a duct is formed between the earth and a hundred or so feet above the earth. If the wavelength of the energy being propagated is small compared to the height of this duct, the waves are trapped in the duct and reflections occur as shown in Fig. 14. Thus at v.h.f., u.h.f., and s.h.f., ranges far exceeding the optical line of sight are obtained; for example, irregular reception of London television programmes in South Africa have been attributed to this phenomenon which is known as *super refraction*. An adverse effect of super refraction is encountered in radar range measuring equipments, when echoes are received from ranges far exceeding the design maximum range of the equipment. This results in false targets being indicated.

Weather conditions which produce super refraction exist mainly in tropical and sub-tropical climates. In temperate climates, the weather must be fine and settled, when centimetric waves are most likely to be affected.
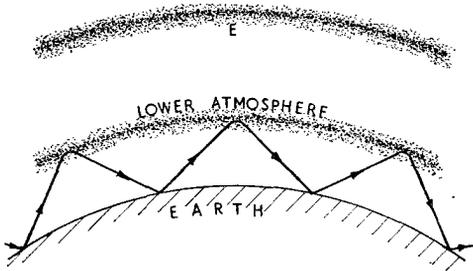
Fig. 14. SUPER REFRACTION.

Effects of Obstructions at v.h.f.

31. The surface wave and the sky wave are only slightly affected by obstructions. At very high frequencies and above however, the space wave is seriously attenuated by buildings, trees and mountainous terrain; these cause absorption, reflection and diffraction, and cause ' blind spots ' at comparatively short distances from the transmitter. Moving the transmitter or receiver a few yards out of these blind spots often results in good reception.

Summary

32. To summarise the different means of propagation, Table 2 has been constructed. This divides radio frequencies from 10 kc/s to 300 Mc/s into five bands; the main features of propagation in each band are summarised under these band headings.

Band	Frequency	Wavelength
A	10 kc/s to 300 kc/s	30,000m to 1000m
B	300 kc/s to 3 Mc/s	1000m to 100m
C	3 Mc/s to 30 Mc/s	100m to 10m
D	30 Mc/s to 300 Mc/s	10m to 1m
E	Above 300 Mc/s	Below 1m

TABLE 2

Band A. This is the long wave band and propagation by ground wave gives reliable and stable reception for ranges up to 1000 miles. At these long wavelengths the ionised layer may be considered as an almost perfect reflecting surface and greatly increased ranges are obtained from a ray reflected between the lower edge of the ionosphere and the surface of the earth. Frequencies in band A are used for very high grade transoceanic telegraph communications and are practically free from fading.

Communication in this band is subject to atmospheric interference caused by electrical discharges from clouds, and from man-made interference caused by unsuppressed electrical apparatus.

Band B. This is the broadcast band and provides high quality reception within a limited area. A station of medium power has a service radius of the order of 100 miles, the signal being provided by the ground wave. Beyond this is a zone of fading where signals are not normally received in daylight beyond about 150 miles from the transmitter over land, or 600 miles over water. During night time, when ionisation is low, the sky wave is only slightly attenuated and reception (subject to severe fading) is possible up to 1000 miles. The portion of the band between 1·5 Mc/s and 3 Mc/s is erratic and unsatisfactory for distant communication because of large losses in the ionosphere.

Frequencies in this band are subject to the same type of interference that affects band A frequencies.

Band C. The short-wave band is used for long distance communication. The sky wave is used, and frequencies are selected by means of m.u.f. prediction charts. The part of the band from 3 Mc/s to 6 Mc/s is used for communication within the limits of a continent, and above 6 Mc/s, by reason of the longer skip distances, for inter-continental communications. Short ranges inside the skip distance are covered by the ground wave.

Atmospheric and man-made interference is not nearly so severe on frequencies in this band. However, above 20 Mc/s, radiation from the sun and from nearby galaxies produces a ' hissing ' interference noise.

Band D. For frequencies above 30 Mc/s, the sky wave is not regularly returned to the ground, and the ground wave is rapidly attenuated. Communication is maintained within, or somewhat beyond the optical horizon, by means of the space wave. Short range reception within built up areas is possible, making this band usable for short range communication services, television and v.h.f. broadcasting. Greatly increased line-of-sight ranges are obtained in communication between high-flying aircraft and ground stations.

Frequencies up to 60 Mc/s are used for ionospheric scatter propagation and ranges of the order of 1000 miles are obtained.

Band E. Frequencies in this band provide line-of-sight reception between two stations, and are widely used for radar, navigational aids and landing aids. Development of u.h.f. t.v. and f.m. sound broadcast transmissions using frequencies in this band is progressing. Both bands D and E are virtually free from atmospheric interference, but they are affected by solar and galactic interference.

Above 500 Mc/s tropospheric scatter provides limited reception up to ranges of 400 miles.

33. This Chapter has dealt with some of the basic aspects of wave propagation of interest to the radio technician. It has touched on some of the techniques employed to obtain more effective and stable means of communication by using the ionisation layers. Considerable information about the ionosphere still remains to be learnt and improved techniques may be expected.

SECTION 18

RADIO MEASUREMENTS

SECTION 18

RADIO MEASUREMENTS

CURRENT, VOLTAGE AND FREQUENCY MEASUREMENTS

Introduction

1. The requirements of instruments capable of measuring current, voltage and power were discussed in Book 1, Section 6 and a selection of such instruments was considered in detail. It was noted that certain instruments were used for measuring d.c. quantities only, while others could measure d.c. and a.c. quantities. In radio engineering, a.c. quantities vary in frequency from a few cycles per second to radio frequencies of 10,000 Mc/s and above. Instruments to measure quantities over this range of frequencies are available to the radio technician and a knowledge of their function and applications is essential. The main measurements necessary to assist in servicing and repairing modern radio equipment are measurement of current, voltage, frequency, modulation, standing wave and power. This Section will deal with instruments which make these measurements, and how these instruments are used.

Current Measurement

2. The most commonly used meter for measuring currents at radio frequencies is the thermo-junction meter. It is capable of measuring very small currents over a wide range of radio frequencies, and its current range can be conveniently adjusted by means of shunts. Its predecessor, the hot wire ammeter, is rarely used nowadays because of its comparative insensitivity.

Voltage Measurement

3. Any instrument which is required to measure voltage must have a high internal resistance compared to the resistance of the source of voltage to be measured. The internal resistance of a multimeter increases as the voltage range is increased. For example a multimeter with a rating of 1000 ohms per volt has an internal resistance of 10 kΩ on the 10 volt range and 1 MΩ on the 1000 volt range. Thus in measuring a p.d. of 5 volts across a 50 kΩ resistor using the 10 volt range, the multimeter would heavily shunt the resistor and a low reading would result. If the 1000 volt range were used the internal resistance of the multimeter would be sufficiently high to have little shunting effect, but the meter deflection would be so small as to be unreadable. Thus for reading small voltages across large value resistors a multimeter is of little use. A far more accurate instrument for such measurements is the valve voltmeter.
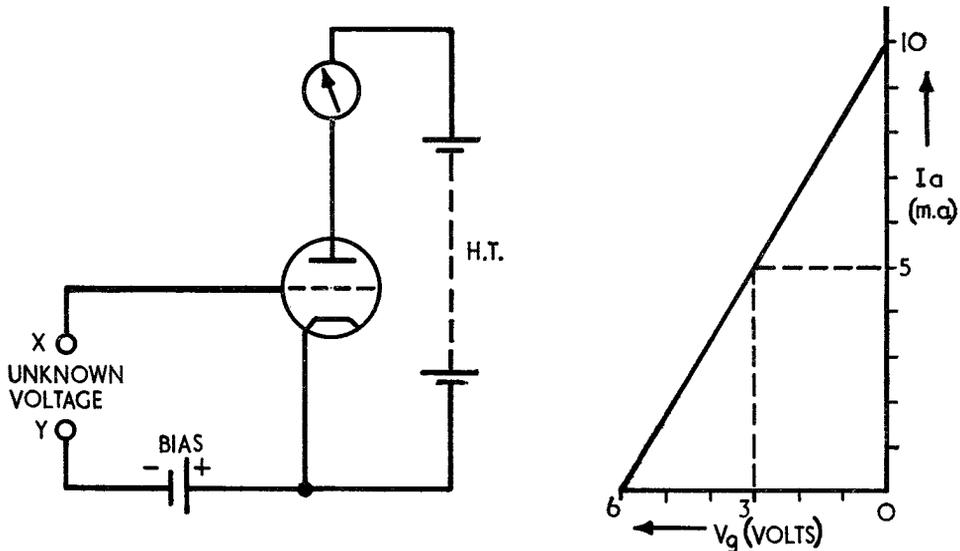
Fig. 1. PRINCIPLE OF THE VALVE VOLTMETER.

SECTION 18

CHAPTER 1

CURRENT, VOLTAGE AND FREQUENCY MEASUREMENTS

4. Valve Voltmeter. This instrument is widely used to measure d.c. voltages and a.c. voltages of frequencies up to 50 Mc/s. It has a very high internal resistance and therefore places very little load on the source of the voltage being measured.

The principle of the valve voltmeter is illustrated in Fig. 1. With zero voltage across the input terminals XY, the valve is biased to cut-off (–6V) and as no valve current flows, the meter in the anode circuit will read zero. If a d.c. voltage of 3V is applied to XY an anode current of 5 m.a. will flow through the meter. The value of current flowing through the meter is directly proportional to the applied voltage, and the meter can be calibrated directly in volts. To measure an alternating voltage, either the meter could be calibrated to read peak voltage, or a capacitor, which would charge up to the peak applied voltage, can be placed in the grid lead.

5. A practical circuit for a valve voltmeter is shown in Fig. 2. The input voltage to be measured is developed across the potential divider which has tappings for the different voltage ranges. This voltage is applied to the grid of a cathode follower V_1, which has a d.c. current measuring meter in its cathode circuit. Valve current proportional to the change in grid voltage will flow, and the meter which is calibrated in volts will give a

measure of the input voltage. The three input terminals 'LOW,' 'D.C. HIGH' and 'A.C. HIGH' are used for measuring direct and alternating voltages, the 'LOW' terminal being connected to the low potential side of the circuit being tested (usually the chassis).

A bias voltage in the cathode of V_1, in conjunction with a potential divider across h.t. provides a means of reducing the standing anode current to zero when the input terminals are shorted. Thus before making any voltage measurement the appropriate 'HIGH' terminal must be shorted to the 'LOW' terminal and the SET ZERO control adjusted to give zero reading in the meter.

6. For d.c. voltage measurement the 'HIGH' terminal is connected directly to the top of the potential divider network. To measure a.c. voltages up to a frequency of 10 kc/s a diode rectifier followed by a filter network is built into the meter. The d.c. blocking capacitor C_1 charges to the peak input voltage and after smoothing out the a.c. ripple by means of the filter $R_1 C_2$, the rectified voltage is applied across the potential divider and fed to the grid of V_1.

To measure r.f. voltages of frequencies up to 300 Mc/s an external probe is usually available as part of the instrument. The probe consists of a diode rectifier and filter
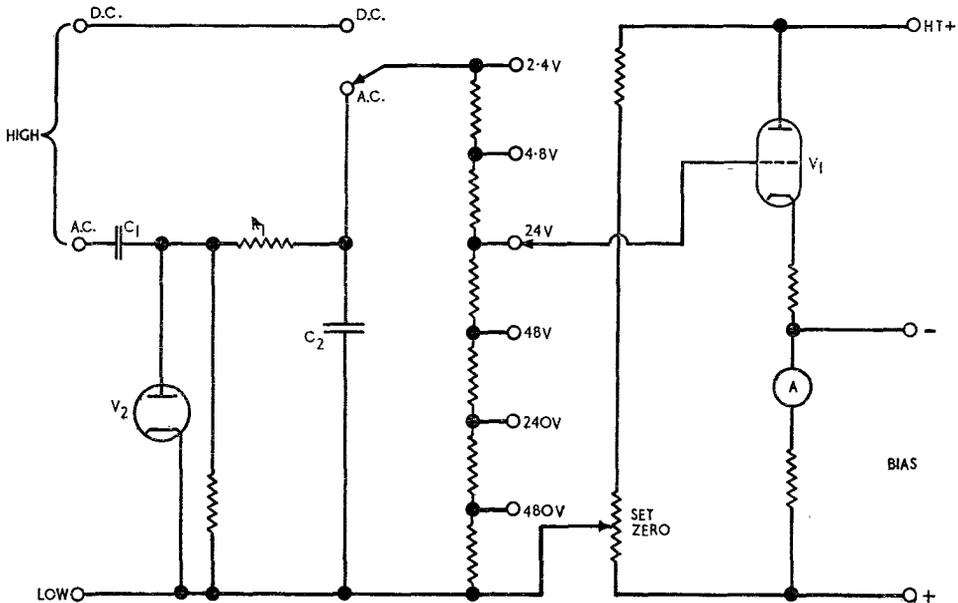
Fig. 2. BASIC CIRCUIT OF A VALVE VOLTMETER.

network similar to that shown in Fig. 2, but in order to reduce the resistance and capacitance of the test leads, it is placed at the end of the ' HIGH ' lead remote from the valve voltmeter. Thus the probe is placed directly at the point of measurement and the only limitation to the frequency of the voltages which can be measured, is set by the input capacitance of the diode. By using a germanium or silicon diode and reducing the values of the input filter capacitors, voltages at frequencies up to 500 Mc/s can be accurately measured.

The valve voltmeter is a versatile instrument capable of giving accurate voltage measurements over a wide frequency range. It is the only suitable instrument for making r.f. voltage measurements.

FREQUENCY MEASUREMENT

Introduction

7. One of the most important and critical measurements necessary in radio engineering is the measurement of frequency. Transmitters and receivers are normally required to operate on a specific frequency and there is a variety of instruments which enables the radio technician to set the equipment to the desired frequency. A selection of these instruments will now be considered.

Absorption Type Wavemeter

8. This instrument is normally used when making preliminary adjustments to high power transmitters, or for determining their approximate fundamental frequency. It consists of a simple oscillatory circuit with a variable capacitor and a device to indicate maximum circuit current. It takes power from the circuit under test, hence its name.

The circuit of Fig. 3 shows a typical absorption type wavemeter. The coil L is externally mounted and the wavemeter
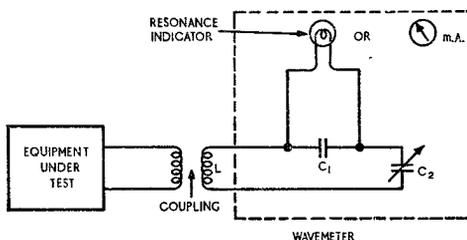
Fig. 3. ABSORPTION TYPE WAVEMETER

is placed so that there is magnetic coupling between it and the equipment under test. C_2 has a calibrated dial and is varied until the circuit LC_2 resonates, when the voltage across C_1 is maximum and the lamp lights. The capacitor C_1 has a much larger value than C_2 and therefore a much lower reactance to any frequency. Thus it has little effect on the calibration of LC_2. As LC_2 approaches resonance the lamp will glow dimly, and the coupling between the wavemeter and the equipment being tested must be decreased (i.e. the distance between them increased) to prevent the lamp burning out. Since the wavemeter is loading the circuit under test, maximum accuracy will be obtained when coupling between the two is at a minimum.

Reaction Type Wavemeter

9. This instrument is similar to the absorption type wavemeter, but uses a resonance indicator in the circuit under test. The coil L is externally mounted and, in use, is coupled to the output circuit of the equipment whose frequency is being measured (Fig. 4). The wavemeter is then tuned by varying C until resonance is indicated in
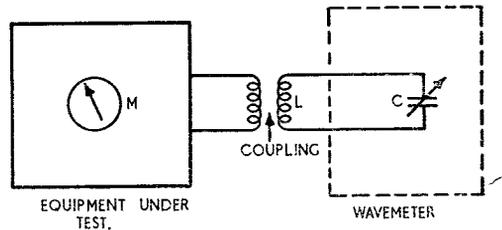
Fig. 4. REACTION TYPE WAVEMETER.

meter M, either by a maximum or minimum reading depending on the position of M in the circuit. The capacitor C usually has a vernier calibration, and the frequency can be read from either graphs or charts issued with the wavemeter and bearing its serial number. In order to get a sharp indication of resonance, the coupling should be reduced to an absolute minimum.

This type of wavemeter absorbs little power from the circuit under test and is therefore used for checking low power equipments. It is not as accurate as the absorption type wavemeter but is useful in checking that an oscillator is on the correct fundamental frequency.

INPUT SIGNAL
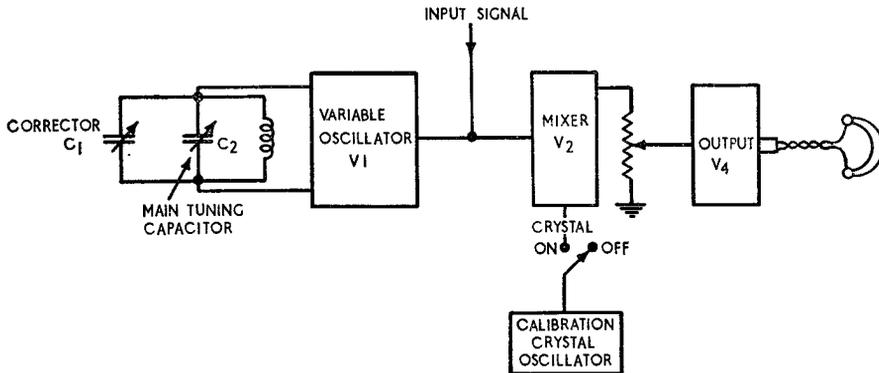
Fig. 5. BLOCK DIAGRAM OF A HETERODYNE FREQUENCY METER.

Heterodyne Frequency Meter

10. This type of frequency measuring instrument is used when a high degree of accuracy is required, such as when setting a transmitter or a receiver to a required frequency or when measuring the frequency of a transmitter. It consists of a calibrated tunable oscillator, the output of which is mixed with the frequency under test. When a zero beat is produced the two frequencies are the same and can be read from a chart supplied with the meter.

11. **Calibration of the variable oscillator.** The variable oscillator is usually a conventional circuit based on one of the types described in Book 2, Section 12, Chapter 1, and while its tuning dial will have been accurately calibrated by the manufacturer, its accuracy will vary with time and working conditions. For example a change in temperature will vary the LC values; a replaced valve will introduce different values of stray capacitances, and of course a change in power supplies will cause the frequency of the oscillator to change from the calibrated frequency. To provide the required degree of accuracy the frequency of the variable oscillator must therefore be checked against the frequency from an accurate stable oscillator, such as a crystal oscillator, *whenever a new frequency has to be set up*. Before making any tuning adjustments to the heterodyne frequency meter, the instrument should be switched on and allowed to warm up for at least 30 minutes. It will then be at its normal working temperature and the frequency chosen will be stable and will not drift. The calibrated crystal oscillator is sometimes built into the meter, and works

on a typical fundamental frequency of 1000 kc/s. Harmonics of this fundamental are used to obtain a zero beat with the variable oscillator by varying the corrector capacitor C_1 (see Fig. 5). These harmonic settings are called crystal check points and serve to calibrate the variable oscillator at frequent intervals throughout the tuning range.

For example, to check the calibration of the variable oscillator against a check point of 7000 kc/s. The main dial setting for C_2 is found in the calibration book, and C_2 is set to this reading. The crystal oscillator is switched into circuit, and will produce oscillations on the 7th harmonic of the fundamental, i.e. it will produce an output at 7000 kc/s. Assuming that the setting of C_2 causes the variable oscillator to oscillate at $7000 \cdot 5$ kc/s a beat note of 500 c/s will be heard in the headset. By adjusting the corrector control C_1 to produce zero beat, the variable oscillator will oscillate at exactly 7000 kc/s. The crystal oscillator is now switched off and the variable oscillator is accurately calibrated.

12. **Operation of heterodyne frequency meter.** In the simple block schematic diagram of Fig. 5, V_1 is the variable oscillator valve. The output from this oscillator is fed, with the input frequency under test, to the mixer valve V_2. If there is any frequency difference between the two, a beat note will be heard in the headset, which is in the output of V_4, the audio amplifier. By adjusting for zero beat with the main tuning capacitor C_2, the capacitor dial reading will give, in conjunction with the calibration book, the exact frequency of the input.

13. To set the master oscillator of a transmitter to a specific frequency, the variable oscillator is checked against a crystal check point near the required frequency, and after adjusting the corrector control, the main tuning capacitor C_2 is set by means of the calibration chart, to the required frequency. The output from the transmitter is then loosely coupled, usually by means of a short aerial, into the mixer. The transmitter master oscillator control is then adjusted for zero beat, when the transmitter will be on the required frequency.

14. A receiver can be accurately set to a particular frequency by setting the heterodyne frequency meter to the required frequency and coupling the output of V_1, again by means of a short aerial, to the receiver input. The receiver tuning control is then adjusted for zero beat in the receiver loudspeaker or headphones, when the receiver is tuned to the required frequency.

Counter Method of Frequency Measurement

15. The accuracy of the heterodyne frequency meter is limited by the need to interpolate between harmonics of the calibration oscillator. An instrument which is superseding the heterodyne frequency meter is the *electronic counter*. This instrument uses a standard frequency source in conjunction with counter circuits, and provides a means of making rapid frequency measurements which are accurate to 1 part in 10^7. Its operation requires little skill.

16. **Principle.** The principle is simply that of counting the number of cycles of the unknown frequency that occur in an accurately measured period of time. Fig. 6 shows a block diagram of the basic circuit arrangement. A standard frequency oscillator, usually a temperature controlled crystal oscillator with a fundamental frequency of 5 Mc/s, controls the accuracy of the instrument. The frequency of the output from this oscillator is reduced by a frequency divider chain and used to operate a " gate " circuit. This circuit is simply an electronic switch which allows signals to pass for an accurately measured period of time. Voltage pulses, each pulse representing one cycle of the unknown frequency, pass through the gate circuit during this period and are counted by the counter circuits and registered as a number on a digital display. Thus if the gate is operated for exactly one second, and the frequency being measured is 825,265 c/s the counter will display these figures.

17. **Heterodyne measurements.** The counting circuits in the system described can handle frequencies up to 20 Mc/s, but its frequency range can be considerably extended by using a heterodyne frequency converter in conjunction with the frequency counter. Fig. 7 is a block diagram of a heterodyne frequency converter. The unknown frequency is found approximately by tuning the r.f. input circuit (the pre-selector) for maximum deflection on the tuning indicator. The unknown frequency is then mixed in the following stage with the output from a locked oscillator. This is a

Fig. 6. BLOCK DIAGRAM OF A BASIC COUNTER FREQUENCY METER.

Fig. 7. BLOCK DIAGRAM OF HETERODYNE FREQUENCY CONVERTER.

conventional oscillator which is locked to a harmonic of the standard frequency. The frequency at which the locked oscillator functions is switch-selected to the nearest harmonic below the frequency indicated by the pre-selector. The difference frequency output from the mixer stage is fed to the counter circuits and is indicated on the display. Then, by adding the displayed frequency to that of the locked oscillator, the unknown frequency is obtained.

For example, if the pre-selector stage indicates an approximate input frequency of 31 Mc/s and the standard frequency crystal is 5 Mc/s, the 6th harmonic (30 Mc/s) would be selected to control the locked oscillator. If the difference frequency displayed is 1·125 mc/s, the unknown frequency would be 1·125 + 30 mc/s = 31·125 mc/s.

Frequency Monitor
18. International regulations lay down certain tolerances outside which the frequency of a transmission must not drift. The heterodyne principle is employed in instruments, called *frequency monitors*, to indicate the amount a transmitter drifts off frequency.

The monitor consists of an independent accurate and stable oscillator, operating on a frequency 1000 c/s above or below that of the transmitter. Thus with the transmitter on frequency a beat of 1000 c/s is produced. This is fed into a direct reading A.F. meter designed to produce mid-scale deflection at 1000 c/s. When the transmitter frequency drifts, the beat frequency will change, either up or down. The change is indicated on the meter and the percentage deviation can be read directly.

Crystal Calibrator
19. This is a high precision crystal controlled oscillator used to provide a spot frequency which may be required as a reference frequency. Usually three internal crystals, with fundamental frequencies of 10 kc/s, 100 kc/s and 1000 kc/s are used, with provision for

plugging in an external crystal when some other fundamental frequency is required. Harmonics generated by these crystals provide an output signal, and in some calibrators a mixer/audio amplifier enables an external signal to be set to a zero beat against a selected crystal, or one of its harmonics.

The crystal calibrator is used for general frequency measurement, or to supplement a receiver or wavemeter without a built-in crystal calibrator. It can also be used as a direct frequency check of any r.f. source. If there is any doubt as to whether the correct harmonic is being employed, it should be used in conjunction with an absorption type wavemeter, when the doubt can be resolved.

Grid-dip Meter
20. This is a simple yet versatile instrument which finds applications in many fields of radio engineering. It consists of a multi-range oscillator with a milliammeter in the grid circuit, and externally mounted plug-in coils. When an external circuit absorbs energy from the grid circuit the drive is reduced and grid current falls, producing a dip in the milliammeter.

A typical circuit shown in Fig. 8, consists of a Colpitts oscillator with a grid current indicating meter, and provision for a plug-in
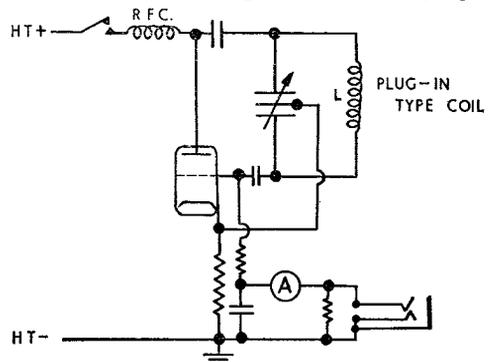
Fig. 8. GRID-DIP METER.

headset. To determine the frequency of an *unenergised* tuned circuit, the coil L is coupled to the circuit under test and the frequency of the meter adjusted until a dip in the milliammeter is obtained. The oscillator tuned circuit is then at the same frequency as the circuit under test and from the dial reading and frequency charts the frequency can be found.

To determine the frequency of an *energised* external r.f. source, the heterodyne method

can be employed. The r.f. source under test is coupled to the oscillator and a headset plugged in. The grid-dip meter circuit is then tuned for zero beat in the headphones, when the external frequency and the grid-dip meter frequency are the same and can be determined from the meter dial reading and frequency charts.

When the h.t. supply to the oscillator valve is switched off, the grid/cathode circuit acts as a diode detector, and the meter can then be used as an absorption type wavemeter. Energy taken from the source is rectified by the grid/cathode of the valve and a reading will result in the milliammeter. At resonance energy absorbed is maximum and therefore the grid-dip meter should be tuned for a *maximum* reading in the milliammeter in this context.

The grid-dip meter is a small portable instrument, and the dial calibration is not accurate. Therefore, for precise frequency measurement the grid-dip meter should be checked against a more accurate frequency standard.

Field Strength Measurements

21. The amplitude of the electric field of a wave at a distance from the transmitter is known as the field strength of the wave at that point. By measuring the field strength around a transmitter aerial information concerning the amount of energy actually being radiated can be obtained.

A simple form of meter for measuring relative field strength is shown in Fig. 9. It consists of an aerial which excites a tuned circuit: the voltage developed across LC
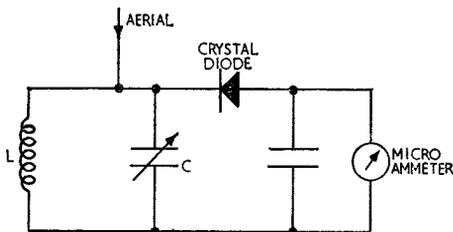
Fig. 9. SIMPLE RELATIVE FIELD STRENGTH METER.

is rectified by a crystal diode and the resultant d.c. voltage used to operate a sensitive microammeter. The frequency range of the meter can be extended by plug-in coils.

More complex field strength meters in the form of a superheterodyne receiver with built-in attenuators and a reference oscillator are available for measuring absolute field strength. When using a field strength meter it must be remembered that reflections from near-by objects or persons can cause erratic readings. If a telescopic aerial is fitted, it should be fully extended when in use.

U.H.F. and S.H.F. Measurements

22. It is not possible to design a conventional type of wavemeter for making frequency measurements in the u.h.f. and s.h.f. bands. This is because of the very small values of capacitance and inductance necessary to obtain resonance at these frequencies. Such components would have to be physically small, and the design of a wavemeter would be further complicated by the unavoidable presence of stray reactances due to connecting leads and valves. Many different instruments for measuring frequencies in these bands do exist, however, and a typical example suitable for use in each band will now be considered.

23. **Lecher bar method of frequency measurement.** A simple and accurate method of u.h.f. measurement makes use of the standing waves set up along a short-circuited section of transmission lines (*see* Section 15 Chapter 4). A typical arrangement is shown in Fig. 10, and consists of a folded bar forming a loop at one end and short-circuited by a
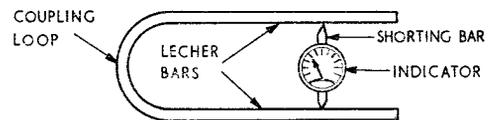
Fig. 10. LECHER BAR FREQUENCY METER.

movable shorting bar at the other. In series with the shorting bar is a sensitive thermo-junction meter.

The loop end is magnetically coupled to the source of frequency to be measured and standing waves are set up along the lecher bars. As the shorting bar is moved, a series of well defined maximum and minimum readings are indicated in the meter. These correspond to antinodes and nodes of the current standing wave and any two consecutive points giving similar readings must therefore be half a wavelength apart for the particular frequency being measured. The bars are calibrated in centimetres, and the

wavelength can be obtained simply by taking the reading between two similar points and multiplying it by two. The frequency can be calculated from the formula.

$$f(Mc/s) = \frac{30,000}{\lambda \text{ (cms)}}$$

The measurements should be taken between two nodes as these are more sharply defined, and give more accurate results, than the antinodes.

24. Resonant cavity wavemeter. At s.h.f. (3,000 Mc/s to 30,000 Mc/s), the tuned circuit takes the form of a hollow box or cavity
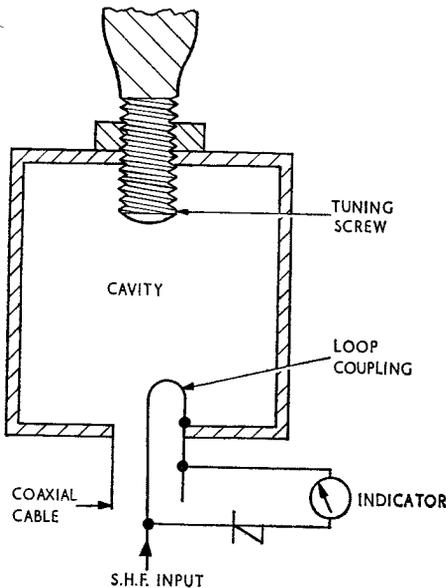
Fig. 11. RESONANT CAVITY WAVEMETER

(Fig. 11). It possesses inductance and capacitance, the values of which are dependent mainly on the internal dimensions of the cavity. Thus there are a number of frequencies at which it will resonate. By inserting a metal screw into the wall of the cavity the inductive component is reduced, and by varying the amount by which the screw projects into the cavity, the resonant frequency of the cavity can be varied.

Energy at the frequency to be measured is fed into the cavity via a coaxial loop. The screw has a calibrated vernier adjustment, and resonance is indicated by a dip in the indicating meter. The frequency is then read on the vernier dial.

Summary

25. The frequency measuring instruments discussed in this chapter form a representative cross section of those which will be encountered in practice.

Of the wavemeters discussed the counter type is the most accurate. It eliminates the interpolation errors which are possible with the heterodyne frequency meter. The accuracy of the absorption type wavemeter is comparatively low and it should only be used where approximate measurements are required. It can be used in conjunction with the high-accuracy crystal calibrator to find the approximate frequency, the exact frequency being found with the calibrator. The grid-dip meter has many uses, but again, in frequency measurement, its accuracy is low. Wavemeters for use in the u.h.f. and s.h.f. bands are normally associated with a particular equipment and often form part of a test set.

SECTION 18

CHAPTER 2

SIGNAL GENERATORS

SIGNAL GENERATORS

Introduction

1. One of the most useful of the wide range of test equipments available to the radio technician is the signal generator. It will be found in various forms in every radio workshop and a knowledge of its function and applications is essential to those responsible for testing and servicing radio equipments.

The signal generator is a source of oscillation whose output frequency and amplitude can be varied. Since the range of frequencies dealt with in radio engineering is so wide, one signal generator cannot be designed to cover the whole range; signal generators can be divided into two categories; those which generate audio frequencies in the band 20 c/s to 20 kc/s and those which produce outputs at radio frequencies, 10 kc/s to 10,000 Mc/s. Further, each of these can be sub-divided according to their output waveform.

A.F. signal generators are designed to produce sine wave outputs or square wave outputs; r.f. signal generators are available with constant amplitude continuous wave output, amplitude-modulated c.w. output, phase-modulated output and frequency-modulated output. In some signal generators several of these functions are combined and are available in one instrument, although generally, phase-modulated and frequency-modulated signal generators are designed for one function only.

2. Some of the more common uses of the r.f. signal generator are alignment of tuned circuits, fault tracing, receiver sensitivity measurements, field intensity measurements and approximate frequency measurements. In this last application it must be stressed that the signal generator is not a standard frequency measuring device and its frequency accuracy must not be relied upon.

An important section of most signal generators is the attenuator; the construction and types of attenuator will now be discussed.

Attenuators

3. The loss of power in an electrical system is known as *attenuation*. It is often required to attenuate currents and voltages at certain stages and attenuators meet this requirement. To avoid distortion, all frequencies of a complex waveform must be attenuated to the same degree and so the attenuation network must be purely resistive. A fixed attenuator is sometimes known as a " pad."

The requirements of an attenuator are as follows (a) it must have the correct input impedance (b) it must have the correct output impedance (c) it must give the required attenuation.

Attenuation is usually measured in decibels but can be measured in nepers (see Book 1, Section 6, Chapter 3).

4. Types of Attenuator.

Symmetrical-T. This is one of the most common types of attenuator pad and consists of a series arm and a central shunt arm (Fig. 1 (a)). The series arm is divided into two equal parts ($R_A = R'_A$) and by suitably choosing the values of R_A and R_B any required attenuation and impedance match can be obtained.
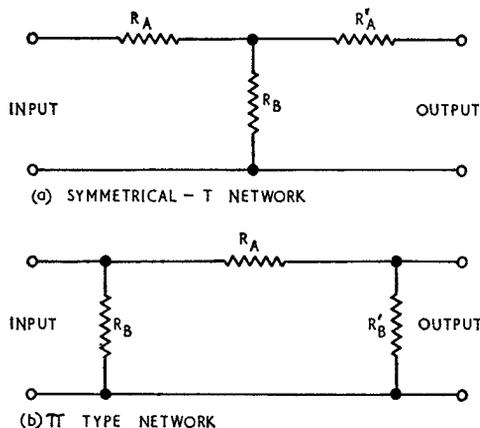
Fig. I. BASIC ATTENUATION NETWORKS.

π *type attenuator.* This is another common type of attenuator and as shown in Fig. 1 (b) consists of one series and two shunt arms. When it is used as an attenuator between two equal impedances the two shunt arms must be equal ($R_B = R'_B$). The only difference

P

between the T and π type attenuator pads is in the value of the resistors of which they are composed.

Variable attenuators. These are designed to have a constant input and output impedance but a variable attenuation. There are various forms, three of which are shown in Fig. 2. The types shown in Fig. 2 (*a*) and (*b*) are simple T and π type pads with all the resistors variable. The variable resistors are all ganged together so that at different positions the input and output impedance is unaltered although the attenuation is varied.

The bridge-T type attenuator shown in Fig. 2 (*c*) has the advantage that only two resistors have to be varied; this simplifies the ganging.

Another form of variable attenuator consists of a number of pads of equal impedance but different attenuation, connected in series. Each pad may be switched in or out to vary the attenuation as required.

Audio Frequency Signal Generator

5. This instrument, often called an audio oscillator, is used for tests and measurements on a.f. amplifiers, a.f. modulators and other a.f. equipments. In order to avoid using heavy and expensive coils and transformers necessary to produce the audio frequencies directly, the beat frequency principle is often employed. Alternatively, resistance capacitance oscillators can be used.

A typical audio frequency signal generator employing the b.f.o. principle is shown in block diagram form in Fig. 3. The output from a fixed frequency oscillator oscillating at a frequency of 185 kc/s, is fed, via a buffer amplifier which prevents frequency pulling at the low difference frequencies, to a mixer stage. Here it beats with the output from an oscillator which can be tuned over the frequency range 155 to 185 kc/s, to produce a

(a) VARIABLE T TYPE

(b) VARIABLE π TYPE

(c) VARIABLE BRIDGED T TYPE
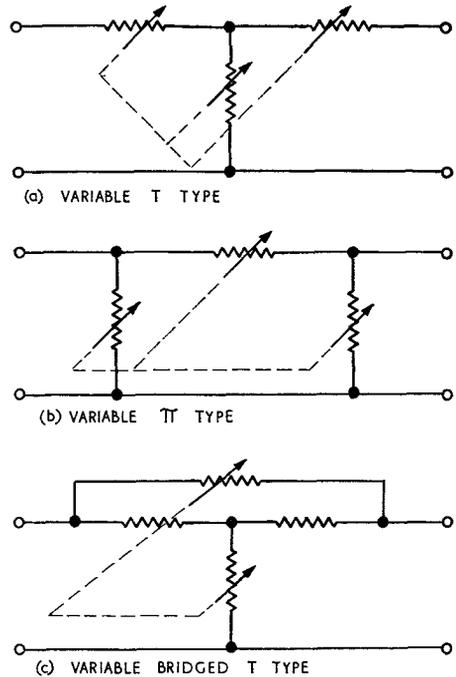
Fig. 2. VARIABLE ATTENUATORS.

beat frequency of between 0 and 30 kc/s. This signal is then fed through a low-pass filter which eliminates any harmonics present, to a wide-band a.f. amplifier. The amplitude of the output from the signal generator can be controlled by a calibrated attenuator.

The frequency calibration of this type of a.f. signal generator is easily checked, since a zero beat should result when the calibrated dial of the variable oscillator is set to zero.

6. The circuit described in para. 5 can be easily adapted to produce a square wave output instead of a sine wave output (Fig. 4).
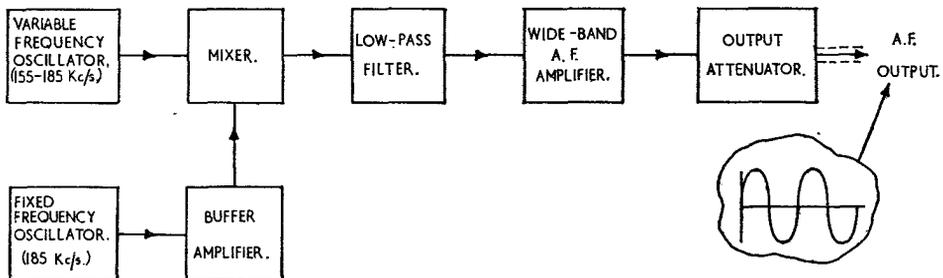
Fig. 3. BLOCK DIAGRAM OF AN A.F. SIGNAL GENERATOR WITH SINE WAVE OUTPUT.
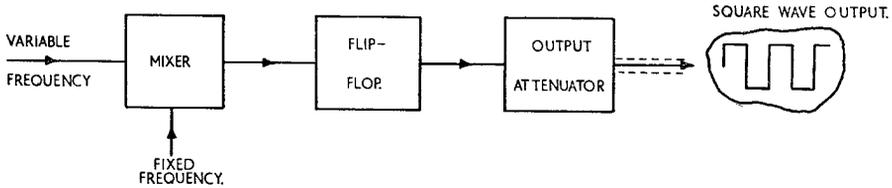
Fig. 4. BLOCK DIAGRAM OF AN A.F. SIGNAL GENERATOR WITH SQUARE WAVE OUTPUT.

In this case the output from the mixer is used to switch on and off (i.e. "trigger") a square wave oscillator similar to the multivibrator discussed in Book 2, Section 12, Chapter 3. The oscillator used is called a *flip-flop* and differs from the multivibrator in that it does not oscillate unless it is triggered by an input signal voltage. The output from the flip-flop will be symmetrical square waves at the frequency of the a.f. input. Again the amplitude of the square waves can be varied by means of the output attenuator.

With suitable switching arrangements the same signal generator can be used to produce either sine wave or square wave output.

7. An a.f. signal generator can be used to measure a number of important characteristics of a.f. equipment. It can be used to measure the gain of an a.f. amplifier, to check its frequency response and to determine any phase distortion present. One of these applications will now be considered.

8. **Use of an audio signal generator to determine the frequency response of an a.f. amplifier.** For this test the necessary equipment and method of connection is shown in Fig. 5. The method employed is that of feeding into the amplifier under test, voltages of fixed amplitude and at spot frequencies

throughout the pass-band of the amplifier (typically between 30 c/s and 1600 c/s). The relative gain of the amplifier at each spot frequency is noted and plotted to obtain the response curve.

In Fig. 5 the output of the a.f. signal generator is connected to the input of the a.f. amplifier under test. An a.c. valve voltmeter monitors the amplitude of signal generator output. An output meter can be switched to either the signal generator output, or to the output from the a.f. amplifier via an attenuator. With the switch as shown in Fig. 5, the signal generator is set to 1000 c/s and the output meter reading is adjusted to a reference point (usually mid scale) by means of the output meter sensitivity control. The reading on the a.c. valve voltmeter is noted. The output meter is now switched to the attenuator output, and the *attenuator* is adjusted to give the original reading in the output meter.

The frequency of the signal generator is now varied in steps over the required frequency range, the amplitude, as indicated on the valve voltmeter, being maintained constant at each change of frequency. The reading on the output meter is noted at each step, and these readings are plotted against frequency to produce the response curve for the amplifier under test.
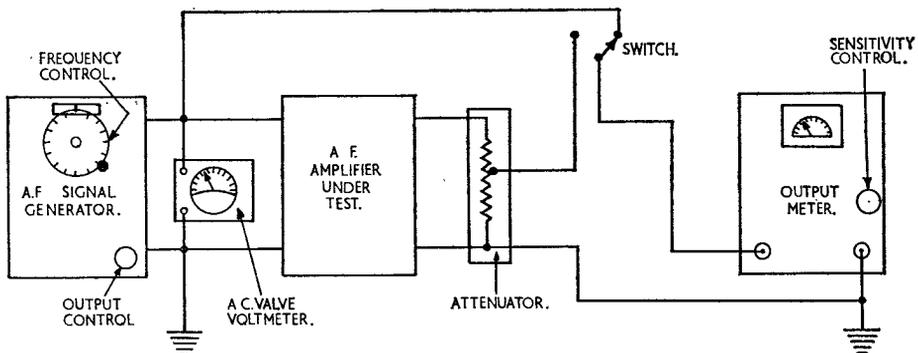
Fig. 5. CIRCUIT TO DETERMINE A.F. AMPLIFIER FREQUENCY RESPONSE.

Radio Frequency Signal Generator

9. R.F. signal generators covering the frequency range 10 kc/s to 10,000 Mc/s are available to the radio technician. They vary in appearance and design according to the frequency range they cover and their applications. The block diagram of Fig. 6 shows the essential circuits of a typical r.f. signal generator. It consists of the r.f. oscillator circuit, the modulator circuit and the output circuit.

10. **R.F. oscillator circuit.** At the lower end of the frequency spectrum this circuit would be a conventional LC oscillator covering the required frequency range, possibly in several bands. For signal generators designed to operate in the v.h.f. band a lecher-bar oscillator could be used, while s.h.f. signal generators would employ a tunable cavity as the resonant element.

To enable the *amplitude* of the oscillator output to be kept constant for different frequencies, an output (or " set R.F.") control is usually included in the oscillator circuit.

11. **Modulator circuit.** The purpose of this circuit is to enable the amplitude of the r.f. oscillations to be varied at an audio frequency, i.e. to provide a modulated r.f. output. Either sine wave or square wave modulation may be required, and some signal generators provide both. The depth of modulation can usually be controlled in conjunction with a meter which indicates the percentage modulation.

For square wave modulated outputs the frequency, the amplitude and the time duration of the square wave, can on some instruments, be made variable. In addition to any build-in modulation circuits, most r.f. signal generators have provision for applying modulation from an external source, so making the instrument more versatile.

12. **Output circuit.** This consists of a calibrated attenuator, and an output level meter which indicates the level of the oscillator output. The attenuator thus selects the required amount of output, and when the oscillator output is adjusted to a calibration mark on the output level meter, a direct reading of the output level in μV or mV is available.

Typical R.F. Signal Generator

13. A typical r.f. signal generator covers the frequency range 10 Mc/s to 300 Mc/s in four bands. The output can be c.w., sine wave modulated, or square wave modulated. Internal or external modulation can be used. Internal modulation frequencies are 400 c/s, 1000 c/s or 5000 c/s. The percentage modulation can be controlled and is indicated on a meter. The level of r.f. output is controlled by varying the oscillator h.t., and for correct frequency indication, must be set to a fixed calibration mark on the output level meter. The output attenuator has coarse and fine controls which control the output in steps of 1 db down to 100 db below 100 mV, i.e. down to 1 μV.

Frequency-Modulated Signal Generator

14. A frequency-modulated signal is one which varies in frequency above and below a centre frequency at a predetermined rate according to the modulating signal. The
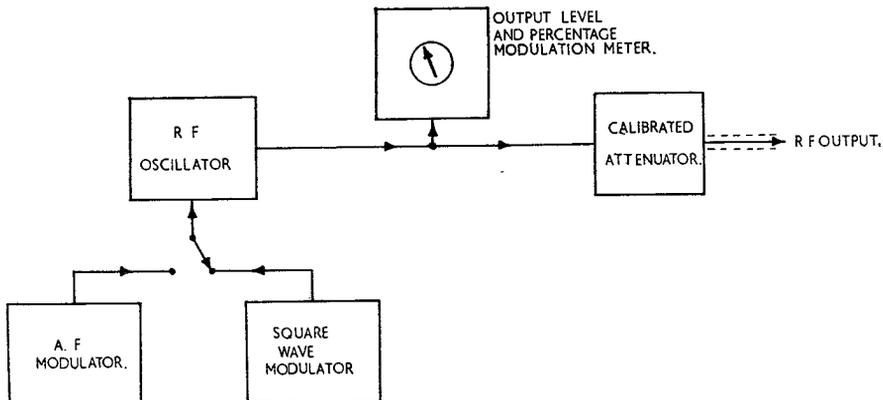
Fig. 6. ESSENTIALS OF AN R.F. SIGNAL GENERATOR.

overall frequency variation is called the *frequency deviation* or *frequency sweep*. An f.m. signal can be produced by continuously varying the inductance or capacitance of an oscillatory circuit, or by varying the reactance of a valve connected to the circuit.

15. **Frequency Sweep Generator, Reactance Valve Type.** This is one form of f.m. signal generator a simple block diagram of which is shown in Fig. 7.
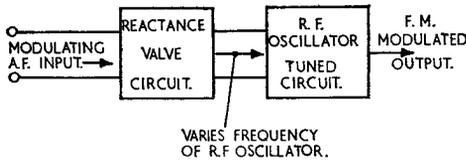
Fig. 7. BLOCK DIAGRAM OF A REACTANCE VALVE F.M. OSCILLATOR.

The valve, which is in parallel with the oscillator tuned circuit, presents either an inductive or a capacitive reactance to the circuit. This reactance will cause the frequency of oscillation to alter. The value of the reactance presented to the circuit depends on the value of the current flowing through the valve; this can be varied by varying the voltage applied to the control grid or to the suppressor grid. Thus by applying a modulating signal to either or both of these grids, the output of the oscillator will be frequency-modulated in sympathy with the modulating signal. The rate of frequency deviation is controlled by the frequency of the modulating signal, and the amount of frequency deviation by controlling the amplitude of the modulating signal.

16. **Ferrite Modulator.** Another method of frequency-modulating the output of an

oscillator is by varying the value of the tuned circuit inductance. The inductor of the tuned circuit is wound with a ferrite core which is placed in the magnetic field of an electro-magnet. The ferrite core is biased by means of a permanent magnet. The permeability of the ferrite varies as the strength of the magnetic flux through it varies. The permeability of the core governs the value of the inductance, which in turn governs the frequency of the tuned circuit. Thus by feeding a modulating current to the electro-magnet, the output of the oscillator will be frequency-modulated in sympathy with the frequency of the current through the electro-magnet. A block diagram of such an arrangement is given in Fig. 8. The stability of a ferrite modulator is better than that of a reactance valve type modulator, and much wider frequency bands can be covered.
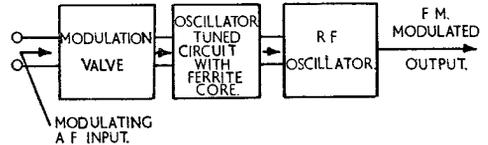
Fig. 8. BLOCK DIAGRAM OF A FERRITE F.M. OSCILLATOR.

Use of a Frequency Sweep Generator to Measure the Frequency Response of an I.F. Amplifier

17. Perhaps the best known application of the frequency sweep generator is in the checking of the tuned circuits of the i.f. amplifier of a superheterodyne receiver, for optimum output over the frequency band.

The block diagram of Fig. 9 shows the necessary equipment and connections for
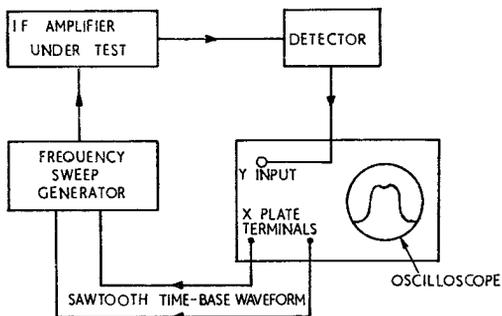
Fig. 9. CIRCUIT TO DETERMINE I.F. AMPLIFIER FREQUENCY RESPONSE.
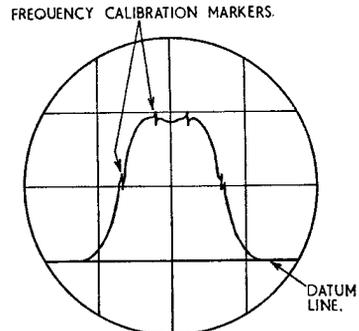
such a test. The modulating signal for the frequency sweep generator is obtained from the time-base circuits of the oscilloscope. Thus the rate of frequency deviation and the oscilloscope time-base are synchronised and points along the horizontal axis of the c.r.t. display will indicate the frequency being swept. An output from the frequency sweep generator is fed to the amplifier under test. The output from the i.f. amplifier is detected and fed as a d.c. voltage to the Y plates of the oscilloscope.

18. The centre frequency of the frequency-sweep generator is set to the mid-band frequency of the r.f. amplifier and the frequency deviation adjusted with the time-base amplitude control, to cover the bandwidth of the amplifier. Thus as the frequency is swept through this band, the output from the i.f. amplifier rises from minimum to maximum and back to minimum again. As this output is fed to the Y plates of the oscilloscope as a d.c. voltage, the display of Fig. 9 will be shown on the c.r.t. This is a graph of frequency (horizontal) against gain (vertical), and is therefore the frequency response of the amplifier.

19. To increase the accuracy of measurements made in the test just described the x and y axes of the graph produced can be calibrated. Frequency calibration markers produced in a marker oscillator are mixed with the output from the frequency sweep generator and displayed on the c.r.t. as shown in the inset of Fig. 9.

The y axis can be calibrated by producing a datum line corresponding to zero amplifier gain. This is done by pulse modulating the frequency sweep generator so that it is switched off during each alternative oscilloscope scan. Thus the response curve will be displayed on one scan and the datum line on the next. From this datum line the oscilloscope graticule can be calibrated in decibels.

ALIGNMENT OF SUPER-HETERODYNE RECEIVERS

Introduction

20. Because the values of fixed components in a superheterodyne receiver can vary with age, small trimming components are included in the tuned circuits to enable minor adjustments to be made. Main tuning capacitors have small pre-set variable capacitors in series or parallel: inductors have brass slugs or iron-dust cores which can be screwed into or out of the inductors so varying their values by a small amount.

Circuit alignment is the process of adjusting the tuned circuits by means of these trimmers so that the circuits resonate at the required frequency.

The Superheterodyne Receiver

21. The basic circuit of a superheterodyne receiver is given in Fig. 44 of Section 14 Chapter 2 (Book 2) and is reproduced in Fig. 10. For alignment, the circuit can be divided into two parts, the r.f. tuner and the i.f. amplifier. The r.f. tuner in Fig. 10 consists of V_1 and V_2 and their associated components. The main tuning capacitors and their trimmers and padders are listed under the diagram.

22. When aligning the receiver, the shunt trimmers are always adjusted at the high frequency end of the tuning range. Thus stray capacitances which have most effect at high frequencies can be compensated for in the most critical region of the tuning range. Series padders are adjusted at the low end of the frequency range.

23. The i.f. amplifier consists of V_3 and V_4 and their associated circuits. Low frequency i.f. amplifiers are coupled by double-tuned transformers, as in Fig. 10 and if the pass-band is not too wide these can be adjusted for maximum gain at the centre frequency of the band. Broad pass-band i.f. amplifiers employ staggered tuning and greater-than-critical coupling (see Book 1, Section 7, Chapter 1): the instructions in the appropriate equipment A.P. must be followed when aligning. Slugcore adjustment of the inductors is often used for aligning i.f. amplifiers.

Circuit Alignment—General

24. Adjustments should start with the tuned circuit nearest the detector stage and proceed towards the aerial, the aerial circuit being adjusted last. Before starting the alignment the a.g.c. voltage should be switched off or disconnected, since it would mask any change in amplifier gain. The b.f.o. should also be switched off as this produces a bias voltage in the final detector which would also mask changes in the output voltage.

25. If the r.f. and i.f. stages have a manual gain control, this should be set to the normal operating position, since this setting will affect

Fig. 10. BASIC CIRCUIT OF SUPERHETERODYNE RECEIVER.

the input impedance of the valves controlled and may detune them.

Before making any adjustment to the receiver the signal generator to be used and the receiver itself should be switched on and allowed to warm up to their normal working temperature.

I.F. Amplifier Alignment

26. The i.f. amplifier is aligned first, and so the signal generator is set to the intermediate frequency of the receiver and connected to the grid of the mixer valve V_2 (Fig. 10). The signal level at the final detector can be used to measure circuit alignment by connecting a high impedance valve voltmeter across the diode load. The stage before the signal injection (the r.f. amplifier) should have the valve removed to prevent unwanted signals and noise interfering with the alignment signal.

27. Starting with the i.f. transformer before the final detector (i.f.t.$_3$) the trimming capacitors are adjusted for maximum reading in the valve voltmeter. If the i.f. amplifier has a wide band-width, " peaking " with the trimmers may be necessary at one or more frequencies, off the centre band frequency, as specified in the equipment A.P.

The signal generator input level is then reduced and i.f.t.$_2$ trimmers are adjusted for maximum in the valve voltmeter. The trimmers of i.f.t.$_1$ are then adjusted. This procedure should be repeated, starting with i.f.t.$_3$ again, and it may be advisable to re-peak for a third time.

R.F. Tuner Alignment

28. The aerial is disconnected, V_1 replaced, and a dummy aerial or matching pad is used to provide a correct match to the receiver input. The signal generator should be accurately set to the upper alignment frequency for the receiver, and connected via the matching pad to the receiver input. The receiver main tuning dial is then set carefully to the signal generator frequency. This is indicated by a maximum reading in the valve voltmeter. If the receiver dial reading differs from the signal generator reading, the receiver dial should be set to the correct reading and the shunt trimmer of the local oscillator tuned circuit (C_{17}) adjusted for maximum reading in the valve voltmeter. Care must be taken not to adjust the oscillator trimmer such that the local oscillator frequency is on the wrong side of the signal frequency at this high end of the band.

The inter stage trimmer (C_9) and the aerial trimmer (C_3) are next adjusted for maximum output, the signal generator level being reduced to prevent saturation.

29. Alignment at the lower end of the band is usually done by adjusting the inductance of the r.f. coils. The signal generator is set to the lower alignment frequency and L_3/L_4 and L_1/L_2 adjusted for maximum output in the valve voltmeter. The oscillator series padder (C_{15}) is then adjusted for maximum output reading while the main receiver tuning dial is rocked back and forth through the region of best response.

After this the signal generator is retuned to the high frequency calibration point and the shunt trimmers adjusted for maximum output and correct dial calibration. Finally the signal generator is retuned to the low frequency calibration point and the local oscillator series padder is re-adjusted for maximum output and correct receiver dial calibration.

Conclusion

30. This chapter has dealt briefly with some types of signal generator in common use, and with some of their applications. Detailed information concerning the operation and functions of particular signal generators can be found in the appropriate publication which the technician should consult before using the equipment.

SECTION 18

CHAPTER 3

WAVEFORM MEASUREMENTS

WAVEFORM MEASUREMENTS

Introduction

1. Previous Sections of these notes have dealt with ammeters and voltmeters which give certain limited information about alternating currents and voltages. Such meters record either r.m.s. or peak values but give no indication of the frequency or phase relationship between various quantities recorded. The cathode-ray oscilloscope (c.r.o.) overcomes these disadvantages and thus greatly increases the amount of information available to the radio technician about the behaviour of a circuit.

The Cathode-ray Oscilloscope

2. The cathode-ray oscilloscope provides a complete graphical representation of an alternating quantity. The heart of a c.r.o. is the cathode-ray tube discussed in Book 2, Section 8, Chapter 5. Basically the c.r.t. provides a means of controlling a beam of electrons which produces a visible trace on a fluorescent screen. An oscilloscope incorporating the basic circuits necessary for its operation is shown in block form in Fig. 1. The circuits involved will now be considered.

Basic C.R.O. Circuits

3. (a) **Time-base Generator.** This produces a special type of voltage waveform that is applied to the X deflection plates of the c.r.t. through an amplifier (the horizontal amplifier). This time-base voltage produces the horizontal trace on the c.r.t. screen.

(b) **Vertical amplifier.** The voltages applied to the Y deflection plates of the c.r.t. must be fairly large in order to produce significant vertical deflection. They are therefore amplified in the vertical amplifier, which is sometimes called the signal amplifier.

(c) **Synchronisation.** The waveform displayed would drift along the time-base if it
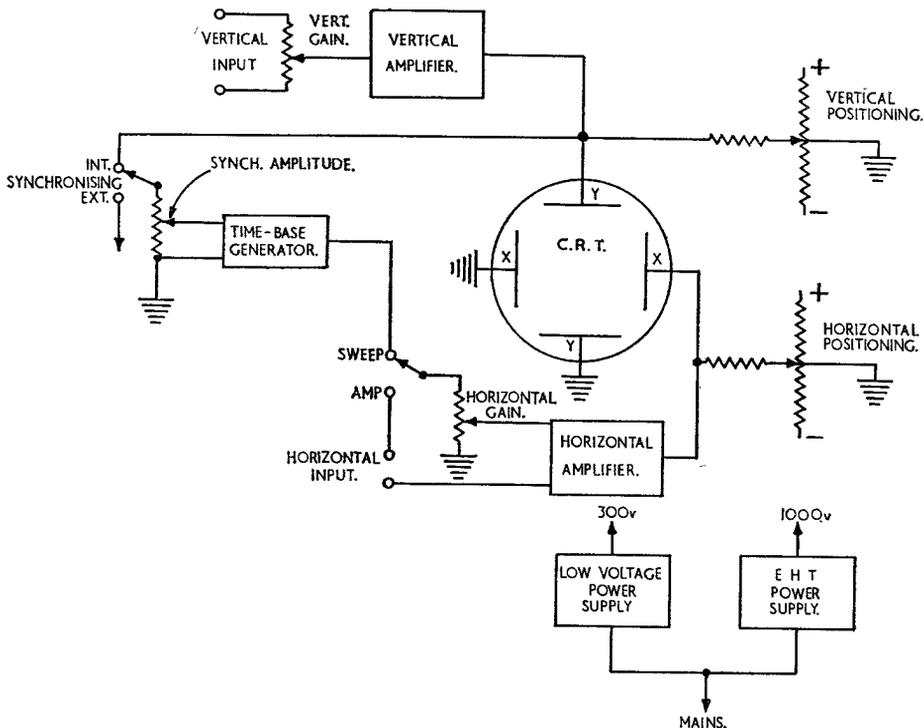
Fig. I. BASIC CATHODE-RAY OSCILLOSCOPE.

were not synchronised with the time-base frequency. The synchronising control provides a mean of obtaining a stationary display.

