1 Principles

Alan Betts, G0HIQ

A good understanding of the basic principles and physics of matter, electronics and radio communication is essential if the self-training implicit in amateur radio is to be realised. These principles are not particularly difficult and a good grasp will allow the reader to understand the following material rather than simply accepting that it is true but not really knowing why. This will, in turn, make more aspects of the hobby both attractive and enjoyable.

STRUCTURE OF MATTER

All matter is made up of atoms and molecules. A molecule is the smallest quantity of a substance that can exist and still display the physical and chemical properties of that substance. There is a very great number of different sorts of molecule. Each molecule is, in turn, made up of a number of atoms. There are about 102 different types of atom which are the basic elements of matter. Two atoms of hydrogen will bond with one atom of oxygen to form a molecule of water for example. The chemical symbol is H_2O. The H stands for hydrogen and the subscript 2 indicates that two atoms are required; the O denotes the oxygen atom.

A more complex substance is H_2SO_4. Two hydrogen atoms, one sulphur atom and four oxygen atoms form a molecule of sulphuric acid, a rather nasty and corrosive substance used in lead-acid batteries.

Atoms are so small that they cannot be seen even under the most powerful optical microscopes. They can, however, be visualised using electronic (not electron) microscopes such as the scanning tunnelling microscope (STM) and the atomic force microscope (AFM). **Fig 1.1** shows an AFM representation of the surface of a near-perfect crystal of graphite with the carbon atoms in a hexagonal lattice.

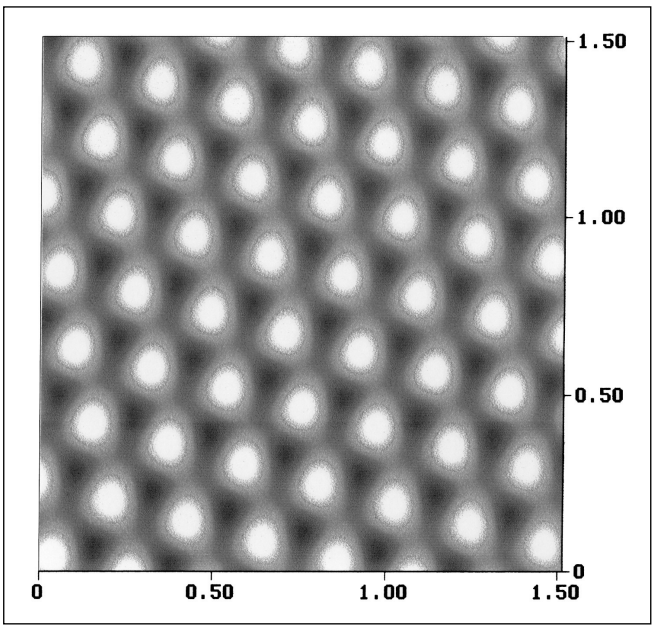


Fig 1.1: Image of the atoms in a piece of high-purity graphite (distances in nanometres). The magnification is approximately 45 million times. Note that no optical microscope can produce more than about 1000 times magnification

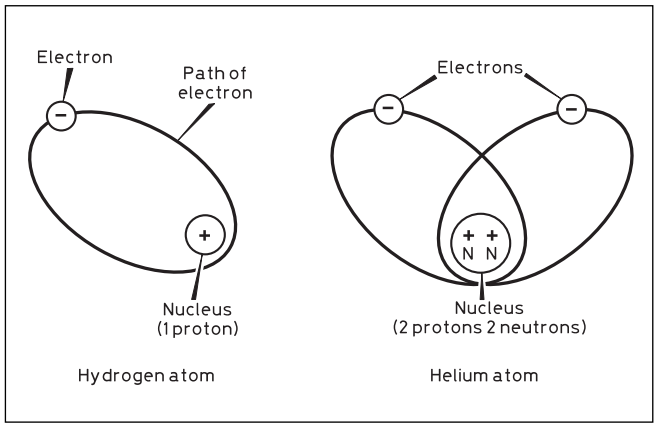


Fig 1.2: Structure of hydrogen and helium atoms

Atoms are themselves made up of yet smaller particles; the electron, the proton and the neutron, long believed to be the smallest things that could exist. Modern atomic physics has shown that this is not so and that not only are there smaller particles but that these particles, energy and waves are, in certain scenarios, indistinguishable from each other. Fortunately we need only concern ourselves with particles down to the electron level but there are effects, such as in the tunnel diode, where the electron seems to 'disappear' and 'reappear' on the far side of a barrier.

The core of an atom comprises one or more protons and may include a number of neutrons. The electrons orbit the core, or 'nucleus' as it is called, rather like the planets orbit the sun. Electrons have an electrical charge that we now know to be a negative charge. Protons have an equal positive charge. The neutrons are not charged.

A hydrogen atom has a single proton and a single orbiting electron. A helium atom has a nucleus of two protons and two neutrons with two orbiting electrons; this is shown in **Fig 1.2**. An atom is electrically neutral; the positive charge on the nucleus is balanced by the negative charge on the electrons. The magnitude of the charge is tiny; it would require 6,250,000,000,000,000,000 ($6{\cdot}25 \times 10^{18}$) electrons to produce a charge of 1 coulomb, that is 1A flowing for 1s.

Conductors and Insulators

The ease with which the electrons in a substance can be detached from their parent atoms varies from substance to substance. In some substances there is a continual movement of electrons in a random manner from one atom to another and the application of an electrical force or potential difference (for example from a battery) to the two ends of a piece of wire made of such a substance will cause a drift of electrons along the wire called an electric current; electrical conduction is then said to take place. It should be noted that if an electron enters the wire from the battery at one end it will be a different electron which immediately leaves the other end of the wire.

To visualise this, consider a long tube such as a scaffold pole filled with snooker balls. As soon as another ball is pushed in one end, one falls out the other but the progress of any particular ball is much slower. The actual progress of an individual electron along a wire, the drift velocity, is such that it could take some minutes to move even a millimetre.

The flow of current is from a point of positive charge to negative. Historically, the decision of what represented 'positive' was arbitrary and it turns out that, by this convention, electrons have a negative charge and the movement of electrons is in the opposite direction to conventional current flow.

Materials that exhibit this property of electrical conduction are called conductors. All metals belong to this class. Materials that do not conduct electricity are called insulators, and **Table 1.1** shows a few examples of commonly used conductors and insulators.

Conductors	Insulators
Silver	Mica
Copper	Quartz
Gold	Glass
Aluminium	Ceramics
Brass	Ebonite
Steel	Plastics
Mercury	Air and other gasses
Carbon	Oil
Solutions of salts or acids in water	Pure water

Table 1.1: Examples of conducting and insulating materials

ELECTRICAL UNITS

Charge (q)

Charge is the quantity of electricity measured in units of coulombs. **Table 1.2** gives the units and their symbol.

One coulomb is the quantity of electricity given by a current of one ampere flowing for one second.

Charge q = current (A) x time (s), normally written as: q=I x t

Current Flow, the Ampere (A)

The ampere, usually called amp, is a fundamental (or base) unit in the SI (System International) system of units. It is actually defined in terms of the magnetic force on two parallel conductors each carrying 1A.

Energy (J)

Energy is the ability to do work and is measured in joules. One joule is the energy required to move a force of one Newton through a distance of one metre. As an example, in lifting a 1kg bag of sugar 1m from the floor to a table, the work done or energy transferred is 9·81 joules.

Power (W)

Power is simply the rate at which work is done or energy is transferred and is measured in watts, W.

$$\text{Power} = \frac{\text{energy transferred}}{\text{time taken}}$$

For example, if the bag of sugar was lifted in two seconds, the power would be 9·81/2 or approximately 5W.

Potential Difference, Voltage (V)

If a source of electrical energy has a Potential Difference of 1 volt, each coulomb of electricity, ie charge, that flows has an energy of 1 joule. If one coulomb of charge flows in, for example, a bulb, and 12 joules of energy are transferred into heat and light, the potential difference across the bulb is 12 volts.

The definition of the volt is the number of joules of energy per coulomb of electricity.

Quantity	Symbol	Unit	Abbreviation
Charge	q	coulomb	C
Conductance	G	Siemen	S
Current	I	Ampere (Amp)	A
Voltage*	E or V	volt	V
Time	t	second	s or sec
Resistance	R	ohm	Ω
Capacitance	C	farad	F
Inductance	L	henry	H
Mutual inductance	M	henry	H
Power	P	watt	W
Frequency	f	hertz	Hz
Wavelength	λ	metre	m

** 'Voltage' includes 'electromotive force' and 'potential difference'.*

Since the above units are sometimes much too large (eg the farad) and sometimes too small, a series of multiples and sub-multiples are used:

Unit	Symbol	Multiple
Microamp	μA	1 millionth (10^{-6}) amp
Milliamp	mA	1 thousandth (10^{-3}) amp
Microvolt	μV	10^{-6}V
Millivolt	mV	10^{-3}V
Kilovolt	kV	10^{3}V
Picofarad	pF	10^{-12}F
Nanofarad	nF	10^{-9}F
Microfarad	μF	10^{-6}F
Femtosecond	fs	10^{-15}s
Picosecond	ps	10^{-12}s
Nanosecond	ns	10^{-9}s
Microsecond	μs	10^{-6}s
Millisecond	ms	10^{-3}s
Microwatt	μW	10^{-6}W
Milliwatt	mW	10^{-3}W
Kilowatt	kW	10^{3}W
Gigahertz	GHz	10^{9}Hz
Megahertz	MHz	10^{6}Hz
Kilohertz	kHz	10^{3}Hz
Centimetre	cm	10^{-2}m
Kilometre	km	10^{3}m

Note: The sub-multiples abbreviate to lower case letters. All multiples or sub-multiples are in factors of a thousand except for the centimetre.

Table 1.2: Units and symbols

$$1\,\text{volt} = \frac{1\,\text{joule}}{1\,\text{coulomb}}$$

Historically, voltage was viewed as a force but that is not strictly true although the term electromotive force (emf) is still in use. The better term is Electrical Potential.

Resistance

Resistance restricts the flow of charge, the current. In forcing electrons through a conductor, some energy is lost as heat. A longer, thinner conductor will have a greater loss, that is a higher resistance.

Different materials have differing resistivities, that is, a wire of the same dimensions will have different resistances depending on the material. The conductors in the list in Table 1.1 are in conductivity (inverse of resistivity) order.

Materials such a Nichrome, Manganin and Eureka are alloys with a deliberately high resistivity and are used in power resistors and wire-wound variable resistors. Tungsten has a relatively high

resistance but its key property is a high melting point and relative strength when white hot. It is used to make the filament of incandescent light bulbs.

Specific resistance: This is simply the resistance of a standard size piece of the subject material, normally a 1 metre cube and is quoted in units of ohm metre, Ωm and has the symbol ρ, the Greek letter rho. Its purpose is to compare the resistivity of different materials.

The unit of resistance is the ohm, symbol Ω, the Greek upper case letter omega. It is defined as the ratio of the applied EMF and the resulting current.

$$\text{Resistance R ohms} = \frac{\text{applied EMF}}{\text{current flowing}}$$

Ohm's Law

Ohm's Law is simply a restatement of the definition of resistance. In words it reads:

In a circuit at constant temperature the current flowing is directly proportional to the applied voltage and inversely proportional to the resistance. The reference to temperature is important. In a practical test, the energy will be converted to heat and most materials change their resistance as the temperature changes.

In algebraic form, where V is the applied voltage or potential difference, I is the current flowing and R is the resistance:

$$V = I \times R \qquad I = \frac{V}{R} \quad \text{and} \quad R = \frac{V}{I}$$

Example.

Consider the circuit shown in **Fig 1.3** which consists of a 4V battery and a resistance R of 8Ω. What is the magnitude of the current in the circuit?

Here V = 4V and R = 8Ω. Let I be the current flowing in amperes. Then from Ohm's Law:

$$I = \frac{V}{R} = \frac{4}{8} = \frac{1}{2} = 0 \cdot 5\text{A}$$

It should be noted that in all calculations based on Ohm's Law care must be taken to ensure that V, I and R are in consistent units, ie in amperes, volts and ohms respectively, if errors in the result are to be avoided. In reality , typical currents may be in mA or μA and resistances in kΩ or MΩ.

Conductance

Conductance is simply the inverse or reciprocal of resistance. Many years ago it was measured in a unit called the mho; today the correct unit is the Siemen. A resistance of 10 ohms is the same as a conductance of 0·1 Siemen.

EMF, PD and Source Resistance

Sources of electrical energy, such as batteries, hold a limited amount of energy and there is also a limit to the rate at which this energy can be drawn, that is the power is limited. A battery or power supply can be considered as a perfect source (which can supply any desired current) together with a series resistance. The series resistance will limit the total current and will also result in a drop in the potential difference or voltage at the terminals. This is shown in **Fig 1.4**.

The source V has a voltage equal to the open circuit terminal voltage and is called the electromotive force (EMF) of the device. On load, that is when a current is being drawn, the potential difference at the terminals will drop according to the current drawn. The drop may be calculated as the voltage across the internal resistor 'r' using the formula:

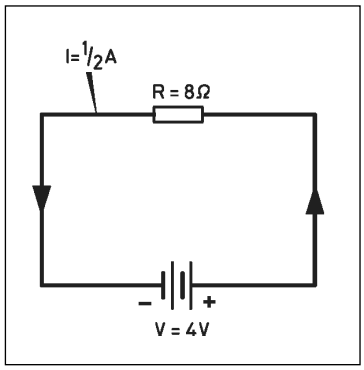


Fig 1.3: Application of Ohm's Law

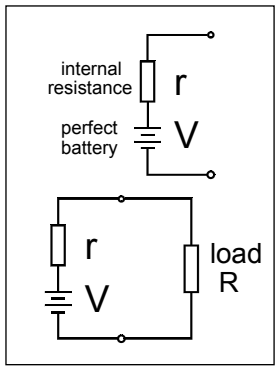


Fig 1.4: A real battery

Voltage drop = current drawn (I) x rΩ and the terminal voltage will be

$$V_{Terminal} = V_{Supply} - I \times r$$

Maximum Power Transfer

It is interesting to consider what is the maximum power that can be drawn from a particular source. As the load resistance decreases, the current drawn increases but the potential difference across the load decreases. Since the power is the product of the load current and the terminal voltage, there will be a maximum point and attempts to draw more power will be thwarted by the drop in terminal voltage.

Fig 1.5 shows the power in the load as the load resistance is varied from 0 to 10Ω, connected to a source of EMF, 10V and internal resistance 2Ω.

Maximum power transfer occurs when the load resistance is the same as the source resistance. For DC and power circuits this is never done since the efficiency drops to 50% but in RF and low level signal handling, maximum signal power transfer may be a key requirement.

Sources of Electricity

When two dissimilar metals are immersed in certain chemical solutions, or electrolytes, an electromotive force (EMF or voltage) is created by chemical action within the cell so that if these pieces of metal are joined externally, there will be a continuous flow of electric current. This device is called a simple cell and such a cell, comprising copper and zinc rods immersed in diluted sulphuric acid, is shown in **Fig 1.6(a)**. The flow of current is from the copper to the zinc plate in the external circuit; ie the copper forms the positive (+) terminal of the cell and the zinc forms the negative (-) terminal.

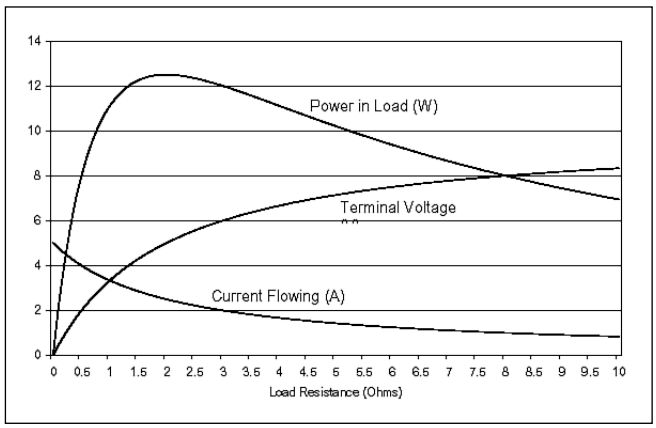


Fig 1.5: Power dissipated in the load

In a simple cell of this type hydrogen forms on the copper electrode, and this gas film has the effect of increasing the internal resistance of the cell and also setting up within it a counter or polarising EMF which rapidly reduces the effective EMF of the cell as a whole.

This polarisation effect is overcome in practical cells by the introduction of chemical agents (depolarisers) surrounding the anode for the purpose of removing the hydrogen by oxidation as soon as it is formed.

Primary Cells

Practical cells in which electricity is produced in this way by direct chemical action are called primary cells; a common example is the Leclanché cell, the construction of which is shown diagrammatically in **Fig 1.6(b)**. The zinc case is the negative electrode and a carbon rod is the positive electrode. The black paste surrounding the carbon rod may contain powdered carbon, manganese dioxide, zinc chloride, ammonium chloride and water, the manganese dioxide acting as depolariser by combining with hydrogen formed at the anode to produce manganese oxide and water.

Alkaline cells are now more common and are interchangeable with zinc-carbon since both produce about 1.5V PD. They should not be mixed due to differing capacities. The steel can, the cathode (positive), is filled with a paste of manganese dioxide and carbon powder to improve conductivity. The centre has a porous separator filled with zinc powder in a potassium hydroxide paste electrolyte. This can leak a caustic gel at the end of its life forming a crystalline powdery coating which corrodes metal tracks and is a skin and respiratory irritant. Most cells are now leak proof, but attempts at recharging may accelerate any tendency to leak.

Silver oxide cells, typically button cells produce about 1.8V and have a silver oxide/ zinc chemistry, again with an alkaline electrolyte, typically sodium or potassium hydroxide.

Lithium batteries are now more common. These were relatively expensive and could suffer from catastrophic failure modes including fire, but technology improvements have largely overcome those difficulties. Their main advantage is much higher energy density, by volume or by weight. As with all batteries the manufacturer's instructions do need to be followed, particularly regarding removal when exhausted, safe disposal and the caution against attempted recharging.

Cells may be connected in series to form a battery with a higher voltage. The cells should be of the same type, capacity and age. With batteries of several cells it is quite possible for fresh cells to continue to cause a current which will reverse charge an old or dud cell. That may cause a chemical leak, swelling and physical damage and, possibly, overheating.

The symbol used to denote a cell in a circuit diagram is shown in **Fig 1.6(c)**. The long thin stroke represents the positive terminal and the short thick stroke the negative terminal. Several cells joined in series to form a battery are shown; for higher voltages it becomes impracticable to draw all the individual cells involved and it is sufficient to indicate merely the first and last cells with a dotted line between them with perhaps a note added to state the actual voltage. The amount of current which can be derived from a dry cell depends on its size and the life required, and may range from a few milliamperes to an amp or two.

