A Level Science Applications Support Booklet: Physics

Updated October 2009

Contents List

Introduction 1
Gathering and Communicating Information 2
28. Direct Sensing 2
29. Remote Sensing 13
30. Communicating Information 27

A LEVEL SCIENCE APPLICATIONS SUPPORT BOOKLET: PHYSICS

Introduction

Too often the study of Physics at A level can seem theoretical and abstract. The connections between Physics and real life can seem remote.

In reality, Physics is not a purely abstract subject. Like other science subjects, Physics has a "pure" theoretical side and an applied side. The principles of Physics are applied in a vast range of contexts, from the building of bridges to the design of integrated circuits. Much of the technological revolution has its foundations in applied Physics.

To ensure that the syllabus retains a balance between pure and applied Physics, there is a whole section at the end of the syllabus on Applications of Physics. It is at the end of the syllabus because the theoretical principles have to be learned and understood first if the applications are to be understood.

This booklet has been written to support teachers and students as they follow the Applications of Physics part of the syllabus.

In the booklet, each learning outcome is printed in italics and is followed by a detailed explanation. These explanations have been written by examiners and it is hoped that they will help to illustrate the level of detail that students are expected to master.

It should be stressed that this booklet is not a replacement to the syllabus. While it is hoped that the booklet will help to make the syllabus content clearer to students and teachers, it should not be read as an authoritative guide as to what is and is not included in the syllabus. The examination papers will assess the syllabus, not this booklet.

The sections of the booklet are numbered in the same way as the sections of the syllabus, so that the first part of this booklet is (perhaps rather unusually) section 28. The learning outcomes are covered in syllabus order.

The overarching theme of the Applications of Physics part of the syllabus is "Gathering and Communicating Information". This is in three sections, as follows.

- Section 28, *Direct Sensing*, covers the electronics necessary to measure temperature, light intensity or strain; to detect sound signals; to amplify signals; and to connect sensors to circuits.
- Section 29, *Remote Sensing*, covers some of the ways in which medical physicists obtain information about the inside of the body without surgery, by using X-rays, ultrasound and NMR; and how the information can be converted into images of the inside of the body.
- Section 30, *Communicating Information*, covers some of the ways in which information is communicated using radio waves, optic fibres, satellites and mobile phones.

These three sections are interconnected. The information that is communicated from one place to another can come from sensors, microphones or scanners. Communications systems contain amplifier circuits. Ultrasound scanners and microphones both use piezoelectric transducers. The equations for the attenuation of X-rays in matter and the attenuation of a signal in a wire are equivalent. Because of the many links, examination papers will often contain questions assessing more than one section.

The Applications of Physics section of the syllabus forms approximately 12% of the full Advanced Level course. **It follows that teachers should dedicate about one-eighth of their teaching time to the topics outlined in the following pages.** Experience shows that students who are left to work through this booklet on their own, without support, supervision or tuition from their teacher, usually do not perform well in the examination.

It is hoped that this booklet will be used in conjunction with a variety of other sources of information, perhaps including visits to hospitals and communications centres, guest speakers, practical work, the internet, textbooks and videos.

Gathering and Communicating Information

28. Direct Sensing

(a) Candidates should be able to show an understanding that an electronic sensor consists of a sensing device and a circuit that provides an output voltage.

Electronic sensors have many different applications in modern-day life. Frequently we take these sensors for granted. For example, the small red indicator lamp fitted to an electrical appliance that glows when the mains supply has been switched on. Other sensor circuits are more sophisticated and could, for example indicate a temperature or a light intensity level.

An electronic sensor consists of a sensing device and, usually, some form of electrical circuit connected to it. The sensing device could be for example, a light-dependent resistor (LDR) so that light intensity may be monitored (see 28(b)) or a strain gauge so that the strain experienced by a sample of material may be measured (see 28(e)). The sensing device changes one of its physical properties (e.g. resistance) with a change in whatever is to be monitored.

In order that the information gathered by the sensing device may be communicated, the change in its physical property must be processed so that an output device will indicate this change. This output device could be, for example, a simple indicator lamp or a digital meter. The output device will respond to a change in voltage. Consequently, the sensing device is connected to the output device via an electrical circuit (a processing unit) that will provide a voltage as its output. This is illustrated in Fig. 1.1.

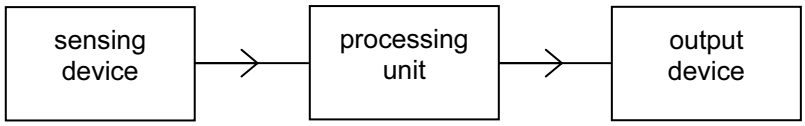


Fig. 1.1

(b) Candidates should be able to show an understanding of the change in resistance with light intensity of a light-dependent resistor (LDR).

A light-dependent resistor (LDR) may be made by sandwiching a film of cadmium sulphide between two metal electrodes. Typically, in moonlight, its resistance is about 1 MΩ and in sunlight, 100 Ω. The symbol for an LDR is shown in Fig. 1.2.

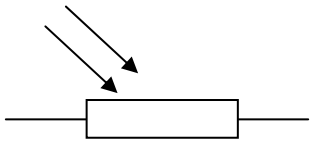


Fig. 1.2

The resistance of an LDR is constant at constant light intensity.

An LDR is sensitive to changes in light intensity. Note that the change in resistance with change in light intensity is not linear.

(c) Candidates should be able to sketch the temperature characteristic of a negative temperature coefficient thermistor.

The resistance of most substances does change slightly with a change in temperature. However, a thermistor is a device that is manufactured in various shapes and sizes using the oxides of different metals so that there is a significant change in resistance with temperature. The symbol for a thermistor is shown in Fig. 1.3.

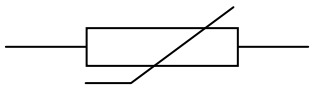


Fig. 1.3

Negative temperature coefficient thermistors have a resistance that becomes less as the temperature of the thermistor rises. The change in resistance R with temperature θ for a typical thermistor is illustrated in Fig. 1.4.

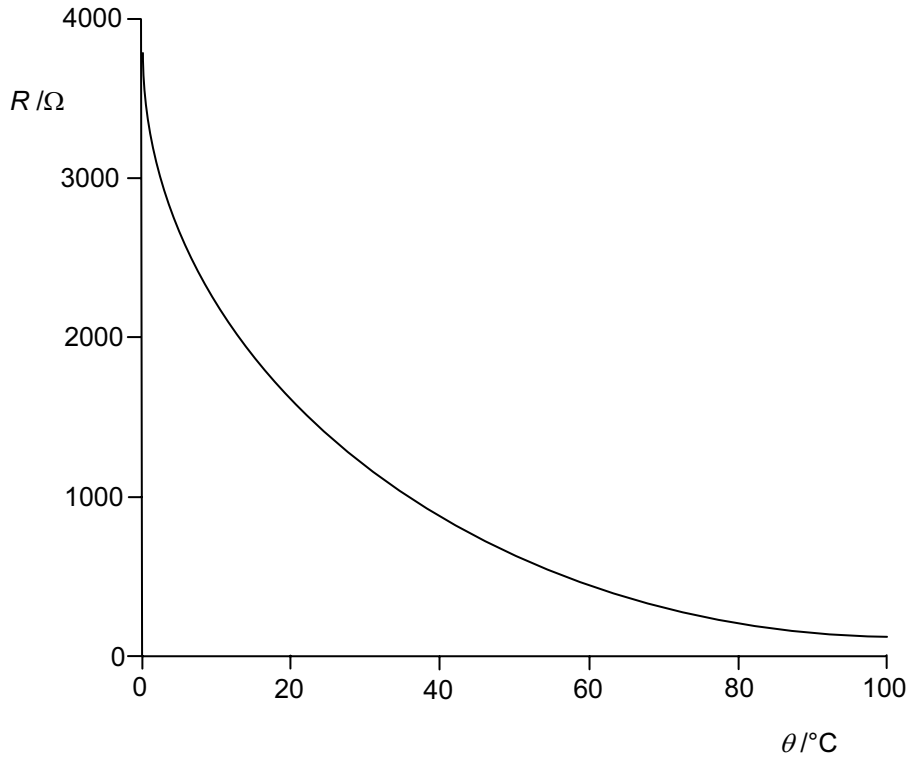


Fig. 1.4

It can be seen that there is a comparatively large change in resistance with temperature but this change is non-linear.

(d) Candidates should be able to show an understanding of the action of a piezo-electric transducer and its application in a simple microphone.

A transducer is any device that converts energy from one form to another.

Piezo-electric crystals such as quartz have a complex ionic structure. When the crystal is unstressed, the centres of charge of the positive and the negative ions bound in the lattice of the piezo-electric crystal coincide. If, however, pressure is applied to the crystal, the crystal will distort and the centres of charge for the positive and negative ions will no longer coincide. A voltage will be generated across the crystal. The effect is known as the piezo-electric effect (see also the section on 29*(h)*).

Electrical connections can be made to the crystal if opposite sides of the crystal are coated with a metal. The magnitude of the voltage generated depends on the magnitude of the pressure applied to the crystal. The polarity of the voltage depends on whether the crystal is compressed or expanded (increase or decrease in the applied pressure).

A sound wave consists of a series of compressions and rarefactions. If the wave is incident on a piezo-electric crystal, a varying voltage across the crystal will be produced. This voltage can be amplified. The crystal and its amplifier act as a simple microphone.

(e) Candidates should be able to describe the structure of a metal-wire strain gauge.

(f) Candidates should be able to relate extension of a strain gauge to change in resistance of the gauge.

A strain gauge is made by sealing a length of very fine wire in a small rectangle of thin plastic, as shown in Fig. 1.5.

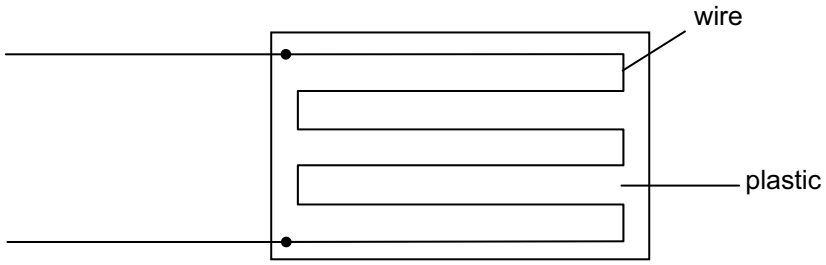


Fig. 1.5

When the plastic is stretched (the plastic experiences a strain), the wire will also be stretched. This causes the wire's length to increase and its cross-sectional area to decrease slightly. Both these changes cause the resistance of the wire to increase. Strain gauges are usually glued very securely to the material that is under test.

The resistance R of a wire of length L and of uniform cross-sectional area A is given by the expression

$$R = \rho L / A,$$

where ρ is the resistivity of the material of the wire.

Assuming that, when the wire extends by a small amount ΔL, the change in the cross-sectional area is negligible, the new resistance will be given by

$$(R + \Delta R) = \rho(L + \Delta L) / A,$$

where ΔR is the change in the resistance.

Subtracting these two expressions,

$$\Delta R = \rho \Delta L / A$$

or,

$$\Delta R \propto \Delta L.$$

Thus the strain which is proportional to the extension ΔL is also proportional to the change in resistance ΔR. Note that the cross-sectional area A is assumed to be constant.

(g) Candidates should be able to show an understanding that the output from sensing devices can be registered as a voltage.

In 28*(a)*, it was stated that a sensing device is usually connected to an electrical circuit. This circuit is designed to provide a voltage that will control an output device (see 28*(n)(o)* and *(p)*).

Where a sensing device gives rise to a change in resistance, this change in resistance can be converted into a voltage change using a potential divider, as shown in Fig. 1.6.

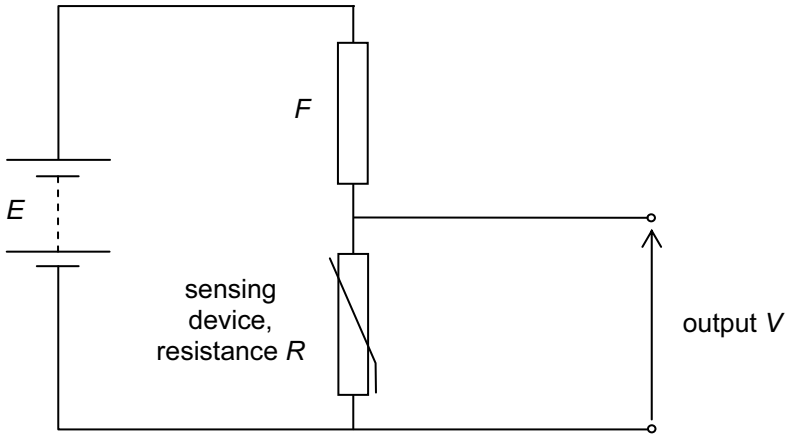


Fig. 1.6

The cell of e.m.f. E and negligible internal resistance is connected in series with a fixed resistor of resistance F and the sensing device of resistance R. The output voltage V is given by

$$V = \frac{R}{F+R} \times E\,.$$

The magnitude of the output voltage V at any particular value of resistance R of the sensing device is dependent on the relative values of R and F. A change in R will give rise to a change in V. If the resistance R decreases, then the output V will also decrease. However, connecting the output across the fixed resistor would mean that V increases when R decreases.

(h) Candidates should be able to recall the main properties of the ideal operational amplifier (op-amp).

In some applications, the change in output voltage from the potential divider (see the section on 28*(g)*) may be small. Any small change can be amplified using an electrical circuit incorporating an operational amplifier (op-amp).

An operational amplifier is an integrated circuit of about twenty transistors together with resistors and capacitors, all formed on a small slice of silicon. The slice is sealed in a package from which emerge connections to the external circuit. Some of these connections and the op-amp symbol are shown in Fig. 1.7.

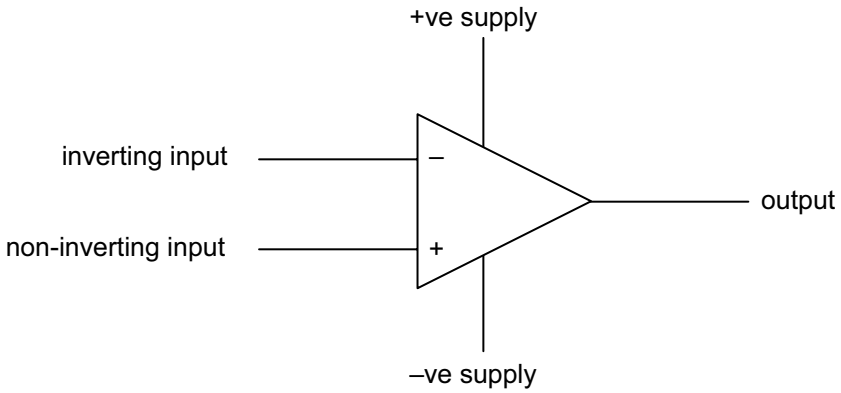


Fig. 1.7

When connected to appropriate power supplies, an op-amp produces an output voltage V_{out} that is proportional to the difference between the voltage V^+ at the non-inverting input and the voltage V^- at the inverting input.

$$V_{out} = A_0 (V^+ - V^-),$$

where A_0 is the open-loop gain of the op-amp.

The ideal operational amplifier (op-amp) has the following properties:

- infinite input impedance (i.e. no current enters or leaves either of the inputs);
- infinite open-loop gain (i.e. if there is only a very slight difference between the input voltages, the output will be saturated - the output will have the same value as the supply voltage);
- zero output impedance (i.e. the whole of the output voltage is provided across the output load);
- infinite bandwidth (i.e. all frequencies are amplified by the same factor);
- infinite slew rate (i.e. there is no delay between changes in the input and consequent changes in the output).

Real operational amplifiers do deviate from the ideal. In practice, the input impedance is usually between $10^6\ \Omega$ and $10^{12}\ \Omega$ and the output impedance is about $10^2\ \Omega$. The open-loop gain is usually about 10^5 for constant voltages. The slew rate (about 10 $V\,\mu s^{-1}$) and bandwidth are not infinite.

(i) Candidates should be able to deduce, from the properties of an ideal operational amplifier, the use of an operational amplifier as a comparator.

When an operational amplifier is used in a circuit, it is usually connected to a dual, or split, power supply. Such a supply can be thought to be made up of two sets of batteries, as shown in Fig. 1.8.

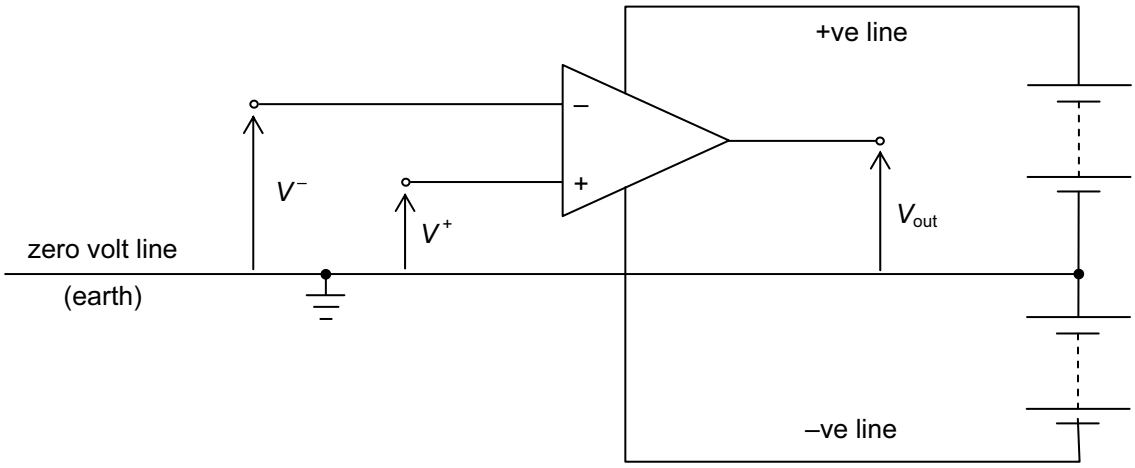


Fig. 1.8

The common link between the two sets of batteries is termed the zero-volt, or earth, line. This forms the reference line from which all input and output voltages are measured. Connecting the supplies in this way enables the output voltage to be either positive or negative.

Fig. 1.8 shows an input V^- connected to the inverting input and an input V^+ connected to the non-inverting input. The output voltage V_{out} of the op-amp is given by

$$V_{out} = A_0 (V^+ - V^-),$$

where A_0 is the open-loop gain (typically 10^5 for d.c. voltages).

Consider the examples below.