(d) **Positioning (or shift) controls.** These enable the display to be positioned on the c.r.t. screen as required.

(e) **Power supplies.** These provide the e.h.t., h.t. and l.t. voltages necessary for the function of the c.r.o.

The Time-base

4. To study an unknown voltage waveform using a c.r.o., a voltage which is changing uniformly with time is applied to the horizontal deflection system, and the unknown voltage is applied to the vertical deflection system of the c.r.t. The voltage waveform necessary to produce uniform movement of the electron spot on the c.r.t. screen is shown in Fig. 2 and is known as a *sawtooth wave-form*. For accurate measurement, the rise
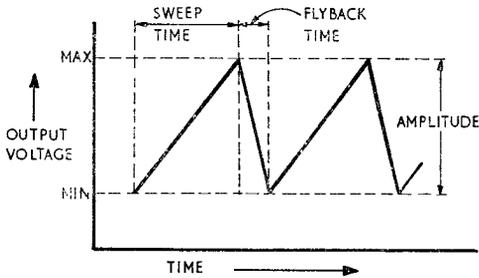
Fig. 2. IDEAL TIME-BASE VOLTAGE WAVEFORM.

in voltage must be linear. To obtain the maximum number of sweeps per second for a given sweep time, the time taken for the voltage to fall from maximum to minimum (the *flyback time*) must be as short as possible. The *amplitude* of the waveform is the difference between minimum and maximum potential and governs the length of trace on the c.r.t. screen. The velocity at which the spot moves across the screen depends on the *slope* of the voltage rise.

5. The principle of a circuit which produces a sawtooth voltage is illustrated in Fig. 3. The circuit consists of a capacitor C connected in series with a resistor R_1 across a d.c. supply. In parallel with the capacitor is a switch S.

If C is initially uncharged, and the switch S open, then when the d.c. supply is connected, current will flow through R_1 into C; as C charges, the voltage across its plates will rise exponentially (A to B in Fig. 3 (b)).

The time taken for the potential across C to reach the applied voltage V will depend upon the value of C and R_1 (refer to Book 1, Section 4, Chapter 3). If the value of either of these components is *reduced* the time taken for the capacitor voltage to reach the applied voltage will be *less*, i.e. the velocity of the voltage rise will be *increased*. Conversely, if the value of C or R_1 is increased the velocity will decrease.

(a)   (b)
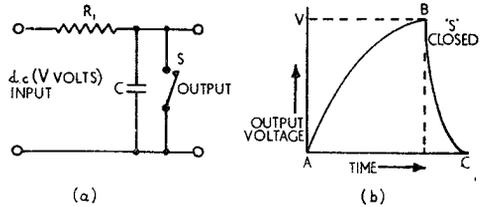
Fig. 3. PRINCIPLE OF THE SAWTOOTH VOLTAGE GENERATOR.

When C is fully charged switch S is closed: C will rapidly discharge through the low resistance path of the switch S and the voltage across C falls to zero (B to C). If S is now opened C will start to charge again. The output voltage taken across C will be as shown in Fig. 3 (b).

6. The circuit of Fig. 3(a) can be made fully automatic by replacing S with a cold cathode gas-filled diode (see Book 2, Section 8, Chapter 4, para. 16). The circuit then becomes that of Fig. 4(a) and when the voltage across C reaches the striking potential

(a)

(b)
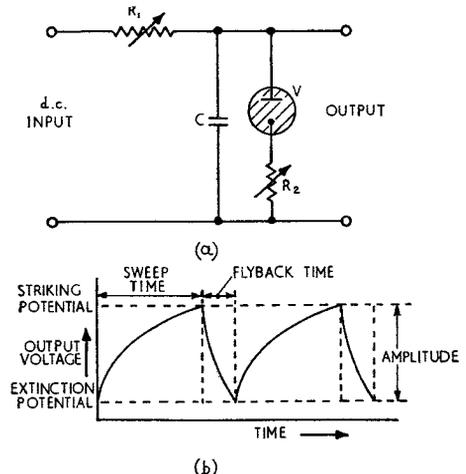
Fig. 4. BASIC CIRCUIT OF A SAWTOOTH VOLTAGE GENERATOR.

of the soft diode, the gas will ionise and C will discharge rapidly through V and R_2. The resistor R_2 is included to limit the current flowing through V. When the potential across C reaches the extinction voltage of the diode, V will de-ionise and C will commence to charge again. The circuit will thus produce a succession of sawtooth waveforms of the shape shown in Fig. 4(b).

By making R_1 variable the velocity of the output waveform can be varied; the value of R_2 will control the flyback time.

Synchronisation of Time-bases
7. In order to obtain a stationary trace on the c.r.t. screen the frequency of the time-base and the frequency of the waveform being examined must be locked or synchronised in some way. The method of synchronisation will vary with different types of time-base generator, but in general a sample of the waveform under examination is fed to the time-base generator such as to cause the scan time to start at the beginning of one cycle of the waveform being examined.

A typical time-base circuit suitable for use in a c.r.o. will now be discussed.

Thyratron Time-base Circuit with Pentode Charger
8. The simple time-base circuit of Fig. 4(a) has a serious disadvantage. The voltage across the capacitor during the sweep time does not rise *uniformly* with time. To produce an output waveform approaching that of the ideal waveform of Fig. 2, the circuit must be arranged such that the current flowing into the capacitor remains constant. This is the main object in the design of any time-base circuit.

The circuit shown in Fig. 5 uses a pentode valve to replace the resistor R_1 of Fig. 4,

and since the anode current of a pentode valve is independent of anode voltage provided the voltage is not too low, the current flowing into the capacitor C_1 will remain constant. The thyratron V_2 provides the discharge path for C_1.

When the h.t. is switched on C_1 is uncharged, and cannot charge instantaneously; the anode potential of V_1 and the cathode potential of V_2 are therefore initially at h.t. positive. The grid of V_2 is at a lower potential than h.t. positive and therefore V_2 is cut off with its grid negative with respect to its cathode. C_1 commences to charge through V_1. Since this charging current is constant, the voltage across C_1 will rise linearly. As this voltage rises, the potential at V_1 anode and V_2 cathode falls and the p.d. between grid and cathode of V_2 will eventually come within V_2 cut on. When this happens V_2 will ionise and provide a discharge path for C_1. As C_1 discharges V_1 anode will rise taking with it V_2 cathode. When the p.d. between V_2 grid and cathode is insufficient to maintain ionisation in V_2, V_2 will cut off, allowing C_1 to start its charge again. R_1, R_2 and R_3 are control resistors while R_4 and R_5 are safety resistors.

9. **Time-base controls.** An oscilloscope is required to display waveforms of widely differing frequencies and some control of the time-base functions must be included in the time-base circuit. The following controls are incorporated in the circuit of Fig. 5 and their function is common to most c.r.o. timebases.

Coarse velocity control. The switch S_1 selects capacitors of different values (C_2 or C_3) so enabling a range of velocities to be chosen.
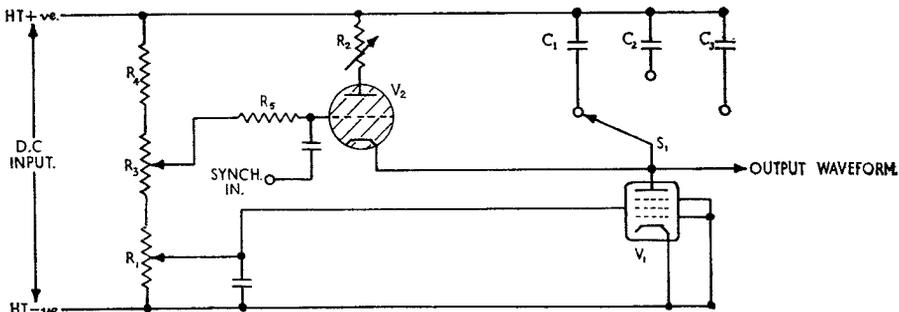
Fig. 5. THYRATRON WITH PENTODE CHARGER.

Fine Velocity Control. By varying the potential on the screen grid of V_1 with R_1 the value of V_1 anode current and hence the rate of rise of capacitor voltage can be varied. This control has the same effect as the variation of R_1 in Fig. 4(a).

Flyback or fine frequency control. R_2 varies the *discharge* time of the capacitor (*i.e.* the flyback time) thus enabling a small variation in the frequency of the output waveform to be made without altering the velocity.

Amplitude Control. Since R_3 enables the grid potential of the thyratron to be varied, it varies the striking potential and so controls the maximum potential of the output voltage.

Synchronising terminal. By feeding a positive going pulse of slightly higher frequency than the free-running frequency of the time-base to the grid of V_2, the valve will ionise slightly before it would normally. Thus the time-base frequency will be synchronised with the frequency of the input pulses.

10. The time-base circuit described in para. 8 is one of many types of circuit designed to produce a linear time-base waveform. In general there are two main groups of time-base circuits; soft valve circuits such as the one described provide a time-base waveform with a maximum frequency of about 2500 c/s; hard valve time-base circuits can operate at frequencies up to 250,000 c/s.

The Signal Amplifier
11. As the deflection system of a c.r.t. is relatively insensitive, voltages of the order of several hundred volts are required to obtain full scale deflection. Thus the input signals must be fed through an amplifier before being applied to the Y plates. The amplifier should have a wide frequency response in order that the signals are not distorted, and often has a step gain control.

Power Supplies
12. Both low voltage and e.h.t. voltage are required. The low voltage (about 300V) is necessary for the time-base and amplifying circuits and a conventional full-wave rectifier with associated smoothing circuits is normally employed. The e.h.t. is applied so that the cathode is 1000 volts negative and the final anode at or near the earth potential.

Oscilloscope Calibration Circuits
13. For accurate measurement of the waveform under examination the X and Y axes

of the c.r.t. should be calibrated. When used for measuring time, the horizontal axis can have calibration markers superimposed upon it. The markers can be obtained from a crystal controlled oscillator, the output of which, after being passed through suitable shaping circuits, can be applied to the Y plates to produce accurately spaced calibration markers on the time-base.

14. The vertical axis can be calibrated in terms of voltage either by comparing the input with that of a sine wave of known amplitude or by using a built-in centre zero reading voltmeter. With the latter device the peak to peak voltage of the input is measured by moving the waveform up and down with the Y shift control so that the positive and negative peaks sit on a graticule reference line. The voltmeter indicates the value of the shift potential and so the difference between the voltmeter readings, when calibrated in conjunction with the amplifier, gives the peak to peak value of the input waveform.

Oscilloscope Operating Controls and Terminals
15. The basic controls and terminals of an oscilloscope are shown in Fig. 6 and summarised as follows.
Brilliance. A variable resistor changes the voltage between the grid and cathode of the c.r.t. and so controls the intensity of the trace.
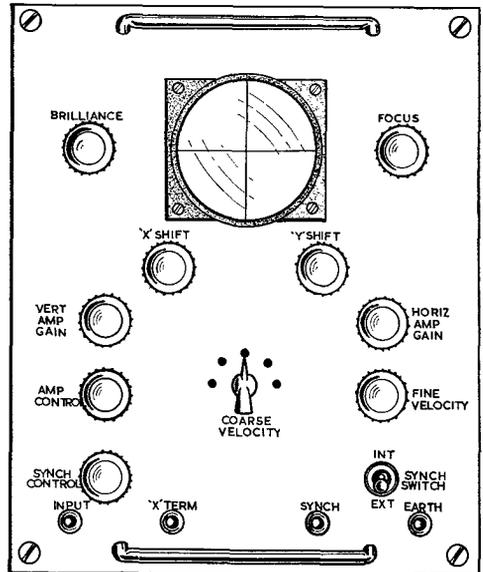
Fig. 6. OSCILLOSCOPE CONTROLS.

Focus. This control changes the voltage on the second anode of the c.r.t. and so adjusts the focal point of the electron beam.

X Shift. A variable resistor changes the d.c. potential on one horizontal deflection plate and so positions the trace horizontally.

Y Shift. This changes the d.c. potential on one vertical deflection plate and so positions the trace vertically.

Amplifier gain. This is usually a step control which can vary the bias on the amplifier valves, thus varying the gain of the amplifier.

Amplitude control. This varies the amplitude of the time-base waveform and thus the length of trace on the c.r.t.

Coarse (or Step) Velocity Control. This switches different value capacitors into the time-base circuit so varying its frequency in steps.

Fine Velocity Control. This varies the resistance of the time-base charging circuit so giving fine velocity variation.

Synchronising Control. This varies the amplitude of the applied synchronising signal so enabling a stationary trace to be obtained.

INT/EXT. Synchronising Switch. This allows a synchronising signal to be switched from either an internal or external source.

Input terminal. Connects the input signal to the vertical amplifier.

X Plate Terminal. Provides a means of feeding an external time-base voltage to the X plates in place of the internal time-base. Alternatively the internal time-base can be used with other associated equipments.

Synchronising terminal. This is used with the synchronising switch in the external position and provides a means of connecting an external synchronising signal to the oscilloscope.

Earth terminal. This gives a means of earthing the oscilloscope.

Operating Precautions

16. The following precautions should be observed when using or servicing an oscilloscope.

(*a*) Never use an oscilloscope with the case removed as high voltages are exposed which could cause fatal shock.

(*b*) Ensure that the display is not affected by stray magnetic fields.

(*c*) Do not allow a bright spot to remain at one point on the c.r.t. screen. This could cause the screen to be burnt at this point.

(*d*) If voltages in excess of those specified are applied to the c.r.o. they will damage the circuit components.
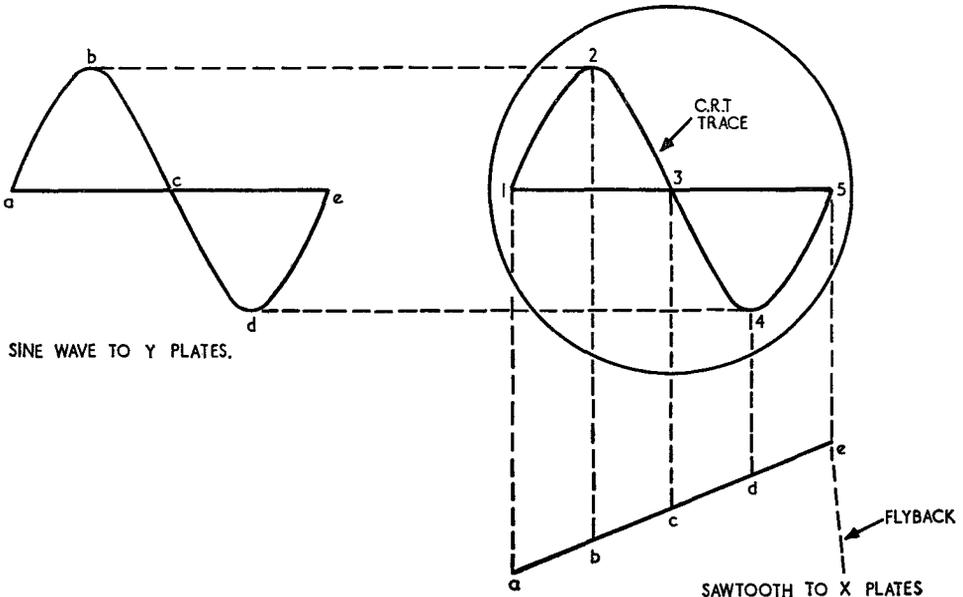
Fig. 7. FORMATION OF A SINE WAVE TRACE.

Formation of an Oscilloscope Pattern

17. If a d.c. potential is applied to the X plates of the c.r.t. and another d.c. potential applied to the Y plates, the resultant position of the electron spot on the c.r.t. screen will depend upon the relative amplitudes and polarities of the two voltages.

If a sawtooth waveform is applied to the X plates and a sine wave waveform to the Y plates the resultant waveform traced on the c.r.t. screen will be as shown in Fig. 7. The position of the spot at any instant can be deduced from the X and Y plate potentials at that instant. Thus at instant b in Fig. 7, the sawtooth voltage will have deflected the beam to the right and the sine wave voltage will have deflected it upwards, the resultant position of the spot being at 2 on the c.r.t. screen.

Lissajous Figures

18. **Frequency measurement.** The c.r.o. can be used to measure the frequency of a signal by comparing it with a reference signal of known frequency. The accuracy of such a measurement when correctly made, is limited only by the accuracy of the reference frequency.

If a sine wave voltage (instead of the usual sawtooth waveform) is fed to the horizontal deflection plates, and another sine wave is applied to the vertical deflection system, the resultant pattern traced on the c.r.t. screen is called a *Lissajous figure.*

19. The Lissajous figure formed by two sine waves of different frequencies (2:1 ratio) is shown in Fig. 8. This pattern can be deduced by projecting the instantaneous voltage of waveform X on to the horizontal axis of the c.r.t. display and projecting the corresponding reference voltage on to the vertical axis of the display as illustrated in Fig. 8.

The ratio of the frequencies of the two sine waves can be found by counting the number of loops along the top or bottom edge of the pattern, and the number of loops down the side of the pattern. Then

$$\frac{\text{Frequency on horizontal axis}}{\text{Frequency on vertical axis}} = \frac{\text{Number of loops on side of pattern}}{\text{Number of loops on top of pattern}}$$

20. Thus if a variable calibrated reference frequency is applied to one deflection system and the unknown frequency signal to the other, a rapid and accurate method of frequency measurement is available. Ratios of up to 10:1 can be measured in this way but
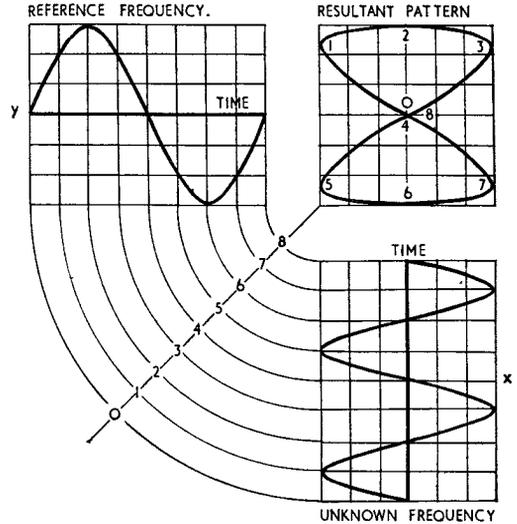
Fig. 8. FORMATION OF A LISSAJOUS PATTERN FROM TWO FREQUENCIES OF 2:1 RATIO.

due to the difficulty in counting the loops, the accuracy of ratios greater than this decreases. Lissajous patterns can only be used for frequency measurement if one frequency is an *exact* simple ratio of the other.

Lissajous patterns for frequency ratios commonly encountered in frequency measurement are shown in Fig. 9.
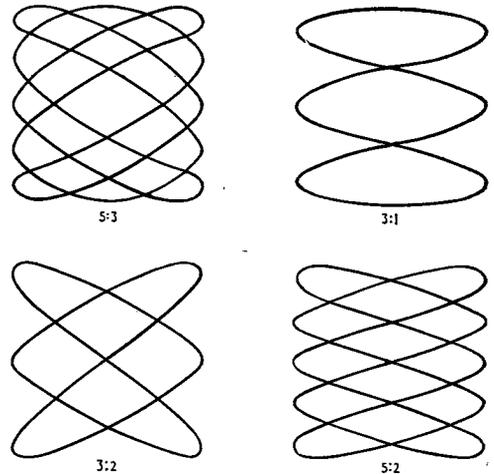
Fig. 9. COMMON LISSAJOUS FIGURES.

21. **Phase measurements.** Lissajous figures can also be used to measure the phase difference between two sine wave voltages of the same frequency. Fig. 10 shows two voltages of equal amplitude and frequency but differing in phase by 90°, applied to the deflection
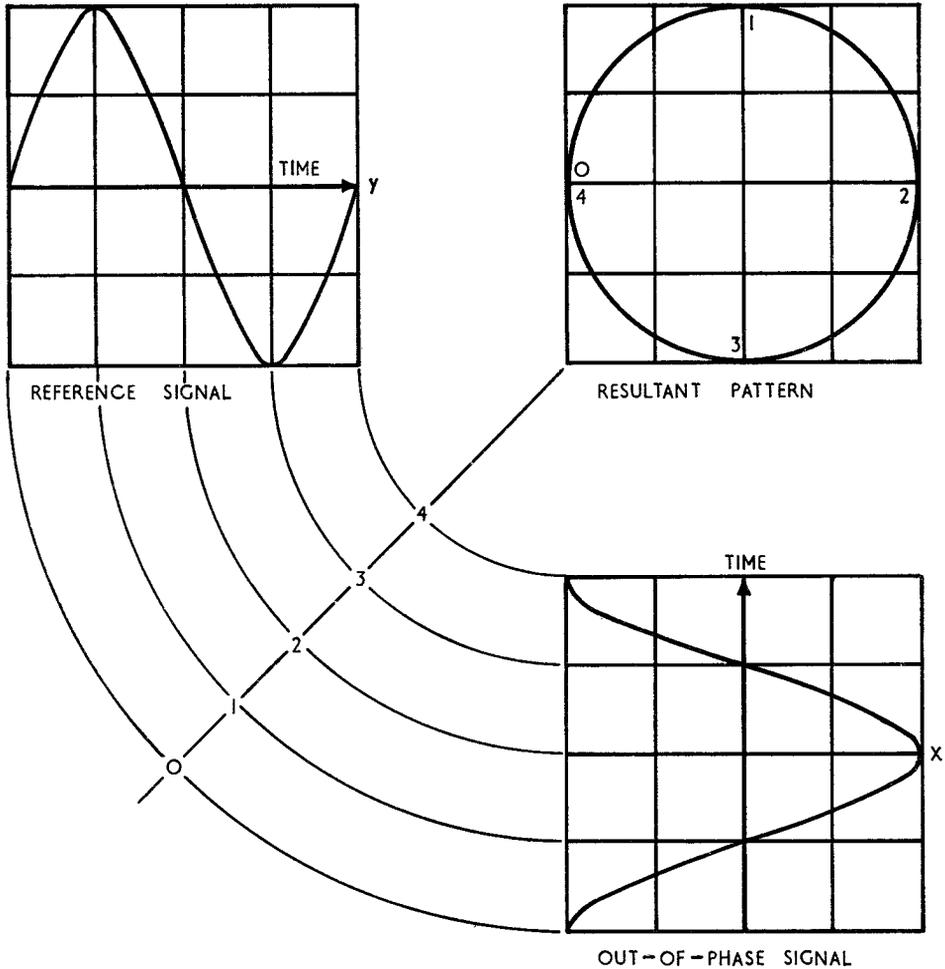
Fig. 10.   FORMATION OF LISSAJOUS FIGURE FROM TWO SIGNALS
WITH 90° PHASE DIFFERENCE.

system of a c.r.o.   By making plots of the
instantaneous voltages from waveforms X and
Y on the corresponding X and Y axes of the
c.r.t. display, a circular pattern, as shown in
Fig. 10 will be displayed.

22.   If the phase difference between the two
signals varies, the pattern formed will change.
When the two signals are in phase the pattern
is a straight line as shown in Fig. 11(a).   This
broadens to an ellipse at 45° (Fig. 11(b)),
then to a circle at 90° and so on as shown.

23.   To measure the angle of phase dis-
placement the c.r.t. graticule must be used
to provide the X and Y ordinates.   The
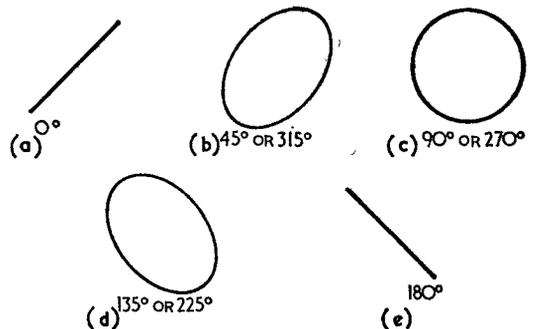
Fig. 11.   PATTERNS ILLUSTRATING PHASE
DIFFERENCE.

vertical and horizontal gain controls are set to zero and the spot is centered at the intersection of the X and Y axes by means of the shift controls. The test signals are then applied to the vertical and horizontal inputs and the gain controls adjusted until the pattern extends equally in both the X and Y directions as shown in Fig. 12.
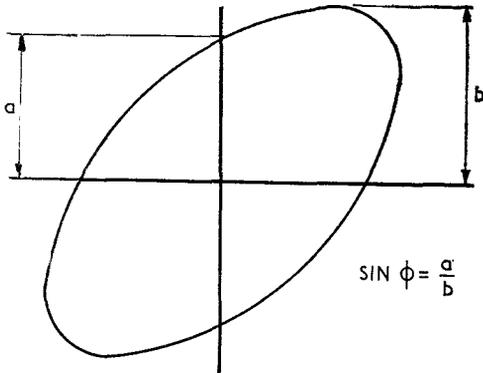
Fig. 12. PHASE ANGLE MEASUREMENT.

The point where the pattern cuts the Y axis (the Y axis intercept) is measured ('a' in Fig. 12) and the maximum vertical amplitude of the pattern is also measured ('b' in Fig. 12). The angle of phase difference ϕ, is then found from the following formula.

$$\text{Sin } \phi = \frac{\text{Y axis intercept}}{\text{Y axis maximum}} = \frac{a}{b}$$

Interpretation of Patterns

24. To gain any information about an unknown signal using an oscilloscope, the unknown signal must be plotted against a signal of known characteristics. If a sawtooth voltage waveform is fed to the horizontal deflection system and the unknown signal to the vertical deflection system, then the resultant pattern is a plot of the voltage of the unknown signal against time.

An important factor to be remembered when interpreting a waveform is the type and amount of distortion introduced by the oscilloscope circuits. With complex waveforms, e.g. square waves or triangular waves, the frequency response of the c.r.o. amplifiers determines the amount and type of distortion introduced into the waveform by the c.r.o. amplifier.

25. When RC coupling into the c.r.o. is employed, phase distortion of a complex waveform (see Book 2, Section 10, Chapter 1) may be introduced. A further cause of distortion of an unknown signal is stray pick-up. This can be reduced by using short coaxial input leads and by connecting the circuit under test and the c.r.o. to a good physical earth.

Procedure for Waveform Observation

26. The oscilloscope is especially useful when tracing a signal through an equipment. The technician should have a good knowledge of the equipment being tested and should know the approximate waveforms to expect at various points in the circuit. The following procedure should be adopted for viewing waveforms.

(a) Connect the Y amplifier terminal of the oscilloscope to the circuit under test and connect the oscilloscope earth terminal to the earth side of the circuit. Ensure a good physical earth to both oscilloscope and equipment.

(b) To observe one complete cycle of waveform the time-base velocity control should be set to the same frequency as the waveform. If more than one cycle is to be viewed, the oscilloscope time-base frequency should be decreased. The time-base fine frequency control and synchronising amplitude control should then be adjusted until the waveform is stationary.

Conclusion

27. The foregoing paragraphs have indicated the circuits, controls and a few of the uses of a cathode-ray oscilloscope. There are many different types of c.r.o. in Service use all of which are based on the principles discussed in this Chapter.

SECTION 18

CHAPTER 4

MODULATION MEASUREMENTS

MODULATION MEASUREMENTS

Introduction

1. Modulation is a process whereby intelligence is impressed upon an r.f. carrier. During the process of modulation the characteristics of the carrier wave are altered in sympathy with the intelligence being conveyed. The characteristics which can be varied are the phase, frequency and amplitude of the carrier; these types of modulation are the types most commonly used. To reduce distortion of the intelligence and to avoid interference with stations on nearby frequencies, measurements of the modulated wave must be taken when testing or servicing transmitters.

Amplitude Modulation Measurements

2. The average depth of modulation of an amplitude modulated wave is often expressed as a percentage (Section 13, Chapter 1), and with reference to Fig. 1, is given by the formula

$$\text{Average percentage modulation} = \frac{E_{max}-E_{min}}{2E_c} \times 100\%$$

This formula gives the average percentage modulation of a carrier modulated by a pure sine wave of constant amplitude. The mean carrier amplitude is constant over a whole number of modulation cycles and the depth of modulation also remains constant.

(a) AMPLITUDE MODULATED WAVE

(b) RECTIFIED AMPLITUDE MODULATED WAVE
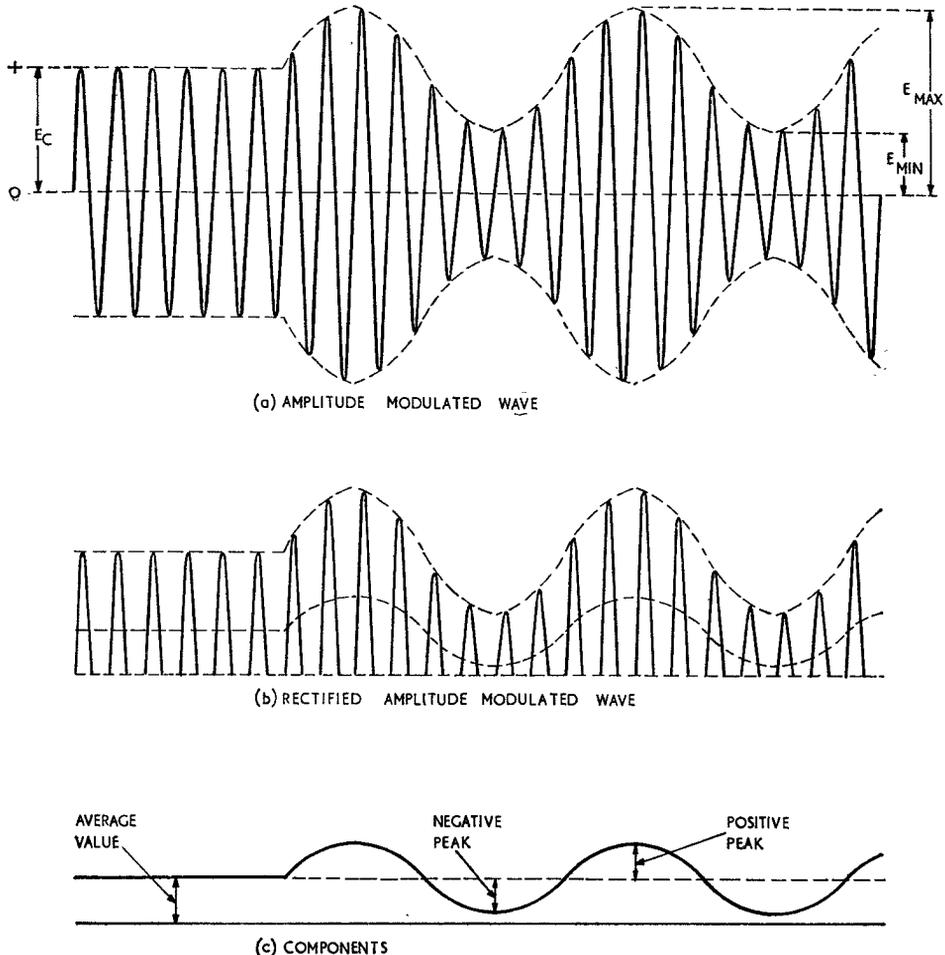
(c) COMPONENTS

Fig. I. COMPONENTS OF AN AMPLITUDE MODULATED WAVE.

With normal speech modulation however, the maximum carrier amplitudes are not the same value and the percentage modulation varies considerably. Thus a method of reading the *average* percentage modulation would not disclose momentary over-modulation. In this case a meter which responds to the modulation peaks and troughs is required.

Measurement of Average Percentage Modulation

3. When a carrier is modulated by a pure sine wave of constant amplitude (Fig. 1(*a*)) the average amplitude of the carrier is the same as that of the unmodulated carrier. Furthermore, if the modulated carrier of (*a*) is rectified and the output passed through a r.f. filter, the a.c. component produced will be proportional to the amplitude of the modulating signal; the d.c. component produced will be proportional to the carrier amplitude and is independent of the percentage modulation. If these two amplitudes are measured, the percentage modulation of the original carrier can be calculated.

4. Fig. 2 shows a circuit suitable for providing these measurements. The modulated carrier is fed into the tuned circuit LC, which is tuned to the carrier frequency. During the negative half cycles C_1 charges through the diode with the polarity shown, and only the positive half cycles are applied across the remainder of the circuit. The waveform of the voltage across AB is that of Fig. 1 (*b*). The r.f. filter smooths out the carrier frequency and a d.c. voltage varying at the modulation frequency, as in Fig. 1(*c*), appears across DE. The d.c. component is fed through the d.c. voltmeter M_1, and so this meter will read a voltage proportional to the *mean* carrier amplitude. The a.c. component passes through M_2, a high resistance a.c. voltmeter, and so M_2 will read a voltage proportional to the r.m.s. value of the original modulating signal. By multiplying the M_2 reading by $1 \cdot 414$ to give the peak value, the percentage modulation can be calculated.

The Double Rectifier Method

5. With normal speech modulation the depth of modulation varies. To obtain the percentage modulation of such a signal the meter must be able to read either the positive or negative peak voltage of the carrier. In the circuit of Fig. 3, V_1 is a negative detector and so across the high resistance voltmeter, M_1, a varying d.c. voltage as shown in Fig. 1(*c*) will be developed. The polarity of this voltage will be as shown in Fig. 3. In series with this voltage between the points A and C is a d.c. voltage supplied from the battery. The value of this voltage is deter-

Fig. 2. SIMPLE MODULATION METER CIRCUIT.
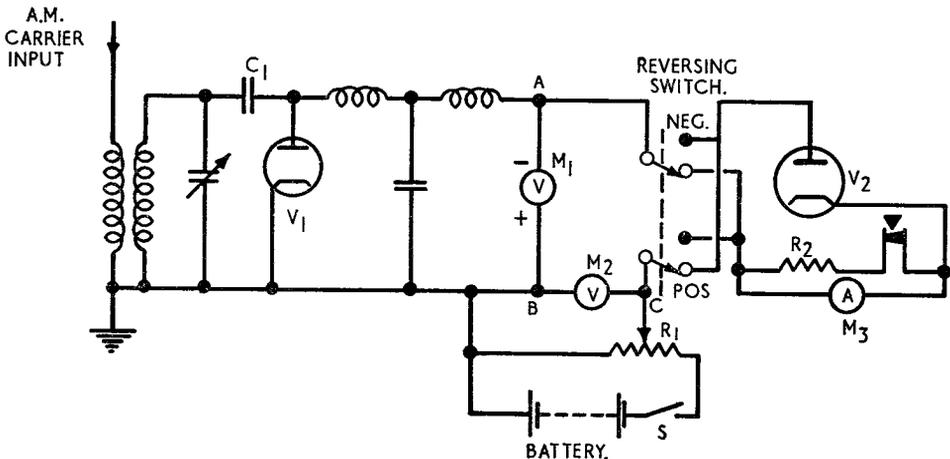
Fig. 3. CIRCUIT OF A DOUBLE-RECTIFIER MODULATION METER.

mined by the setting of R_1 and is indicated on the meter M_2. The polarity of the voltage is opposite to that across AB and so the *difference* between those two voltages will appear across AC.

6. These points A and C are also connected via a reversing switch to a second rectifier circuit consisting of diode V_2, meter M_3 and the press button shunting circuit. Thus the difference voltage across AC is applied to V_2. With the reversing switch in the position shown, when point A is negative with respect to point C, current will flow through M_3. This condition exists when the varying d.c. voltage exceeds the battery voltage. If R_1 is adjusted until current just begins to flow through M_3, then M_2 will read the maximum value of the rectified modulated wave. This is the positive modulation peak.

When the reversing switch is placed in the opposite position to that shown in Fig. 3, i.e. in the NEG position, and the same adjustment made to R_1, then M_2 will indicate the negative modulation peak.

7. The meter M_1, connected across AB will indicate the carrier voltage. By putting these meter readings in the following formulae, the positive percentage modulation, the negative percentage modulation and the average percentage modulation of the carrier can be obtained.

Positive percentage modulation

$$= \frac{E_p - E_c}{E_c} \text{ x } 100$$

Negative percentage modulation

$$= \frac{E_c - E_a}{E_c} \text{ x } 100$$

Average percentage modulation

$$= \frac{E_p - E_a}{2E_c} \text{ x } 100$$

Where E_p = positive modulation peak (meter M_2 reading)

E_a = negative modulation peak (meter M_2 reading)

E_c = carrier voltage (meter M_1 reading)

8. When taking measurements the meter M_3 must be initially protected by the shunt resistance R_2. After preliminary adjustment of R_1 the shunt resistance can be removed and final adjustment of R_1 can then be made.

Monitor Meter Method

9. To monitor the percentage modulation of a transmitter, the meter must give a direct and continuous indication of the percentage modulation. A suitable circuit for such a meter is shown in Fig. 4(a). This circuit is
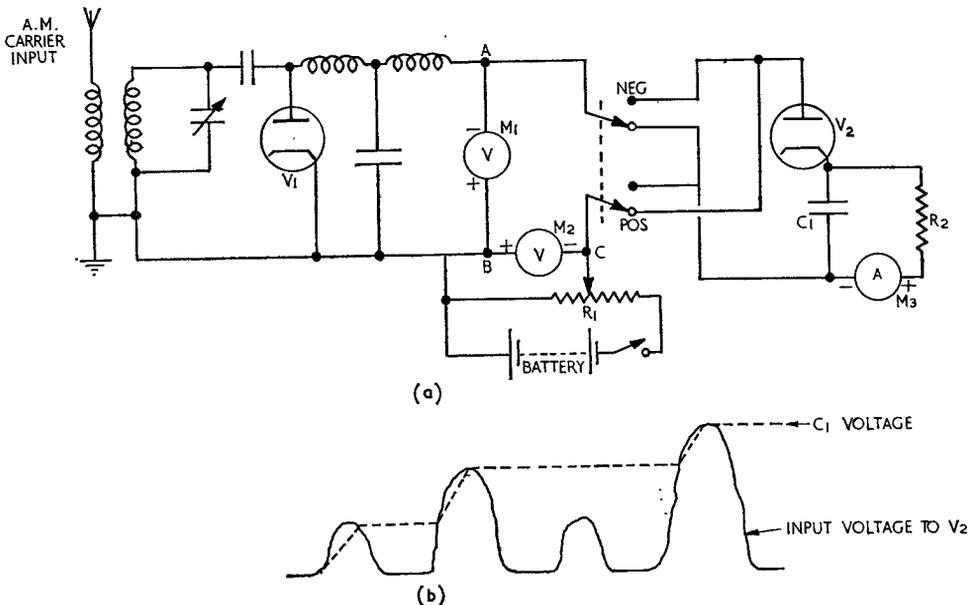
Fig. 4. MONITOR METER.

similar to that of Fig. 3. The main difference is that a long CR circuit (C_1 and R_2) is included in the cathode of V_2. With this arrangement R_1 is adjusted to give equal readings in M_1 and M_2. With a modulated input the voltage across AC varies in accordance with the modulation signal. This voltage is applied via the reversing switch to V_2.

When the anode of V_2 is positive with respect to the cathode, the capacitor C_1 charges quickly through the diode, to the peak voltage of the modulating signal (Fig. 4(b)). When V_2 is cut off C_1 will discharge slowly through M_3 and R_2; as R_2 is a large value resistor, the voltage across C_1 is maintained almost equal to the peak of the modulation signal. The meter M_3 is calibrated directly in percentage modulation. By means of the reversing switch, M_3 will read either the negative modulation peaks or the positive modulation peaks.

Oscilloscope Methods of Measuring Percentage Modulation

10. The c.r.o. provides a convenient means of indicating the percentage modulation of a transmitter, either as a monitor, giving a continuous indication, or as a means of measuring the percentage modulation.

When the c.r.o. is coupled to the r.f. output tank circuit of the transmitter in the manner shown in Fig. 5, the pattern on the c.r.t. screen shows the actual shape of the modulation envelope, if the usual linear sawtooth waveform is applied to the horizontal plates. This pattern will vary contin-
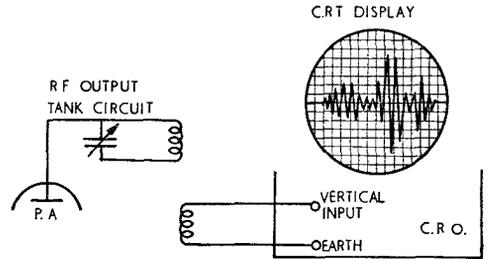
Fig. 5. METHOD OF MONITORING A.M. CARRIER.

uously with the degree of modulation, making it difficult to measure. However, this type of presentation provides a quick means of detecting overmodulation in single-tone modulated outputs.

Trapezoidal Pattern

11. By connecting the c.r.o. into the transmitter circuit in the manner shown in Fig. 6, a means of accurately measuring the percentage modulation is available.

The modulating voltage is fed to the horizontal deflection system of the c.r.o. instead of the linear time-base and the output from the tank circuit of the transmitter is fed to the vertical deflection system. With the modulator switched off, a single vertical trace as shown in Fig. 6(a) appears on the c.r.t. screen. The height (X) of this trace is proportional to the unmodulated carrier voltage. This trace is centred on the c.r.t. screen by means of the horizontal and vertical shift controls and the height is adjusted to a convenient number of graticule divisions, by means of the vertical gain control.

(a) UNMODULATED CARRIER   (b) 50% MODULATION   (c) 90% MODULATION   (d) OVER 100% MODULATION

(e) METHOD OF CONNECTING C.R.O.
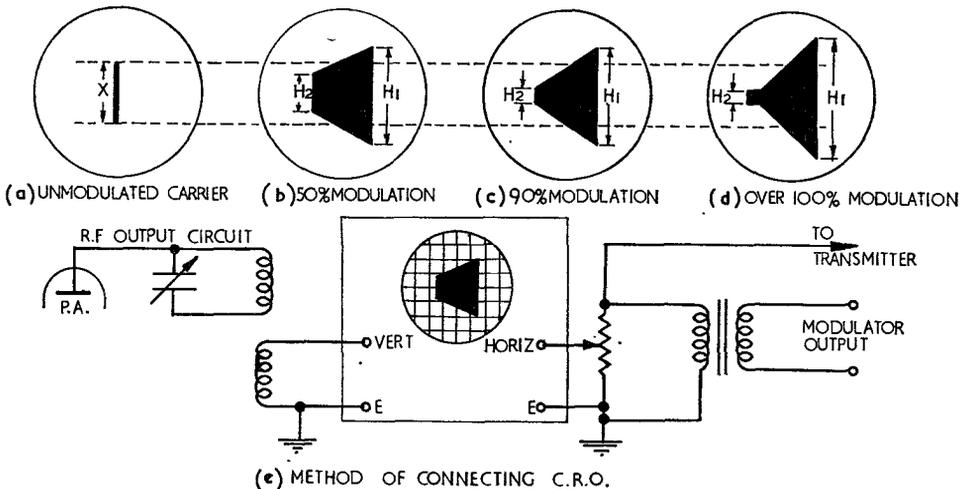
Fig. 6. OSCILLOSCOPE METHOD OF MEASURING PERCENTAGE MODULATION.

12. When the modulator is operating, a trapezoidal pattern is formed and the horizontal gain is adjusted to produce a pattern of convenient size. The longer side of the pattern (H$_1$) indicates the modulation peaks and the shorter side (H$_2$) indicates the modulation troughs. By taking these two measurements from the c.r.t. graticule, the percentage modulation can be calculated thus.

Percentage modulation

$$= \frac{H_1 - H_2}{H_1 + H_2} \times 100$$

At 100% modulation a triangular pattern is formed (H$_2$ = 0 and H$_1$ = 2X); modulation in excess of 100% causes the pattern to form a tail and H$_1$ becomes greater than 2X (Fig. 6(d)).

13. This type of pattern provides an easy method of measuring the percentage modulation of a speech modulated carrier, and clearly indicates over-modulation.

Frequency Modulation Measurements

14. When a wave is frequency modulated the carrier amplitude remains constant and the *frequency* of the wave is made to vary about the carrier frequency at a *rate* corresponding to the audio frequencies of the modulating signal.

The *amount* by which the frequency changes in one direction from the unmodulated frequency is dependent on the *amplitude* of the modulating signal and is called the *frequency deviation*. It corresponds to the change of carrier amplitude in an amplitude modulated wave. Frequency deviation is usually expressed in kilocycles, and is equal to the difference between the carrier frequency and either the highest or lowest frequency reached by the carrier in its excursions when modulated. There is no modulation percentage

in the normal sense, but the maximum deviation is determined by the width of the band assigned to the station. In f.m. broadcasting the band is limited to 75 kc/s either side of the carrier centre frequency, and 100% modulation is said to occur when the frequency deviation equals this predetermined maximum. If the modulation exceeds 100%, distortion does not occur but interference with other stations may result.

Frequency Deviation Measurement

15. The simplest method of measuring the frequency deviation of a f.m. transmitter is to apply a d.c. voltage to the grid of the frequency modulator (Fig. 7). By varying the value of this d.c. voltage, the oscillator frequency can be varied. A heterodyne frequency meter is set to the maximum or minimum specified oscillator frequency and the d.c. voltage is adjusted to produce a zero beat. The value of the d.c. voltage is then equal to the peak a.c. modulating voltage which would produce the equivalent of 100% modulation. The audio modulating signal is then arranged not to exceed this peak.

If frequency multiplication stages follow the oscillator stage, the final carrier frequency is of course the oscillator frequency times the multiplication factor of the subsequent multiplying stages. Thus the final carrier frequency deviation is the oscillator deviation times the same multiplication factor.

16. If the transmitter circuit is such that the above method is not practicable, frequency deviation can be measured by means of a deviation meter. This consists of a low sensitivity receiver, incorporating a frequency discriminator, an audio amplifier and a valve voltmeter (Fig. 8). The valve voltmeter is calibrated directly in frequency deviation against a signal with known amounts of frequency deviation superimposed.
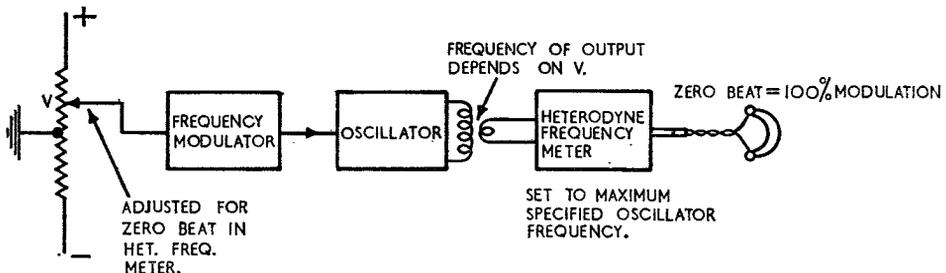
Fig. 7. BASIC METHOD OF MEASURING FREQUENCY DEVIATION.

FREQUENCY MODULATED SIGNAL → RECEIVER → DISCRIMINATOR → A.F. AMPLIFIER → VALVE VOLTMETER

Fig. 8.   BASIC FREQUENCY DEVIATION METER.

Summary

17.   This chapter has outlined the basic methods of making modulation measurements.   It has dealt with simple circuits for measuring average and peak amplitude modulation, and has described the use of an oscilloscope to make these measurements. Frequency modulated transmitters often have built-in frequency deviation meters designed for use with the particular transmitter, but the basic principles of these meters are as described in this Chapter.

SECTION 18

CHAPTER 5

POWER MEASUREMENTS AND STANDING WAVE MEASUREMENTS

POWER MEASUREMENTS AND STANDING WAVE MEASUREMENTS

Introduction

1. In radio equipment servicing, the frequencies at which power measurements have to be made vary from the mains frequency (50 c/s) to the output frequency of a radar transmitter which can be 10,000 Mc/s or greater. Thus the instruments and methods used to measure power over this frequency range vary considerably.

The instrument used to measure mains power, the dynamometer, was discussed in Book 1, Section 6, Chapter 3, and the units used in power measurements, the watt and the decibel, have also been discussed. In this chapter, power measurements at frequencies from audio frequencies to super high frequencies will be considered.

Audio-frequency Power Measurement

2. The unit which is commonly used to measure the a.f. power of an amplifier or receiver is the decibel (see Book 1, Section 6, Chapter 3). This unit is a measure of the ratio of two powers P_1 and P_2: the a.f. power output P_2 of an amplifier or receiver can be measured by comparing it with a reference power level P_1. The reference power used is often 1 milli-watt when the unit is written as dbm.

The instrument used for making a.f. power measurements is the decibel-meter and is usually a rectifier type a.c. voltmeter (possibly a valve voltmeter) with a scale calibrated in decibels. It may be self-contained, with built-in loads to dissipate the power being measured, or a voltmeter calibrated in decibels may be used to measure the power dissipated in an external load. When the decibel meter is calibrated, a reference point based on a specific power, or value of voltage across a specific resistance, is selected to read zero decibels. Based on this reference point various voltage readings are made on the same voltage scale and after conversion, these readings are marked on the decibel scale.

A decibel scale is often incorporated in a multimeter for power comparison measurements, and is used in conjunction with the voltage or current ranges. When changing from one range to another, the decibel indication will change, although the input is

kept constant. Consequently, if readings on the two ranges are to be compared, a correction to the new reading is necessary. The amount of this correction (plus or minus) depends on the ratio of the two ranges, and is given in the instrument handbook.

Receiver Power Output Measurement

3. At frequencies below the u.h.f. band, power measurements are usually made by measuring the current flowing through, or the voltage developed across, a resistor of known value.

A simple method of measuring the audio power output of a receiver uses this principle. A resistor of value as specified in the receiver handbook, is connected across the output terminals of the receiver in place of the earphones or speaker, and an a.c. voltmeter is connected in parallel with it. With a signal generator delivering an input voltage to the receiver aerial terminals, the voltmeter measures the output voltage across the resistor. The power output can then be calculated from the formula:

$$\text{Receiver power output} = \frac{E^2}{R} \text{ watts}$$

where E = Voltage across standard resistor
and R = Standard resistor value.

Transmitter Power Measurement—LF to VHF

4. At frequencies from the l.f. band to the v.h.f. band, the power output of a transmitter can be measured by connecting a dummy aerial and a thermo-ammeter into the transmitter aerial circuit as shown in Fig. 1. The
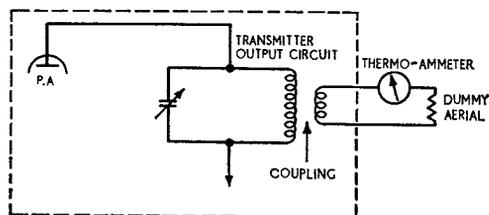
Fig. I. MEASUREMENT OF TRANSMITTER OUTPUT POWER.

dummy aerial consists of non-inductive resistors of a value to suit the output impedance of the transmitter being tested. These resistors replace the aerial during this test and absorb the r.f. energy from the transmitter.

The thermo-ammeter is connected in series with the dummy aerial and the transmitter is correctly tuned. The power absorbed by the dummy aerial depends on the coupling between it and the transmitter and so the coupling must be adjusted to give maximum power output. The meter reading is then noted and the power output of the transmitter calculated from the formula:

Transmitter power output $= I_2R$ (watts) where R is the terminal resistance of the dummy load and I is the thermo-ammeter reading.

R.F. Power Meter

5. This is a test instrument which gives a *direct* reading of r.f. power. It can be designed to operate on frequencies up to 300 Mc/s and can be made compact and portable.

The power to be measured is dissipated in a suitable resistor and the voltage developed across the resistor is rectified and used to operate a meter calibrated directly in watts.

6. The circuit of a typical r.f. power meter is shown in Fig. 2. When the meter is connected to the transmitter output, the load resistor R_L takes all the r.f. output and
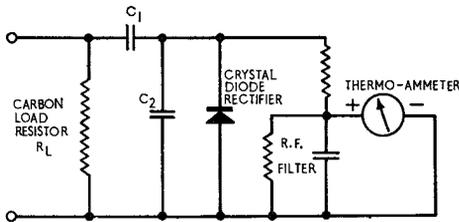
Fig. 2. R.F. POWER METER.

must be capable of dissipating all the energy supplied to it. It is often in the form of a carbon pile, consisting of carbon discs held together by spring pressure. The value of the load resistor must be equal to the output impedance of the transmitter output circuit, if the meter is connected directly to it; if it is connected to the output circuit through a transmission line, it must equal the characteristic impedance of the line. Under these

conditions maximum power is developed in the load resistor.

7. In the circuit of Fig. 2, C_1 has a small capacitance (about 5 pF) and C_2 a relatively large capacitance (about 0.25 μF). Thus the capacitive reactance across the load resistance is high, and will not cause a noticeable mis-match. Further, the r.f. voltage developed across C_2 is small and easily handled by the crystal diode rectifier. The r.f. component of the rectified voltage is by-passed by the r.f. filter and the d.c. component is used to operate the meter.

High powers must not be applied to the meter for any longer than is necessary to take a reading, otherwise the heat generated will cause the value of the load resistor to change and the meter calibration will become inaccurate. The highest frequency at which the meter can be used is limited by the shunting effect of the capacitors C_1 and C_2.

U.H.F. and S.H.F. Power Measurement

8. Because of difficulties in measuring voltage, current and resistance at ultra-high frequencies and above, the method of power measurement previously described cannot be used for frequencies in the u.h.f. and s.h.f. bands. At these frequencies, a *bolometer* is often used to measure power. Two common types of bolometer, the thermistor and the barretter are illustrated in Fig. 3. The bolometer has a resistive element that is temperature sensitive. When r.f. power is absorbed by this element, the resultant temperature change causes a change of resistance which can be used to produce a meter reading in an external circuit.

9. The thermistor shown in Fig. 3(a) consists of a bead of nickel oxide with manganese, uranium or copper oxide as a binding agent. It has a *negative* temperature coefficient, i.e. its resistance decreases with an increase in temperature. Thus when the thermistor bead absorbs r.f. energy, its temperature rises and its resistance falls.

The barretter of Fig. 3(b) has a *positive* temperature coefficient and consists of a very thin tungsten wire supported inside a quartz capsule. As the power absorbed by the barretter increases, its resistance increases.

The Thermistor Bridge

10. The bridge circuit shown in Fig. 4 depends for balance upon the ratio of the resistances. Resistors R_1, R_2, and R_3 are

GLASS
CAPSULE

THERMISTOR BEAD

CONNECTING WIRES.

(a) BEAD TYPE THERMISTOR.

TUNGSTEN WIRE

QUARTZ CAPSULE
METAL ENDS
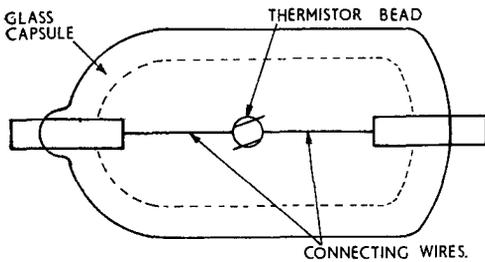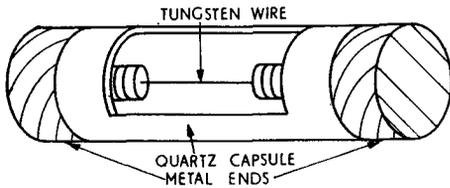
(b) BARRETTER.

Fig. 3. TYPES OF BOLOMETER.

equal in value; the balance control alters the current flowing in the circuit and this current is adjusted until the resistance of the thermistor equals that in each of the other arms. The bridge is then balanced and the meter reads zero.

THERMISTOR

R1

R3   R2

BALANCE
CONTROL
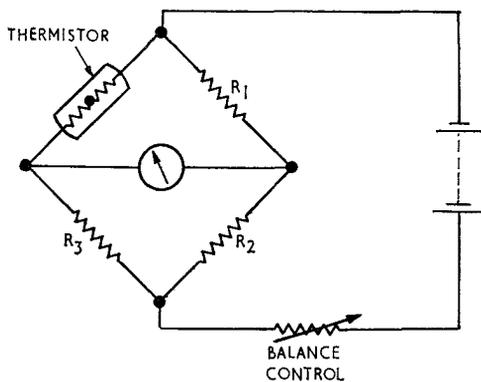
Fig. 4. BASIC THERMISTOR BRIDGE CIRCUIT.

11. To measure u.h.f. power the thermistor, connected to the bridge circuit, is inserted into the output circuit of the equipment under test in such a way that the thermistor bead absorbs r.f. energy. The manner of insertion will depend upon the equipment being tested. The resistance of the thermistor decreases in proportion to the amount of r.f. energy absorbed and the bridge becomes unbalanced. The bridge meter will then

give a reading from which the power output can be calculated.

The thermistor (and the barretter) can measure low powers only, and so must be used in conjunction with a sampling device such as an attenuator, r.f. probe or directional coupler. Usually the thermistor is built into a coupling device of known attenuation thus enabling the total power output of the equipment to be calculated from the measured power.

STANDING WAVE MEASUREMENTS

Introduction

12. A transmission line terminated with an impedance which is not equal to the characteristic impedance of the line will cause partial reflection of an incident wave, and standing waves will be set up along the line. The manner in which these standing waves are formed was discussed in detail in Chapter 3 of Section 15 and there it was shown that the standing wave ratio (s.w.r.) is given by

$\dfrac{E_{max}}{E_{min}}$ $\left(\text{or } \dfrac{I_{max}}{I_{min}}\right)$ where E_{max} is the

maximum amplitude of the s.w. voltage (antinode) and E_{min} is the minimum amplitude. The s.w.r. therefore gives an indication of the degree of mismatch of the line. Mismatch can also be caused by incorrect construction of a transmission line and by faulty components in the line.

Thus by measuring the s.w.r. on a transmission line the degree of mismatch can be assessed, and an indication of any faults in the line can be obtained. Methods used to make standing wave measurements will now be discussed.

Standing Wave Measurements

13. The type of test equipment suitable for measuring standing waves varies with the type of line involved. For an open-wire line a single-loop coil mounted a fixed distance below one of the wires will couple with the magnetic field around the wire. The e.m.f. induced in the loop will cause a current to flow in a suitable meter connected to the loop. By moving the coupling loop along the wire the meter reading will vary in proportion to the value of the current standing wave on the line (see Fig. 5). The maximum and minimum values indicated can be used to calculate the s.w.r.
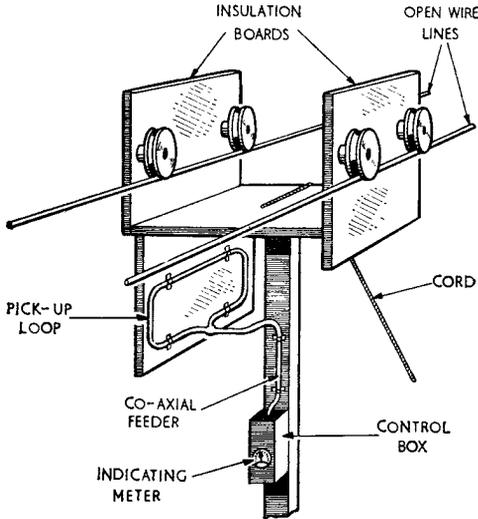
Fig. 5. STANDING WAVE INDICATOR SUITABLE FOR OPEN LINES.
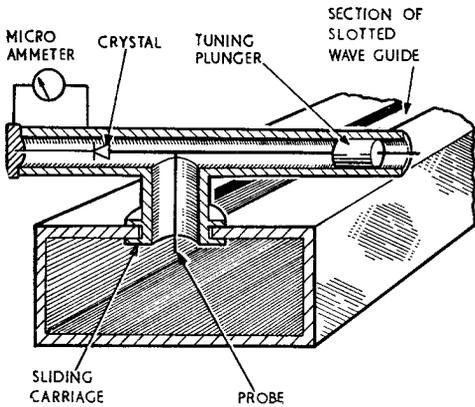
Fig. 6. SECTION THROUGH A WAVEGUIDE STANDING WAVE DETECTOR.