Secondary Cells

In primary cells some of the various chemicals are used up in producing the electrical energy - a relatively expensive and wasteful process. The maximum current available also is limited. Another type of cell called a secondary cell or accumulator offers the advantage of being able to provide a higher current and is capable of being charged by feeding electrical energy into the cell to be stored chemically, and be drawn out or discharged later as electrical energy again. This process of charging and discharging the cell is capable of repetition for a large number of cycles depending on the chemistry of cell.

A common type of secondary cell is the lead-acid cell such as that used in vehicle batteries. Vehicle batteries are of limited use for amateur radio for two reasons. Firstly they are liable to leak acid if tipped and give off hydrogen (explosive in confined spaces) and secondly, they are designed to float charge and start vehicle engines with very high current surges rather than undergo deep discharge.

Sealed or 'maintenance free' types are available but they must still be used the correct way up or leakage will occur as gas is generated and the pressure increases.

Deep cycle batteries (sometimes called leisure batteries for caravans) are available that are designed to be fully charge/discharge cycled but are considerably more expensive. These batteries are often available at amateur rallies. It is necessary to check which type they are and their origin. Those removed from alarm systems or uninterruptible power supplies (UPS) are likely to have been changed at the five-year maintenance review. They may well have a couple of years' service left and perhaps considerably more. It is a risk but can be a cheap way of obtaining otherwise expensive batteries.

Nickel-metal-hydride (NiMh) is now replacing Nickel-cadmium as the ubiquitous rechargeable battery, the European Union having banned NiCd imports due to the toxic nature of cadmium. Not that any battery technology should be considered entirely safe. They don't have quite the energy capacity per unit volume (energy density) but are much easier to handle. Lithium-ion batteries are still relatively expensive and offer a better energy density. All these types can be damaged by misuse, especially overcharging or the use of the wrong type of charger. As always, check with the manufacturer.

The chapter on Power Supplies discusses batteries in more detail.

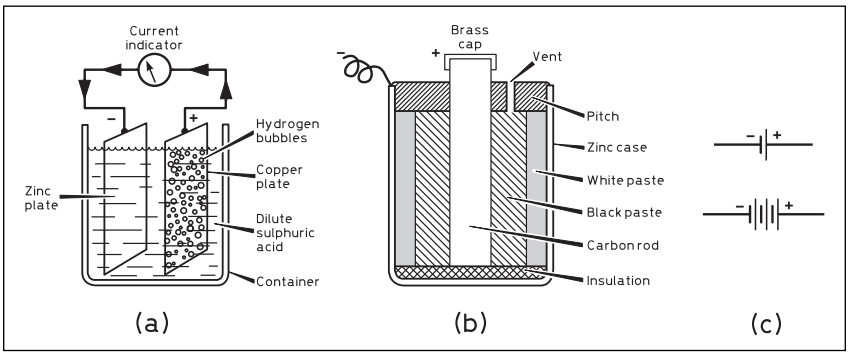


Fig 1.6: The electric cell. A number of cells connected in series is called a battery. (a) A simple electric cell consisting of copper and zinc electrodes immersed in dilute sulphuric acid. (b) Sectional drawing showing construction of a dry cell. (c) Symbol used to represent single cells and batteries in circuit diagram

Mechanical Generators

Mechanical energy may be converted into electrical energy by moving a coil of wire in a magnetic field. Direct-current or alternating-current generators are available in all sizes but the commonest types likely to be met in amateur radio work are:

(a) Petrol-driven AC generators of up to 1 or 2kW output such as are used for supplying portable equipment; and

(b) Small motor generators, sometimes called dynamotors or rotary converters, which furnish up to about 100W of power and comprise a combined low-voltage DC electric motor and a high voltage AC or DC generator. These have two origins; ex military devices providing high voltage DC for use in valve transmitter/receiver equipments and those supplied for use in mobile caravans to provide mains voltage AC from the 12V DC battery. The latter function is now normally achieved by wholly electronic means (commonly known as inverters) but the waveform can be a compromise.

ELECTRICAL POWER

When a current flows through a resistor, eg in an electric fire, the resistor gets hot and electrical energy is turned into heat. The actual rise in temperature depends on the amount of power dissipated in the resistor and its size and shape. In most circuits the power dissipated is insignificant but it is a factor the designer must consider, both in fitting a resistor of adequate dissipation and the effect of the heat generated on nearby devices.

The unit of electrical power is the watt (W). The amount of power dissipated in a resistor is equal to the product of the potential difference across the resistor and the current flowing in it. Thus:

Power (watts) = Voltage (volts) x Current (amperes)

$W = V \times I$

Ohm's Law states: $V = I \times R$ and $I = V/R$

Substituting V or I in the formula for power gives two further formulas:

$$W = I^2R$$

and

$$W = \frac{V^2}{R}$$

These formulas are useful for finding, for example, the power input to a transmitter or the power dissipated in various resistors in an amplifier so that suitably rated resistors can be selected.

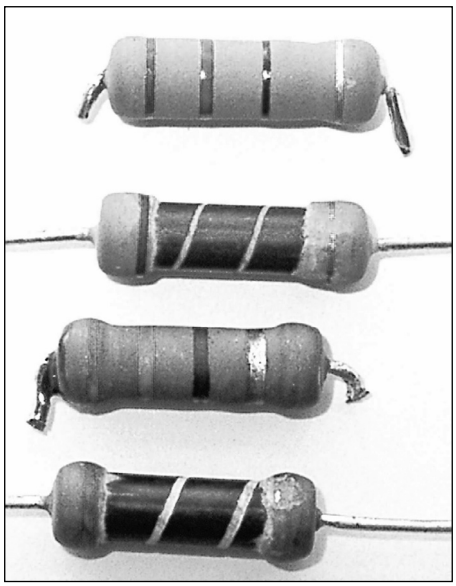


Fig 1.7: Metal film resistors revealed. The metal film is the dark coating on the outside of the white ceramic former, laser cut to make a broad spiral strip

Colour	Value (numbers)	Value (multiplier)	Value (tolerance)
Black	0	1	-
Brown	1	10	1%
Red	2	100	2%
Orange	3	1000 (10^3)	-
Yellow	4	10^4	-
Green	5	10^5	-
Blue	6	10^6	-
Violet	7	-	-
Grey	8	-	-
White	9	-	-
Silver	-	0.01	10%
Gold	-	0.1	5%
No Colour	-	-	20%

Note: A pink band may be used to denote a 'high-stability' resistor. Sometimes an extra band is added to give three figures before the multiplier.

Table 1.3: Resistor colour code

To take a practical case, consider again the circuit of Fig 1.3. The power dissipated in the resistor may be calculated as follows:

Here V = 4V and R = 8Ω. so the correct formula to use is

$$W = \frac{V^2}{R}$$

Inserting the numbers gives

$$W = \frac{4^2}{8} = \frac{16}{8} = 2W$$

It must be stressed again that the beginner should always see that all values are expressed in terms of volts, amperes and ohms in this type of calculation. The careless use of megohms or milliamperes, for example, may lead to an answer several orders too large or too small.

RESISTORS

Resistors Used in Radio Equipment

As already mentioned, a resistor through which a substantial current is flowing will get hot. It follows therefore that in a piece of radio equipment the resistors of various types and sizes that are needed must be capable of dissipating the power as required without overheating.

Generally speaking, radio resistors can be divided roughly into two classes, (a) low power up to 3W and (b) above 3W. The low-power resistors are usually made of carbon film or metal film and may be obtained in a wide range of resistance values from about 10Ω to 10MΩ and in power ratings of 0·1W to 3W. The film usually has a spiral grove to narrow and lengthen the track to obtain the desired resistance, (see **Fig 1.7**) but the inductance introduced is low, typically a few nanohenries. Carbon composition resistors are no longer available. For higher powers, resistors are usually wire-wound on ceramic formers and the very fine wire is protected by a vitreous enamel coating. Typical resistors are shown in the Passive Components chapter.

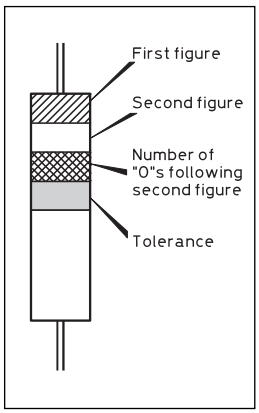


Fig 1.8: Standard resistance value markings

Resistors, particularly the small carbon types, are usually colour-coded to indicate the value of the resistance in ohms and sometimes also the tolerance or accuracy of the resistance. The standard colour code is shown in **Table 1.3**.

The colours are applied as bands at one end of the resistor as shown in **Fig 1.8** As an example, what would be the value of a resistor with the following colour bands: yellow, violet, orange, silver?

The yellow first band signifies that the first figure is 4, the violet second band signifies that the second figure is 7, while the orange third band signifies that there are three zeros to follow; the silver fourth band indicates a tolerance of ±10%. The value of the resistor is therefore 47,000Ω ±10% (47kΩ ±10%).

So far only fixed resistors have been mentioned. Variable resistors, sometimes called potentiometers or volume controls, are also used. The latter are usually panel-mounted by means of a threaded bush through which a quarter inch or 6mm diameter spindle protrudes and to which the control knob is fitted. Low-power high-value variable resistances use a carbon resistance element, and high-power lower-resistance types (up to 10,000Ω) use a wire-wound element. These are not colour coded but the value is printed on the body either directly, eg '4k7', or as '472' meaning '47' followed by two noughts.

Resistors in Series and Parallel

Resistors may be joined in series or parallel to obtain a specific value of resistance. Some caution may be called for since the tolerance will affect the actual value and close tolerance resistors may be needed if the value is critical such as a divider used in a measuring circuit.

When in series, resistors are connected as shown in **Fig 1.9(a)** and the total resistance is equal to the sum of the separate resistances. The parallel connection is shown in **Fig 1.9(b)**, and with this arrangement the reciprocal of the total resistance is equal to the sum of the reciprocals of the separate resistances.

Series connection:

$$R_{total} = R1 + R2 + R3 + etc$$

Parallel connection

$$\frac{1}{R_{total}} = \frac{1}{R1} + \frac{1}{R2} + \frac{1}{R3} \quad etc$$

If only two resistors are in parallel, an easier calculation is possible:

$$\frac{1}{R_{total}} = \frac{1}{R1} + \frac{1}{R2} = \frac{R2}{R1 \times R2} + \frac{R1}{R1 \times R2} = \frac{R1 + R2}{R1 \times R2}$$

Inverting this and writing in 'standard form' (omitting multiplication signs) gives

$$R_{total} = \frac{R1\,R2}{R1 + R2}$$

This is a useful formula since the value of two resistors in parallel can be calculated easily.

It is also useful to remember that for equal resistors in parallel the formula simplifies to

$$R_{total} = \frac{R}{n}$$

where R is the value of one resistor and n is the number of resistors.

Example 1.

Calculate the resistance of a 30Ω and a 70Ω resistor connected first in series and then in parallel.

In series connection:
R = 30 + 70 = 100Ω

In parallel connection, using the simpler formula for two resistors in parallel:

$$R_{total} = \frac{R1\,R2}{R1 + R2} = \frac{30 \times 70}{30 + 70} = \frac{2100}{100} = 21\Omega$$

These two calculations are illustrated in **Fig 1.9(c)**.

Example 2.

Three resistors of 7Ω, 14Ω and 28Ω are connected in parallel. If another resistor of 6Ω is connected in series with this combination, what is the total resistance of the circuit?

The circuit, shown in **Fig 1.9(d)**, has both a series and parallel configuration. Taking the three resistors in parallel first, these are equivalent to a single resistance of R ohms given by:

$$\frac{1}{R_{total}} = \frac{1}{R1} + \frac{1}{R2} + \frac{1}{R3} = \frac{1}{7} + \frac{1}{14} + \frac{1}{28}$$

$$= \frac{4}{28} + \frac{2}{28} + \frac{1}{28} = \frac{7}{28}$$

Inverting (don't forget to do this!)

$$R_{total} = \frac{28}{7} = 4\Omega$$

This parallel combination is in series with the 6Ω resistor, giving a total resistance of 10Ω for the whole circuit.

Note: Often the maths is the most awkward issue and numerical mistakes are easy to make. A calculator helps but whether done with pencil and paper or by calculator, it is essential to estimate an answer first so mistakes can be recognised.

Let us consider the parallel calculation. The answer will be less than the lowest value. The 14Ω and 28Ω resistors will form a resistor less than 14Ω but greater than 7Ω since two 14Ω parallel resistors will give 7Ω and one of the actual resistors is well above 14Ω.

We now have this 'pair' in parallel with the 7Ω resistor. By the same logic, the answer will be less than 7Ω but greater than

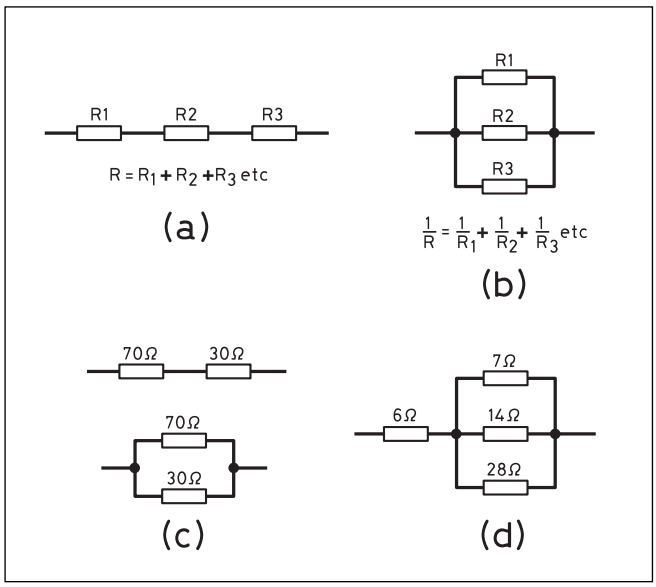


Fig 1.9: Resistors in various combinations: (a) series, (b) parallel, (c) series and parallel, (d) series-parallel. The calculation of the resultant resistances in (c) and (d) is explained in the text

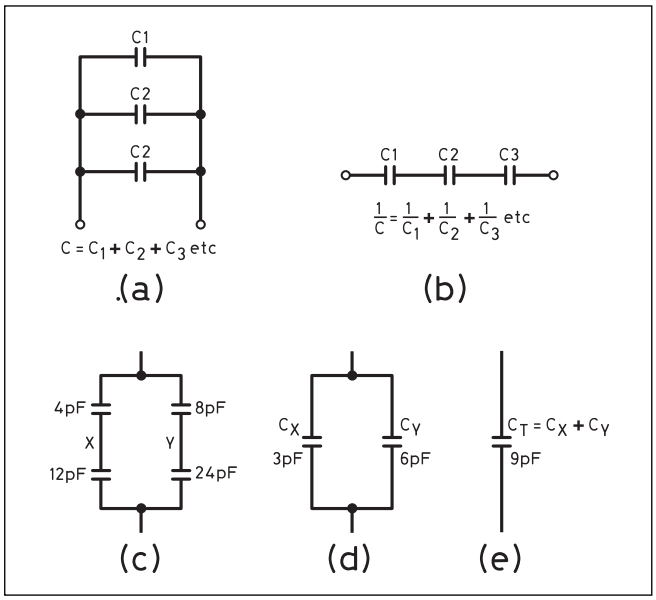


Fig 1.10: Capacitors in various combinations: (a) parallel, (b) series, (c) series-parallel. The calculation of the resultant capacitance of the combination shown in (c) required first the evaluation of each series arm X and Y as shown in (d). The single equivalent capacitance of the combination is shown in (e)

3·5Ω. The calculated answer of 4Ω meets this criterion so stands a good chance of being correct. Results outside the 3·5-7Ω range must be wrong.

This may seem a bit long winded. In words it is. In practice, with a little bit of experience it takes moments.

Capacitors and Capacitance

Capacitors have the property of being able to store a charge of electricity. They consist of two parallel conducting plates or strips separated by an insulating medium called a dielectric. When a capacitor is charged there is a potential difference between its plates.

The capacitance of a capacitor is defined in terms of the amount of charge it can hold for a given potential difference. The capacitance 'C' is given by:

$$C = \frac{q}{V}$$

in units of Farads, F. (q in coulombs and V in volts)

A larger capacitor can hold a greater charge for a given voltage. Or to put it another way, a smaller capacitor will need a greater voltage if it is to store the same charge as a large capacitor.

The Farad is too large a unit for practical use, and typical values range from a few picofarads (pf) to some thousands of microfarads (µF). Unusually the convention is to write 20,000µF rather than 20mF which is, technically, correct notation.

The area of the two plates and the distance between them determines the capacitance. The material of the dielectric also has an effect; materials with a high dielectric constant can considerably enhance the capacitance.

The capacitance can be calculated from the formula

$$C = \frac{\varepsilon_0 \varepsilon_r A}{d}$$

where:

C is in Farads

ε_0 is a natural constant, the permitivity of free space

ε_r is the relative permitivity (or dielectric constant K)

A is the area of the plates in m^2 and

d is their separation in m.

Example:

A tuning capacitor has six fixed plates and five movable plates meshed between them, with a gap of 1mm. Fully meshed at maximum capacitance the area of overlap is $8cm^2$. The dielectric is air, ε_r=1, and ε_0 is 8·85 x 10^{-12} F/m. What is the capacitance?

Firstly, it will be numerically easier to re-arrange the formula so the separation may be given in centimetres and the area in cm^2 and the answer in picofarads. The formula becomes

$$C = \frac{0{\cdot}0885\,A}{d} = \frac{0{\cdot}0885 \times 8 \times 10}{0{\cdot}1} = 71pF$$

The factor 10 above comes from the fact that there are 10 'surfaces' to consider in the meshed capacitor.

Such a capacitor will probably be acceptable for receiving and transmitting up to about 10-20W of power, depending in part on the matching and voltages involved. Higher powers and voltages will require considerably more separation between the plates, reducing the capacitance.

The factor by which the dielectric increases the capacitance compared with air is called the dielectric constant ε_r (or relative permitivity K) of the material. Physicists tend to use the symbol ε_r and engineers the symbol K. You may meet both, depending on which texts you are reading.

Typical values of K are: air 1, paper 2·5, glass 5, mica 7. Certain ceramics have much higher values of K of 10,000 or more. If the dielectric is a vacuum, as in the case of the inter-electrode capacitance of a valve, the same value of K as for air may be assumed. (Strictly, K = 1 for a vacuum and is very slightly higher for air.) The voltage at which a capacitor breaks down depends on the spacing between the plates and the type of dielectric used. Capacitors are often labelled with the maximum working voltage which they are designed to withstand and this figure should not be exceeded.

Capacitors Used in Radio Equipment

The values of capacitors used in radio equipment extend from below 1pF to 100,000µF. They are described in detail in the Passive Components chapter.