Example 1: +ve supply line = +9.0 V
–ve supply line = –9.0 V
V^+ = 1.4 V
V^- = 1.3 V

Substituting into the above equation,

$$V_{out} = 10^5 \times (1.4 - 1.3) = 10\,000\,V$$

Obviously, this answer is not possible because, from energy considerations, the output voltage can never exceed its power supply voltage. The output voltage will be 9.0 V. The amplifier is said to be *saturated*.

Example 2: +ve supply line = +6.0 V
–ve supply line = –6.0 V
V^+ = 3.652 V
V^- = 3.654 V

Substituting,

$$V_{out} = 10^5 \times (3.652 - 3.654) = -200\,V$$

Again, the amplifier will be saturated and the output will be –6.0 V.

The examples show that, unless the two inputs are almost identical, the amplifier is saturated. Furthermore, the polarity of the output depends on which input is the larger.

If $V^- > V^+$, the output is negative.

The circuit incorporating the op-amp compares the two inputs and is known as a *comparator.*

A comparator for use with an LDR is shown in Fig. 1.9.

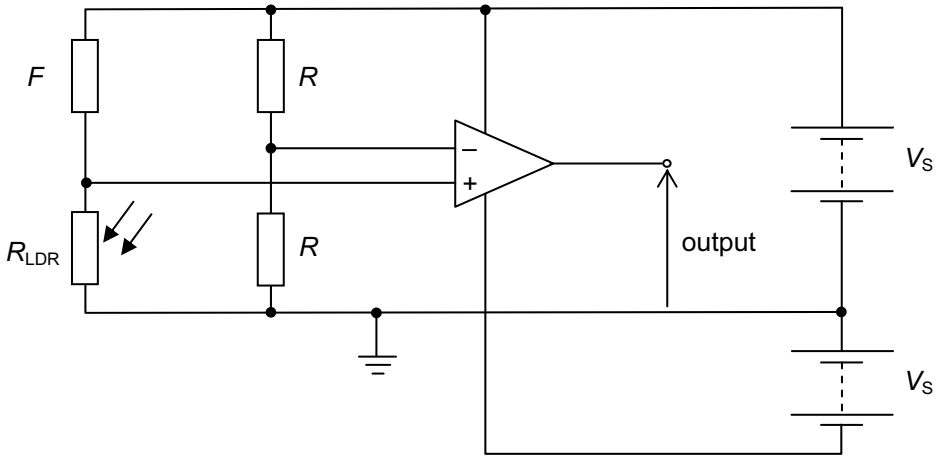


Fig. 1.9

It is usual to connect a potential divider to each of the two inputs. One potential divider provides a fixed voltage at one input while the other potential divider provides a voltage dependent on light intensity.

In Fig. 1.9, the resistors of resistance R will give rise to a constant voltage of $\frac{1}{2}V_S$ at the inverting input. The LDR, of resistance R_{LDR} is connected in series with a fixed resistor of resistance F.

If $R_{LDR} > F$ (that is, the LDR is in darkness), then $V^+ > V^-$ and the output is positive.

If $R_{LDR} < F$ (that is, the LDR is in daylight), then $V^+ < V^-$ and the output is negative.

It can be seen that by suitable choice of the resistance F, the comparator gives an output, either positive or negative, that is dependent on light intensity. The light intensity at which the circuit switches polarity can be varied if the resistor of resistance F is replaced with a variable resistor.

The LDR could be replaced by other sensors to provide alternative sensing devices. For example, use of a thermistor could provide a frost-warning device.

(j) Candidates should be able to show an understanding of the effects of negative feedback on the gain of an operational amplifier.

The process of taking some, or all, of the output of the amplifier and adding it to the input is known as *feedback*. The basic arrangement is illustrated in Fig. 1.10.

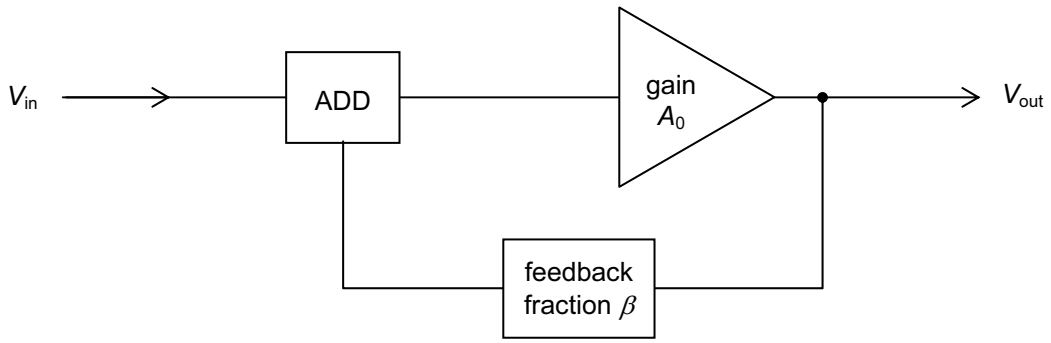


Fig. 1.10

A fraction β of the output voltage of the amplifier is fed back and added to the input voltage.

The amplifier itself amplifies by an amount A_0 whatever voltage is present at its input.

The output voltage V_{out} is given by

$$V_{out} = A_0 \times (\text{input to amplifier})$$
$$= A_0 \times (V_{in} + \beta V_{out}).$$

Re-arranging, $V_{out}(1 - A_0\beta) = A_0 \times V_{in}.$

The overall voltage gain of the amplifier with feedback is then given by

$$\frac{V_{out}}{V_{in}} = \frac{A_0}{(1 - A_0\beta)}$$

If the fraction β is negative, then the denominator must be greater than unity. This produces an amplifying system with an overall gain that is smaller than the open-loop gain A_0 of the op-amp itself. This can be achieved by feeding back part of the output to the inverting input, as illustrated in Fig. 1.11.

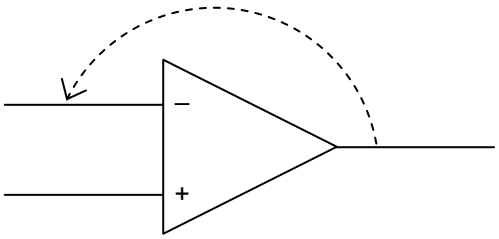


Fig. 1.11

Although negative feedback may seem to make the process of amplification rather fruitless, there are some important reasons for using negative feedback. The reduction in amplification is a small price to pay for the benefits. These benefits include

- an increase in the range of frequencies over which the gain is constant (increased bandwidth),
- less *distortion*,
- greater operating stability.

(k) Candidates should be able to recall the circuit diagrams for both the inverting and the non-inverting amplifier for single signal input.

(l) Candidates should be able to show an understanding of the virtual earth approximation and derive an expression for the gain of inverting amplifiers.

(m) Candidates should be able to recall and use expressions for the voltage gain of inverting and of non-inverting amplifiers.

In order to simplify the analysis of the circuits, the power supplies to the op-amps have not been shown. It is assumed that the op-amps are not saturated.

The inverting amplifier

The circuit for an inverting amplifier is shown in Fig. 1.12.

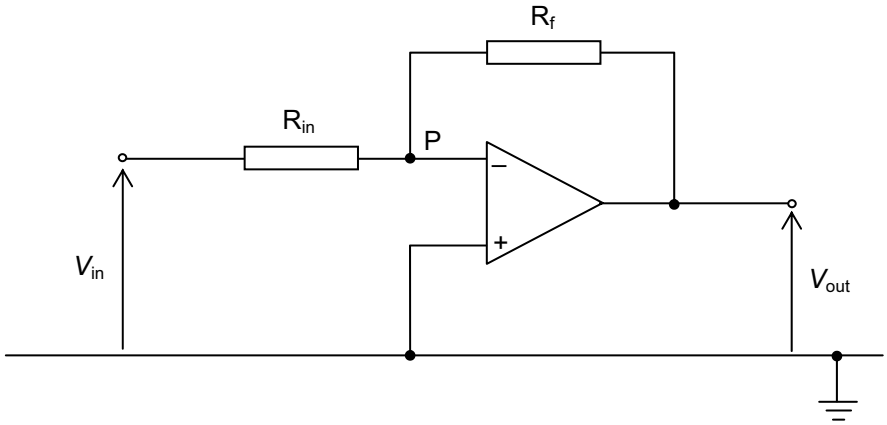


Fig. 1.12

An input signal V_{in} is applied to the input resistor R_{in}. Negative feedback is applied by means of the resistor R_f. The resistors R_{in} and R_f act as a potential divider between the input and the output of the op-amp.

In order that the amplifier is not saturated, the two input voltages must be almost the same. The non-inverting input (+) is connected directly to the zero-volt line (the earth) and so it is at exactly 0 V. Thus, the inverting input (–) must be virtually at zero volts (or earth) and for this reason, the point P is known as a *virtual earth*.

The input impedance of the op-amp itself is very large and so there is no current in either the non-inverting or the inverting inputs. This means that the current from, or to, the signal source must go to, or from, the output, as shown in Fig. 1.13.

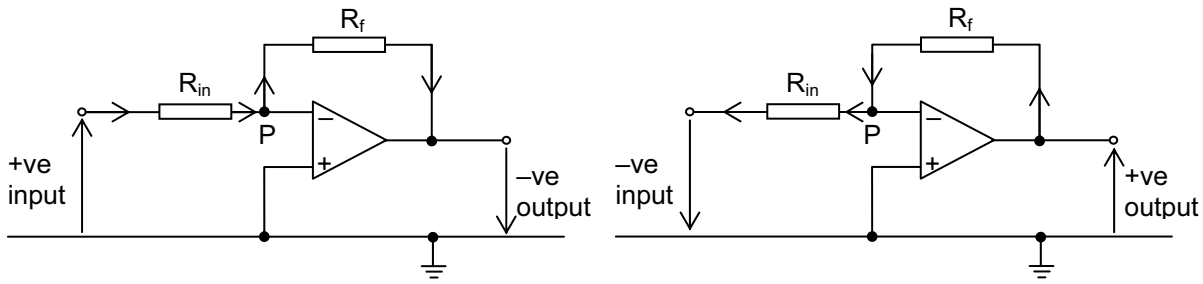


Fig. 1.13

Because the inverting input is at zero volts, a positive input gives rise to a negative output and *vice versa*. This is why the arrangement is given the name *inverting amplifier*.

Referring to Fig. 1.13, since the input resistance of the op-amp is infinite,

$$\text{current in } R_{in} = \text{current in } R_f$$

and

$$\frac{\text{p.d. across } R_{in}}{R_{in}} = \frac{\text{p.d. across } R_f}{R_f}$$

where R_{in} and R_f are the resistances of R_{in} and R_f respectively.

The potential at P is zero (virtual earth) and so

$$\frac{V_{in} - 0}{R_{in}} = \frac{0 - V_{out}}{R_f}$$

The overall voltage gain of the amplifier circuit is given by

$$\text{voltage gain} = \frac{V_{out}}{V_{in}} = -\frac{R_f}{R_{in}}$$

The non-inverting amplifier

The circuit for a non-inverting amplifier incorporating an op-amp is shown in Fig. 1.14.

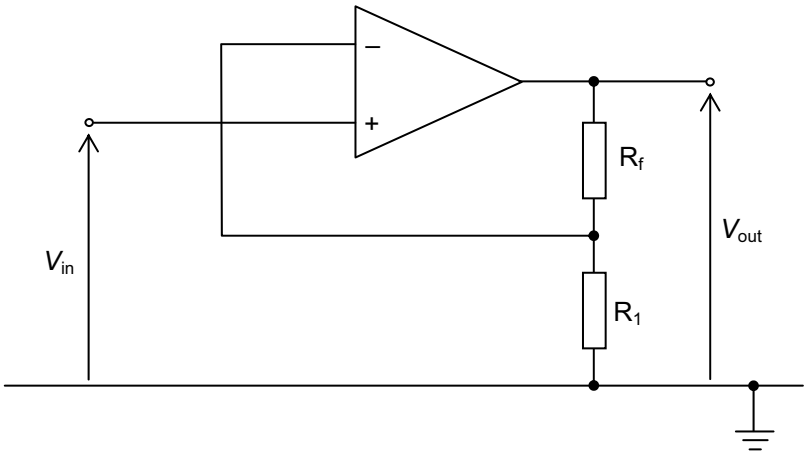


Fig. 1.14

The input signal V_{in} is applied directly to the non-inverting input. Negative feedback is provided by means of the potential divider consisting of resistors R_1 and R_f.

The voltage gain of the amplifier circuit is given by

$$\text{voltage gain} = \frac{V_{out}}{V_{in}} = 1 + \frac{R_f}{R_1}$$

where R_1 and R_f are the resistances of R_1 and R_f respectively.

The non-inverting amplifier produces an output voltage that is in phase with the input voltage.

(n) Candidates should be able to show an understanding of the use of relays in electronic circuits.

Circuits incorporating op-amps produce an output voltage. This output voltage can be used to operate warning lamps, digital meters, motors etc. However, the output of an op-amp cannot exceed a current of more than about 25 mA. Otherwise, the op-amp would be destroyed. In fact, op-amps generally contain an output resistor so that, should the output be 'shorted', the op-amp will not be damaged. In order that electronic circuits may be used to switch on and off appliances that require large currents to operate them, a *relay* may be used.

A relay is an electromagnetic switch that uses a small current to switch on or off a larger current. The small current energises an electromagnet that operates contacts, switching on or off the larger current. The symbol for a relay is shown in Fig. 1.15.

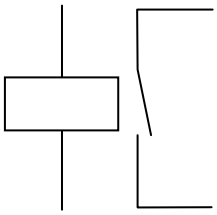


Fig. 1.15

The connection of a relay to the output of an op-amp circuit is shown in Fig. 1.16.

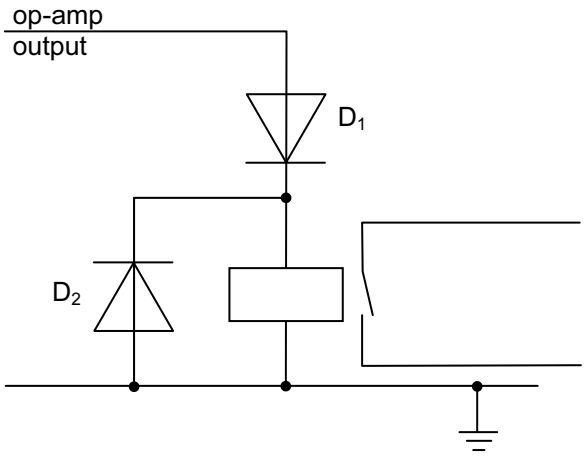


Fig. 1.16

The diode D_1 conducts only when the output is positive with respect to earth and thus the relay coil is energised only when the output is positive. When the current in the relay coil is switched off, a back e.m.f. is generated in the coil that could damage the op-amp. A diode D_2 is connected across the coil to protect the op-amp from this back e.m.f.

(o) Candidates should be able to show an understanding of the use of light-emitting diodes (LEDs) as devices to indicate the state of the output of electronic circuits.

A light-emitting diode (LED) is a diode that emits light only when it is forward biased. The symbol for an LED is shown in Fig. 1.17.

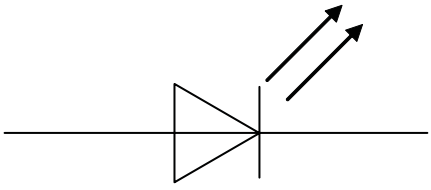


Fig. 1.17

LEDs are available that emit different colours of light, including red, green, yellow and amber. They are commonly used as 'indicators' because they have a low power consumption. Furthermore, since LEDs are solid-state devices, they are much more robust than filament lamps.

A resistor is frequently connected in series with an LED so that, when the LED is forward biased (the diode is conducting), the current is not so large as to damage the LED. A typical maximum forward current for an LED is 20 mA. Furthermore, the LED will be damaged if the reverse bias voltage exceeds about 5 V.

Fig. 1.18 is a circuit using two diodes to indicate whether the output from an op-amp is positive or negative with respect to earth.

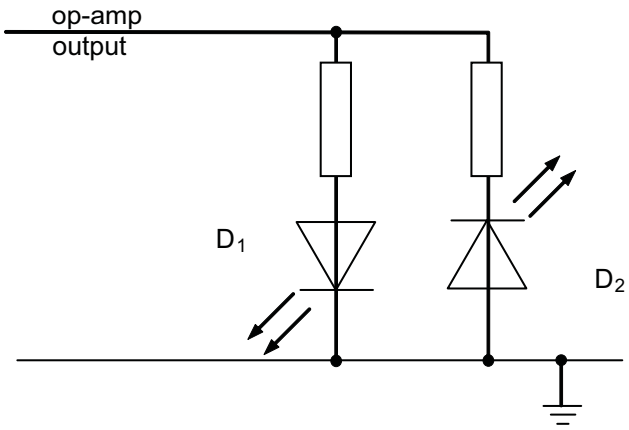


Fig. 1.18

When the output is positive with respect to earth, diode D_1 will conduct and emit light. Diode D_2 will not conduct because it is reverse biased. If the polarity of the output changes, then D_2 will conduct and emit light and D_1 will not emit light. The state of the output can be seen by which diode is emitting light. The diodes can be chosen so that they emit light of different colours.

(p) Candidates should be able to show an understanding of the need for calibration where digital or analogue meters are used as output devices.

An LED may be used to indicate whether an output is positive or negative. If the output is from a comparator, then LEDs can give information as to, for example, whether a temperature is above or below a set value. However, the LED does not give a value of the temperature reading.

Many sensors, for example, a thermistor or an LDR, are non-linear. It was seen in 28*(g)* that the sensor could be connected into a potential divider circuit so that the output of the potential divider varied with some property, for example temperature or light intensity. This variable voltage could be measured using an analogue or a digital voltmeter.

The reading on the voltmeter would vary with the property being monitored. However, the reading on the voltmeter would not vary linearly with change in the property. In order that the property can be measured, a calibration curve is required.

The reading on the voltmeter is recorded for known values of the property X. A graph is then plotted showing the variation with the property X of the voltmeter reading. The value of the property X can then be read from the graph for any particular reading on the voltmeter.

29. Remote Sensing

(a) Candidates should be able to explain in simple terms the need for remote sensing (non-invasive techniques of diagnosis) in medicine.

Historically, diagnosis consisted of two techniques – observing the patient outwardly for signs of fever, vomiting, changed breathing rate etc, and observing the patient inwardly by surgery. The first technique depended greatly on experience but was still blind to detailed internal conditions. The second quite often led to trauma and sometimes death of the patient. In earlier times there was also the significant risk of post-operative infection.

Modern diagnostic techniques have concentrated on using externally placed devices to obtain information from underneath the skin. X-rays have been used for a century. More recently, ultrasound has been used, especially in cases of pregnancy. Magnetic resonance imaging (MRI) is now becoming a frequently-used technique. Other techniques involve lasers that can shine through a finger or can be used in a very narrow tube that can be inserted into the body through various orifices.