14. With coaxial cable or waveguide the standing wave detector consists of a length of suitable coaxial cable (or waveguide) with a slot cut along its axis (Fig. 6). An r.f. probe is mounted in the slot such that the probe can be moved along the slot. The probe is coupled to an indicating meter either directly or through an amplifier. When in use the standing wave detector forms part of the transmission line system under test and the probe is moved along the slot, which must be greater than half a wavelength long. The meter will then indicate the maximum and minimum value of standing wave voltage and from these readings the s.w.r. can be calculated.

15. The block diagram of Fig. 7 shows the arrangement of the test equipment necessary to measure a standing wave ratio. The output from an r.f. signal generator is fed, via an attenuator and standing wave detector, into the transmission line under test. For correct indications the input to the transmission line must be matched to the line characteristic impedance, and the line must be correctly terminated with a dummy load to dissipate the r.f. energy. With this arrangement any component in the transmission line system, can be adjusted to give optimum s.w.r.

16. In some ground radar installations a standing wave measuring device is permanently fitted into the transmission line system. It takes the form of a short length of transmission line inserted in series with the main transmission line, and fitted with several vertical neon tubes along its length. Each tube is excited by an r.f. probe projecting into
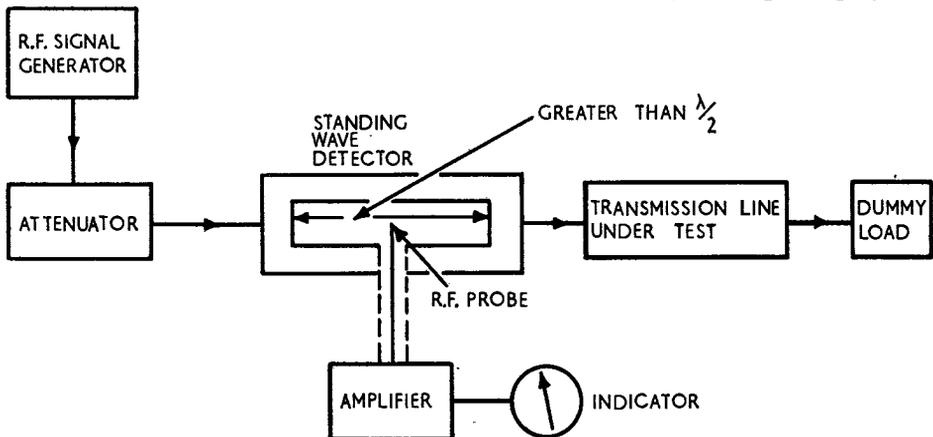
Fig. 7. EQUIPMENT FOR TESTING TRANSMISSION LINE COMPONENTS.

the main transmission line, and is so designed that the height of the neon glow up the tube is proportional to the r.f. voltage exciting the tube. A calibrated scale placed behind the tubes enables the relative amplitudes of the nodes and antinodes to be measured, and from these readings the s.w.r. can be obtained.

Impedance Measurement Using S.W.R.

17. It was shown in Section 15 Chapter 3, that the s.w.r. on a transmission line depends on the terminating impedance. Thus the value of an unknown terminating impedance can be found by measuring the s.w.r. and applying the relationship

$$\text{S.W.R.} = \frac{Z_0}{Z_T}$$

where Z_0 = line characteristic impedance

Z_T = terminating impedance

Then $Z_T = \dfrac{Z_0}{\text{s.w.r.}}$

Wavelength or Frequency Measurement

18. The principle of wavelength measurement using a short-circuited length of transmission line was discussed in Chapter 1 Paragraph 17. By measuring the distance between adjacent nodes of either current or voltage, the wavelength, and hence the frequency of the energy on the line, can be determined. Since the velocity of propagation of the r.f. energy down the transmission line is always less than that in free space, the half wavelength measured in the line will be less than that corresponding to the same frequency in free space. Thus it is necessary to divide the measured wavelength by the velocity factor of the line. With open-air spaced lines such as the lecher bars, this factor is almost unity, but with coaxial lines it may be as low as $0 \cdot 6$.

The Reflectometer

19. Standing waves are set up on a transmission line as a result of the combination of incident and reflected waves. Thus standing wave measurements are really an indirect measurement of these two waves, and when the line is almost correctly matched, the effect of the greatly reduced reflected wave becomes very difficult to detect in standing wave measurements. A more accurate method of measuring the s.w.r. under well-matched conditions is to measure the amplitudes of the incident and reflected waves separately and from these calculate the s.w.r. thus:

$$\text{v.s.w.r.} = \frac{V_i - V_r}{V_i + V_r}$$

where V_i = amplitude of incident wave and

V_r = amplitude of reflected wave.

20. The instrument used for making such measurements is called a *reflectometer*. It consists of a directional coupler which is inserted into the transmission line under test. A small coupling loop inside the directional coupler can be rotated through $180°$, and can sample the incident and reflected waves. The voltages induced in the loop are suitably amplified and detected and then applied to a sensitive meter. The instrument is adjusted so that the voltage due to the incident wave brings the meter pointer to a setting mark, and then, when the loop is reversed the meter reading due to the reflected wave is read against a scale calibrated in voltage standing wave ratio.

Conclusion

21. This Section has discussed the methods of making a few common measurements necessary in servicing and testing radio equipment. It has given a background of the theory of some measuring instruments, thus enabling the technician to correctly interpret the readings he takes. To obtain results of any value the technician must be fully familiar with any instrument he uses; for detailed information concerning any test equipment, the relevant handbook or publication should be consulted. For more detailed information on radio measurements reference should be made to A.P. 2900 C (Handbook of Electronic Test Methods and Practices).

SECTION 19

CONTROL SYSTEMS

SECTION 19

CONTROL SYSTEMS

SECTION 19

CHAPTER 1

REMOTE INDICATION AND CONTROL

REMOTE INDICATION AND CONTROL

Introduction

1.   It is often necessary to note or record the value or any change in value of physical quantities (e.g., angular position, speed, temperature, etc.) at a point remote from the physical quantity itself.

There are many instances in radio engineering where angular movement of an input shaft must be reproduced accurately by motion of a second shaft, often at some considerable distance from the first. For example, in an aircraft, the loop aerial used for direction finding is placed at a suitable point in the aircraft and is often not readily accessible to the operator. A bearing indicator, if mounted at the base of the aerial shaft, would be inconvenient. This difficulty is overcome by ensuring that the movement of the loop aerial shaft is reproduced accurately by the motion of a second shaft in a *remote* indicator, placed at some convenient point in the aircraft.

A direct mechanical linkage, such as a flexible drive, between the two shafts is possible, but because of their separation distance there are practical difficulties of installation: in addition, there are inherent inaccuracies, and the efficiency of the system is poor. Much more satisfactory results are obtained by using electrical remote indication systems.

Electrical remote indication systems are sometimes referred to as '*data transmission systems*'. This term, however, is nowadays normally taken to have a much wider meaning: it is used, for example, to describe the method by which information is fed to a computer. Because of this, in order to avoid confusion, the term 'data transmission' is not used in this Chapter.

2.   In electrical remote indication systems, the movements of the input shaft are translated into suitable electrical signals by a device known as a *transducer* or *transmitter unit*. A transducer (not to be confused with the transductor in a magnetic amplifier) measures the physical movement in terms of some electrical quantity whose magnitude is a strict measure of the movement. The electrical signals from this transmitter unit are then transmitted through wire links (and in certain cases, radio links) to appropriate receiver units located at any desired position: the received signals are used to turn a shaft which gives the remote indication or the required movement.

In the simple system outlined above, no torque amplification is provided: the torque developed in the output shaft is therefore less than that developed in the input shaft and the power required is provided by the input. Thus only moderate torques can be developed—in many cases only sufficient to move a light pointer over a graduated scale. For remote *indication* of such things as D/F bearings, or the position of a radar scanner, this system is normally adequate (Fig. 1).

There are many occasions, however, when accurate remote control of the *position* of
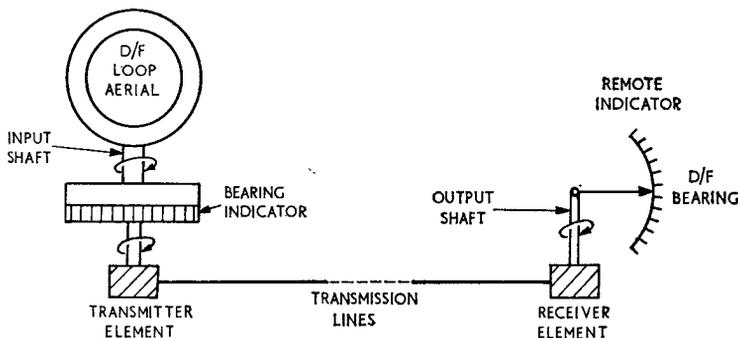
Fig. I.   ELECTRICAL REMOTE INDICATION.

a heavy load is required (e.g. remote rotation of a radar scanner). Torque amplification is now necessary and to provide the required torque, use is made of hydraulic or electric amplifiers.

Many different devices are used to give remote indication of angular position or to control the movement of heavy loads from a distance. Some are operated from a d.c. supply; others from an a.c. supply. Some of the methods used in the Service are considered in the following paragraphs.

D.C. SYSTEMS

Desynn

3. **Introduction.** The Desynn system of transmission is a simple system which, because of its low torque characteristic, is useful only for remote *indication* of angular position. It is ideal where a simple pointer and scale indicator is adequate. Aircraft applications include remote indication of flap, rudder and elevator positions, and of D/F loop and compass readings. It is also used in ground installations to repeat the reading of an instrument at a remote point. The accuracy of the system is of the order of $\pm$ 2°, and this is sufficient for the applications mentioned.

4. **Circuit.** As in all electrical remote indication systems, the input shaft is connected to a transmitter element, and the output shaft, which operates the remote indicator, is driven by a receiver element: the transmitter and receiver are connected by electrical lines.

In the Desynn system (Fig. 2) the transmitter is a continuous resistance ring or toroidal potentiometer, which has three fixed tappings A, B, C spaced 120° apart and connected to the receiver. A rotating

spring-loaded mechanism mounted on the input shaft, carries two sliding contacts or wipers that are at diammetrically opposite points on the toroid. The wipers are fed, via slip rings and brushes, from the positive and negative lines of the d.c. supply.

The receiver has three high resistance coils with axes at 120° in space (like the star-connected stator winding of an a.c. induction motor): within them is a permanent magnet rotor which is capable of rotation and which carries a pointer over a calibrated scale. The three coils in the receiver are connected to the tapping points A, B, C on the transmitter by the three lines as shown in Fig. 2.

5. **Operation.** When a d.c. supply is connected to the transmitter wipers, the voltages at the tapping points A, B, C cause currents to flow through the three stator coils in the receiver, and a resultant magnetic field is produced. The rotor magnet aligns itself with this field. The magnitude and polarity of the voltage at each tapping point in the transmitter vary according to the position of the wipers. Thus, if the input shaft is rotated, the variation of voltage at A, B, C produces changes in the currents flowing in the stator coils and a magnetic field rotating in sympathy with the input shaft is produced. The rotor magnet remains aligned with this field at all times and so rotates in synchronism with the input shaft.

6. This action is illustrated in Fig. 3, where the wipers of the transmitter are connected to opposite poles of a 24V d.c. supply. With the input shaft in the position shown in Fig. 3(a), the voltage distribution round the toroid is such that point A is 24 volts positive with respect to supply negative, while B and C are both 8 volts
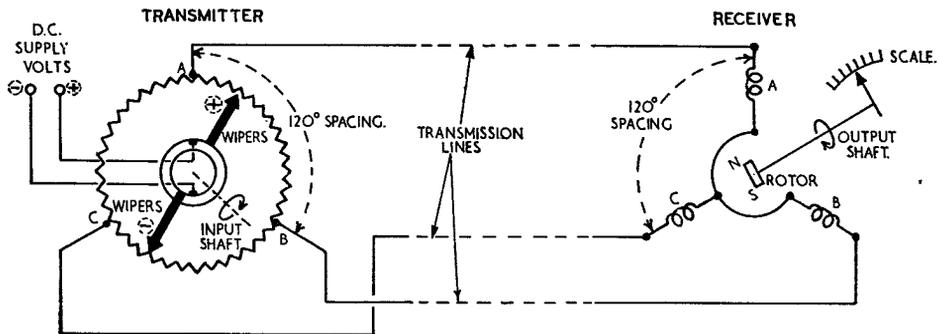
Fig. 2. THE DESYNN SYSTEM.
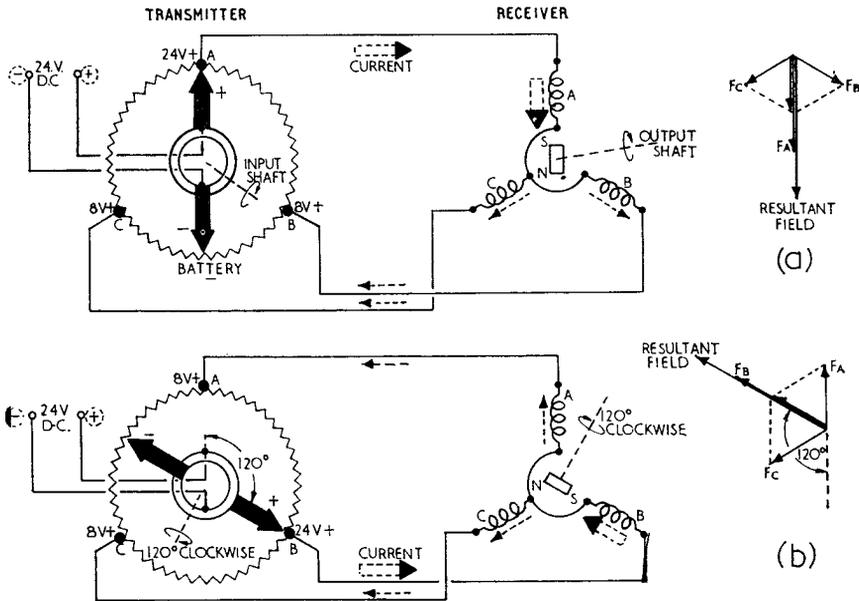
Fig. 3. OPERATION OF DESYNN.

positive with respect to supply negative. With A positive by the same amount to both B and C, current flows from A through coil A in the receiver; it then divides equally at the star point and half the total current flows through coil B and half through coil C back to transmitter. The individual and resultant magnetic fields produced by these currents are indicated by vectors, and with the input shaft in the position shown in Fig. 3(a), the axis of the resultant field lies vertically: the rotor magnet aligns itself with this axis.

If the input shaft is rotated 120° clockwise from its initial position, as shown in Fig. 3(b), the voltage distribution round the toroid is such that current flows from B through coil B in the receiver; it then divides equally and flows through coils A and C back to the transmitter. The vectors show that the resultant magnetic field has also rotated through 120° clockwise from its initial position and the rotor magnet aligns itself along this new axis.

7. **Summary.** Enough has been said to show that if the wipers in the transmitter are placed in any position by the input shaft, the resultant field at the receiver and hence the rotor magnet take up corresponding

positions. Thus as the input shaft rotates through 360°, the rotor follows this movement *in the same direction;* if a pointer, moving over a calibrated scale, is attached to the rotor, remote indication of the position of the input shaft is immediately available. A typical example of the use of the Desynn is remote indication of bearing from the D/F loop—the loop shaft acting as the 'input' shaft.

M-type Transmission System

8. **Introduction.** Many practical indicators take the form of light geared mechanisms which are required to rotate fairly substantial drum-type indicators or comparable devices. For example, it may be required to rotate the deflection coils in a magnetic c.r.t. in synchronism with a radar aerial to produce the necessary trace on the screen. This requires a larger torque than is available with the Desynn system: in such circumstances, an M-type or 'step-by-step' transmission system can provide the moderate torque required.

9. **Circuit.** A toroidal potentiometer type of transmitter cannot control large currents to the receiver (i.e. allow the receiver to

TRANSMITTER   RECEIVER
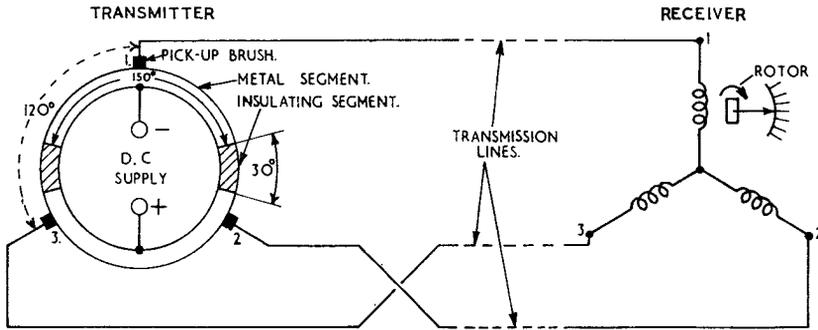
Fig. 4. M-TYPE TRANSMISSION.

develop much torque) because a low-resistance transmitter would be needed and this would overheat and tend to burn out. In the M-type transmission system, therefore, the transmitter unit is modified considerably from that in the Desynn system: the receiver, however, operates on the same principle.

The essential features of a simple M-type transmission system are shown in Fig. 4 The transmitter is basically a drum-type switch, the metal drum of which consists of two segments each spanning an arc of 150°, separated by two sections of insulating material each extending over 30°. The two metal segments are connected to

opposite poles of a suitable d.c. supply, and three 'pick-up' brushes are disposed round the drum at intervals of 120°.

The receiver unit (more than one unit may be operated from a single transmitter to give multiple indication) is similar to that in the Desynn system, although the rotor may be either a permanent magnet or a laminated soft-iron core. The outer end of each coil in the receiver is connected, via a transmission line, to one of the three pick-up brushes in the transmitter.

10. **Operation.** The action of the M-type transmission system is illustrated in Fig. 5.

In position 1, the input shaft driving the
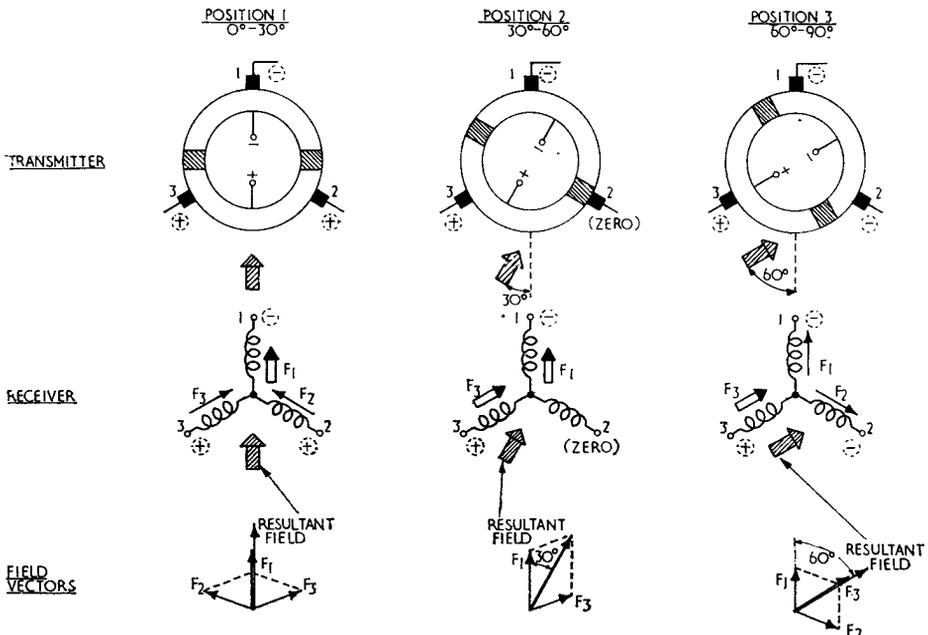
Fig. 5. OPERATION OF M-TYPE TRANSMISSION SYSTEM.

transmitter drum is in such a position that brush 1 is connected to − of supply and brushes 2 and 3 to +. These polarities are applied to the three coils in the receiver, with the result that the current divides through coils 2 and 3, *all* the current flowing through coil 1. Magnetic fields F_1, F_2, F_3 are produced and resolving these three fields by vector gives a resultant in the direction shown.

If now the input shaft rotates through 30° clockwise (position 2), brush 1 is −, brush 2 is disconnected by the insulating segment and brush 3 is +. At the receiver, equal currents flow through coils 1 and 3, while coil 2 carries no current. By resolving the magnetic fields F_1 and F_3 produced by these currents, the resultant field is seen to have rotated through 30° in a clockwise direction in sympathy with the input shaft.

The condition after a further 30° rotation of the input shaft is shown at position 3. The resultant field at the receiver has now rotated through 60° clockwise from its initial position, thus following the input shaft.

In each case, the receiver rotor aligns itself with the axis of the resultant field and therefore follows the angular movement of the input shaft, *but only in discrete steps of 30°*.

11. With the arrangement shown, there is a change of pick-up brush polarity at one or other of the brushes each time the input shaft is turned through 30°. The complete pattern showing how in turn, the brushes are connected to +, −, 0 when the drum is rotated through 360° is given in Table 1. There are, of course, 12 steps in a complete rotation: the first 3 steps correspond to position 1, 2 and 3 in Fig. 5.

For certain purposes, the 30° step is too large, and in such cases a modified transmission system giving 24 steps of 15° each may be used. The principle remains the same, but accuracy is improved.

The maximum operating rate of the M-type transmission system is dependent on the inductive time constant of the receiver coils. The system in general use is designed to give practical operating speeds of up to 180 steps per second.

12. **Drum transmitter.** The basic principle of the M-type transmission system has been

Transmitter Position	Pick-up Brush Polarity		
	1	2	3
0° − 30°	−	+	+
30° − 60°	−	0	+
60° − 90°	−	−	+
90° − 120°	0	−	+
120° − 150°	+	−	+
150° − 180°	+	−	0
180° − 210°	+	−	−
210° − 240°	+	0	−
240° − 270°	+	+	−
270° − 300°	0	+	−
300° − 330°	−	+	−
330° − 360°	−	+	0

TABLE I.
POLARITY CHANGES DURING
360° ROTATION OF TRANSMITTER.

described in a previous paragraph. A practical example of the transmitter unit used in this description is illustrated in Fig. 6. It is so constructed that the d.c. supply brushes are in contact at all times with opposite metal segments, whilst the pick-up brushes are in contact with either of the metal segments or with the insulated segment, depending on the position of the input shaft (i.e., connected to +, −, or 0 as previously explained).

13. **Commutator transmitter.** A schematic diagram of a commutator transmitter designed to give 24 step (15°) M-type sequence is shown in Fig. 7. It is similar to the drum type but much more compact. It consists of a thin commutator face plate rotating against five fixed brushes that are held in a support plate. The commutator is made up of three concentric rings of metal, separated by rings of insulating material. The outer and inner rings are both electrically continuous throughout and two of the five

3 OUTPUT BRUSHES
120° APART

INSULATING SEGMENT

SUPPLY BRUSH
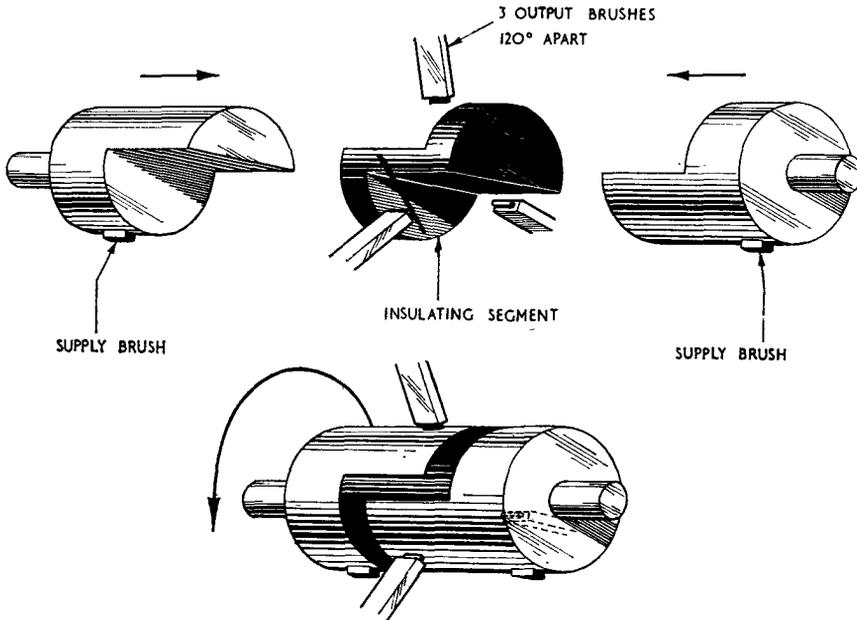
SUPPLY BRUSH

Fig. 6. M-TYPE DRUM TRANSMITTER.

brushes, one connected to supply + and the other to supply −, bear on the inner and outer rings respectively. The centre ring is divided into four equal portions by four 'islands' of insulation, each extending over an arc of 15°: the segments so formed are connected to the outer and inner rings alternately. In the middle (divided) ring there is, therefore, a regular +, 0, − sequence.

The remaining three brushes are positioned to bear on the middle ring, each of the outer brushes being displaced from the centre brush by 60°. With this arrangement there is a change of polarity at one or other of the three pick-up brushes each 15°

D.C. SUPPLY BRUSHES

INSULATED SEGMENT.

PICK-UP BRUSHES 60° SPACING

OUTER RING(−)

INNER RING(+)

15°

Fig. 7. COMMUTATOR TRANSMITTER.

rotation of the face plate: a 24 (15°) step M-type sequence is thus produced at the output brushes.

14. **Cam-type transmitter.** There are two main types of cam-operated transmitter, but the more common of the two types, and the one considered here, is the eccentric cam transmitter. It consists of a single circular cam, mounted *eccentrically* on the input shaft and operating three pairs of two-position switches, via push rods. The switches are fitted radially around the cam with 60° spacing as shown in Fig. 8, and the inner contact of each switch is permanently energised, + or − as illustrated. The two switches in a pair are diametrically opposed and control the polarity on one line feeding the receiver.

With the input shaft in the position shown at (a) of Fig. 8, with maximum eccentricity opposite switch 2, switches 1, 2 and 3 are open; switches 4, 5 and 6 are closed. Thus line 1 is −, line 2 is +, and line 3 is −.

On turning the input shaft through 30° in either direction from the initial position, the maximum eccentricity of the cam wheel is midway between two of the switches. Thus, at position (b) of Fig. 8, switches 1, 2, 3 and 4 are open, and switches 5 and 6
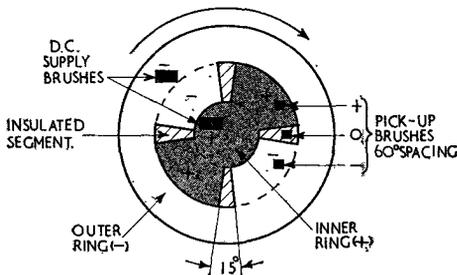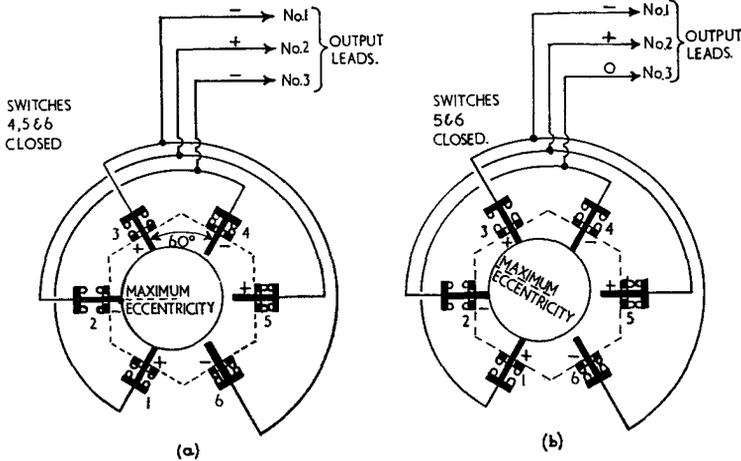
Fig. 8. CAM-TYPE TRANSMITTER.

only are closed. In this condition, line 1 is −, line 2 is +, and line 3 is 0.

A standard M-type sequence of line polarity changes at 30° steps is therefore established, 12 steps being provided for each complete revolution.

The drum and commutator transmitters both suffer from brush wear. The eccentric cam transmitter does not. Because of this, the eccentric-cam type is used to a greater extend than the others where a high operating rate has to be maintained.

15. M-type receivers.
As has already been stated, the stator of an M-type receiver is similar to that of an a.c. induction motor and to that in a Desynn receiver. The rotor, however, may be of the soft iron (inductor) type or it may be a permanent magnet.

The inductor rotor is built up of iron and aluminium laminations, and the laminated rotor continuously re-aligns itself with the resultant field axis of its stator to offer the path of lowest reluctance—i.e., when the laminations are in line with the resultant flux. Since this type of rotor is non-polarized it is possible for it to align itself in either of two positions 180° apart.

The permanent magnet rotor is more common. A disadvantage of the earlier types of rotor magnet was that they tended to become demagnetised after a time. This does not happen with modern materials such as Alnico. Because of the relatively strong magnetic field produced by the magnet,

the rotor torque is considerably higher than that of the indicator rotor unit. Higher stepping rates are also achieved for the same reason and, being polarized, the rotor lines up in one position only.

16. Transmitter and receiver synchronisation.
The fact that the receiver in an M-type transmission system moves in 30° (or 15°) steps is a disadvantage. If greater accuracy is required, the input shaft can be geared up to the transmitter shaft: a 60 : 1 gear system is common, the transmitter shaft completing 60 revolutions for each revolution of the input shaft. The receiver is equally geared *down* to the output shaft, if a 1 : 1 input-to-output ratio is required. Although the accuracy of the system is improved 60 times by this means, there is now the possibility of ambiguity. This can be seen as follows:—

The switching sequence is completed in 12 (30°) steps, and with a 60 : 1 gear system between the input shaft and the transmitter, the sequence is completed for a rotation of only $\frac{360°}{60} = 6°$ of the primary drive, i.e. in 12 steps of $\frac{1}{2}°$ each. Since the transmitter completes 60 revolutions for each revolution of the input shaft, there are 60 different positions in the full 360° movement of the *primary* drive, each separated by 6°, into which the receiver rotor can 'lock' and still follow the M-type sequence. However, in all but one of these positions, the
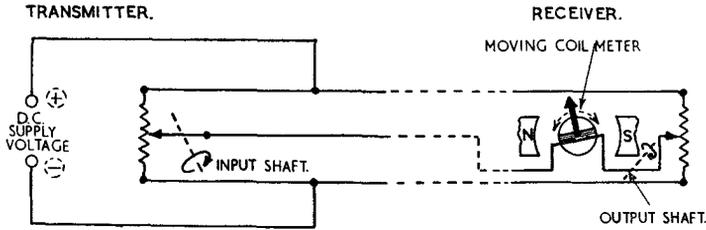
Fig. 9. SIMPLE WHEATSTONE DRIVE.

output shaft will be out of synchronisation with the input shaft.

In such circumstances it is necessary to connect to the receiver a local manually-controlled transmitter acting as a 'coarse' control. The coarse control is needed to bring the output shaft into synchronism with the input shaft by comparison of dial readings. When this has been done, the local transmitter is disconnected and the receiver again connected to the transmission line and hence to the remote transmitter.

Wheatstone Bridge System

17. **Introduction.** The accuracies of the Desynn and M-type transmission systems can never be better than about $\pm$ 2° because of frictional and resistive losses. Where greater accuracy is required, an error-operated or follow-up system can be used. The Wheatstone bridge system is a typical example for d.c., when continuous rotation is not a requirement.

18. **Circuit and action.** A circuit arrangement is illustrated in Fig. 9. The transmitter and receiver are both potentiometers connected to form a Wheatstone bridge. The

wiper on the transmitter potentiometer is controlled by the input shaft. A moving coil meter, acting as the remote indicator is connected between the two wipers, and the arrangement is such that rotation of the moving coil moves the wiper on the receiver: the sense of rotation tends to reduce the coil current to zero.

If the input shaft is rotated, the bridge becomes unbalanced and current flows through the moving coil, which also rotates, moving the pointer and the receiver wiper. This movement continues until the receiver wiper reaches the position at which the bridge becomes balanced. In this way, the receiver wiper (and pointer) copy exactly the movements of the transmitter wiper (and input shaft). The accuracy of this system is better than $\pm$ 1°, but the torque developed at the receiver is very small.

Greater output driving torque can be developed by modifying the basic Wheatstone bridge system. The moving coil meter can be replaced by a polarized relay which is used to switch a d.c. supply to a small motor, as shown in Fig. 10.

At balance, the relay is de-energised and the supply to the motor is disconnected.
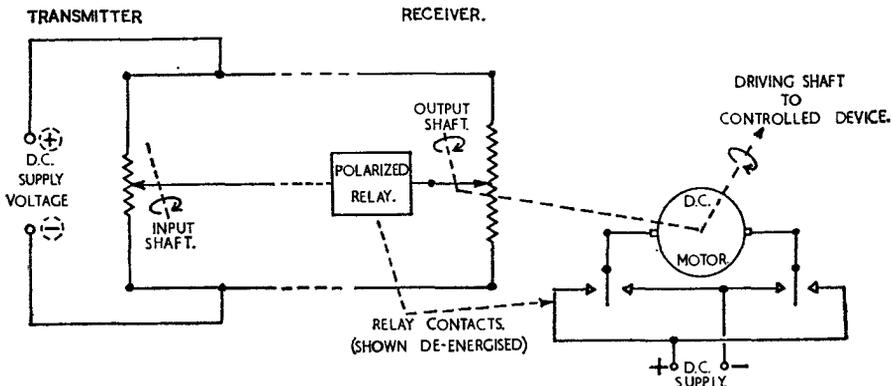
Fig. 10. TORQUE-AMPLIFIED WHEATSTONE DRIVE.

If the input shaft is rotated, the bridge becomes unbalanced and the relay is energised. The sense of unbalance determines the direction in which the relay is energised: this, in turn, determines which set of contacts is closed and thus the direction of rotation of the motor. The motor rotates in such a direction that the receiver wiper moves towards the balance condition: at the same time the controlled device is moved to the desired position.

This system is used for remote control and tuning of radio equipment.

A.C. SYSTEMS

Introduction

19. The d.c. transmission systems described earlier are limited in their practical application to remote *indication* of position of a shaft and the transmission of *moderate* torque to a remote device: also, in general, the degree of accuracy is of the order of $\pm 2°$.

Where a remote transmission system is required to operate efficiently with a high degree of accuracy, or where it is required to be used as part of a servomechanism to move a heavy load at a remote point, an a.c. system is generally preferred.

20. A.C. transmission systems have many applications: they include:—

(*a*) Remote indication of position or movement.

(*b*) The transmission of moderate torque to a remote device with a high degree of accuracy.

(*c*) The controlling element in a servo-mechanism system used to control the position and speed of heavy loads at a remote point.

(*d*) The summation of two or more mechanical movements.

(*e*) Analogue computation.

21. A.C. transmission systems are referred to generally as '*synchros*' because of the self-synchronous characteristic of the systems, i.e., any movement of the input shaft connected to the transmitter is exactly reproduced in the angular movement of the remote output shaft connected directly or indirectly to the receiver.

Synchros are manufactured by many firms and are known by various trade names such as Selsynn, Magslip, Synchrotic, Autosyn, Aysynn, and Telesyn. All of them, however, work on the same basic principles.

22. Synchro systems can be divided into several categories according to their function. The three main categories are as follows:—

(*a*) **Torque synchros.** These are the simplest form of synchro: *no torque amplification* is provided. They are used for the transmission of angular position information by electrical means and for the reproduction of this information by the position of the shaft of the receiver element. Moderate torque only is developed in the output shaft and the main use of torque synchros is in instrument repeater systems.

(*b*) **Control synchros.** These normally form part of a power amplifying servomechanism system. Such a system can provide almost any degree of torque amplification and can therefore be designed to handle heavy loads such as a directional aerial array on a turntable. The control synchro, in effect, provides the data on which the servomechanism acts.

(*c*) **Resolver synchros.** These are used extensively in computers to convert voltages, which represent the cartesian co-ordinates of a point into a shaft position and a voltage which together represent the polar co-ordinates of the point. They can also be used for conversion from polar to cartesian co-ordinates.

Although the construction of these different categories differs in detail, their action can be appreciated by considering the elements used in a basic synchro system.

Basic Synchro System

23. The basic synchro system for transmission of continuous rotation is illustrated in schematic form in Fig. 11. The transmitter and receiver are electrically similar. Each consists of a rotor carrying a single winding, round which is a stator on which are three windings arranged with their axes at 120° in space.

The principle physical difference between transmitter and receiver is that the receiver is usually fitted with low-friction ball bearings on the rotor and a mechanical damper to

TRANSMITTER.   RECEIVER.
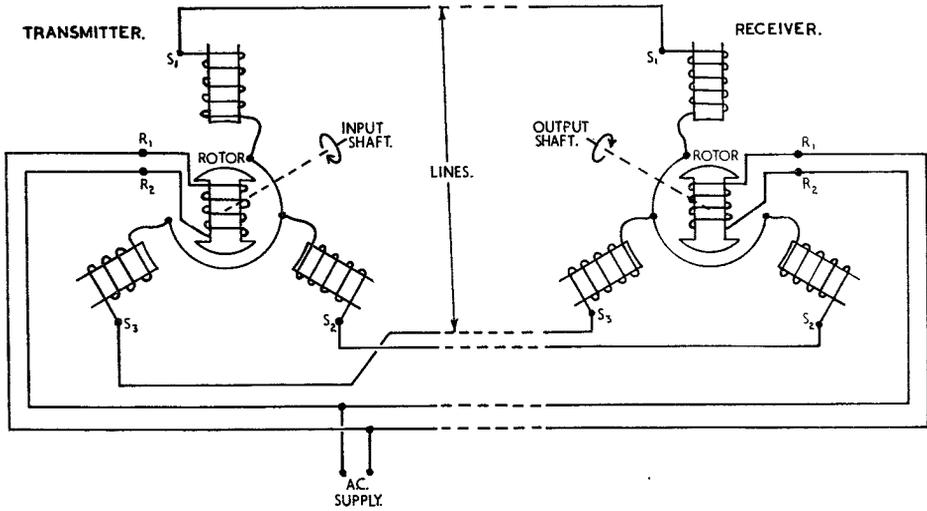
Fig. 11.   BASIC SYNCHRO SYSTEM.

reduce oscillation, whereas the transmitter is not damped and often has no ball bearings.

In operation, both rotors are energised from an a.c. supply (often 115V 400 c/s) and the corresponding stator connections are joined together by transmission lines.

24.   The schematic representation of the synchro unit used in Fig. 11 and shown at (a) of Fig. 12 is useful for certain purposes, but the simplified version of the same symbol, shown at (b) of Fig. 12, is generally sufficient. For detailed circuit diagrams, the symbol shown at (c) is used.

A synchro unit is said to be positioned at *electrical zero* when the axis of the rotor is in line with the axis of the S_1 winding of the stator.   This same position is indicated in Fig. 12 (c), when the two arrows are lined up.

Construction of Synchro Unit

25.   The construction of a simple form of synchro transmitter is shown in Fig. 13. The construction of a synchro receiver is similar with the addition of an oscillation damper.

The stator body is made up of internally-slotted laminations, in the slots of which are fitted the three sets of windings S_1, S_2, S_3 spaced 120° apart and arranged in a similar fashion to the stator coils of a small three-phase induction motor.   Despite the similarity, the stator windings of a synchro unit must not be confused with normal three-phase windings.   In a three-phase machine, the voltages in the three windings of the stator are equal in magnitude and 120° apart in phase: in a synchro unit, the voltages are *not equal* in magnitude and
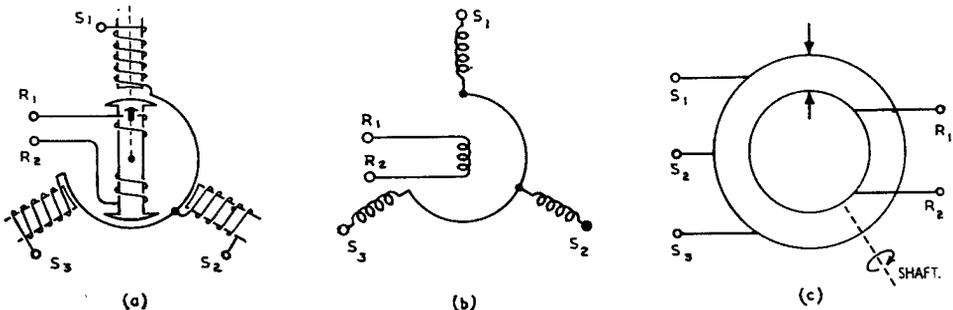
Fig. 12.   TORQUE SYNCHRO SYMBOLS.
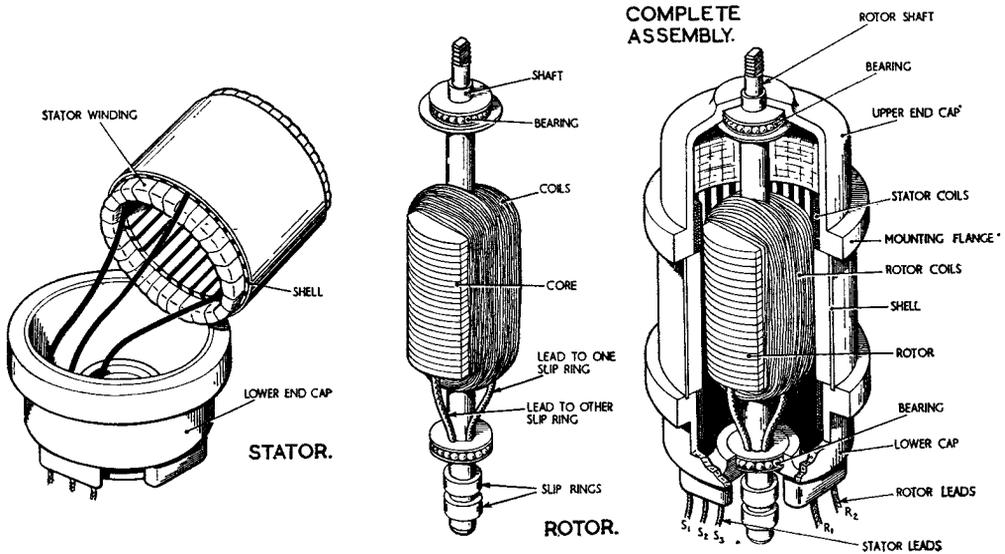
Fig. 13. CONSTRUCTION OF TORQUE SYNCHRO.

are either in phase or 180° out of phase with each other: in a synchro unit, the stator windings are said to be "space-phased".

The laminated rotor carries a single winding, the two ends of which are connected, via slip rings and brushes, to two terminals R_1 and R_2, which are in turn connected to the a.c. supply.

Principle of Torque Synchro System

26. **Basic operation.** A simple torque synchro system is illustrated in Fig. 14. The rotors of transmitter (TX = torque trans-

mitter) and receiver (TR = torque receiver) are both energised from the a.c. supply and produce an alternating flux which links with their corresponding stators. Should the relative dispositions of rotor to stator in the two elements be different, the three voltages induced in each of the two stator windings by the alternating fluxes differ: currents then flow between the two stators and a torque is produced in each synchro which is so directed as to eliminate the discrepancy; thus, in effect, to align the two rotors.

Normally, the transmitter rotor is held mechanically by the input shaft and the
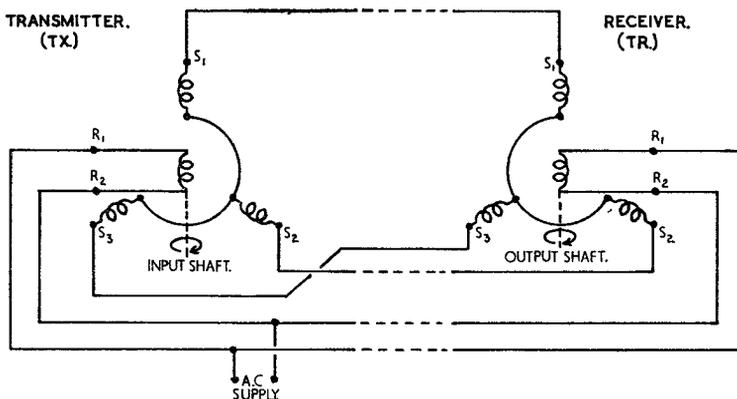
Fig. 14. TORQUE SYNCHRO SYSTEM.

receiver rotor is free to turn so that it aligns itself with the transmitter rotor. Thus, in Fig. 14 any movement of the transmitter rotor is repeated synchronously by the movement of the receiver rotor.

27. Operation of transmitter. To understand why the receiver rotor follows the transmitter rotor, it is necessary to consider the manner in which the transmitter stator voltages change as the input shaft is turned.

Consider Fig. 15, in which the transmitter rotor is connected to the a.c. supply but the stator windings are disconnected from the receiver. A current flows in the rotor and
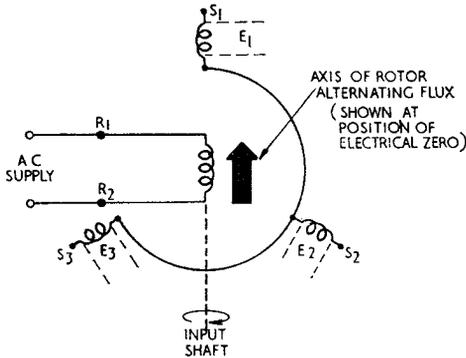
Fig. 15. OPERATION OF SYNCHRO TRANSMITTER (TX)

sets up an alternating magnetic flux in the transmitter: the direction of this flux is along the axis of the rotor winding. The flux links with the stator windings, and induces alternating voltages in each of them. The synchro unit is therefore, like a single-phase transformer in which the rotor winding is the primary and the stator windings are three secondary coils.

28. The magnitudes of the voltages induced in the stator windings depend on the relative number of turns in rotor and stator coils and on the orientation of the rotor.

For the position of the rotor in Fig. 15, voltage E_1 has its maximum value (the axis of the rotor winding and the axis of stator coil S_1 are in line): E_1 is also *in phase* with the applied voltage E. As the rotor is turned clockwise from this position, E_1 decreases and becomes zero when the rotor has turned through 90°. Further turning of the rotor causes a voltage of *reversed* phase to appear across the stator coil S_1.

29. Similar reasoning applies for the voltages E_2 and E_3 induced in the stator coils S_2 and S_3 respectively. A graph showing the variations in the magnitudes of the stator voltages with the angular position of the rotor is plotted in Fig. 16.
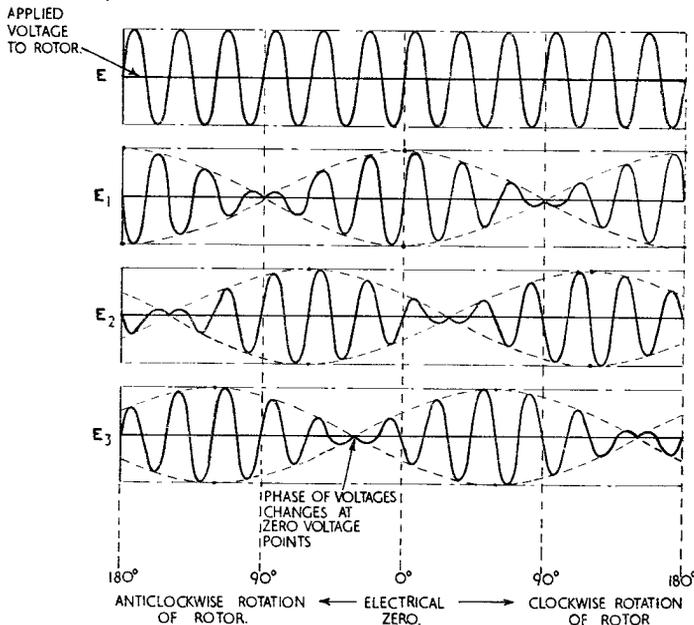
Fig. 16. OUTPUT FROM TRANSMITTER (TX)
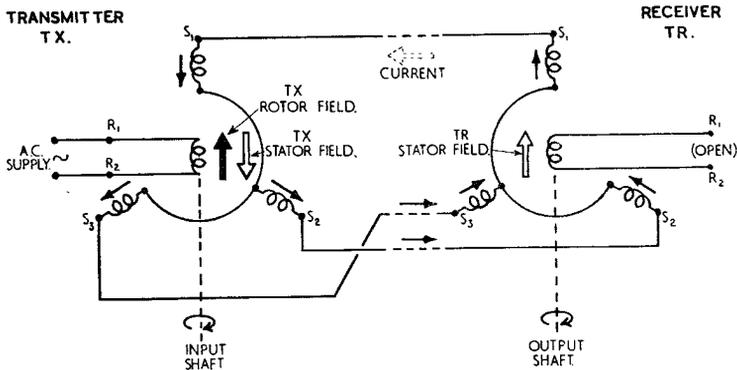
Fig. 17. PRODUCTION OF RECEIVER (TR) STATOR FIELD.

It is again emphasised that the stator voltages do *not* constitute a set of three-phase voltages. The stator voltages, in effect, are modulated with sinusoidal envelopes that differ in phase by 120° as the rotor is turned at a constant speed. The individual cycles of the stator voltages, however, are either in phase or 180° out of phase with each other and with the rotor voltage.

30. Consider now what happens when the stator coils in the transmitter and receiver are connected: for the moment, the receiver rotor is disconnected from the supply. The arrangement is shown in Fig. 17.

The voltages induced in the transmitter stator coils by the alternating rotor flux are applied to the receiver stator coils and currents flow through the closed circuits formed by the two sets of stator windings. The magnitude and phase of the current through each coil of the *transmitter* stator depend on the magnitude and phase of the induced voltage in each coil, and this in turn depends on the orientation of the transmitter rotor. The current through each coil of the transmitter stator produces a magnetic field and the three fields combine to form a single resultant magnetic field inside the stator. Since the rotor winding is, in effect, the primary winding of a transformer with the stator windings acting as three secondary coils, the resultant flux produced by the currents in the stator coils must at all times balance that produced by the rotor current: this is normal transformer action. The directions of the transmitter rotor and stator magnetic fields are therefore *opposite* as shown in Fig. 17.

31. **Operation of receiver.** The current flowing through each coil of the *receiver* stator is of the same magnitude as that flowing through the corresponding coil of the transmitter stator. It is however, in the *opposite* direction. Thus, since the transmitter and receiver stators are identical in form, the resultant magnetic field established inside the receiver stator is equal in magnitude but opposite in direction to that produced in the transmitter stator. The receiver stator field thus coincides, both as regards axis line and direction of flux, with that of the *transmitter rotor*. This is also indicated in Fig. 17.

32. A bar of soft iron placed in a magnetic field tends to align itself parallel to the field. Thus the rotor of the synchro receiver tends to turn into alignment with the receiver stator field, even though the rotor winding is open. Operation with open rotor winding is undesirable however, because for a given position of the transmitter rotor, the receiver rotor can take up one of two positions 180° apart: in addition, the torque developed by the receiver is small.

33. These difficulties are avoided if the rotor winding of the receiver is connected to the a.c. input as in Fig. 18. The rotor is now an electromagnet excited by alternating current and is said to be *polarized*. The magnetic field produced by the current in the two rotors are, of course, identical both in magnitude and direction.

The receiver rotor is normally free to turn under the influence of any applied magnetic force: such a force is developed by interaction of receiver rotor and stator fields
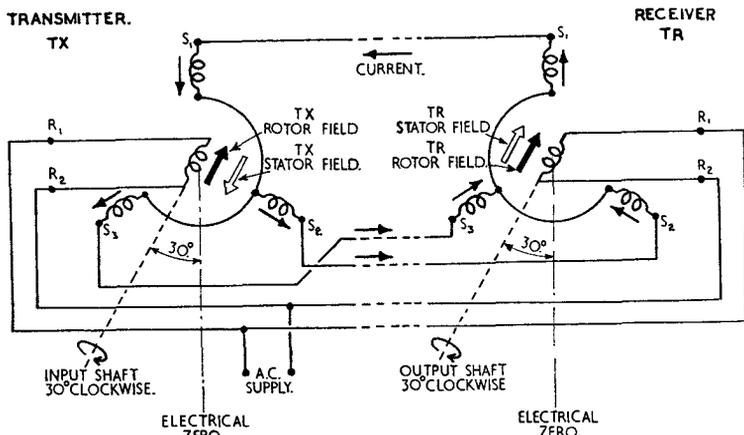
Fig. 18. ALIGNMENT OF OUTPUT AND INPUT SHAFTS.

and the receiver rotor turns until the two fields are in alignment. It has already been stated that the receiver stator field coincides with the field of the *transmitter rotor:* hence the receiver rotor synchronously follows any movement of the remote transmitter rotor.

34. Accuracy and efficiency. The torque synchro draws negligible current from the supply. This is because the voltages induced in each set of stator windings from their respective rotors are in opposition: thus, when the receiver rotor is correctly aligned with the input shaft, the two sets of voltages balance and negligible stator current is drawn. The only current drawn from the supply under this condition is that required to energise the rotors.

Because the stator currents decrease as the receiver rotor comes into alignment, the torque developed by the receiver rotor also decreases and this produces an inherent inaccuracy in the system: the accuracy is of the order of $\pm 1°$. As the load increases so does the degree of error and also the current drawn from the supply. The torque synchro is, therefore, suitable only for driving relatively light loads such as those used in instrument repeaters.

35. Magslip. It is possible to connect several receivers in parallel and drive them from a single transmitter, to give indication in a number of places simultaneously. But in this case, all the units react on each other, so that if one receiver develops a fault all the other receivers give false indi-
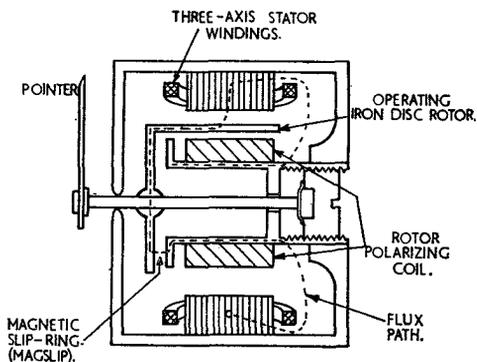
Fig. 19. MAGSLIP.

cations. The interaction can be reduced by using a 'magslip' as the synchro receiver (Fig. 19). The magslip has the normal three-coil stator, but the rotor is a disc of iron having a single tongue projecting under the stator winding. The 'rotor' coil does not in fact rotate, which removes any slip ring friction and the extra friction on the bearings due to the weight of the coil. It is fed from the a.c. supply, and the flux it produces is coupled to the stator by the magnetic slip ring (magslip) between the rotor disc and the rotor polarizing coil. The rotor disc turns until the current in the transmitter and receiver stator coil is zero; the rotor is then in alignment.

It will be seen that there is a long air-gap in the rotor flux path; consequently, any rotor misalignment has little reaction on the stator flux distribution, and therefore there is little reaction between receivers connected
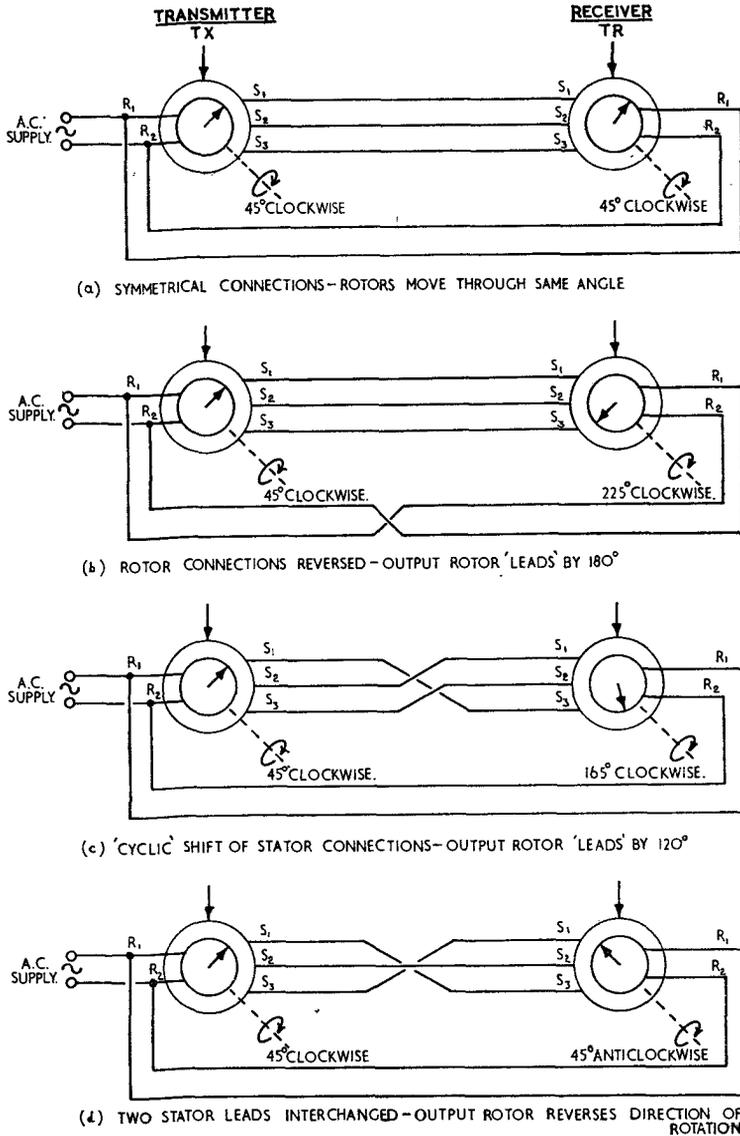
Fig. 20. VARIOUS INTERCONNECTIONS IN TORQUE SYNCHRO SYSTEM.

in parallel. But the long gap means that the rotor flux is weak and the torque available is very small.

36. Torque synchro connections. So far, only symmetrical connections between transmitter and receiver windings have been considered. However, re-arrangement of rotor and stator connections between transmitter and receiver produces different results. The receiver rotor still moves synchronously with the transmitter rotor, but it can do so from a different reference position or in the reverse direction. Fig. 20 illustrates the possibilities.

Torque Differential Synchro Systems

37. Introduction. In the torque synchro transmission system so far considered the 'output' as represented by the angular move-
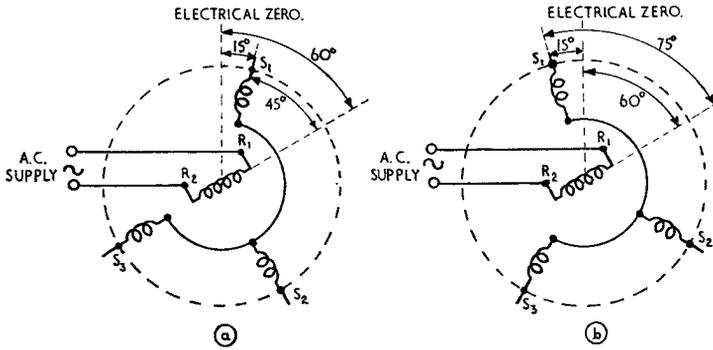
Fig. 21. DIFFERENTIAL ACTION IN TORQUE SYNCHRO SYSTEM.

ment of the receiver rotor, is simply a reproduction of a single 'input', i.e., the angular movement of the transmitter rotor.

Under certain conditions, however, it is necessary to transmit *two* angular positions, the synchro receiver indicating the difference or the sum of the two angles.