Capacitors in Series and Parallel

Capacitors can be connected in series or parallel, as shown in **Fig 1.10**, either to obtain some special capacitance value using a standard range of capacitors, or perhaps in the case of series connection to obtain a capacitor capable of withstanding a greater voltage without breakdown than is provided by a single capacitor. When capacitors are connected in parallel, as in **Fig 1.10(a)**, the total capacitance of the combination is equal to the sum of the separate capacitances. When capacitors are connected in series, as in **Fig 1.10(b)**, the reciprocal of the equivalent capacitance is equal to the sum of the reciprocals of the separate capacitances.

If C is the total capacitance these formulas can be written as follows:

Parallel connection

$$C = C1 + C2 + C3 \text{ etc}$$

Series connection

$$\frac{1}{C_{total}} = \frac{1}{C1} + \frac{1}{C2} + \frac{1}{C3} \quad \text{etc}$$

Similar to the formula for resistors in parallel, a useful equivalent formula for two capacitors in series is

$$C_{total} = \frac{C1\,C2}{C1+C2}$$

The use of these formulas is illustrated by the following *example:*

Two capacitors of 4pF and 12pF are connected in series; two others of 8pF and 24pF are also connected in series. What is the equivalent capacitance if these series combinations are joined in parallel?

The circuit is shown in **Fig 1.10(c)**. Using the formula for two capacitors in series, the two series arms X and Y can be reduced to single equivalent capacitances C_X and C_Y as shown in **Fig 1.10(d)**.

$$C_X = \frac{4\times12}{4+12} = \frac{48}{16} = 3\text{pF}$$

$$C_Y = \frac{8\times24}{8+24} = \frac{96}{32} = 6\text{pF}$$

These two capacitances are in parallel and may be added to give the total effective capacitance represented by the single capacitor C_T in **Fig 1.10(e)**.

$C = C_X + C_Y = 3 + 6 = 9\text{pF}$

The total equivalent capacitance of the four capacitors connected as described is therefore 9pF.

This is not just an academic exercise or something to do for an examination. Consider the circuit in **Fig 1.11**. C is a variable capacitor covering the range 5-100pF, a 20:1 range. It will be seen later that, in a tuned circuit, this will give a tuning range of about 4·5:1 which may well be rather greater than required and cramping the desired range over a relatively small part of the 180 degree rotation normally available. A 5pF capacitor C1 is connected in parallel with C, giving a capacitance range for the pair of 10-105pF by addition of the capacitor values. This is now in series with a 200pF capacitor, so the capacitance range becomes 9·5-69pF, a ratio of 7·3 and a tuning range of 2·7:1.

Many variable capacitors have a small parallel trimmer capacitor included in their construction. Often that is used to set the highest frequency of the tuning range and a series capacitor or a variable inductor (see later in this chapter) used to set the lower end.

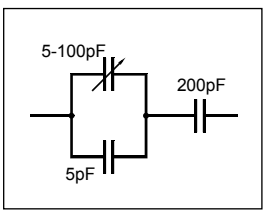


Fig 1.11: Padding a variable capacitor to change its range

MAGNETISM

Permanent Magnets

A magnet will attract pieces of iron towards it by exerting a magnetic force upon them. The field of this magnetic force can be demonstrated by sprinkling iron filings on a piece of thin cardboard under which is placed a bar magnet. The iron filings will map out the magnetic field as sketched in **Fig 1.12** and the photograph **Fig 1.13**. It will be seen that the field is most intense near the ends of the magnet, the centres of intensity being called the poles, and lines of force spread out on either side and continue through the material of the magnet from one end to the other.

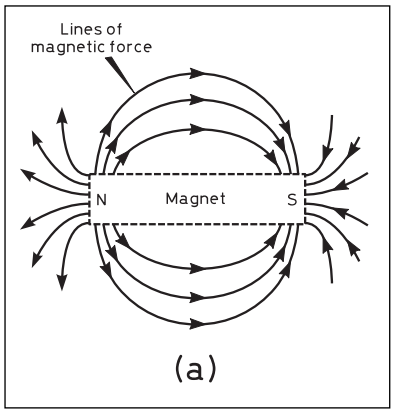


Fig 1.12: magnetic field produced by a bar magnet

If such a magnet is suspended so that it can swing freely in a horizontal plane it will always come to rest pointing in one particular direction, namely towards the Earth's magnetic poles, the Earth itself acting as a magnet. A compass needle is simply a bar of magnetised steel. One end of the magnet (N) is called the north pole, which is an abbreviation of 'north-seeking pole' and the other end (S) a south pole or south-seeking pole. It is an accepted convention that magnetic force acts in the direction from N to S as indicated by the arrows on the lines of force in Fig 1.12.

If two magnets are arranged so that the north pole of one is near the south pole of another, there will be a force of attraction between them, whereas if similar poles are opposite one another, the magnets will repel: see **Fig 1.14**.

Permanent magnets are made from certain kinds of iron, nickel and cobalt alloys and certain ceramics, the hard ferrites (see Passive Components chapter) and retain their magnetism more or less indefinitely. They have many uses in radio equipment, such as loudspeakers, headphones, some microphones, cathode-ray tube focusing arrangements and magnetron oscillators.

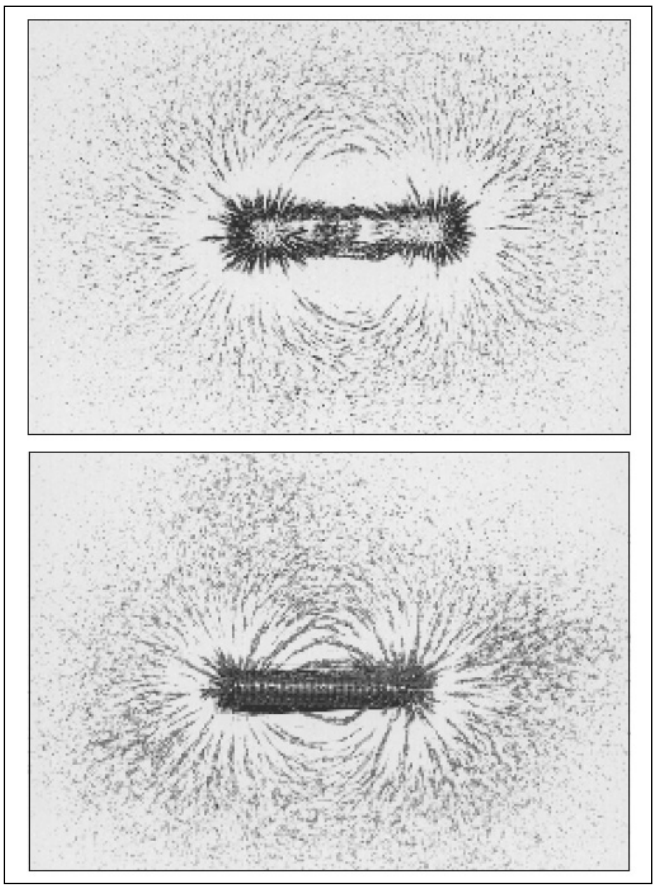


Fig 1.13: Iron filings mapping out the magnetic field of (top) a bar magnet, and (bottom) a solenoid carrying current

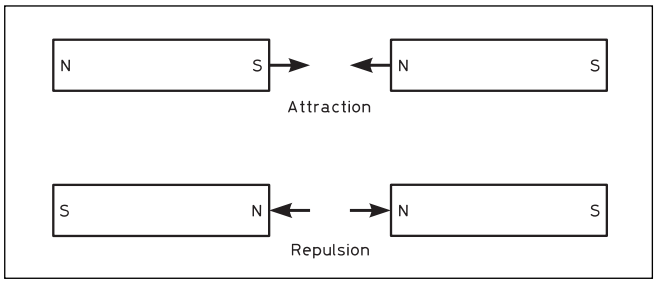


Fig 1.14: Attraction and repulsion between bar magnets

Other types of iron and nickel alloys and some ceramics (the soft ferrites), eg soft iron, are not capable of retaining magnetism, and cannot be used for making permanent magnets. They are effective in transmitting magnetic force and are used as cores in electromagnets and transformers. These materials concentrate the magnetic field by means of a property called permeability. The permeability is, essentially, the ratio of the magnetic field with a core to that without it.

Electromagnets

A current of electricity flowing through a straight wire exhibits a magnetic field, the lines of force of which are in a plane perpendicular to the wire and concentric with the wire. If a piece of cardboard is sprinkled with iron filings, as shown in **Fig 1.15**, they will arrange themselves in rings round the wire, thus illustrating the magnetic field associated with the flow of current in the wire. Observation of a small compass needle placed near the wire would indicate that for a current flow in the direction illustrated the magnetic force acts clockwise round the wire. A reversal of current would reverse the direction of the magnetic field.

The corkscrew rule enables the direction of the magnetic field round a wire to be found. Imagine a right-handed corkscrew being driven into the wire so that it progresses in the direction of current flow; the direction of the magnetic field around the wire will then be in the direction of rotation of the corkscrew.

The magnetic field surrounding a single wire is relatively weak. Forming the wire into a coil will combine the field of each turn producing a stronger field. A much greater increase in the magnetic field can then be achieved by inserting a piece of soft iron, called a core, inside the coil.

Fig 1.16 and the bottom photograph in Fig 1.13 show the magnetic field produced by a coil or solenoid as it is often called. It will be seen that it is very similar to that of a bar magnet also shown in Fig 1.13. A north pole is produced at one end of the coil and a south pole at the other. Reversal of the current will reverse the polarity of the electromagnet. The polarity can be deduced from the S rule, which states that the pole that faces an observer looking at the end of the coil is a south pole if the current is flowing in a clockwise direction; see Fig 1.16. The current is the conventional current, not the direction of flow of electrons.

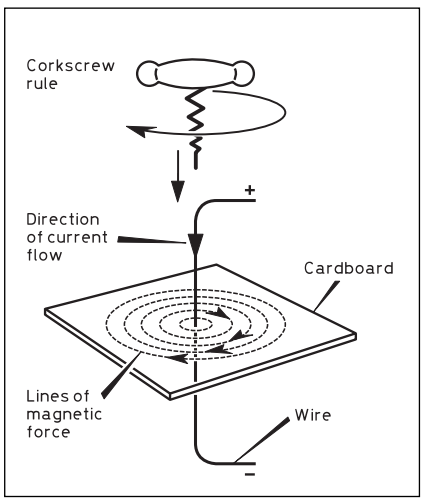


Fig 1.15: Magnetic field produced by current flowing in a straight wire

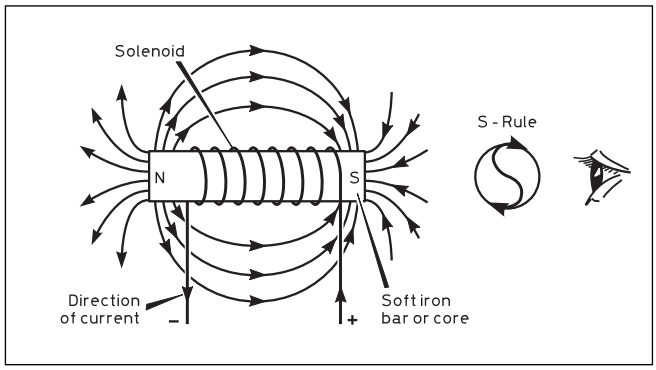


Fig 1.16: The S rule for determining the polarity of an electromagnet

The strength of a magnetic field produced by a current is directly proportional to the current, a fact made use of in moving coil meters (see Test equipment chapter). It also depends on the number of turns of wire, the area of the coil, and the permeability of the core.

Interaction of Magnetic Fields

Just as permanent magnets can attract or repel, so can electromagnets. If one of the devices, a coil for example, is free to move, then a current will cause the coil to move with a force or at a rate related to the magnitude of the current. The moving coil meter relies on this effect, balancing the force caused by the current against a return spring so that the movement or deflection indicates the magnitude of the current.

Electromagnetic Induction

If a bar magnet is plunged into a coil as indicated in **Fig 1.17(a)**, the moving-coil microammeter connected across the coil will show a deflection. The explanation of this phenomenon, known as electromagnetic induction, is that the movement of the magnet's lines of force past the turns of the coil causes an electromotive force to be induced in the coil which in turn causes a current to flow through the meter. The magnitude of the effect depends on the strength and rate of movement of the magnet and the size of the coil. Withdrawal of the magnet causes a reversal of the current. No current flows unless the lines of force are moving relative to the coil. The same effect is obtained if a coil of wire is arranged to move relative to a fixed magnetic field. Dynamos and generators depend for their operation on the principle of electromagnetic induction.

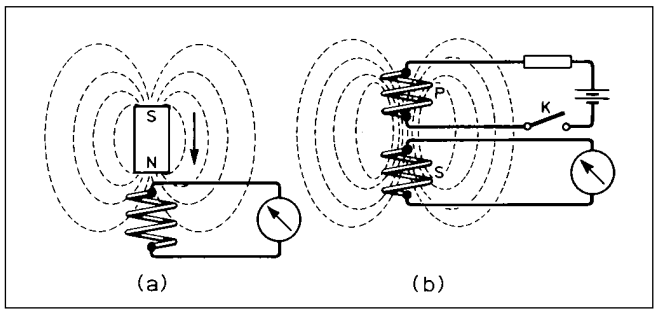


Fig 1.17: Electromagnetic induction. (a) Relative movement of a magnet and a coil causes a voltage to be induced in the coil. (b) When the current in one of a pair of coupled coils changes in value, current is induced in the second coil

Consider a pair of coils of wire are arranged as shown in **Fig 1.17(b)**. When the switch K is open there is no magnetic field from the coil P linking the turns of the coil S, and the current through S is zero. Closing K will cause a magnetic field to build up due to the current in the coil P. This field, while it is building up, will induce an EMF in coil S and cause a current to flow through the meter for a short time until the field due to P has reached a steady value, when the current through S falls to zero again. The effect is only momentary while the current P is changing.

The fact that a changing current in one circuit can induce a voltage in another circuit is the principle underlying the operation of transformers.

Self-inductance

Above we considered the effect of a change of current in coil P inducing a voltage in coil S. In fact the changing field also induces an EMF in coil P even though it is the current in coil P that is causing the effect. The induced EMF is of a polarity such that it tends to oppose the original change in current.

This needs some care in understanding. On closing the switch K the EMF induced in P will tend to oppose the build up of current, that is it will oppose the voltage from the battery and the current will build up more slowly as a result. However if K is opened the current will now fall and the EMF induced in P will be of opposite polarity, trying to keep the current flowing. In reality the current falls more slowly.

This effect of the induced EMF due to the change in current is known as *inductance*, usually denoted by the letter L. If the current is changing at the rate of one amp per second (1A/s), and the induced EMF is one volt, the coil has an inductance of one Henry (1H). A 2H coil will have 2V induced for the same changing current. Since the induced voltage is in the coil containing the changing current this inductance is properly called *self-inductance*.

$$\text{Inductance } L = \frac{V}{dI/dt}$$

where dI is a small change in current and dt is the time for that change.

That is

$$dI/dt$$

is the rate of change of current in A/s

The inductance values used in radio equipment may be only a very small fraction of a henry, the units millihenry (mH) and microhenry (µH) meaning one thousandth and one millionth of a henry respectively are commonly used.

The inductance of a coil depends on the number of turns, on the area of the coil and the permeability of the core material on which the coil is wound. The inductance of a coil of a certain physical size and number of turns can be calculated to a fair degree of accuracy from formulas or they can be derived from coil charts.

Mutual Inductance

A changing current in one circuit can induce a voltage in a second circuit: as in **Fig 1.16(b)**. The strength of the voltage induced in the second circuit depends on the closeness or tightness of the magnetic coupling between the circuits; for example, if both coils are wound together on an iron core practically all the magnetic flux from the first circuit will link with the turns of the second circuit. Such coils would be said to be tightly coupled whereas if the coils were both air-cored and spaced some distance apart they would be loosely coupled.

The mutual inductance between two coils is also measured in henrys, and two coils are said to have a mutual inductance of one henry if, when the current in the primary coil changes at a rate of one ampere per second, the voltage across the secondary is one volt. Mutual inductance is often denoted in formulas by the symbol M.

Inductors in Series and Parallel

Provided that there is no mutual coupling between inductors when they are connected in series, the total inductance obtained is equal to the sum of the separate inductances. When they are in parallel the reciprocal of the total inductance is equal to the sum of the reciprocals of the separate inductances.

If L is the total inductance (no mutual coupling) the relationships are as follows:

Series connection

$$L_{total} = L1 + L2 + L3 \text{ etc}$$

Parallel connection

$$\frac{1}{L_{total}} = \frac{1}{L1} + \frac{1}{L2} + \frac{1}{L3} \quad \text{etc}$$

These two formulas are of the same format as the formula for resistors and the special case of only two resistors is also true for inductors. In reality paralleled inductors are uncommon but may be found in older radio receivers that are permeability tuned. That is where the inductance of a tuned circuit (see later) is varied rather than the more common case of variable capacitors.

CR Circuits and Time Constants

Fig 1.18(a) shows a circuit in which a capacitor C can either be charged from a battery of EMF E, or discharged through a resistor R, according to whether the switch S is in position a or b.

If the switch is thrown from b to a at time ta, current will start to flow into the capacitor with an initial value E/R. As the capacitor charges the potential difference across the capacitor increases, leaving less PD across the resistor, and the current through the circuit therefore falls away, as shown in the charging portion of **Fig 1.18(b)**. When fully charged to the voltage E the current will have dropped to zero.

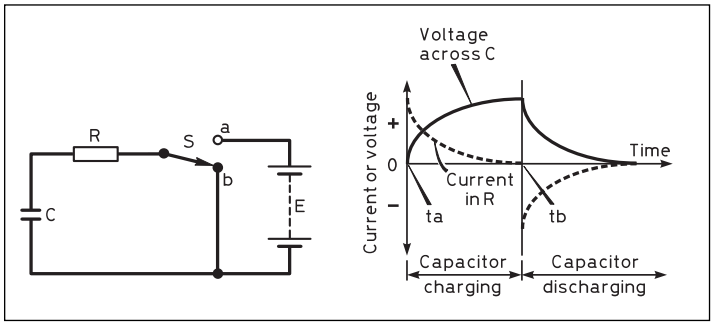


Fig 1.18: In (a) a capacitor C can be charged or discharged through the resistor R by operating the switch S. The curves of (b) show how the voltage across the capacitor and the current into and out of the capacitor vary with time as the capacitor is charged and discharged. The curve for the rise and fall of current in an LR circuit is similar to the voltage curve for the CR circuit

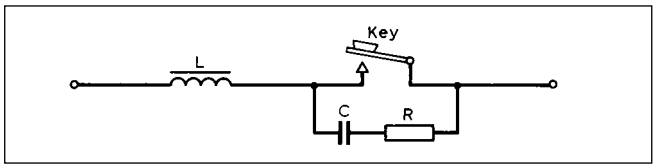


Fig 1.19: Typical key-click filter. L serves to prevent a rise of current. C, charging through R, serves to continue flow of current briefly when key contacts open. Typical values: L = 0.01 to 0·1H, C = 0·01 to 0·1µF, R = 10 to 100 ohms

At time tb, the switch is thrown back to b, the capacitor will discharge through the resistor R, the current being in the opposite direction to the charging current, starting at a value -V/R and dying away to zero. As the capacitor discharges, the PD across its plates falls to zero as shown in the discharge portion of Fig 1.18(b).