In all these situations, the aim is to obtain detailed information concerning internal structures. This may be concerned, for example, with the functioning of an organ or the search for abnormalities. This is achieved without the need of investigative surgery and is described as a *non-invasive* technique. Non-invasive techniques are designed to present a much smaller risk than surgery and are, in general, far less traumatic for the patient.

(b) Candidates should be able to explain the principles of the production of X-rays by electron bombardment of a metal target.

X-rays are produced by bombarding metal targets with high-speed electrons. A typical spectrum of the X-rays produced is shown in Fig. 2.1.

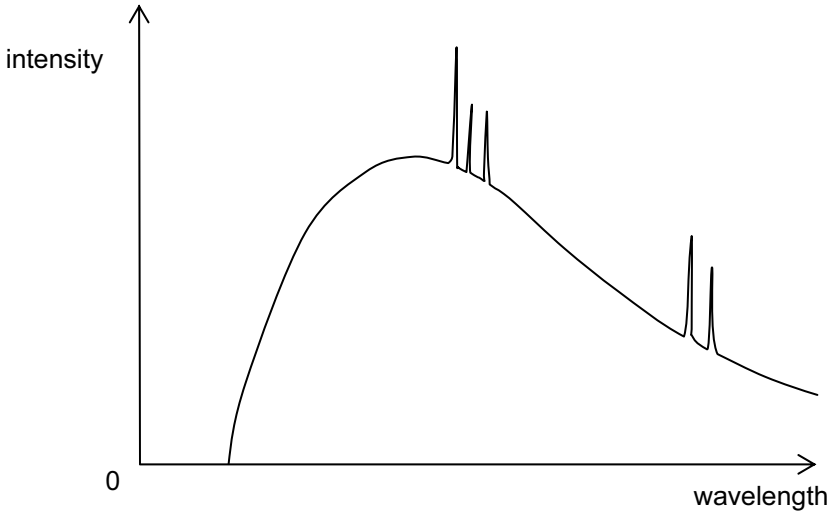


Fig. 2.1

The spectrum consists of two components. There is a continuous distribution of wavelengths with a sharp cut-off at short wavelength and also a series of high-intensity spikes that are characteristic of the target material.

Whenever a charged particle is accelerated, electromagnetic radiation is emitted. The greater the acceleration, the shorter is the wavelength of the emitted radiation. This radiation is known as Bremmstrahlung radiation. When high-speed electrons strike a metal target, large accelerations occur and the radiation produced is in the X-ray region of the electromagnetic spectrum. Since the electrons have a continuous distribution of accelerations, a continuous distribution of wavelengths of X-rays is produced. There is a minimum wavelength (a cut-off wavelength) where the whole of the energy of the electron is converted into the energy of one photon. That is,

$$\text{kinetic energy of electron} = eV = hc / \lambda,$$

where e is the charge on the electron that has moved through a potential difference V, h is the Planck constant, c is the speed of light and λ is the wavelength of the emitted X-ray photon.

As well as the continuous distribution of wavelengths, sharp peaks are observed. These peaks correspond to the emission line spectrum of the atoms of the target. The electrons that bombard the target excite orbital electrons in the lower energy levels and the subsequent de-excitation of electrons gives rise to the line spectrum.

(c) Candidates should be able to describe the main features of a modern X-ray tube, including control of the intensity and hardness of the X-ray beam.

A simplified diagram of a modern form of X-ray tube is shown in Fig. 2.2.

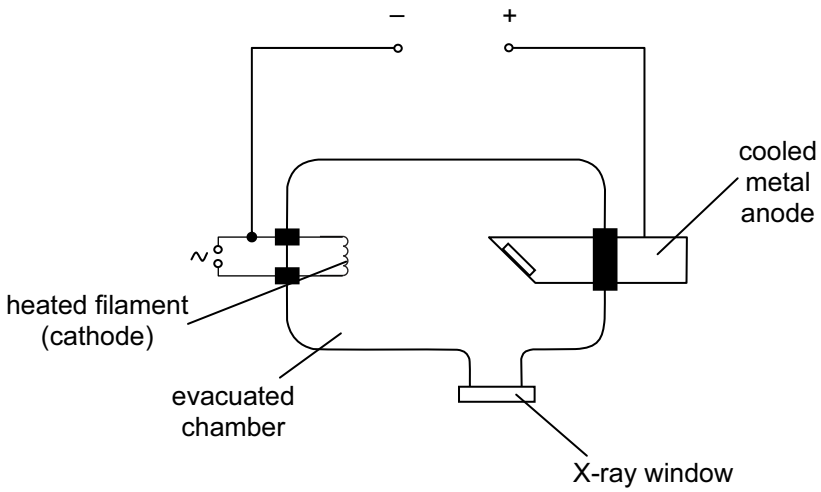


Fig. 2.2

Electrons are emitted from the heated cathode (thermionic effect). The electrons are accelerated through a large potential difference (20 kV $\rightarrow$ 100 kV for diagnosis) before bombarding a metal anode. The X-rays produced leave the tube via a 'window'. Since the majority of the energy of the electrons is transferred to thermal energy in the metal anode, the anode is either water-cooled or is made to spin rapidly so that the target area is increased. The anode is held at earth potential.

The intensity of the X-ray beam is determined by the rate of arrival of electrons at the metal target, that is, the *tube current*. This tube current is controlled by the heater current of the cathode. The greater the heater current, the hotter the filament and hence the greater the rate of emission of thermo-electrons.

The hardness of the X-ray beam (the penetration of the X-rays) is controlled by the accelerating voltage between the cathode and the anode. More penetrating X-rays have higher photon energies and thus a larger accelerating potential is required. Referring to Fig. 2.1, it can be seen that longer wavelength X-rays ('softer' X-rays) are always also produced. Indeed some X-ray photons are of such low energy that they would not be able to pass through the patient. These 'soft' X-rays would contribute to the total radiation dose without any useful purpose. Consequently, an aluminium filter is frequently fitted across the window of the X-ray tube to absorb the 'soft' X-ray photons.

(d) Candidates should be able to show an understanding of the use of X-rays in imaging internal body structures, including a simple analysis of the causes of sharpness and contrast in X-ray imaging.

X-ray radiation affects photographic plates in much the same way as visible light. A photographic plate, once exposed, will appear blackened after development. The degree of blackening is dependent on the total X-ray exposure.

X-ray photons also cause fluorescence in certain materials. The mechanism is similar to that by which visible light is produced on the screen of a cathode-ray oscilloscope.

X-ray beams are used to obtain 'shadow' pictures of the inside of the body to assist in the diagnosis or treatment of illness. If a picture is required of bones, this is relatively simple since the absorption by bone of X-ray photons is considerably greater than the absorption by surrounding muscles and tissues. X-ray pictures of other parts of the body may be obtained if there is sufficient difference between the absorption properties of the organ under review and the surrounding tissues.

The quality of the shadow picture (the image) produced on the photographic plate depends on its sharpness and contrast. Sharpness is concerned with the ease with which the edges of structures can be determined. A sharp image implies that the edges of organs are clearly defined. An image may be sharp but, unless there is a marked difference in the degree of blackening of the image between one organ and another (or between different parts of the same organ), the information that can be gained is limited. An X-ray plate with a wide range of exposures, having areas showing little or no blackening as well as areas of heavy blackening, is said to have good contrast.

In order to achieve as sharp an image as possible, the X-ray tube is designed to generate a beam of X-rays with minimum width. Factors in the design of the X-ray apparatus that may affect sharpness include

- the area of the target anode, as illustrated in Fig. 2.3,

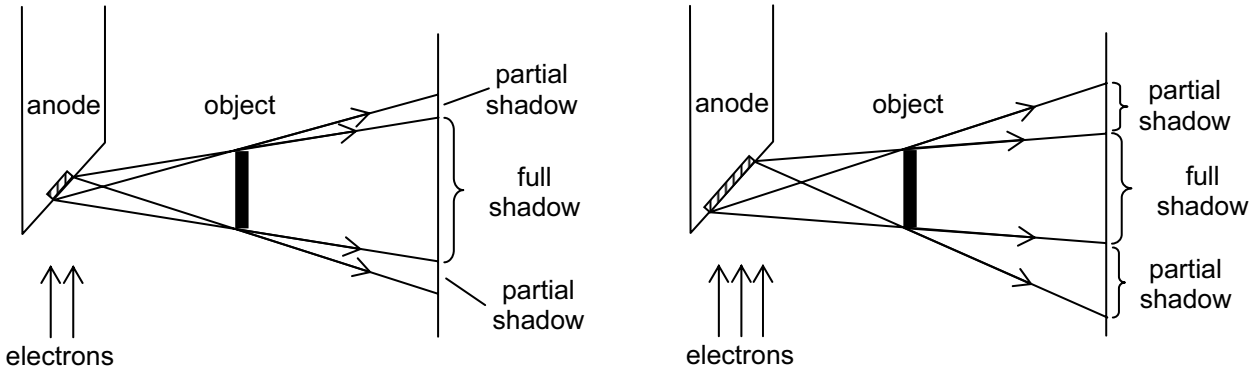


Fig. 2.3

- the size of the aperture, produced by overlapping metal plates, through which the X-ray beam passes after leaving the tube (see Fig. 2.4),

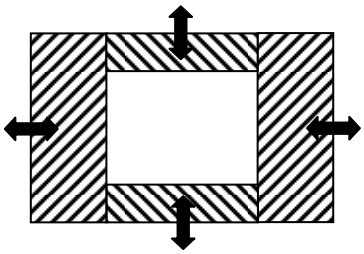


Fig. 2.4

- the use of a lead grid in front of the photographic film to absorb scattered X-ray photons, as illustrated in Fig. 2.5.

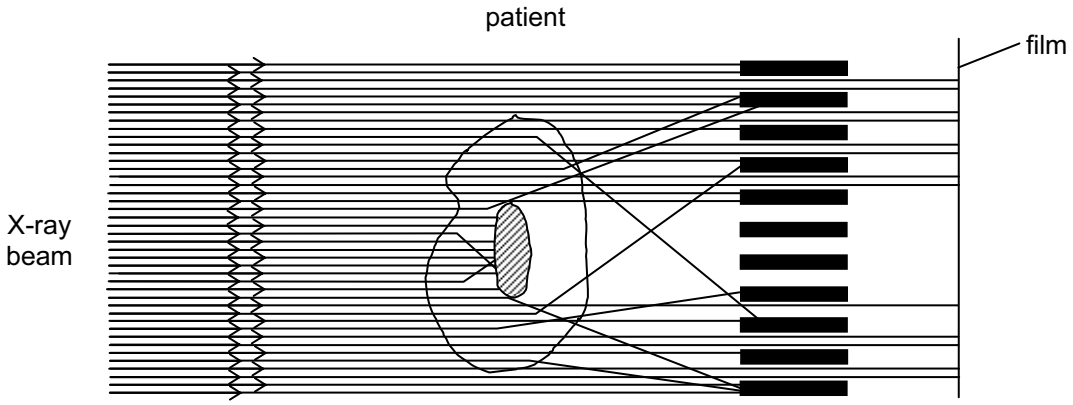


Fig. 2.5

In order to improve contrast, a 'contrast medium' may be used. For example, the stomach may be examined by giving the patient a drink containing barium sulphate. Similarly, to outline blood vessels, a contrast medium that absorbs strongly the X-radiation would be injected into the bloodstream.

The contrast of the image produced on the photographic film is affected by exposure time, X-ray penetration and scattering of the X-ray beam within the patient's body. Contrast may be improved by backing the photographic film with a fluorescent material.

(e) Candidates should be able to show an understanding of the purpose of computed tomography or CT scanning.

The image produced on an X-ray plate as outlined in the section on 29*(d)* is a 'flat image' and does not give any impression of depth. That is, whether an organ is near to the skin or deep within the body is not apparent. Tomography is a technique by which an image of a slice, or plane, of the object may be obtained.

In this technique, a series of X-ray images are obtained from different angles through one section, or slice, of the object to be examined. The images are all in the plane of the slice, as illustrated in Fig. 2.7.

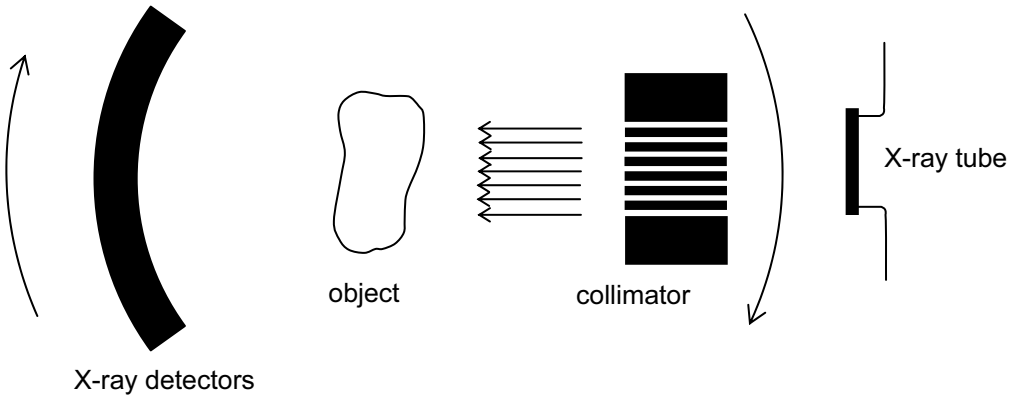


Fig. 2.7

Computer techniques make it possible to combine these images to give an image of the slice. The technique is called computed (axial) tomography or *CT scanning*.

Images of successive slices can be combined to give a three-dimensional image. The three-dimensional image can be rotated and viewed from any angle.

(f) Candidates should be able to show an understanding of the principles of CT scanning.

(g) Candidates should be able to show an understanding of how the image of an 8-voxel cube can be developed using CT scanning.

The aim of CT scanning is to make an image of a section (or slice) through the body from measurements made about its axis, as illustrated in Fig. 2.7.

The section (or slice) through the body is divided up into a series of small units called voxels. The image of each voxel would have a particular intensity, known as a pixel. The pixels are built up from measurements of X-ray intensity made along a series of different directions around the section of the body.

Suppose a section consists of four voxels with intensities as shown in Fig. 2.8.

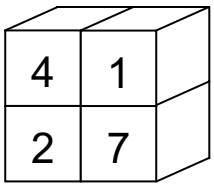


Fig. 2.8

The number on each voxel is the pixel intensity that is to be reproduced.

If a beam of X-rays is directed from the left, then detectors will give readings of 5 and 9. This allows the four voxels to be "reconstructed", as shown in Fig. 2.9.

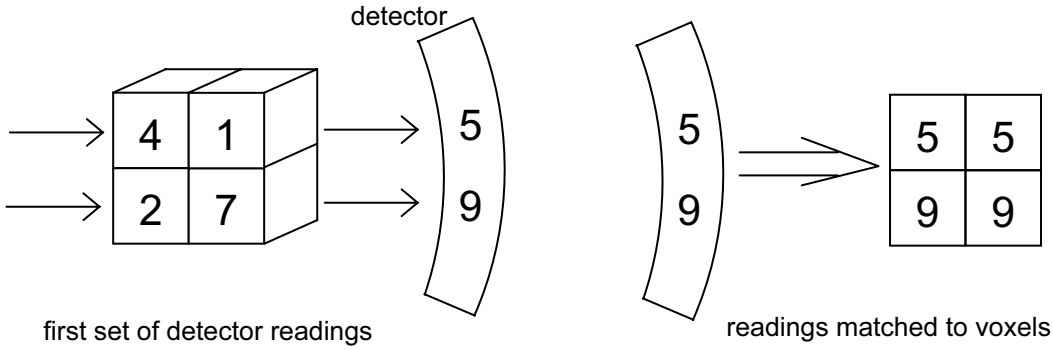


Fig. 2.9

The X-ray tube and detectors are now rotated through 45° and new detector readings are found, as shown in Fig. 2.10. These new detector readings are added to the readings already obtained for the voxels.

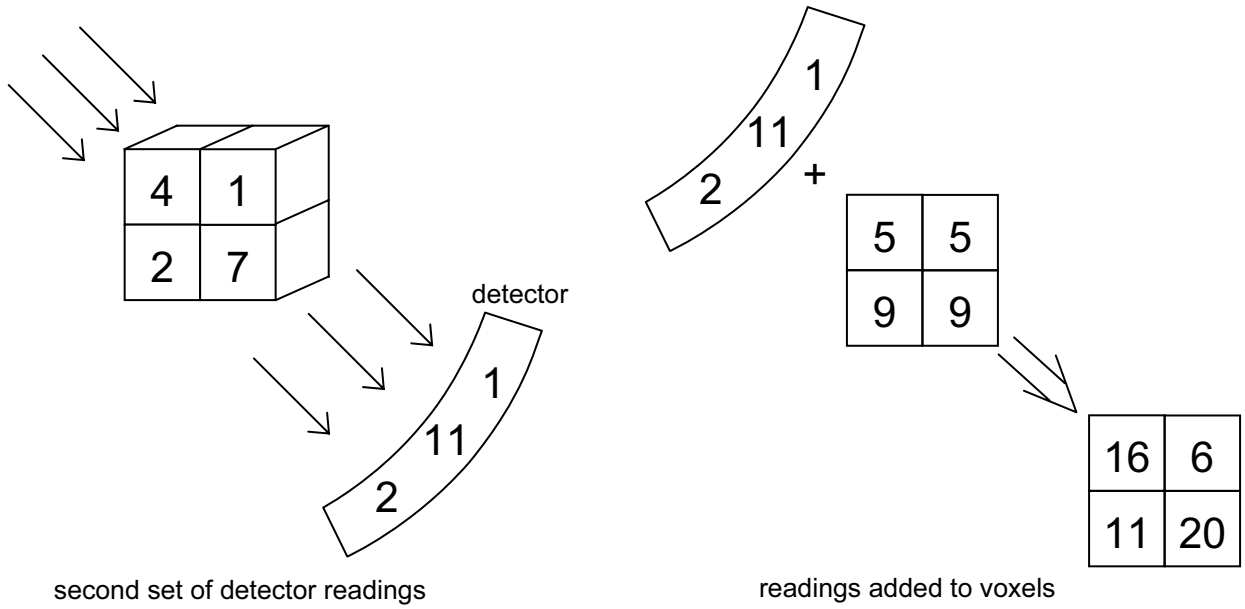
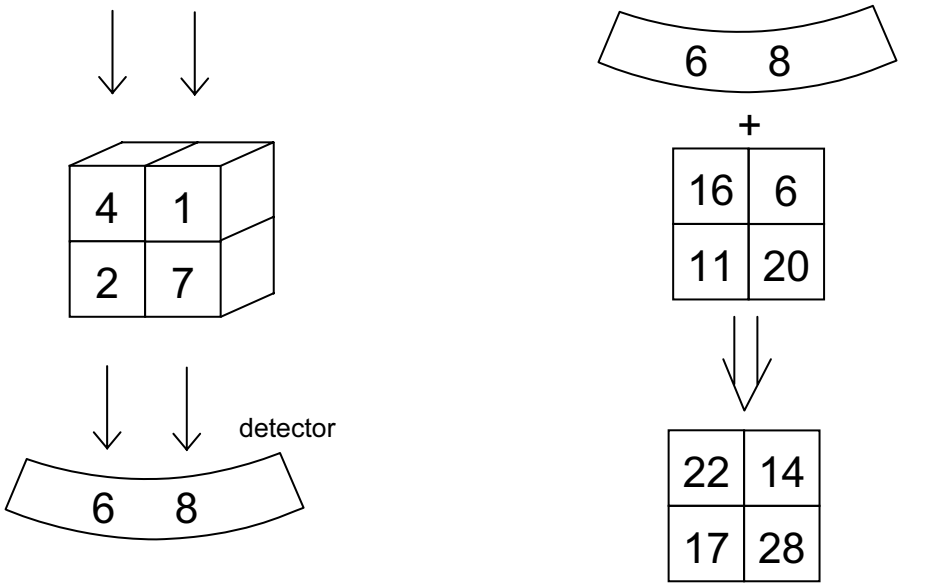


Fig. 2.10

The procedure is repeated after rotating the X-ray tube and the detectors through a further 45°. The result is shown in Fig. 2.11.