38. One simple way of achieving this is to rotate the *stator* coils of the synchro transmitter through one angle and the *rotor* through the other angle. This is indicated in Fig. 21. In (a) the rotor is rotated through 60° clockwise from electrical zero and the stator is rotated 15° *in the same direction:* the *relative* angle between rotor and stator is the *difference* between the two angles,

namely 45°, and the electrical output of the transmitter is such that the receiver turns 45° clockwise. In (b) the addition of two angles is shown.

Mechanical displacement of both rotor and stator in a synchro transmitter is not generally practicable. It is normally better to insert a torque differential synchro in the transmission chain. This differential synchro can operate either as a transmitter or as a receiver. Since the torque differential transmitter is more common, it will be considered first.

39. **Torque differential transmitter (TDX).** The differential transmitter has a stator identical with that of a synchro transmitter
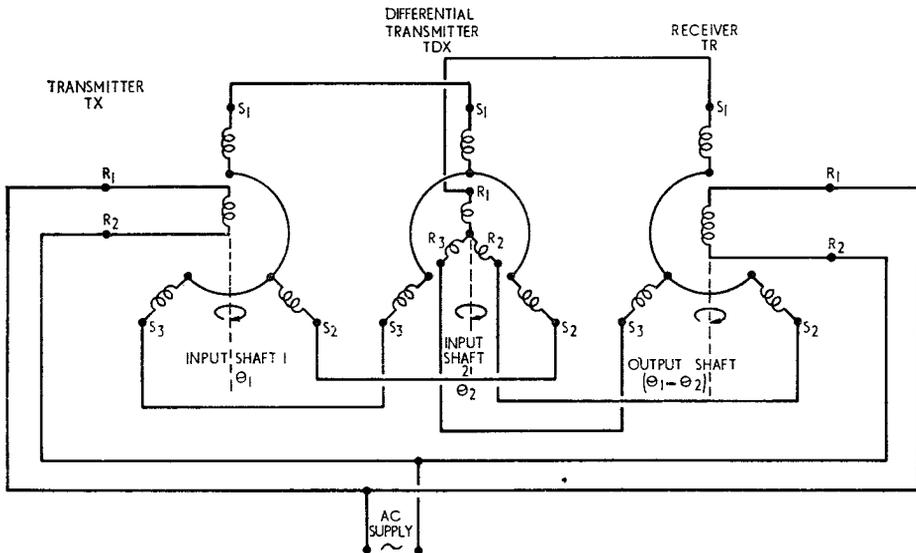
Fig. 22. TORQUE DIFFERENTIAL SYNCHRO SYSTEM.

or receiver. It differs from the transmitter or receiver in that it has a *cylindrical*, instead of a two-pole rotor; and the rotor like the stator, has three distributed windings spaced 120° apart.

40. The circuit shown in Fig. 22 is that of a differential synchro system set up for the *subtraction* of two inputs. The arrangement is such that one input shaft turns the transmitter (TX) rotor and a second input shaft drives the differential transmitter (TDX) rotor. The differential transmitter receives an electrical signal corresponding to a certain angular position of the transmitter (TX) rotor: it modifies this signal by an amount corresponding to the angular position of its own rotor: it then transmits the modified electrical output signal to the receiver (TR).

This modified output signal produces an angular position of the flux in the receiver which, for Fig. 22, is the *difference* of the rotor angles of the two transmitters (TX and TDX).

The circuit symbol for a differential synchro transmitter is shown in Fig. 23.
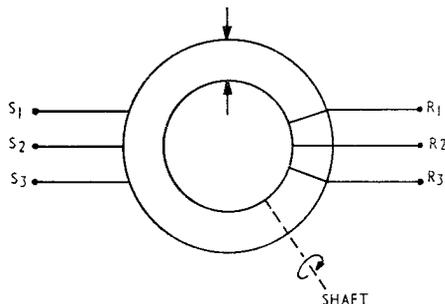
Fig. 23. SYMBOL FOR TORQUE DIFFER-
ENTIAL TRANSMITTER (TDX).

In the differential synchro system, the rotors of the normal transmitter (TX) and receiver (TR) are supplied in parallel with single-phase a.c. The stator windings of the transmitter are connected to the stator windings of the differential transmitter (TDX) and the three windings on the rotor of the latter are connected to the windings of the receiver stator. Note that the rotor of the differential transmitter is *not* connected to the a.c. supply.

41. **Action of TDX.** The action of the torque differential synchro system set up for subtraction is illustrated in Fig. 24.

In (a) both input shafts are at electrical zero and the distribution of current throughout the system is such that the magnetic fields are in the direction shown. Thus the output (TR) rotor also takes up the position of electrical zero.

In (b) shaft 1 is rotated through 60° clockwise and shaft 2 remains at electrical zero. All magnetic fields rotate as shown and the output (TR) rotor also rotates 60° from electrical zero.

In (c) shaft 1 is at electrical zero and shaft 2 is rotated 15° clockwise. The magnetic fields of the transmitter (TX) and the differential transmitter (TDX) remain in the electrical zero position because their position is determined by the orientation of the transmitter (TX) rotor. However a 15° clockwise rotation of the TDX rotor without a change in the position of its field is equivalent to moving the rotor field 15° *anti-clockwise* whilst leaving the rotor at electrical zero. This shift in the position of the TDX rotor field relative to the rotor itself is duplicated in the receiver (TR) stator windings and the output rotor aligns itself with its stator field. Thus the output rotor moves 15° *anti-clockwise* for a 15° *clockwise* movement of the differential (TDX) rotor.

42. It is now easy to see that if both input shafts are rotated simultaneously in a clockwise direction, the receiver rotor turns through an angle equal to the *difference* between the two input angles, i.e., a clockwise movement of the TX rotor gives a clockwise movement of the TR rotor, whereas a clockwise movement of the TDX rotor gives an anti-clockwise movement of the TR rotor. This is illustrated in Fig. 24(d) where the TR rotor turns through $(60° - 15°) = 45°$ clockwise.

43. The differential effect is of course reversed when the differential rotor is moved in the opposite direction to the transmitter rotor. The receiver rotor now moves through an angle equal to the *sum* of the two input angles. However, it is more usual to rotate the input shafts in the same direction and to alter the connections between the various elements to obtain the required output. Various possibilities are illustrated in Fig. 25.

44. **Torque differential receiver (TDR).** A torque differential synchro system can use a
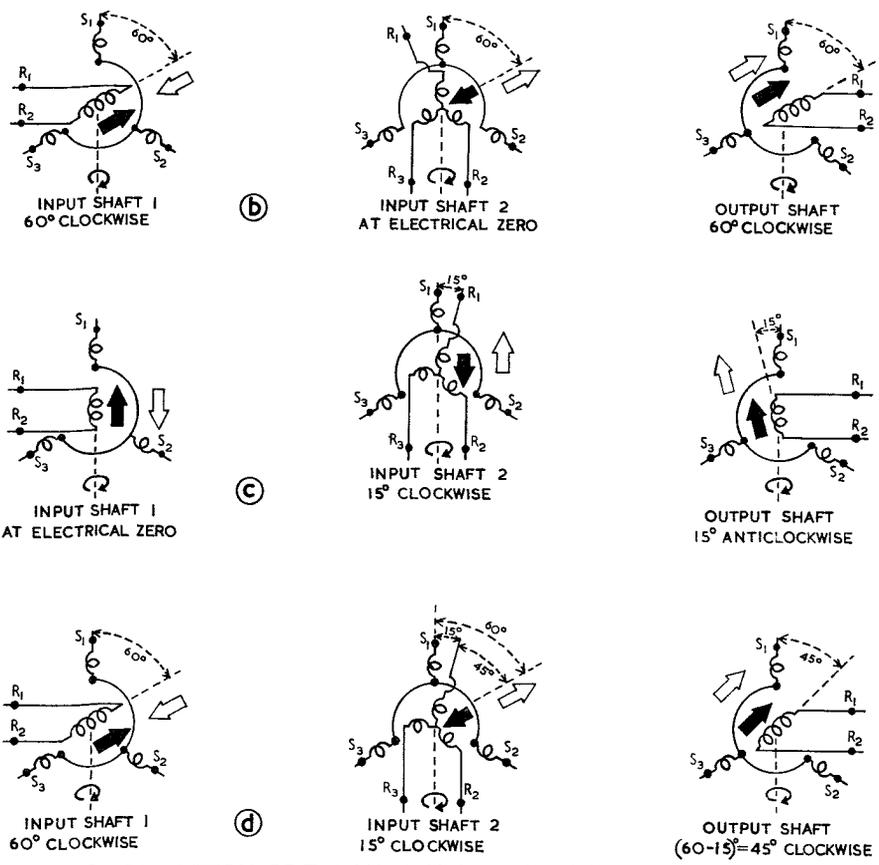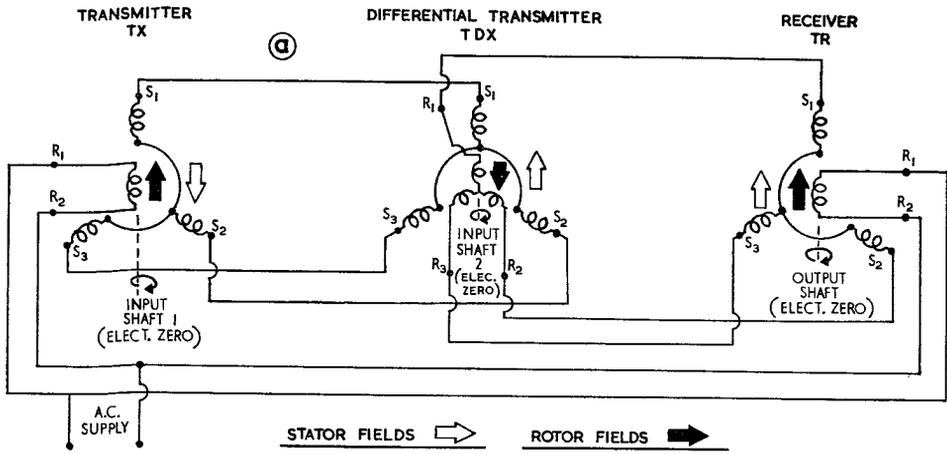
Fig. 24. ACTION OF TORQUE DIFFERENTIAL TRANSMITTER SYSTEM.

differential *receiver* in conjunction with two synchro transmitters as shown in Fig. 26.

Voltages indicating a 60° clockwise rotation from electrical zero are applied from transmitter A to the stator windings of the differential receiver, and a magnetic field Φ_1 is created along an axis 60° clockwise from electrical zero. The 15° clockwise electrical signal from transmitter B is applied to the rotor windings of the differential receiver and establishes a magnetic field Φ_2 that is 15° clockwise from the rotor electrical
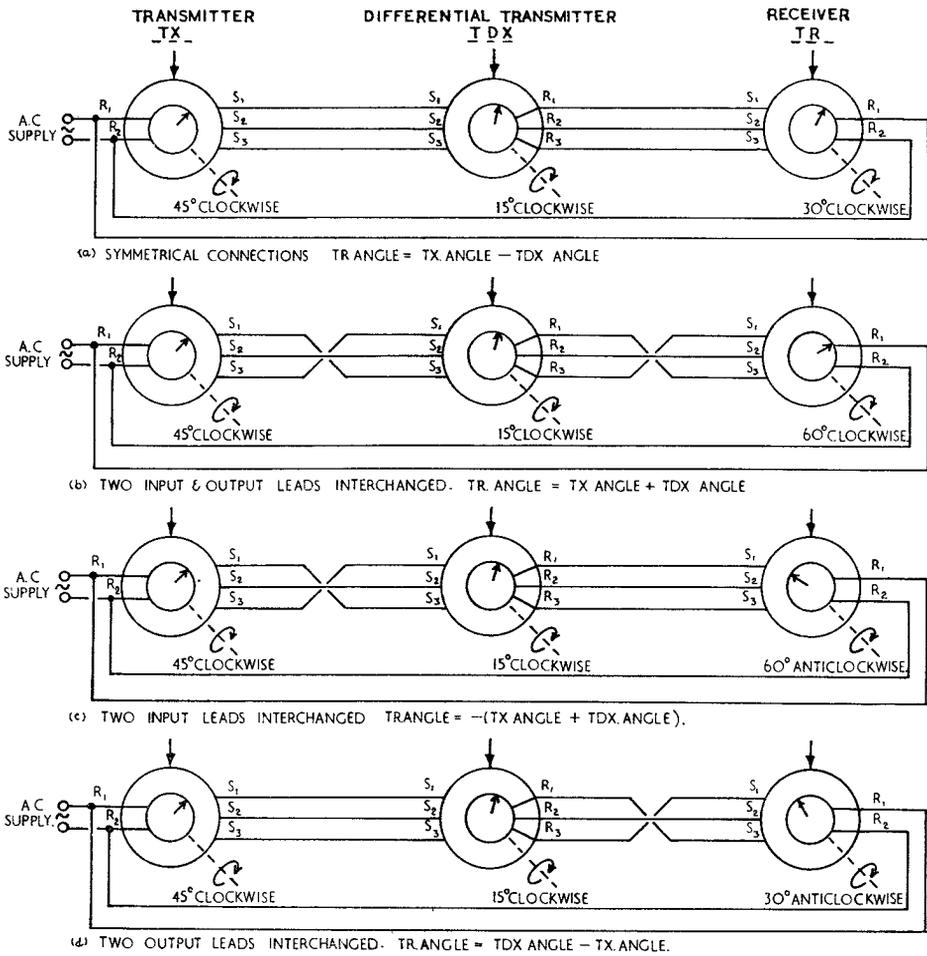
Fig. 25. VARIOUS INTERCONNECTIONS IN TORQUE DIFFERENTIAL SYSTEM.

zero point (R_1 in line with S_1). The differential rotor, if free to turn assumes the position in which Φ_1 and Φ_2 are aligned: this requires a clockwise movement of only 45°. Thus, the output shaft indicates the *difference* (60° clockwise minus 15° clockwise) in the angular displacement of the two input shafts connected to the transmitter rotors.

For the connections shown, the output angle is the angle of transmitter A *minus* the angle of transmitter B. Reversing pairs of connections at the differential receiver can change the relative directions of motion in much the same way as illustrated in Fig. 25 for the differential transmitter.

Control Synchro Systems

45. Introduction. In a torque synchro system, the output element exerts a torque which tends to align its rotor with the angle of the input shaft. The action is similar in torque differential synchro systems.

In *control* synchro systems, however, the rotor of the output element does not exert any such torque. Instead it produces a voltage, sometimes called an *error signal*, which indicates the error of alignment between the input shaft and the output shaft: this has important practical applications in electrical servomechanisms.
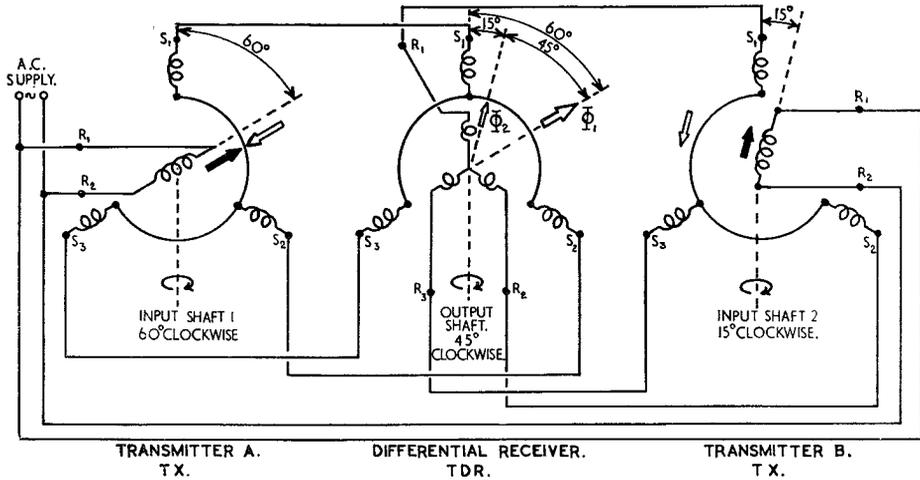
Fig. 26. TORQUE DIFFERENTIAL RECEIVER (TDR) SYSTEM.

46. Circuit. The basic control synchro system has two elements—a synchro transmitter (CX) and a synchro control transformer (CT) connected as shown in Fig. 27. The transmitter is similar to that used in torque systems: it consists of three stator coils spaced 120° apart, inside which a single-winding, two-pole rotor, energised from the a.c. supply, can be rotated by the input shaft.

The control transformer has a stator similar in design and appearance to that of other synchro units, but with high impedance coils to limit the alternating currents through the windings. The rotor, like that of a normal synchro receiver, carries a single winding which is brought out via slip rings and brushes, to terminals R_1 and R_2. Unlike the receiver rotor, the winding of the control transformer rotor is wound on a *cylindrical* former, thereby ensuring that the rotor is not subjected to any torque when the magnetic field of the transformer stator is displaced. In addition, the rotor of the control transformer is *not energised:* it acts merely as an inductive 'error detector.'

47. Action. The control transformer operates as a single-phase transformer having three stationary primary windings and one movable secondary winding.

When the rotor of the synchro transmitter (CX) is energised, voltages are induced in the transmitter stator windings and applied to the stator windings of the control transformer. The resultant magnetic fields produced when the input shaft is in such a position that the transmitter rotor is at electrical zero are illustrated at (a) of Fig. 28. The alternating stator flux of the control
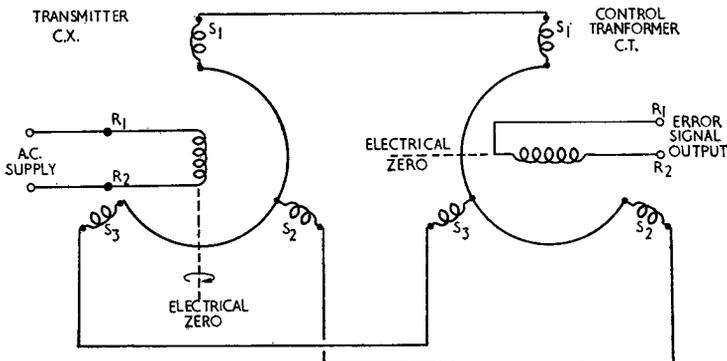
Fig. 27. CONTROL SYNCHRO SYSTEM.
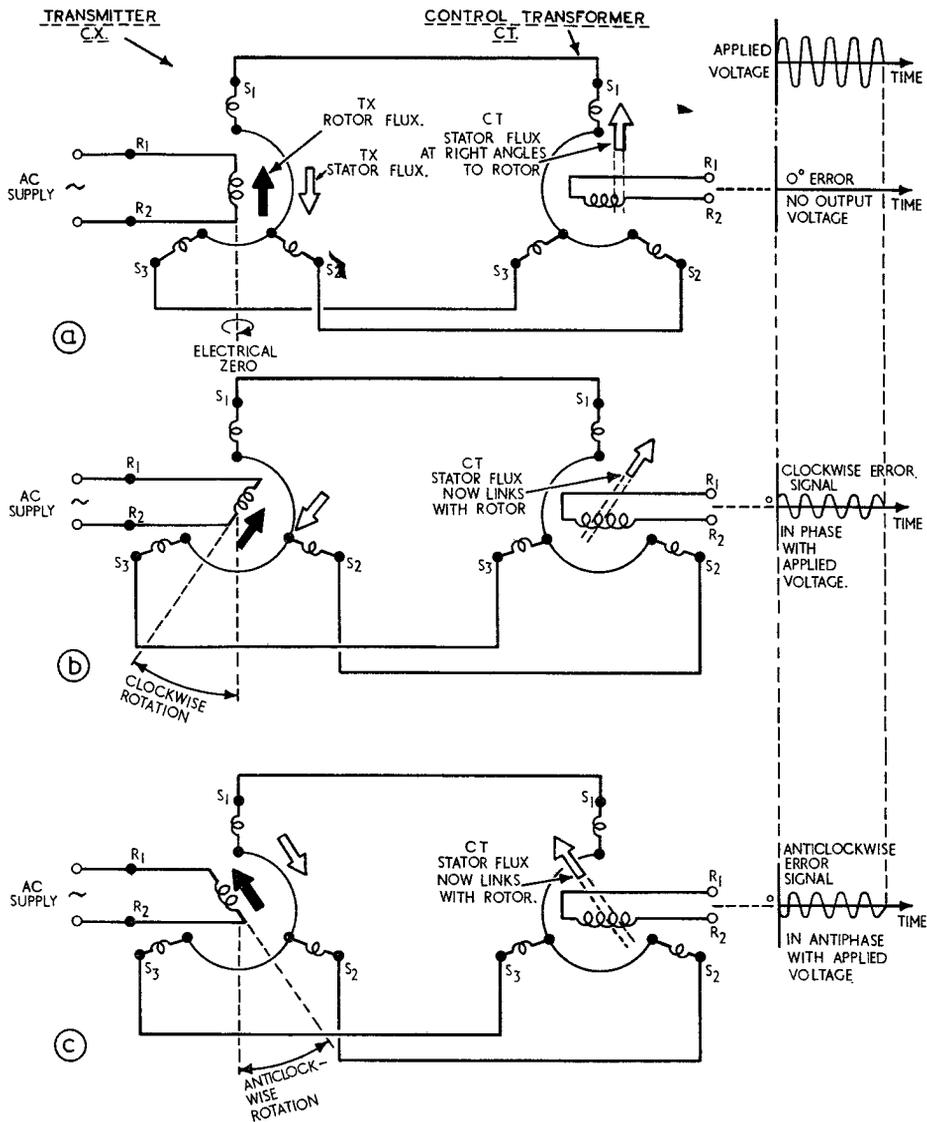
Fig. 28. ACTION OF CONTROL SYNCHRO SYSTEM.

transformer induces a voltage in its rotor, the magnitude of which depends upon the position of the rotor relative to the flux: that is, when the rotor axis is at 90° to the flux direction as shown at (a), the induced voltage is zero. Thus note that the electrical zero point of a control transformer is at 90° to the zero points of a synchro transmitter and receiver.

48. If now the input shaft is rotated clockwise from the electrical zero position, the resultant flux in the control transformer stator is displaced from its datum point by the same angle, and the magnitude of the voltage induced in the transformer rotor increases from zero. For the connections shown, the voltage is also *in phase* with the line voltage applied to the transmitter rotor (Fig. 28(b)).

For an anti-clockwise rotation of the input shaft from the electrical zero position, the transformer rotor voltage again increases in magnitude, but this time it is in *anti-phase*
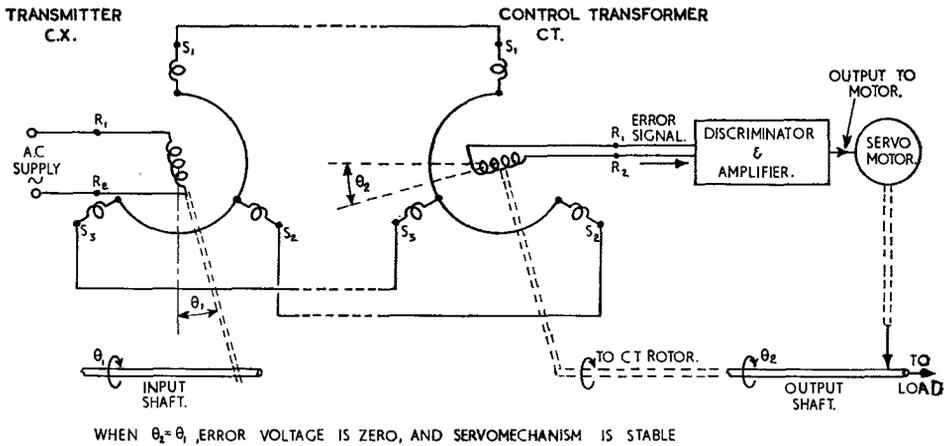
Fig. 29. APPLICATION OF CONTROL SYNCHRO SYSTEM

with the line voltage applied to the transmitter rotor (Fig. 28(c)).

49. **Use.** The error voltage derived from the control transformer rotor varies with the misalignment between the input shaft and the rotor of the transformer (remembering that, in any case, the electrical zero points are displaced from each other by 90°). When the two are 'aligned', there is no error voltage: a misalignment in one sense provides an in-phase error voltage: a misalignment in the other direction produces an anti-phase error voltage: the magnitude of the error voltage in each case depends on the degree of misalignment.

As commonly used in electrical servomechanisms, the synchro control transformer supplies an error signal from its rotor winding to an amplifier that controls a d.c. or a.c. motor. The circuit (Fig. 29) is such that the speed of the motor is proportional to the magnitude of the error voltage, and the direction of rotation is determined by the phase of the error voltage with respect to the applied a.c.

In normal operation, the servo motor drives the mechanism being controlled (e.g. a radar scanner) and also turns the rotor of the control transformer. The circuit arrangement is such that the transformer rotor is turned into alignment with the input shaft, thereby reducing the error voltage to zero, at which point the servomechanism is stable.

The use of an amplifier and rotor makes the
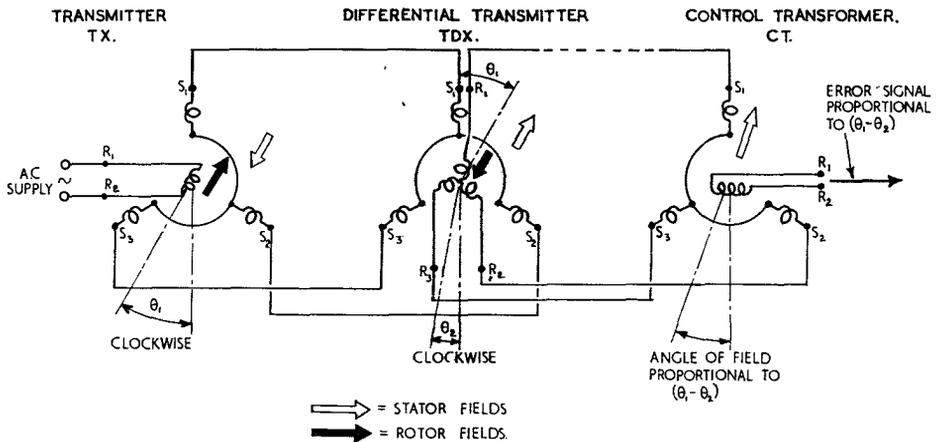
Fig. 30. CONTROL DIFFERENTIAL SYNCHRO SYSTEM.

system *torque amplifying* and the output of such a system depends solely upon the power output of the amplifier and servo motor. By means of control synchros, very small units and light controlling forces can operate heavy mechanisms remote from the control point.

50. In the same way that differential synchros can be used in torque transmission systems, so they can be used in control synchro systems to transmit information on the sum or difference of two angles. A simple arrangement is illustrated in Fig. 30.
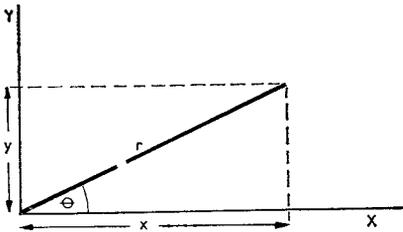
Fig. 31. CO-ORDINATES OF A POINT.

Resolver Synchro System

51. **Introduction.** A vector representing an alternating voltage (Fig. 31) can be defined in terms of the *modulus* or length of the vector r, and the *argument* or angle θ it makes to the X axis: these are the polar co-ordinates $r \diagup \theta$ of a vector. This same vector can be defined in terms of x and y where $x = r \cos \theta$ and $y = r \sin \theta$: these expressions give the cartesian co-ordinates of the vector.

52. Resolver synchros are employed, generally in analogue computers (see Sect. 20), to convert voltages which represent the cartesian co-ordinates of a point, into a shaft position and a voltage which together represent the polar co-ordinates of that point. They are also used in the reverse manner for conversion from polar to cartesian co-ordinates.

53. **Construction of resolver synchros.** Outwardly, a resolver synchro looks like all the other synchros already dealt with. It has however, four stator and four rotor windings, arranged as shown in Fig. 32. Stator windings S_1 and S_2 are in series and have a common axis which is at right angles to that formed by S_3 and S_4 in series: similarly, rotor windings R_1 and R_2 in series have a common axis which is at right angles to that formed by R_3 and R_4 in series.

54. **Resolution from polar to cartesian co-ordinates.** In Fig. 33 an alternating voltage of magnitude r applied to one of the rotor windings represents the modulus of polar co-ordinate, and the angle θ through which the rotor shaft has turned represents the argument. In this application, only one of the rotor windings is used and the unused winding is normally short-circuited to improve the accuracy, and limit spurious response.

The alternating flux produced by the rotor current links with the stator windings, and voltages are induced in each of the stators. In the position shown in Fig. 33, maximum voltage is induced across that stator coil which is aligned with the rotor in use,
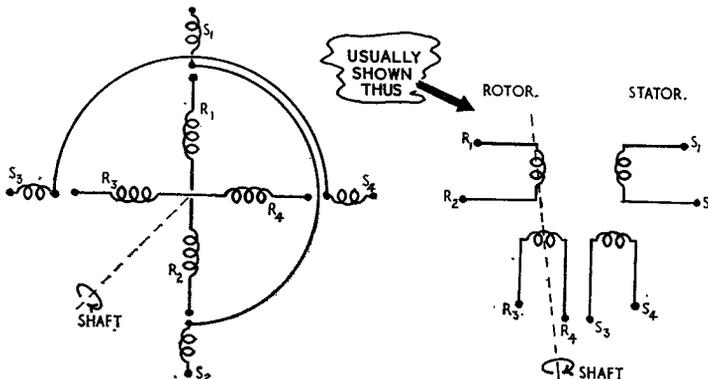
Fig. 32. RESOLVER SYNCHRO.
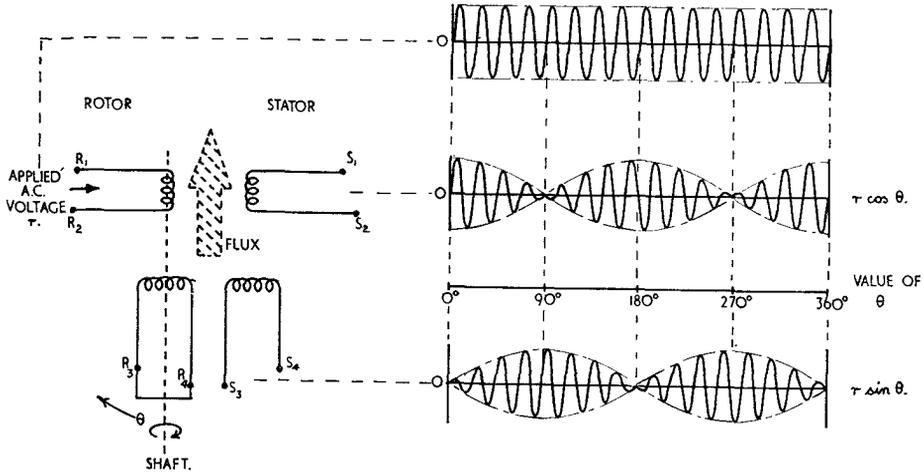
Fig. 33. CONVERSION FROM POLAR TO CARTESIAN CO-ORDINATES.

i.e., $S_1 S_2$ in line with $R_1 R_2$. No voltage is induced in the other stator coil which is at right angles to the rotor flux. Movement of the rotor at a constant speed will induce voltages across the two stator coils which will vary sinusoidally.

The voltage across that stator coil which is aligned with the rotor at electrical zero will be a maximum at that position and will fall to zero after 90° displacement: this voltage is therefore a measure of the *cosine* of the angle of displacement (cos θ). It is *in phase* with the energising voltage during the first 90° displacement and *in anti-phase* from 90° to 270°, finally rising from zero at 270° to maximum in-phase at 360°. Any angle of displacement can therefore be identified by the amplitude and phase of the induced stator voltages.

Similarly, the stator coil which at datum is at right angles to the energised rotor

coil will at that point have zero voltage induced in it. Through a displacement of 90°, this voltage will rise to maximum *in-phase* sinusoidally and is therefore directly proportional to the sine of the displacement angle (sin θ). Again, the phase depends on the angle of displacement and any angle can be identified by the amplitude and phase of the induced stator voltages.

The output from one stator is of the form r cos θ and from the other r sin θ: the sum of these two defines in cartesian co-ordinates the input voltage and shaft rotation $r \angle \theta$.

55. Resolution from cartesian to polar co-ordinates. To convert from cartesian to polar co-ordinates a zero-nulling device is required. One arrangement is illustrated in the circuit of Fig. 34. An alternating voltage $V_x = r \cos \theta$ is applied to the cos

STATOR.   ROTOR.
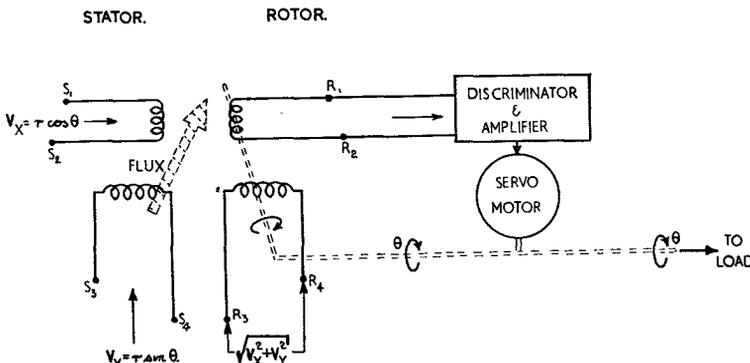
Fig. 34. CONVERSION FROM CARTESIAN TO POLAR CO-ORDINATES.
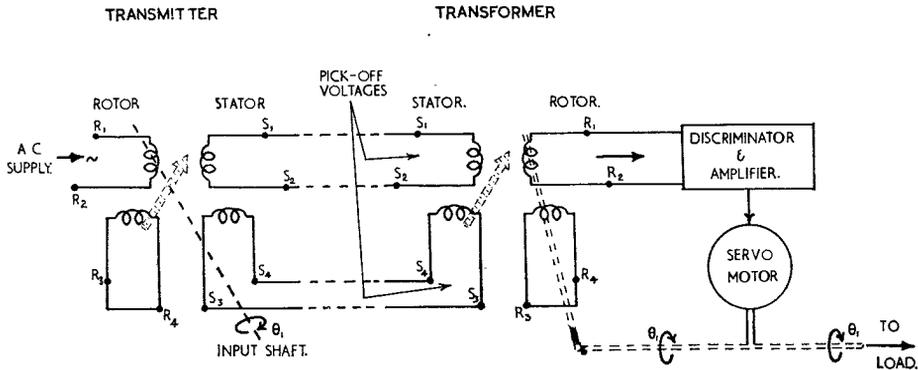
Fig. 35. RESOLVER SYNCHRO AS A REMOTE INDICATOR.

stator winding S_1 S_2, and $V_y = r \sin \theta$ is applied to the sin stator winding S_3 S_4. An alternating flux of amplitude and direction dependent upon these voltages representing the cartesian co-ordinates, is therefore produced inside the stator.

One of the rotor windings R_1 R_2 is connected to an amplifier and servo motor which drives the output load and also the rotor in such a direction as to return the rotor to a null position: the motor then stops.

The other rotor winding R_3 R_4 has induced in it a voltage porportional to the amplitude of the alternating flux, i.e., proportional to $\sqrt{V_x^2 + V_y^2}$. This voltage represents the modulus r. The shaft position of the rotors represents the argument θ. Thus the input defined in cartesian co-ordinates has been converted to an output in terms of the polar co-ordinates.

56. Remote indication by resolver synchros.

In some instances it is more convenient to have positional information transmitted in cartesian co-ordinates. The information is then readily available for application to the horizontal and vertical plates of a c.r.t., or for modification by other computer elements. Such instances occur, for example, when transmitting the position of a radar scanner. A typical arrangement is illustrated in Fig. 35. It will be seen that this application is very similar to that of a *control* synchro system with the added advantage that voltages corresponding to the cartesian co-ordinates can be picked off the resolver transformer stator windings.

The voltages induced in the transmitter stator windings from the energised rotor are transmitted via the connecting leads to the stator coils of the transformer, and the resulting magnetic flux of the transformer stator lines up with that of the transmitter rotor. By normal transformer action a voltage is induced in the transformer rotor windings and the output of one of them is fed to an amplifier which, in turn, controls a servo motor. The motor drives the load and at the same time turns the rotor to the null position, i.e., at right angles to the axis of the stator magnetic field. Thus, note that the electrical zero of a resolver transformer, like that of a control transformer, is displaced 90° from that in other synchros. When the rotor is in the null position, the servo motor stops, having turned the output shaft through the same angle as the input shaft. Transmission of position has therefore been achieved.

57. Resolver synchro as a phase-shift device.

A resolver synchro can be used in conjunction with a resistor and a capacitor connected in series across both stator windings, as shown in Fig. 36: this arrangement gives a phase-shifting device. If the rotor is energised, a voltage can be obtained between

Fig. 36. RESOLVER SYNCHRO AS A PHASE-SHIFT DEVICE.

TRANSMITTER.   TRANSFORMER.
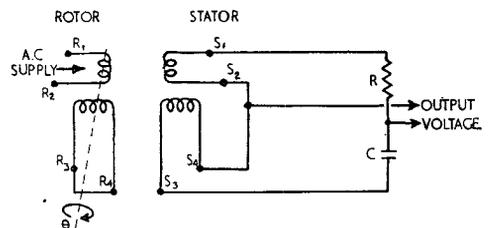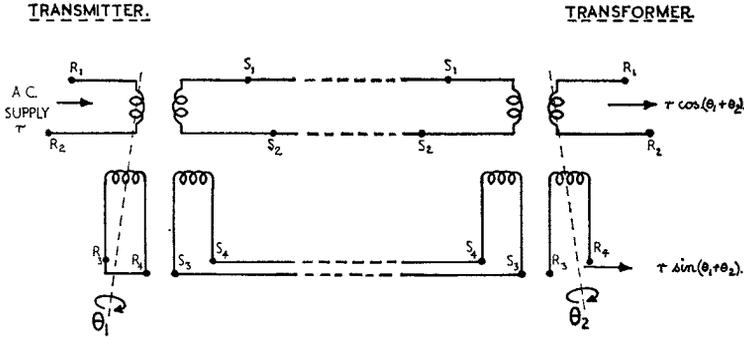
Fig. 37. RESOLVER DIFFERENTIAL SYNCHRO SYSTEM.

the common point of both stator windings and a point between the resistor and the capacitor. The phase of this output voltage relative to the input depends on the orientation of the rotor relative to the stator windings: the phase can in fact be varied through 360° by turning the rotor through one complete revolution.

58. **Differential resolution.** It is sometimes necessary to obtain the sine and cosine values of the sum or difference of two inputs. A resolver synchro system arranged to give this is illustrated in Fig. 37.

One input shaft, connected to the transmitter rotors, turns through the angle θ_1: the other input shaft, connected to the transformer rotors, turns *in the same direction* through the angle θ_2. Since the axes of the two rotor windings on the transformer are at right angles to each other, one will

give a cos output and the other a sin output. The magnitude and phase of the voltage induced in each output rotor depends on the orientation of each set of rotors relative to their respective set of stator windings, i.e., on the angles through which the input shafts have turned. With the connections shown, the outputs are $r \cos (\theta_1 + \theta_2)$ and $r \sin (\theta_1 + \theta_2)$. A re-arrangement of the connections between transmitter and transformer will give the *difference* of two angles.

59. An alternative arrangement that gives similar results is shown in Fig. 38. In this circuit, a synchro known as a *resolver differential synchro* is used in conjunction with a normal synchro transmitter. In the resolver differential synchro there are two stator windings at right angles to each other but the rotor is a three space-phased winding. The differential rotor produces a magnetic
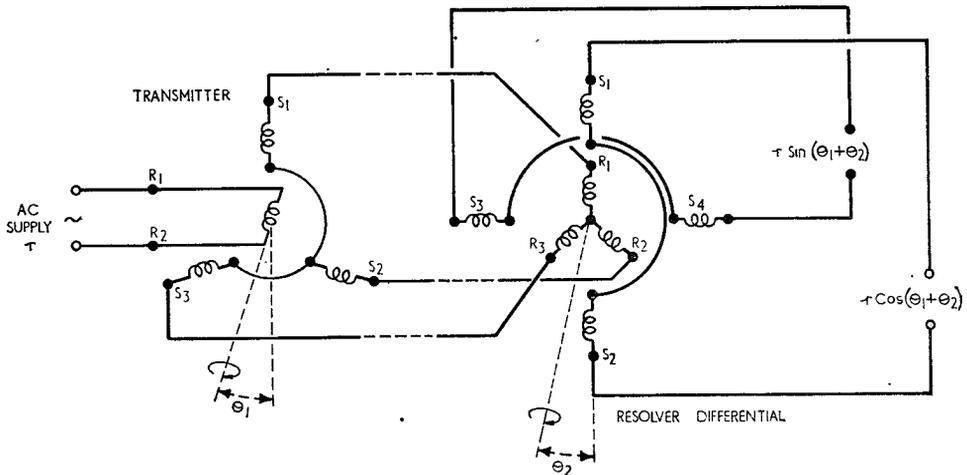
Fig. 38. RESOLVER DIFFERENTIAL SYNCHRO.

field in accordance with the electrical signals received from the transmitter.

Due to the normal differential action described earlier, the amplitude and phase of the voltage induced in each stator winding of the differential depends on the relative directions of the stator windings and the rotor flux. The stator windings are arranged with their axes at right angles to give cos and sin outputs, and the position of the rotor flux is determined by the angles of the two input shafts − θ_1 to the transmitter

Systems	Remarks
D.C. SYSTEMS	
Desynn	Provides only sufficient torque to operate small instruments: gives remote indication of dial readings to an accuracy of about ±2°.
M-type	Provides moderate torque, sufficient to drive small mechanisms: accurate to about ±2°. Typical use is to rotate the scanning coils in a c.r.t. in synchronism with a radar aerial.
Wheatstone bridge	An error-operated system, accurate to within ±1°. Does not provide continuous rotation and gives very little torque: can be used as the controlling element in a torque-amplifying system, e.g., remote tuning of radio equipment.
A.C. SYSTEMS	
Torque synchro	Provides only sufficient torque to operate small instruments: efficient and accurate to within ±1°: often used to transmit data such as radar bearings to the place where the information is required.
Torque differential synchro ..	As for the torque synchro, but provides summation of two input shaft angles: used, for example, to combine a D/F loop reading and a compass reading to give a true bearing.
Control synchro	Gives an electrical output that is dependent on the error in alignment between driving shaft and load shaft. The error signal is normally used as the input to a control system driving a heavy load.
Control differential synchro..	As for control synchro, but provides summation of two input shaft angles.
Resolver synchro	Used in computers to give either cartesian or polar co-ordinates of an input, and for conversation of one to the other: can also be used in a manner similar to that of a control synchro.
Resolver differential synchro	Gives an electrical output in the form of sine and cosine values of the sum or difference of two inputs.

TABLE 2—SUMMARY OF REMOTE INDICATION SYSTEMS

rotor and θ_2 to the differential rotor. Thus, outputs of r cos $(\theta_1 + \theta_2)$ and r sin $(\theta_1 + \theta_2)$ or of r cos $(\theta_1 - \theta_2)$ and r sin $(\theta_1 - \theta_2)$ are obtained, depending on the connections between the synchros.

Summary

60. This chapter has considered electrical remote indication systems in a general way. It has shown how changes in a physical quantity at one point, as represented by rotation of an input shaft, can be accurately reproduced at a remote point. It has also shown how transmission devices can be incorporated into control systems to give accurate remote control of heavy loads from small units and light operating forces. These control systems will be considered in greater detail in the next chapter (servo-mechanisms). Table 2 lists the devices and systems discussed in this Chapter.

SECTION 19

CHAPTER 2

SERVOMECHANISMS

SERVOMECHANISMS

Introduction

1. The discovery that heat (from coal or oil) can be converted into mechanical energy brought about the industrial revolution, and machines which could use this energy to produce useful results were quickly invented and improved. By controlling such machines, man was able to release large quantities of energy with very little expenditure of energy on his own part.

At first, machines were simple and a human being was quite capable of controlling in detail the various operations that went to complete any process. But as time has passed, machines and processes have become more complicated and results have had to be produced more quickly and more accurately. Thus it has come about that in many cases, man has proved to be an imperfect controller of the machines he has created. It is natural therefore that, wherever possible, the human controller should be replaced by some form of automatic controller.

2. Automatic control systems can include electronic, electro-mechanical, pneumatic, hydraulic and mechanical devices. Such devices are used to perform diverse functions: for example, the automatic piloting of an aircraft, the control of a guided missile, the movement of a radar aerial, keeping a telescope trained on a star, and so on. Nevertheless, regardless of the nature of the quantities handled, the resulting arrangements have a strong family likeness to each other and behave in very similar ways. A common theory is therefore applicable to all forms of automatic control.

The title of this chapter is 'Servomechanisms'. A servomechanism, in fact, is merely a particular type of automatic control system whose output is the position of a shaft. It is however the most common type of control system in radio engineering, and since the theory which has been developed in its design is now used for all types of control systems, it is convenient to confine the discussion to servomechanisms.

3. A complete treatment of servomechanisms is far beyond the scope of these Notes.

Only an elementary outline of the basic principles involved and a general idea of the purpose and applications of control systems can be attempted in this chapter. Further information is given in Part 3 of these Notes.

4. Chapter 1 has shown that d.c. remote indication and a.c. synchro systems can operate between shafts separated by a considerable distance, but cannot supply torque amplification: the torque delivered to the load can never exceed the input torque. For this reason, and because the error increases when large torques are transmitted, remote indication and synchro systems are employed to turn dials and pointers, to move control valves or to actuate similar low-torque loads.

Automatic control systems, on the other hand, can supply the large torques required to move heavy loads, *and only a very small torque need be applied to the input shaft.* Remote operation is not inherent in servomechanisms but it can be obtained if synchro devices are made part of the system.

5. From any process there is an end-product which can be called the *output*. The production of the output depends on the process, and how the process is affected by the *input*. A control system acts in such a way that the output can be controlled in the optimum manner to give a desired result which bears a definite relationship to the input to the system.

Speed Control of a D.C. Motor

6. Fig. 1 shows an arrangement (known as the 'Ward-Leonard' system) that is used for controlling the speed of a d.c. motor driving a load.

The motor armature current is supplied by a d.c. generator which, in turn, is driven at constant speed. The d.c. motor is separately excited, the field current being held constant. The generator is also separately excited, but its field current can be varied by adjusting the controller potentiometer. Any variation in the generator field current varies the generator output
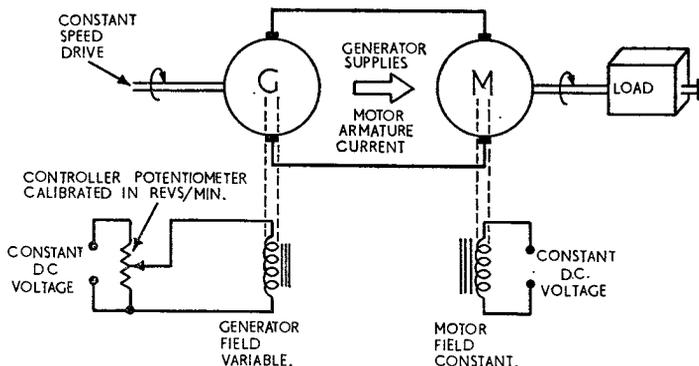
Fig. I. SIMPLE SPEED CONTROL OF A D.C. MOTOR.

voltage and hence the armature current to the motor: thus the *speed* of the motor is varied. If the generator field current is increased, the generated voltage increases, as does the armature current and hence the motor speed. Therefore the controller potentiometer could be calibrated with a scale in revolutions per minute and set for whatever motor speed is required.

With this arrangement, the speed of the output shaft represents the 'output': the 'input' is the setting of the potentiometer. The input can therefore be set to give the desired output which the system should hold constant.

7. In practice, however, speed is not held exactly constant even though ideally the speed of a separately-excited motor is determined by the voltage applied to its armature. Variations in speed arise from a variety of reasons. In particular, variations of *load* conditions will cause varying motor speeds and the output is no longer that demanded by the input.

This system is not good enough if speed control to within a fraction of one per cent is required.

Action of Human Operator

8. It is interesting to discuss at this point the actions that a human operator would take to maintain a constant speed (Fig. 2).

The first action of a human operator is to collect the information or data on which he is to act. He has in mind a picture of the output speed required and at the same time he notes the *actual* output speed. His sole function is to compare the two impressions and so to adjust the system as to reduce the difference, or error, between them: he does this of course by adjusting the controller potentiometer. He is thus, in this connection, primarily an *error-measuring device*, and the amount of error determines how he causes the motor to use energy from the generator to produce the required output speed. Note that the human operator provides a *feedback* link between output and input.

An Improved Method of Speed Control

9. In practice, a more effective and efficient control of output speed can be obtained by replacing the human operator with an automatic control system as shown by the arrange-
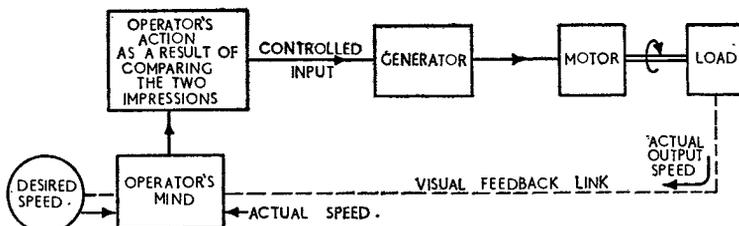
Fig. 2. HOW A HUMAN OPERATOR CONTROLS MOTOR SPEED.
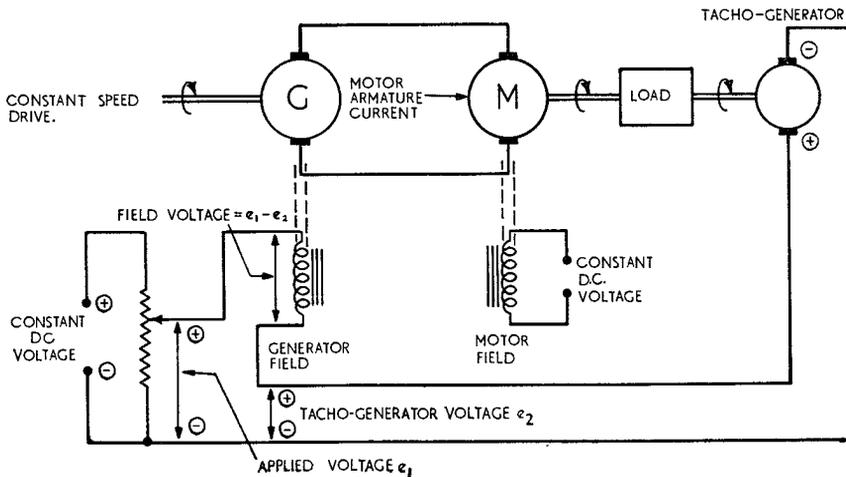
Fig. 3. AN IMPROVED METHOD OF SPEED CONTROL.

ment of Fig. 3. The response of the automatic system is better than that of a human operator and the automatic arrangement is not subject to fatigue.

10. **Tachometer-generator.** In Fig. 3, the actual motor speed is measured by connecting a device known as a tachometer-generator or tacho-generator to the output shaft: this produces a voltage proportional to the speed at which it is driven, i.e. the actual output speed. The tacho-generator in the circuit of Fig. 3 is a separately-excited d.c. generator with a constant field and it is so constructed that it produces a generated e.m.f. which is exactly proportional to the rotational speed of its armature. On load, the terminal voltage is still proportional to speed, provided the load current is small enough for armature reaction to have no effect. D.C. tacho-generators usually have quite a small maximum-load current, because the armature is wound with many turns of fine wire, and the commutator has a large number of segments to ensure a relatively smooth d.c. output.

11. The output voltage from the tacho-generator representing the actual speed of the load is compared with the voltage across the controller potentiometer representing the demanded speed: the *difference* between these two voltages causes the flow of generator field current. The connections are such that if the load speed is less than that demanded, the opposition voltage produced

by the tacho-generator decreases and the generator field current increases; the generated voltage thus increases as does the motor armature current, and the motor speeds up until it reaches the demanded speed. It follows that if for any reason the load speed tends to change from that demanded, the correct restoring action will automatically be taken.

12. The operations of the circuits shown in Fig. 1 and Fig. 3 differ considerably in detail.

In Fig. 1 the output depends primarily on the input demand, but the accuracy of control is limited because there are no means of controlling other factors that affect the output (such as variation in output load). The accuracy of control therefore depends on the linearity of the system. Such systems are referred to as *open-loop* control systems. An open loop control system is characterized by the lack of error comparison; that is, there is *no feedback* of information from output to input. Because of their limited accuracy, open loop systems are hardly ever used.

In Fig. 3, there is feedback of information from output to input so that the input demand and the output can be compared. The feedback is in opposition to the input and tends to reduce the net input to the system as the output follows the input demand more closely: it is therefore, *negative* feedback. The system is automatically adjusted such as to reduce the error between

the input demand and the output: it is therefore an *error-actuated* device and is referred to as a *closed-loop* control system: such systems are the only means of obtaining accurate and predictable control of output.

Both the open-loop and the closed-loop systems discussed above are power amplifying; the energy expended in adjusting the controller to the desired output speed setting is only a small fraction of that expended in turning the load. The amplification comes of course by choosing a motor that is powerful enough to drive the load.

13. Note again the main features of a closed-loop control system: there is an input demand and an output; there is negative feedback from the output which is compared with the input demand; the resulting error is amplified and used to control the power into a servomotor; the servomotor turns the load in such a direction as to reduce the error and ensure that the output follows the input demand.

Position Control Systems

14. For the closed loop speed control system, the quantity fed back and compared with the input demand is a voltage proportional to the *speed* of the output shaft.

A more common type of control system of particular interest in radio engineering is one which controls the angular or transverse *position* of the output shaft. In closed loop position control systems therefore, the quantity to be fed back must be a measure of the output shaft *position*. One of the most convenient ways of providing this feedback is to produce a voltage that is proportional to the position of the shaft at any instant. This can be done by some of the data transmission devices discussed in Chapter 1. However, one of the easiest ways in d.c. systems is to use a potentiometer as shown in Fig. 4. This method will be assumed in subsequent paragraphs.

15. Consider Fig. 5 which shows in block form, the essential elements in a closed loop system for position control. Comparison with Fig. 3 shows that the controller and amplifier are together equivalent to the potentiometer and generator in the Ward-Leonard system: the servomotor is equivalent to the d.c. motor: and the output potentiometer replaces the tacho-generator because here *position* is being measured.

The input is the angular position of the input shaft and the output is the angular position of the output shaft connected to the load. The requirement is that the position of the output shaft should conform with the input demand, i.e. that the output shaft follows precisely the movements of the input shaft.

The input shaft controls a potentiometer that provides a voltage proportional to the angle θ_i of the input shaft. The output shaft similarly controls a potentiometer that provides a voltage proportional to the angle θ_0 of the output shaft. The voltage proportional to θ_0 is fed back to an error-measuring device where it is compared with the voltage proportional to θ_i. The feedback is such that it is in *opposition* to the input voltage (i.e., negative feedback) and the output from the error-measuring device is an error voltage e proportional to the *difference* between θ_i and θ_0: that is, $e = \theta_i - \theta_0$, and it can be *positive* or *negative*.

In Each Case, The Potential Of 'A' Relative To Earth Gives A Measure Of The Shaft Position.
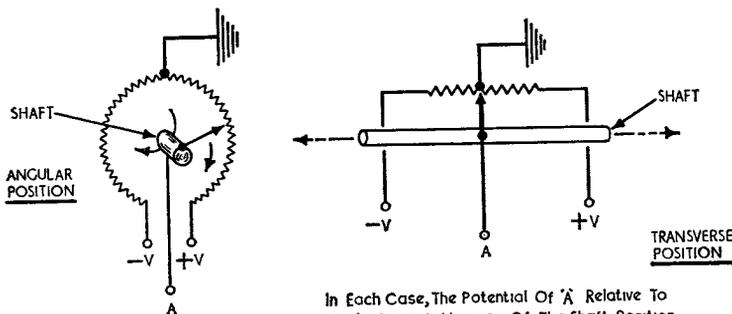
Fig. 4. PRODUCTION OF VOLTAGE PROPORTIONAL TO SHAFT POSITION.
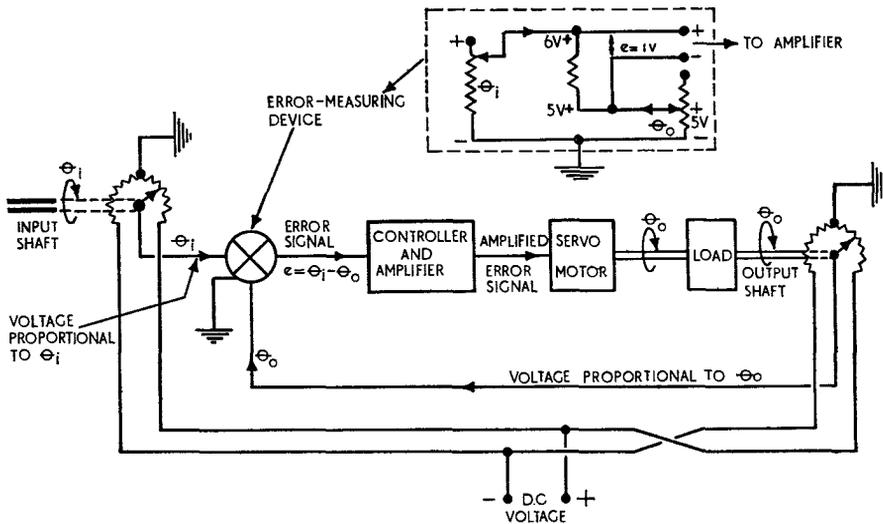
Fig. 5. CLOSED LOOP POSITION CONTROL SYSTEM.

16. This error signal is amplified and applied to the motor which then turns the load in a direction depending on the sense of the error signal. The direction of rotation is always such as to tend to reduce the error voltage to zero; that is, to drive the output shaft into alignment with the input shaft. When the voltage proportional to θ_0 equals that due to θ_i, the error signal is zero: the motor stops at this point, with input and output shafts aligned.

17. This particular type of closed loop automatic control system defines a true *servomechanism*—an error-actuated, power-amplifying, position control system. For a servomechanism to fulfil its function it must have 'follow-up' properties, i.e. the output must be capable of following random variations of input demand over a very wide range. Note again that the final net input voltage is an error voltage, and *not* the simple voltage proportional to the input demand θ_i: this is the first improvement of a servomechanism over an open loop system.

A servomechanism has many applications. It is used, for example to make a searchlight follow its sighting mechanism, to rotate a radar scanner to a desired position, to control aerodynamic surfaces in aircraft and missiles, and so on.

Behaviour of a Simple Servomechanism

18. Consider Fig. 6 which shows in block form the elements in a simple d.c. servomechanism. It is assumed that the output shaft is driving a load, such as a radar scanner, and that it has taken up a position which agrees with the position demanded by the input shaft, i.e., $\theta_0 = \theta_i$. The resultant
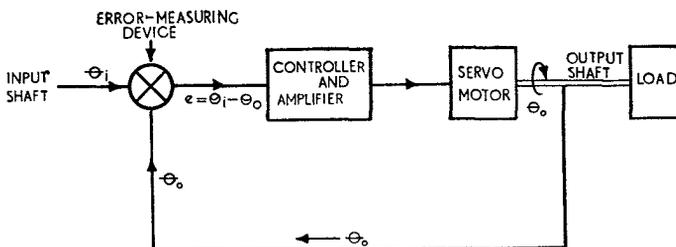
Fig. 6. ELEMENTS IN A SIMPLE D.C. SERVOMECHANISM.

error signal ($e = \theta_i - \theta_0$) is zero and the motor is stationary in a steady state condition.

Now suppose that the azimuth bearing of the radar scanner is to be changed from its initial angle to another angle; then the input demand is suddenly changed. The output shaft cannot immediately follow this change in demand because of the inertia of the load. There is therefore now a difference between θ_0 and θ_i and the resulting error signal, after amplification, causes the motor to accelerate in an attempt to bring the output shaft to the new demanded position.