The voltage at any point during the charge cycle is given by the formula

$$V = V_b\left(1 - e^{-\frac{t}{CR}}\right)$$

Where V_b is the battery voltage and t is the time, in seconds, after the switch is thrown. C and R are given in Farads and Ohms respectively and e is the base of natural logarithms.

This is known as an exponential formula and the curve is one of a family of exponential curves.

As the capacitor approaches fully charged, or discharged, the current is very low. In theory an infinite time is needed to fully complete charging. For practical purposes the circuit will have charged to 63% (or 1/e) in the time given by CR and this time is known as the *time constant* of the circuit. On discharge the voltage will have fallen to 37% of the initial voltage (1-1/e) after one time constant. Over the next time constant it will have fallen to 37% of its new starting voltage, or 14% of the original voltage. Five time constants will see the voltage down to 0·7%.

Time constant τ = CR seconds

As an example, the time constant of a capacitance of 0·01µF (10^{-8}F) and a resistance of 47kΩ (4·7 x $10^4\Omega$) is:

$\tau = 10^{-8} \times 4{\cdot}7 \times 10^4$

$= 4{\cdot}7 \times 10^{-4}$s = 0·47ms

High voltage power supplies should have a bleeder resistor across the smoothing capacitor to ensure lethal voltages are removed before anyone can remove the lid after switching off, thinking it is safe! The time constant here may be a few tens of seconds. In audio detector circuits, the capacitor following the detector diode is chosen so that, along with the load resistance, the time constant is rather longer than the period of the intermediate frequency to give good smoothing or filtering. However it also needs to be rather shorter than the period of the highest audio frequency, or unwanted attenuation of the higher audio notes will occur.

A similar constraint applies in AGC (automatic gain control) circuits where the receiver must be responsive to variation caused by RF signal fading without affecting the signal modulation.

Digital and pulse circuits may rely on fast transient waveforms. Here very short time constants are required, unwanted or stray capacitance and inductance must be avoided.

LR Circuits

Inductors oppose the change, rise or fall, of current. The greater the inductance, the greater the opposition to change. In a circuit containing resistance and inductance the current will not rise immediately to a value given by V/R if a PD V is applied but will rise at a rate depending on the L/R ratio. LR circuits also have a time constant, the formula is:

τ = L/R where L is in henries and R in ohms

Fig 1.18 showed the rise in capacitor voltage as it charged through a resistor; the curve of the rise in current in an LR circuit is identical. In one time constant, the current will have risen to 63%, (1 - 1/e) of its final value, or to decay to 37% of its initial value.

An example of the use of an inductor to slow the rise and fall of current is the Morse key filter shown in **Fig 1.19**. Slowing the rise of current as the key is depressed is simple enough but when the key is raised the circuit is cut and a high voltage can be developed which will cause sparking at the key contacts. The C and R across the key provide a path for the current to flow momentarily as it falls to zero. The reason for the use of this filter is discussed in the Morse chapter.

ALTERNATING CURRENT

So far we have concerned ourselves with uni-directional current flow or direct current (DC). Audio and radio circuits rely heavily on currents (and voltages) that change their polarity continuously; alternating currents (AC).

Fig 1.20(a) shows a battery and reversing switch. Continually operating the switch would cause the current in the resistor to flow in alternate directions as shown by the waveform in **Fig 1.20(b)**. The waveform is known as a square wave.

Alternating current normally has a smoother waveform shown in **Fig 1.20(d)** and can be produced by an electrical generator shown in **Fig 1.20(c)**. Consider for a moment a single wire in the coil in Fig 1.19(c) and remember that if a wire moves through a magnetic field an EMF is induced in the wire proportional to the relative velocity of the wire.

At the bottom the wire is travelling horizontally and not cutting through the field; no EMF is induced. After 90 degrees of rotation the wire is travelling up through the field at maximum velocity and the induced EMF is at a maximum. At the top the EMF has fallen to zero and now reverses polarity as the wire descends, falling to zero again at the bottom where the cycle begins again. The vertical component of the velocity follows a sinusoidal pattern, as does the EMF; it is a sine wave. The precise shape can be plotted using mathematical 'sine' tables used in geometry.

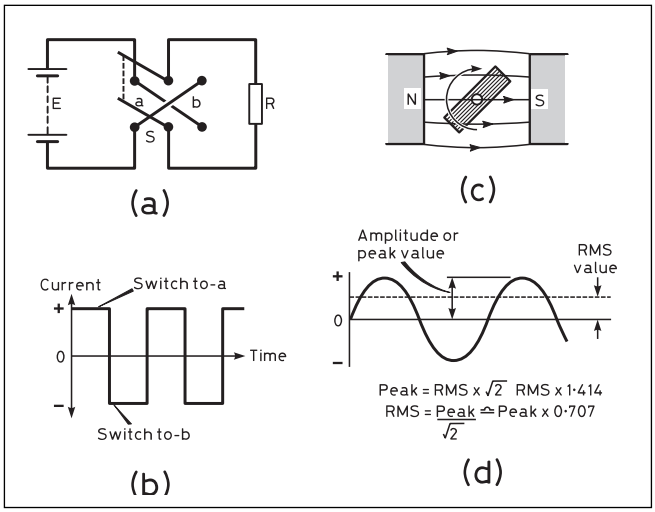


Fig 1.20: Alternating current. A simple circuit with a current-reversing switch shown at (a) produces a square-wave current through the resistor R as shown in (b). When a coil is rotated in a magnetic field as in (c) the voltage induced in the coil has a sinusoidal waveform (d)

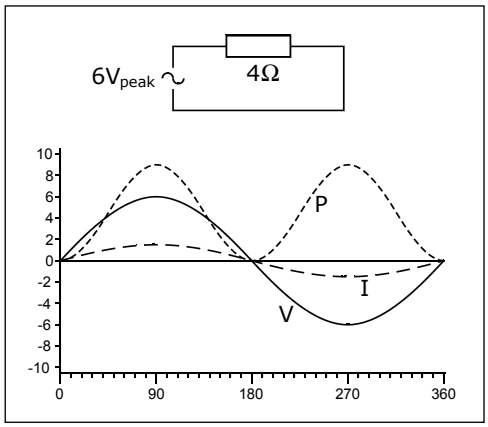


Fig 1.21: The power dissipated in a resistor with a sine wave of current

Specifying an AC Waveform

For AC there are two parameters that must be quoted to define the current or voltage. Like DC we must give the magnitude or *amplitude* as it is called and we must also say how fast the cycles occur.

If t is the time for one cycle (in seconds) then 1/t will give the number of cycles occurring in a second; this is known as the frequency.

t = 1/f and f=1/t

The unit of frequency is cycles per second and is given the name Hertz, abbreviation Hz.

Example: the UK mains has a frequency of 50Hz, what is the periodic time?

f =50Hz so the time for 1 cycle t =1/f = 1/50 second or 0·02s.

RMS Values

Specifying the amplitude is more interesting. It would seem sensible to quote the maximum amplitude and allow the reader to find the amplitude at other parts of the cycle by looking up the values in sine tables. In actual fact the RMS (root mean square) value is used but what does that mean?

Consider passing an AC current through a resistor. The power dissipated in the resistor will vary over the cycle as shown in **Fig 1.21**.

The peak current is 6V/4Ω = 1·5A and peak power is 6V x 1·5A = 9W. The power curve is symmetrical about 4·5W and the average power over the cycle is indeed 4·5W. The DC voltage that would produce 4·5W in a 4Ω resistor is just over 4V (4·24V). This is the RMS value of the 6V peak waveform. It is that value of DC voltage (or current) that would produce the same heating effect in a resistor.

The RMS value is related to the peak value by

$$\text{RMS} = \frac{\text{Peak}}{\sqrt{2}}$$

or

$$\text{Peak} = \text{RMS} \times \sqrt{2}$$

It may help to remember that

$$\sqrt{2} = 1 \cdot 4142$$

and

$$1/\sqrt{2} = 0 \cdot 707$$

Example: suppose your mains supply is 240V RMS, what is the peak value?

$$\text{Peak} = \text{RMS} \times \sqrt{2} \quad = 240 \times \sqrt{2} \quad = 340\text{V}$$

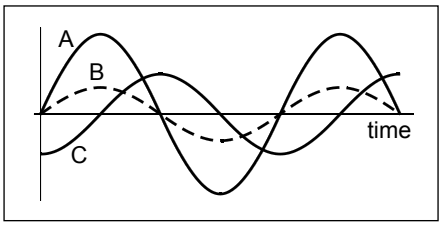


Fig 1.22: Relative phase

Other Frequencies

Audible frequencies range from around 50-100Hz up to 20kHz for a young child, rather less for adults. Some animals can hear up to around 40-50kHz. Dogs will respond to an ultrasonic whistle (above human hearing) and bats use high pitch sounds around 50kHz for echo location.

Radio frequencies legally start at 9kHz, and 20kHz is used for worldwide maritime communication. Below 30kHz is known as the VLF band (very low frequency; LF is 30-300kHz; MF (medium frequencies) is 300-3000kHz. The HF, high frequency, band is 3 to 30MHz, although amateurs often refer to the 1·8MHz amateur band as being part of HF. VHF is 30-300MHz and UHF (ultra-high frequencies) is 300-3000MHz. Above that is the SHF band 3-30GHz, super-high frequencies. Also, by common usage, the 'microwave band' is regarded as being above 1GHz.

Phase and Harmonics

Two waveforms that are of the same frequency are *in phase* if they both start at the same point in time. If one waveform is delayed with respect to the other then they are not in phase and the phase difference is usually expressed as a proportion of a complete cycle of 360°. This is shown in **Fig 1. 22**. Waveforms A and B are in phase but are of different amplitudes. Waveform C is not in phase with A, it lags A by 1/4 cycle or 90°, or A leads C by 90°. It lags because the 'start' of the cycle is to the right on the time axis of the graph, that is, it occurs later in time. The 'start' is conventionally regarded as the zero point, going positive.

If two waveforms are of different frequencies then their phase relationships are continuously changing. However, if the higher frequency is an exact multiple of the lower, the phase relationship is again constant and the pattern repeats for every cycle of the lower frequency waveform. These multiple frequencies are known as harmonics of the lower 'fundamental' frequency. The second harmonic is exactly twice the frequency and the third, three times the frequency.

Fig 1.23 shows a fundamental and a third harmonic at 1/6 amplitude. The heavy line is the sum of the fundamental and the harmonic. It shows a phenomenon known as harmonic distortion.

If a sine wave is distorted, perhaps by over-driving a loudspeaker or by imperfections (often deliberate) in electronic circuits, then the distortion can be regarded as the original sine wave plus the right amplitudes and phases of various harmonics required to produce the actual distorted waveform. The procedure to determine the required harmonics is known as Fourier analysis.

It is important to appreciate that in distorting an otherwise clean sine wave, the harmonics really are created; new frequen-

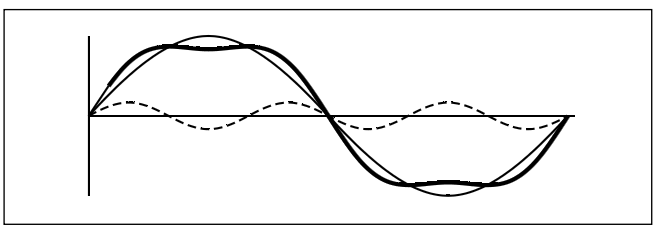


Fig 1.23: Harmonics of a sine wave

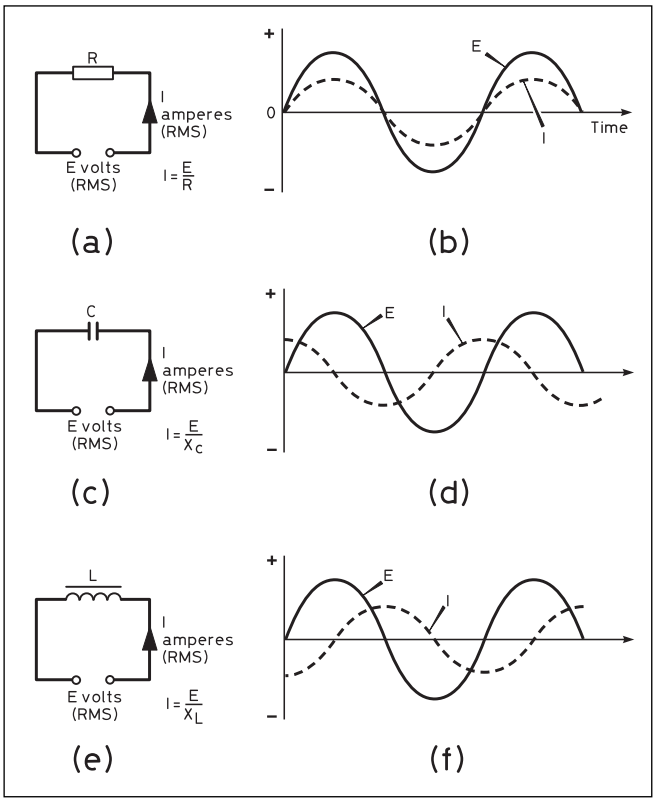


Fig 1.24: Voltage and current relationships in AC circuits comprising (a) resistance only, (c) capacitance only and (e) inductance only

cies are now present that were not there prior to the distortion. If a distorted sine wave is observed using a spectrum analyser, the harmonic frequencies can be seen at their relative amplitudes. Moreover, if they can be filtered out, the distortion will have been removed. This will prove a useful technique later in the discussions on radio receivers, intermediate frequencies and AGC (automatic gain control).

AC Circuit Containing Resistance Only

Ohm's Law is true for a resistor at every point in a cycle of alternating current or voltage. The current will flow backwards and forwards through the resistor under the influence of the applied voltage and will be in phase with it as shown in **Fig 1.24(a)**.

The power dissipated in the resistor can be calculated directly from the power formulas, provided that RMS values for voltage and current are used.

$$P = VI \qquad P = \frac{V^2}{R} \qquad P = I^2R$$

If peak values of voltage and current are used, these formulas become:

$$P = \frac{V_P}{\sqrt{2}} \times \frac{I_P}{\sqrt{2}} = \frac{V_P I_P}{2} \qquad P = \frac{V_P{}^2}{2R}$$

AC Circuit Containing Capacitance Only; Reactance of a Capacitor

If an alternating current is applied to a capacitor the capacitor will charge up. The PD at its terminals will depend on the amount of charge gained over the previous part cycle and the capacitance of the capacitor. With an alternating current, the capacitor will charge up first with one polarity, then the other. **Fig 1.24(c)** shows the circuit and **Fig 1.24(d)** shows the voltage (E) and current (I) waveforms. Inspection of **Fig 1.24(d)** shows that the voltage waveform is 90 degrees lagging on the current. This is explained by remembering that the voltage builds up as more charge flows into the capacitor. When the current is a maximum, the voltage is increasing at its maximum rate. When the current has fallen to zero the voltage is (momentarily) constant at its peak value.

There is no such thing as the resistance of a capacitor. Consider **Fig 1.24(d)** again. At time zero the voltage is zero yet a current is flowing, a quarter cycle (90°) later the voltage is at a maximum yet the current is zero. The resistance appears to vary between zero and infinity! Ohm's Law does not apply.

What we can do is to consider the value of V_{rms}/I_{rms}. It is not a resistance but it is similar to resistance because it does relate the voltage to the current. The quantity is called *reactance* (to denote to 90° phase shift between V and I) and is measured in ohms, Ω.

Remember that for the same charge, a larger value capacitor will have a lower potential difference between the plates. This suggests a larger value capacitor will have a lower reactance. Also, as the frequency rises, the periodic time falls and the charge, which is current x time, also falls. The reactance is lower as the frequency rises, it varies with frequency. The proof of this requires integral calculus which is beyond the scope of the book but the formula for the reactance of a capacitor is

$$X_C = \frac{1}{2\pi fC}$$

where f is the frequency in Hertz and C the capacitance in Farads. You may also meet this formula written using the symbol ω instead of 2πf. The lower case Greek letter omega, ω, gives the frequency in radians per second - remember there are 2π radians in 360° and ω is called the angular frequency.

Example:

A capacitor of 500pF is used in an antenna matching unit and is found to have 400V across it when the transmitter is set to 7MHz. Calculate the reactance and current flowing.

$$X_C = \frac{1}{2\pi fC} = \frac{1}{2\pi \times 7\times10^6 \times 500\times10^{-12}}$$

$$X_C = \frac{1}{7\pi\times10^{-3}} = \frac{1000}{7\pi} = 45\cdot5\Omega$$

With 400V applied the current is

$$I = \frac{V}{X_C} = \frac{400}{45\cdot5} = 8\cdot8A$$

AC Circuit Containing Inductance Only; Reactance of a Coil

The opposition of an inductance to alternating current flow is called the inductive reactance of the coil: **Fig 1.24(e)** shows the circuit and the current and voltage waveforms. If a potential difference is applied to a coil, the current builds up slowly, how slowly depending on the inductance. Similarly, if the PD is reduced, the current in the coil tries to keep flowing. With AC the current waveform lags the voltage waveform by 90°. As the frequency increases the current in the coil is being expected to change more rapidly and the opposition is greater, the magnitude of the current is less and the reactance higher.

This leads to the formula for the reactance of a coil as:

$$X_L = 2\pi fL$$

As before X_L is in ohms, L in henries and f in Hertz.

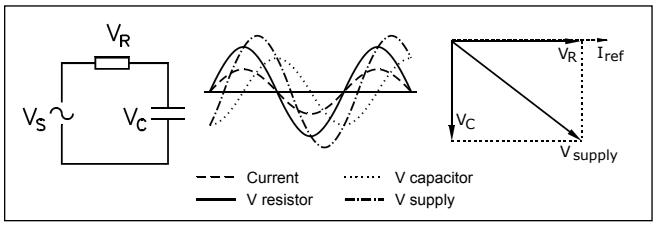


Fig 1.25: Voltage and current in a circuit containing a resistor and capacitor

Example. A coil has an inductance of 5µH, calculate its reactance at 7MHz.

$$X_L = 2\pi fL = 2\pi \times 7\times10^6 \times 5\times10^{-6}$$
$$= 70\pi = 220\Omega$$

AC Circuits with R and C or R and L

Consider **Fig 1.25**, a resistor and capacitor in series. The same current flows through both R and C and is shown by the dashed waveform. The voltage across R will be in phase with the current, shown solid, and the voltage across C will lag by 90°, shown dotted. The supply voltage is the sum of the voltages across R and C. Since the voltages across R and C are not in phase, they cannot simply be added together; they must be added graphically, shown by the dash-dot line. For example when V_R is at a maximum, V_C is zero, so the sum should be coincident with V_R.