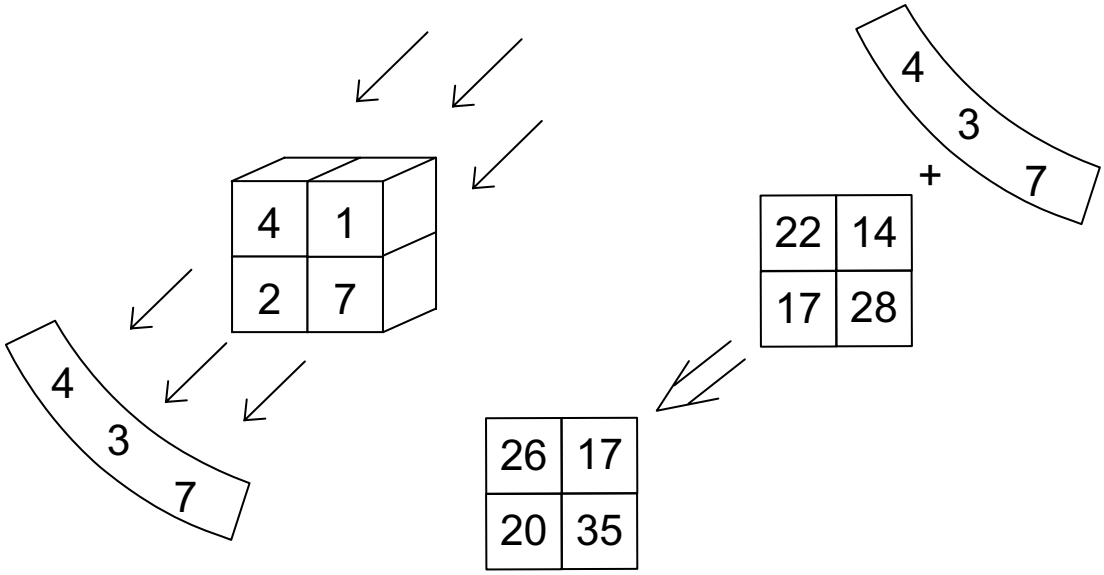


third set of detector readings

readings added to voxels

Fig. 2.11

The final images are taken after rotating the X-ray tube and the detectors through a further 45°. The result is shown in Fig. 2.12.



fourth set of detector readings

readings added to voxels

Fig. 2.12

The final pattern of pixels is shown in Fig. 2.13.

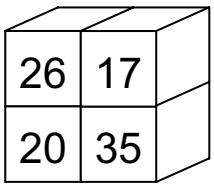


Fig. 2.13

In order to obtain the original pattern of pixels, two operations must be performed.

1. The 'background' intensity must be removed. The 'background' intensity is the total of each set of detector readings. In this case, 14 is deducted from each pixel.
2. After deduction of the 'background', the result must be divided by three to allow for the duplication of the views of the section.

These processes are illustrated in Fig. 2.14.

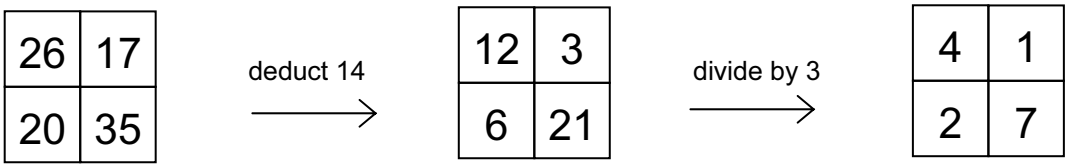


Fig. 2.14

The pattern of pixels for the section now emerges.

In practice, the image of each section is built up from many small pixels, each viewed from many different angles. The collection of the data and its construction into a display on a screen requires a powerful computer and complicated programmes. In fact, the reconstruction of each pixel intensity value requires more than one million computations. The contrast and brightness of the image of the section as viewed on the TV screen can be varied to achieve optimum results.

In order to build up an image of the whole body, the procedure would be repeated for further sections (or slices) through the body. All the data for all the sections can be stored in the computer memory to create a three-dimensional image. Views of the body from different angles may be constructed.

(h) Candidates should be able to explain the principles of the generation and detection of ultrasonic waves using piezo-electric transducers.

Ultrasonic waves may be produced using a piezo-electric transducer. The basis of this is a piezo-electric crystal such as quartz. Two opposite sides of the crystal are coated with thin layers of silver to act as electrical contacts, as illustrated in Fig. 2.15.

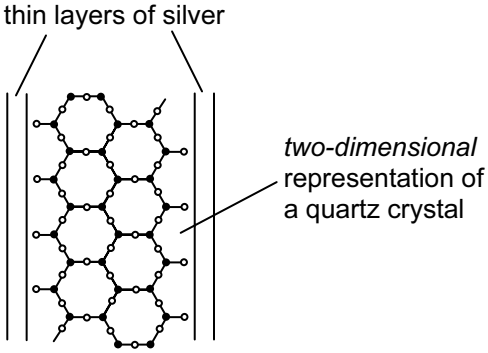


Fig. 2.15

Quartz has a complex structure made up of a large number of repeating tetrahedral silicate units, as illustrated in Fig. 2.16.

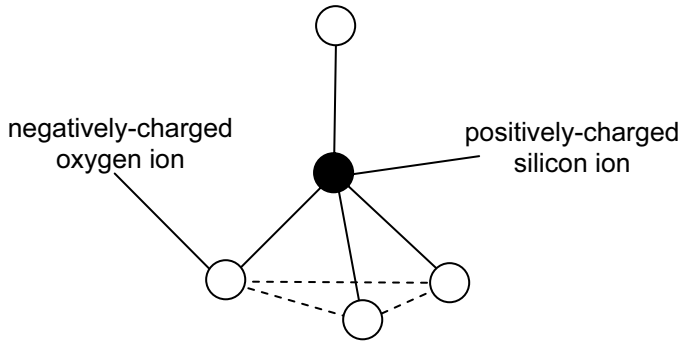
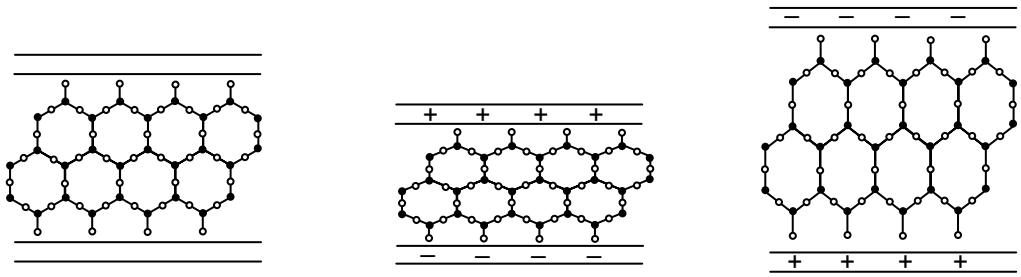


Fig. 2.16

The positions of the oxygen links are not rigidly fixed in these units, or lattices, and since the oxygen ions are negatively charged, movement can be encouraged by applying an electric field.

When the crystal is unstressed, the centres of charge of the positive and the negative ions bound in the lattice of the piezo-electric crystal coincide, so their effects are neutralised, as shown in Fig. 2.17(a).



(a) unstressed (b) compressed (c) extended

Fig. 2.17

If a constant voltage is then applied across the electrodes (i.e. across the layers of silver), the positive silicon ions are attracted towards the cathode and the negative oxygen ions towards the anode. This causes distortion of the silicate units. Depending on the polarity of the applied voltage, the crystal becomes either thinner or thicker as a result of the altered charge distribution. These effects are illustrated in Fig. 2.17(b) and Fig. 2.17(c).

An alternating voltage applied across the silver electrodes will set up mechanical vibrations in the crystal. If the frequency of the applied voltage is the same as the natural frequency of vibration of the crystal, resonance occurs and the oscillations have maximum amplitude. The dimensions of the crystal can be such that the oscillations are in the ultrasonic range (i.e. greater than 20 kHz), thus producing ultrasonic waves in the surrounding medium.

Ultrasonic transducers can also be used as receivers. When an ultrasonic wave is incident on an unstressed piezo-electric crystal, the pressure variations alter the positions of positive and negative ions within the crystal. This induces opposite charges on the silver electrodes, producing a potential difference between them. This varying potential difference can then be amplified and processed.

A simplified diagram of a typical piezo-electric transducer/receiver is illustrated in Fig. 2.18.

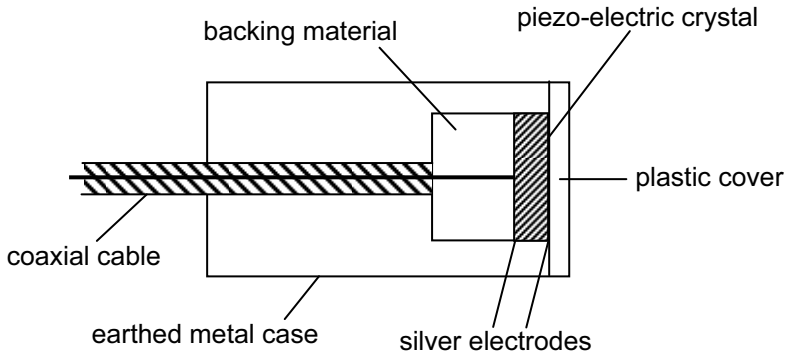


Fig. 2.18

Such devices operate in the MHz frequency range, up to a maximum of about 600 MHz.

(i) Candidates should be able to explain the main principles behind the use of ultrasound to obtain diagnostic information about internal structures.

(j) Candidates should be able to show an understanding of the meaning of specific acoustic impedance and its importance to the intensity reflection coefficient at a boundary.

In order to be able to explain the principles of the use of ultrasound in diagnosis, it is necessary to have an understanding of the reflection of ultrasound at boundaries and its absorption in media.

Ultrasound obeys the same laws of reflection and refraction at boundaries as audible sound and light. When an ultrasound wave meets the boundary between two media, some of the wave energy is reflected and some is transmitted, as illustrated in Fig. 2.19.

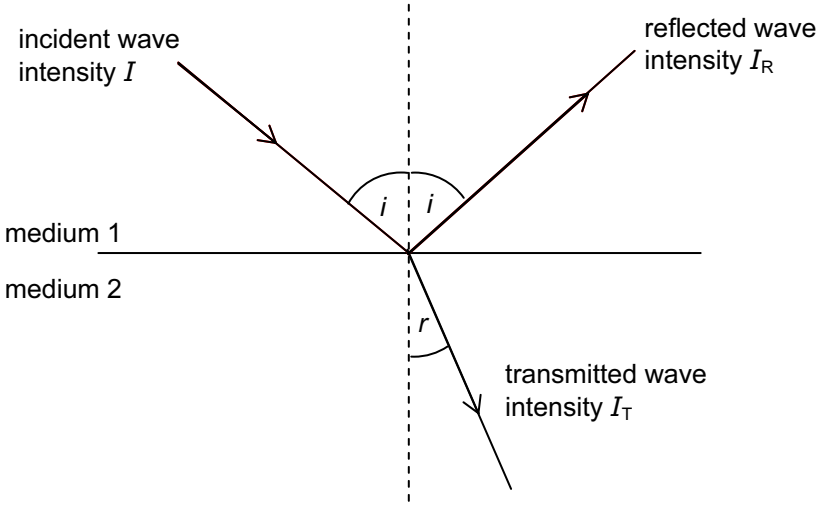


Fig. 2.19

For an incident intensity I, reflected intensity I_R and transmitted intensity I_T, then from energy considerations,

$$I = I_R + I_T.$$

The relative magnitudes of the reflected and transmitted intensities depend not only on the angle of incidence but also on the two media themselves.

For any medium, a quantity known as the *specific acoustic impedance Z* is defined as

$$Z = \rho c,$$

where c is the speed of the wave in the medium of density ρ. When a wave is incident normally on a boundary between two media having specific acoustic impedances of Z_1 and Z_2, the ratio I_R / I of the reflected intensity to the incident intensity is given by the expression

$$\frac{I_R}{I} = \frac{(Z_2 - Z_1)^2}{(Z_2 + Z_1)^2}$$

The ratio I_R / I is known as the *intensity reflection coefficient* for the boundary and is usually given the symbol α. Clearly, the value of α depends on the difference between the specific acoustic impedances of the media on each side of the boundary. Some approximate values of specific acoustic impedance Z are given in Fig. 2.20.

medium	$Z = \rho c$ / kg m^{-2} s^{-1}
air	430
quartz	1.52×10^7
water	1.50×10^6
blood	1.59×10^6
fat	1.38×10^6
muscle	1.70×10^6
soft tissue	1.63×10^6
bone	$(5.6 - 7.8) \times 10^6$

Fig. 2.20

It can be seen that the intensity reflection coefficient is very large for ultrasound entering or leaving the human body (a boundary between air and soft tissue). In order that ultrasound waves may be transmitted from the transducer into the body (and also return to the transducer after reflection from the boundaries of body structures), it is important to ensure that there is no air trapped between the transducer and the skin. This is achieved by means of a coupling medium such as a gel that fills any spaces between the transducer and the skin.

A second factor that affects the intensity of ultrasonic waves passing through a medium is absorption. As a wave travels through a medium, energy is absorbed by the medium and the intensity of a parallel beam decreases exponentially. The temperature of the medium rises. The heating effect caused by ultrasound of suitable frequencies is, in fact, used in physiotherapy to assist with recovery from sprained joints.

Fig. 2.21 illustrates a parallel beam of ultrasound waves of intensity I_0 incident on a medium of thickness x.

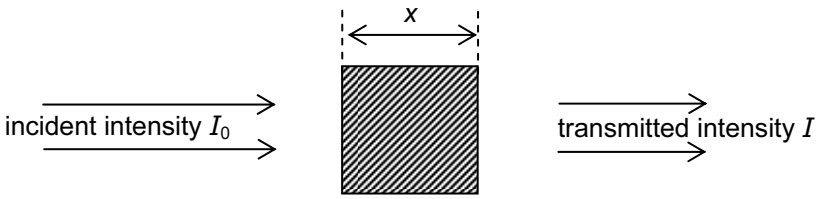


Fig. 2.21

The intensity I of the beam after passing through the medium is related to the incident intensity by the expression

$$I = I_0 e^{-kx},$$

where k is a constant for the medium referred to as the *absorption coefficient*. This coefficient is dependent on the frequency of the ultrasound. Fig. 2.22 gives some values for ultrasound of frequency 1 MHz.

medium	absorption coefficient / m^{-1}
air	120
water	0.02
muscle	23
bone	130

Fig. 2.22

In order to obtain diagnostic information about internal body structures, the transducer is placed in contact with the skin, with a gel acting as a coupling medium. The gel reduces the size of the impedance change between boundaries at the skin and thus reduces reflection at the skin. Short pulses of ultrasound are transmitted into the body. These pulses are partly reflected and partly transmitted at boundaries between media in the body (e.g. a fat – muscle boundary). The reflected pulses return to the transducer where they are detected and transformed into voltage pulses. These voltage pulses can then be amplified and processed to give an image on an oscilloscope screen. Two techniques, A-scan and B-scan, are in common use for the display of an ultrasound scan.

The A-scan system basically measures the distance of different boundaries from the transducer, with the transducer held in one position. A short burst of ultrasound is transmitted to the body through the coupling medium. At each boundary between different media in the body, some ultrasound is reflected and some is transmitted. The reflected pulse is picked up be the transducer which now acts as a receiver. The signal is amplified and displayed on a cathode-ray oscilloscope (c.r.o.). The reflected pulse also meets boundaries as it returns to the transducer. This causes some of the energy of the reflected pulse to be lost and energy is also lost due to absorption in the media. Consequently, echoes from deeper in the body tend to be of lower intensity. To compensate for this, the later an echo is received at the transducer, the more it is amplified before display on the c.r.o. A vertical line appears on the screen each time an echo is received. The time-base on the X-plates is adjusted so that all of the reflections are seen on the screen for one scan (pulse). The distance between boundaries can be calculated if the speed of ultrasound in the various media is known. An example of an A-scan for the brain is shown in Fig. 2.23.

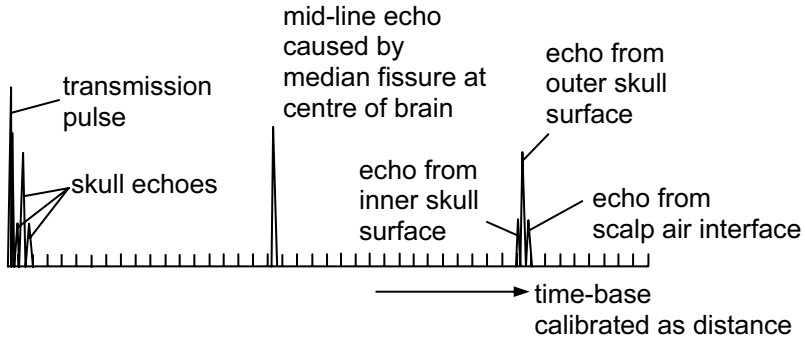


Fig. 2.23

The B-scan technique basically combines a series of A-scans, taken from a range of different angles, to form a two-dimensional picture. As before, each A-scan corresponds to a single ultrasound pulse being emitted by the transducer and producing a series of reflected pulses from boundaries within the body.