As θ_0 approaches alignment with θ_i, the error signal and the motor acceleration are progressively reduced until the condition is reached where θ_0 again equals θ_i and the error signal is zero: the motor then stops. This is the stable condition of the servomechanism, with the output shaft in the position required by the input demand. It obviously takes time.

19. The period during which the output is changing in response to the change of demand is called the *transient period*. When this period has been completed, the system is said to have reached a *steady state*. The time taken by the system to reach a new steady state after a change of demand (i.e., the time occupied by the transient period) is called the *response time* or the *time lag* of the servomechanism.

Response and Stability of Servomechanisms

20. The change in the value of θ_i if the input demand changes instantaneously from one fixed value to another fixed value in a remote position control system can be represented by a '*step input*' as shown by the graph of Fig. 7(a).

As explained in the preceding paragraphs, initially the system is at rest with $\theta_0 = \theta_i$, and at a θ_i suddenly changes to a new value: θ_0 cannot follow immediately and the error therefore, increases from zero to θ_i (Fig. 7(c) and a large torque is applied to the load. As the load accelerates and θ_0 increases, so the error and torque are reduced until at b, θ_0 reaches the required value and they become zero.

However, unless special precautions are taken, a servomechanism will oscillate readily. Thus, by the time θ_0 reaches the required value at b, the load has acquired considerable momentum and consequently overshoots. The error now increases in the *opposite* sense and a reverse torque is applied which eventually brings the load to rest at c, and then accelerates it back again until once more it passes through the required position at d. But again it has acquired kinetic energy in the period c to d and another overshoot occurs at d.

21. This process can continue indefinitely if the frictional losses in the system are negligible, and the system oscillates continuously, being unstable and useless: it is said to '*hunt*'.

Where there are frictional losses, a damped train of oscillations results as shown in Fig. 7(b): in this case, the output shaft oscillates several times about its new position before coming to rest.
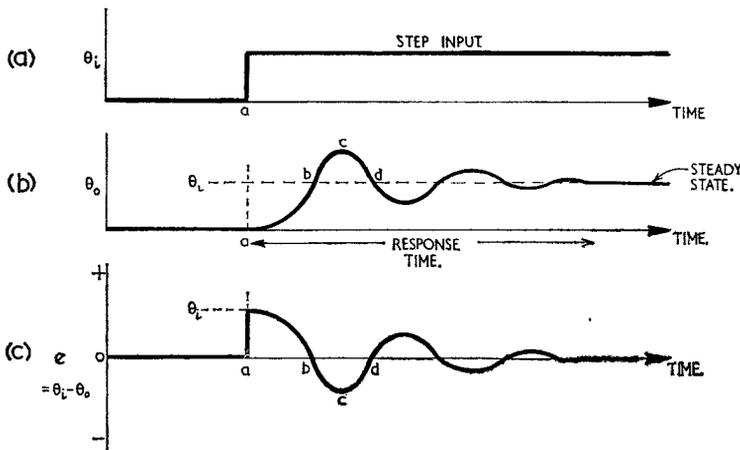
Fig. 7. RESPONSE OF A SERVOMECHANISM TO A STEP INPUT.

a = ———— = UNDERDAMPED.
b = ·· ··· = REQUIRED DAMPING.
c = —·—·— = CRITICAL DAMPING.
d = — — — = OVERDAMPED.
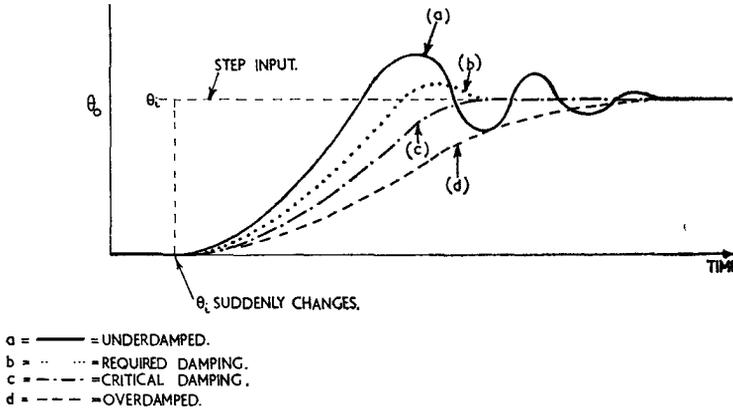
Fig. 8.   EFFECT OF DIFFERENT DEGREES OF DAMPING.

To avoid oscillations and subsequent hunting, friction or damping is necessary. As will be seen later, the effect of frictional damping can be given by electrical means; hence *damping* is a more general term than friction.

22.   Different degrees of damping produce different response curves.   Fig. 8 illustrates typical response curves: where there is overshooting and transient oscillation, as in (a), the system is *under-damped:* where there is no overshooting or oscillation, as in (d), the system is *over-damped: critical damping*, as in (c), marks the boundary between a non-oscillatory and an oscillatory response.

The response reaches its final value more quickly if there is under-damping: if there is too much under-damping, however, the response is oscillatory.   For these reasons, practical servomechanisms are designed to have *slightly less* than critical damping (about $0·75$ times critical damping): this is illustrated in curve (b) which shows only one overshoot.

Viscous Damping

23.   The main requirement in a remote position control (*r.p.c.*) system is that the output shaft should follow the input demand precisely and with minimum time lag.   It has been shown that for a rapid response time the damping of the system should be slightly less than critical damping.   The frictional losses inherent in a servomechanism produce some damping, but usually very much less than that required to ensure a rapid response time and a short settling-down period.   It is therefore necessary to introduce additional damping into the system to obtain the required transient performance.

24.   One obvious method is to increase the friction in the system by inserting some form of *brake* on the output shaft, as indicated in Fig. 9.   This can be achieved either by putting a mechanical friction plate device on the motor shaft or by causing a copper disc, mounted on the output shaft, to rotate between the poles of a horseshoe permanent magnet (eddy current damping).

With a suitable amount of such damping the required response can be obtained; it is necessary, however, that the amount of damping be accurately adjustable.

25.   Viscous friction damping, as this method is called, is not a good method and is used
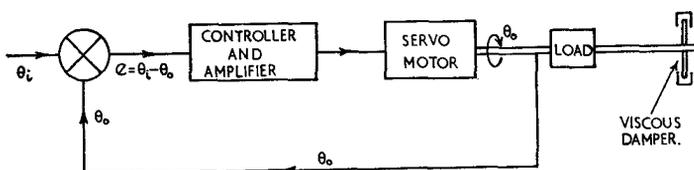
Fig. 9.   VISCOUS DAMPING.
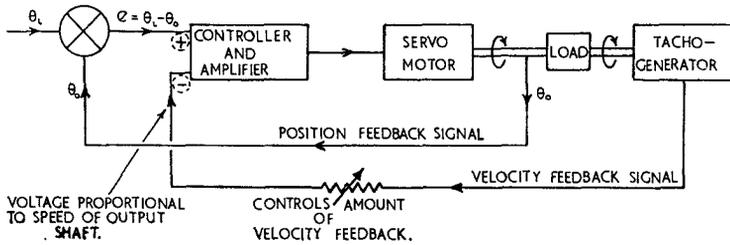
Fig. 10. ARRANGEMENT FOR VELOCITY FEEDBACK DAMPING.

only on very *small* servomechanisms. One obvious disadvantage is that it dissipates energy and therefore reduces the efficiency of the control system. Also, the energy absorbed by the viscous damper is dissipated in the form of heat, and in large servomechanisms elaborate and expensive cooling arrangements would have to be provided to get rid of the heat from the brake.

In fact any form of damping on the output side of the control system has serious disadvantages because of the much higher power levels at this end of the system. Because of this it is usual to insert the required damping on the input side of the servomechanism. In this case, the damping must be *electrical* in nature and takes the form of a modification to the error signal.

Velocity Feedback Damping

26. Since a r.p.c. servomechanism is self-correcting, it tends to remain quite stable when the input shaft is stationary. Thus, damping is required only during the *transient* period which follows a change in input demand. As previously explained, the instability is produced by the load's acquired momentum, resulting in an overshoot.

27. It is interesting to see how a human controller, faced with the same task of causing a motor to move a load from one position to another, is able to do so without causing instability or wasting energy. On receipt of his instructions, corresponding to a step input of position, the human controller will cause the driving motor to apply a torque accelerating the load. As the load gathers speed and approaches the required position, the controller anticipates that it will overshoot and therefore *reverses* the torque.

Under this condition, the load is driving against the motor and no energy is being dissipated in the load: there is therefore no

power loss such as is obtained with viscous damping.

If the controller is skilful, the result is that the load comes to rest just as it reaches the required position: overshooting with resultant instability is therefore prevented.

28. In the case of the servomechanism, this behaviour is imitated by attaching a tacho-generator to the output shaft as in Fig. 10.

The tacho-generator produces a voltage proportional to the angular velocity of the output shaft, and a suitable fraction of this voltage is fed back to the input of the amplifier in *opposition* to the error signal (negative feedback); this is known as *velocity feedback*.

29. It has been shown earlier that the error signal is $e = (\theta_i - \theta_0)$, where θ_i is a voltage proportional to the input demand and θ^0 is a voltage proportional to the output shaft position.

With velocity feedback, a voltage proportional to the *speed* of the output shaft is fed back to the input of the amplifier in opposition to the error signal. Thus, the *net input* to the amplifier is a voltage proportional to the error *minus* a voltage proportional to the speed of the output shaft.

The aim with velocity feedback is to reduce the net input to the amplifier to zero and then to reverse it *before* the output shaft reaches its final position: if the amount of feedback is correctly adjusted the result is that the momentum of the load, acting against the reversed torque, causes the load to come to rest just as it reaches the required position: this obviously reduces the risk of overshooting and subsequent instability.

30. This action is illustrated in Fig. 11. Initially, the error signal predominates and the load is accelerated. As the load velocity
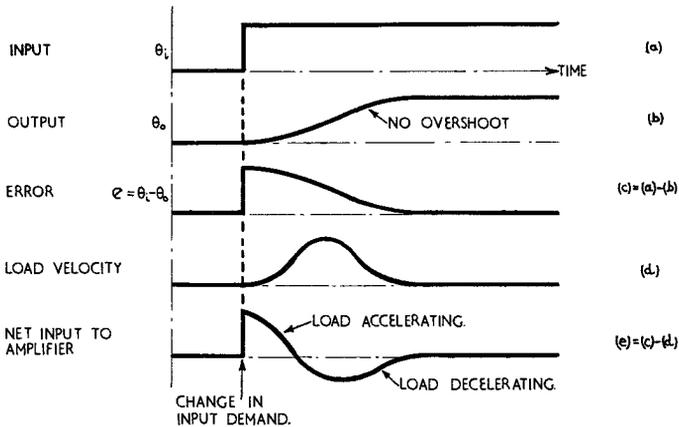
Fig. 11. PRINCIPLE OF VELOCITY FEEDBACK DAMPING.

rises and the error falls (i.e., input and output shafts coming into alignment), the net input to the amplifier drops rapidly and then increases in the *opposite* sense, so that a decelerating torque is applied to the load before it reaches the required position.

31. In addition to the advantage over a physical damping system of not causing a waste of energy, the velocity feedback method of damping possesses the important practical advantage that the amount of voltage fed back, and hence the degree of damping, can be simply controlled by inserting a potentiometer in the velocity feedback path. This is a method of damping frequently used in r.p.c. systems.

Velocity Lag

32. Velocity feedback provides a satisfactory means of obtaining the required response in r.p.c. systems because in the steady state such systems are quite stable. However, where the servomotor is required to rotate a load with a constant angular *velocity*, the disadvantage of velocity feedback becomes apparent.

Suppose the servomechanism load is a radar aerial that is required to rotate with a constant velocity. In such a system, a *ramp function input* is used as a means of investigating the servomechanism behaviour, in the same way that a step input is appropriate in r.p.c. systems.

A ramp input is illustrated in Fig. 12(a): it corresponds to the input shaft suddenly being rotated with a constant angular velocity, i.e., θ_i increasing linearly with time.

33. For a servomechanism of this type, with velocity feedback, the transient period is as already discussed. However the final result, after the initial transients have died out, is that the output shaft rotates *at the same speed* as the input shaft but lags behind it by some constant angle (see Fig. 12(b)).
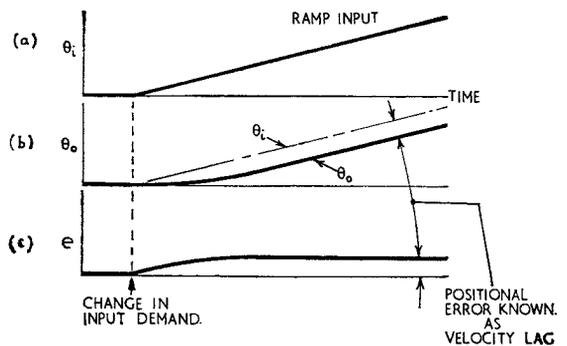
Fig. 12. RESPONSE OF A SERVOMECHANISM TO A RAMP INPUT.

The resultant instantaneous *positional* error between input and output shafts is known as '*velocity error*' or '*velocity lag*'. Its effect can be quite serious; it can result, for instance, in wrong radar bearings of a target.

34. In a velocity feedback system, velocity lag arises in the following way. Since the output shaft is rotating, the tacho-generator is producing a velocity feedback voltage

input to the amplifier. On the other hand, since the load is being rotated with a constant velocity, it is neither being accelerated nor decelerated, so *no torque* is required from the motor; that is, the *net input* to the amplifier must be zero (neglecting friction at bearings and wind resistance).

It has been shown that the net input, with velocity feedback, is a voltage proportional to the error *minus* a voltage proportional to the speed of the output shaft. Thus, if the net input is to be zero, and a voltage proportional to output speed is being fed back from the tacho-generator in opposition to the input, *there must be a balancing error signal*. There is, therefore, always an error signal in this system (and consequently an error) to compensate for the velocity feedback signal. The error signal, if it is to cancel that due to velocity feedback, must be proportional to the speed of the output shaft, and so velocity lag is proportional to output velocity.

Error-rate Damping

35. It has been shown that viscous damping improves the *transient* response of servo-mechanisms, but because of power losses it is used only on small servo systems. Velocity feedback similarly improves the transient response without introducing power losses and is, therefore, generally preferable in all but very small servo systems.

In r.p.c. systems, either of these two forms of damping can produce the required result, because it is only the *transient* performance that is important: the steady state error is zero in any case.

However, in angular velocity control systems, although velocity feedback improves the transient performance, it also unfortunately gives rise to a steady state error known as velocity lag. Steps must therefore be taken to reduce this error in servomechan-

isms required to rotate a load at constant speed.

36. Velocity lag is proportional to the speed of the output (and input) shaft. Therefore, if some other signal proportional to speed can be used to offset the velocity feedback, the error can be made zero. This could be done with the arrangement shown in Fig. 13.

One tacho-generator is mounted on the *output* shaft and produces a voltage proportional to the speed of this shaft. A second tacho-generator mounted on the *input* shaft produces a voltage proportional to input speed. There are therefore *three* input signals to the amplifier, and for the connections shown, the combined input is a voltage proportional to error *plus* a voltage proportional to input speed *minus* a voltage proportional to output speed.

In the steady state, the input and output shafts in a velocity feedback system rotate at the same speed, and in this condition the velocity lag is caused by the constant signal produced by the output tacho-generator. In the system of Fig. 13, this signal is exactly cancelled in the steady state, by the signal from the input tacho-generator. Therefore since the *net input* to the amplifier is required to be zero, the error signal (e) itself can be zero; that is, the system will have zero velocity lag.

37. The system outlined in para. 36 is not, in fact, a practicable proposition because of the difficulty of ensuring that the outputs from the two tacho-generators remain constant with time. Fortunately, a simplification is possible in which both tacho-generators can be dispensed with.

For velocity damping of *step position* inputs, a feedback voltage proportional to
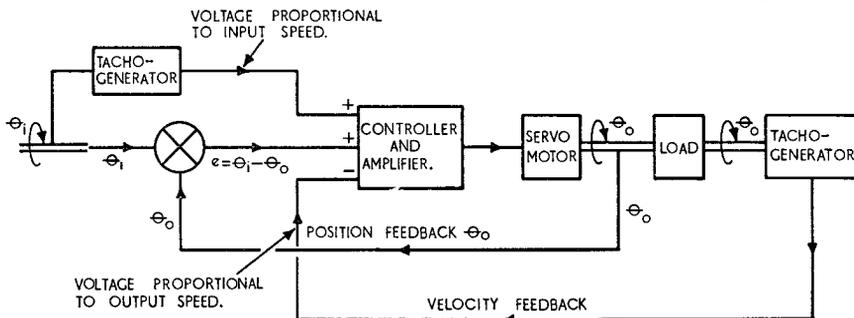
Fig. 13. METHOD FOR REDUCING VELOCITY LAG.

the speed of the output shaft is required. This, however, introduces velocity lag in servomechanisms required to rotate a load at constant speed, and to avoid or reduce this form of error, a voltage proportional to the speed of the input *minus* the speed of the output is required; that is, a voltage proportional to speed of (input—output) or proportional to speed of (error signal).

In the same way as the velocity of the output shaft equals the rate of change of θ_0 with time, so velocity of error equals rate of change of *error* with time. This is obtained by *differentiating* the error with respect to time.

Thus by differentiating the error signal and combining the derivative with the actual error at the input to the amplifier, the *net input* to the amplifier is a voltage proportional to the error *plus* a voltage proportional to speed of (input—output). This is the same form as that given in para. 36. The result will therefore be the same, so that in the steady state the system has zero velocity lag.

38. An arrangement for providing this is illustrated in Fig. 14. This method of stabilization is called 'derivative of error compensation' or *error-rate damping*.

Use of Stabilizing Networks

39. Stabilization of a servomechanism to obtain a good *transient* response in a r.p.c. system and a good *steady state* response in a velocity system can also be obtained by inserting a suitable network in the input to the servo amplifier. A typical circuit, known as a *phase-advance* network, is illustrated in Fig. 15.

For a *position control* system, if the servomechanism is subjected to a step position input, the error jumps immediately to its maximum value, because the output shaft momentarily will not move. Initially, therefore, since the capacitor cannot charge instantaneously, the full error voltage is developed across R_2 and applied to the amplifier, and the motor accelerates rapidly. As the capacitor charges, the voltage across it rises and the input to the amplifier *falls;* the motor torque, therefore, also drops. The effect of the network is initially to cause the load to accelerate quickly. The resistor R_1 is inserted to allow C to discharge on a pre-arranged time constant.

40. As the load approaches the required position, the error voltage falls. However, if the values of the components have been correctly chosen, the charge acquired by the capacitor during the initial period now causes the voltage across it to exceed the error voltage. Thus the voltage applied to the amplifier is now *negative* even though the error voltage is still slightly positive. In other words, a retarding torque is applied to the load *before* it reaches the required position: overshooting is prevented and stability during the transient period consequently improved.

41. For a *ramp* input, this system has almost zero error in the steady state: that is, velocity lag has been virtually eliminated. In the steady state, zero torque is required (no acceleration or deceleration) and for this condition to be satisfied, the input to the amplifier must also be zero. The net-
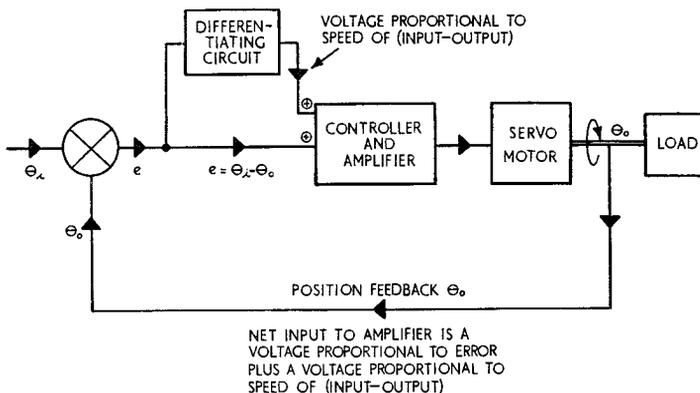
Fig. 14. ERROR-RATE DAMPING ARRANGEMENT.
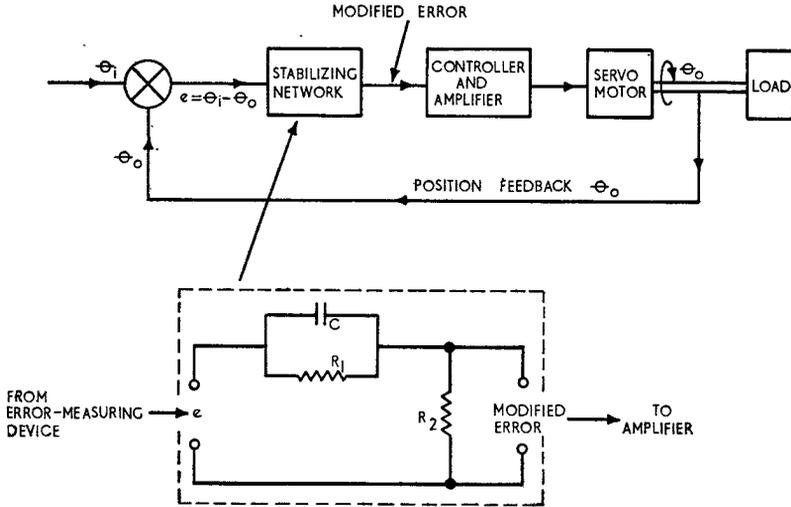
Fig. 15.   USE OF STABILIZING NETWORK TO REDUCE VELOCITY LAG.

work however will only supply zero voltage to the amplifier if the input to the network is zero, i.e., if the error is zero.

Transient Velocity Damping

42.   In an angular velocity control system, velocity feedback is used to improve the transient response: the only effect it has once the steady state is reached, is to introduce velocity lag.   During a state of steady rotation, the velocity feedback signal, which is the cause of velocity lag, is itself constant and therefore provides no damping.   It is only during the transient period that the velocity feedback signal is changing and therefore providing damping.

If it were possible to use only the *changing* part of the velocity feedback signal during the transient period and not the *constant* part during steady rotation, velocity lag would be reduced.   If this could be arranged, there

would be no velocity feedback signal at the amplifier input under conditions of steady rotation and no error signal would be required to offset it i.e., the velocity lag would be zero (neglecting inherent friction).

43.   A simple method of achieving this is shown in Fig. 16.   A voltage proportional to the speed of the output shaft is produced by the tacho-generator and this is applied to a CR network: that part of the velocity feedback signal appearing across R is applied to the amplifier input in *opposition* to the error signal.

The time constant of the CR circuit is such that only *variations* of voltage applied across the combination are developed across R and applied to the amplifier.   If the velocity feedback signal is constant (as it will be during steady rotation) then the voltage across R falls to zero with a time constant
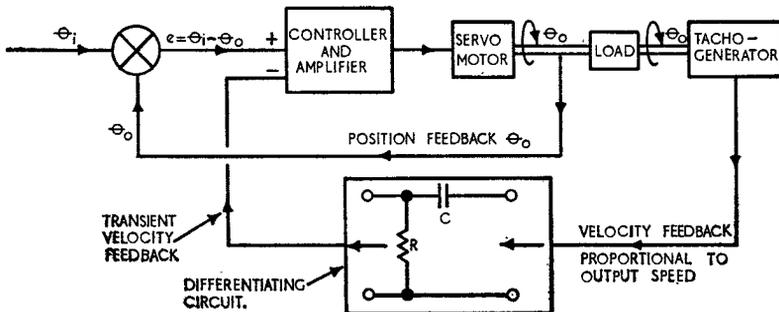
Fig. 16.   TRANSIENT VELOCITY FEEDBACK.
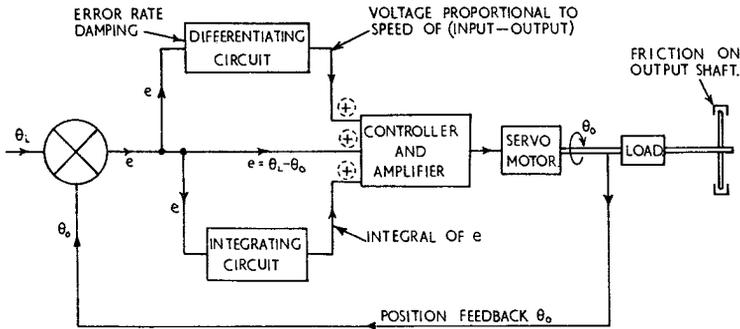
Fig. 17. INTEGRAL OF ERROR COMPENSATION.

of CR and the capacitor charges to this voltage. In other words, the tacho-generator output is *differentiated* by CR so that velocity feedback damping is effective only during the time that the output velocity is changing, i.e., during the transient period. This is the requirement.

This method of stabilization is known as *transient velocity feedback damping:* an alternative name is *acceleration feedback damping*, because the damping is effective only when the load velocity is changing.

Integral of Error Compensation

44. It has been assumed so far that the inherent damping and frictional losses in a servomechanism are so small that they can be neglected. Because of this, it was stated that in the steady state with a ramp input, no acceleration or deceleration was required and the input to the amplifier under such conditions was required to be zero.

However, in practice, there will always be a small amount of damping due to bearing and commutator friction; also in large aerials driven by a servomechanism, wind friction will introduce damping: the damping is quite insufficient to produce a stabilized performance and for this reason additional damping, of one of the forms described in the preceding paragraphs, has to be provided.

Nevertheless, the fact that there is inherent damping in these systems means that some torque is required on the output shaft even in the steady state for a ramp input, and a finite error will exist.

45. One method of reducing the steady state error in a velocity control system is illustrated in Fig. 17.

It is assumed that some inherent damping is present on the output shaft but that it is insufficient to provide a correctly damped response: error-rate damping is, therefore, provided as well. To reduce the velocity lag that results from the inherent frictional damping, the amplifier is also supplied with the *integral* of the error.

46. Neglecting the operation of the integrator for the moment, it has been shown earlier that in the steady state for a ramp input, there is a steady error known as
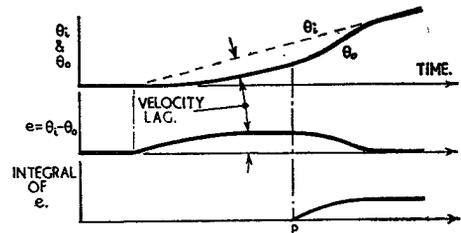
Fig. 18. PRINCIPLE OF INTEGRAL OF ERROR COMPENSATION METHOD.

velocity lag due to the friction on the output shaft. This is illustrated in Fig. 18.

Suppose now that the error signal is applied to the integrator and that this circuit becomes effective at point P in Fig. 18. The operation of the integrator is such that its output increases continuously in the presence of an error. Thus the total input to the amplifier starts to increase and the output torque consequently becomes greater than that absorbed by friction in the system: the excess torque accelerates the load and the output shaft starts to catch up on the

input shaft. As it does so, the error is reduced and the integrator output increases more slowly with time.

Nevertheless, as long as there is a finite error, the integrator output continues to increase with time. Thus equilibrium is established and a steady state attained only when the error signal has fallen to zero, at which point the integrator output remains constant at the required level.

47. This method of stabilization obviously has tremendous advantages, but unless the integrating circuit is carefully designed, over-correction can result and the error-rate damping provided may be insufficient to prevent oscillation. Under-correction is just as bad, because the effect then is that the output shaft catches up on the input shaft only very slowly. Design is therefore very critical.

The integrator circuit commonly takes the form of a series CR network connected in a phase-lag circuit, the output being derived from the voltage across C. The time constant is very important.

Summary of Stabilization Methods

48. The effects of the various methods used for obtaining stability in servomechanisms are illustrated in Fig. 19. The approximate response to step input and ramp input functions are shown and appropriate remarks are included.

Power Requirements of a Servomechanism

49. In the preceding paragraphs a broad outline of the basic principles involved in the operation of servomechanisms and of the steps taken to improve stability have been considered. To complete the picture it is necessary to have another look at the essential elements of a servomechanism so that some idea of the range of applications can be obtained.

50. It has already been stated that in a servomechanism much greater power is associated with the output than is available from the source of the input signal: in this sense therefore a servomechanism can be looked upon as a power amplifier.

The power requirements may be small or may be very great, depending on the job the servomechanism has to do. In what may be termed 'instrument servos', the drive power required is only a few watts. In other systems that are required to control the movement of large radar scanners, the output power required may be of the order of kilowatts.

51. Some idea of this range of application is given in the illustration of Fig. 20. In (a) the servomechanism is required to rotate the large radar scanner: the size of the load is such that the power output required is of the order of 100 kilowatts.

In (b) on the other hand, the equipment shown is carried in an aircraft and is used

METHOD OF DAMPING.	RESPONSE TO STEP INPUT.	RESPONSE TO RAMP INPUT.	REMARKS.
INHERENT FRICTION. (UNDERDAMPED)			FAST OSCILLATORY RESPONSE, (CONSIDERABLE HUNTING): SMALL VELOCITY LAG
VISCOUS DAMPING.			SLOW DAMPED RESPONSE LARGE VELOCITY LAG USEFUL ONLY ON SMALL SERVOS BECAUSE OF POWER LOSSES.
VELOCITY FEEDBACK DAMPING.			FAST DAMPED RESPONSE: LARGE VELOCITY LAG SATISFACTORY FOR MOST R.P.C SYSTEMS.
ERROR-RATE DAMPING PHASE-ADVANCE NETWORK TRANSIENT VELOCITY FEEDBACK DAMPING.			FAST DAMPED RESPONSE: SMALL VELOCITY LAG:
INTEGRAL OF ERROR.			FAST DAMPED RESPONSE PRACTICALLY ZERO VELOCITY LAG DIFFICULT TO MAINTAIN STABILITY

Fig. 19. COMPARISON OF DAMPING METHODS.
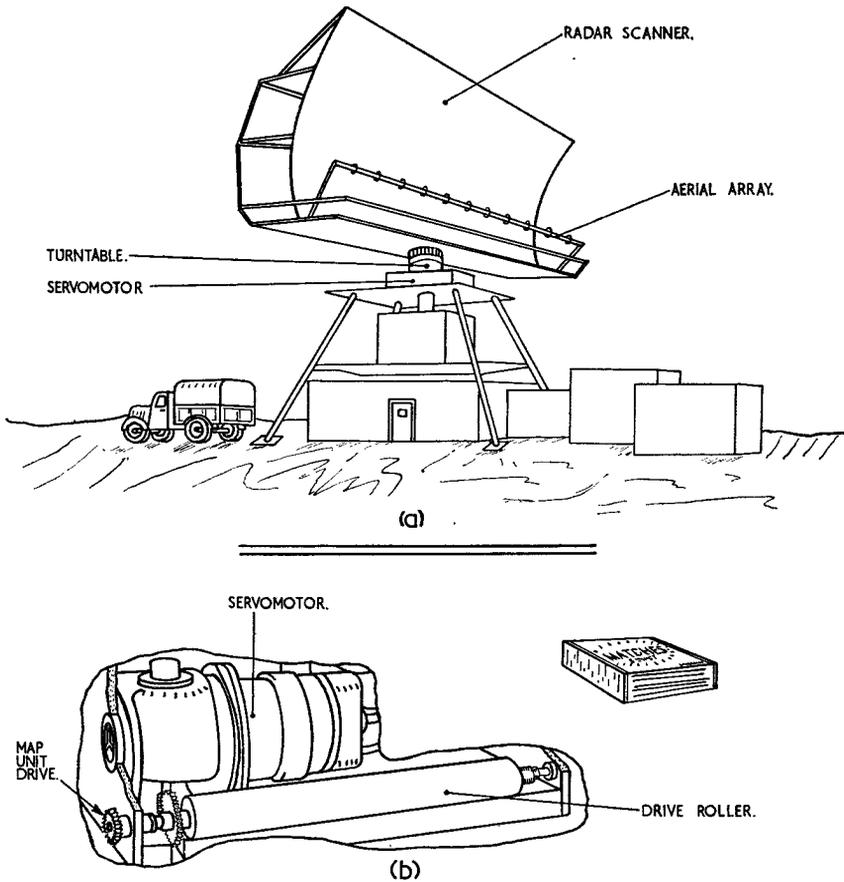
Fig. 20.   RANGE OF APPLICATION OF SERVOMECHANISMS.

to produce a map of the ground over which the aircraft is flying: a servomechanism is required to drive the roller at a speed proportional to the ground speed of the aircraft: it is a small mechanism, the power output requirements being of the order of a few watts.

Components in a Servomechanism

52.   Fig. 21 shows the essential elements in a r.p.c. servomechanism.   Brief notes on each of these elements for varying applications and power requirements are given in the following paragraphs.

53.   **Error-measuring device.** In d.c. systems, the input and output shafts are commonly connected to potentiometer wipers that pick off d.c. voltages proportional to θ_i and θ_0 respectively.   The error-measuring device in this case is merely the arrangement of the connections at the amplifier such that

Fig. 21.   ESSENTIAL ELEMENTS IN A R.P.C. SERVOMECHANISM.
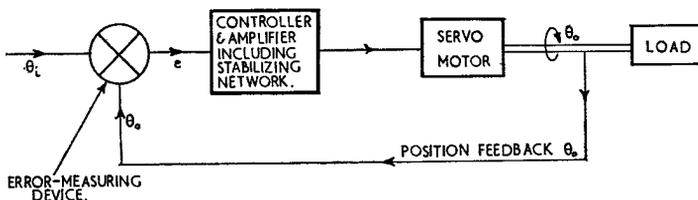
SYNCHRO CONTROL
TRANSMITTER

SYNCHRO CONTROL
TRANSFORMER
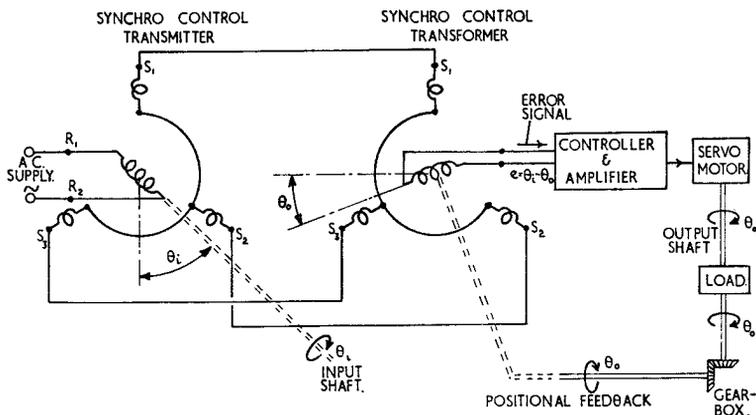
Fig. 22.   A.C. ERROR-MEASURING DEVICE.

the input to the amplifier is the error voltage proportional to $\theta_i - \theta_0$.

In a.c. systems, a control synchro system is usual as the error-measuring device. This has been dealt with in Chapter 1 and a typical arrangement is illustrated in Fig. 22. This shows that the input to the amplifier is an a.c. error signal proportional to $\theta_i - \theta_0$.

54. Controller, amplifier and servomotor.

The components used in these stages, and their size and complexity, are determined by the power output requirements of the servomechanism.

(a) *Low-power d.c. servos.* In small d.c. servomechanisms where the power output required is of the order of a few watts, the error signal is applied to a static d.c. amplifier or several such amplifiers in cascade. Either thermionic hard valve or magnetic amplifiers can be used and the circuit will include the usual stabilizing arrangements to ensure the required response. The d.c. output of the amplifier controls the armature current of a separately-excited d.c. motor, thereby controlling the speed and direction of the output load. Where velocity feedback or transient velocity feedback is required, the feedback voltage can be obtained from a small d.c. tacho-generator mounted on the output shaft which produces a voltage proportional to speed. In certain cases, the back e.m.f. of the motor itself can be utilized.

(b) *Lower-power a.c. servos.* In small a.c. systems, the a.c. error signal from the control synchro is amplified through hard

valve or magnetic amplifiers: the amplified signal can then be applied to one of the field windings of the driving motor—usually a two-phase induction motor (see Book 1, Section 5, Chapter 4). It will be remembered that the magnitude and phase of the output voltage from a control synchro depends on the magnitude and sense of the error and is either in phase or in anti-phase with the reference alternating voltage. If the reference voltage is applied to one field winding of the two-phase induction motor through a 90° phase shifting network, and the amplified error signal is applied to the quadrature field winding, the motor rotates. Since the field due to the error is now either 90° leading or 90° lagging on the reference field, the speed and direction of rotation of the motor depends on the magnitude and sense of the error.

An *a.c. tacho-generator* is used to give velocity or transient velocity feedback. The a.c. tacho-generator produces an alternating voltage at a constant frequency, but the *magnitude* of the generated voltage is directly proportional to the speed at which the generator is driven, i.e., the output speed of the servomechanism. The rotor is a 'drag-cup' or hollow brass cylinder, but the principles of operation can be more easely explained by considering a squirrel-cage rotor, i.e., made up of a large number of uninsulated coils as illustrated in Fig. 23(a). As the rotor rotates, voltages are induced in the coils by interaction with the primary field: at any instant, maximum e.m.f. is induced in the coil passing through the primary axis, i.e.,
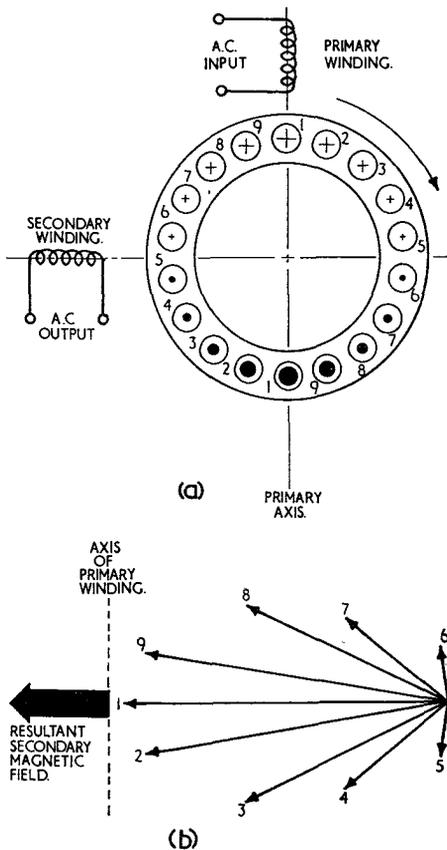
Fig. 23.   A.C. TACHO-GENERATOR PRINCIPLE.

in coil 1 of Fig. 23 (a).   The e.m.f.s induced in the other coils are progressively less as the axis of the coil departs from the primary axis.   However, the resultant of all the secondary fields produced by the currents induced in the coils from the primary field is, at any instant, at *right angles* to the axis of the primary field (Fig. 23(b)).   This secondary field oscillates at the frequency of the supply current but its *magnitude* is proportional to the speed of the rotor.   The secondary stator winding of the tacho-generator is at right angles to the primary winding and so has a voltage induced in it by the *secondary* field only: this is the output voltage.   A typical a.c. tacho-generator provides a signal output of approximately $0 \cdot 5V$ per 1000 r.p.m. of rotor.

(c) *Moderate-power servos*.   In most r.p.c. systems where the power requirements

are in excess of about 100 watts, a separately-excited d.c. motor is used to drive the load.   The input to the system, on the other hand, is usually from a synchro control transformer and is therefore an a.c. error signal.   There must then be conversion from a.c. to d.c. in the servomechanism.   The small a.c. error signal is amplified in a static hard valve or magnetic amplifier and is then applied to a special type of rectifier known as a *phase-sensitive rectifier*.   It was noted earlier that the magnitude and phase of the voltage from a synchro control transformer depends on the magnitude and sense of the error.   A simple rectifier takes no account of phase difference so that a rectifier circuit sensitive to phase is required.   A phase-sensitive rectifier is supplied with a reference alternating voltage against which to compare the phase of the error voltage, and it produces an output of the correct polarity depending on that comparison (Fig. 24).

The output from the phase sensitive-rectifier is, in effect, a d.c. voltage of magnitude and polarity depending on the magnitude and sense of the error.   This voltage is further amplified through hard valve or magnetic amplifiers before being applied to the power-amplifying stage.

The maximum power available at the output of a hard valve amplifier is usually limited by economic considerations to about 20 watts.   Magnetic amplifiers or gas-filled valve amplifiers (e.g. thyratron) can, however, be designed to produce outputs of the order of a few kilowatts. Where this is sufficient, the power amplifier controls the power into the armature of a separately-excited d.c. motor that drives the load.   The speed and direction of rotation of the load depend on the magnitude and sense of the error signal.   The circuit will include the usual stabilizing and feedback circuits to improve the response of the servomechanism.

(d) *High-power servos*.   If the power required is of the order of that needed to rotate the radar scanner of Fig. 20(a) (i.e., about 100 kilowatts) it is more practical to use rotating machinery in the power-amplifying stage.   A typical arrangement is shown in Fig. 25.

The output from the static valve or
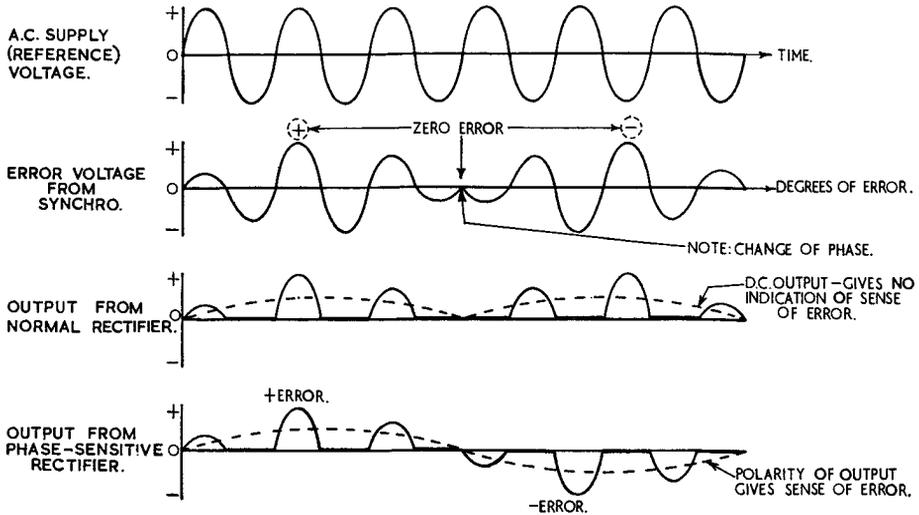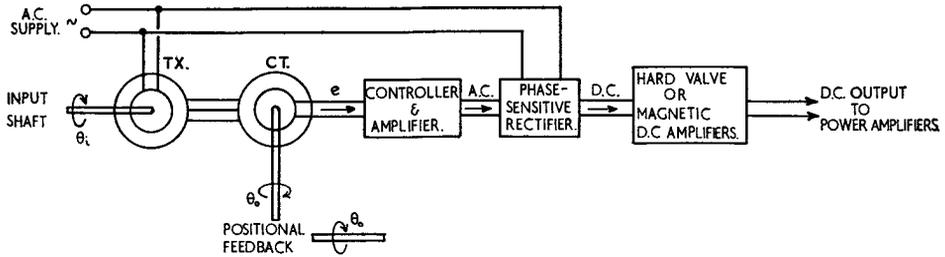
Fig. 24.   OPERATION OF PHASE-SENSITIVE RECTIFIER.

magnetic amplifier energises the field of a small exciter generator G_1 which is capable of producing an output of, say, 100 watts. This in turn is sufficient to energise the field of the main generator G_2 which may produce an output of, say, 10 kilowatts to drive the separately-excited motor M. Power amplification has thus been achieved.

This system can be extended to produce outputs of the order of several hundreds of kilowatts by connecting a number of generators in cascade. It is, however, more economical to group two or even three generators inside one casing. This is the method adopted in the Amplidyne and Metadyne generator systems.
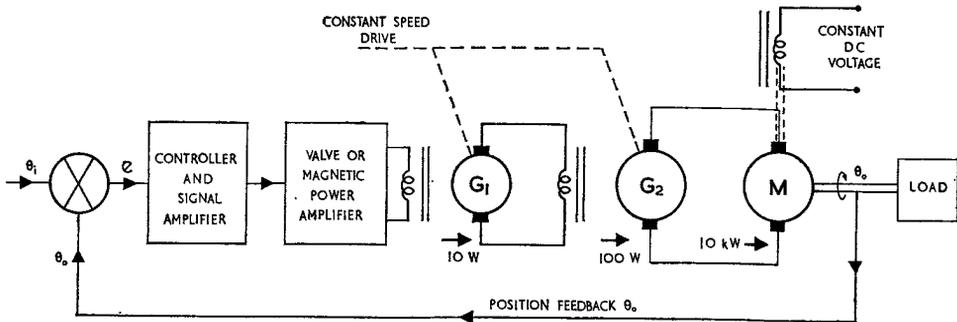
Fig. 25.   USE OF ROTARY POWER AMPLIFIERS.
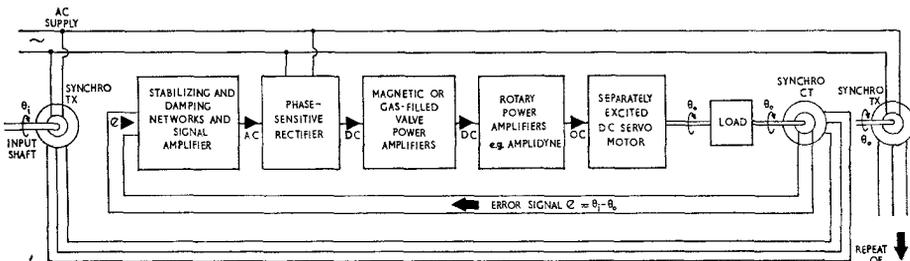
Fig. 26.   SOME TYPICAL SERVOMECHANISM ARRANGEMENTS.

Summary

55. This chapter has considered briefly the operation of a basic servomechanism and its behaviour to step input and ramp input functions: the steps taken to improve stability were also discussed. Mention was made of the range of applications of servo-mechanisms, and a general outline of the methods adopted to cover this range was given. These specimen methods are illustrated in the block diagrams of Fig. 26.

It should be noted however that there is considerable room for manœuvre on the part of the designer within this broad area. Because of this, discussion has intentionally been kept at block diagram level.

Circuit details of servo amplifiers (including magnetic and thyratron amplifiers), phase-sensitive rectifiers and generator power-amplifier systems of the Amplidyne type are considered in the more appropriate part of these Notes—Part 3.

SECTION 20

COMPUTING PRINCIPLES AND CIRCUITS

.

SECTION 20

COMPUTING PRINCIPLES AND CIRCUITS

SECTION 20

CHAPTER 1

ANALOGUE COMPUTERS

ANALOGUE COMPUTERS

Introduction

1. Modern electronic calculating machines, or *computers*, are widely used in science, industry and commerce to give accurate and quick answers to complex and often continuously varying problems. A computer is, in effect, a machine which accepts information presented to it in its required form, carries out arithmetical and other calculations automatically, and supplies the results in a form that can be readily understood. Computers are of interest to the technician because he is concerned with their maintenance.

The very wide field of use for computers means that there are considerable differences in design for each application. It is therefore not possible in these notes to give anything more than a general introduction to the subject and an elementary idea of the principles and circuits used.

2. In these days of high-speed aircraft and guided weapons, computers are becoming increasingly important in the Service. For example, to enable a modern bomber to reach a position from which it can bomb a target accurately and quickly, many factors have to be considered: these factors include airspeed and aircraft heading, windspeed and direction, bomb characteristics, height above the target and distance flown: all these things have to be computed, and the computation must be done accurately and quickly: because many of the factors are subject to random and continuous variation, the necessary calculations can be done efficiently only by using an electronic computer.

Conversely, when a ground radar is tracking a hostile target (e.g., a guided weapon) information about the target's height, bearing, range, speed and track must be computed accurately, quickly and continuously if efficient counter-measures are to be taken. This can be done adequately only by using an electronic computer.

Two Kinds of Computer

3. All computers can be divided into two distinct kinds—*analogue* and *digital*. The *analogue computer* operates by a process of *measurement* of various physical quantities (voltages, currents, shaft rotations, etc.) which are in turn used to represent numerical values: the physical quantities are said to be *analogues* of the respective numerical values.

Probably the best known example of an analogue computer is the slide rule, in which distances along the stock and slide correspond to logarithms of numbers; since adding the logarithms of two numbers gives the logarithm of the product, multiplication can be done simply by adding together lengths of scale. Thus in a slide rule, '*length along the scale*' is the *analogue* of the numerical value.

4. *Digital computers* are basically arithmetic machines: that is, they operate by a process of *counting* numbers or *digits* (whence their name). The fundamental operations that a digital computer can perform are addition, subtraction, and determination of which of two given numbers is the greater.

The best known and simplest example of a digital computer (apart from one's fingers) is the bead frame or abacus that young children use for simple addition and subtraction: in the abacus the digits are beads: in an electronic digital computer the digits are electrical *pulses*.

Comparison Between Analogue and Digital Computers

5. **Accuracy.** *Analogue machines* are usually much less accurate than digital computers because of the difficulty of setting the analogue quantity accurately, and of measuring the answer to the same accuracy. For example, it would be practically impossible to construct a slide rule that was accurate to one part in a million. In general, therefore, the accuracy of analogue computers is of the order of 1%.

Digital computers on the other hand are capable of extreme accuracy, because they handle numbers expressed in finite digital form: by arranging for an extra place of figures in the computer, the accuracy can be increased. The accuracy obtained, therefore, is limited only by the size of the machine and modern digital computers have accuracies better than 0.001%.

6. **Physical size.** A *digital computer* needs a fairly large amount of equipment before it will work at all, and the size of the machine cannot usefully be reduced below a certain minimum. Thus, even simple electronic digital computers are fairly elaborate and costly devices, containing many components. However, there is a trend towards reducing the physical size of digital computers by miniaturisation and microminiaturisation using transistors and 'packaged' circuits. In addition, many digital computers use a *time-sharing* process where one circuit, by correct timing, can handle a number of different calculations in sequence: this obviously gives a large economy in components. These trends mean that digital computers are becoming more practical for most aircraft applications.

With *analogue computers*, simple problems can be handled with relatively little equipment, and the size and complexity of the machine only increases directly as the complexity of the problem.

7. **Programming.** To perform complicated operations, *digital computers* must be provided with a '*programme*' of instructions in which multiplication, division, integration and so on are broken down into a series of additions and other simple operations. For complex problems, programming can be difficult and lengthy.

In *analogue machines*, the 'programme' is *built in* by the manner in which the various units are interconnected and no special programme need be prepared.

8. **Form of output.** In a *digital computer*, the required result is extracted by reading a 'counter', the answer being presented directly as a number or as a set of numbers.

In an *analogue machine* there must be a process of measurement of the magnitude of the physical quantity which is the analogue of the solution.

It may be of interest at this point to note that a motor car contains both analogue and digital instruments: the speedometer uses pointer angle as the analogue of speed, and the mileage indicator, which is a counter, reckons distances directly in terms of a set of numbers.

9. **Applications.** From what has been said it will be clear that *digital computers* are used when the problem is basically of an arithmetical nature and an exact answer is required. For *specific* tasks, the programme can be 'built in', in much the same way as that in an analogue machine, and digital machines of this type have many aircraft applications. A general purpose digital computer, on the other hand, requires programme preparation and such machines are limited to use in ground installations where the cost can be justified.

Analogue computers are used if *continuously varying* quantities are to be dealt with and great precision is not necessary: since small analogue computers can adequately deal with relatively simple problems, this type is in general preferred for aircraft applications in which accuracies of the order of 1% are acceptable.

Analogue computers are considered further in this Chapter: digital computers are dealt with in Chapter 2.

Analogues

10. It has been stated that in analogue computing one quantity is used to represent another: it was shown that for a slide rule, 'length along the scale' is an analogue of a number. Another example of the use of an analogue quantity is the altimeter, which uses the measurement of barometric pressure to represent height; in this case, the pressure measured is the analogue of height. Also, in radar measurement of range, the time delay between the transmission of a pulse and the reception of a response is used to measure the distance between the transmitter and the target; time delay is an analogue of distance in this case.

11. One problem in analogue computing is to select appropriate *scales* for the analogues. For example, height may be represented by a voltage to a scale such as 5 volts for every 1,000 feet of height: or if the computing is mechanical, a shaft position such as 5 degrees of rotation for every 1,000 feet of height can be used as the analogue.

It is sometimes necessary to change the *scale factor* at various points in the computer so that the stages can adequately handle the range of quantities applied. However, provided all analogues are changed in the same ratio and due allowance is made for the change, the final result is not affected. Thus, computation of 4 volts as an analogue of height and 2 volts as an analogue of range will produce the same answer as computation of 2 volts and 1 volt respectively,

provided the circuit is arranged in such a way that this change of scale factor is accounted for.

Conversion of Analogues

12. The individual stages in a computer are used to carry out specific processes on the analogues fed to them; that is, addition, subtraction, multiplication, division, differentiation, integration and so on. It is often necessary to change from one type of analogue quantity to another type to enable different mathematical operations to be carried out. Some of the more important conversion processes are considered in the following paragraphs.

13. **Conversion of a shaft angle to a voltage.** The conversion of a given shaft rotation to an alternating voltage analogue can be obtained with the control synchro system discussed in Sect. 19, Chap. 1, Fig. 28.

However, a simple system that is commonly used and is equally suited to both d.c. and a.c. is illustrated in Fig. 1. A high-grade, linear potentiometer (known as a '*R-pot*')
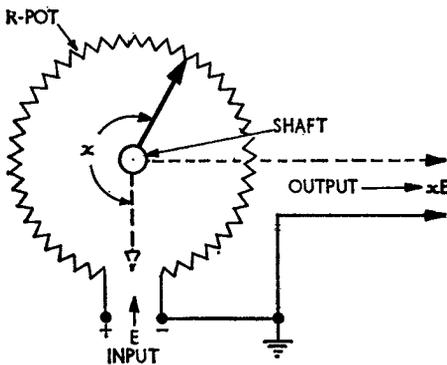
Fig. 1. USE OF R-POT TO CONVERT A SHAFT ANGLE TO A VOLTAGE.

is connected across a voltage E, and the wiper of the R-pot is set by the shaft angle. Thus, if the wiper is rotated by a fraction x from one end, the output voltage is xE, i.e., directly proportional to x, the shaft angle. In this way the shaft angle has been converted to a voltage analogue. To ensure accuracy:—

(*a*) the magnitude of the supply voltage E must be *constant*:

(*b*) The R-pot must be *linear*:

(*c*) there must be very little loading on the R-pot, i.e., it must feed into a *high impedance* (see Paragraph 20).

14. **Conversion of a voltage to a shaft angle.** The reverse process of converting a voltage to a shaft angle analogue is carried out as shown in Fig. 2. The given voltage xE is compared in an error-measuring device
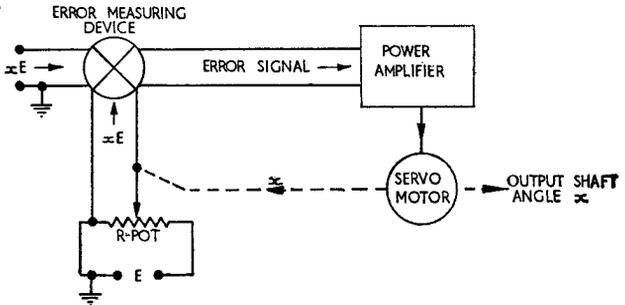
Fig. 2. CONVERSION OF A VOLTAGE TO A SHAFT ANGLE.

with the voltage produced at the wiper of a R-pot connected across a voltage E. If there is any difference between these two voltages, the resultant error signal is amplified and used to control the power into a servo motor. The servo motor drives the wiper of the R-pot (and also the output shaft) in such a direction as to reduce the error. Provided attention is paid to problems of stability (see Sect. 19, Chap. 2) the result is that the input to the error-measuring device from the R-pot can be made equal to xE: with zero error, the motor stops, having driven the output shaft through the angle x. Thus, x is a shaft angle analogue representing the input voltage xE.

15. **Conversion of voltage to current and vice versa.** This can be done with an *impedance:* a given current passed through an impedance results in a voltage across it which is directly proportional to the current: conversely, a given voltage across an impedance produces a current which is directly proportional to the voltage. In this way, a voltage analogue of a current or a current analogue of a voltage can be produced. With d.c., the impedance used must be a resistance: with a.c., inductance or capacitance can be used instead of resistance.

16. Conversion of alternating voltages to direct voltages. With direct voltages the *sign* of the analogue is determined by the polarity of the voltage with respect to a reference point (usually earth). Similarly, with shaft rotations, the sign of the analogue is indicated by clockwise or anti-clockwise rotations from a datum point.

With alternating voltages, the sign of the analogue is indicated by the *phase* of the voltage with respect to a reference voltage. The indication of phase must therefore be preserved when converting from an alternating voltage to a direct voltage analogue. A normal rectifier takes no account of phase. Therefore, a special kind of rectifier circuit, known as a *phase-sensitive rectifier*, is used. This is supplied with the reference alternating voltage against which to compare the input alternating voltage and it produces a d.c. output which is + or − according to the phase of the applied voltage (see Sect. 19, Chap. 2). In this way a direct voltage analogue of an alternating voltage can be produced.

17. Conversion of direct voltages to alternating voltages. The arrangement illustrated in Fig. 3 may be used to convert a direct voltage to an alternating voltage analogue. The direct voltage xE_D is first converted to a shaft rotation x as previously explained in Para. 14. The servo motor drives the wipers on both R-pots and when the shaft has been rotated through the angle x, the error signal into the amplifier is zero and the motor stops, both wipers having been turned through the angle x. The output voltage is then xE_A and the production of an alternating voltage analogue of a direct voltage has been achieved. Note that both the direct and alternating voltages applied to the R-pots are centre tapped to earth to

permit positive and negative signs to be handled.

18. Conversion of angular velocity to a voltage. It is sometimes necessary to produce a voltage analogue of the speed of rotation of a shaft.

For *direct voltages*, a d.c. tacho-generator can be mounted on the shaft. This is a d.c. generator constructed in such a way that it produces an output voltage directly proportional to the speed at which it is driven (see Sect. 19, Chap. 2, Para. 10).

For *alternating voltages*, an a.c. tacho-generator (or drag-cup generator) can be mounted on the shaft. This is a specially-constructed a.c. generator in which the frequency of the generated voltage remains constant, but the *magnitude* of the voltage is directly proportional to the speed at which it is driven (see Sect. 19, Chap. 2, Para. 54.).

19. Conversion of angular velocity to a direct current. To produce a direct current analogue of the speed of rotation of a shaft, a device known as a *capacity commutator* can be mounted on the shaft. The principle is illustrated in Fig. 4.

When the switch is in position 1, C_1 charges completely to the supply voltage E volts. When the switch is moved to position 2, C_1 discharges virtually completely into C_2, since C_2 has a much larger value of capacitance than C_1.

For each complete cycle of the switching from position 1 to position 2, a charge of $Q (= C_1E)$ coulombs is transferred to C_2. The switch is controlled by a shaft and the frequency of switching (f cycles per second) is proportional to the speed of rotation of this shaft.
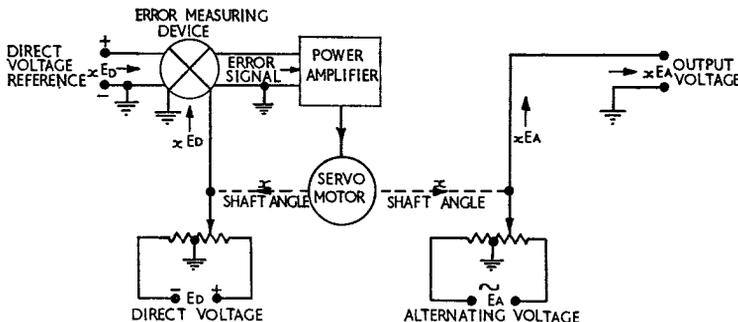
Fig. 3. CONVERSION OF DIRECT VOLTAGES TO ALTERNATING VOLTAGES.
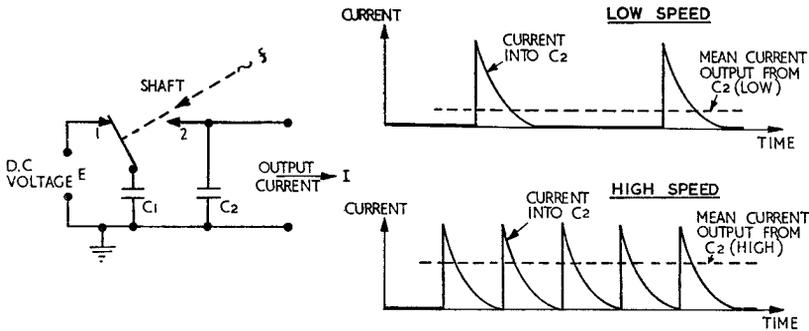
Fig. 4. CONVERSION OF ANGULAR VELOCITY TO A DIRECT CURRENT.