The waveform diagram is messy to say the least, it has been shown once to illustrate the situation. Fortunately there is a simpler way, the phasor diagram shown to the right in Fig 1.25. By convention the current, which is common to both components, is drawn horizontally to the right. The vector or phasor representing the voltage across the resistor is drawn parallel to it and of a length representing the magnitude of the voltage across R. V_C lags by 90°, shown as a phasor downwards, again of the correct length to represent the magnitude of the voltage across C. The vector sum, the resultant, gives the supply voltage. The length of the arrow represents the actual voltage and the angle will give its phase.

Since V_R and V_C are at right angles we can use Pythagoras' Theorem to obtain a formula for the voltage.

$$V_{supply} = \sqrt{V_R{}^2 + V_C{}^2}$$

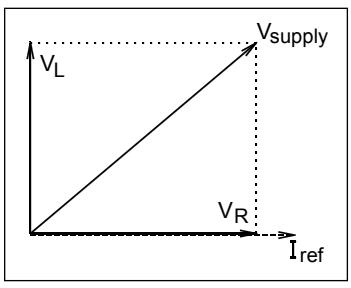
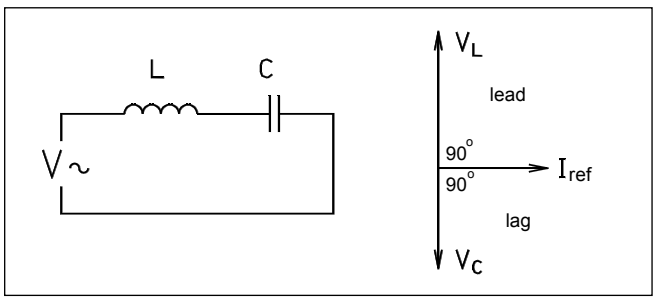


Fig 1.26: Phasor diagram for R and L

(below) Fig 1.27: L and C in series



By a similar argument we can also obtain:

$$Z = \sqrt{R^2 + X_C{}^2}$$

where Z is the impedance of the CR circuit, R is the resistance and X_C the reactance.

Impedance (symbol Z) is the term used to denote the 'resistance' of a circuit containing both resistance and reactance and is also measured in ohms. The convention is that *resistance* means V and I are in phase; *reactance* means 90° (capacitive or inductive) and *impedance* means somewhere between, or indeterminate.

Fig 1.26 shows the phasor diagram for R and L. Since the coil voltage now leads the current, it is drawn upwards but the geometry and the formula are the same.

$$V_{supply} = \sqrt{V_R{}^2 + V_L{}^2}$$

$$Z = \sqrt{R^2 + X_L{}^2}$$

Capacitance and Inductance, Resonance

In the series circuit containing C and L the current is still the common factor and the voltage across C will lag the current by 90° whilst the voltage across the inductor will lead by 90°. Consequently these two voltages will be 180° out of phase; that is, in anti-phase. The two voltages will tend to cancel rather than add. **Fig 1.27** shows the circuit and the phasor diagram. As drawn, the voltage across the capacitor is less than the voltage across the inductor, indicating the inductor has the greater reactance. In general the total reactance is given by:

$X = X_L - X_C$.

However the reactance varies with frequency as shown in **Fig 1.28**. If the reactances of C and L are equal, the voltages V_C and V_L will also be equal and will cancel exactly. The voltage across the circuit will be zero. This occurs at a frequency where the curves intersect in Fig 1.28, it is called the resonant frequency.

At resonance the total reactance $X = X_L - X_C$ will be zero and if $X_L = X_C$ then:

$$2\pi fL = \frac{1}{2\pi fC}$$

Rearranging this equation gives:

$$f = \frac{1}{2\pi\sqrt{LC}} \quad \text{or} \quad C = \frac{1}{4\pi^2 f^2 L} \quad \text{or} \quad L = \frac{1}{4\pi^2 f^2 C}$$

Fig 1.29 shows a circuit containing a capacitor, inductor and resistor. This is representative of real circuits, even if the resistance is merely that of the coil. This may be higher than expected due to 'skin effect'; a phenomenon explained later in the chapter. The overall impedance of the circuit is high when away from resonance but it falls, towards 'R' the value of resistance, as resonance is approached. The series resonant circuit is sometimes called an acceptor circuit because it accepts current at resonance. The current is also shown in Fig 1.29.

Example 1:

A 50µH inductor and a 500pF capacitor are connected in series; what is the resonant frequency?

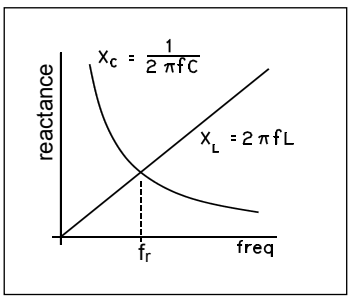


Fig 1.28: Reactance of L and C

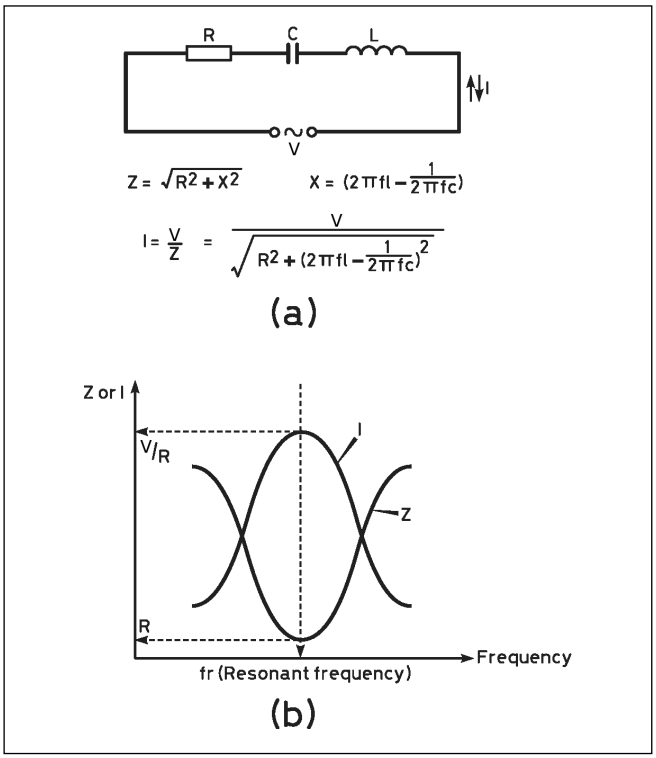


Fig 1.29: The series-resonant circuit. The curves shown at (b) indicate how the impedance and the current vary with frequency in the type of circuit shown at (a)

The formula is:

$$f = \frac{1}{2\pi\sqrt{LC}}$$

So inserting the values:

$$f = \frac{1}{2\pi\sqrt{50\times10^{-6}\times500\times10^{-12}}} = \frac{1}{2\pi\sqrt{5\times10^{-5}\times5\times10^{-10}}}$$

$$f = \frac{1}{2\pi\sqrt{25\times10^{-15}}} = \frac{1}{2\pi\sqrt{2\cdot5\times10^{-14}}}$$

$$f = \frac{1}{2\pi\times1.58\times10^{-7}} = \frac{10^7}{3\cdot16\pi} = 1\text{MHz}$$

Note: The calculation has been set out deliberately to illustrate issues in handling powers and square roots. In line 1 the numbers were put in standard form, that is a number between 1 and 10, with the appropriate power of 10. This then resulted in 25 x 10^{-15} inside the square root sign. The root of 25 is easy but the root of an odd power of 10 is not. Consequently the sum has been changed to 2·5 x 10^{-14}. It so happens that this is also then in standard form but the real purpose was to obtain an even power of ten, the square root of which is found by halving the 'power' or exponent.

There is another method which may be easier to calculate; namely to use the formula for f^2 but remembering to take the square root at the end.

The formula is:

$$f^2 = \frac{1}{4\pi^2LC}$$

An advantage is that π^2 is 9·87 which is close enough to 10, remembering the error is much less than the tolerance on the components.

Example 2:

A vertical antenna has a series inductance of 20µH and capacitance of 100pF. What value of loading coil is required to resonate at 1·8MHz?

The formula is

$$L = \frac{1}{4\pi^2f^2C} = \frac{1}{40\times(1\cdot8\times10^6)^2\times10^{-10}}$$

$$L = \frac{1}{40\times3\cdot24\times10^{12}\times10^{-10}} = \frac{1}{12960}$$

which is 77µH.

The antenna already has 20µH of inductance, so the loading coil needs to add another 57µH. A coil some 10cm diameter of 30 turns spread over 10cm could be suitable.

Parallel Resonance

The L and C can also be connected in parallel as shown in **Fig 1.30**. Resistor 'r' is the internal resistance of the coil which should normally be as low as reasonably possible. The voltage is common to both components and the current in the capacitor will lead by 90° while that in the coil lags by 90°.

This time the two currents will tend to cancel out, leaving only a small supply current. The impedance of the circuit will increase dramatically at resonance as shown in the graph in Fig.1.30.

This circuit is sometimes called a rejector circuit because it rejects current at resonance; the opposite of the series resonant circuit.

Magnification Factor, Q

Consider again the series resonant circuit in Fig 1.29. At resonance the overall impedance falls to the resistance R and the current is comparatively large. The voltage across the coil is still given by V = I x X_L and similarly for the capacitor V = I x X_C. Thus the voltages across L and C are very much greater than the voltage across the circuit as a whole which is simply V = I x R. Remember X_C and X_L are very much larger than R.

The magnification factor is the ratio of the voltage across the coil (or capacitor) to that across the resistor.

The formula are:

$$Q = \frac{X_L}{R} \text{ or } \frac{X_C}{R}$$

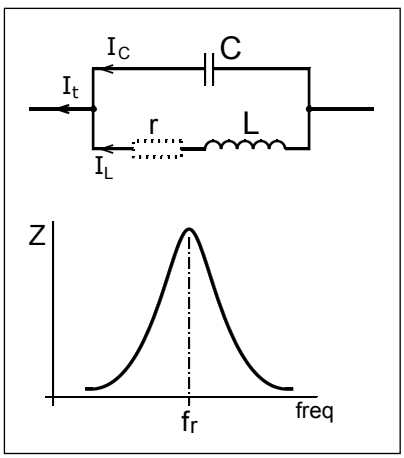


Fig 1.30: Parallel resonance. The resistor 'r' may simply be the resistance of the coil L, but this is the resistance at the resonant frequency. See Skin Effect

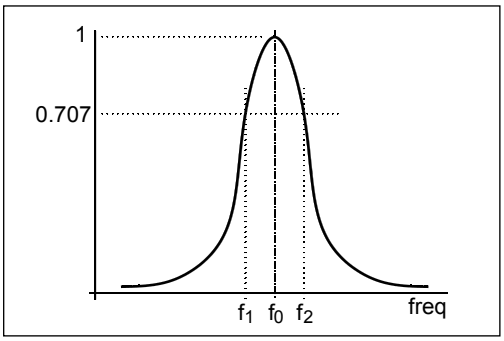


Fig 1.31: Bandwidth of a tuned circuit

that is:

$$Q = \frac{2\pi fL}{R} \quad \text{or} \quad \frac{1}{2\pi fCR}$$

This can also be written as:

$$Q = \frac{\omega L}{R} \quad \text{or} \quad \frac{1}{\omega CR} \qquad \text{where } \omega = 2\pi f.$$

Resonance Curves and Selectivity

The curves in Figs 1.29 and 1.30 showed how tuned circuits were more responsive at their resonant frequency than neighbouring frequencies. These resonance curves show that a tuned circuit can be used to select a particular frequency from a multitude of frequencies that might be present. An obvious example is in selecting the wanted radio signal from the thousands that are transmitted.

The sharpness with which a tuned circuit can select the wanted frequency and reject nearby ones is also determined by the Q factor. **Fig 1.31** shows a resonance curve with the centre frequency and the frequencies each side at which the voltage has fallen to $1/\sqrt{2}$ or 0·707 of its peak value, that is the *half power bandwidth*. The magnification or Q factor is also given by:

$$Q = \frac{\text{resonant frequency}}{\text{bandwidth}}$$

where the resonant frequency is f_0 and the bandwidth $f_2 - f_1$ in Fig 1.31.

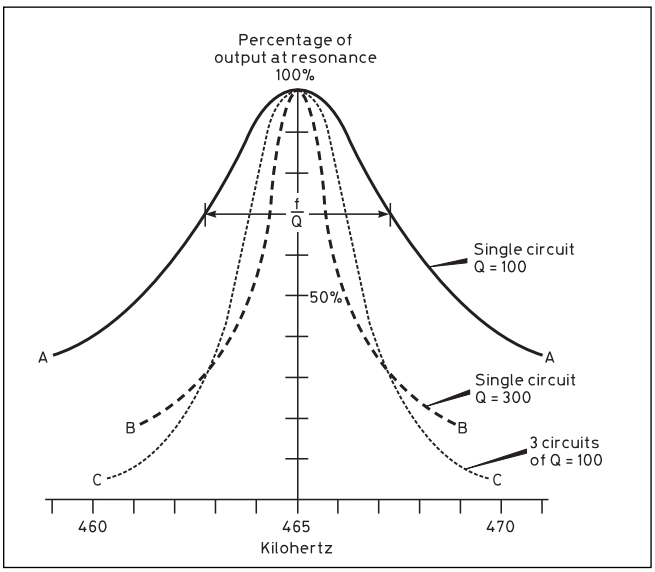


Fig 1.32: Selectivity curves of single and cascaded tuned circuits

Relative bandwidth	Percentage output (voltage)	Loss (dB)
f/3Q	95	0·45
f/2Q	90	0·92
f/Q	70	3
2f/Q	45	6.9
4f/Q	24	12.4
8f/Q	12	18.4

Table 1.4: Selectivity of tuned circuits

Example:

A radio receiver is required to cover the long and medium wave broadcast bands, from 148·5kHz to 1606·5kHz. Each station occupies 6kHz of bandwidth. What Q factors are required at each end of the tuning range.

For long wave the Q factor is 148·5/6 ≈ 25. At the top of the medium wave broadcast band the required Q factor is 1606·6/6 ≈ 268. It is not realistic to achieve a Q factor of 268 with one tuned circuit.

Fig 1.32 shows several resonance curves offering different selectivities. Curve A is for a single tuned circuit with a resonant frequency of 465kHz and a Q factor of 100. The width of the curve will be 4·65kHz at 70·7% of the full height. An unwanted signal 4kHz off-tune at 461kHz will still be present with half its voltage amplitude or one quarter of the power.

The dashed curve, B, has a Q of 300; the higher Q giving greater selectivity. In a radio receiver it is common to have several stages of tuning and amplification, and each stage can have its own tuned circuit, all contributing to the overall selectivity. Curve C shows the effect of three cascaded tuned circuits, each with a Q of 100. In practice this is better than attempting to get a Q of 300 because, as will be seen in later chapters, the usual requirement is for a response curve with a reasonably broad top but steeper sides. For a single tuned circuit the response, as a proportion of the bandwidth, is shown in **Table 1.4**. The values are valid for series circuits and parallel circuits where the Q-factor is greater than 10, as is normally the case.

Dynamic Resistance

Fig 1.33 shows a parallel tuned circuit with the resistance of the coil separately identified. This resistance means that the currents I_L and I_C are not exactly equal and do not fully cancel. The impedance of the circuit at resonance is high, but not infinite. This impedance is purely resistive, that is V and I are in phase, and is known as the dynamic impedance or resistance.

To find its value, the series combination of R and L must be transformed into its parallel equivalent.

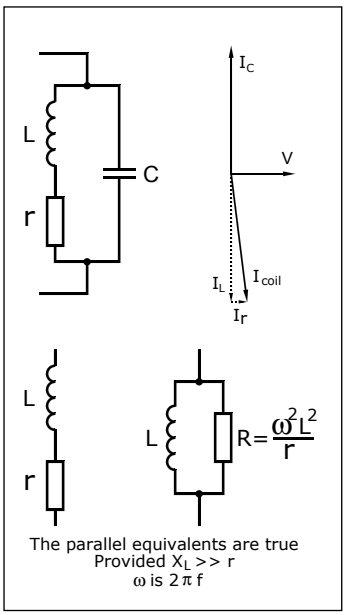


Fig 1.33: A parallel LCR circuit. The phasor diagram shows that the current in the C and LR branches do not cancel exactly due to the phase change caused by r. At resonance the parallel tuned circuit behaves as a high value resistor called R_D, the dynamic resistance

$$R_P = \frac{R_S^2 + X_S^2}{R_S} \quad \text{and} \quad X_P = \frac{R_S^2 + X_S^2}{X_S}$$

The proof of these transformations is outside the scope of this book.

If the resistance is considerably lower than the reactance of L; normal in radio circuits, then this simplifies to

$$R_P = \frac{X_S^2}{R_S} \quad \text{and} \quad X_P = X_S$$

Remembering that $X_S=2\pi fL$ and that, at resonance, $2\pi fL = 1/2\pi fC$, then the dynamic resistance, normally written R_D can be written as

$$R_D = \frac{L}{Cr}$$

where L and C are the coil and capacitor values and r is the series resistance of the coil.

Remembering also that $Q=2\pi fL/r$, allows another substitution in the formula to get:

$Q=2\pi fCR_D$.

The significance of this is to note that a high R_D implies a high Q.

L/C Ratio

From the formula for R_D above, it can be seen that as well as keeping r as low as practicable, the values of L and C will influence R_D. The value of L x C is fixed (from the formula for the resonant frequency) but we can vary the ratio L/C as required with the choice determined by practical considerations.

A high L/C ratio is normal in HF receivers, allowing a high dynamic resistance and a high gain in the amplifier circuits. Stray capacitances limit the minimum value for C. In variable frequency oscillators (VFOs) on the other hand a lower ratio is normal so the capacitor used swamps any changes in the parameters of the active devices, minimising frequency instability.

External components in the rest of the circuit may appear in parallel with the tuned circuit, affecting the actual R_D and reducing overall Q factor. The loading of power output stages has a significant effect and the L/C ratio must then be a compromise between efficiency and harmonic suppression.

Skin Effect

Skin effect is a phenomenon that affects the resistance of a conductor as the frequency rises. The magnetic field round the conductor also exists inside the conductor but the magnetic field at the surface is slightly weaker than at the centre. Consequently the inductance of the centre of the conductor is slightly higher than at the outer surface. The difference is small but at higher frequencies is sufficient that the current flows increasingly close to the surface. Since the centre is not now used, the cross-section of the conductor is effectively reduced and its resistance rises.