The ultrasound probe for a B-scan consists of a series of small crystals, each having a slightly different orientation. The signals received from the crystals in the probe are processed by a computer. Each reflected pulse is shown as a bright spot in the correct orientation of the crystal on the screen of a c.r.o. Consequently, the completed pattern of spots from all the crystals in the probe builds up into a two-dimensional representation of the boundary positions in the body being scanned. This image may be photographed or stored in the computer memory.

The main advantage of ultrasonic scanning is that the health risk factor to human patients, and to those operating the system, is considered to be very much less than in X-ray diagnosis. Other advantages are that the equipment may be portable and is relatively simple to use. With higher frequencies, smaller features within the body can be identified. Modern techniques allow low intensity echoes to be detected and as a result, boundaries between soft tissues, as well as between hard and soft tissues, may be detected.

(k) Candidates should be able to recall and solve problems by using the equation $I = I_0 e^{-\mu x}$ for the attenuation of X-rays and of ultrasound in matter.

When the energy of an X-ray beam radiates from the source in all directions in a vacuum, the intensity decreases in proportional to the inverse of the square of the distance from the source. This is a consequence of the energy being 'spread' over the surface of a sphere of radius *r* having surface area $4\pi r^2$. Thus, in a vacuum, $I \propto I_0/r^2$. The law also applies approximately to X-rays in air since there is little absorption of X-rays by air.

In a medium where absorption processes are occurring, the intensity I of a parallel beam decreases by a constant fraction in passing through equal small thicknesses of the medium. This gives rise to an exponential decrease in the intensity of the transmitted beam. For a parallel beam of radiation of initial intensity I_0 passing through a thickness *x* of a medium, then the transmitted intensity I is given by

$$I = I_0 e^{-\mu x},$$

where μ is a constant for the medium that is dependent on photon energy. The unit of μ is mm^{-1} or cm^{-1} or m^{-1}. μ is referred to as the *linear absorption coefficient* or *linear attenuation coefficient*.

The variation with thickness *x* of an absorber of the percentage transmission of a parallel beam of X-ray radiation is illustrated in Fig. 2.6.

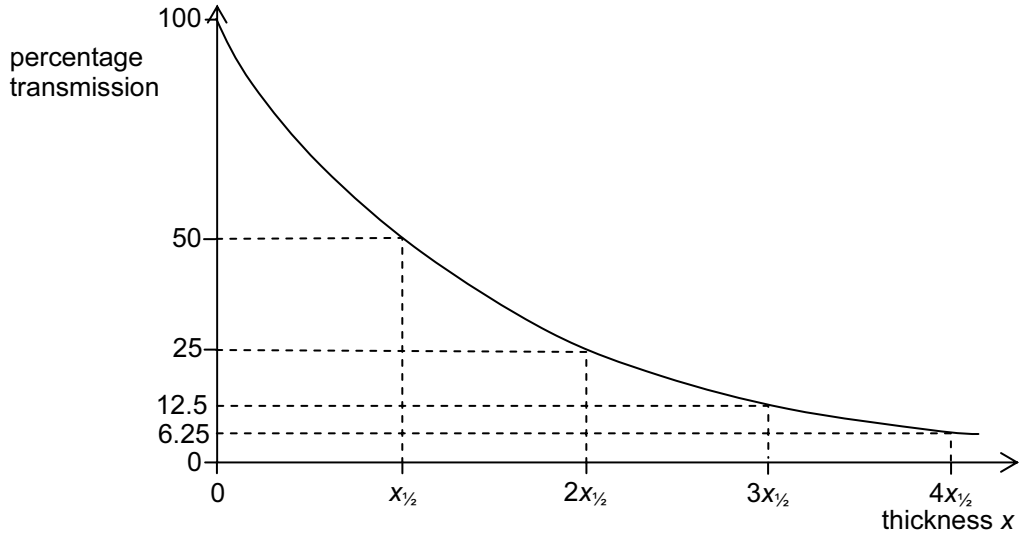


Fig. 2.6

The thickness of the medium required to reduce the transmitted intensity to one half of its initial value is a constant and is known as the *half-value thickness* $x_{½}$ or *HVT*. The half-value thickness $x_{½}$ is related to the linear absorption coefficient μ by the expression

$$x_{½} \times \mu = \ln 2.$$

In practice, $x_{½}$ does not have a precise value as it is constant only when the beam has photons of one energy only.

The intensity of a parallel beam of ultrasound passing through a medium also decreases exponentially with thickness. Thus, ultrasound has a linear absorption coefficient (and half-value thickness) that is dependent on the medium and also the frequency of the ultrasound. The intensity I decreases according to the expression $I = I_0 e^{-\mu x}$.

(l) Candidates should be able to explain the main principles behind the use of magnetic resonance to obtain diagnostic information about internal structures.

(m) Candidates should be able to show an understanding of the function of the non-uniform magnetic field, superimposed on the large constant magnetic field, in diagnosis using magnetic resonance.

Many atomic nuclei behave as if they possess a 'spin'. Such nuclei have an odd number of protons and/or an odd number of neutrons. Their 'spin' causes the nuclei of these atoms to behave as tiny magnets. If an external magnetic field is applied to these atoms, they will tend to line up in the magnetic field. This alignment is not perfect and the nuclei rotate about the direction of the field as they spin. This type of motion is referred to as *precession*. The motion is similar to the motion of a top spinning in a gravitational field.

The frequency of precession (the Lamour frequency) depends on the nature of the nucleus and the strength of the magnetic field. The Lamour frequency is found to lie in the radio-frequency (RF) region of the electromagnetic spectrum.

If a short pulse of radio waves of frequency equal to the Lamour frequency is applied, the atoms will resonate, absorbing energy. When the pulse ends, the atoms will return to their original equilibrium state after a short period of time, called the *relaxation time*. In so doing, RF radiation is emitted by the atoms. There are, in fact, two relaxation processes and it is the times between these that forms the basis of magnetic resonance imaging (MRI).

Examples of nuclei that show this effect include hydrogen, carbon and phosphorus. Because of its abundance in body tissue and fluids, hydrogen is the atom used in this scanning tehnique.

A schematic diagram of a magnetic resonance (MR) scanner is shown in Fig. 2.24.

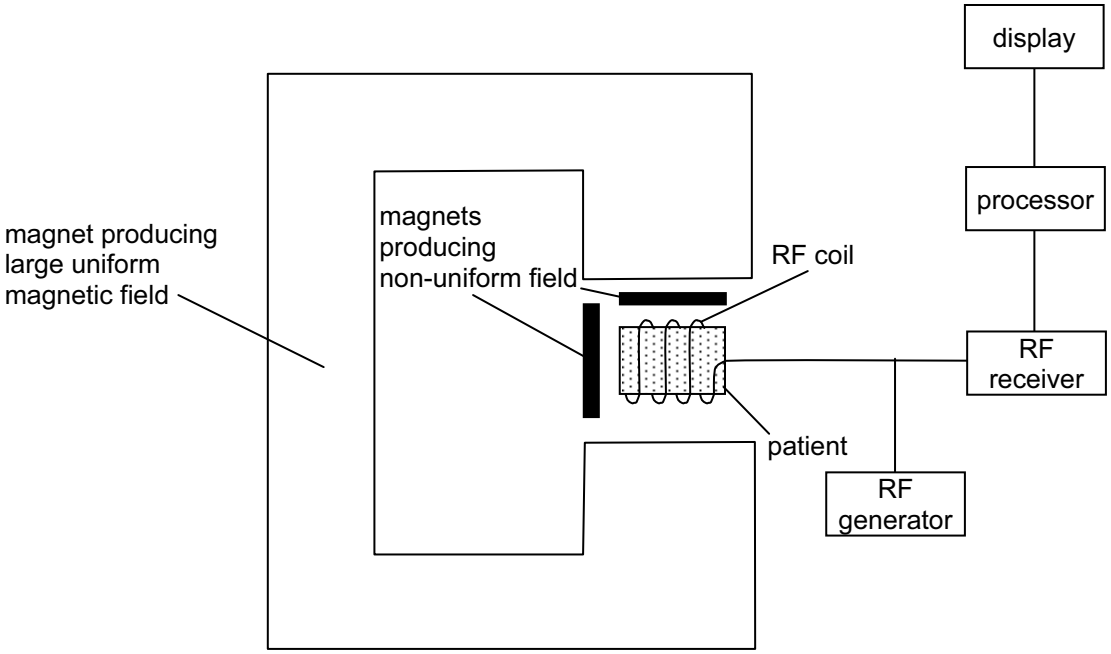


Fig. 2.24

The person under investigation is placed between the poles of a very large magnet that produces a uniform magnetic field in excess of 1 tesla. All the hydrogen nuclei within the person would have the same Lamour frequency because this frequency is dependent on the magnetic field strength. In order to locate a particular position of hydrogen atoms within the person, a non-uniform magnetic field is also applied. This non-uniform field is accurately calibrated so that there is a unique value of magnetic field strength at each point in the person. This value, coupled with the particular value of the Lamour frequency, enables the hydrogen nuclei to be located.

Radio-frequency pulses are transmitted to the person by means of suitable coils. These coils are also used to detect the RF emissions from the patient. The received emissions are processed in order to construct an image of the number density of hydrogen atoms in the patient. As the non-uniform magnetic field is changed, then atoms in different parts of the person will be detected. One such MR scan, which shows a section through the spine and back muscles, is shown in Fig. 2.25.

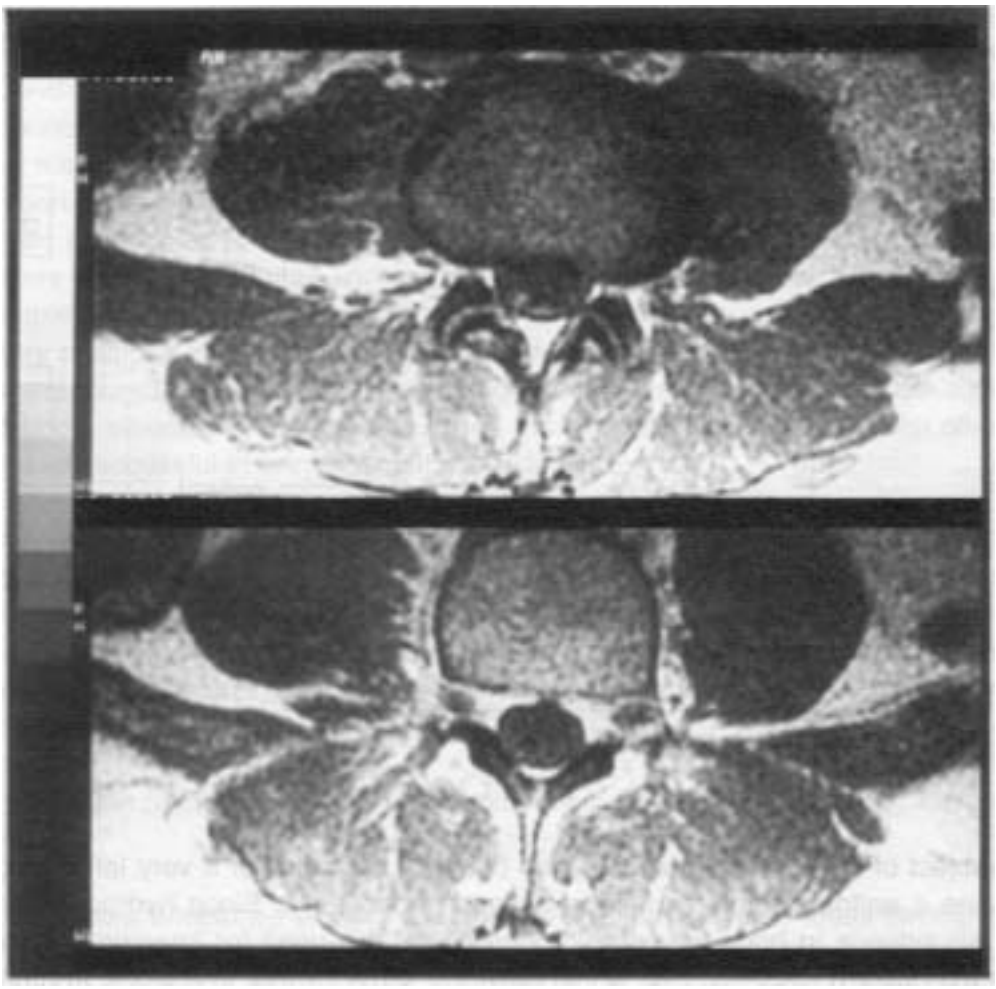


Fig. 2.25

30. Communicating Information

(a) Candidates should be able to understand the term modulation and be able to distinguish between amplitude modulation (AM) and frequency modulation (FM).

All communication systems require a source and a receiver. Three such systems are illustrated in Fig. 3.1.

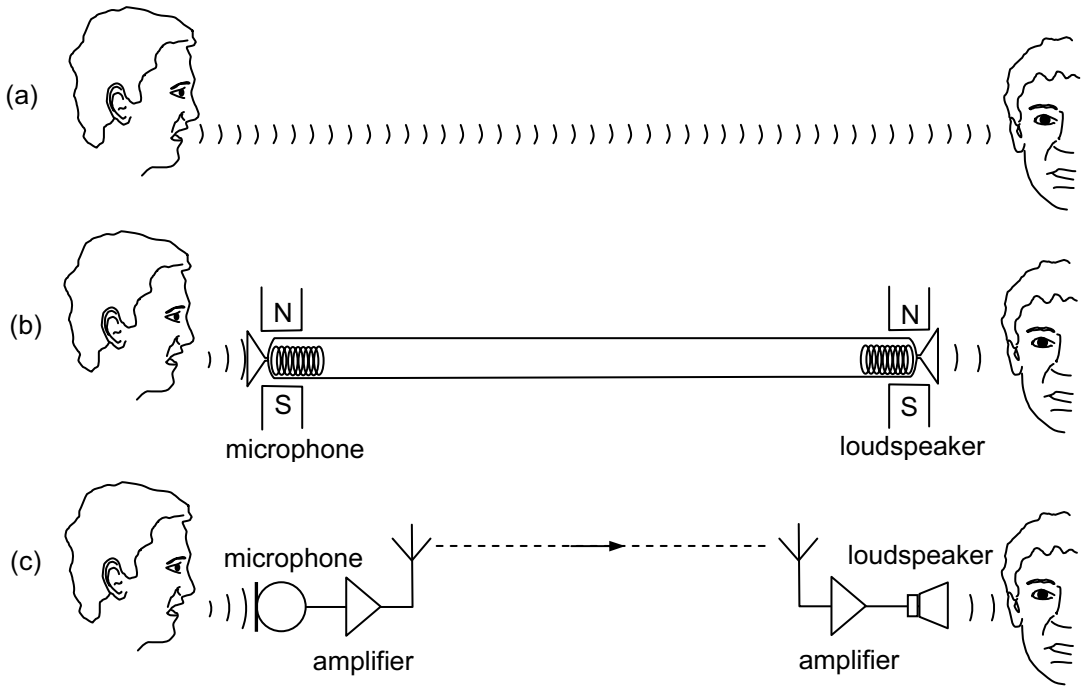


Fig. 3.1

Sound can be transmitted either directly as in (a) or via the alternating currents induced in a moving-coil microphone and received by a moving coil speaker as illustrated in (b).

It is also possible to communicate using radio waves by simply amplifying the audio signal and applying it to a suitable aerial as illustrated in (c). However, there are two fundamental problems with this system.

1. Only one radio station can operate in the region because the wave from a second operating station would interfere with the first.
2. The aerial required to transmit frequencies in the audio range (20 Hz to 20 kHz) would be both very long and inefficient (the radio waves would not travel very far unless huge powers were used).

Both of these problems are solved by the process of *modulation*, the principle of which is illustrated in Fig. 3.2. In modulation, a high frequency wave known as the *carrier wave* has either its amplitude or its frequency altered by the information signal in order to carry the information.

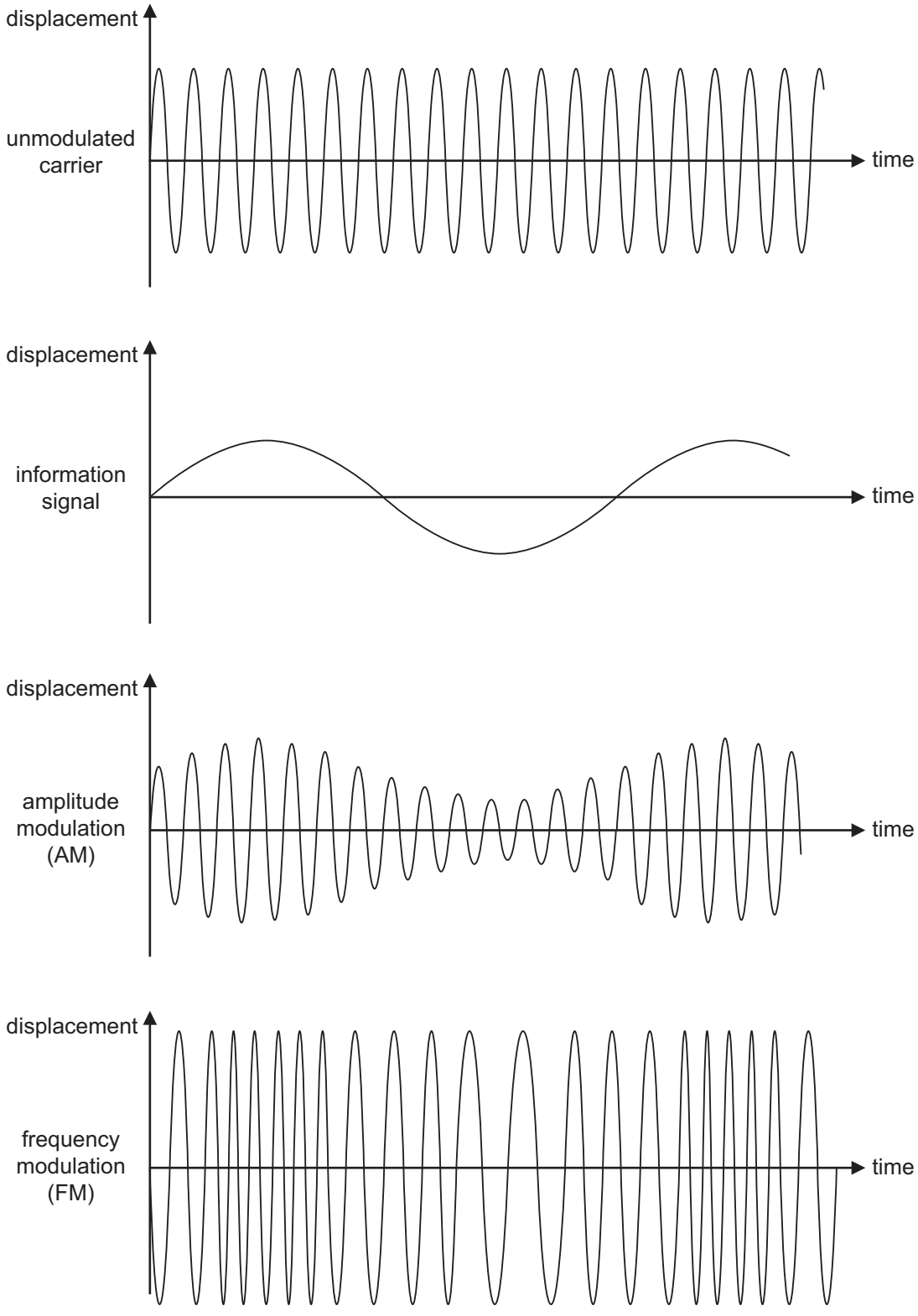


Fig. 3.2

For *amplitude modulation* (AM), the amplitude of the carrier wave is made to vary in synchrony with the displacement of the information signal. The variation in the amplitude of the carrier wave is a measure of the displacement of the information signal and the rate at which the carrier amplitude varies is equal to the frequency of the information signal.