Thus, the mean current I from C_2 equals C_1Ef amperes. Since E and C_1 are constant then the current is directly proportional to f, the speed of rotation of the shaft, as indicated in Fig. 4.

Although the current into C_2 is in the form of pulses, the output current is virtually steady because C_2 acts as a smoothing capacitor. For accuracy, E must be constant (usually a stabilized supply) and f, the frequency of switching, must be low enough to allow full charge and discharge of C_1. Since the output is a current, it must feed into a *low impedance*.

The simple system described has the disadvantage that the output has the same polarity whatever the direction of rotation of the shaft. This disadvantage can be overcome by using an arrangement such as that shown in Fig. 5.
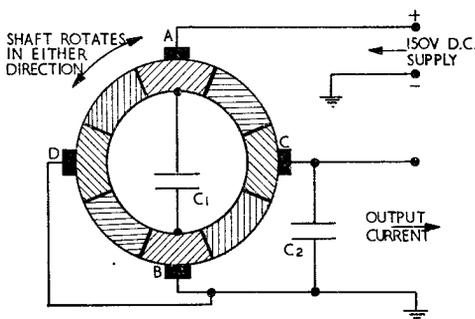
Fig. 5. CAPACITY COMMUTATOR.

A capacitor is connected between each opposite pair of segments of the commutator. When the capacitor C_1 is connected between A and B, it charges to 150 volts. If the shaft now rotates *clockwise*, then after 90 degrees rotation the positive plate of C_1 will be connected to the output and the negative plate will be earthed. If the shaft had rotated *anti-clockwise*, then after 90 degrees rotation the positive plate of C_1 would have been connected to earth and the negative plate connected to the output. Thus for a *clockwise* rotation there is a *positive* output: for an *anti-clockwise* rotation there is a *negative* output. The *sign* of the output therefore depends on the *direction* of rotation and the *magnitude* depends on the *speed* of rotation. A current analogue of speed of rotation of a shaft has therefore been obtained.

Need for Negative Feedback Amplifiers

20. In analogue computers it is important that each of the units performing the various mathematical processes do not react upon each other, otherwise inacuracies in the computation would result. The problem is to take the analogue quantity (a voltage, say) which is the output from one unit, without imposing any load upon that unit, and to 'copy' it at a lower impedance to provide the input to the next unit.

For example, it was noted earlier that in the conversion of a shaft angle to a voltage analogue, the load on the R-pot must be reduced to a minimum. If the R-pot feeds directly into a low input impedance the current drawn will upset the voltage distribution on the R-pot and the output voltage will no longer be accurately proportional to the shaft angle.

This loading effect is clearly undesirable but at the same time it is necessary to provide sufficient current to operate succeeding units in the computer. These conflicting requirements can be met by inserting a buffer amplifier with *high input impedance* and *low output impedance* between the R-pot and its load. A cathode follower is an amplifier

X

having these characteristics (see Book 2, Sect. 10, Chap. 3, Para. 32): it can therefore be used for this purpose. The low output impedance can provide the current necessary to operate the next stage, while the high input impedance ensures negligible load on the R-pot.

Conversely, a capacity commutator which provides a *current* analogue of the speed of rotation of a shaft must feed into an amplifier that has a *low input* impedance and a *low output* impedance.

21. Since the gain of the amplifier used will modify the analogue scale factor, it is necessary to know the gain accurately. If the gain of the amplifier is unity, there is no change of scale factor: a gain of two provides an output voltage scale which is double the input scale and allowance must be made for this change of scale factor.

It is also essential that the gain of the amplifier should remain *constant*, i.e., that it should not be altered appreciably by changes in the characteristics of the components. This is particularly difficult when thermionic valves are employed because their characteristics change gradually with age. Fortunately these difficulties can be considerably reduced by using *negative feedback amplifiers*.

22. Negative feedback amplifiers are discussed in detail in Book 2, Sect. 10, Chap. 3 of these notes. The important effects of negative feedback on amplifier performance for computer applications are summarised below:—

(a) The overall gain of the amplifier is reduced from A to A' which equals $\dfrac{A}{1 + \beta A'}$ where A is the inherent gain without negative feedback, A' is the gain with negative feedback and β is the feedback factor which determines the fraction of the output which is fed back.

(b) If the inherent gain A of the amplifier is very large, then the gain A' with negative feedback is approximately equal to $\dfrac{1}{\beta}$: in this case the overall gain of the system is independent of the characteristics of the amplifier and is determined only by the feedback factor β: the gain of the system is then practically constant irrespective of any changes in the characteristics of the amplifier.

(c) If the inherent gain A of the amplifier is very large and β is made equal to unity (i.e., *all* the output quantity is fed back to the input as negative feedback) then the gain A' with negative feedback is approximately unity: this is often a requirement in computers.

(d) The input and output impedances of a negative feedback amplifier are determined by the form that the feedback takes. For computers there are two main requirements:—

(i) amplifiers with *high input* impedance and *low output* impedance:

(ii) amplifiers with *low input* impedance and *low output* impedance:

Voltage negative feedback is general, so the feedback loop is connected in *parallel* with the *output* load: this *reduces* the output impedance and by suitable choice of the feedback loop it can be made a *low value*.

To get a *high input* impedance, the voltage negative feedback must be applied in *series* with the *input*: for a *low input* impedance it is applied in *parallel* with the *input*.

Arrangements which meet these requirements are illustrated in the block diagram of Fig. 6.
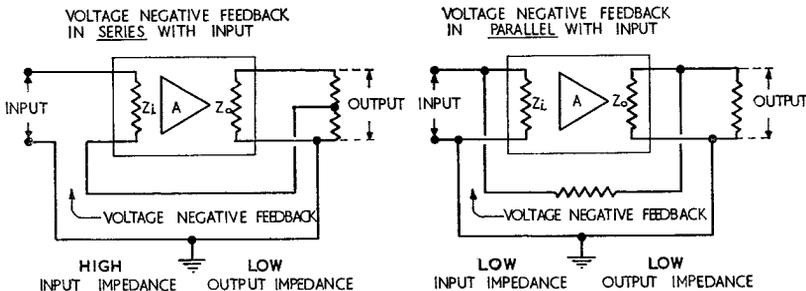
Fig. 6. APPLICATIONS OF VOLTAGE NEGATIVE FEED-BACK IN AMPLIFIERS.

Simple Mathematical Processes in Analogue Computing

23. In the computation of analogues, mechanical or electrical stages are required which can perform the normal algebraic processes of addition, subtraction, multiplication and division.

For example, if an electronic computer was required to compute the voltage amplification factor of an amplifier stage with different values of load resistance R_L and different types of valve, it would have to work out the equation:—

$$V.A.F. = \frac{\mu\, R_L}{R_L + r_a},$$

where μ = amplification factor of valve
r_a = anode slope resistance of valve.

Thus it would require stages capable of multiplying, adding and dividing analogues as shown in Fig. 7.
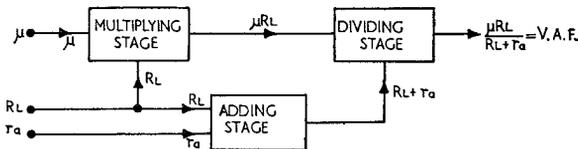
Fig. 7. SIMPLE MATHEMATICAL PROCESSES IN ANALOGUE COMPUTING.

There are many ways in which the necessary mathematical processes can be carried out in analogue computers. The following paragraphs consider some of the more common methods.

Addition and Subtraction of Shaft Rotations

24. **Differential gear.** One way of obtaining a shaft rotation analogue which is equal to the sum of two input shaft rotations is to use a differential gear (Fig. 8).

The gears on the two input shafts A and B both mesh with two idler gears C which are housed in a casing free to rotate about shafts A and B. Another gear D, integral with the casing, transmits any casing movement to the output shaft E with a step-up ratio of 2 : 1.

If gear B is held stationary, any rotation made by gear A will cause the idler gears to roll round gear B and thus impart a rotary movement to the casing. The angular movement made by the casing will be *one*
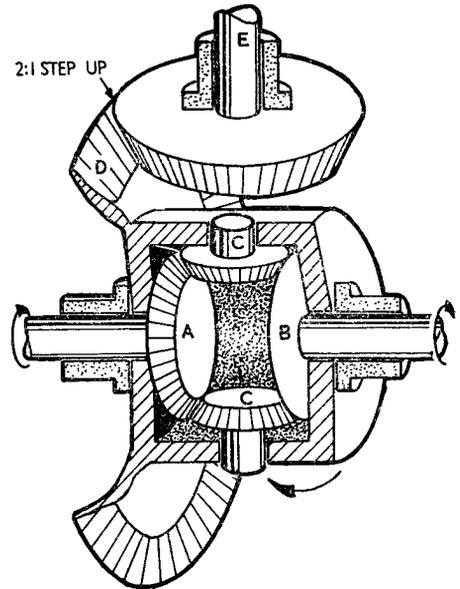
Fig. 8. DIFFERENTIAL GEAR.

half of that made by gear A. This may be seen as follows.

Consider a plank AB, one foot long, resting on a cylinder that is free to roll on a flat surface (Fig. 9). Rolling the plank in the direction of the arrow will cause the
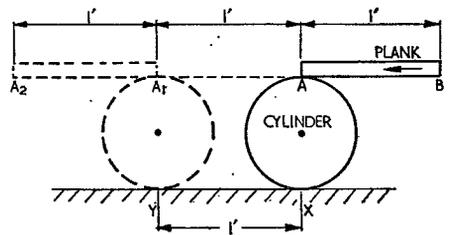
Fig. 9. PRINCIPLE OF DIFFERENTIAL GEAR.

cylinder to roll in the same direction. Thus, if the plank is moved through one foot, any point on the periphery of the cylinder moves one foot, causing the cylinder to move from X to Y. However, in imparting the motion to the cylinder, the plank also moves *relative* to the cylinder and therefore the total linear movement made by the plank (A to A_2) is *twice* that made by the cylinder.

Applying this principle to the differential gear of Fig. 8, if gear A is moved through θ degrees then gear C rolls round gear B causing the casing and gear D to move

TRANSMITTER   DIFFERENTIAL TRANSMITTER   RECEIVER
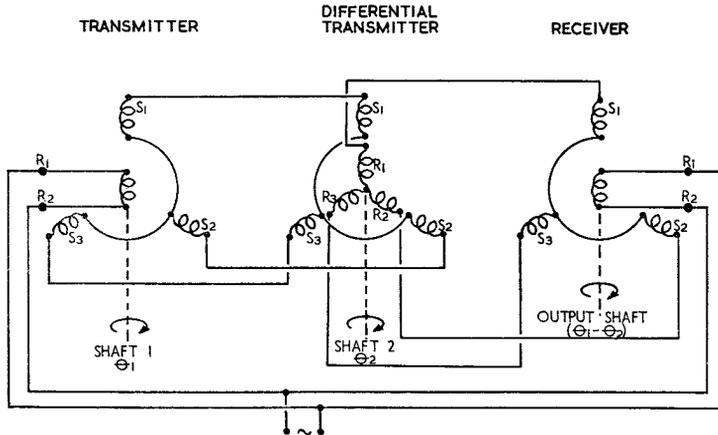
Fig. 10. TORQUE DIFFERENTIAL SYNCHRO.

through $\frac{\theta}{2}$ degrees. The output shaft E will thus move through θ degrees because of the 2 : 1 step-up ratio. Similarly, if B is moved through ϕ degrees, A being kept fixed, then the idler gears will roll round A and impart $\frac{\phi}{2}$ degrees rotation to the casing and gear B: the output shaft E will therefore move through ϕ degrees. Thus the total rotation of the output shaft E is $\theta + \phi$, the *sum* of the two input shaft rotations.

Subtraction can be obtained by *reversing* the direction of rotation of *one* of the input shafts.

The similarity between the differential gear discussed in this paragraph and that used in the rear axle of a motor car may be noticed. The mode of operation is, however, inverted, i.e., in a motor car rear axle differential, shaft E is the *driving* shaft while shafts A and B are the driven shafts connected to the rear wheels.

25. Torque differential synchro. This component is considered in detail in Sect. 19, Chap. 1. There it is shown that if the rotor of the transmitter is rotated through one angle, and the rotor of the differential transmitter is rotated through another angle, then the output rotor of the receiver moves through an angle equal to the sum or the difference of the two input angles, depending on the interconnections between the synchro elements. A simple arrangement for the subtraction of two angles of shaft rotations is shown in Fig. 10.

26. Conversion to electrical form. It is often preferable to convert the mechanical analogues to electrical form and then to add the electrical components. This provides an electrical analogue equal to the sum or difference of input shaft rotations. A typical arrangement is shown in Fig. 11.
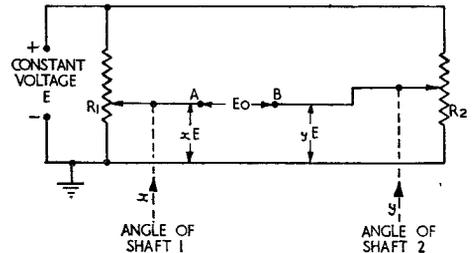
Fig. 11. ADDITION OF MECHANICAL ANALOGUES BY ELECTRICAL MEANS.

Two identical R-pots, R_1 and R_2, are each connected to a constant voltage source E. The wiper of R_1 is rotated by a fraction x from the zero datum position and the voltage tapped off is xE volts, directly proportional to x, the input shaft angle. Similarly, the voltage tapped off R_2 is yE volts, directly proportional to y, the other input shaft angle. The voltage analogue output E_0 from terminals AB is a voltage proportional to the *difference* between xE and yE: that is:—

$$E_0 = kE(x - y),$$

where k is the circuit constant.

The circuit therefore provides a voltage analogue of the *difference* between two shaft rotations.

If the constant voltage applied to R_2 is *reversed* in polarity and becomes $-E$, the voltage tapped off R_2 is $-yE$ and the voltage analogue output is then:—

$$E_0 = kE(x + y),$$

where k is the circuit constant.

The circuit now provides a voltage analogue of the *sum* of the two input shaft rotations.

As noted in an earlier paragraph, the output terminals must feed into a *high* impedance to reduce the loading on the R-pots.

27. Control differential synchro. This can also be used to provide an alternating voltage analogue output representing the sum or difference of two input shaft rotations. The rotor of the transmitter is turned through the angle determined by one of the input shafts whilst the rotor of the differential transmitter is turned through the angle determined by the second input shaft. The resultant orientation of the field inside the stator of the transformer is dependent on the interconnections between the synchro elements and is such that the voltage induced in the rotor is proportional to the sum or the difference of the two input shaft rotations. A typical arrangement is shown in Fig. 12.

Addition and Subtraction of Alternating Voltages

28. The addition or subtraction of alternating voltage analogues is performed by using Kirchhoff's second law. The voltage sources are connected in series and the voltage appearing across the combination is the algebraic sum of the individual voltages.

With alternating voltages no difficulty arises since the individual analogues can be connected to the primary windings of 1 : 1 transformers whose secondary windings can be connected in series as shown in Fig. 13. The output voltage analogue E_0 is then the *algebraic* sum of E_A, E_B and E_C.

Addition and Subtraction of Direct Voltages and Currents

29. **Addition of two voltages.** With direct voltages, transformers cannot be used to provide addition or subtraction. Furthermore, direct voltage analogues normally have one terminal common (earth) so that no more than *two* voltages can be connected in series for the purpose of adding or subtracting. Where only two voltages are to be added or subtracted, they can be applied to a simple resistive network as illustrated in Fig. 14.

Fig. 12. CONTROL DIFFERENTIAL SYNCHRO.

Fig. 13. ADDITION OF ALTERNATING VOLTAGES.
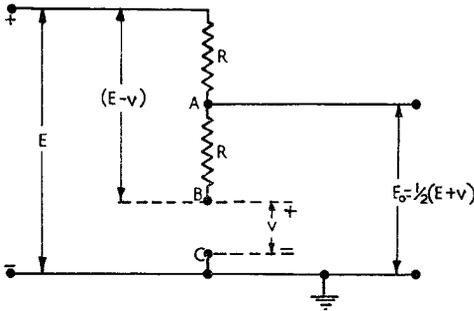
Fig. 14. ADDITION OF TWO DIRECT VOLTAGES.

The voltage across the two equal matched resistors R is $E - v$, where E is one voltage analogue and v is the other. Thus the voltage at A with respect to B, by simple potential divider action, is $\dfrac{R}{R + R} (E - v)$: since the resistors are matched, this reduces to $\frac{1}{2} (E - v)$.

Also, the voltage at A with respect to C (earth) is equal to that at A with respect to B plus (v): that is, $\frac{1}{2} (E - v) + v$, or $\frac{1}{2}(E + v)$. This is the output voltage analogue E_0. This arrangement therefore gives addition of two direct voltage analogues: at the same time, the analogue scale factor is changed and account must be taken of this by succeeding stages.

Subtraction of the two voltage analogues E and v will be obtained if the polarity of *one* of the inputs is reversed.

Again it is emphasised that for accurate computation, the loading on the output R-pot must be negligible. Thus the output voltage analogue E_0 must feed into an amplifier with a *high* input impedance.

30. **Addition of several voltages.** When more than two voltages are to be added or subtracted simultaneously, a star-point adding circuit (Fig. 15) is used. With this method, the direct voltage analogues are converted into proportionate *currents* which are then added at the star point to give a resultant current: this action is based on Kirchhoff's first law which states that the sum of the currents flowing to a junction (i.e., the star point) equals the sum of those flowing from it. The resultant current at the star point can then be converted back into an output voltage analogue which represents the sum of the input voltage analogues.

This principle is illustrated in Fig. 15 where all the resistors have the same value R. If point A is considered to be at zero volts (virtual earth) then I_1 equals $\dfrac{E_1}{R}$, I_2 equals $\dfrac{E_2}{R}$ and I_3 equals $\dfrac{E_3}{R}$.

The current I flowing from the star point A equals the *sum* of I_1, I_2 and I_3.

Thus, $I = \dfrac{E_1}{R} + \dfrac{E_2}{R} + \dfrac{E_3}{R}$

$\therefore IR = E_1 + E_2 + E_3$

But $IR = E_0$, the output voltage across the load.

$\therefore E_0 = E_1 + E_2 + E_3$,

provided A remains at zero volts.

This circuit therefore provides a means of obtaining a voltage which is the algebraic sum of any number of voltages all of which have one connection in common.

31. However, for point A to remain virtually at zero volts, the resistance of the load must be negligible: therefore E_0 will be extremely small and of little use for direct application to another circuit. This is overcome by using an amplifier to increase
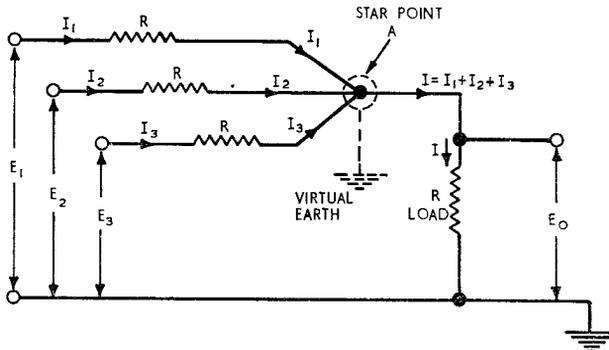
Fig. 15. STAR-POINT ADDING CIRCUIT.

the magnitude of E_0. The amplifier used employs parallel (shunt) voltage negative feedback and therefore has a *very low input impedance* which replaces the load resistance and maintains A at a potential very close to earth. Such an amplifier is termed a *see-saw summing amplifier* (see Para. 36).

The output from the star-point adding circuit represents the algebraic sum of the inputs and the circuit can add and subtract simultaneously, i.e., both positive and negative analogue inputs can be handled.

Furthermore, the circuit can handle analogues on different scales. In para. 30 it was assumed that all resistors had the same value so that the input currents were scaled differently according to the relationship $\frac{E_1}{R}$, $\frac{E_2}{R}$ and $\frac{E_3}{R}$, i.e., proportional to the appropriate voltage analogue input. This is not necessarily the requirement (see para. 36). By choosing *different* values of input resistors, all the input currents can be made to have the *same value* (i.e., $\frac{E_1}{R_1} = \frac{E_2}{R_2} = \frac{E_3}{R_3}$). Thus, by converting all the inputs to a *common* current scale, it is possible to sum input voltage analogues that have been computed to different scales.

32. It is often convenient to compute certain quantities as direct current instead of direct voltage analogues. The current analogue (e.g., from a capacity commutator) can be fed directly into the star point without any input resistor provided that the other inputs are converted to the same current scale: as before, the star point is maintained

virtually at earth potential by a circuit such as a see-saw summing amplifier.

See-saw Amplifier

33. For star-point addition of voltage analogues (and also for other applications), an amplifier with a *very low input impedance* is required. Such an amplifier is the see-saw amplifier, the basic circuit of which is illustrated in Fig. 16.

The circuit consists of a voltage amplifier using voltage negative feedback in *parallel* with the input, i.e., shunt voltage negative feedback. A circuit of this type has very low values of input and output impedances.

34. If the inherent amplifier gain A is assumed to be very high (in practice a two- or three-stage amplifier having a gain of the order of 25,000 is used) and the amplifier is also phase inverting, the point Y remains virtually at earth (zero volts) since the super-imposed effects of e_i and $- e_0$ very nearly cancel each other at the amplifier input. Since the impedance between point Y and earth is very small, point Y is referred to as '*virtual earth*'. The actual input impedance Z_i measured across the input terminals is also very low and, in fact, is approximately equal to R_i (provided the inherent gain A of the amplifier is very high).

35. Since point Y is at virtual earth, the voltage v_g between Y and earth is so small that the current flowing into the amplifier from Y is negligible and can be neglected at this stage. Thus the current I in the input
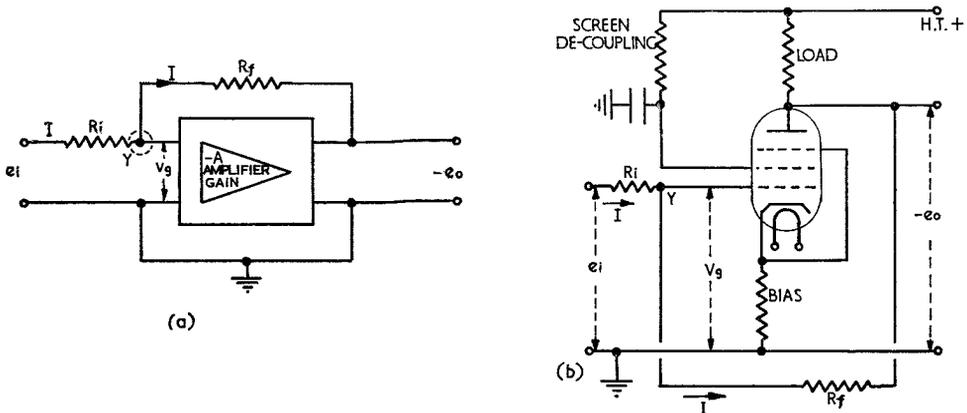
Fig. 16. BASIC SEE-SAW AMPLIFIER.

resistor R_i must also flow through the feed-back resistor R_f. Also, for the feedback to be negative, the output must be of opposite polarity to the input. Therefore:—

Current I in $R_i \simeq \dfrac{e_i}{R_i}$

(since Y is at virtual earth)

Current I in $R_f \simeq -\dfrac{e_0}{R_f}$

Thus, $I = \dfrac{e_i}{R_i} = -\dfrac{e_0}{R_f}$

$\therefore -\dfrac{R_f}{R_i} = \dfrac{e_0}{e_i} = A'$

(gain with feedback).

If the feedback resistor R_f and the input resistor R_i are equal in value then the gain $A' = -1$ and $e_0 = -e_i$: that is, the output voltage is *equal* in magnitude but of *opposite* polarity to the input voltage. This circuit can therefore be used as an 'inverter' but in addition, by adjusting the ratio $\dfrac{R_f}{R_i}$, the gain A' with feedback can be altered. This provides a method of changing the scale of a computed voltage.

Note that the potential division across R_i and R_f in series maintains the point Y virtually at earth potential.

See-saw Summing Amplifier

36. This circuit is a combination of a star-point adding circuit and a see-saw amplifier. The amplifier maintains the star-point at virtual earth. A typical arrangement is illustrated in Fig. 17.

If the inherent gain A of the amplifier is high, the point Y remains at a potential very close to zero volts and supplies negligible current to the amplifier.

Thus, $I_f = i_1 + i_2 + i_3$

$\therefore -\dfrac{e_0}{R_f} = \dfrac{e_1}{R_1} + \dfrac{e_2}{R_2} + \dfrac{e_3}{R_3}$

$\therefore -e_0 = \dfrac{R_f}{R_1} e_1 + \dfrac{R_f}{R_2} e_2 + \dfrac{R_f}{R_3} e_3.$

Now $\dfrac{R_f}{R_1} = $ the gain A'_1 of the see-saw amplifier to input e_1.

Also $\dfrac{R_f}{R_2} = A'_2$ and $\dfrac{R_f}{R_3} = A'_3$.

Thus $-e_0 = e_1 A'_1 + e_2 A'_2 + e_3 A'_3.$

It can be seen that if R_1, R_2 and R_3 are of different values, then the amplifier provides a different gain to each input. Thus the resistors can be chosen to alter the analogue voltage scale of an input as required.

37. It should be noted that a see-saw amplifier or a see-saw summing amplifier cannot be supplied directly with a voltage analogue from a R-pot, because the low input impedance of the see-saw amplifier would heavily load the R-pot. Where several voltages have to be computed, each voltage analogue derived from a R-pot is first applied to an amplifier having a *high input impedance* and a *low output impedance* (a *series* voltage negative feedback amplifier): the voltage analogue output of this amplifier (e.g., a cathode follower) can then be applied to the see-saw summing amplifier as one input, so that the effect of loading is negligible.

Multiplication by a Constant

38. Multiplication of an analogue quantity by a constant, including changing the sign
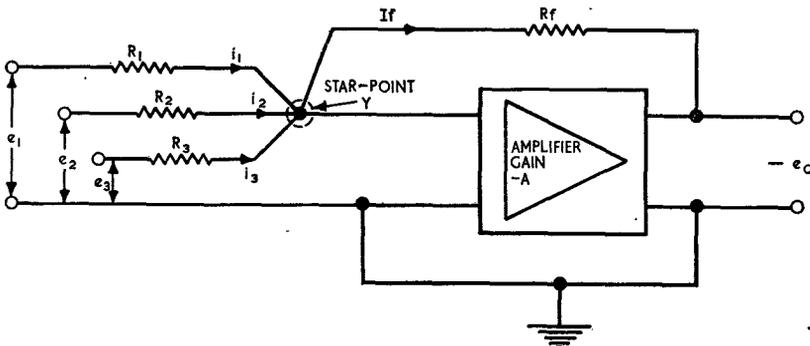
Fig. 17. SEE-SAW SUMMING AMPLIFIER.

of the analogue (i.e., multiplying by -1) can be obtained fairly easily.

With *alternating voltages* a transformer of the required turns ratio is all that is needed for multiplication by a constant: if the turns ratio is 3 : 1 step-up, an input analogue voltage V applied to the primary will produce 3V across the secondary. To change the sign of an alternating voltage analogue, the connections to *either* the primary *or* the secondary winding are changed over.

Shaft rotations can be multiplied with simple gear trains: an idler gear is inserted if change of sign of the output is required.

For *direct voltages*, change of sign can be obtained by applying the voltage analogue to the input of a see-saw amplifier in which the ratio of feedback resistor to input resistor is unity (i.e., $\dfrac{R_f}{R_i} = 1$): under this condition, as shown earlier, the output voltage e_0 is of the same magnitude but of *opposite* polarity to the input voltage e_i (i.e., $e_0 = -e_i$). If the feedback resistor R_f is not equal to the input resistor R_i, the input voltage analogue is multiplied by the factor $-\dfrac{R_f}{R_i}$: the output voltage e_0 then equals $-\dfrac{R_f}{R_i} e_i$, and multiplication by a constant has been achieved.

Multiplication of Two Variables

39. The direct multiplication of two analogue quantities of the same kind to give a product also of the same kind is neither a simple nor an accurate process. It is usually necessary that the variables to be multiplied should be of *different* kinds and this is one reason for the conversion of analogues.

40. With *direct voltages* it is usual to use a R-pot. One of the variables to be multiplied is applied as a voltage across the R-pot: the other variable is a shaft rotation moving the wiper: the voltage at the wiper then represents the product of the two variables. This is illustrated in Fig. 18. An analogue voltage E is applied across the R-pot whose wiper is moved a fraction x from the zero datum position by the shaft rotation analogue x: the output is then xE, the product of the two input analogues. The output of
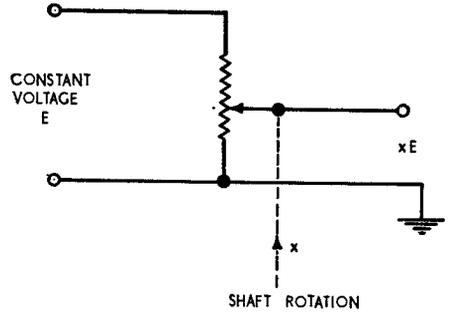
Fig. 18. MULTIPLICATION, DIRECT VOLTAGE SYSTEM.

course cannot exceed E and a change of scale factor by subsequent amplification may be necessary.

41. With *alternating voltages* the principle is similar but it is usual to replace the R-pot by an inductance potentiometer or *I-pot*. An I-pot is, in effect, a variable auto-transformer of the form indicated in Fig. 19. It is wound toroidally with copper wire on a ring of mumetal. Since the arrangement is a transformer, the ratio of output voltage to input voltage is determined by the turns ratio. The input voltage is applied between
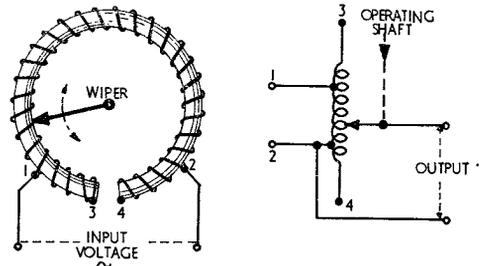
Fig. 19. MULTIPLICATION USING I-POT, ALTERNATING VOLTAGE SYSTEM.

tappings 1 and 2 and as in any auto-transformer, the whole winding is excited even though the input voltage is applied across only part of it. It follows that the output voltage between tapping 2 and the wiper can be *greater* than the input voltage, or even *reversed* in sign if the wiper lies between tappings 2 and 4. This is one great advantage of the I-pot over the R-pot: another is that the accuracy of the I-pot is not affected by loading to the same extent as the R-pot.
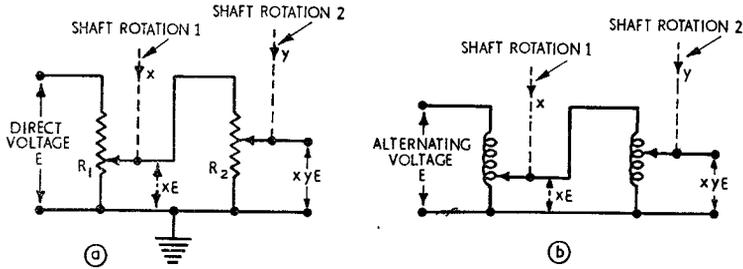
Fig. 20. MULTIPLICATION OF SHAFT ROTATIONS.

42. *Shaft rotations* can be multiplied by converting the mechanical analogues to voltage analogues. A simple system using d.c. voltages and R-pots is illustrated in Fig. 20(a). R_1 has voltage E applied across it and the wiper is rotated through angle x by the first shaft to give an output voltage xE. This voltage is applied across R_2 whose wiper is rotated through angle y by the second shaft to give an output voltage xyE. This voltage is an analogue of the products of the two shaft rotations x and y. Normally, buffer amplifiers would be required between R_1 and R_2 and between R_2 and the output, otherwise the loading on the R-pots would be such that results would be completely inaccurate.

This difficulty is overcome if alternating voltages in conjunction with I-pots are used, as illustrated in Fig. 20(b).

In both cases, the product is in the form of a voltage so that subsequent analogue conversion by means of a servomechanism is necessary if the output is required as a shaft rotation.

Division

43. Division of one analogue quantity by another is obtained in several ways. Some examples are illustrated in Fig. 21.

In (a) the two analogue inputs are E_1 and E_2 and it is required to produce an output voltage proportional to the quotient $\dfrac{E_2}{E_1}$. In a negative feedback amplifier, the closed loop gain A′ is controlled by the feedback, being *inversely* proportional to the feedback voltage. Thus, if the feedback voltage is made *proportional* to E_1 the gain A′ is then *inversely* proportional to E_1, i.e., A′ = $\dfrac{k}{E_1}$, where k is a constant. The output voltage E_0 is equal to A′E_2: that is, $E_0 = \dfrac{k}{E_1} E_2$. The output voltage is therefore proportional to the *quotient* $\dfrac{E_2}{E_1}$.

In (b) an I-pot is used, the analogue inputs being the alternating voltage E applied to the I-pot and the shaft rotation x of the wiper. This arrangement gives an alternating voltage analogue output of the form $E_0 = \dfrac{E}{x}$. If x is a small value, the wiper is almost at the zero datum position and the high step-up ratio gives a large value for E_0. The linearity of the I-pot ensures accuracy. It is necessary to make sure that the output voltage never becomes too great, either because E is too great or x is too small, otherwise the high voltage may damage the I-pot.
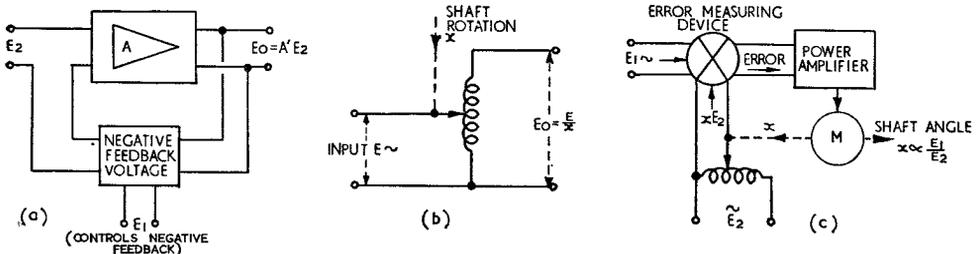
Fig. 21. DIVISION IN ANALOGUE COMPUTERS.

In (c) an I-pot is used in conjunction with a servomechanism. The analogue inputs are the alternating voltages E_1 and E_2 and the output is in the form of a shaft rotation x which is proportional to $\dfrac{E_1}{E_2}$. The analogue E_2 is applied across the I-pot, the wiper of which is controlled by the servo motor. The output from the I-pot is compared in an error-measuring device with the other analogue E_1 and when $E_1 = xE_2$ the error is zero and the motor stops. The shaft has then rotated x degrees which is proportional to $\dfrac{E_1}{E_2}$. Thus a shaft rotation analogue proportional to the *quotient* of two voltage analogues has been produced.

Production of an Analogue Proportional to Rate of Change with Time of an Input

44. It is often required to produce a voltage or a current that is proportional to the rotational speed of a shaft; that is, proportional to the *rate of change* of the *position* of the shaft with *time*. This process is known as *differentiation*.

If a *direct voltage* proportional to the rotational speed of the shaft is required, a d.c. tacho-generator connected to the shaft can produce the required output. Similarly, if an *alternating voltage* analogue is required, an a.c. tacho-generator (drag-cup generator) can be used. A direct current analogue can be produced by a capacity commutator mounted on the shaft. All these methods have been considered previously in paragraphs 18 and 19.
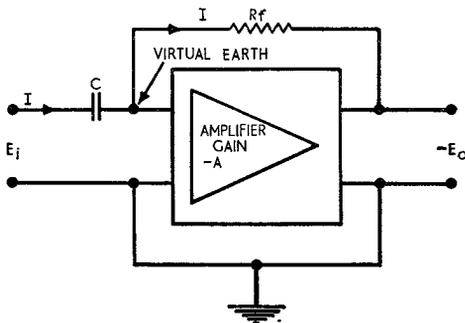
Fig. 22. DIFFERENTIATION IN ANALOGUE COMPUTERS.

45. It may be required to produce an output voltage that is proportional to the rate of change with time of another (input) voltage. This can be obtained with the simple see-saw amplifier arrangement of Fig. 22, in which the input resistor has been replaced by a capacitor of capacitance C.

Due to the normal shunt voltage negative feedback arrangement of the see-saw amplifier, the junction of C and R_f is at virtual earth. There is therefore no current into the amplifier from this junction and the supply current I flows through both C and R_f.

It will be remembered that the current I through a capacitor is the rate of change of charge Q with respect to time. In symbols:—

$$I = \frac{dQ}{dt}.$$

However, the charge Q on the capacitor equals CE_i (remember $Q = CV$) where C is the capacitance (a constant) and E_i is the voltage across it (a variable).

In symbols:—

$$\frac{dQ}{dt} = C\frac{dE_i}{dt}$$

$$\therefore I = C\frac{dE_i}{dt}.$$

Turning now to the feedback resistor R_f, the voltage across this is $-E_0$ (negative because of the reversal of polarity by the amplifier). Thus, the current I through R_f is:—

$$I = -\frac{E_0}{R_f}.$$

Now the current I is the same through both C and R_f.

$$\therefore I = C\frac{dE_i}{dt} = -\frac{E_0}{R_f}.$$

And from this,

$$E_0 = -\,C\,R_f\,\frac{dE_i}{dt}.$$

This relationship shows that the output voltage E_0 is proportional to the rate of change of input voltage E_i with respect to time t $\left(\text{i.e., } E_0 \propto \dfrac{dE_i}{dt}\right)$ multiplied by a constant $(-CR_f)$. This is the requirement.

Analogue of Sum of Instantaneous Changes over a Time Period

46. It is often required to produce an output voltage which changes with time but always represents the sum of all the instantaneous changes occurring in the input analogue quantity during the time interval under consideration. This process is known as *integration*.

Integration is, in effect, the reverse process to differentiation. If a quantity x equals the rate of change of another quantity y with respect to time (i.e., x equals the *differential* of y with time) then y equals the total change in x during this time (i.e., y equals the *integral* of x with time).

For example, on a long journey, a car travels at varying speed and possibly stops for short intervals. The total distance travelled in a given time is found by multiplying the mean speed by the given time interval. This is of course done automatically by the mileage indicator which keeps a continuous record of the distance travelled up to the present time. The distance travelled on a particular journey is found by taking the *difference* between the indicator readings before and after the journey. This is integration.

47. **Miller integrator.** To produce a voltage output which represents the sum of all the instantaneous changes occurring at the input over a given period of time, an *integrating amplifier* can be used. This is a see-saw amplifier in which the feedback resistor has been replaced by a capacitor C as shown in Fig. 23(a). A simplified circuit of this arrangement is illustrated in Fig. 23(b), and is referred to as a *Miller integrator*.

The circuit is so biased that the anode is at a high potential when the input to the grid is zero. If a positive voltage is applied to the grid, the valve current increases and the anode voltage falls. Since the capacitor cannot change its charge instantaneously, this fall of anode potential is fed back via C to the grid, thus returning the grid to earth potential. As the capacitor charges, the grid tends to go positive thereby causing the anode potential to fall. This tendency is counteracted however by the feedback action of C which conveys any *change* of anode potential back to the grid and holds the junction of R_i and C at virtual earth.

48. With the grid at virtual earth, there is no current into the amplifier from the junction of R_i and C so that the supply current I flows through R_i and C in series. This current I through R_i is given simply by $\dfrac{E_i}{R_i}$.

The current through the capacitor is also I, which equals the rate of change of charge Q with respect to time t: in symbols, $I = \dfrac{dQ}{dt}$.

However, $Q = - C E_0$. where C is the capacitance (a constant), E_0 is the voltage across it (a variable) and the negative sign
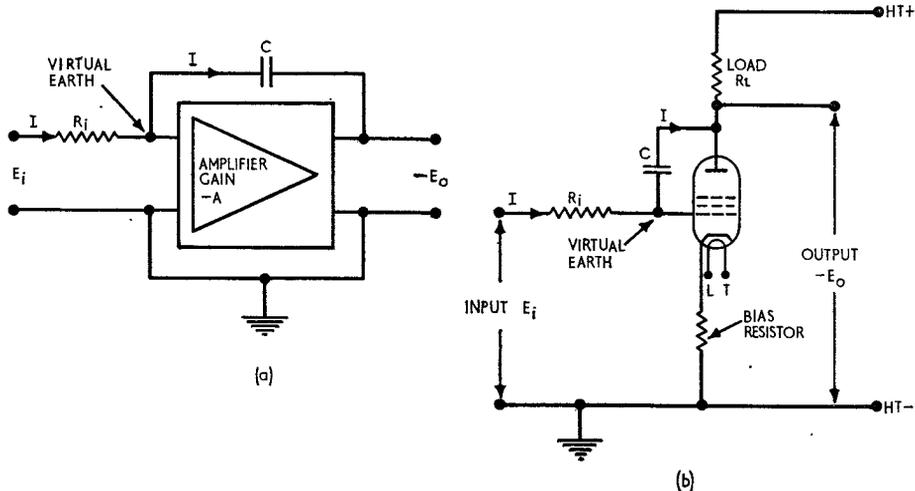
Fig. 23.   THE MILLER INTEGRATOR.

indicates the reversal of polarity by the amplifier. Thus:—

$$\frac{dQ}{dt} = - C\frac{dE_0}{dt}$$

$$\therefore I = - C\frac{dE_0}{dt}.$$

The current through R_i equals that through C and the relationship is given in symbols by:—

$$\frac{E_i}{R_i} = - C\frac{dE_0}{dt}$$

$$\therefore E_i = - C R_i\frac{dE_0}{dt}.$$

It was stated earlier that if x equals the differential of y with time, then y equals the integral of x with time. Since E_i is proportional to the differential of E_0 with time, then E_0 is proportional to the integral of E_i with time: that is:—

$$E_0 = - \left(\frac{E_i}{C\,R_i} \times t\right).$$

As commonly used in computers, E_i is an analogue voltage representing speed. In this case, the output voltage E_0 is equal to a constant $\left(-\dfrac{1}{C\,R_i}\right)$, multiplied by the speed (E_i) multiplied by the time (t) over which integration has been performed. Now, speed x time gives distance travelled. So the output voltage is an analogue of *distance*, in the time of integration, multiplied by a constant.

49. The Miller integrator suffers from the disadvantage that integration can take place over a limited period of time only. Integration proceeds, with C continuing to charge, until the anode potential falls to the level at which 'bottoming' of the valve occurs: no further changes can then take place and integration ceases. Where there is a requirement for continuous integration

over long periods of time, the velodyne integrator is used.

50. **Velodyne integrator.** The velodyne is a velocity servomechanism or 'rate servo' in which the motor is controlled to run at a speed proportional to the input voltage: in essence, it is similar to the speed control system of Sect. 19, Chap. 2, Fig. 3.

The basic block diagram of a velodyne is illustrated in Fig. 24. The input voltage analogue representing the demanded speed is compared in an error-measuring device with the output voltage from a tacho-generator mounted on the output shaft: the feedback voltage from the tacho-generator (d.c. or a.c., depending on the system) is directly proportional to the speed of rotation of the output shaft. Any difference between the voltages representing the demanded speed and the actual speed produces an error signal which, after amplification, accelerates the motor until the feedback voltage is equal and opposite to the input voltage. At this point, the actual speed of the servo motor is the same as that demanded and it will continue to run at this speed until there is a change in input demand. If the input increases, the motor speeds up: if the input decreases, the motor slows down: if the input reverses in sign the motor runs in the reverse direction.

In practice of course there is always a small error signal (velocity lag) because some driving torque is required to overcome the inherent frictional and damping losses in the system: an error signal is therefore needed to produce this torque. However, by using a high gain amplifier, the error can be kept very small.

51. If a velodyne can be made to follow the input speed demand accurately, it can be used as a continuous integrator. This may be seen as follows:—
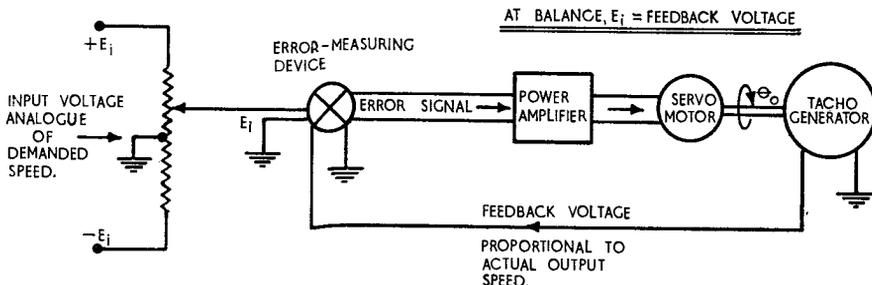
Fig. 24. THE VELODYNE.

At balance, the velodyne is running at the same speed as that demanded: so the input voltage analogue E_i is equal to the feedback from the tacho-generator (neglecting frictional losses). The voltage from the tacho-generator is proportional to the speed of rotation of the output shaft and this, in turn, can be defined as the rate of change of angular position of the output shaft with respect to time: in symbols, $\dfrac{d\theta_0}{dt}$, where θ_0 is the angular position of the output shaft and t is time.

Thus, at balance, input voltage analogue equals feedback voltage.

$$\text{i.e., } E_i = k\frac{d\theta_0}{dt}$$

(where k is a constant).

Since E_i is proportional to the differential of θ_0 with time, then θ_0 is proportional to the time integral of E_i: that is:—

$$\theta_0 = \frac{1}{k}E_i \times t.$$

Thus the angular position of the output shaft at any instant is proportional to the input voltage multiplied by the time for which integration has taken place. It is therefore a measure of the sum of all the instantaneous changes in the input over this period of time.

52. The velodyne integrator can provide reasonably accurate integration without any time limit: that is, unlike the Miller integrator, it can provide continuous integration.

As commonly used in aircraft analogue computers, the input voltage is an analogue of ground speed: the angular position of the output shaft at any instant is then an analogue of the *integral* of ground speed with respect to time, i.e., an analogue of ground distance flown.

Trigonometrical Operations

53. It is shown in Sect. 19, Chap. 1, Para. 51, that a vector can be defined in terms of its polar co-ordinates $r \diagup \theta$, where r is the modulus or length of the vector and θ is the argument or angle it makes to the X axis. This same vector can also be defined in terms of its cartesian co-ordinates $x = r \cos \theta$ and $y = r \sin \theta$.

It is often necessary in analogue computing to convert from one form to the other. For example, given a voltage E and a shaft angle θ, it is sometimes necessary to form voltages proportional to E sin θ and E cos θ: conversion from cartesian to polar form may also be required. These trigonometrical operations are carried out by the components and circuit arrangements discussed in the following paragraphs.

54. **Sine-cosine potentiometers.** For *direct voltages*, a simple slab type sin-cos potentiometer can be used to convert a vector $E \diagup \theta$ into its cartesian form E sin θ and E cos θ. The arrangement is illustrated in Fig. 25.
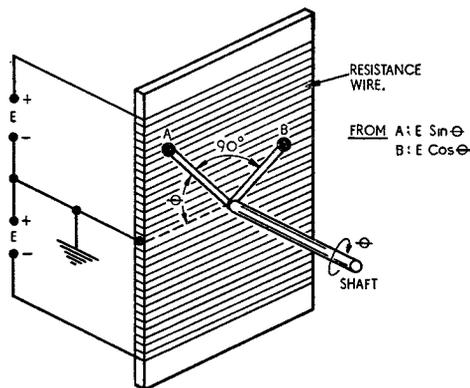
Fig. 25. SLAB TYPE SIN-COS POTENTIOMETER.

Since the resistance winding is uniform, there is a uniform variation in voltage from the bottom to the top of the slab. Therefore the voltage at a wiper is proportional to its linear displacement from the dotted centre line (earth). A supply voltage centre-tapped to earth is required to provide the + and − signs in the output voltage. For wiper A, the displacement is proportional to sin θ; for wiper B, displaced from A by 90°, it is proportional to sin $(\theta + 90°)$ which equals cos θ. Thus from wipers A and B respectively, voltages of E sin θ and E cos θ with respect to earth are obtained.

An alternative form of sin-cos potentiometer is the shaped-former potentiometer or *graded R-pot*. In this, the former on which the resistance wire is wound is so shaped that when a voltage is applied to the R-pot, the wipers pick off voltages proportional to the sine or the cosine of the angle of rotation θ. The arrangement is illustrated in Fig. 26(a). Voltages corresponding to + E and − E, both with respect to earth, are applied across the R-pot as
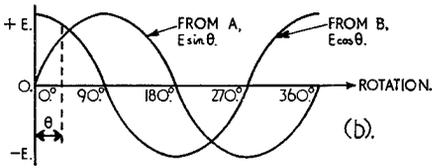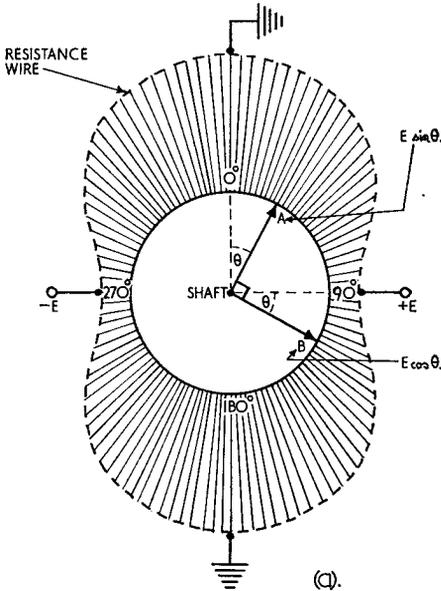
Fig. 26. GRADED SIN-COS R-POT.

Fig. 27—SCOTCH LINK TYPE SIN-COS POTENTIOMETER
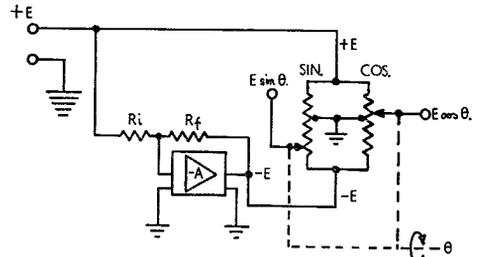
Fig. 28. USE OF PARAPHASE SEE-SAW AMPLIFIER.

shown. Wiper B is displaced from wiper A by 90°, and when the shaft is rotated through the angle θ from the zero datum position, the voltage picked off by wiper A rises sinusoidally from zero: at wiper B, the voltage falls co-sinusoidally. This is illustrated in the graph of Fig. 26(b) which shows the variation in voltage at A and at B through 360° rotation. The output from A is thus of the form E sin θ and that from B is E cos θ. Thus the cartesian co-ordinate of the vector E$\diagup\theta$ has been obtained.

A more accurate form of instrument uses two normal R-pots and derives sin θ and cos θ mechanically by means of a device known as a *Scotch link* (Fig. 27).

In all the arrangements for sin-cos potentiometers discussed in this paragraph, equal and opposite voltages are applied to each end of the potentiometer and the centre is earthed. To obtain these voltages balanced about earth, a *paraphase see-saw*

amplifier is used as shown in Fig. 28. From previous notes on the see-saw amplifier, it will be remembered that when the input resistor R_i is equal to the feedback resistor R_f, the amplifier gain is unity, and phase inversion is also obtained. Thus, if the input to the amplifier is $+ E$ volts with respect to earth, the output is $- E$ volts with respect to earth. These equal voltages of opposite polarity are applied across the sin-cos potentiometer as shown in Fig. 28 and outputs of E sin θ and E cos θ are obtained from a vector input expressed in terms of a voltage of magnitude E and the shaft rotation θ of the wipers.

55. Resolver synchros. For alternating voltages, a synchro resolver of the type discussed in Sect. 19, Chap. 1, Para. 54 can be used to convert an alternating voltage E and a shaft angle θ into corresponding cartesian co-ordinates E sin θ and E cos θ. The arrangement described in Sect. 19 is repeated in Fig. 29. Only one rotor winding is energised, the other being short-circuited
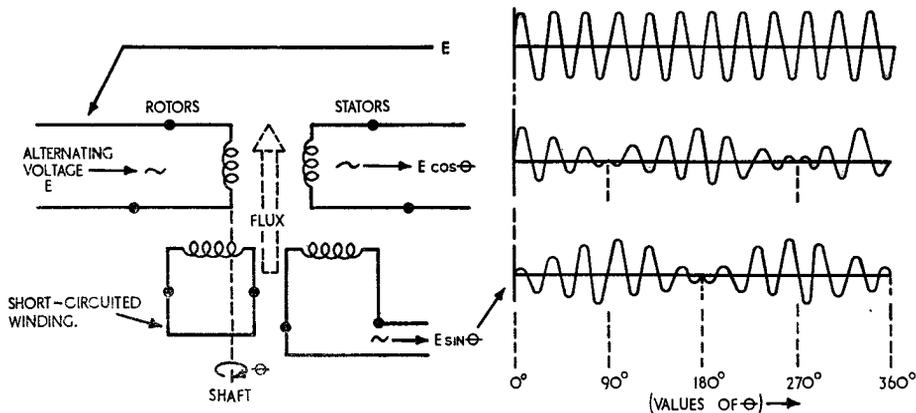
Fig. 29. USE OF RESOLVER SYNCHRO TO GIVE SIN-COS OUTPUT.

to reduce inaccuracies. An alternating voltage E is applied to the rotor winding in use and this is also rotated through the angle θ by the input shaft. The resulting alternating flux induces voltages in the stator windings and the arrangement is such that the output from one winding is an alternating voltage of the form E sin θ and from the other E cos θ.

56. Vector compounding, sin-cos potentiometers. Vector compounding is the name given to the process of converting a vector from its cartesian form to its polar form. To do this, a zero-nulling device (usually a small servo motor) is required. For *direct voltages*, sin-cos potentiometers are used arranged as shown in Fig. 30.

The cartesian co-ordinates of the vector to be compounded are E cos θ and E sin θ.

These input voltages are applied to separate see-saw paraphase amplifiers to give equal and opposite voltages E cos θ and $-$ E cos θ across one sin-cos potentiometer, and E sin θ and $-$ E sin θ across the other.

All the wipers are rotated through the same angle ϕ by a servo motor. The output from wiper A (E cos θ sin ϕ) is combined with the output from wiper D ($-$ E sin θ cos ϕ) at the input of a summing amplifier: this latter provides an output equal to the sum of the two inputs, i.e., equal to (E cos θ sin ϕ $-$ E sin θ cos ϕ) and this is the trigonometrical relationship E sin ($\theta - \phi$).

The voltage E sin ($\theta - \phi$) is amplified and applied as the input to the servo motor which rotates, moving the wipers on the potentiometers and varying the angle ϕ. The rotation is always in such a direction
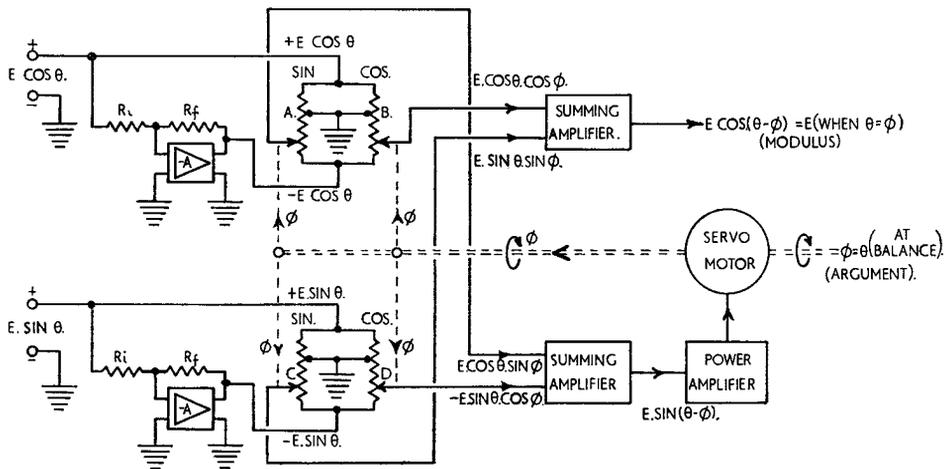
Fig. 30. VECTOR COMPOUNDING, SIN-COS POTENTIOMETERS.

that the error $(\theta - \phi)$ tends towards zero. Thus, when $\phi = \theta$, the error is zero and the voltage E sin $(\theta - \phi)$ is also zero: the motor therefore stops with its output shaft rotated through the angle $\phi = \theta$. The argument or angle θ is therefore known.

The output from wiper B (E cos θ cos ϕ) is combined with the output from wiper C (E sin θ sin ϕ) at the input of the top summing amplifier: this provides an output equal to the sum of the two inputs, i.e., equal to (E cos θ cos ϕ + E sin θ sin ϕ) and this is the trigonometrical relationship E cos $(\theta - \phi)$. Since the servomotor has rotated the wipers in such a direction that $\theta = \phi$, then the output from the top summing amplifier is E cos (0°) which equals E.

Thus the vector whose cartesian co-ordinates are E cos θ and E sin θ has been converted to its polar form: the output from the top summing amplifier represents the modulus E and the shaft position of the servomotor represents the argument θ.

57. Vector compounding, resolver synchros.

For *alternating voltages*, resolver synchros can be used to convert a vector from its cartesian form to its polar form. The principle is discussed in Sept. 19, Chap. 1, Para. 55 and the arrangement used in that discussion is again illustrated in Fig. 31.

The cartesian co-ordinates of the vector to be compounded are represented by the alternating voltages E cos θ and E sin θ. These inputs are applied to the cos and sin stator windings respectively of the resolver synchro. An alternating flux of amplitude and direction dependent upon these voltages is produced inside the synchro.

One of the rotor windings is connected to an amplifier and servo motor which drives the output shaft and also the rotor in
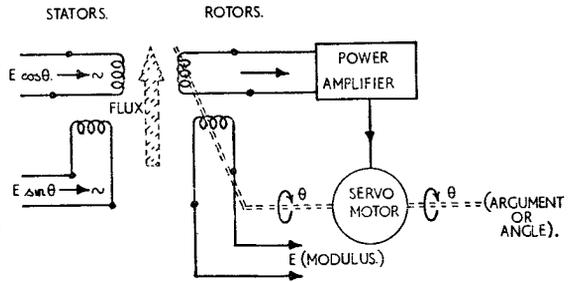
Fig. 31. VECTOR COMPOUNDING, RESOLVER SYNCHROS.

such a direction as to return the rotor to a null position: the motor then stops, having turned the rotor through the angle θ.

The other rotor winding then has induced in it a voltage proportional to the amplitude of the alternating flux, i.e., proportional to the modulus E.

Thus the vector whose cartesian co-ordinates are E cos θ and E sin θ has been converted to its polar form: the output from one rotor winding represents the modulus E and the shaft position of the servo motor represents the argument θ.

58. Typical application.

Fig. 32 illustrates the sort of trigonometrical problem that an analogue computer may be required to handle. It is a simple aircraft navigation problem: the known factors are airspeed and aircraft heading, and windspeed and direction: from these factors the ground-speed and track of the aircraft have to be calculated: the computation must be done accurately and quickly and, in fact, since the input analogues are continuously changing, continuous computation is required.