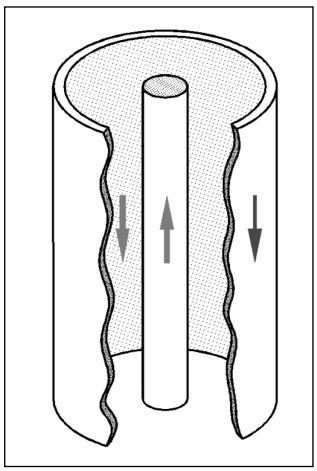


Fig 1.34: Currents in a coaxial cable

The skin depth is the depth at which the current has fallen to 1/e of its value at the surface. (e is the base of natural logarithms). For copper this works out as:

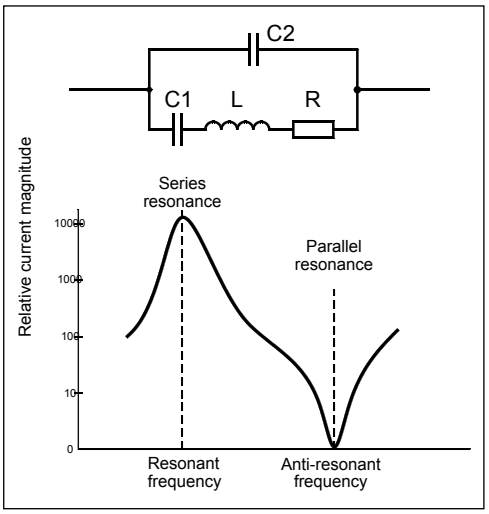


Fig 1.35: Typical variation of current through a quartz crystal with frequency. The resonant and anti-resonant frequencies are normally separated by less than 0·1%; a few hundred hertz for a 7MHz crystal

$$\text{skin depth } \delta = \frac{66 \cdot 2}{\sqrt{f}} \quad \text{mm}$$

At 1MHz this is 0·066 mm. 26SWG wire, which might be used on a moderately fine coil, has a diameter of 0·46mm, (0·23mm radius) so current is flowing in approximately half the copper cross-section and the resistance will be roughly doubled.

At UHF, 430MHz, however, a self-supporting coil of 16SWG coil (wire radius 0·8mm) the skin depth is 0·003mm and only the surface will carry a current, resulting in a considerable increase in resistance. If possible it will be better to use an even larger wire diameter which has a larger surface or use silver plated wire because the resistivity of silver is lower.

A coaxial cable has three conducting surfaces when carrying RF currents. This is shown in **Fig 1.34**. Current flows mainly along the surface of the inner conductor, depicted by the 'up' arrow. An exactly equal and opposite current flows along the inside surface of the outer conductor. The outside surface of the outer conductor is an entirely separate conducting surface and may have no current at all or even an extraneous interfering signal unrelated to the signal inside the coax.

Quartz Crystals

A quartz crystal is a very thin slice of quartz cut from a naturally occurring crystal. Quartz exhibits the piezo-electric effect, which is a mechanical-electrical effect. If the crystal is subjected to a mechanical stress, a voltage is developed between opposite faces. Similarly, if a voltage is applied then the crystal changes shape slightly. When an AC signal is applied to a crystal at the correct frequency, its mechanical resonance produces an electrical resonance. The resonant frequency depends on the size of the crystal slice. The electrical connections are made by depositing a thin film of gold or silver on the two faces and connecting two very thin leads.

Below 1MHz, the crystal is usually in the form of a bar rather than a thin slice. At 20kHz the bar is about 70mm long. Up to 22MHz the crystal can operate on its fundamental mode; above that a harmonic or overtone resonance is used. This is close to a multiple of the fundamental resonance but the term 'overtone' is used since the frequency is close to odd multiples of the fundamental but not exact.

Fig 1.35 shows the equivalent circuit of a crystal and the two resonances possible with L and either C1 or C2. The two frequencies are within about 0·1% of each other. The crystal is supplied and calibrated for one particular resonance, series or parallel and should be used in the designed mode.

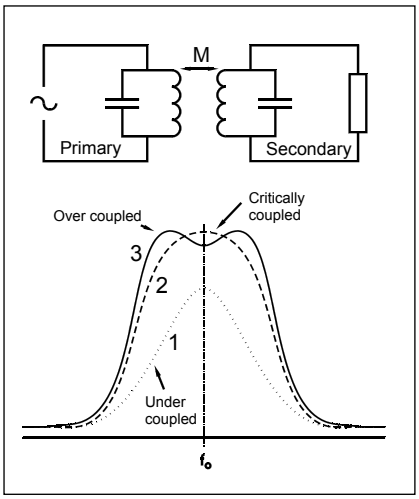


Fig 1.36: Inductively coupled tuned circuits. The mutual inductance M and the coupling increase as the coils are moved together. The response varies according to the degree of coupling with close coupling giving a broad response

The key advantage of using a crystal is that its Q factor is very much higher than can be achieved with a real LC circuit. Qs for crystals are typically around 50,000 but can reach 1,000,000. Care must be taken to ensure circuit resistances do not degrade this. There are more details in the oscillators chapter.

Coupled Circuits

Pairs of coupled tuned circuits are often used in transmitters and receivers. The coupling is by transformer action (see later in this chapter), and the two tuned circuits interact as shown in **Fig 1.36**.

When the coupling is loose, that is the coils are separated, the overall response is as shown in curve 1. As the coils are moved closer, the coupling and the output increase until curve 2 is reached which shows *critical coupling*. Closer coupling results in curve 3 which is *over-coupling*. Critical coupling is the closest coupling before a dip appears in the middle of the response curve. Often some over-coupling is desirable because it causes a broader 'peak' with steep sides which rapidly attenuate signals outside the desired frequency range.

Two tuned circuits are often mounted in a screening can, the coils generally being wound the necessary distance apart on the same former to give the required coupling. The coupling is then said to be fixed.

Some alternative arrangements for coupling tuned circuits are shown in the Building Blocks chapter.

Filters

Filters may be 'passive', that is an array of capacitors and inductors or 'active' where an amplifier and, usually, capacitors and resistors are used. Mostly active filters are used at lower frequencies, typically audio, and passive LC filters at RF. In both cases the aim is to pass some frequencies and block or attenuate others.

Fig 1.37 shows the four basic configurations; low pass, high pass, band pass and band stop or notch; together with the circuit of a typical passive filter, and the circuits for passive filters. The Building Blocks chapter discusses the topic further.

Transformers

Mutual inductance was introduced in Fig 1.17(b); a changing current in one coil inducing an EMF in another. This is the basis of the transformer. In radio frequency transformers the degree of coupling, the mutual inductance, was one of the design features. In power transformers, the windings, the primary and secondary, are tightly coupled by being wound on a bobbin sharing the same laminated iron core to share and maximise the magnetic flux (the name for magnetic 'current'). The size of core used in the transformer depends on the amount of power to be handled.

A transformer can provide DC isolation between the two coils and also vary the current and voltage by varying the relative number of turns. **Fig 1.38** shows a transformer with n_P turns on the primary and n_S turns on the secondary. Since the voltage induced in each coil will depend on the number of turns, we can conclude that:

$$V_S = V_P \times \frac{n_S}{n_P}$$

With a load on the secondary (the output winding) a secondary current. I_S will flow. Recognising that, neglecting losses, the power into the primary must equal the power out of the secondary, we can also conclude that

$$I_P = I_S \times \frac{n_S}{n_P}$$

It is simple to say that if power is drawn from the secondary, the primary current must increase to provide that power, but that does not really provide an explanation. When current flows in the secondary the magnetic field due to the secondary current will weaken the overall field, resulting in a reduced back EMF in the primary. This allows more primary current to flow, restoring the field. The primary current does not fall to zero with no

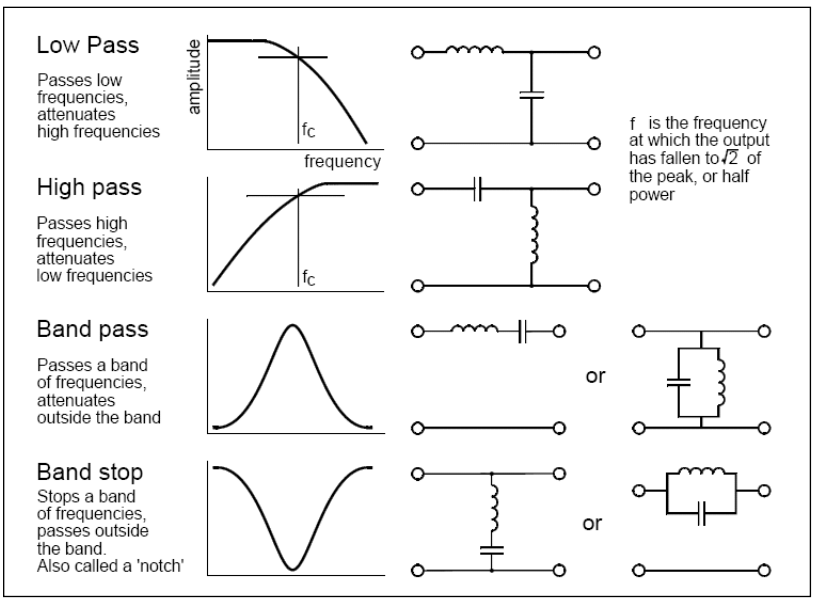


Fig 1.37: The circuits and frequency response of low pass, high pass, band pass and band stop (or notch) filters

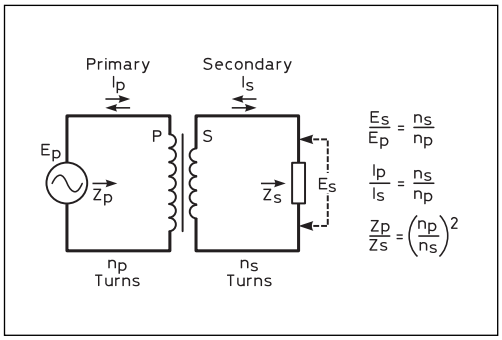


Fig 1.38: The low-frequency transformer

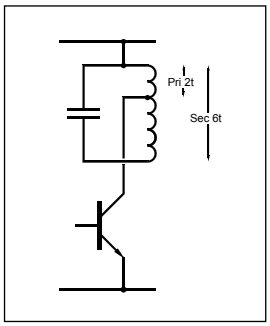


Fig 1.39: An autotransformer in a parallel tuned circuit. Transformer tapping will reduce the loading of the transistor on the tuned circuit R_D

secondary load, the residual current, known as the *magnetising current*, is that which would flow through the inductance of the primary. Normally this current and any small losses in the transformer can be neglected but it should be noted that the windings do rise in temperature during use and an overload could cause breakdown of the internal insulation or start a small fire before the fuse blows.

Since the current and voltages can be changed, it follows that the impedances, given by $Z_P = V_P/I_P$ and $Z_S = V_S/I_S$ will also change. If, for a step-down transformer, the secondary voltage is halved, the secondary current will be double the primary current. The impedance has reduced to a quarter of the primary value.

$$Z_P = Z_S\left(\frac{n_P}{n_S}\right)^2$$

The transformation of impedance is a valuable property and transformers are widely used for impedance matching. Several examples will be seen in the transmitters chapters.

Auto-transformers

If DC isolation between the primary and secondary is not required, an auto-transformer can be used, which has a single winding. It is tapped at an appropriate point so that, for a step-down transformer, the whole winding forms the primary but only a few of those turns form the secondary. An example is a mains transformer with 1000 turns, tapped at 478 turns to give a 110V output from the 230V supply. The advantage of the auto-transformer is partly simplified construction but mainly space and weight saving.

Fig 1.39 shows a step-up transformer in a tuned circuit in the collector of a transistor. The loading effect of the transistor will be considerably reduced, thereby preserving the Q of the unloaded tuned circuit.

Screening

When two circuits are near one another, unwanted coupling may exist between them due to stray capacitance between them, or due to stray magnetic coupling.

Placing an earthed screen of good conductivity between the two circuits, as shown in **Fig 1.40(b)**, can eliminate stray capacitance coupling. There is then only stray capacitance from each circuit to earth and no direct capacitance between them. A useful practical rule is to position screens so that the two circuits are not visible from one another.

Stray magnetic coupling can occur between coils and wires due to the magnetic field of one coil or wire intersecting the other. At radio frequency, coils can be inductively screened (as well as capacitively) by placing them in closed boxes or cans made from material of high conductivity such as copper, brass or aluminium. Eddy currents are induced in the can, setting up a field which opposes and practically cancels the field due to the coil beyond the confines of the can.

If a screening can is too close to a coil the performance of the coil, ie its Q and also its inductance, will be considerably reduced. A useful working rule is to ensure the can in no closer to the coil than its diameter, see **Fig 1.40(c)**.

At low frequencies eddy current screening is not so effective and it may be necessary to enclose the coil or transformer in a box of high-permeability magnetic material such as Mumetal in order to obtain satisfactory magnetic screening. Such measures are not often required but a sensitive component such as a microphone transformer may be enclosed in such a screen in order to make it immune from hum pick-up.

It is sometimes desirable to have pure inductive coupling between two circuits with no stray capacitance coupling. In this case a Faraday screen can be employed between the two coils in question, as shown in **Fig 1.40(d)**. This arrangement is sometimes used between an antenna and a receiver input circuit or between a transmitter tank circuit and an antenna. The Faraday screen is made of stiff wires (**Fig 1.40e**) connected together at one end only, rather like a comb. The 'open' end may be held by non-conductive material. The screen is transparent to magnetic fields because there is no continuous conducting surface in which eddy currents can flow. However, because the screen is connected to earth it acts very effectively as an electrostatic screen, eliminating stray capacitance coupling between the circuits.

SEMICONDUCTORS

Silicon forms the basis of most transistors and diodes. Specialist devices may be formed of the compound gallium arsenide. Early semiconductors were based on germanium and this is now making a come back in microwave transistors with silicon-germanium junctions (Si-Ge), offering similar performance to GaAsFets but at a lower cost. A simplified picture of the silicon atom is shown in **Fig 1.41**. Silicon has an atomic number of 14 indicating it has 14 electrons and 14 protons. The electrons are arranged in shells that must be filled before starting the next

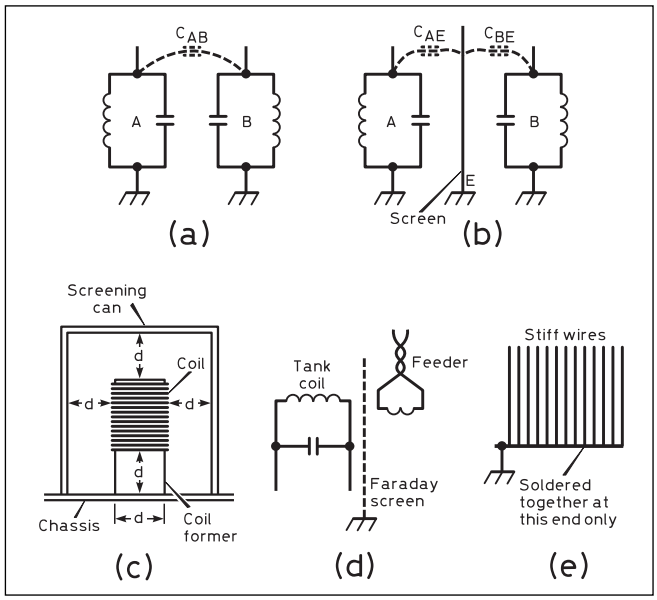
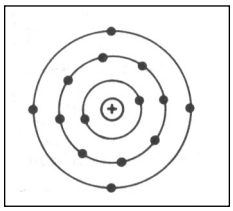


Fig 1.40: (a) Stray capacitance coupling C_{AB} between two circuits A and B. The introduction of an earthed screen E in (b) eliminates direct capacitance coupling, there being now only stray capacitance to earth from each circuit C_{AE} and C_{BE}. A screening can (c) should be of such dimensions that it is nowhere nearer to the coil it contains than a distance equal to the diameter of the coil d. A Faraday screen between two circuits (d) allows magnetic coupling between them but eliminates stray capacitance coupling. The Faraday screen is made of wires as shown at (e)

Fig 1.41: The silicon atom



outermost layer. It has four electrons in its outer shell, which are available for chemical bonding with neighbouring atoms, forming a crystal lattice. A single near perfect crystal can be grown from a molten pool of pure material. As a pure crystal, the four outer electrons are all committed to bonding and none are available to support the flow of current. Pure crystaline silicon is an insulator. To obtain the semiconductor, carefully controlled impurities are added. This is known as doping.

If a small quantity of an element with five outer electrons, for example phosphorus, antimony or arsenic, is added at around 1 part in 10^7, then some of the atoms in the lattice will now appear to have an extra electron. Four will form bonds but the fifth is unattached. Although it belongs to its parent nucleus, it is relatively free to move about and support an electric current. Since this material has mobile electrons it is called an n-type semiconductor.

If an element with three outer electrons is added to the crystal, eg indium, boron or aluminium, these three electrons form bonds but a hole is left in the fourth place. An electron that is part of a nearby bond may move to complete the vacant bond. The new hole being filled, in turn, by another nearby electron. It is easier to consider the hole as being the mobile entity than to visualise several separate electrons each making single jumps. This material is a p-type semiconductor.

A *diode* is formed from a p-n junction and will only allow current to flow in one direction. This is used in rectification, changing AC into unidirectional half-cycles which can then be smoothed to conventional DC using a large capacitor.

A transistor uses a three layer semiconductor device formed either as n-p-n or, perhaps less common, as p-n-p sandwich. The transistor may act as an amplifier, increasing the amplitude or power of weak signals, or it can act as a switch in a control circuit. These uses, and more detailed discussion of the operation of transistors and diodes, are contained in this book.

BALANCED AND UNBALANCED CIRCUITS

Fig 1.42 shows a balanced and unbalanced wire or line. The balanced line (top) has equal and opposite signals on the two conductors, the voltages and currents are of equal magnitude but opposite in direction. For AC, this would mean a 180 degree phase difference between the two signals. The input and output from the line is taken between the two conductors. If each conductor had a 2V AC signal, the potential difference between the two conductors would be 4V.