For *frequency modulation* (FM), the frequency of the carrier wave is made to vary in synchrony with the displacement of the information signal. The amplitude of the carrier wave does not vary. The change in frequency of the carrier wave is a measure of the displacement of the information signal. The rate at which the carrier wave frequency is made to vary is equal to the (instantaneous) frequency of the information signal.

Note: The use of a carrier wave allows different radio stations in the same locality to transmit simultaneously. Each station transmits on a different carrier frequency and consequently the carrier waves do not, in effect, interfere with one another. This is because any one receiver is tuned to the frequency of a particular carrier wave. The receiver then responds to, and gives an output based on, the differences in displacement, or frequency, between the actual waveform and the 'underlying' carrier wave. In other words, the receiver recognises the information signal and rejects others.

(b) Candidates should be able to recall that a carrier wave, amplitude modulated by a single audio frequency, is equivalent to the carrier wave frequency together with two sideband frequencies.

(c) Candidates should be able to understand the term bandwidth.

Fig. 3.3 shows the waveform resulting from the amplitude modulation of a high frequency carrier wave by a signal that consists of a single audio frequency.

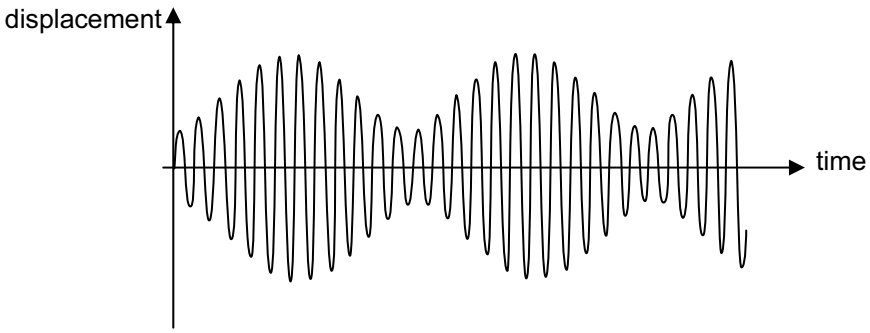


Fig. 3.3

When this waveform is analysed, it is seen to be composed of the sum of three waves of three separate frequencies. These waves are illustrated in the frequency spectrum of Fig. 3.4.

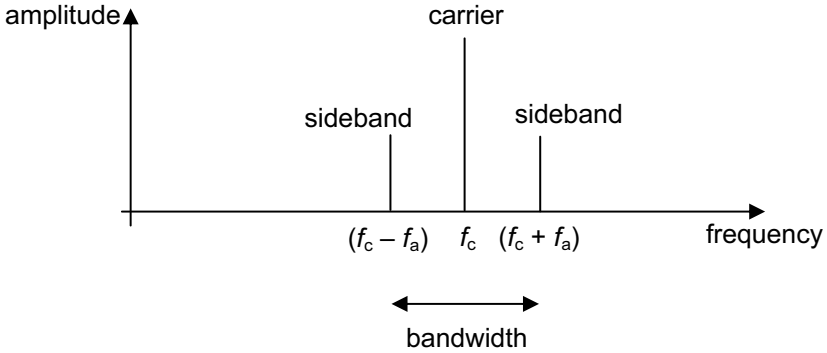


Fig. 3.4

The central frequency f_c is that of the high-frequency carrier wave. The other two are known as *sidebands* and for the AM waveform, they occur at frequencies given by $f_c \pm f_a$, where f_a is the frequency of the audio signal.

The relative amplitude of the sidebands and the carrier depends on the relative amplitudes of the audio and the carrier waveforms. If there is no audio frequency signal, there are no sidebands!

Bandwidth is the frequency range occupied by the AM waveform. This is equal to $2f_a$.

Fig. 3.5 illustrates the AM waveform and the corresponding frequency spectrum for a voice signal.

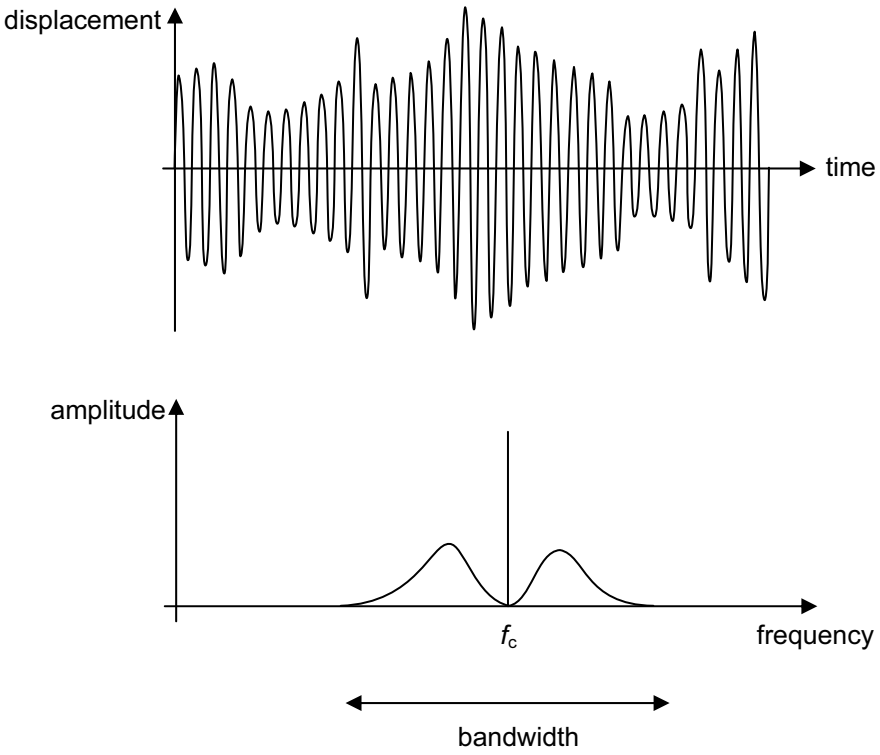


Fig. 3.5

Many audio frequencies are involved. It can be seen that the bandwidth for an AM waveform is the range of frequencies from the lowest to the highest component in the sidebands.

Note that the frequency spectrum of an FM waveform is not the same as that for an AM waveform because further side frequencies that are multiples of the audio frequencies are produced.

(d) Candidates should be able to demonstrate an awareness of the relative advantages of AM and FM transmissions.

An aerial receiving electromagnetic waves cannot distinguish between a genuine radio signal and, say, the interfering radiation from the ignition system of a passing motorbike. If the radio signal is AM, the interference would be considered to be part of the modulation and so it becomes audible in the output produced by the receiver. If, however, the radio signal is FM, the interference will not be picked up by the receiver because it is only variations in frequency that are important, not variations in amplitude. Thus, the quality, in terms of interference, of AM reception is generally poorer than that of FM.

On the long wave (LW) and medium wave (MW) wavebands, the bandwidth on an AM radio station is 9 kHz. This means that the maximum audio frequency that can be broadcast is 4.5 kHz. This frequency is well below the highest frequency audible to the human ear (about 15 kHz) and therefore such broadcasts lack higher frequencies and thus quality.

On the very-high frequency (VHF) waveband, the bandwidth of an FM radio station is about 200 kHz and the maximum audio frequency broadcast is 15 kHz. Thus, the quality of music received on AM is poorer than that of FM but in this case, on the basis of bandwidth.

The LW waveband occupies a region of the electromagnetic spectrum from 30 kHz to 300 kHz. The number of separate AM radio stations that could share this waveband is, theoretically, 270 / 9 = 30. However, the number of separate FM stations would be only 270 / 200 = 1. So, more AM radio stations than FM radio stations can share any waveband. For this reason, FM is used only at frequencies in excess of 1 MHz.

The AM transmissions on the LW, MW and SW (short-wave) wavebands are propagated very large distances so that broadcasts can be made to a very large area from only one transmitter. FM transmissions have a range of only about 30 km by line-of-sight. To broadcast to a large area, many FM transmitters are required. It is, therefore, much cheaper and simpler to broadcast by AM than by FM.

AM transmitters and receivers are electronically simpler and cheaper and they also occupy a much smaller bandwidth than those of FM.

(e) Candidates should be able to recall the advantages of the transmission of data in digital form, compared to the transmission of data in analogue form.

Much of the information that is to be communicated in the real world is analogue information (e.g. the voltage output of a microphone that varies with time in a similar manner to the sound waveform that caused it). If this analogue signal is to be transmitted over a large distance (either by radio or by cable) it will be attenuated and it will pick up noise.

Attenuation is a gradual reduction in signal power. This could be, for example, ohmic losses in a metal cable. In any electrical system there is always unwanted power present that adds itself in a random manner to the signal. This unwanted random power is called *noise* and it causes distortion of the signal. There are several sources of noise. One arises from the thermal vibrations of the atoms of the material through which the signal is passing. As a result, noise power cannot be totally eliminated.

Attenuation will mean that, eventually, the signal will have to be amplified so that it can be distinguished from the background noise. This is achieved using a repeater amplifier that amplifies the signal before passing it further on. The amplifier will, however, amplify the noise as well as the original signal. After several of these repeater amplifications (required for transmission over long distances), the signal will become very 'noisy'. This effect is illustrated in Fig. 3.6.

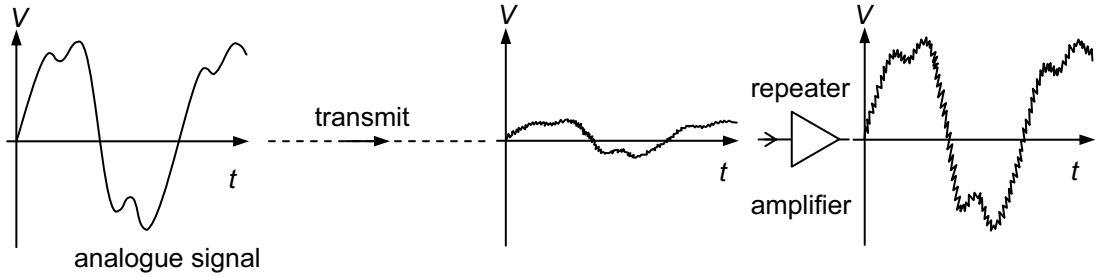


Fig. 3.6

If the signal is transmitted in digital form, then it also suffers from attenuation and the addition of noise. However, the amplifiers that are used for amplifying digital signals are required only to produce a 'high' voltage or a 'low' voltage. They are not required to amplify small fluctuations in amplitude, as is the situation for amplification of an analogue signal. Since noise consists, typically, of small fluctuations, the amplification of a digital signal does not also amplify the noise. Such amplifiers are called *regenerator amplifiers* and are able to reproduce the original digital signal and, at the same time, 'filter out' the noise. This is illustrated in Fig. 3.7.

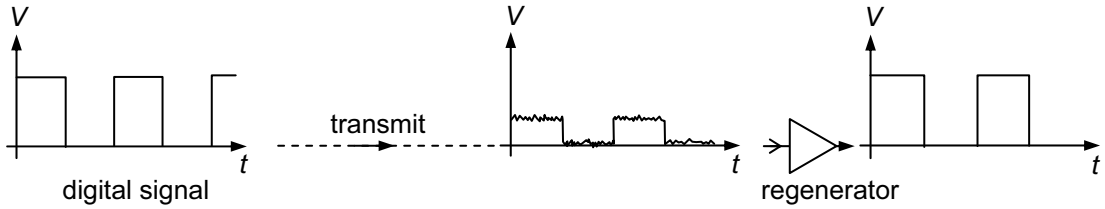


Fig. 3.7

As a result, a digital signal can be transmitted over very long distances with regular regenerations without becoming increasingly noisy, as would happen with an analogue signal.

A further advantage of digital transmissions is that they can have extra information – extra bits of data – added by the transmitting system. These extra data are a code to be used by the receiving system to check for errors and to correct them before passing the information on to the receiver.

Nowadays, digital circuits are generally more reliable and cheaper to produce than analogue circuits. This is, perhaps, the main reason why, in the near future, almost all communication systems will be digitally based.

(f) Candidates should be able to understand that the digital transmission of speech or music involves analogue-to-digital conversion (ADC) on transmission and digital-to-analogue conversion (DAC) on reception.

(g) Candidates should be able to show an understanding of the effect of the sampling rate and the number of bits in each sample on the reproduction of an input signal.

The electrical signals derived from speech or music are analogue audio-frequency signals. The voltage generated varies continuously. To convert an analogue signal into a digital signal involves taking samples of the analogue waveform (i.e. measuring its instantaneous voltage) at regular intervals of time. The instantaneous or sample voltage is converted into a binary number that represents its value.

For example, if the instantaneous value of the analogue signal is 6 V, the binary number could be 0110. For an instantaneous value of 13 V, the binary number could be 1101.

Note that a binary digit is referred to as a *bit*. The most significant bit (MSB) – the bit representing the largest decimal number is written first. The bit representing the lowest decimal number (1) is known as the least significant bit (LSB) and is written last.

A digital signal consists of a series of 'high' and 'low' voltages. A 1 represents a 'high' voltage and a 0 represents a 'low' voltage. A 4-bit system is used in the examples in this booklet. In reality, 8 or more bits would be used for any sampling.

Fig. 3.8 (a) shows an analogue signal of frequency 1 kHz. This signal is sampled every 125 μs (a sampling frequency of 8 kHz). The sample voltages are shown in Fig. 3.8 (b). It should be noted that the value given to the sampled voltage is always the value of the nearest increment *below* the actual sample voltage. In this particular example, an analogue signal of 14.3 V would be sampled as 14 V and one of 3.8 V would be sampled as 3 V. The resulting digital signal is shown in Fig. 3.8 (c). Each number is a group of 4 bits and these groups are separated in time by 125 μs.

The choice of sampling frequency is important. A lower sampling frequency means that less information can be gathered from the analogue signal. More than eighty years ago, it was shown by Nyquist that, in order to be able to recover the analogue signal from its digital conversion, the sampling has to occur at a frequency greater than twice the highest frequency component in the original signal. As a result, in the telephone system, the highest frequency is restricted to 3.4 kHz because the sampling frequency is 8 kHz. In the manufacture of compact discs, the highest frequency is 20 kHz and the sampling frequency is 44.1 kHz.

After the analogue signal has been converted to a 4-bit digital signal by the analogue-to-digital converter (ADC), the digital signal is transmitted. The original signal can be recreated by passing the 4-bit numbers into a digital-to-analogue converter (DAC). This is illustrated in Fig. 3.8 (d) where the original analogue signal of Fig. 3.8(a) has been recreated.

The output of the DAC is 'grainy' and is not smooth because the number of bits limits the number of possible voltage levels (with 4 bits there are 2^4 = 16 levels; with 8 bits, there are 2^8 = 256 levels). As described above, a higher sampling frequency also enables more detail of the analogue signal to be recovered.

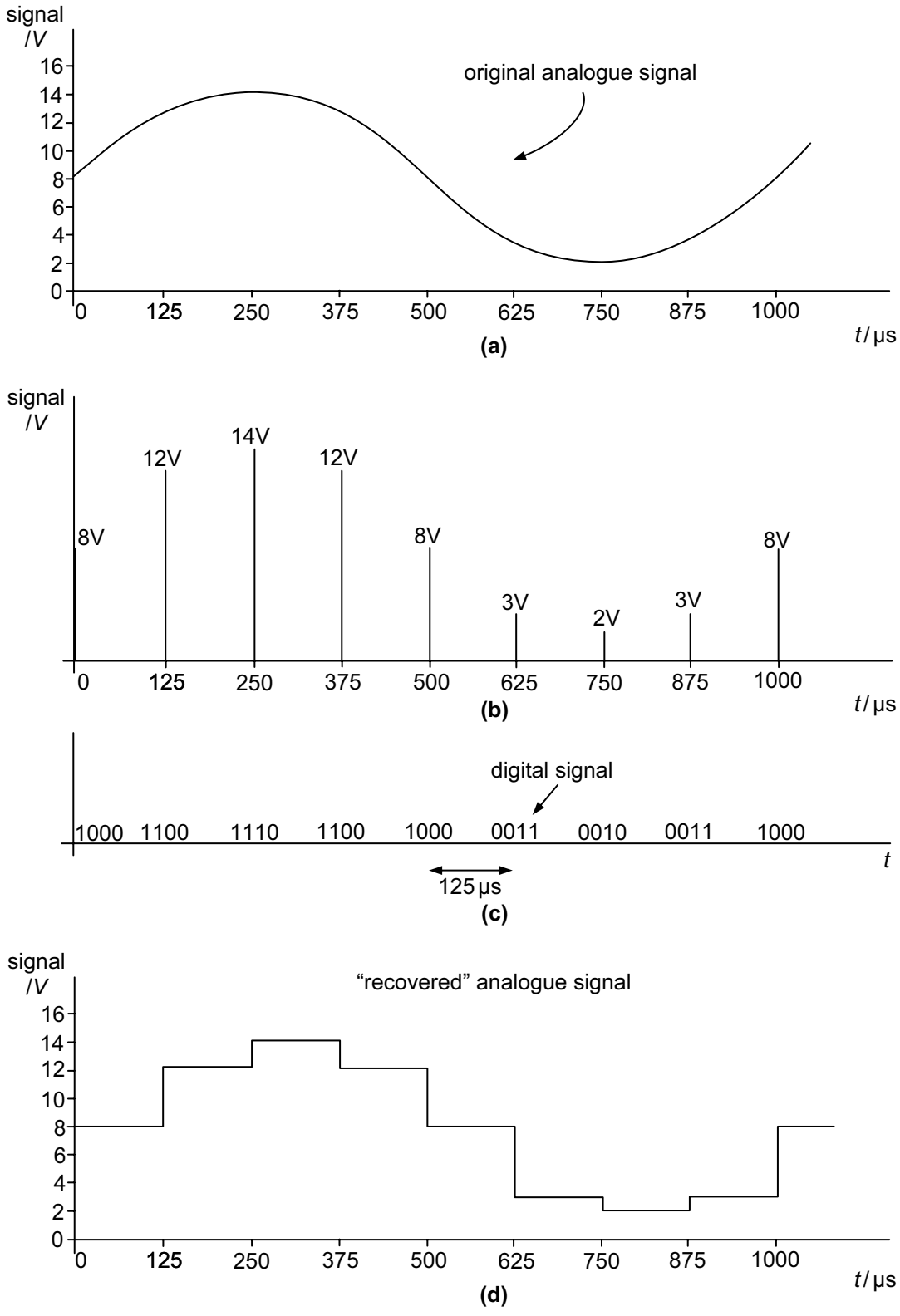


Fig. 3.8

(h) Candidates should be able to appreciate that information may be carried by a number of different channels, including wire-pairs, coaxial cables, radio and microwave links, and optic fibres.