This is done with the vector resolver and vector compounder circuits discussed in
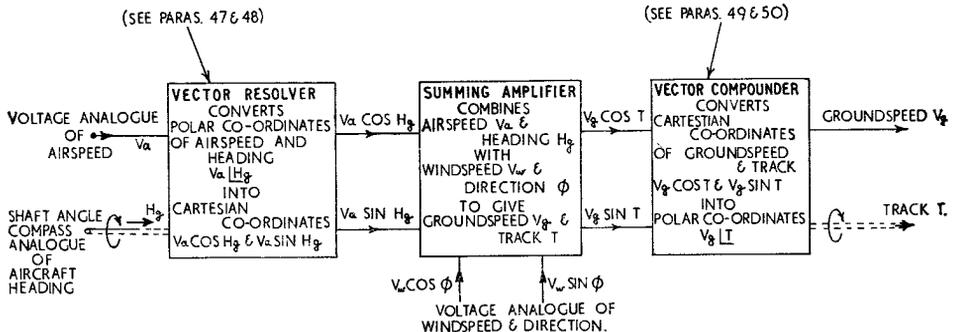
Fig. 32. TYPICAL TRIGONOMETRICAL PROBLEM.

the preceding paragraphs. The resolvers and compounders will consist of sin-cos potentiometer or resolver synchro arrangements depending on whether the system is d.c. or a.c. respectively.

The airspeed and aircraft heading together represent a vector of polar form $V_a \angle H_g$. These are converted to cartesian co-ordinates $V_a \cos H_g$ and $V_a \sin H_g$ in the vector resolver circuit and then added to the cartesian co-ordinates of windspeed and direction, $V_w \cos \phi$ and $V_w \sin \phi$. The resultant computation produces the cartesian co-ordinates of groundspeed and track, namely $V_g \cos T$ and $V_g \sin T$ which are converted in the vector compounder into their polar form $V_g \angle T$. Groundspeed is then available as a voltage analogue V_g and track is available as a shaft angle T on the servomotor.

Analogue Computer Amplifiers

59. It has already been stated that the amplifiers used in analogue computing are circuits having a high inherent gain but employing a large amount of negative feedback to ensure accuracy: the overall gain is therefore fairly small.

The form that the negative feedback takes depends on whether amplifiers with high or low input and output impedances are required. For example, where loading of potentiometers must be prevented, an amplifier with a high input impedance and a low output impedance is required (e.g., a cathode follower). For star point addition or subtraction of several input analogues, an amplifier with a low input impedance and a low output impedance is required (e.g., see-saw summing amplifier).

60. The arrangement of the amplifier also depends on whether the system is a.c. or d.c. If the analogue quantities are alternating voltages or currents, normal RC coupled negative feedback amplifiers can be used.

If the system uses direct voltages or currents as the analogue quantities a problem arises. The amplifiers must be capable of handling very slow changes in the analogue quantities with time, from a few cycles per second down to zero cycles per second. RC coupled amplifiers can no longer be used because of the high reactance of the coupling capacitor at these very low frequencies. The amplifiers used are therefore direct coupled (d.c.) amplifiers. The main problem in d.c. valve amplifiers, as pointed out in Book 2, Section 10, Chapter 1, Para. 25, is that of controlling 'drift': if a d.c. amplifier is adjusted by means of bias arrangements to give zero output for zero input, it will tend to drift slowly from this setting because of changes in component values (e.g., due to heating), valve characteristics and power supplies. The amplifier must be re-set at regular intervals to maintain accuracy.

In this respect, the *magnetic amplifier* has a decided advantage over valve d.c. amplifiers. A magnetic amplifier has operating characteristics that remain constant over long periods of time. It is therefore well suited for amplification of analogue quantities which change slowly with time (see Book 2, Sect. 10, Chap. 4).

In some d.c. analogue computers, the problem of the long term stability of d.c. amplifiers is overcome by converting the d.c. analogues into an a.c. signal, amplifying this through a conventional RC coupled amplifier, and then converting the amplified signal back again to d.c. Amplifiers using this principle are sometimes referred to as 'chopper' amplifiers, the name being derived from the fact that the d.c. input is, in effect, 'chopped' by a switching relay to produce an a.c. signal.

Summary

61. This chapter has shown in broad, general terms what a computer is and how an analogue computer differs from a digital computer.

The chapter has also attempted to show what an analogue is, how analogues are produced and how they are converted from one form to another.

The sort of mathematical calculations that an electronic analogue computer may be required to do have been discussed and some indication has been given of typical applications of the analogue computer.

It must of course be realized that this does no more than give a general introduction to a subject in which new techniques are being constantly introduced and in which rapid expansion is occurring.

Thus it is essential that a technician employed on the maintenance of analogue computers be completely familiar with all circuit details appropriate to that equipment, and the relevant Air Publication should always be consulted.

SECTION 20

CHAPTER 2

DIGITAL COMPUTERS

DIGITAL COMPUTERS

Introduction

1. Chapter 1 of this Section considered in broad general terms the essential differences between analogue and digital computers. It was there stated that in analogue computers the quantities or numbers to be calculated are represented by something else—a voltage, a shaft angle, length along a scale, and so on. In digital computers on the other hand, each number or digit is coded as a succession of signals or pulses and each digit is a separate or '*discrete*' term. The result is that digital computers are more accurate than their analogue counterpart, although in general the increase in accuracy is obtained only at the expense of more complicated equipments.

2. An outline of the basic operation of analogue computers is given in Chapter 1. It is proposed to do the same for digital computers in this Chapter. However it will be appreciated that there is a very wide range of variation in digital computers, depending on the sort of job they do. In fact, most digital computers are designed and built to perform a specific task or set of tasks: this may be data processing at the head office of a large commercial firm, where records and files have to be kept up-to-date, wage rates for employees worked out, and so on. On the other hand, the digital computer may be required in industry where it may form part of an automation process, or it may be used to work out stock-holdings and indicate stores requirements. Digital computers are also used to assist in research at universities and in industry, where the requirement is for a machine to provide accurate and quick answers to complex problems fed to it.

In the Service, digital computers are finding increasing use both in ground installations and in aircraft. On the ground they are used, for example, in conjunction with the long-range radar chains to break down target information and supply it quickly and in a clear and concise form to the master controller so that the necessary action may be taken. Also because of the increasing use being made of sub-miniature components in conjunction with transistors, digital computers can be used in aircraft in automatic pilots and to provide accurate and quick answers to navigational problems.

3. The wide range of applications is easily seen, and there are considerable differences in the physical size, the complexity, power requirements, circuits and arrangements according to the task. Nevertheless, although digital computers differ so much in general they operate on the *same basic principles* and it is these aspects which are to be considered in this Chapter. Naturally with such a subject, it is only the general outline of the scheme of operation of the system and a summary of the function of the main units that can be attempted in these Notes.

Processes in Simple Arithmetic

4. The digital computer can be said to do basic arithmetic operations of add, subtract and so on. To see how it does this, and to discover the units required in a computer and the function of each unit, it is necessary to set down in detail the various processes that a human being performs when he solves a simple arithmetical problem.

Consider first the very simple problem of adding 49 to 37. This is set down as follows:—

$$
\begin{array}{r}
49 \\
37 \\
\hline
86 \\
\hline
\end{array}
$$

To solve it one could say 7 plus 9 equals 16; that is, 6 in the units column and '*carry* 1' to the tens column. Turning to the tens column, the answer there is 3 plus 4 plus '*the carry* 1' equals 8: the combined answer is 86.

Even in this very simple example, there are several well-defined processes which the human being must be capable of carrying out:—

(*a*) He must *plan* the operation and decide on the best method of tackling the problem: even the simple problem given can be solved in several ways.

(*b*) He must ensure that the plan is *executed* in the *correct sequence*.

(*c*) He must be capable of *carrying out* the actual *arithmetical operations* of add, subtract and so on.

(*d*) He must provide the means for '*storing*' numbers either as part of a 'carry' operation or for partial results: in the simple example given, the 6 in the units column was 'stored' either in the person's memory or on a piece of paper until the operation on the tens column was completed, and the final result then obtained.

5. Of these *four* distinct processes, *three* can be 'built in' to an electronic digital computer:—

(*a*) A *control unit* ensures that the predetermined plan is carried out automatically and in the correct sequence.

(*b*) An *arithmetic unit* performs the actual arithmetical operations of add, subtract and so on.

(*c*) A *storage unit* is included in which numbers are stored until required.

The only process that the machine is incapable of performing is that of *planning* the whole operation. This involves thought, and that the machine cannot provide. Therefore a detailed set of instructions or a '*programme*' must be prepared by a human being in which the problem is posed and the machine is 'told' exactly what to do, step by step, in order to solve the problem. These basic processes are shown in block form in Fig. 1.

6. The necessary processes can perhaps be made clearer by considering another simple problem. How would a digital computer cope with the calculation given below?

$$\frac{527}{28} + 45\cdot3(628 + 213).$$

The first essential is to prepare a *programme* in terms that the machine can understand. The human operator must therefore *plan* the whole operation and decide on the sequence to be adopted. The instructions given to the computer in coded form might then run as follows:—

(*a*) perform 527 ÷ 28:

(*b*) store the quotient:

(*c*) perform 628 + 213:

(*d*) store the sum:

(*e*) multiply the sum at (d) by 45·3:

(*f*) store the product:

(*g*) add the product at (*f*) to the quotient stored at (b). This last step provides the answer.

7. The digital computer is therefore provided with a *programme* of instructions: this is fed to the *control unit* which routes the instructions to the other units in the computer in the correct sequence: the actual calculations are performed in the *arithmetic unit:* and *the storage unit* is used for the transfer of numbers to and from the store, according to the instructions received from the control unit.

Elementary Digital Computer

8. A human being performing any calculation conveys the initial problem and data to his mind by one of his senses (often visual): also, he will normally *write down* the final answer in some form. The digital computer must also be provided with a means for converting the input information presented by the programme into terms on which it can work (usually voltage pulses): this requires another unit—the *input unit*. Also, like the human operator, the digital computer must be able to provide the final answer in terms understandable to men: an *output unit* is provided for this purpose.

9. The block diagram of a basic electronic digital computer can be now be given (Fig. 2). The function of each unit is summarized below:—

(*a*) **Input unit.** This is the means by which the required information (the pro-
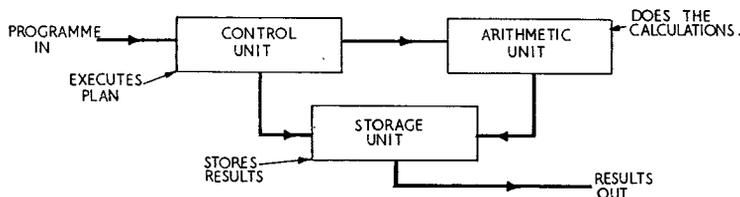
Fig. I. ELEMENTARY PROCESSES IN A DIGITAL COMPUTER.
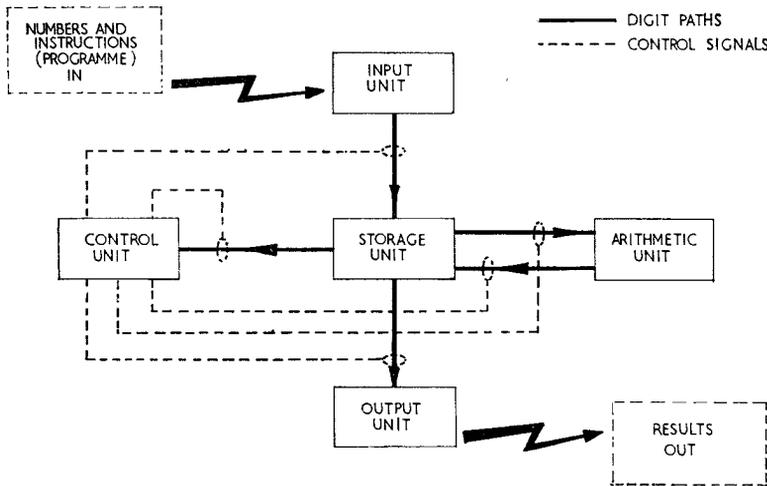
Fig. 2. BLOCK DIAGRAM OF A DIGITAL COMPUTER.

gramme of instructions and numerical data) is communicated from outside and converted into suitable electronic pulse trains.

(b) **Control unit.** The control unit is the directing force of the computer—the automatic controller. It determines the sequence of operations of the computation according to the instructions received and provides the means of communication within the machine. It connects the separate units in the computer into a whole system by arranging for the transfer of information from one unit to another.

Physical connection between the control unit and the other units is by means of wires along which the electrical pulses are carried. The wires form a channel, known as the '*highway*', which has a series of switches or '*gates*' to be opened or shut under the direction of the control unit.

(c) **Storage unit.** All information (i.e., numbers and instructions in the form of numbers) is stored in storage devices. This information is held available for the other units in the computer as required and it is under the direction of the control unit. Many different kinds of devices can be used as storage units, but in digital computers there are two main classes, known respectively as *working store* and *backing store* (including *external storage*). Each has special characteristics depending mainly on the quantity of information stored and on the speed of access to the information.

(d) **Arithmetic unit.** The circuits in the arithmetic unit carry out the arithmetical operations of add, subtract, multiply and divide under the direction of the control unit. During the operation, pulses (representing numbers) are transferred to and from the storage unit as required until the operation is completed. The operation is at all times under the direction of the control unit.

(e) **Output unit.** This is the unit which '*writes down*' the final answer worked out by the computer. The output unit gathers the computed results from the storage unit and delivers them under the direction of the control unit to the processing devices, of which a wide variety are in use (see later paragraphs).

10. The operation of the computer as discussed so far can be summed up as follows. The programme causes all the various numbers representing the instructions, and also those on which calculations are required, to be assembled into the storage unit under the direction of the control unit. The control unit takes the numbers from the store in the correct sequence according to the programme and transfers them to the arithmetic unit where the calculations are carried out. The results are transferred back to storage and the programme, through the control unit, transfers the output data from storage to the output unit where it is printed or recorded as required.

This gives a broad picture, but to get a reasonable understanding of the operation

of digital computers a more detailed examination is necessary.

Disadvantages of Decimal Scale for Digital Computers

11. In an electronic digital computer, the calculations are carried out in the arithmetic unit which consists of a number of circuits using electronic devices. These devices may be thermionic valves, transistors, semiconductor diodes, cold-cathode diodes and so on.

The arithmetical scale in everyday use is a *decimal scale* where numbers are expressed on a scale-of-ten system: that is, there are ten basic digits (0 to 9) and the same digits are used over again after the tenth (i.e., after 9). If, say, a valve is being used to count on this scale there would have to be ten different distinguishable levels of anode current: for example, 0 mA made equivalent to an input of 0, 1 mA to an input of 1, 2 mA to an input of 2 and so on. Such a system would be extremely difficult to operate and would require stabilized power supplies, negative feedback arrangements and careful design to avoid drift and resultant inaccuracies. In mathematical calculations, precision is all-important, and it is taking a risk to expect the circuit to distinguish infallibly between *ten* different magnitudes of current. In practice manufacturing tolerances in components make this difficult especially if the current is of the order of low milliamperes or microamperes: accuracy in such cases requires precision components and high cost.

12. Where accuracy is important, there are only *two* possible states in an electronic device on which complete reliance can be placed: these states are ON and OFF. So long as the circuit conditions are such that the electronic device can be considered to be off or on, variations in the characteristics of the components or in the power supplies are of no importance. On the 'on-off' basis, the system is as reliable as it possibly can be: and this is where the scale-of-two, or the *binary scale*, comes in.

Binary Numbers

13. The binary scale has only two basic digits, 0 and 1: its whole scale can therefore be fully represented by the two conditions

'off' and 'on': that is, when the electronic device is *off* it represents the digit 0, and when it is *on* it represents the digit 1.

For any arithmetical scale, the term *'radix value'* is used to denote the number on which the system is based. In the decimal system the radix is 10: in the binary system it is 2. In the decimal scale where 5 is *half* the radix, 5 + 5 equals the radix 10. In effect, the *'carry'* point (i.e., 0 carry 1, written as 10) at which the same symbols are used over again occurs *at the radix*.

Similar remarks apply for the binary scale. *Nought* is denoted by the symbol 0, and *one* is denoted by the symbol 1 (which is *half* the radix on this scale). There is no symbol 2 in the binary scale and *one* added to *one* causes a *'carry'* to occur because half the radix is being added to half the radix. Thus, in the same way as 5 + 5 is written as '0, carry 1' (or more conventionally, '10') on the decimal scale, so 1 + 1 is written as 10 on the binary scale. Therefore the binary numbers 0, 1 and 10 correspond to the decimal numbers 0, 1 and 2.

14. On the decimal scale (radix 10) the value of a number is denoted by the columns in which the various digits appear. These columns are based on powers of 10 and are of course the familiar columns of *units* (10^0), *tens* (10^1), *hundreds* (10^2), *thousands* (10^3) and so on. Thus the number 77 in decimals means 7 tens plus 7 units: that is, $(7 \times 10^1) + (7 \times 10^0)$. Note that any radix to the power 0 is 1.

On the binary scale (radix 2) the columns in which the two digits 0 and 1 appear are based on *powers of* 2 and each column has *double* the value of the previous one. The columns are shown in Table 1, which also shows the relationship between decimal and binary numbers.

15. Any number can be given in terms of binary digits. The number of digits used to represent a number is much greater with the binary scale than with decimals: also, the binary scale is much less familiar than the decimal scale. Nevertheless, because electronic devices give their most reliable operation as *two-state* devices, the practical advantages to be gained by using binary digits are so great that the system is almost universally applied.

POWERS OF 2

DECIMAL NUMBER	$2^6=64$	$2^5=32$	$2^4=16$	$2^3=8$	$2^2=4$	$2^1=2$	$2^0=1$	BINARY EQUIVALENT
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	1
2	0	0	0	0	0	1	0	10
3	0	0	0	0	0	1	1	11
4	0	0	0	0	1	0	0	100
5	0	0	0	0	1	0	1	101
6	0	0	0	0	1	1	0	110
7	0	0	0	0	1	1	1	111
8	0	0	0	1	0	0	0	1000
9	0	0	0	1	0	0	1	1001
10	0	0	0	1	0	1	0	1010
11	0	0	0	1	0	1	1	1011
12	0	0	0	1	1	0	0	1100
13	0	0	0	1	1	0	1	1101
14	0	0	0	1	1	1	0	1110
15	0	0	0	1	1	1	1	1111
16	0	0	1	0	0	0	0	10000
⋮ 77 ⋮	1	0	0	1	1	0	1	1001101

TABLE 1. BINARY NUMBERS.

Binary Arithmetic

16. **Addition.** The rules of addition with binary digits are similar to those used with decimals. In binary, of course, there are only two symbols (0 and 1) so that:—

$0 + 0 = 0$: $0 + 1 = 1$:

$1 + 0 = 1$: $1 + 1 = 0$, *carry* 1 (or 10).

For example, from Table 1 the binary equivalent of the decimal number 31 can be worked out; it is 11111: similarly decimal 21 equals binary 10101. To add these two numbers in the binary scale, they are set down as shown below and the rules of addition applied:—

$$
\begin{array}{r}
11111 \\
+ \ 10101 \\
\end{array}
$$

Carry digit: 11111

Sum: 110100

Starting at the right-hand (units) column, $1 + 1 = 0$, carry 1:

for the second column,
$0 + 1 +$ carried $1 = 0$, carry 1:

for the third column,
$1 + 1 +$ carried $1 = 1$, carry 1:

for the fourth column,
$0 + 1 +$ carried $1 = 0$, carry 1:

for the fifth column,
$1 + 1 +$ carried $1 = 1$, carry 1:

for the sixth column, the carried 1.

This gives the sum **110100**, the value of which in decimals can be calculated from Table 1: it is $32 + 16 + 4 = 52$ in decimal

$= 31 + 21$ in decimal.

17. **Subtraction.** The subtraction of binary numbers is similar to decimal subtraction: the rules in binary are:—

$0 - 0 = 0$: $1 - 0 = 1$:

$1 - 1 = 0$: $0 - 1 = 1$, carry 1.

The last rule is the same as that adopted for the subtraction $0 - 9$ in decimal notation, namely $0 - 9 = 1$, with 1 carried to the next column.

For example, $27 - 22$ is written as follows in binary notation:—

```
   11011
 − 10110
```

Carry digit: 1

Result: $00101 = 5$ in decimal notation.

However in electronic digital computers, *adding* circuits are much easier to construct than circuits which subtract. Therefore it is usual to convert the number that is to be subtracted into what is known as its *complement* and to use this for *addition*.

Taking the above example $(27 - 22)$, this is written in binary as:—

```
   11011
 − 10110
```

The complement of the negative number is obtained simply by *reversing* all its digits to give 01001, i.e., writing 1 where there is a 0, and 0 where there is a 1. The procedure is then as follows:—

```
 27 ──►  11011 ─────────────────────►  11011
−22 ──► −10110 ──► Complement ──► + 01001
                        Add ──► (1) 00100
"Surplus" digit carried to units column ──┘ ──► 1
                        Add ──►    00101 = 5 in decimal
```
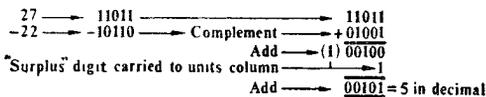
In practice it is a simple matter to obtain the complement by connecting an '*inverter*' in one input of the adding circuit, thereby reversing all the digits of the negative number. Transferring or *shifting* the 'surplus' digit from the left to the units column and adding it is equally straight forward.

18. **Multiplication.** The process of multiplication can be performed by *repeated addition* of the number that is to be multiplied. For example, to multiply $5 \times 4 = 20$, the binary equivalent of 5 (0101) is applied to the adding circuit and the circuit is made to perform *four* addition cycles:—

```
0101 + 0101 + 0101 + 0101 =  0101      .. (1)
                             0101      .. (2)
                             ────
                             1010
                             0101      .. (3)
                             ────
                             1111
                             0101      .. (4)
                             ────
                            10100 = 20 in
                             ────     decimal
```

A more common method of multiplication in digital computers is based on the fact that moving the digits of a binary number one place to the *left* multiplies that number by *two*. Thus, 0101 (or 5 in decimal) moved one place to the left becomes 1010 (or 10 in decimal). Multiplication in the binary system can therefore be reduced to the operation of *adding*, combined with *shifting to the left*. Taking the example 9×6 (or 1001×0110 in binary) this is worked out in long multiplication as follows:—

```
                       1001
                       0110

                    ┌─ 0000
Moved to the left←──┤─  1001
                    │  1001
                    └_0000

                   0110110 = 54 in
                    ─────     decimal.
```

This example shows that the operation consists of successively moving the number that is being multiplied one place to the left for every digit in the multiplier and adding where the multiplier digit is a 1 and *not* adding where the multiplier digit is a 0.

This is a comparatively easy extension of the basic adding circuit. In the computer, the numbers are usually represented by electrical pulses and the arithmetical process of 'moving one place to the left' is carried out by '*shift*' circuits (see later paragraphs).

19. **Division.** Division is very much the reverse of multiplication, and just as there are two methods of multiplication so there are two methods of division.

In the first method, one number is *repeatedly subtracted* from the other and the number of times a subtraction is successfully achieved is counted. For example, the decimal $12 \div 3$ is performed as $12 - 3 - 3 - 3 - 3$, the answer being *four* subtractions of 3. The action is stopped as soon as 12 is reduced to zero, or in the case of a number that cannot be divided perfectly, when the answer becomes negative.

In the second and more common method, just as multiplication is achieved by shifts to the *left* followed by *addition*, so division is by shifts to the *right* followed by *subtractions*. In practice, of course, the complement method will also be used to turn subtraction into addition.

20. **Summary.** Almost all practical mathematical operations may be converted into *arithmetical* operations, and as shown in the preceding paragraphs, all arithmetical operations can be considered to be extensions of addition. Thus a device which adds can with a few modifications, be converted into a device which can also do most mathematical operations.

The basic operations that the arithmetic unit of a digital computer must perform are therefore *addition, shift* and the *production of complements*. All calculations are broken down by the programme so that only these operations are used.

Representation of Binary Digits

21. Because only two symbols (0 and 1) are used in binary notation, binary digits (or '*bits*') can be represented by any of the conditions 'on-off', 'go-nogo', 'positive-zero', 'negative-zero', 'present-not present', and so on. Bits can therefore be represented electrically in a computer as a voltage being either present or absent at some point in a circuit, or by a switch being either closed or open, or by any other *two-state* system.

There are two main forms of representation —*dynamic* and *static*—and these are considered in the following paragraphs.

Dynamic Representation of Bits

22. In digital computers, binary numbers are required to be represented in dynamic form as a series of voltage pulses occurring at precise instants of time. This is the way by which numbers are transferred from one unit to another inside the computer. The master voltage pulses (called '*clock pulses*') are derived from a pulse generator which is accurately controlled in frequency (and hence in *time*) usually by a crystal oscillator.

The number of pulses occuring in a given period depends on the magnitudes of the numbers to be handled by the machine (generally referred to as the '*word length*'). Many digital computers work with a word length of 32 binary digits so that in the required period (called the '*number period*') 32 pulses are produced. The 32 pulses are repeated cyclically and a new number can be represented during each repetition.
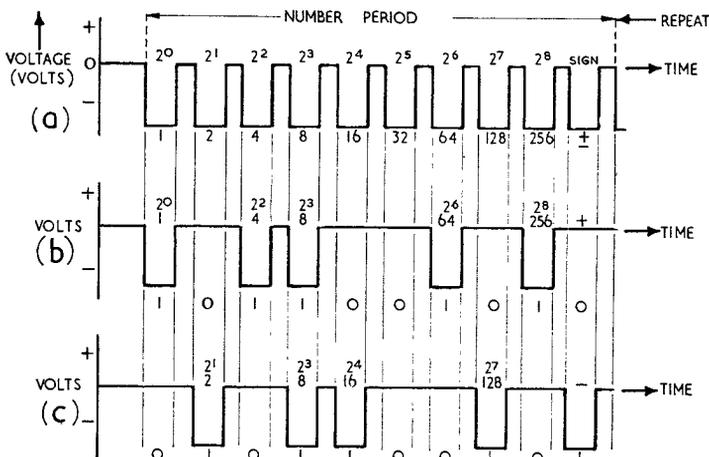
Fig. 3. DYNAMIC REPRESENTATION OF BINARY DIGITS.

23. A word length of 32 bits is not convenient to illustrate. However, Fig. 3(a) shows how a 10 *bit word* can be represented. *Ten* negative going pulses occur in the number period. Each succeeding pulse in time (starting from the *left*) represents a higher power of 2 up to 2^8: the final pulse in the number period is often used as a '*sign*' digit to indicate whether the binary number is positive or negative.

A binary number can be represented by the presence or absence of a pulse at the appropriate intervals of time. Thus a negative-going pulse represents the bit 1 and the absence of a pulse, the bit 0: for the sign digit, the bit 1 (a pulse) is used to indicate a *negative* number, and 0 (no pulse) indicates a *positive* number.

It will be noted that in this method of representation, the least significant digit (i.e., the 2^0 or units column) appears on the *left* instead of on the right as it would be if the number were *written down*. In fact, of course, the digits are *not* reversed on the time scale because the least significant digit appears first in *time;* so that if this series of pulses were applied to a circuit, the least significant digit would be applied *first*. The pulses would appear the right way round if the pulse train is displayed, for example, on a c.r.t.

Thus in Fig. 3(b), the pulses represent the binary number + 101001101 (333 in decimal) and Fig. 3(c) illustrates the binary number − 010011010 (− 154 in decimal). The highest number that can be represented by ten pulses in a number period, one pulse being the sign pulse, is + 111111111 (511 in decimal). *Twenty* pulses allow for numbers up to more than a million, and *thirty-two* pulses for numbers up to 10^{10}.

Static Representation of Bits

24. In a digital computer it is also required to store or register a binary number in static form at various stages of the computation. Thus the binary number 1101 (13 decimal) can be registered in static form by a row of four switches as shown in Fig. 4, where the presence of a voltage +V represents a 1 bit and zero voltage a 0 bit. In practice, the switches could be replaced by a set of electromagnetic relays with holding contacts. Computers with this form of storage have been used in the past but where high-speed operation is required some form of electronic device is used. The electronic device must have *two stable states*, to represent the 0 or the 1, and it must be capable of holding either state indefinitely. Many such two-state (or *bistable*) devices are used for storage of binary numbers: the more common are considered in the following paragraphs.

Storage Devices

25. **Eccles-Jordan trigger circuit.** Probably one of the best known bistable electronic circuits is the Eccles-Jordan, the basic circuit of which is illustrated in Fig. 5. The circuit is a form of multivibrator but the time constants of the coupling components and the biasing arrangements are such that, unlike the multivibrator discussed in Book 2, Sect. 12, Chap. 3, it is *not* 'free-running'. The circuit is stable with *either*
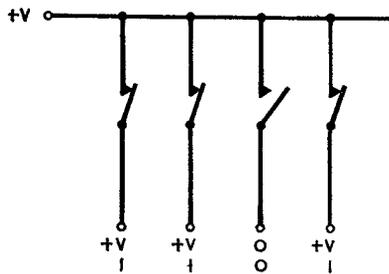
Fig. 4. STATIC REPRESENTATION OF BINARY DIGITS.
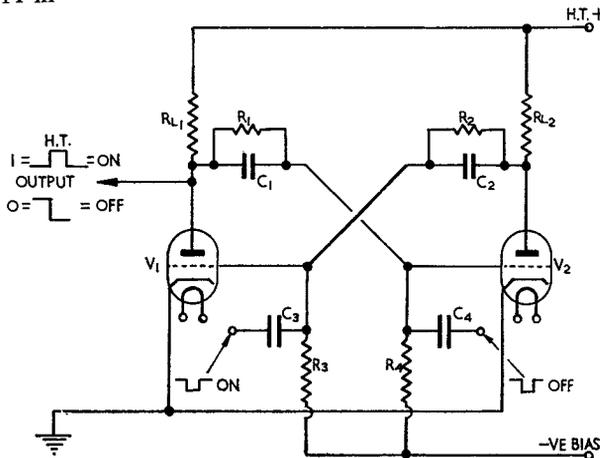
Fig. 5. ECCLES-JORDAN TRIGGER CIRCUIT.

V_1 conducting *or* V_2 conducting: transition from one state to the other is obtained by applying negative pulses to the grids.

Assume V_1 is conducting: then a negative pulse applied to the terminal marked ON drives V_1 grid negative, effectively cutting this valve off: its anode potential rises to h.t. potential as does the output voltage: at the same time this rise in voltage is transferred through C_1 and drives V_2 into conduction. The anode potential of V_2 falls to a low value and this fall in voltage is applied through C_2 to V_1 grid to hold this valve cut off and maintain the stable state.

Applying a negative pulse to the terminal marked OFF cuts off V_2 and the circuit is unbalanced into the other stable state where V_2 is cut off and V_1 is conducting heavily: V_1 anode potential now falls to a very low value as does the output voltage.

Thus when the circuit is triggered *ON* it registers a 1 *digit* as a *high voltage* at the output and when it is triggered *OFF* it registers a 0 *digit* as a voltage approaching *earth* potential. Either stable state can of course be maintained indefinitely, thereby providing a means of storing a binary digit. To store a binary number of, say, 32 bits then thirty-two such circuits are required. Such an arrangement is usually referred to as a '*trigger store*' or '*trigger register*'.

26. Transistor bistable circuit. A typical basic bistable circuit using p-n-p transistors is illustrated in Fig. 6. The two stable states of this circuit are defined by the condition of the transistor TR_1. The circuit is said to be ON when TR_1 conducts (TR_2 cut off) and it is OFF when TR_1 is cut off (TR_2 conducting). The circuit is triggered ON by a *negative* pulse applied to the *base* of TR_1 and it is triggered OFF by a *positive* pulse to TR_1 base.
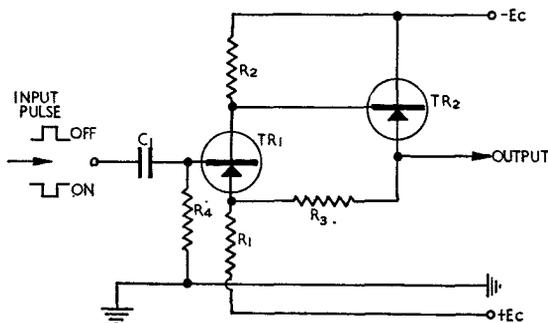
Fig. 6. TRANSISTOR BISTABLE CIRCUIT.

Assume TR_1 is conducting (i.e., the circuit is ON): the potential difference between TR_1 collector and emitter is then very small and the resulting current flowing in the resistor R_3 and in the emitter base circuit of TR_2 is so small as to be negligible. Also when TR_1 is conducting, its emitter potential is approximately zero because the base of the transitor is returned to earth through the low value resistor R_4. The output therefore under this condition is *zero*.

When the transistor TR_1 is cut off by the application of a *positive* pulse to its base, its collector potential is equal to the supply voltage $-E_c$. The base of TR_2 is also at this potential so that TR_2 is biased in the forward direction and will therefore conduct. Because TR_2 is conducting, its emitter-to-base potential will be small and its emitter is also at approximately the supply voltage $-E_c$. The output under this condition is therefore $-E_c$ *volts*.

Thus the application of a *negative* pulse to TR_1 base gives *zero* output and the bit 0 can be said to be stored: a *positive* pulse to TR_1 base gives an output of $-E_c$ *volts* and the bit 1 can be said to be stored or registered. Either state can, of course, be maintained indefinitely.

There are extensions to this basic circuit but enough has been said to show how the binary digits 0 or 1 can be stored in static form. For storage of a binary number, several such circuits are required, the number depending on the word length. The arrangement is then called a *register*.

27. Magnetic tape storage. This is a popular method of storing binary digits in static form. It is probably one of the easiest methods to understand because it is very similar to that used in domestic tape recorders.

The recording is done by applying digit pulses to the energizing coil of a recording (writing) head as shown in Fig. 7(a). The pole pieces of the head are arranged so that the small magnetic fields induced in the magnetic surface of the tape have their flux lines running along the *length* of the tape: thus the tape can accommodate a number of parallel tracks.

The digit pulses are applied to the head as an a.c. square waveform (Fig. 7(b)) so that the 1 pulses produce currents in one direction in the energizing coil and the 0 pulses produce currents in the opposite
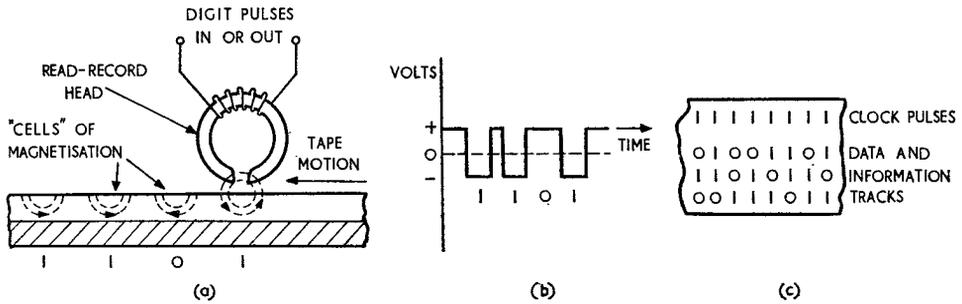
Fig. 7. MAGNETIC TAPE STORAGE.

direction. Each current pulse magnetizes the tape to saturation and as the tape moves past the head, small 'cells' of magnetization are induced, the polarity depending on the direction of the current pulse. Magnetization in one direction represents a stored 1 bit, and in the other direction a 0 bit.

Reading out is done by drawing the magnetized tape across the pole pieces of a similar head (the reading head). The coil then has voltage pulses induced in it in response to the changes in the flux linking the pole pieces. This converts the static representation of bits into dynamic form.

The magnetic tape storage system has virtually unlimited capacity. One commercial system uses a 2400-feet tape which stores 1½ million binary digits. It is also easy to erase information and substitute new data as required. However in spite of tape speeds as high as 100 feet per second,

it takes a relatively long time (in terms of computer time) to locate and extract specific information on the spool. Thus magnetic tape storage is said to have a poor 'access' time, and is used mainly as external storage.

28. **Magnetic drum storage.** The magnetic drum system of storing binary numbers in static form operates on the same principle as that used for magnetic tape. The drum (typically 12 in. long and 6 in. in diameter) is coated with magnetic iron oxide and is caused to rotate past reading and writing heads at speeds of up to 6000 r.p.m. The arrangement is shown in Fig. 8.

Magnetic fields of polarity depending on the binary digits 0 and 1, are impressed on the magnetic coating of the drum by the recording heads: when reading out (and converting from static form into dynamic form) voltage pulses are induced in the reading heads, the direction of the voltage
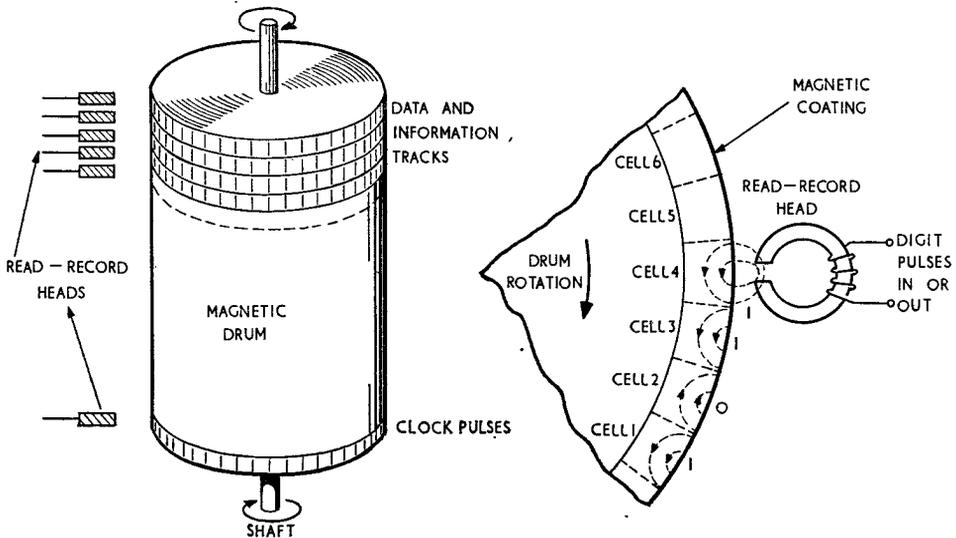
Fig. 8. MAGNETIC DRUM STORAGE.

pulses (representing 1 or 0) depending on the polarity of the magnetic field.

The drum is run at constant speed and each 'cell' of magnetization is presented to the read/write heads with a frequency which coincides with the clock pulses controlling the computer. In this way, *synchronization* is achieved.

Like the magnetic tape system, the drum has a large storage capacity but with a drum the access time is shorter. It is used mainly for external storage.

29. Ferrite core storage. Small ferrite cores can be used to store binary digits because they possess two distinct states of *remanent magnetism*. One core is required for each digit in a binary number and each core is provided with at least three windings as shown in Fig. 9(a): one winding is used to magnetize the core 'positively', another

(a)

(b)

Fig. 9. FERRITE CORE STORAGE.

magnetizes it negatively and the third winding is an output winding which has a voltage induced in it by the change of flux when the core changes its state of magnetization.

Ferrite possesses an almost rectangular hysteresis loop as shown in Fig. 9(b). If a

current is applied to winding A in a 'positive' direction such that the core becomes magnetically saturated to $+ B_m$, then the remanent magnetism ($+ B_r$) when the current is disconnected is *not much less* than $+ B_m$. Similar remarks apply to $- B_m$ and $- B_r$ when the core is magnetized in the other direction. Thus these two distinct states of remanence ($+ B_r$ and $- B_r$) can be used to represent the binary digits 1 and 0 respectively.

To read out the contents of the ferrite store it is necessary to re-apply a magnetizing force. This is done by applying a pulse representing a 1 or a 0 to winding B. Assume a 1 pulse is applied: if a 1 is *already* stored in the core there is very little flux change ($+ B_r$ to $+ B_m$ only) and the voltage induced in the output coil C is negligible. If however the core is storing a 0, the flux change is considerable ($- B_r$ to $+ B_m$) and a voltage pulse is induced in winding C. Thus if a 1 pulse is applied to determine the state of the core, an output pulse indicates a stored 0 and no pulse indicates a stored 1. It will be noticed that read-out is destructive in that it can reverse the state of magnetization of the core. However this bistable action means that ferrite cores can be used in a manner similar to that of an Eccles-Jordan trigger or transistor bistable circuit. In practice the current output of the ferrite core is amplified by transistors.

Where a large number of binary digits are to be stored in ferrite cores, the cores are arranged in a *matrix system* as illustrated in Fig. 10. The matrix consists of rows of horizontal and vertical wires with tiny ferrite cores threaded on to them so as to enclose all the intersections. Each of the wires forms a one-turn magnetizing winding where it passes through a core and it requires currents through *both wires* at an intersection to take the relevant core to the remanent magnetic condition. Current in one wire only is insufficient to do this. Thus it is possible to store a binary digit, 0 or 1, on any of the cores by applying current pulses to the two wires which intersect at that core: whether a 0 or a 1 is stored depends on the direction of the current as indicated in Fig. 10. Read-out of the stored information is also achieved by applying current pulses to the appropriate pairs of wires. If a given core is in the 1 state, the application of a 0 pulse to the two wires
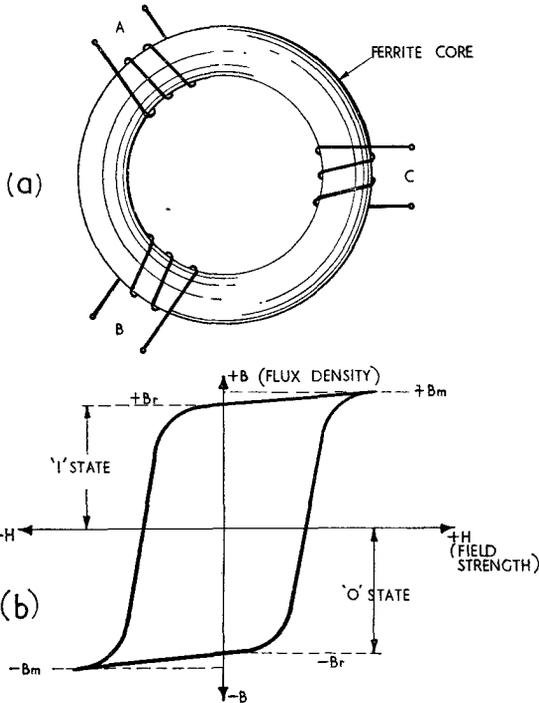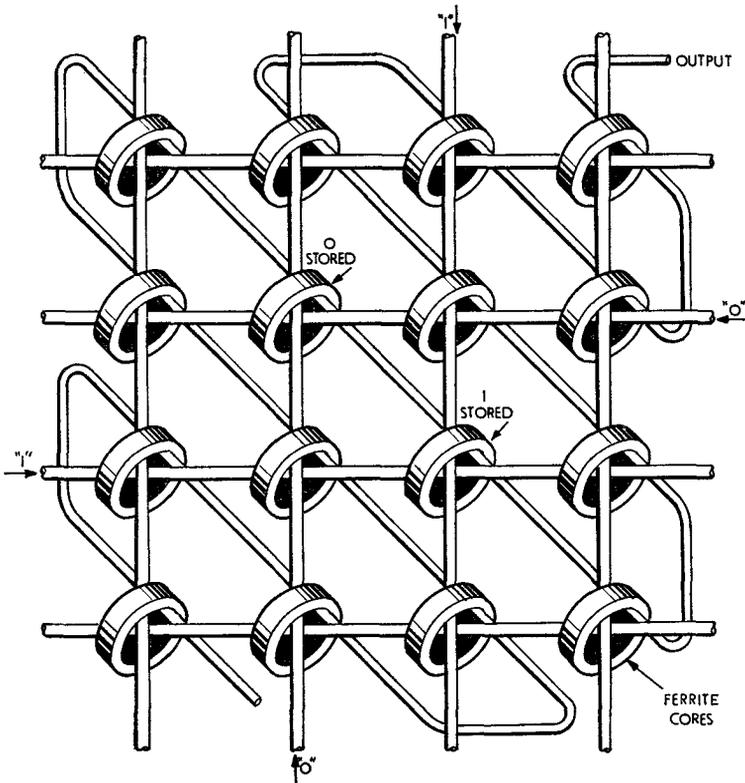
Fig. 10. FERRITE MATRIX STORE.

intersecting at that core causes the core to be switched over, and the resultant change of flux induces a pulse in the output wire. If of course a 0 had been stored, no output would have been obtained.

Using very small ferrite cores, the storage capacity can be made very large: one commercial system uses 10,000 cores. Access time to the stored information is also very short. The main disadvantage is that rather heavy currents are required through the wires to switch the cores from one state of magnetization to the other, so that the incoming digit pulses have first to be amplified.

Ferrite cores are used as temporary storage units or registers in a computer and can also be used as the backing-up store.

30. Another type of ferrite storage system uses *magnetic film* as the storage medium. Small spots of magnetic material, a few millimetres in diameter and very thin, are deposited in regular arrays on a glass base. Each spot acts in much the same way as a ferrite core in the matrix type of store (Fig.

10). The magnetic film has a rectangular hysteresis loop and it can be switched from one direction of magnetization to the other by currents passing through adjacent conductors which are *printed* on both sides of the glass base. Coincident-current methods can be used for the driving system in much the same way as currents through two wires intersecting at a core provide switching of a ferrite matrix store. Compared with the ferrite toroidal core system, magnetic film storage devices are much smaller and have a more rapid access time.

31. **Ferroelectric storage cells.** The ferroelectric cell provides a means of storage in much the same way as a ferrite core: that is, it is based on a rectangular hysteresis loop, this time in a *dielectric*. Certain dielectrics, such as barium and strontium, display marked hysteresis effects and a capacitor using such a dielectric can be used to store a binary digit. A typical circuit arrangement capable of storing one bit is illustrated in Fig. 11(a).

When a positive voltage pulse is written in to the cell, the capacitor is left with a

(a)

(b)

(c)

Fig. 11. FERROELECTRIC STORAGE CELLS.

positive remanent induced charge, indicating that a 1 digit is stored (Fig. 11(b)).

To read out the information, a *negative* pulse is applied to the other terminal. When a 1 is stored, the reversal of polarity caused by the read-out pulse produces a negative pulse at the output: if a 0 had been stored, the read-out pulse would have produced no output.

Many of these cells may be incorporated in a matrix of electrodes separated by a sheet of ferroelectric dielectric (Fig. 11(c)). This is similar to the ferrite core matrix: thus both have large storage capacities, both have quick access times and in both the information is destroyed on reading out. However, unlike the ferrite matrix, the ferroelectric system requires *very little power* for operation.

32. Cryotron storage devices. The cryotron is based on the phenomenon of superconductivity, or zero electrical resistance, which occurs in some metals (e.g., niobium) at extremely low temperatures. The specific resistance of such a metal falls linearly with decrease in temperature until a critical temperature is reached (normally a few degrees above 0° Absolute): at this critical point the specific resistance falls abruptly to zero and the material is superconductive.

The temperature at which this occurs for a given material is affected by the presence of a magnetic field. If a magnetic field is present, the temperature must be *further reduced* before the material becomes superconductive. This is illustrated in Fig. 12(a).

Thus if the superconductor is held at a *constant low temperature* at which it is 'normally' superconductive, then the application of a magnetic field causes the material to be switched to a resistive state. This is illustrated in Fig. 12 (b). This gives a bistable device suitable for storing binary
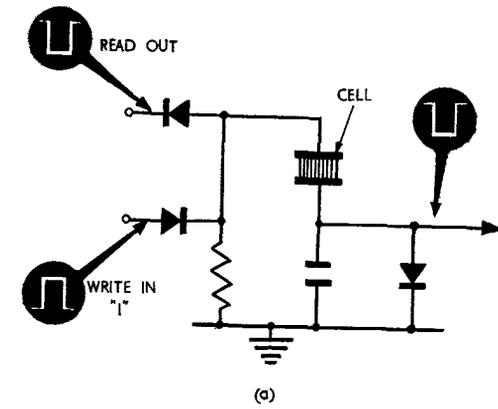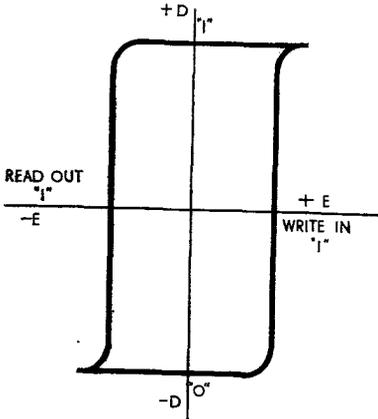
(a)
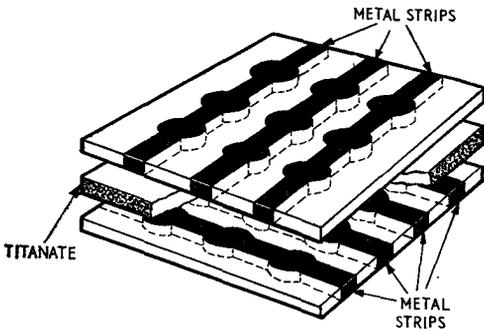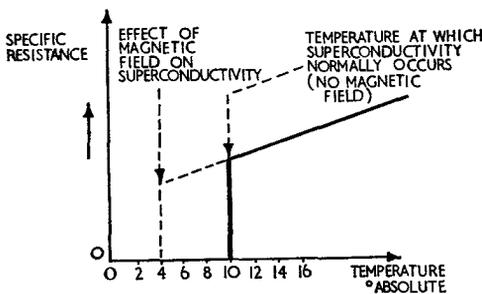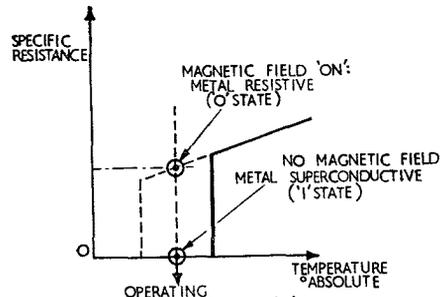
(b)

Fig. 12. CRYOTRON PRINCIPLE.

digits, the magnetic field being used to switch from one state to the other: in the superconductive state a 1 is said to be stored: in the resistive state a 0 is stored.

The cryotron takes the form of a short wire core of suitable material carrying a small coil: it has the advantages of extremely small size, simple construction, low power dissipation and very fast switching speeds. A bistable circuit using two cryotrons is shown in Fig. 13. If a current is established
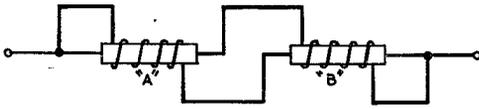
Fig. 13.   CRYOTRON CIRCUIT.

in core A (a 1 state) it passes through the coil of B, driving B into normal resistivity (a 0 state). This condition is maintained (because the coils themselves are superconductive) until a pulse of current is applied to core B. This switches the bistable circuit. It will be realized of course that the cryotron must be held at a *very low temperature.*

33.   **Delay line storage devices.** The storage devices considered in the preceding paragraphs stored the binary digits in *static* form. Delay line storage devices, on the other hand, store binary digits in *dynamic* form: they do this by causing the train of

digit pulses to be *circulated* in a loop consisting of the delay device with its output connected back to its input. The general arrangement is illustrated in Fig. 14.

Pulses of current representing the digits 0 or 1 are converted into mechanical pulses by the sending circuit and applied into the delay line. After a time equal to the period of the complete pulse train (i.e., equal to the *number period*) the mechanical pulses reach a receiving circuit which converts them back into electrical pulses. After amplification, the pulses are used to '*gate*' fresh pulses from the clock generator back to the sending circuit. This action continues for as long as the circuit is switched on so that the binary digits are always available in dynamic form.

The delay line itself can employ solid material, such as quartz, in which multiple paths, and consequently considerable delay, are possible. More usually however the delay line uses either the delay properties of a *column of mercury* or the *magnetostriction* properties of a nickel wire. In the former, quartz crystals are used to convert the electrical pulses into acoustic pulses at the sending end, conversion back into electrical pulses occurring at the receiving end. The acoustic pulses are delayed in the line by an amount proportional to the length of the line.

In the magnetostriction delay line, pulses of current representing the digits are passed
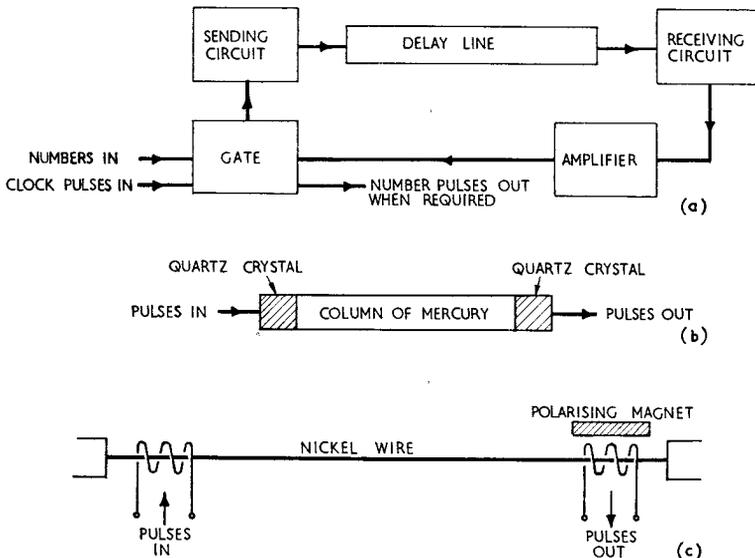
Fig. 14.   DELAY LINE STORAGE DEVICES.

through a coil surrounding one end of the wire. Pulses of mechanical stress are therefore set up in the nickel wire by magnetostriction effects (see Book 1, Sect. 2, Chap. 1) and the pulses travel along the wire at the speed of sound. A receiving coil, with a polarizing magnet, is placed at the other end of the line and the arriving mechanical pulses produce changes of permeability in the nickel wire by magnetostriction: these in turn produce changes in the magnetic flux linking the receiving coils and voltage pulses are therefore induced in the receiving circuit.

When a binary number is circulating in a delay line, the pattern of pulses is repeated over and over again. The line itself is made to have sufficient delay to ensure that the first pulse of the train does not emerge at the receiving circuit until after the last pulse has been passed into the line by the sending circuit. In practice it is not possible to adjust the line mechanically to the degree of accuracy required and so at each circulation the pulses emerging from the output of the line (with slight timing errors) are used to gate fresh and accurately timed clock pulses into the input in the same digit pattern.

A series of delay lines can be used to form a working or backing-up store and this is their main use in computers to date.

34. **Summary.** There are other forms of storage device, and new ideas are constantly being developed. However enough has been said to show the various possibilities. *Capacity* and *access time* are the chief conflicting requirements and these have not yet been successfully incorporated together in a single storage device. However by using, say, magnetic tape in the main external store. ferrite core devices in the backing-up store, and transistor bistable circuits in the working store and registers, all the requirements of the computer can be met.

Logic Circuits

35. In numerical calculations in a digital computer, it is often required that an output from a unit is obtained only when certain conditions prevail at the input. For example, in binary addition, a *'carry* 1' is obtained only when adding two 1 bits. Thus, it might be desired that A occurs only if B occurs: or A occurs only if B and C occur simultaneously: and so on.

One might say, 'If I had the time *and* the money, I would go to the cinema': going to the cinema is, in this case, dependent on the two conditions—time *and* money. If one said, 'I could see a film at either the George *or* the Odeon', a visit to one *or* the other would produce the desired result. Finally, 'I shall see this film if I do *not* fall asleep in the cinema': here the result depends on something *not* happening.

These three statements describe the three basic *"logical"* operations required in a digital computer, namely:—

(*a*) C occurs if A *AND* B occur simultaneously.

(*b*) C occurs if A *OR* B (*OR* both) occur.

(*c*) C occurs if A does *NOT* occur.

36. These logical operations are illustrated in Fig. 15. In (a), the lamp C lights only if switches A *AND* B are closed. In (b), the lamp C lights if A *OR* B (*OR* both) are closed. In (c) the switch is normally closed: thus the lamp C lights only if A is *NOT* operated.
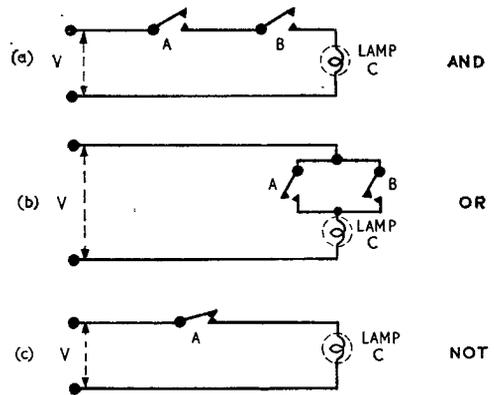
Fig. 15. BASIC LOGICAL OPERATIONS.

37. The circuits that perform these operations in a digital computer are called *logic circuits*. However instead of using switches as in Fig. 15, a great variety of bistable devices can be used (e.g., electromagnetic relays with holding contacts, trigger circuits, ferrite cores, and so on). Such devices are used in special circuit arrangements called *'gates'*. A gate is the name given to a switching circuit which is so designed that the output is obtained only when certain input conditions are met. The three basic gates used in digital computers are those performing the operations described above;

FUNCTIONAL SYMBOL	BRITISH STANDARDS SYMBOL	REMARKS
AND	2	C OCCURS IF THE "2" INPUTS A AND B OCCUR TOGETHER
OR	1	C OCCURS IF ANY "1" INPUT A OR B OCCURS
NOT	1	C OCCURS IF THE "1" INPUT A DOES NOT OCCUR
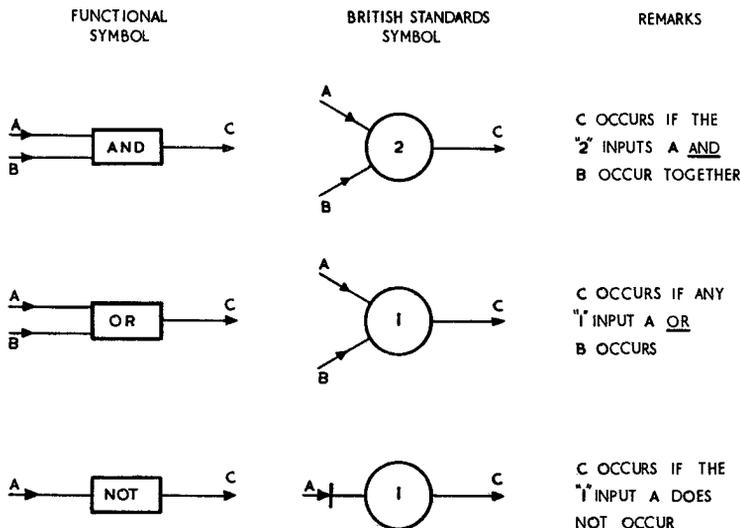
Fig. 16. 'AND', 'OR' AND 'NOT' GATES.

that is, *AND* gates, *OR* gates and *NOT* gates. These are illustrated in Fig. 16.

38. A gating circuit is by no means a new device to those familiar with radar circuits. In particular the use of the suppressor grid of a pentode valve as a 'gating' electrode is common. Fig. 17 illustrates a typical pentode gate. Signal voltages are usually applied to the control grid, whilst gating pulses are applied to the suppressor grid. In the quiescent state the suppressor grid is held at a negative potential, sufficient to cut off anode current. Positive-going pulses are applied to the suppressor grid at precise instants of time, thereby raising the suppressor grid voltage sufficiently to allow anode current to flow—provided that the control grid is also above cut-off potential during this time. Thus the only signals which

produce an output are those which appear at the control grid during the intervals of the positive-going pulses at the suppressor grid. The pentode gate is, in this context, a simple form of *AND* gate.