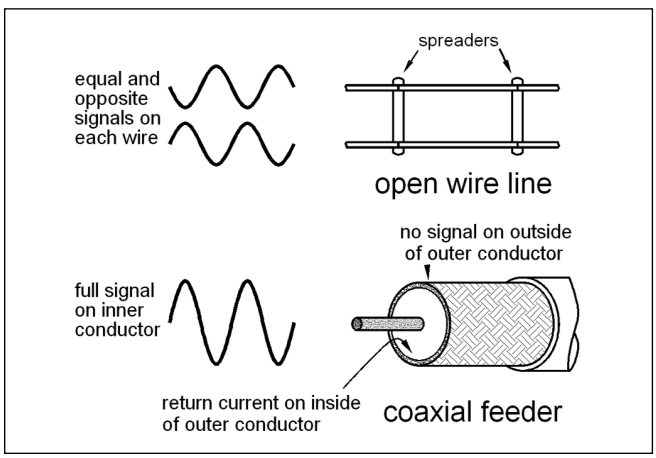


Fig 1.42: Signals on balanced and unbalanced lines

The unbalanced line may have a single conductor with an 'earth return', typically the chassis of an equipment or a real earth in the case of telegraph communication (see below) or may be a coaxial cable as shown in Fig 1.42. The centre conductor carries the signal and the return current path is the inside surface of the braid of the coaxial cable. With a true unbalanced termination at the load end, the braid will be at zero volts and the full magnitude of the signal is on the centre conductor. The outer surface of the braid will also be at zero volts, being earthed at one or both ends. In line telegraph use it was common for there to be a single wire, often alongside railway tracks and the earth literally was used as the return path to complete the circuit.

Any wire carrying a changing current has a tendency to radiate part of the signal as electromagnetic waves or energy. This is the principle of the antenna (aerial). The efficiency with which radiation occurs is a function of the length of the wire as a proportion of the wavelength of the signal. This is covered later in the chapters on HF and VHF/UHF antennas. A radiator can equally pick up any stray electromagnetic signals which will then get added to the wanted signal in the wire.

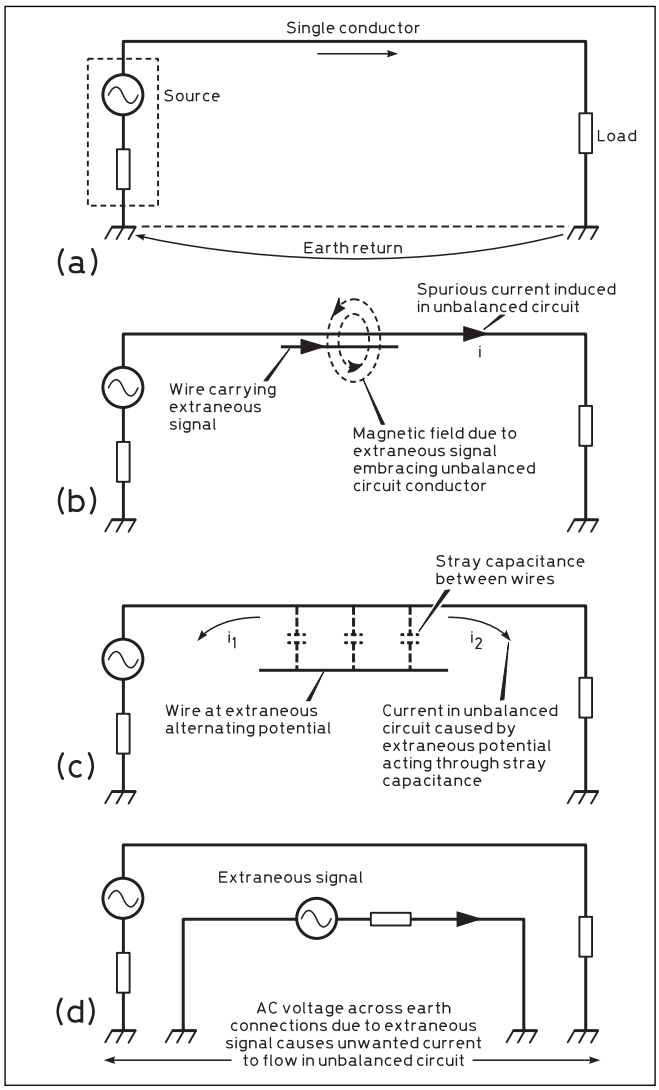


Fig 1.43: The unbalanced circuit. (a) The basic unbalanced circuit showing earth return path. (b) How extraneous signals and noise can be induced in an unbalanced circuit by magnetic induction. (c) Showing how extraneous signals can be induced by stray capacitance coupling. (d) Showing how extraneous signals can be induced due to a common earth return path

Unbalanced circuits are particularly prone to this effect. **Fig 1.43** shows an unbalanced wire with magnetic (mutual inductance) coupling and capacitive coupling to another nearby conductor. Circuits of this type are very commonly used in radio equipment and are perfectly satisfactory provided leads are kept short and are spaced well away from other leads. It is, however, prone to the pick-up of extraneous noise and signals from neighbouring circuits by three means: inductive pick-up, capacitive pick-up and through a common earth return path.

Inductive pick-up, **Fig 1.43(b)**, can take place due to transformer action between the unbalanced circuit wire and another nearby wire carrying an alternating current; a common example is hum pick-up in audio circuits due to the AC mains wiring.

Capacitive pick-up, **Fig 1.43(c)**, takes place through the stray capacitance between the unbalanced circuit lead and a neighbouring wire. Such pick-up can usually be eliminated by introducing an earthed metal screen around the connecting wire.

If the unbalanced circuit has an earth return path that is *common* to another circuit, **Fig 1.43(d)**, unwanted signals or noise may be injected by small voltages appearing between the two earth return points of the unbalanced circuit. Interference of this type can be minimised by using a low-resistance chassis and avoiding common earth paths as far as possible.

A balanced circuit is shown in **Fig 1.44**. As many signal sources, and often loads as well, are inherently unbalanced (ie one side is earthed) it is usual to use transformers to connect a source of signal to a remote load via a balanced circuit. In the balanced circuit, separate wires are used to conduct current to and back from the load; no current passes through a chassis or earth return path.

The circuit is said to be 'balanced' because the impedances from each of the pair of connecting wires to earth are equal. It is usual to use twisted wire between the two transformers as shown in Fig 1.44**.** For a high degree of balance, and therefore immunity to extraneous noise and signals, transformers with an earthed screen between primary and secondary windings are used. In some cases the centre taps of the balanced sides of the transformers are earthed as shown dotted in Fig 1.44.

The balanced circuit overcomes the three disadvantages of the unbalanced circuit. Inductive and capacitive pick-up are eliminated since equal and opposite currents are induced in each of the two wires of the balanced circuit and these cancel out. The same applies to interfering currents in the common earth connection in the case where the centre taps of the windings are earthed.

The argument also applies to radiation from a balanced circuit. The two conductors will tend to radiate equal and opposite signals which will cancel out. Some care is required however because if conducting objects are close to the feeder, comparable in distance to the separation of the two conductors, then the layout may not be symmetrical and some radiation due to the imbalance may occur.

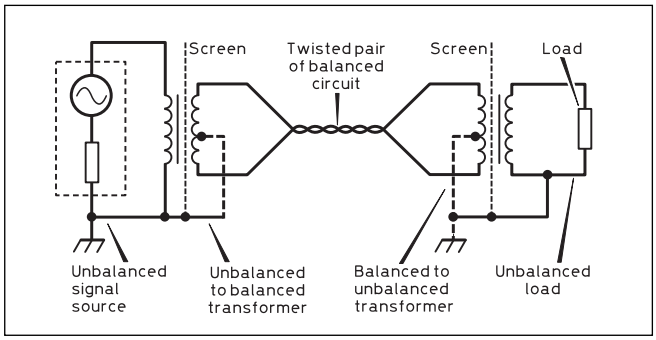


Fig 1.44: The balanced circuit

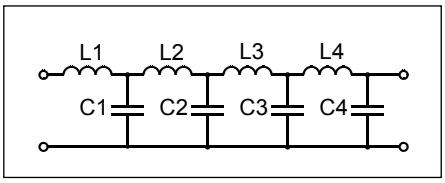


Fig 1.45: A long feeder can be represented by series inductors and parallel capacitors

Looking again at Fig 1.42, the balanced feeder must be symmetrically run, away from walls and other conductors, but the unbalanced feeder has an earthed conducting sleeve, typically an outer braid or metal tape to screen the centre conductor and minimise radiation outside the cable. This is known as coaxial cable. It may be run close to walls and conductors but the manufacturers advice on minimum bending radius should be heeded.

FEEDERS

The feeder is the length of cable from the transmitter/receiver to the antenna. It must not radiate, using the properties of balance or screening outlined above and it must be as loss free as possible. Consider the circuit shown in **Fig 1.45.** Each short length of feeder is represented by its inductance, and the capacitance between the conductors is also shown. Assume now a battery is connected to the input. The first length of cable will charge up to the battery voltage but the rate of rise will be limited by the inductance of L1 and the need for C1 to charge. This will occur progressively down the cable, drawing some current from the battery all the while. Clearly with a short cable, this will take no time at all. However an infinitely long cable will be drawing current from the battery for quite some time. The ratio of the battery voltage to the current drawn will depend on the values of L and C and is the characteristic impedance of the cable.

It is given by:

$$Z_0 = \sqrt{\frac{L}{C}}$$

where L and C are the values per unit length of the feeder.

This argument and the formula assume the series resistance of the feeder is negligible, as is the leakage resistance of the insulation. This is acceptable from the point of view of Z_0 but the series resistance is responsible, in part, for feeder losses; some of the power is lost as heat. The other facet of the loss is the dielectric loss in the insulation, that is the heating effect of the RF on the plastic, easiest viewed as the heating of the plastic in a microwave oven. The loss is frequency dependant and, to a simple approximation, rises as the square root of the frequency. A very long length of feeder will look like a resistance of Z_0. So will a shorter length terminated in an actual resistor of value Z_0. Signals travelling down the feeder will be totally absorbed in a load of value Z_0 (as if it were yet more feeder) but if the feeder is terminated in a value other than Z_0, some of the energy will be absorbed and some will be reflected back to the source. This has a number of effects which are discussed in the chapters on antennas. Properly terminating the feeder in its characteristic impedance (actually a pure resistance) is termed correct matching. It is also, of course, necessary to use a balanced load on a balanced line (twin feeder) and an unbalanced load on an unbalanced line such as coaxial cable. Much more on feeders can be found in the chapter on transmission lines.

THE ELECTROMAGNETIC SPECTRUM

Radio frequencies are regarded by the International Telecommunication Union to comprise frequencies from 9kHz to 400GHz but the upper limit rises occasionally due to advances in technology. This, however is only a small part of the electromagnetic spectrum shown in **Fig 1.46**.

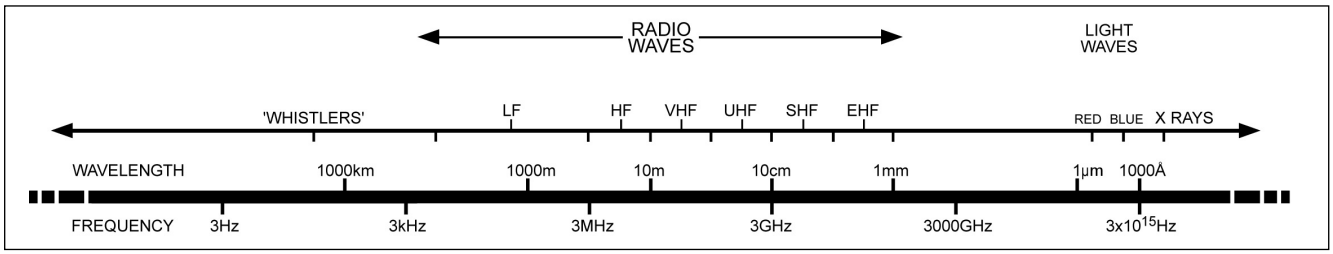


Fig 1.46: Graphical representation of the electromagnetic spectrum

The whole spectrum comprises radio waves, heat, light, ultraviolet, gamma rays, and X-rays. The various forms of electromagnetic radiation are all in the form of oscillatory waves, and differ from each other only in frequency and wavelength. They all travel through space with the same speed, approximately 3 x 10^8 metres per second. This is equivalent to about 186,000 miles per second or once round the world in about one-seventh of a second.

As might be surmised from the name, an electromagnetic wave consists of an oscillating electric field and a magnetic field. It can exist in a vacuum and does not need a medium in which to travel. **Fig 1.47** shows one and a half cycles of an e-m wave. The electric (E) and magnetic (H) fields are at right angles and both are at right angles to the direction of propagation. E and H are in phase and their magnitudes are related by the formula:

$$\frac{E}{H} = \sqrt{\frac{\mu_0}{\varepsilon_0}} \quad = 120\pi\,\Omega$$

where μ_0 is the permeability of free space and ε_0 is the permitivity of free space, both natural physical constants. The quantity E/H is known as the impedance of free space and may be likened to the characteristic impedance of a feeder.

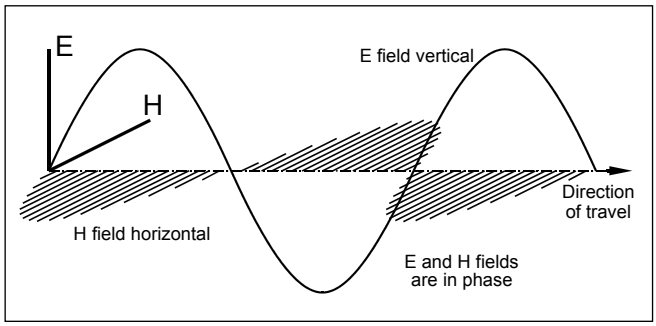


Fig 1.47: An electromagnetic wave. By convention the polarisation is taken from the electric field, so this is a vertically polarised wave

Frequency and Wavelength

The distance travelled by a wave in the time taken to complete one cycle of oscillation is called the *wavelength*. It follows that wavelength, frequency and velocity of propagation are related by the formula

Velocity = Frequency x Wavelength, or c = fλ

where c is the velocity of propagation, f is the frequency (Hz), and λ is the wavelength in metres.

Example:

What are the frequencies corresponding to wavelengths of (i) 150m, (ii) 2m and (iii) 75cm?

From the formula c = fλ, the frequencies are given by:

150m

$$f = \frac{c}{\lambda} = \frac{3\times10^8}{150} = \frac{300\times10^6}{150} = 2\cdot0\text{MHz}$$

2m:

$$f = \frac{c}{\lambda} = \frac{3\times10^8}{2} = \frac{300\times10^6}{2} = 150\text{MHz}$$

75cm:

$$f = \frac{c}{\lambda} = \frac{3\times10^8}{0\cdot75} = \frac{300\times10^6}{0\cdot75} = 400\text{MHz}$$

It is important to remember to work in metres and hertz in these formula. However, it will also be noticed that if the frequency is expressed in MHz throughout and λ in metres, a simplified formula is:

$$f = \frac{300}{\lambda}\,\text{MHz}$$

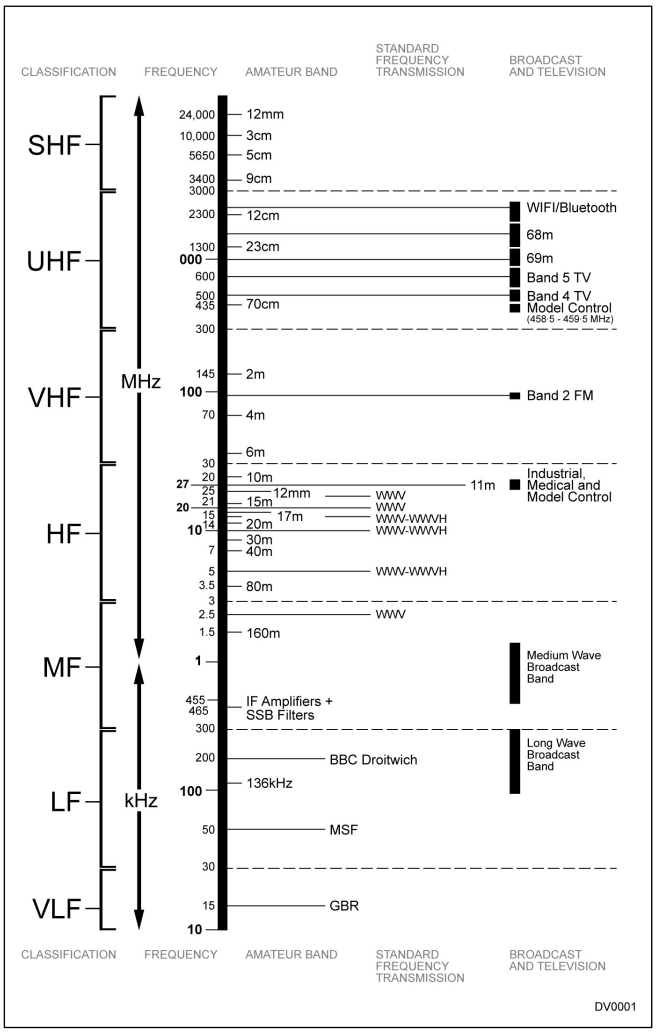


Fig 1.48: The amateur frequency bands in relation to other user services

or

$$\lambda = \frac{300}{f} \text{ metres}$$

Fig 1.48 shows how the radio spectrum may be divided up into various bands of frequencies, the properties of each making them suitable for specific purposes. Amateur transmission is permitted on certain frequency bands in the LF, MF, HF, VHF and UHF, SHF/microwave ranges.

ANTENNAS

An antenna (or aerial) is used to launch electromagnetic waves into space or conversely to pick up energy from such a wave travelling through space. Any wire carrying an alternating current will radiate electromagnetic waves and conversely an electromagnetic wave will induce a voltage in a length of wire. The issue in antenna design is to radiate as much transmitter power as possible in the required direction or, in the case of a receiver, to pick up as strong a signal as possible, very often in the presence of local interference.

Isotropic Radiator

An isotropic radiator is one that radiates equally in all, three-dimensional, directions. In practical terms, it does not exist but it is easy to define as a concept and can be used as a reference.

The power flux density *p* from such an antenna, at a distance *r* is given by:

$$p = \frac{P}{4\pi r^2} \quad W/m^2$$

where P is the power fed to the antenna.

This leads to the electric field strength E, recalling the formula for the impedance of free space above:

$$p = \frac{E^2}{R} = \frac{E^2}{120\pi} = \frac{P}{4\pi r^2}$$

rearranging gives

$$E = \frac{\sqrt{30P}}{r} = \frac{5{\cdot}5\sqrt{P}}{r}$$

Note that this formula is for an isotropic radiator.

Radiation Resistance

The radiation resistance of the antenna can be regarded as:

$$R_r = \frac{\text{total power radiated}}{(\text{RMS current at antenna input})^2}$$

This will vary from one antenna type to another. Electrically short antennas, much less than a wavelength, have very low radiation resistances leading to considerable inefficiency if this approaches the resistance of the various conductors. This resistance is not a physical resistor but the antenna absorbs power from the feeder as if it were a resistor of that value.