(i) Candidates should be able to discuss the relative advantages and disadvantages of channels of communication in terms of available bandwidth, noise, cross-linking, security, signal attenuation, repeaters and regeneration, cost and convenience.

(l) Candidates should be able to recall the frequencies and wavelengths used in different channels of communication.

Wire-pairs

In the early days of electrical communication, a transmitter was connected to a receiver by a pair of insulated copper wires. Fig. 3.9 illustrates an arrangement for transmitting information in digital code (Morse code).

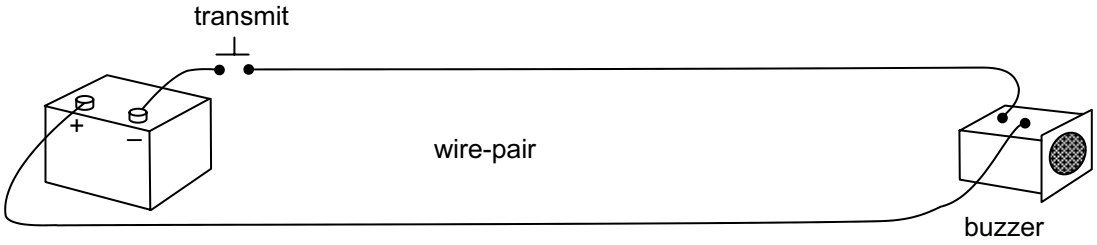


Fig. 3.9

Wire-pairs provide a very simple link. In modern communications, wire-pairs are used mainly for very short distances with low frequencies.

If high frequency signals are transmitted along a pair of wires over an appreciable distance, repeated amplification must be provided at regular intervals. This is due to the very high attenuation of the signal. Energy is lost as heat in the resistance of the wires and also as radiation since the wires act as aerials. A further problem is that the wires easily pick up external interference that degrades the original signal. If several wire-pairs are arranged next to one another, they will pick up each other's signals. This effect is known as *cross-talk* or *cross-linking* and gives very poor security as it is easy to 'tap' a telephone conversation.

The bandwidth of a pair of wires is only about 500 kHz. Consequently, as a means of carrying a large amount of information, it is extremely limited.

Coaxial cable

Coaxial cable is, essentially, a pair of wires arranged so that one wire is shrouded by the other, as illustrated in Fig. 3.10.

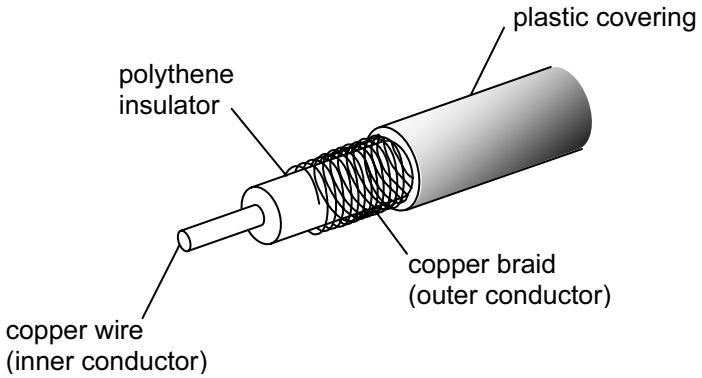


Fig. 3.10

The signal is transmitted down the inner conductor and the outer conductor acts as the return wire and also shields the inner one from external interference. The outer conductor is usually connected to earth.

Coaxial cable is more expensive than wire-pairs but causes less attenuation of the signal. This means that, for long distance communication, repeater amplifiers can be arranged further apart. Coaxial cables are less prone to external interference, though not immune to it, so they do offer slightly greater security.

The bandwidth of coaxial cable is about 50 MHz. It is capable of carrying much more information than a wire-pair.

Radio link

When radio was first developed, an electrical oscillation of a few kilohertz (the carrier wave) was linked to a long wire – the aerial. The oscillations were switched on and off. In this way, information was transmitted from the aerial in digital form (Morse code). It soon became possible to modulate the carrier wave (by AM or by FM) so that information could be sent at a much faster rate. Different carrier frequencies allowed different radio stations to share the same air space (frequency multiplexing).

Energy that is radiated from an aerial is in the form of electromagnetic waves and is propagated at the speed of light. If the frequency of the transmitted waves are somewhere in the range from 30 kHz to 3 GHz, then the waves are known as radio waves.

The electromagnetic radiation that is emitted from a transmitting aerial can be arranged (by suitable choice of the aerial) to radiate in all directions (e.g. for national broadcasting). For point-to-point communications, the aerial can be arranged to radiate mostly in one direction. No matter what aerial is used, there is always energy loss and the power of the signal picked up by a receiving aerial is reduced as the distance between the transmitter and the receiving aerial is increased. The actual distance any particular waves propagate is dependent on frequency, as illustrated in Fig. 3.11.

type of wave	frequency	range
surface wave	below 3 MHz	up to 1000 km
sky wave	3 MHz → 30 MHz	worldwide by means of reflection from ionosphere and ground
space wave	greater than 30 MHz	line of sight – including satellite communication

Fig. 3.11

As a means of communicating from a single transmitter over a large area, the AM broadcasts on the LW and MW are relatively cheap and technically simple, as explained in 30*(d)*.

In modern communication, considerable use is made of the VHF and UHF wavebands for mobile phones, walkie-talkie radio etc. This is due to the fact that, at these frequencies, the wavelength is relatively small and hence the aerial can be made conveniently short.

The part of the electromagnetic spectrum used for radio communication is shown in Fig. 3.12.

	frequency band	frequencies	wavelengths (in a vacuum)
LW radio	low frequencies LF	30 kHz → 300 kHz	10 km → 1 km
MW radio	medium frequencies MW	300 kHz → 3 MHz	1 km → 100 m
SW radio	high frequencies HF	3 MHz → 30 MHz	100 m → 10 m
FM radio	very high frequencies VHF	30 MHz → 300 MHz	10 m → 1 m
TV broadcast	ultra-high frequencies UHF	300 MHz → 3 GHz	1 m → 10 cm
microwave/satellite	super-high frequencies SHF extra-high frequencies EHF	3 GHz → 30 GHz 30 GHz → 300 GHz	10 cm → 1 cm 1 cm → 1 mm

Fig. 3.12

The bandwidth of a radio link increases as the frequency of the carrier wave increases.

Microwave link

Microwaves are radio waves in the SHF waveband from 3 GHz to 30 GHz with wavelengths of only a few centimetres. They are generally used for point-to-point communication, as illustrated in Fig. 3.13.

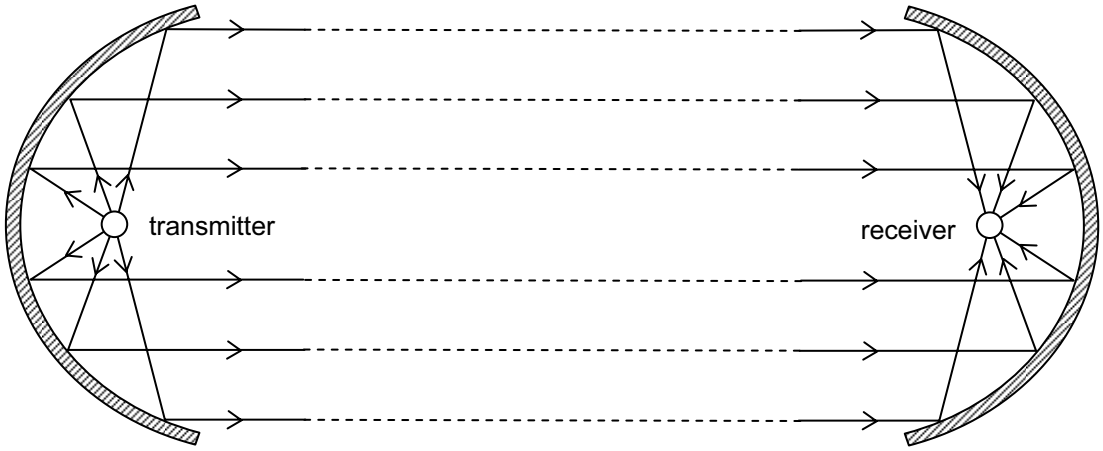


Fig. 3.13

The transmitting element is placed at the focus of a parabolic mirror. This causes the wave power to be radiated in a parallel beam. A parabolic reflector, placed in the path of this beam, reflects and focuses the wave power on to a receiving element. The reflecting parabolic dishes are the not aerials themselves. They are a means of directing as much power as possible into a parallel beam and then collecting this power and directing it to the receiving aerial or element. Parabolic dishes are most useful with short wavelengths where the spread of the waves due to diffraction is less pronounced.

The bandwidth of a microwave link is of the order of GHz. Consequently, microwave links have a very large capacity for carrying information. However, for terrestrial use, the range of the transmissions is limited to line-of-sight. For long-distance transmissions, many repeater stations are required.

Optic fibres

Optic fibres carry digital information in the form of pulses of light or infra-red radiation. These pulses are provided by lasers and the light produced has very high frequencies of the order of 10^8 MHz. In theory, a bit or individual light wave could last for only 10^{-14} s. This would allow hundreds of thousands of individual telephone calls to share the same optic fibre. However, present technology does not allow control at such high frequencies. The duration of a bit is governed by how fast the laser providing light to the fibre can be switched on and off. This is, at present, of the order of GHz but is increasing as technology develops.

The advantages of transmission using optic fibres are indicated below.

- Optic fibres have a wide bandwidth. This gives rise to a large transmission capacity.
- Signal power losses in optic fibres are relatively small. This allows for longer uninterrupted distances between regenerator amplifiers and reduces the costs of installation.
- The cost of optic fibre is much less than that of metal wire.
- The diameter and weight of fibre optic cables is much less that that of metal cables. This implies easier handling and storage.
- Optic fibres have very high security since they do not radiate energy and thus there is negligible ‘cross-talk’ between fibres.
- Optic fibres do not pick up electromagnetic interference. This means they can be used in electromagnetically ‘noisy’ environments, for example alongside electric railway lines. In fact, optic fibre cables are installed along the routes of the National Grid.
- Optic fibre is ideal for digital transmissions since the light is obtained from lasers that can be switched on and off very rapidly.

(j) Candidates should be able to describe the use of satellites in communication.

(k) Candidates should be able to recall the relative merits of both geostationary and polar orbiting satellites for communicating information.

The basic principle of satellite communication is illustrated in Fig. 3.14.

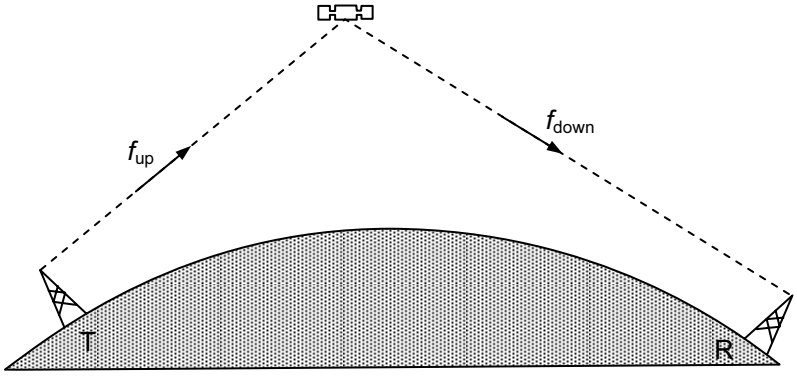


Fig. 3.14

A transmitting station T directs a carrier wave of frequency f_{up} towards the satellite. The satellite receives this signal, amplifies it and changes the carrier frequency to a lower value f_{down} before directing it towards a receiver R back on Earth. Typically the uplink would have a frequency f_{up} of 6 GHz and the downlink would have a frequency f_{down} of 4 GHz (the 6/4 GHz band). Alternatives are the 14/11 GHz band and the 30/20 GHz band. The two carrier frequencies are different to prevent the satellite's high power transmitted signal swamping its reception of the very low power signal that it receives. There is no interference of the actual information being carried by the waves because this is stored as a modulation of the carrier waves.

Although the transmitter in Fig. 3.14 could transmit more or less directly to the receiver without the use of a satellite, it could only do so on the SW or MW wavebands, as described in section 30*(h)*. However, in modern communication systems, this is not done for three reasons.

(i) Long-distance communication on these wavebands is unreliable. Sky waves rely on ionospheric reflection. These layers of ions vary in height and density according to the time of day. In hilly areas, surface waves give rise to regions of poor reception where there are 'shadows'.

(ii) The wavebands are already filled by existing broadcasts.

(iii) The available bandwidths are too narrow to carry the required amount of information.

Satellites may orbit the Earth in polar orbits, as illustrated in Fig. 3.15.

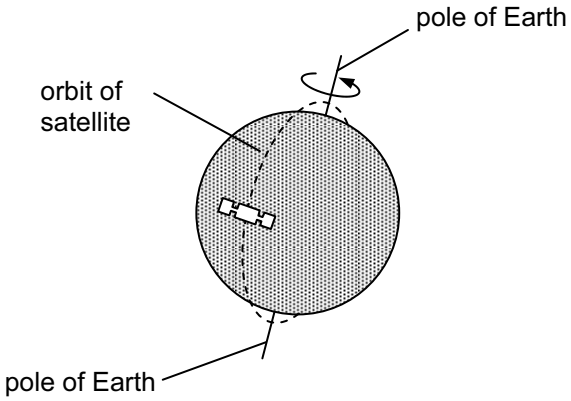


Fig. 3.15

Polar orbits are relatively low with a period of rotation of the order of 90 minutes. Such satellites will, as a result of the rotation of the Earth, at some time each day orbit above every point on the Earth's surface. For a satellite having a period of 90 minutes, each orbit crosses the Equator 23° to the west of the previous orbit.

It is not possible to have continuous communication links with one such satellite because, from Earth, the satellite appears to move rapidly across the sky and, for part of the time, is below the horizon. Polar orbiting satellites are used, as well as for communications, for monitoring the state of the Earth's surface, weather forecasting, spying etc.

Alternatively, satellites may be placed in geostationary orbit, as illustrated in Fig. 3.16.

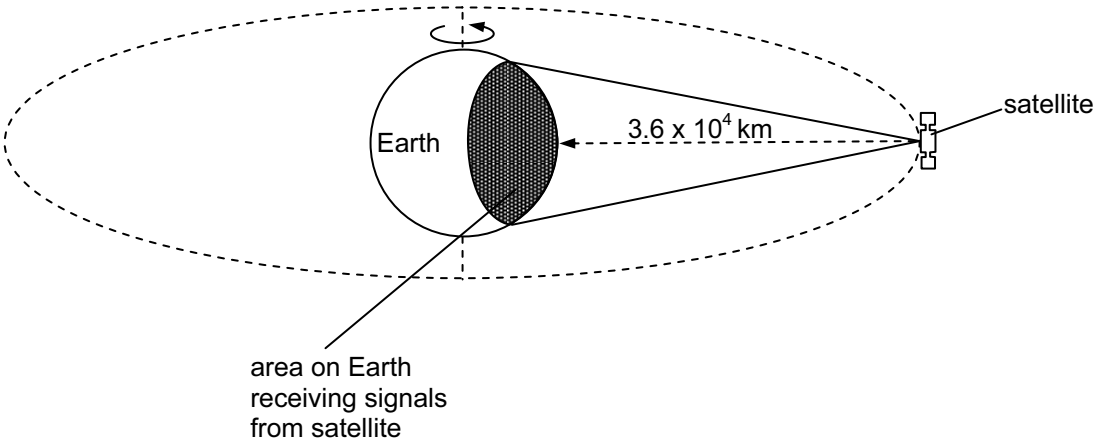


Fig. 3.16

Geostationary satellites orbit the Earth above the Equator with a period of 24 hours at a distance of 3.6×10^4 km above the Earth's surface. If the satellite is orbiting in the same direction as the direction of rotation of the Earth, then, for an observer on the Earth, the satellite will always appear to be above a fixed position on the Equator.

The satellite can allow for continuous communication between a ground station and anywhere on the surface of the Earth that can receive the signal from the satellite. A number of such satellites with over-lapping areas of communication are used for trans-oceanic telephone calls, removing the need for long-distance submarine cables. International television broadcasts are possible, enabling viewers in one country to watch television broadcasts from another.

It should be remembered that geostationary satellites are in Equatorial orbit and thus polar regions may not be in line of sight with a satellite.

Where communication is possible, the height above the Earth's surface of the satellites gives rise to delays in conversation between two people using a satellite link. This delay would be unacceptable where several satellites provide the complete link. For this reason, geostationary satellites may be used in conjunction with optical fibres.

Polar orbiting satellites are used for communication since they are in low orbits, resulting in short time delays between transmission and receipt of a signal. Furthermore, total global coverage is possible. However, a network of such satellites is required in order to maintain continuous links. The satellites must be tracked and the link switched from one satellite to another. Geostationary satellites have the advantage that they do not need to be tracked.

(m) Candidates should be able to understand and use signal attenuation expressed in dB and dB per unit length.

(n) Candidates should be able to recall and use the expression number of dB = $10\lg(P_1/P_2)$ for the ratio of two powers.