39. In digital computers, the inputs to the gates are voltage pulses representing the binary digits 0 or 1. With *two* inputs A and B, there are *four* possible permutations of these digits, namely:—

	A	*B*
Condition 1:	0	0
Condition 2:	0	1
Condition 3:	1	0
Condition 4:	1	1

It is now necessary to see the effect of applying each of these conditions in turn to the three basic gates.

The AND Gate

40. An AND gate is a circuit which provides an output when, and only when, *every* input is present. For *two* inputs A and B, the following table (known as a 'truth' table) can be drawn up for the four conditions mentioned earlier.

	A	*B*	*Output* (*A AND B*)
Condition 1:	0	0	0
Condition 2:	0	1	0
Condition 3:	1	0	0
Condition 4:	1	1	1
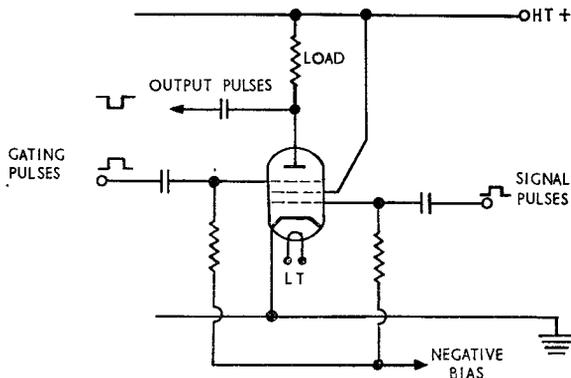
Fig. 17. PENTODE GATE.

(An output digit 1 is obtained only when A *AND* B are present (i.e., when a 1 digit is present at *both* inputs)).

41. A simple *AND* circuit is illustrated in Fig. 18. It consists of two diodes (which can be either thermionic or semiconductor types) connected through a high value
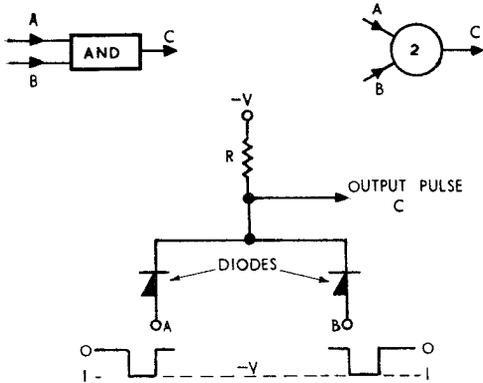
Fig. 18. DIODE 'AND' GATE.

resistor R to a negative supply. The binary digits are applied to the terminals marked A and B, a 0 being represented by *no pulse* and a 1 by a *negative pulse* of voltage.

In the quiescent state, a current is flowing through R and dividing equally through the two diodes. The output C only falls to the negative supply voltage $-V$ (giving an output digit 1) if *both* diodes *cease to conduct;* that is, if there is a negative pulse (a 1 bit) at A *AND* B at the same instant of time. If a negative pulse is applied to only one of the diodes, the other remains conducting and the output is unchanged from the quiescent state.

42. Diodes provide no amplification and when gates are being used in series or cascade there is the need for some active element to provide the necessary gain. Transistors can provide this. A basic AND gate using p-n-p transistors is illustrated in Fig. 19(a). In the quiescent state (zero input) TR_1 and TR_2 are both cut off so that there is no current through the load resistor R and the output at C is at earth potential (zero). A *negative* pulse is needed at the base of *both* transistors to cut them both on. Thus, when 1 digits (i.e., negative pulses) are present together at A *AND* B, transistors TR_1 and TR_2 are forward-biased and so cut on. The current through the load

resistor R causes the voltage at C to fall towards the supply potential $-V$ indicating a 1 digit out. Thus a 1 is obtained at the output only when a 1 is available at both A *AND* B.

A gate circuit that is often used in practice is a form of *NOT-OR* (or *NOR*) gate, an example of which is illustrated in Fig. 19(b). The action is similar to that of Fig. 19(a) but this time the load resistor R is in the *collector* circuit. Thus, when TR_1 and TR_2 are cut off, the output at C has a potential of $-V$ volts (indicating a 1 digit) and only when TR_1 and TR_2 are *both* cut on by negative pulses, or 1 digits applied to their base, does the voltage at C fall towards earth potential because of the voltage drop across R: in other words, an output is available at C if A *or* B *or* both is *NOT* present.

Other devices, including ferrite cores, can be used to provide an AND gate. However, the basic operation of such a gate can be appreciated from the two devices considered and it is therefore not proposed to deal with it further in these notes.
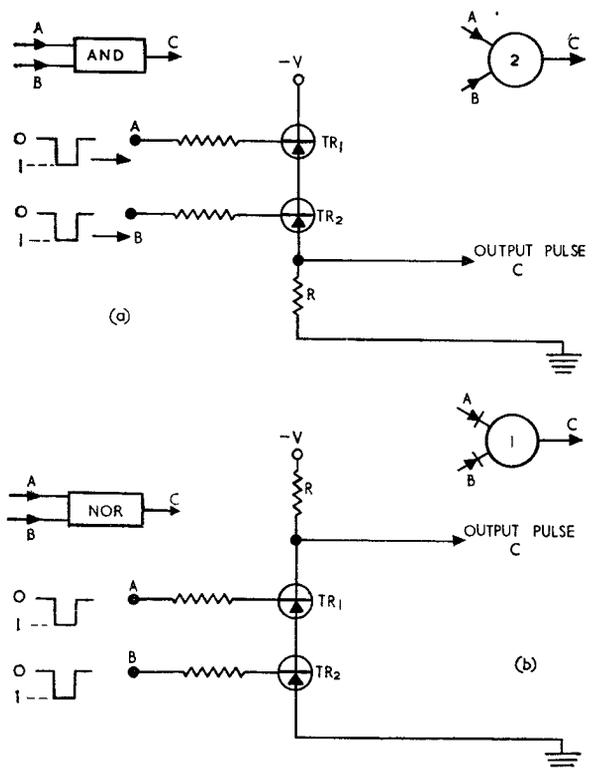
Fig. 19. TRANSISTOR 'AND' AND 'NOR' GATES.

The OR Gate

43. An OR gate is a circuit which provides an output when any *one* or *more* of its inputs is present. For two inputs A and B, a truth table can be drawn up for the four basic conditions.

	A	B	Output (A OR B)
Condition 1:	0	0	0
Condition 2:	0	1	1
Condition 3:	1	0	1
Condition 4:	1	1	1

(An output digit 1 is obtained when A *OR* B *OR* both are present (i.e., when a 1 digit is present at *either* input).)

44. A simple OR circuit is illustrated in Fig. 20. It consists of two thermionic or semiconductor diodes connected through a high value resistor R to a positive supply.
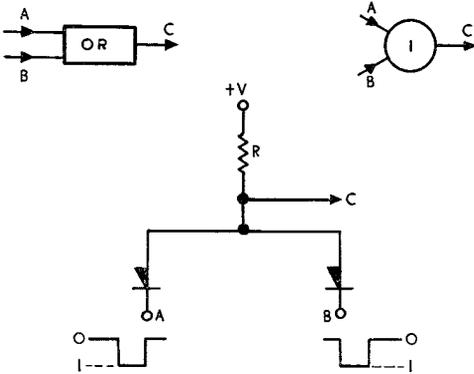
Fig. 20. DIODE 'OR' GATE.

In the quiescent state (zero input) a current is flowing from the positive supply through R and dividing equally through the two diodes in their forward direction. In this condition the output at C is approximately at earth potential (zero output).

When a 1 digit (a negative pulse) is applied to *one* of the inputs, the output point C also goes negative, because the resistance of the diode is low compared with R. Thus an output digit 1 (a negative pulse) is developed at point C when a 1 digit is applied to *either* input terminal.

It should be noted that the OR circuit of Fig. 20 will act as an AND gate for *positive* input pulses. Similarly the AND circuit of Fig. 18 will act as an OR gate for *positive* input pulses. Advantage is often taken of this fact.

45. In the same way that transistors are sometimes required for AND gates, so they are used for OR gates. A basic OR circuit using p-n-p transistors is illustrated in Fig. 21(a). In the quiescent state, TR_1 and TR_2 are both cut off, and with no current through the load resistor R the voltage at C is at earth potential (a 0 digit). If a negative pulse, representing a 1 digit, is applied to *either* A *OR* B, then TR_1 or TR_2 respectively cuts on and the resultant voltage drop across R causes the output at C to go negative. A negative-going output pulse (a 1 bit) is thus obtained if a negative pulse (a 1 bit) is applied to *either* A *or* B. This gives the required result.

If the load is inserted in the common *collector* circuit as in Fig. 21(b), a *NOT-AND* gate is obtained. In this, an output is available at C only when A *and* B are both *NOT* present: with a 1 digit applied to either A or B, the output falls to zero (a 0 bit).
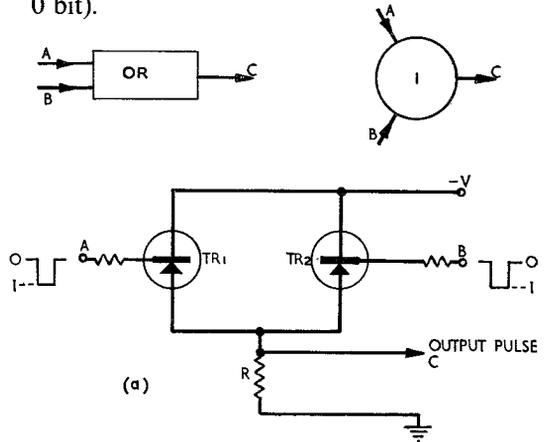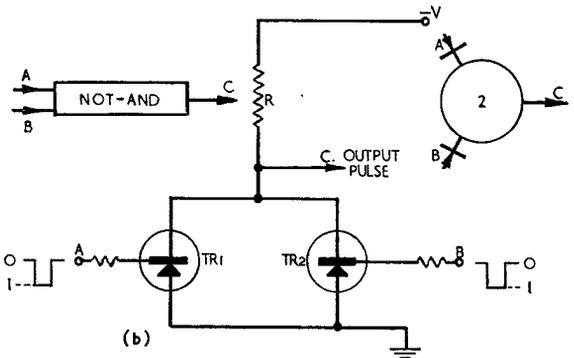
Fig. 21. TRANSISTOR 'OR' AND 'NOT-AND' GATES.

The NOT Gate

46. A NOT gate is a circuit which provides an output only when there is *no input:* alternatively, if there *is* an input, there is *no output.* Various circuits (valve, transistor, ferrite core, cold cathode valves) are used to provide this inversion and examples have already been given in the NOR and the NOT-AND gates. A simple NOT gate using a p-n-p transistor is illustrated in Fig. 22. If a *negative* pulse, representing a 1 digit, is applied to A, TR_1 is forward-
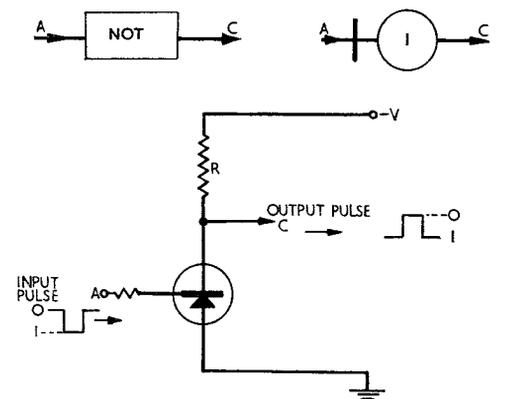
Fig. 22. TRANSISTOR 'NOT' GATE.

biased and conducts: the resultant voltage drop across R is such that the output at C is at approximately earth potential. For the binary digit 0, terminal A has no input and TR_1 is cut off: with no current through R, the output at C falls to the supply potential $-V$ volts. Thus a 1 at A gives a 0 at C: and a 0 at A gives a 1 at C: *inversion* has taken place.

The NOT EQUIVALENT Circuit

47. A combination of AND, OR and NOT gates can be used to provide a circuit known as the NOT EQUIVALENT circuit (also known as a *Binary Comparator*): this is a special circuit which 'recognises' when the input digits are *not the same,* and is an integral part of an *adding* circuit.

For two inputs A and B, a truth table can be drawn up for the four basic conditions.

	A	B	Output (A and B NOT EQUIVALENT)
Condition 1:	0	0	0
Condition 2:	0	1	1
Condition 3:	1	0	1
Condition 4:	1	1	0

(An output digit 1 is obtained when A and B are *different* (i.e., when a 1 digit is present at A and *not* at B, and *vice versa*).)

48. An arrangement that provides the NOT EQUIVALENT output is shown in Fig. 23. If A and B are *both* present (i.e., both 1 digits) there is an output from circuit 1:
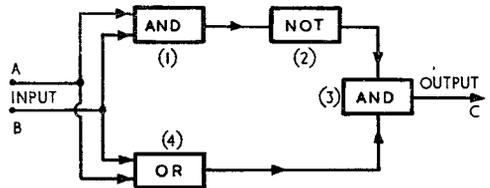
Fig. 23. THE 'NOT EQUIVALENT' CIRCUIT.

this is applied as an input to circuit 2 which, as a NOT circuit, provides *no output:* with no input to circuit 3 from circuit 2 there is no output at C: this gives 'condition 4'.

With no input at either A or B, there is no output from circuit 4 and therefore no output at C from circuit 3: this gives 'condition 1'.

If either A or B (*but not both*) is present, there is no output from circuit 1 and therefore an output from the NOT circuit 2: this is applied to circuit 3. With either A or B energized there is also an output from circuit 4 and this is also applied to circuit 3. With two inputs available at circuit 3, an output is provided at C: this gives 'conditions 2 and 3'.

This arrangement therefore satisfies the required conditions; namely an output from C only when the inputs at A and B are *different.*

The "Half-Adder" Circuit

49. The addition of two binary digits produces a *sum* or output digit and a *carry* digit. For two inputs A and B, a truth table can be drawn up for the four basic conditions.

	A	B	SUM	CARRY
Condition 1:	0 +	0	0	0
Condition 2:	0 +	1	1	0
Condition 3:	1 +	0	1	0
Condition 4:	1 +	1	0 →	1

If the truth tables for the AND gate and the NOT EQUIVALENT circuit are
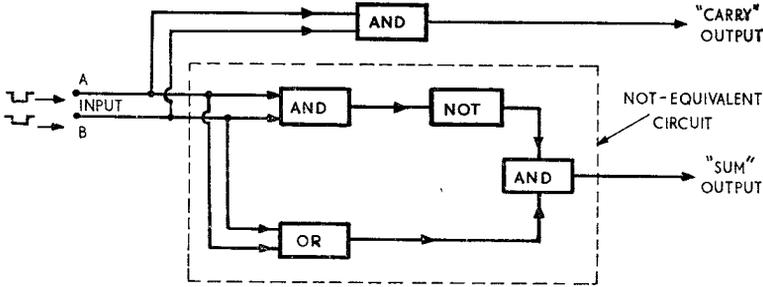
Fig. 24. THE BASIC HALF-ADDER.

examined, it will be seen that the sum column above can be obtained from the NOT EQUIVALENT circuit and the carry column can be obtained from an AND gate. Thus, combination of a NOT EQUIVALENT circuit with an AND gate as in Fig. 24 will satisfy the conditions for basic addition of two binary digits. This can easily be verified by working through the four conditions for A and B.

50. A simplification can be made to the arrangement shown in Fig. 24. Since an AND circuit provides an output only when both inputs are present, the first AND gate in the NOT EQUIVALENT circuit can be used to provide the carry digit as well as giving the input to the NOT gate. The separate AND gate for the carry digit is

Fig. 25. THE MODIFIED HALF-ADDER.

no longer required. The simplified arrangement is illustrated in Fig. 25: such a circuit is known as a 'half-adder'.

51. Although the half-adder satisfies all the conditions for addition as set out in the truth table of para. 49, it is lacking in one respect: it does not have any means of accepting at its input, a *carry* digit which may result from a previous addition. For example, in adding the binary numbers 0111 (7 in decimal) and 0110 (6 in decimal), the procedure on paper is to set them down as follows:—

$$
\begin{array}{r}
0111 \\
+0110 \\
\hline
1101 \quad \text{Sum (13 in decimal)} \\
\hline
11 \quad \text{Carry} \\
\hline
\end{array}
$$

The units (2^0) column gives $0 + 1 = 1$: a half-adder can perform this operation satisfactorily.

The twos (2^1) column gives $1 + 1 = 0$, carry 1: this too the half-adder can cope with.

However, the fours (2^2) column is $1 + 1 +$ '*carried* 1' $= 1$, carry 1: that is, there are now *three* inputs and the half-adder has no means of accepting the carry digit from the previous addition. This disadvantage is overcome in the *full adder* circuit.

Fig. 26. THE FULL ADDER.

52. Complete addition can be achieved by using a second half-adder circuit to add any *carry* digit to the *sum* output of the first half-adder. The basic arrangement is illustrated in Fig. 26(a): an OR gate is included, the reason for this being explained later.

The binary digits at the input terminals A and B are added in the first half-adder, the sum output being applied to the second half-adder. If a carry (a) is produced as a result of the first addition, it is applied through the OR gate to the second half-adder, where it is added to the sum output of the first stage.

The OR gate is required to take care of any carry (b) which may be obtained from the second stage as a result of the previous operation, when A and B by themselves do not produce a carry (a).

Fig. 26(a) therefore represents the basic full adder circuit, sometimes known as a 'three input adder', and this can be drawn in the simplified form shown at Fig. 26(b).

53. The connection between the C terminal of the second half-adder and the 'carry' output of the OR gate is shown dotted in Fig. 26, because there is one other aspect still to be considered.

The procedure for adding the binary digits 0111 and 0110 is repeated below:—

$$
\begin{array}{ll}
0111 & \text{A} \\
0110 & \text{B} \\
\hline
1101 & \text{Sum} \\
0110 & \text{Carry} \\
\end{array}
$$

The units pair of digits presents no problem. With the twos pair of digits however a carry is formed. On paper, this carry digit is merely taken into the *next higher* column and added to the digits already there. However, in a computer, the digits are represented by pulses as explained earlier. Thus, if a carry pulse is formed it has to be *delayed in time* until the pulses representing the next higher value digit pair are presented at terminals A and B. It is therefore necessary to *delay* the carry digit for a time equal to *one digit period* (i.e., the interval between successive pulses) and then present it at terminal C for adding to the sum of the next pair of digit pulses. The delay line, represented by the block D in Fig. 26(c) need consist of nothing more than a suitable inductance-capacitance network, accurately adjusted to delay the pulse by one digit period.

54. A simplified block diagram of the whole system is illustrated in Fig. 27. If the binary numbers 0111 and 0110 are to be added, pulses representing these numbers are applied to the input terminals A and B respectively.

During the first pulse period, the bits 1 and 0 in the units column are applied to AB. Tracing the circuit action through from this point it will be seen that no carry digit pulses are produced and a binary digit 1, representing the sum appears in the units column of the output.

During the second pulse period, the bits 1 and 1 in the twos column are applied to AB. A carry (a) 1 bit is now produced
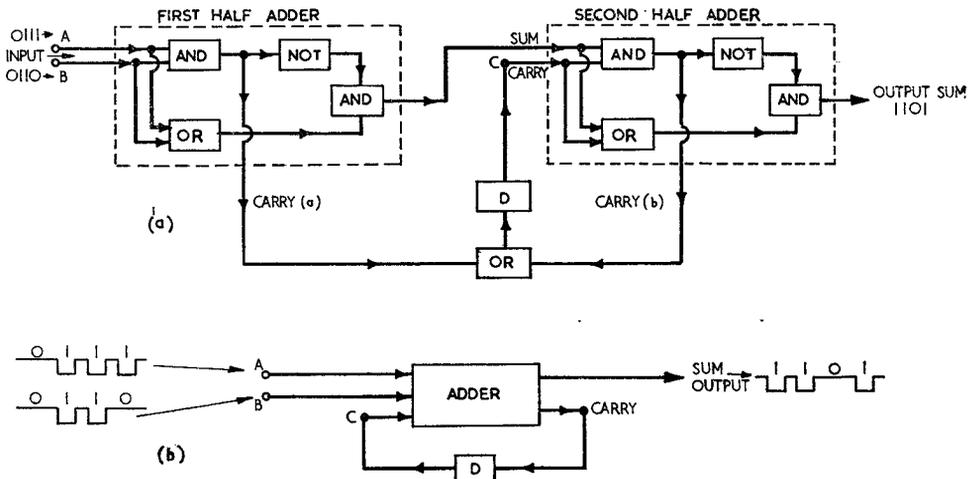
Fig. 27. OPERATION OF ADDER CIRCUIT.

and applied via the OR gate to the delay line D where it is 'stored' for one digit period. There is *no sum output* from the first half-adder, so that with no input to the second stage, the sum digit 0 is produced at the output for the twos column.

During the third pulse period, the bits 1 and 1 in the fours column are applied to AB and, as before, the sum output from the first stage is 0 and a carry (a) 1 digit is applied to the delay line D, where it is stored for one digit period. The original carry (a) 1 digit stored during the *second* pulse period is now applied to terminal C of the second half-adder: the input to this stage is therefore 0 and 1, which produces a sum output of 1 in the fours column: no carry (b) digit is produced.

During the fourth pulse period, there is no input to the first half-adder and therefore no sum or carry (a) output from this stage. However the carry (a) 1 digit stored in D during the *previous* digit period is now applied to terminal C of the second half-adder: the input to this stage is therefore 0 and 1, which produces a sum output of 1 in the eights column: no carry (b) digit is produced.

Arithmetic Unit

55. The arithmetic unit is built round one or more adder circuits of the type described in the preceding paragraphs. However in most digital computers, the two numbers to be added do not become available simultaneously as a matter of course. Usually one of them has to be held ready in *storage* until the other arrives and then both numbers are applied to the adder together. These temporary storage devices in the arithmetic unit are called *registers*.

A register consists of a number of bistable circuits, each circuit being capable of storing a binary digit for as long as required. The actual circuits used may be valve trigger circuits, transistor bistable circuits or ferrite core devices. No matter which device is used, the number of bistable circuits depends on the magnitude of the numbers that the computer is designed to cope with, i.e., on the word length: a register capable of handling thirty-two bit numbers will have thirty-two bistable circuits.

56. If two binary numbers 1101 and 0110 are to be added, pulses representing the first number are fed, digit by digit, into the register during the first number period. This form of operation whereby bit follows bit along the wires (the highway) is referred to as *serial operation* and it is the only method considered in these notes. The units digit enters the *most significant end* of the register first and as each bit is entered, the previous bits are 'shifted' through the register by appropriate pulses until the units digit is registered in the final bistable circuit. The next stage of the operation can now
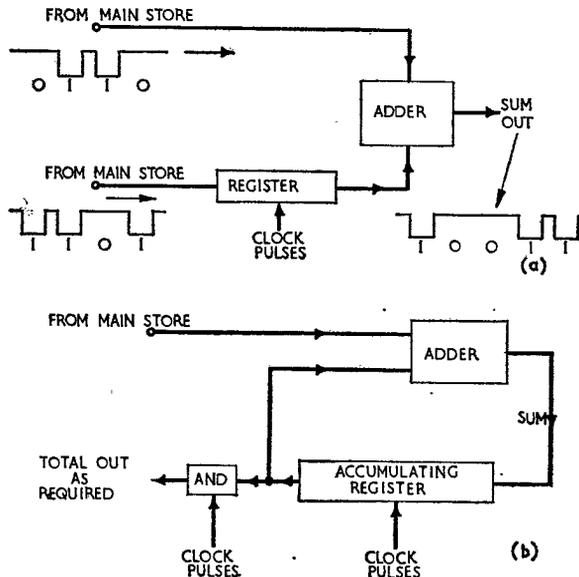
Fig. 28. ARITHMETIC UNIT REGISTERS.

proceed because the first number 1101 is stored in the register (Fig. 28(a)).

The second number 0110 is drawn from the main store and its units digit is applied as one input to the adder at the same instant as the units digit is pulsed out from the register to the adder. *Synchronisation* is important and the clock pulses generated by the master pulse generator, under the control of the control unit, ensure this.

57. In some computers a special type of register is used in the arithmetic unit: this *circulates* its contents through the adder as shown in Fig. 28(b). Initially, the register is 'empty': the first number from the main store is applied to the adder which pulses the number, digit by digit, into the register.

When the second number is applied from the main store in the next number period, it is added digit by digit to the number being pulsed out to the adder from the register: again accurate control and synchronisation is essential.

A new pattern of digits then emerges from the sum output of the adder and this takes the place of the single-number pattern originally stored in the register.

If yet another number is fed in from the main store it will be added to the first total in the register.

Thus the register is arranged so that it *accumulates a running total*, adding new numbers to whatever is already in it. For this reason it is called an accumulating register or *accumulator*. The output can be read out of the accumulator as required by feeding appropriate pulses to an AND gate in synchronism with the digits being pulsed from the accumulator.

58. For the other arithmetical operations of subtract, multiply and divide, *additional* circuits are required in the arithmetic unit. The basic idea of these operations has already been dealt with and it is not proposed in this summary to consider the detailed operation. It is sufficient to know that in most computers, *subtraction* is achieved by the *addition* of the *complements* of negative numbers. Thus in the arithmetic unit, all numbers are applied first to *sign-determining circuits*—usually a bistable trigger circuit: if the number is positive (as indicated by a 'sign' digit of 0) the trigger remains off and the number is applied

directly to the adder. If the number is negative (as indicated by a 'sign' digit of 1), the trigger is pulsed on: this in turn operates a *complementer* circuit which applies to the adder the complement of the negative number (Fig. 29).
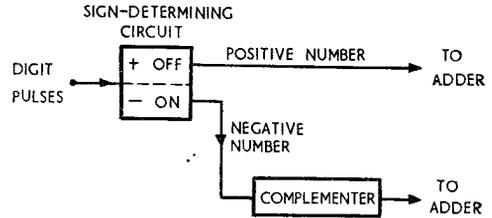
Fig. 29. METHOD OF SUBTRACTION.

Since the complement of a binary number can be achieved by reversing all the digits and adding 1 in the units column, the complementer is merely a form of NOT gate with provision for the addition of binary 1.

59. The procedure usually employed for *multiplication* involves *shifting* the number being multiplied one position to the *left* followed by *addition*: the number of shift and add operations depends on the number of digits in the *multiplier*. Since the numbers are represented by pulse trains, moving one place to the left means *delaying in time* by one digit pulse period. This can be achieved by a suitable bistable trigger circuit, or in practice by a number of such circuits depending on the word length: the combination is then referred to as a *shift register*. Thus the essential parts of a multiplying circuit based on this principle are a shift register and an adder with an accumulator to allow the shifted numbers to be added each time to give a running total: this is shown in Fig. 30. A type of gate circuit is used to test each digit of the multiplier in turn: if it is a 1, the number being multiplied is shifted one place to the left by the shift register and gated through to the adder where it is added to the accumulator: if the multiplier digit is a 0, the number being multiplied is shifted one place to the left by the shift register, but the gate is now inoperative so that the multiplicand is *not added* into the accumulator total.

Division, like multiplication, consists of a succession of shift and add operations although in this case the shift is to the *right* and the addition is of the complements

Fig. 30. SHIFT REGISTERS.

Fig. 31. INTEGRATION BY SUMMATION OF RECTANGLES.

of negative numbers. Thus for division, shift registers are also required along with sign-determining circuits and complementers.

The Digital Differential Analyser (D.D.A.)

60. It was stated earlier in these notes that the input to a digital computer consists of discrete numbers, and a normal digital computer cannot be expected to compute a *continuously varying* quantity: this type of calculation is generally carried out by analogue computers, despite their inherent inaccuracy. However the limitation of the digital computer in this respect has now been overcome by the introduction of a system known as the *digital differential analyser (d.d.a.)*. With this system a digital computer can be used in aircraft for the calculation of navigational problems previously handled exclusively by analogue computers. Such problems, it will be remembered from Chapter 1, involve integration and differentiation processes: for example, given air speed and time as the input quantities, distance flown can be calculated by integrating speed with respect to time.

61. Integration is effectively a summation or adding-up process. Thus, calculation of the area under the curve of Fig. 31 can be obtained as follows. Take two consecutive values of Y separated from each other by a small increment dY: the corresponding X co-ordinates are separated by the increment dX. If the value for Y is multiplied by dX the approximate area of this strip is obtained; and the smaller dX is made,

the more accurate is the calculation. If the area under the curve is divided up into a large number of small strips, then the total area under the curve can be calculated by adding up all the elementary areas. In effect the d.d.a. provides digital integration by the summation of such strips in a step-by-step process.

62. The d.d.a. carries out the calculations given to it in an *incremental* manner: that is, it provides the *up-to-date* value of any given quantity by making a small correction to the immediately preceding value. This mode of computation is particularly applicable to a continuous navigation process where the input quantities are also varying continuously. Provided that the d.d.a. can operate at a rate comparable with the rate of change of the input quantities then the output can be correctly kept up to date.

63. The d.d.a. carries out this function in the following way. The Y number (i.e., the number to be integrated) is represented by a binary number held in a register as shown in Fig. 32. The Y number is kept up to date by the addition of small increments dY which are represented by digital pulses.

The independent variable input dX with respect to which integration takes place is also represented by synchronized digital pulses which cause the number held at that instant in the Y-register to be added into a second register (called the Remainder or R-register) each time a dX pulse is applied. In effect, this is the same as calculating the area of each small strip under the curve of Fig. 31.
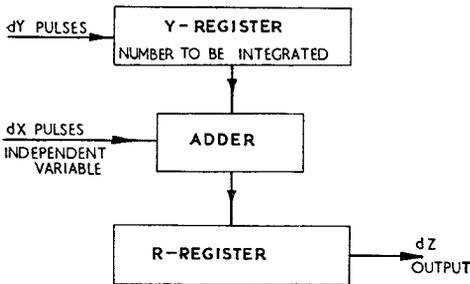
Fig. 32. THE DIGITAL INTEGRATOR.

As integration proceeds the R-register, which has the same capacity as the Y-register, overflows and this overflow represents the increment dZ of the integral. The rate of overflow is proportional to both the fullness of the Y-register and the rate at which Y is added to the R-register (i.e., on the dX pulse rate). Thus dZ is proportional to YdX and integration of Y with respect to dX has taken place. Note that unlike the Miller integrator and velodyne discussed in Chapter 1 where integration is with respect to *time* only, integration with the d.d.a. may be with respect to *any variable X* (including time).

64. A complete d.d.a. consists of a *number* of digital integrators with facilities for interconnection (one d.d.a. in operation uses 50 integrators). There are many ways however in which a machine may be organized in detail. Each integrator may have its own arithmetic unit and be self-contained: alternatively, the Y and R number may be read out in turn from a central store into a single *time-sharing* arithmetic unit. These systems are known as 'simultaneous' and 'sequential' systems respectively. The simultaneous machine does more integration steps per second than the sequential machine, but the sequential machine requires far fewer components because of the time-sharing process and is therefore more economical. The *sequential* machine is therefore the type usually employed for aircraft applications.

Control Unit

65. **Introduction.** It was stated earlier that a general purpose computer is provided with a 'programme' prepared by a human operator. Every operation that the machine can do is given a code number in binary notation, and the programmer inserts the

numbers of the requisite operations into the input unit and thence into the main working store of the computer. The numbers that are to be operated on (i.e., the numerical data) are also fed in and stored in binary form in various registers in the main working store.

The programme consists of a group of coded 'instructions' in the form of binary numbers: each instruction word has two parts—a *command* and an *address*. The command part of the instruction details the operation that is to be performed (e.g., add, shift, transfer, multiply): the address part specifies the *location* of numerical data or of other instructions in storage (e.g., 'address 02' may mean the second register in the main working store). Thus a typical instruction might be '*Add* the contents of *address* 07 into the accumulator'. And of course a programme consists of a whole *group* of such instructions.

It is now necessary to see how the stored instructions are selected in the right order and used to produce the required control pulses which are sent from the central control unit to the remainder of the computer.

Before this is done, however, it is as well to see what the control signals consist of.

66. **Control signals.** In a digital computer, a generator produces clock pulses at precise instants of time, the number of pulses produced in a given number period depending on the word length of the computer (e.g., 32 bits). Furthermore, each succeeding pulse in the number period represents a higher power of 2 (i.e., 2^0, 2^1, 2^2, and so on depending on the word length). These pulses are used for synchronizing the control circuits by opening or closing gates throughout the computer at particular instants of time.

To do this, clock pulses representing particular intervals in the number period must be made available, giving 2^0 pulses on one wire, 2^1 pulses on another, 2^2 pulses on another and so on for the whole number period. This is illustrated in Fig. 33.

67. To produce these gating pulses, a form of electronic distributor is used. The pulses from the clock pulse generator are applied simultaneously to a series of AND gates, as shown in Fig. 34, the number of gates depending on the word length. Each of the gates is then operated in turn (and
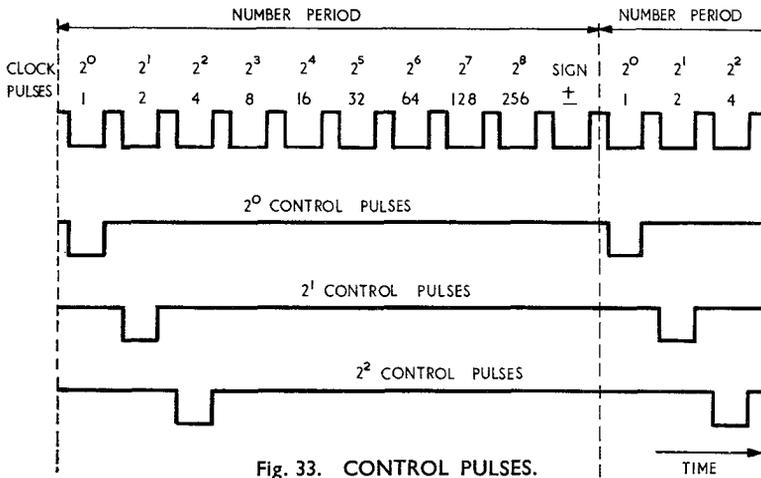
Fig. 33. CONTROL PULSES.

switched off again immediately afterwards) and this is done by using the pulse from each output, *delayed* by one digit pulse period, to operate the next gate.

The output from the clock pulse generator is frequency divided to give a series of initiating pulses at the pulse recurrence frequency of the number period, and these initiating pulses (one during each number period) are applied to the first gate in the chain. Thus the clock pulse allowed through this first gate becomes the first in the number period and takes the value 2^0, whilst the other outputs become 2^1, 2^2, 2^3, and so on. These separate and successive pulses on different wires are available for use by the control unit as required.

68. **Operation.** The basic operation of the control unit and the method by which

the gating pulses are used to control the sequence may be illustrated by considering Fig. 35.

As previously noted, the programme consists of a group of instruction words, each instruction containing an address part and a command (or function) part. These instructions are pulsed into binary code by the input unit and held at successive addresses in the main working store *in the order in which they are required*. The numerical data (i.e., the number to be operated on) are similarly fed into various addresses in the main store.

69. It is necessary for the instructions to be selected *in turn* and the numbers contained therein to be passed along the required channels by opening and closing appropriate gates. This process is initiated by the
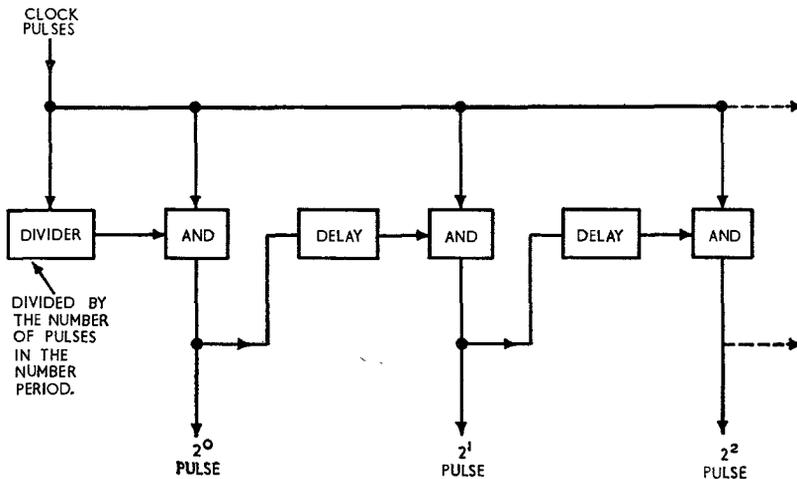
Fig. 34. PRODUCTION OF CONTROL PULSES.
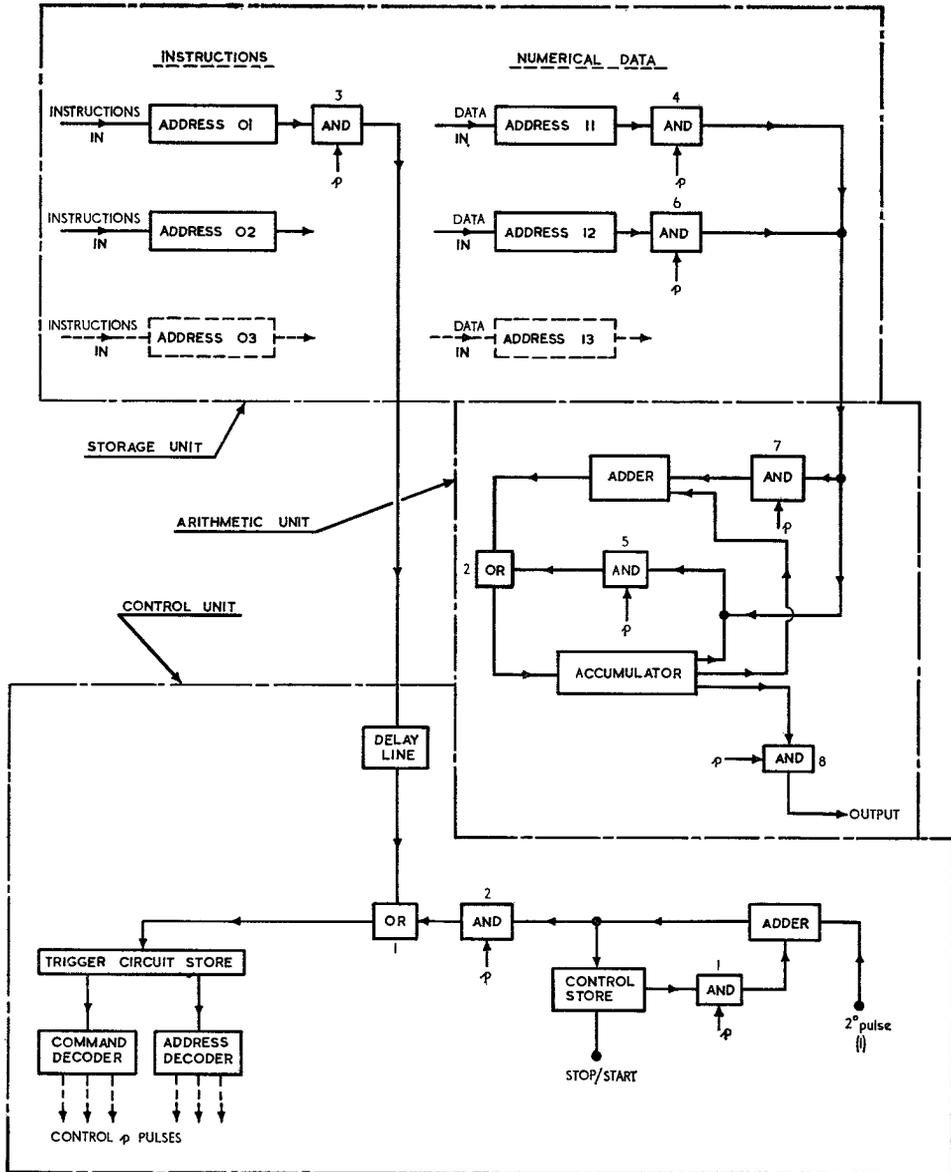
Fig. 35. OPERATION OF CONTROL UNIT.

control store which holds the address of the instruction which has just been obeyed by the computer. The number 1 (from the 2^0 clock pulse) is *added* into the control store to give the address of the *next instruction.*

Initially, on starting, the number held in the control store is 0: by adding 1, the address of the next instruction—in this case the *first* instruction address 01—is obtained. The new address causes the control unit to

produce control pulses which operate the output gate of the main working store register at address 01: this address contains the appropriate instruction which itself passes to the trigger circuit store where both its address and command parts are decoded to produce the desired result.

70. All these operations are performed in an orderly, repetitive cycle which may

extend over, say, four number periods. For example:—

(a) In the first number period, the address number held in the control store is applied through AND gate 1 to the adder where 1 is added to form the address of the *next* instruction. The new address is then held in the control store ready for use during the *next cycle* of operations: at the same time the new address is applied through AND gate 2 and OR gate 1 to the trigger circuit store. On starting up, the new address will be 01, that of the *first* instruction.

(b) During the next number period, the new address (01) is decoded in the trigger circuit section of the control unit and the resultant control pulses operate AND gate 3 to allow the instruction at address 01 to pass out of the main working store. It is desired to place this instruction in the trigger circuit part of the control unit, but as this is in use during the present number period for the selection of the instruction, it is sent to a *delay line* as a temporary measure.

(c) During the next number period of the cycle, the instruction emerges from the delay line and passes into the trigger circuit store via OR gate 1. The address part of the instruction (usually the address of numerical data to be operated on) now takes the place of the instruction address 01 that was there in period (b).

(d) The address and command parts of the first instruction are decoded in the trigger circuit part of the control unit and appropriate clock pulses are produced to open and close gates which cause the arithmetic or other operation specified by the instruction to be carried out.

This cycle (a) to (d) is then *repeated* for as long as required. During (a) of the *second* cycle, the new address 02 will be formed, and the instruction at this address selected and obeyed during the remainder of the cycle. Address 03 will be formed during the third cycle, and so on, until all the required operations have been performed. Finally, a special 'stop' signal is applied to the trigger circuit via the control store and this causes the control cycle to stop.

71. It may be of interest to see how the instructions are obeyed during number period (d) of the operating cycle. Assume the following instructions:—

(a) *At address* 01: 'Take the number at address 11 and store it in the accumulator'.

(b) *At address* 02: 'Add the number at address 12 to the number in the accumulator'.

(c) *At address* 03: 'Transfer the contents of the accumulator to the output'.

(d) *Stop*.

During the first *cycle* of operation, the instruction at address 01 is selected as explained in para. 69 and in number period (d) of this cycle, the control unit produces pulses which operate AND gates 4 and 5 to allow the contents at address 11 to be pulsed into the accumulator via OR gate 2.

During the next cycle, the instruction at address 02 is selected and in number period (d) of this cycle, the control unit produces pulses which operate AND gates 6 and 7 to allow the contents at address 12 to be added into the accumulator via OR gate 2.

During the next cycle, the instruction at address 03 is selected and in number period (d) of this cycle, the control unit operates AND gate 8 to allow the contents of the accumulator to be pulsed to the output unit.

The cycle of operations is then stopped.

72. There are of course *many other* instructions that could be performed, and a practical computer would also contain shift registers, sign-determining circuits, complementers and so on to perform additional operations. However enough has been said to show the basic function of the control unit and to illustrate how it fits in to the general scheme of operation of a computer.

Input and Output Devices

73. The methods used to supply information to computers and getting answers out are varied. The main problem concerns the feed-in of data and the recording of results at a speed comparable with that at which the computer operates. Input and output devices should be able to handle several different kinds of data in large quantities and at high speed: they should also automatically record information in a form suitable for immediate interpretation.

74. **Input devices.** The most widely used methods of input are magnetic tape, punched paper tape and punched cards.

(a) *Magnetic tape*. This has been described in a previous paragraph dealing with storage devices. The tape stores information (programme and numerical data) in binary digital form as a series of small magnetised 'cells' along the length of the tape: each cell is magnetised by a recording head, the direction of the magnetic field depending on whether the stored digit is 0 or 1. The clock pulses are also recorded on the tape and this ensures accurate *synchronization* of the whole computer.

The information recorded on the tape is converted into voltage pulses representing the binary digits (by reading heads, as in a tape recorder). These pulses are fed in to the main working store of the computer as explained earlier.

Magnetic tape has a large capacity: there is usually more than one information track and it can store up to 1000 bits per inch: a feed-in rate of 200 inches per second can be achieved. The main disadvantage of tape (apart from expense) is that specific information on the tape is difficult to locate: this gives a poor access time.

(b) *Punched paper tape*. Information is recorded in a paper tape as a pattern of holes punched along the length of the tape. Holes represent a 1 bit and the absence of holes a 0 bit. Holes are therefore punched in accordance with the binary digits in the information.

The static representation of binary digits in the tape may be converted into corresponding voltage pulses by metal pins making electrical contact and closing a circuit when they penetrate the holes. This gives a pulse for a 1 bit and no pulse for a 0 bit. An alternative method of read-out uses photo-electric cells which are so placed that they emit when light penetrates a hole and cease emitting when paper comes between the source of light and the photocell. This produces the digit pulses, which are fed in to the main working store of the computer.

Although paper tape is cheap and requires very little complex read-in/read-out equipment, it has a low capacity: the paper also tears easily and like magnetic tape it is difficult to locate specific information.

(c) *Punched cards*. This uses high quality cards in which holes are punched to indicate digits. Normally the value of the digit (2^0, 2^1, 2^2, etc.) depends on the distance from the top or bottom edge of the card: the meaning of the digit can also be conveyed by the column which it occupies reading across the card.

Information from the card is read-in to the computer and converted into digital pulses by the same methods as used for paper tape. These pulses are fed into the main working store of the computer.

Punched cards have the advantage of being very useful for documentation: they are easily readable and are convenient for filing: they are also easy to prepare and can be used for several methods of interpretation: they can also be read-in at high speeds. The main disadvantage is that of cost.

Other methods of input can be used and new systems are constantly being developed: for example, electric typewriters can be connected directly to the computer, and character-reading methods of scanning written information to produce tapes and cards are also under development.

75. Output devices. The object of output devices is automatically to record the result of computation in a form suitable for immediate interpretation (preferably in printed form). The three main *direct* methods of output are the same as those used for the input; namely, magnetic tape, punched paper tape and punched cards: in fact, in many cases, these *output* devices are used to provide the *input* to the computer at later stages of the computation.

However, to provide readable information, the three direct methods are used to activate printing mechanisms. Such mechanisms include:—

(a) *Teleprinters:* these give a slow output but the information is immediately available.

(b) *Typewriters:* these 'sense' the information contained in digital form on one of the 'direct' methods and type it at high speed (600 characters per minute is typical).

(c) *Line-at-a-time printers:* these set up a line of type before printing and provide a high speed output (12,000 characters per minute is typical).
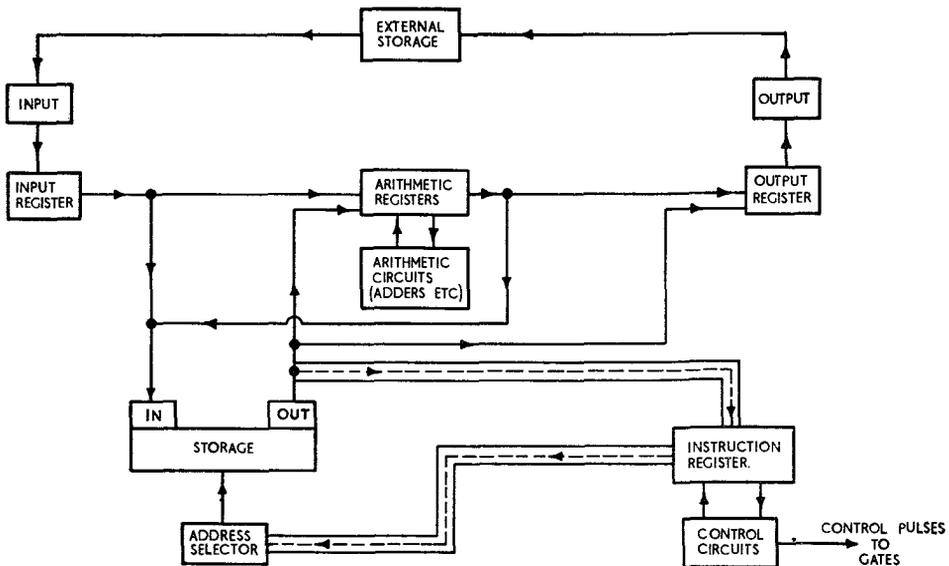
Fig. 36. TYPICAL GENERAL PURPOSE DIGITAL COMPUTER.

(d) *Page printers:* these set up a page of type before printing and provide a high speed output (e.g., 25,000 characters per minute).

(e) *Electrostatic printers:* special devices which provide an ultra high speed output (150,000 characters per minute): however because of cost they are used only for special applications.

Typical General Purpose Computer

76. To conclude this Chapter, a block diagram of a typical general purpose digital computer will be examined: this block diagram is illustrated in Fig. 36.

Both data and instructions are taken from external storage (e.g., magnetic tape or punched cards) and converted by the input unit into digit pulses which are stored temporarily in the input register. From the input register, the programme and data are fed to the main working store to be held at various addresses.

The control unit causes the instructions to be taken out of the main working store in a definite sequence and each instruction is fed to the instruction register and interpreted or decoded by the control circuits: these are converted into control pulses fed to gates throughout the computer to cause

the indicated instruction to be performed. The address part of the instruction is decoded also and the resultant control pulses operate appropriate gates to allow numerical data to leave or enter the main working store as required by the instruction.

Each data 'word' is called into the arithmetic unit from the main working store as needed and results are fed back to storage. After completing the required group of instructions, the result is fed to the output unit from the main working store, the output unit converting the binary pulses into a form suitable for direct interpretation, recording and storage.

Summary

77. This Chapter has, in general way, shown the basic operation of an elementary digital computer. The various units in a computer have been discussed and the methods by which arithmetical operations are performed have been indicated. However, it was stated in the opening paragraphs of this Chapter that nothing more than the outline of digital computer operation could be attempted in these notes. The reasons for this will now be obvious: there are many variations in design and operation: storage devices and logic circuits take many forms and only a selection of the more simple

has been discussed: the tie-up of the control unit depends on the operations to be performed by the computer: and the types of input and output devices used depend on the task being performed.

It is therefore necessary for a technician employed in the maintenance and servicing of a digital computer to be completely conversant with the circuits and operation of that particular machine: the appropriate Air Publication should always be consulted.

INDEX

INDEX TO PART 1A (BOOKS 1, 2 and 3)

This index lists the main contents of Part 1A in alphabetical order. The items in the index appropriate to this book appear in **bold type**.

	BOOK	SECT.	CHAP.	PARA.
Constant-voltage—generator	2	8	2	25
—transformer	1	7	2	39
Constants, valve	2	8	2	11
Control grid, effect of	2	8	2	3
Control synchro	3	19	1	22
Control system, closed loop	3	19	1	12
open loop	3	19	1	12
Control unit, computer	3	20	2	65
Controlling force (instrument)	1	6	1	2
Conventional current	1	1	1	18
Conversion conductance	3	14	2	39
Conversion of analogues	3	20	1	12
Copper losses (transformer)	1	7	2	17
Copper-oxide rectifier	2	9	3	29
Core-type transformer	1	7	2	20
Corkscrew rule	1	2	1	10
Coulomb	1	1	1	20
Coulomb's law	1	4	1	3
Coupled circuits	1	7	1	1
Coupling, band-pass	2	11	1	15
coefficient of	1	7	1	2
critical	1	7	1	13
loose	1	7	1	11
RC	2	10	1	8
r.f.	2	11	1	10
tight	1	7	1	12
transformer	2	10	1	18
Co-valent bond	2	8	7	4
CR circuit	1	5	2	22
Cracked carbon film resistors	1	1	3	5
Critical coupling	1	7	1	13
Critical-distance tetrode	2	8	3	11
Critical frequency (propagation)	3	17	1	13
Cryotron storage device	3	20	2	32
Crystal, quartz	2	12	2	7
Crystal—calibrator	3	18	1	19
—controlled oscillator	2	12	2	10
—detector	3	14	1	15
—filter	3	15	1	28
—holder	2	12	2	6
Current, alternating	1	5	1	1
circulating	1	5	3	17
conduction	1	1	1	17
conventional	1	1	1	18
direct	1	1	1	12
displacement	1	4	2	4
magnetizing	1	7	2	4
primary no-load	1	7	2	9
pulsating	1	1	1	23
wattless	1	5	2	34
Current—measurement	3	18	1	2
—negative feedback	2	10	3	18
—stabilizer	2	9	3	24
Cut-off frequency (filter)	3	15	1	9
C.W. keying	3	13	1	14
C.W. reception	3	14	1	41
Cycle	1	5	1	2
D				
Damped tuned circuit	2	11	1	20
Damping, servomechanism	3	19	2	23
tuned circuit	1	5	3	21
Damping—coefficient	2	12	1	5
—force (instrument)	1	6	1	2
Data transmission	3	19	1	1
Dbm	1	6	3	13
D.C.—	1	1	1	23
generator	1	3	1	1
—motor	1	3	2	9
—remote indicators	3	19	1	3
—tachogenerators	3	19	2	10
—to a.c. conversion	3	20	1	17
Decibel (db)	1	6	3	9
Decibel meter	1	6	3	16
De-coupling	2	10	3	39
Deflecting force (instrument)	1	6	1	2
Deflection, electrostatic c.r.t.	2	8	5	16
magnetic c.r.t.	2	8	5	33
Deflection—defocusing	2	8	5	27
—sensitivity	2	8	5	18
Deflector—coils	2	8	5	33
—plates	2	8	5	9
Delay line storage	3	20	2	33
Delayed a.g.c.	3	14	2	60
Delta—connection	1	5	4	14
—match	3	16	2	14
Depolarizer	2	9	1	9
Depth of modulation	3	13	1	23
Desynn	3	19	1	13
Detector	3	14	1	13
Dia-magnetism	1	2	1	30
Dielectric—	1	4	1	10
—constant	1	4	1	22
—hysteresis	1	4	2	5
—strength	1	4	2	2
Differential gear	3	20	1	24
Differential synchro, control	3	19	1	50
torque	3	19	1	37
Differentiation, analogue	3	20	1	44
Digital computer—	3	20	2	8
—control unit	3	20	2	65
—input devices	3	20	2	74
—logic circuits	3	20	2	35
—output devices	3	20	2	75
—storage devices	3	20	2	25
Digital differential analyser	3	20	2	60
Diode,	2	8	1	18
cold-cathode	2	8	4	16
junction	2	8	7	18
mercury-vapour	2	8	4	4
Diode detector	3	14	1	18
Dipole,	3	16	1	4
folded	3	16	2	15
polar diagrams of	3	16	2	18
Direct current	1	1	1	23
Direct-coupled amplifier	2	10	1	25
Directly heated-cathode	2	8	1	16
Director	3	16	3	3
Discharge of capacitor	1	4	3	12
Discone aerial	3	16	5	19
Displacement current	1	4	2	4
Dissipation, anode	2	8	1	31
power	2	10	2	6
Distortion, attenuation	2	10	1	23
non-linear	2	10	1	23
phase	2	10	1	23
Distortion—in a.f. amplifiers	2	10	2	9
—in c.r.t.	2	8	5	25
—of detector	3	14	1	21
—of waveform	1	7	2	26
Division, analogue	3	20	1	43

	BOOK	SECT.	CHAP.	PARA.
Meters,	1	6	1	1
decibel	1	6	3	16
field strength	3	18	1	21
frequency	3	18	1	10
r.f. power	3	18	5	5
Methods of biasing	2	8	2	23
Mica capacitors	1	4	2	9
Microphone	2	10	1	6
Microwave aerials	3	16	5	22
Miller—effect	2	11	1	5
—feedback	2	12	1	40
—integrator	3	20	1	47
M.K.S. units	1	1	1	5
Modulation—	3	13	1	18
—factor	3	13	1	23
—measurement	3	18	4	2
Modulator, anode	3	13	1	25
ferrite	3	18	2	16
Molecules	1	1	1	8
Monitor, frequency	3	18	1	18
M.O.-P.A. transmitter	3	13	1	9
Morse telegraphy	3	13	1	13
Mosaic telegraphy	3	13	1	13
Motor, commutator	1	5	4	38
d.c.	1	3	2	9
induction	1	5	4	26
losses in	1	3	2	20
speed of	1	3	2	22
starters for	1	3	2	24
synchronous	1	5	4	20
Motor controlled tapped				
transformer	1	7	2	40
Moving coil instrument	1	6	1	5
Moving iron instrument	1	6	1	15
M-type transmission	3	19	1	8
Multimeters	1	6	2	4
Multiple-hop propagation	3	17	1	17
Multiplication (analogue)	3	20	1	38
Multiplicative frequency				
changer	3	14	2	35
Multiplier, electron	2	8	6	7
frequency	2	11	2	26
voltmeter	1	6	1	10
Multi-section filters	3	15	1	32
Multi-stage r.f. amplifiers	2	11	1	14
Multi-unit valves	2	8	3	34
Multivibrator	2	12	3	7
Mutual characteristic, pentode	2	8	3	18
triode	2	8	2	9
Mutual conductance	2	8	2	13
Mutual inductance	1	2	2	13
Mutual inductive coupling	1	7	1	3

N

	BOOK	SECT.	CHAP.	PARA.
Narrow-band amplifiers	2	11	1	1
Negative feedback—	2	10	3	3
—in computers	3	20	1	20
—in transistors	2	10	3	44
Negative resistance—	2	8	3	8
—oscillators	2	12	1	37
Neper	1	6	3	17
Neutralization	2	11	2	6
Neutron	1	1	1	9
Newton	1	1	1	6

	BOOK	SECT.	CHAP.	PARA.
Noise, receiver	3	14	2	26
valve	2	8	3	46
Non-inductive winding	1	2	4	17
Non-linear—device	1	1	2	4
—distortion	2	10	1	23
NOR gate	3	20	2	42
North-south rule	1	2	1	13
NOT-AND gate	3	20	2	44
NOT-EQUIVALENT circuit	3	20	2	47
NOT gate	3	20	2	46
N-P-N transistor	2	8	7	23
N-type semiconductor	2	8	7	7
Number period, computer	3	20	2	22

O

	BOOK	SECT.	CHAP.	PARA.
Octode valve	2	8	3	32
Ohm	1	1	2	5
Ohmmeter	1	6	2	3
Ohm's law	1	1	2	2
Ohms-per-volt rating	1	6	1	11
On-off keying	3	13	1	15
Open-circuited transmission line	3	15	3	12
Open loop control system	3	19	2	12
Open wire feeder	3	15	3	32
Operating conditions, oscillator	2	12	1	21
Operator "j"	1	5	5	1
Optimum working frequency	3	17	1	14
OR gate	3	20	2	43
Oscillator, beat frequency	3	14	2	46
Colpitts	2	12	1	1
crystal	2	12	2	10
dynatron	2	12	1	38
electron-coupled	2	12	1	30
Franklin	2	12	1	31
Hartley	2	12	1	16
local	3	14	2	31
master	3	13	1	9
Meissner	2	12	1	15
negative resistance	2	12	1	37
phase-shift	2	12	3	2
Pierce	2	12	2	13
push-pull	2	12	1	20
RC	2	12	3	2
relaxation	2	12	3	7
squegging	2	12	1	23
transistor	2	12	1	36
transitron	2	12	1	39
tuned anode	2	12	1	12
tuned anode-crystal grid	2	12	2	10
tuned anode-tuned grid	2	12	1	40
tuned grid	2	12	1	14
variable frequency	2	12	1	32
v.h.f.	2	12	1	33
Wien bridge	2	12	3	6
Oscilloscope	3	18	3	2
Out-of-phase current	1	5	2	35
Output devices, computer	3	20	2	75
Overload relay	2	9	3	26
Oxide-coated emitter	2	8	1	13

P

	BOOK	SECT.	CHAP.	PARA.
Padding	3	14	2	21
Paper tape	3	20	2	74