When an electrical signal is transmitted along a metal wire, it gradually loses power, mostly as thermal energy in heating the wire. Similarly, a light pulse travelling along an optic fibre loses power, mostly by absorption due to impurities in the glass and by scattering due to imperfections. Electromagnetic waves lose power by absorption and dispersion. A reduction in signal power is referred to as *attenuation*.

In order that a signal may be detected adequately, its power must be a minimum number of times greater than the power associated with noise (see the section on 30*(e)*). Typically, this signal-to-noise ratio could be 100.

Repeater amplifiers may be required to increase the power of a signal that is being passed along a transmission line. The gain of such an amplifier (the ratio of the output power to the input power) could be 100 000. For a radio link between Earth and a geostationary satellite, the power received by the satellite may be 10^{19} times smaller than that transmitted from Earth.

It can be seen from the above examples that the ratio of the two powers may be very large. Consequently, an extremely convenient unit by which power levels, or any other quantities, may be compared is the bel (B). The number of bels is related to the ratio of two powers P_1 and P_2 by the expression

$$\text{number of bels} = \lg(P_1/P_2).$$

In practice, the ratios are usually expressed in decibels (dB), where 10 dB = 1 B. Consequently,

$$\text{number of decibels} = 10\lg(P_1/P_2).$$

Example

The gain of an amplifier is 45 dB. Calculate the output power P_{out} of the amplifier for an input power P_{in} of 2.0 μW.

number of decibels = $10\lg(P_1/P_2)$

$45 = 10\lg(P_{out} / 2.0 \times 10^{-6})$
$4.5 = \lg(P_{out} / 2.0 \times 10^{-6})$
$10^{4.5} = (P_{out} / 2.0 \times 10^{-6})$
$P_{out} = 10^{4.5} \times 2.0 \times 10^{-6}$
$P_{out} = 6.3 \times 10^{-2}$ W

A transmission line has an input power P_2 and the power at a point distance L along the line is P_1 as illustrated in Fig. 3.17.

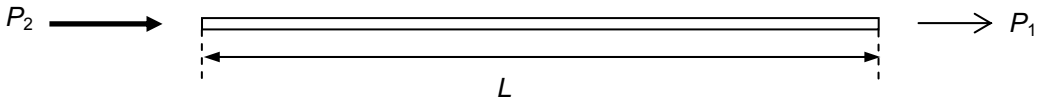


Fig. 3.17

Then, attenuation in the line = $10\lg(P_2 / P_1)$ dB.

Since a transmission line may vary in length, an important feature of a transmission line is its attenuation per unit length.

$$\text{attenuation per unit length} = \frac{1}{L} 10\lg\frac{P_2}{P_1}$$

Example

The input power to a cable of length 25 km is 500 mW. The attenuation per unit length of the cable is 2 dB km^{-1}. Calculate the output power of the signal from the cable.

signal loss in cable = $2 \times 25 = 50$ dB

$50 = 10\lg(500 \times 10^{-3} / P_{out})$, where P_{out} is the output power.

$P_{out} = 500 \times 10^{-3} \times 10^{-5} = 5 \times 10^{-6}$ W.

The signal cannot be allowed to travel indefinitely in the cable because, eventually, it will become so small that it cannot be distinguished from background noise. An important factor is the minimum *signal-to-noise ratio* that effectively provides a value for the lowest signal power allowed in the cable.

In the above example, the background noise is 5×10^{-13} W and the minimum signal-to-noise ratio permissible is 20 dB. Then if P_M is the minimum signal power,

$20 = 10\lg(P_M / 5 \times 10^{-13})$

$P_M = 5 \times 10^{-13} \times 10^2 = 5 \times 10^{-11}$ W.

This enables the maximum uninterrupted length of cable along which the signal can be transmitted to be determined.

Maximum loss in cable = $10\lg(500 \times 10^{-3} / 5 \times 10^{-11}) = 120$ dB

Maximum distance = 120 / 2 = 60 km.

(o) Candidates should be able to understand that, in a mobile-phone system, the public switched telephone network (PSTN) is linked to base stations via a cellular exchange.

In the early days of telephones, each telephone user was connected to all other users by their own cables. This was feasible only where the number of users was small as in, for example, a single building.

As telephones became more popular and widespread, connections between individual users became impractical. Consequently, the telephone exchange was developed. The caller would contact the telephone exchange and, at the exchange, the connection to the other user would be made by a person known as an 'operator'. If the call was not a local call served by that particular exchange, then the local exchange would contact the other user's local exchange via a trunk exchange. Trunk exchanges were connected via trunk lines and hence the expression 'trunk call' for any long-distance call.

In essence, the Public Switched Telephone Network (PSTN) uses the same principles of exchanges but has developed with modern technology and the number of users. For example, switching is no longer achieved using operators. Electromagnetic relays have been replaced by solid-state devices and international 'exchanges', or gateways, have been introduced. The PSTN is illustrated in Fig. 3.18.

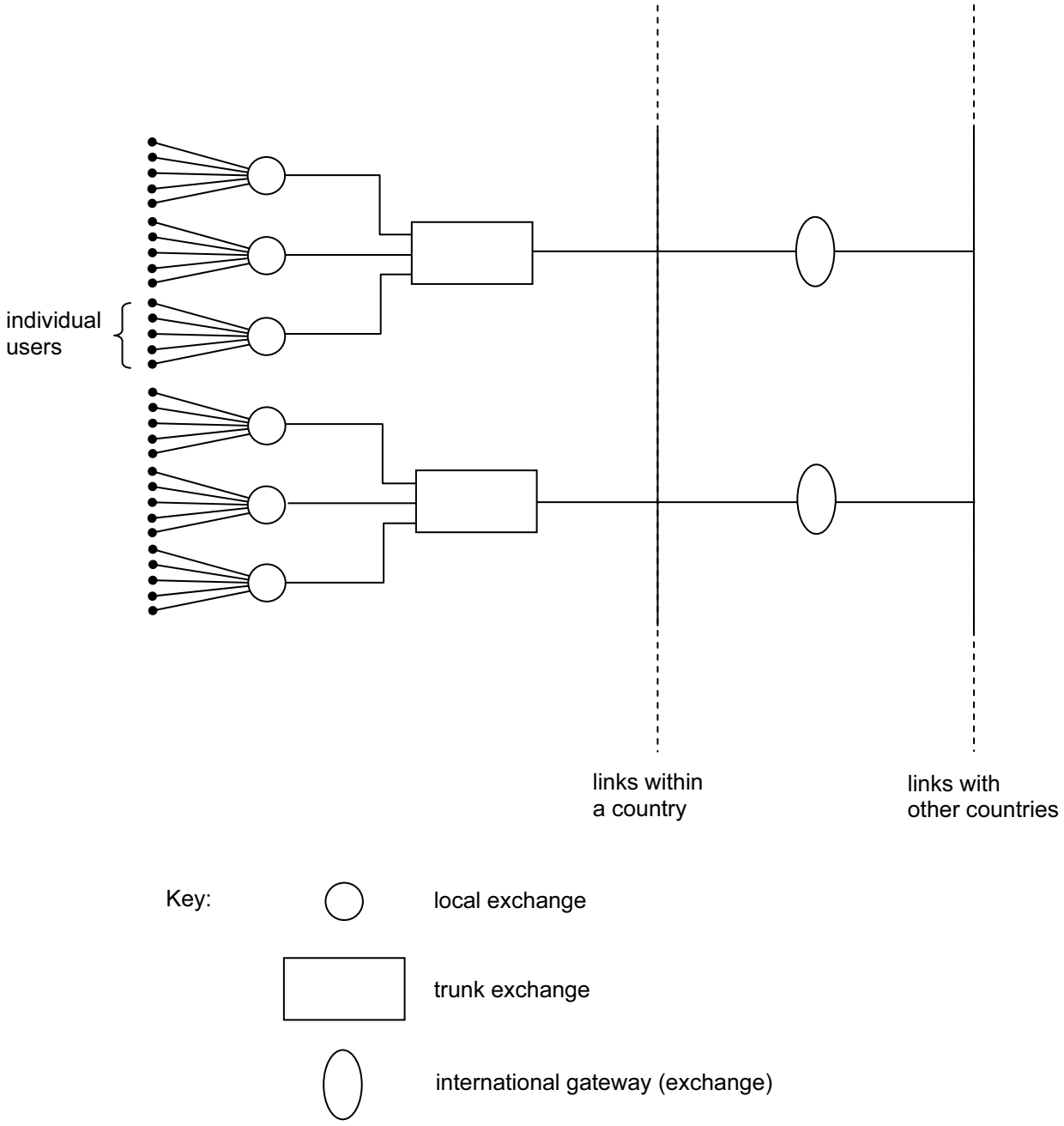


Fig. 3.18

In the system illustrated in Fig. 3.18, the user is connected to the PSTN via the local exchange. Each user has a 'fixed line' to the local exchange, resulting in the user having limited mobility whilst making the call.

During the 1970s and 1980s, mobile phone systems were developed that did not have a permanent link to a local exchange. Basically, a mobile phone is a *handset* that is a radio transmitter and receiver. When a call is to be made, the user makes a radio-wave link with a nearby *base station*. This base station is connected by cable to a *cellular exchange*. The cellular exchange then allows connection to be made to the PSTN. This is illustrated in Fig. 3.19.

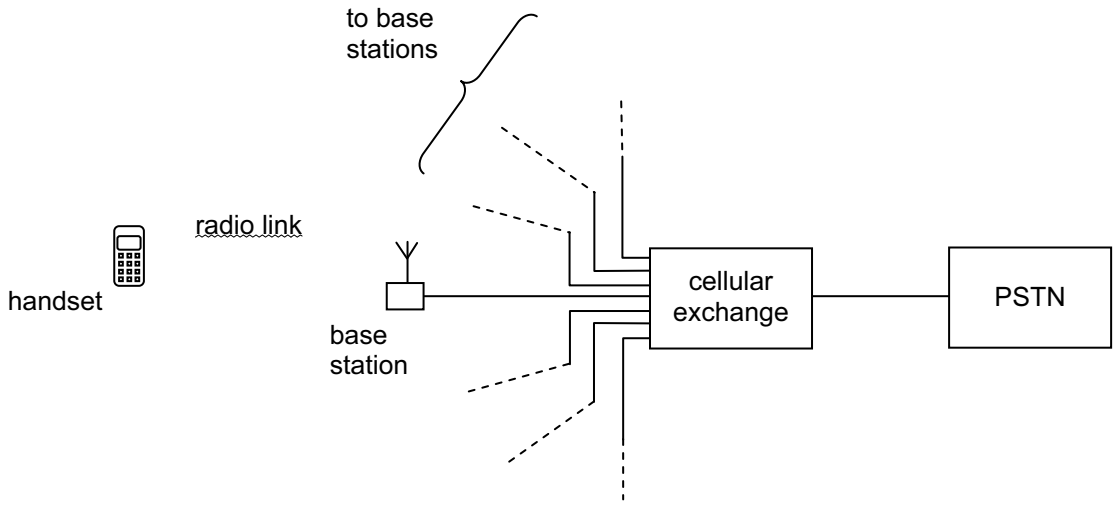


Fig. 3.19

(p) Candidates should be able to understand the need for an area to be divided into a number of cells, each cell served by a base station.

(q) Candidates should be able to understand the role of the base station and the cellular exchange during the making of a call from a mobile phone handset.

The popularity of mobile phones means that large numbers of people use a mobile phone system at the same time. However, the range of carrier-wave frequencies for linking between the mobile phone and the base station is limited. Consequently, each mobile phone cannot have its own carrier frequency. This means that the same carrier frequencies must be used by many mobile phones at the same time. This is achieved using a network of base stations.

The base stations operate on UHF frequencies so that they have a limited range (see the section on 30*(h)*) and are low-power transmitters. The UHF frequencies also mean that the aerial in the mobile phone is conveniently short! A country is divided into areas or *cells*, with each cell having its own base station, usually located near the centre of the cell. The aerial at the base station transmits in all directions so as to cover the whole cell, but not to overlap too far into neighbouring cells. In this way, the whole country is 'covered'. Neighbouring cells cannot use the same carrier frequencies, otherwise interference would occur at the boundaries between cells. A possible arrangement of cells is shown in Fig. 3.20.

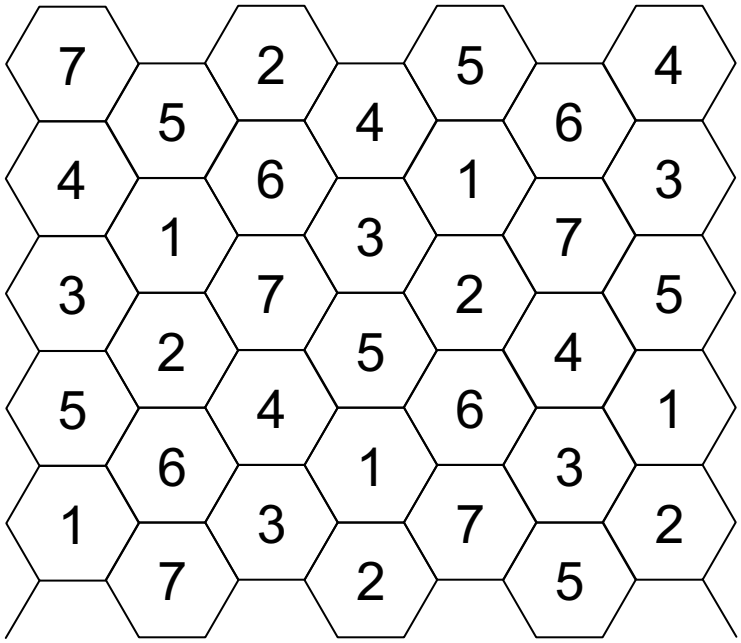


Fig. 3.20

Although each cell is approximately circular (depending on the flatness of the land), the cells are shown as a 'honeycomb' so that the cells fit together. The number in each cell represents a particular range of carrier frequencies that would be allocated to each cell. Neighbouring cells do not have the same range of carrier frequencies.

When a handset is switched on, it transmits a signal to identify itself. This signal is received by a number of base stations, from where it is transferred to the cellular exchange. A computer at the cellular exchange selects the base station with the strongest signal from the handset. The computer also allocates a carrier frequency for communication between the base station and the handset. During communication between the handset and the base station, the computer at the cellular exchange monitors the signal from the handset. If the user of the handset moves from one cell to another, the signal strength changes. The call from the handset is then re-routed through the base station with the greater signal.

(r) Candidates should be able to recall a simplified block diagram of a mobile phone handset and understand the function of each block.

A mobile-phone handset is, in its simplest form, a radio transmitter and receiver. A simplified block diagram of its circuitry is shown in Fig. 3.21.

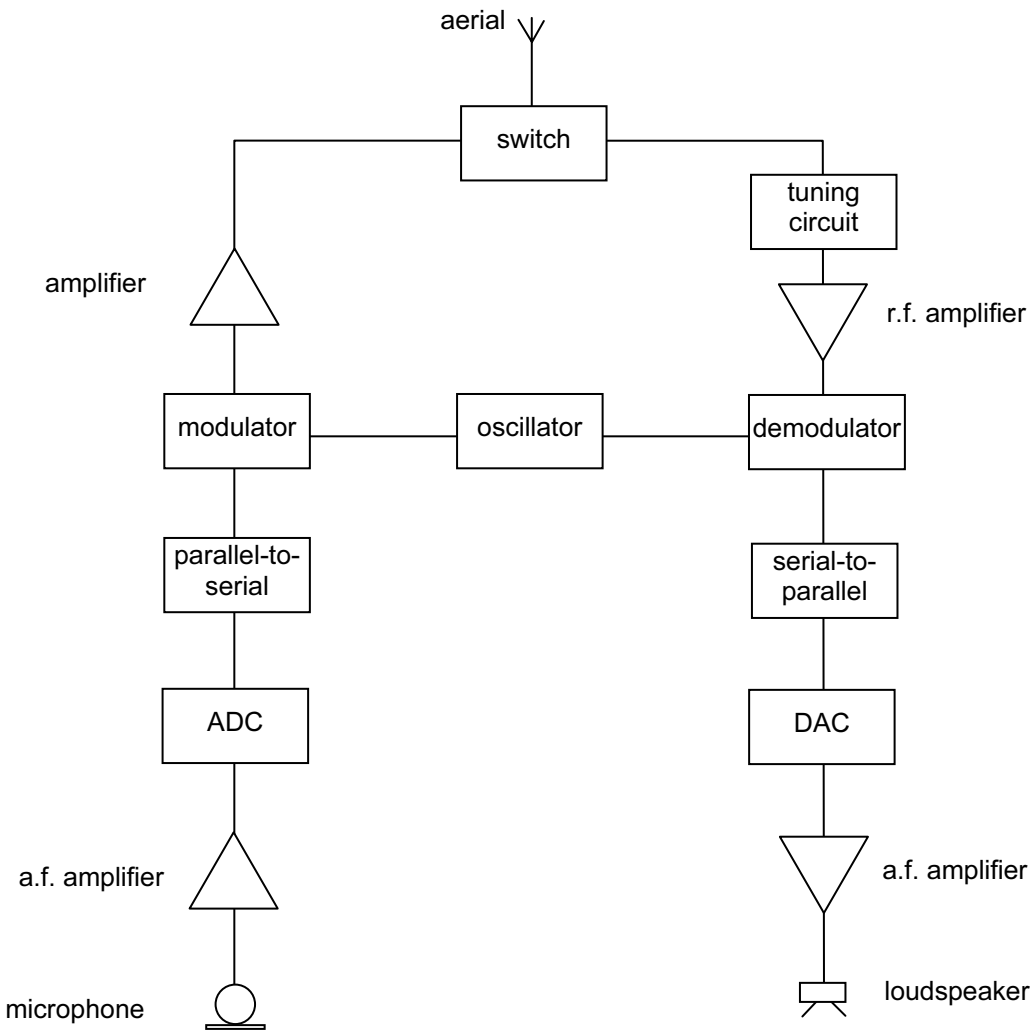


Fig. 3.21

The caller speaks into the microphone. This produces a varying signal voltage that is amplified and converted to a digital signal by means of the ADC. The parallel-to-series converter takes the whole of each digital sample voltage and then emits it as a series of bits. The series of bits is then used to modulate the chosen carrier wave. After further amplification, the modulated carrier wave is switched to the aerial where it is transmitted as a radio wave.

On receipt of a signal at the aerial, the signal is switched to a tuning circuit that selects only the carrier-wave frequency allocated to it by the computer located at the cellular exchange. This selected signal is then amplified and demodulated so that the information signal is separated from the carrier wave. This information signal is in digital form. It is processed in a series-to-parallel converter to produce each sample digital voltage and then in a digital-to-analogue converter (DAC) to provide the analogue signal. After amplification, the analogue signal is passed to a loudspeaker.