Antenna Gain

Antennas achieve gain by focussing the radiated energy in a particular direction. This leads to the concept of Effective Radiated Power (ERP) which is the product of the power fed to the antenna multiplied by the antenna gain. The figure can be regarded as equivalent to the power required to be fed to an antenna without gain to produce the same field strength (in the wanted direction) as the actual antenna. It should be noted that this gain also applies on receive because antennas are passive, reciprocal devices that work identically on transmit or receive.

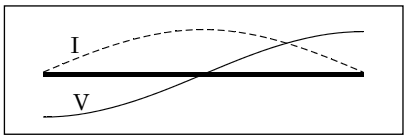


Fig 1.49: Current and voltage distribution on a half-wave dipole

Effective Aperture

Consider a signal of field strength E arriving at an antenna and inducing an EMF. This EMF will have a source resistance of R_r, the radiation resistance and will transfer maximum power to a load of the same value. This power can be viewed as the area required to capture sufficient power from the incident radio wave.

An antenna of greater gain will have a larger effective aperture; the relationship is:

$$A_{eff} = \frac{\lambda^2}{4\pi} \times G$$

where λ is the wavelength of the signal and G is the antenna gain in linear units.

This aperture will be much larger that the physical appearance of a high gain antenna and is an indication of the space required to allow the antenna to operate correctly without loss of gain.

A Dipole

The current and voltage distribution on a dipole are shown in Fig 1.49. The current is a half sine wave. By integrating the field set up by each element of current over the length of the dipole, it can be shown that the dipole has a gain of 1·64 or 2·16dB and a radiation resistance (or feed impedance) of 73 ohms (see below for a discussion of dB, decibels). The dipole is often used as the practical reference antenna for antenna gain measurements. It must then be remembered that the gain of an antenna quoted with reference to a dipole is 1·64 times or 2·16dB less than if the gain is quoted with reference to isotropic.

Effective Radiated Power

The effective radiated power (ERP) of an antenna is simply the product of the actual power to the antenna and the gain of the antenna. Again it is necessary to know if this is referenced to isotropic (EIRP) or a dipole (EDRP). ERP is normally quoted with reference to a dipole. This figure indicates the power that would need to be fed to an actual dipole in order to produce the field strength, in the intended direction, that the gain antenna produces.

The total radiated power is still that supplied to the antenna, the enhancement in the intended direction is only as a result of focussing the power in that direction. A side benefit, but an important one, is that the power in other directions is much reduced.

The field from such an antenna is given by:

$$E = \frac{7{\cdot}01\sqrt{ERP}}{d} \text{ V/m}$$

where d is the distance from the antenna (shown as r above for the isotropic radiator).

The different coefficient of 7·01 rather than 5·5 accounts for the fact that the ERP is reference to a dipole, which already has some gain over isotropic.

Feeding an Antenna

The concept of maximum power transfer applies to an antenna, but there is a additional complication if the antenna is not a resonant length. Fig 1.49 showed the current and voltage distribution on the dipole. The dipole is a half wavelength long and the

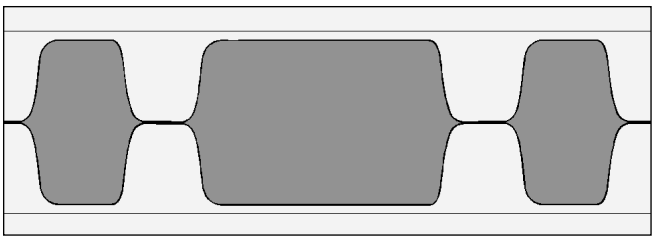


Fig 1.50: The Morse letter R

current can be viewed as millions of electrons 'sloshing' from one end to the other rather like water in a bath.

As a half wavelength (or a multiple) the antenna is resonant and can be regarded as an LC resonant circuit with a resistor, the radiation resistance R_r, to account for the power radiated. Below the resonant frequency the antenna will appear net capacitive (too short in length) and net inductive above the resonant frequency (too long in length).

This means that the antenna will no longer appear as a good match to the feeder and some power will be reflected back down the feeder. This, in turn, will result in the feeder being a poor match to the transmitter resulting in less than maximum power being transferred. To recover this situation, the opposite type of reactance must be inserted to offset that presented to the transmitter. The antenna/feeder/inserted reactance now represents a pure resistance to the transmitter. This function is usually performed by an antenna tuning (or matching) unit and is covered more fully in the chapter on HF antennas.

MODULATION

A simple sine wave radio signal conveys no information except that it is there and perhaps the direction of its origin. In order to convey information we must change the signal in some agreed way, a process called modulation.

The simplest form of modulation is to turn the signal on and off and this was all that early pioneers of radio could achieve. This led to the adoption of the Morse code (originally invented for line telegraphy) which remains very effective at sending text messages and is still practiced by many amateurs although Morse has given way to other techniques in commercial applications. A Morse signal is shown in **Fig 1.50**. It will be noticed that the edges of the carrier waveform are rounded off; the reason will be explained shortly in the section below. It is called CW or carrier wave because the carrier (the sine wave signal) is all that is transmitted. However we now know that is not quite correct.

Amplitude Modulation (AM)

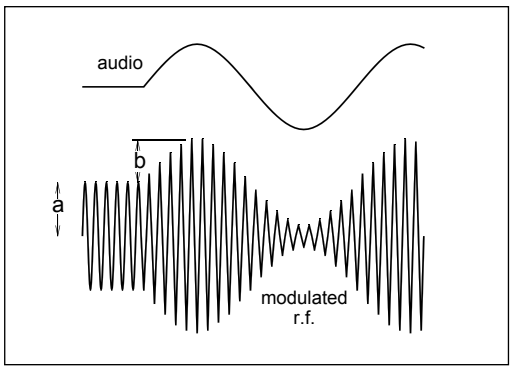
In amplitude modulation, the amplitude of the carrier is varied according to the instantaneous amplitude of the audio modulating signal as shown in **Fig 1.51**. Only the carrier is transmitted in the absence of modulation, the audio (or other) signal increasing and decreasing the transmitted amplitude. The depth of modulation, m, is given by the quantity B/A or B/A x 100%.

It is desirable to achieve a reasonably high level of modulation but the depth must never exceed 100%. To do so results in distortion of the audio signal, the production of audio harmonics and a transmitted signal that is wider than intended and likely to splatter over adjacent frequencies causing interference.

Sidebands

The process of modulation leads to the production of sidebands. These are additional transmitted frequencies each side of the carrier, separated from the carrier by the audio frequency. This is first shown mathematically and then the result used to demonstrate the effect in a more descriptive manner.

Fig 1.51: Amplitude modulation



A single carrier has the form:

$$A \sin \omega_c t$$

where A is the amplitude and ω_c is $2\pi f_c$, the frequency of the signal. When amplitude modulated the formula becomes:

$$A(1 + m \sin \omega_a t) \sin \omega_c t$$

where m is the depth of modulation and ω_a the audio frequency used to modulate the carrier ω_c.

Expanding this gives:

$$A \sin \omega_c t + Am \sin \omega_a t \times \sin \omega_c t$$

The second term is of the form sinA sinB which can be represented as ½(sin(A-B)+sin(A+B)), so our signal becomes:

$$A \sin \omega_c t + \frac{Am}{2}(\sin(\omega_c t - \omega_a t) + \sin(\omega_c t + \omega_a t))$$

Inspection will show that these three terms are the original carrier plus two new signals, one at a frequency $f_c - f_a$ and the other at $f_c + f_a$ and both at an amplitude dependant on the depth of modulation but with a maximum amplitude of half the carrier. This is shown in **Fig 1.52**.

It still might not be all that clear just why the sidebands need to exist. Consider the effect of just the carrier and one of the sidebands but as if they were both high pitch audio notes close together in frequency. A beat note would be heard. Add the other sideband, giving the same beat note. The overall effect is the carrier varying in amplitude. The bandwidth needed for the transmission depends on the audio or modulating frequency. The two sidebands are equidistant from the carrier so the total bandwidth is twice the highest audio frequency.

Modulating the carrier also increases the transmitted power. The formula above gave the amplitude (voltage) of the signal. If the power in the carrier is P_c then the total transmitted power is:

$$P_c + 2 \times P_C \left(\frac{m}{2}\right)^2 = P_c \left(1 + \frac{m^2}{2}\right)$$

For 100% modulation, the total transmitted power is 150% of the carrier power, that is the carrier plus 25% in each sideband.

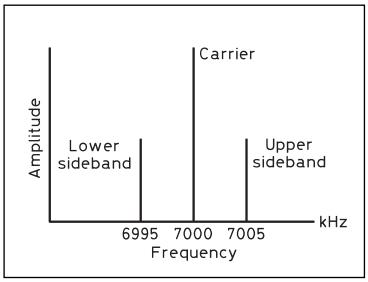


Fig 1.52: Sideband spectrum of a 7MHz carrier which is amplitude modulated by a 5kHz tone

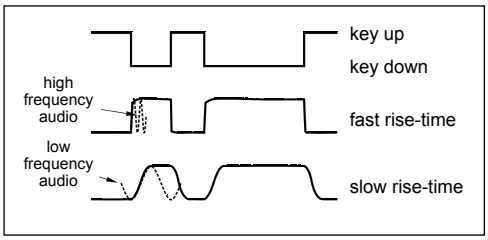


Fig 1.53: Fast and slow rise times on a CW keying signal

It can now be shown why the keying of the Morse signal was rounded. Morse can be regarded as amplitude modulation with only two values, 0 or 100% modulation. Simplistically, there would be sidebands separated from the carrier dependant on the keying speed. That is not the whole story however.

The rise and fall of the transmitted envelope greatly affect the overall bandwidth. **Fig 1.53** shows the Morse keying with fast and slow rises and falls of the keying and thus the transmitted RF envelope. The fast rise can be regarded as a small part of a high frequency audio signal; a slower rise time being part of a lower frequency audio signal. The sidebands at the moment of keying will be those applicable to the 'audio'. The bandwidth will be much wider than that assumed from the Morse speed alone. A well designed keying circuit is inaudible outside the tuning range over which the Morse tone can be heard. Bad keying causes 'key clicks' that can be heard several kHz each side of the Morse signal proper. It should be avoided.

Single Sideband (SSB)

Fig 1.54 shows an AM signal where the audio is speech containing a range of frequencies from just above zero to 3kHz. The upper sideband is a copy of the audio, albeit shifted up in frequency to just above the carrier, and the lower sideband is a mirror image just below the carrier. In fact all the audio information is contained in the sidebands; as stated at the very beginning, the carrier contains no information. We can, therefore, dispense with the carrier and either one of the sidebands. In the figure the lower sideband is shown as filtered out and the carrier has been suppressed.

The modulation and detection techniques needed to do this are shown in the chapters on receivers and transmitters. SSB transmission requires only half the bandwidth and represents a considerable saving in power, as evidenced by the formula for the power in the various components of the AM signal above.

It will be recalled that the frequency of the audio signal was retained in the transmitted signal by the frequency separation between the carrier and the sideband. If the carrier and the other sideband are removed the reference point for determining the audio frequency is lost. It is 're-inserted' at the receiver by the act of tuning in the signal. The receiver must be accurately tuned or all the received frequencies will be offset by the same amount, namely the error in the tuning. This is why the pitch of the audio of an SSB signal varies as the receiver is tuned.

The bandwidth of an SSB signal is simply the audio bandwidth, half that of the AM transmission. The power varies continuously from zero, when unmodulated, to its peak value depending on the audio volume and the designed power of the transmitter.

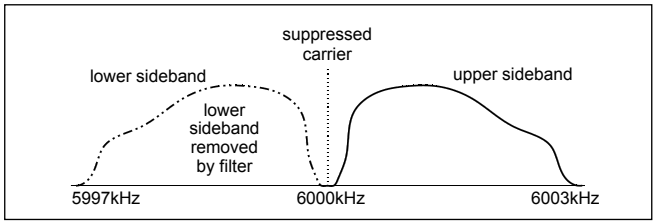
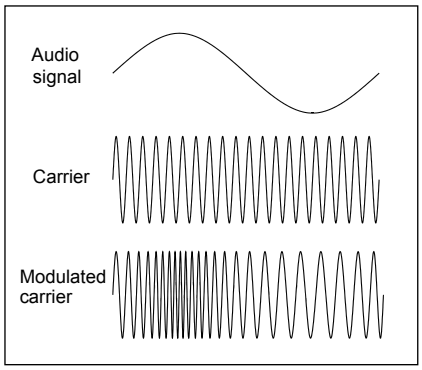


Fig 1.54: SSB signal from a balanced modulator, with one sideband filtered out

Fig 1.55: Frequency modulation



Frequency Modulation (FM)

In FM, the instantaneous frequency of the carrier is varied according to the instantaneous amplitude of the carrier as shown in **Fig 1.55**. Unlike AM there is no natural limit to the amount by which the frequency may be deviated from the nominal centre frequency. The peak deviation is a system design parameter chosen as a compromise between audio quality and occupied bandwidth. The ability of an FM detector or discriminator to combat received noise increases as the cube of the deviation. However, the total received noise increases linearly with the bandwidth so the overall effect is a square law. For instance, doubling the deviation and bandwidth improves the recovered audio signal to noise ratio by a factor of four (6dB).

For FM the deviation ratio, corresponding to the depth of modulation of AM, is defined as:

$$\text{Deviation ratio} = \frac{\text{actual deviation}}{\text{peak deviation}}$$

The modulation index relates the peak deviation to the maximum audio frequency:

$$\text{Modulation index} = \frac{\text{peak deviation}}{\text{audio frequency}}$$

A narrow band FM system is one where the modulation index is about or less than unity. Amateur FM systems are narrow band (NBFM). At 2m, with 12·5kHz channels, the max audio frequency is around 2·8kHz and the peak deviation 2·5kHz; a mod index of 0·9. At 70cm, with 25kHz-spaced channels, the calculation is 5/3·5, a mod index of 1·4.

VHF broadcast stations have an audio range up to 15kHz and a peak deviation of 75kHz, giving a mod index of 5. This is a wide band FM system (WBFM). TV sound has a deviation of 50kHz but good quality TV sound systems now use the NICAM digital sound channel which also offers stereo reproduction. Simplistically, the mod index will give a guide to the audio quality to be expected.

The bandwidth of an FM transmission is considerably wider than twice the peak deviation, as might be expected. The modulation gives rise to sidebands but even for a single audio modulating tone, there may be several sidebands. The mathematical derivation is much more complex and the 'sin' function requires Bessel functions to calculate. The formula is:

$$A\sin(\omega_c t + m\sin\omega_a t)$$

where m is the modulation index.

This gives a transmitted spectrum shown in **Fig 1.56** which has, theoretically, an infinite number of sidebands of decreasing amplitude. The rule of thumb often applied, known as Carson's rule, is that the necessary bandwidth is given by:

Bandwidth = 2(peak deviation + max audio frequency)

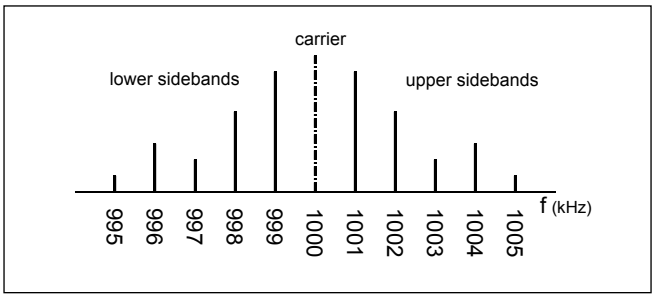


Fig 1.56: The spectrum of an FM signal

For a typical 2m system this works out as:

Bandwidth = 2(2·5 + 2·8) = 10·6kHz

Carson's rule tends to over-estimate the required bandwidth at the lower modulation indices used by 12·5kHz-channel systems. In practice, a receiver filter with a bandwidth around 7·5kHz will prove adequate, and also filter out more of the ever-present RF noise. The number of relevant sidebands depends on the modulation index, according to the graph in **Fig 1.57**. This shows the carrier and the first three sidebands. For those with access to *Microsoft Excel®*, the function BesselJ, available in the maths and statistics pack, will give the values. This pack is provided on the CD but frequently not loaded. The help menu gives guidance on how to load the pack if it is not loaded.

It is interesting to note that the amplitude of the carrier also varies with the modulation index, and goes to zero at an index of 2·4. This provides a very powerful method to set the deviation. Set the transmitter up with an audio signal generator at full amplitude at 1·042kHz (2·5/2·4). Listen to the carrier on a CW narrow bandwidth receiver and adjust the deviation on the transmitter for minimum carrier; this can also be done on a spectrum analyser if one is available. The deviation is now correctly set provided signals of greater amplitude are not fed into the microphone socket. Ideally the transmitter will have a microphone gain control, followed by a limiter and then followed by a deviation control. Thus excessive amplitude signals will not reach the deviation control and the modulator.

Phase Modulation (PM)

This is a subtle variant of frequency modulation. In FM it is the instantaneous frequency of the transmitter that is determined by the instantaneous modulating input voltage. Clearly, as the frequency varies, the phase will differ from what it would have been had the carrier been unmodulated. There are phase changes but these are as a consequence of the frequency modulation. In phase modulation, it is the phase that is directly controlled by the input audio voltage and the frequency varies as a consequence.

Frequency modulation is usually achieved by a variable capacitance diode in the oscillator circuit. This has the side effect of introducing another source of frequency drift or instability. An RC circuit situated after the oscillator can achieve phase modulation by using a variable capacitance diode for C. This way the oscillator is untouched and more stable. Many transceivers do in fact use phase modulation and can pre-condition the audio so the transmitted signal appears frequency modulated. However, there is an advantage in retaining phase modulation because the higher audio frequencies are transmitted with a higher deviation than with FM, helping to combat any noise present in the receiver. Many data transmissions use phase modulation, the digital signal directly changing the carrier phase.

Digital Modulation

The techniques of digital modulation are little different, except the modulating input will normally comprise a data signal having one of two discrete values, normally denoted '0' and '1' but in practice two different voltages. In CMOS digital logic this may be 0V and 10V, or for RS232 signals from a computer ±6V.

Amplitude shift keying, that is two different amplitudes, either of the carrier or of an audio modulating signal is uncommon, partly because nearly all interference mechanisms are amplitude related and better immunity can be achieved by other methods.

Frequency shift keying is popular, the carrier having two frequencies separated by 170Hz for narrow shift or 450 and 750Hz wide shift typically used commercially. The occupied bandwidth is again greater than the frequency shift since each of the two 'carriers' will have sidebands depending on the keying speed and method. The mode can be generated either by direct keying of the modulator or feeding audio tones (1275 and 1455Hz for narrow shift amateur use) into an SSB transmitter.

Phase shift modulation is used in many amateur modes. These modes are discussed in more detail in the data communications chapter.

Two phases, 0° and 180° may be used. This is binary phase shift keying, BPSK. Another method, known as QPSK, quadrature phase shift keying has 4 states or phases; 0°, 90°, 180° and 270°. This allows two binary digits (each with two states 0

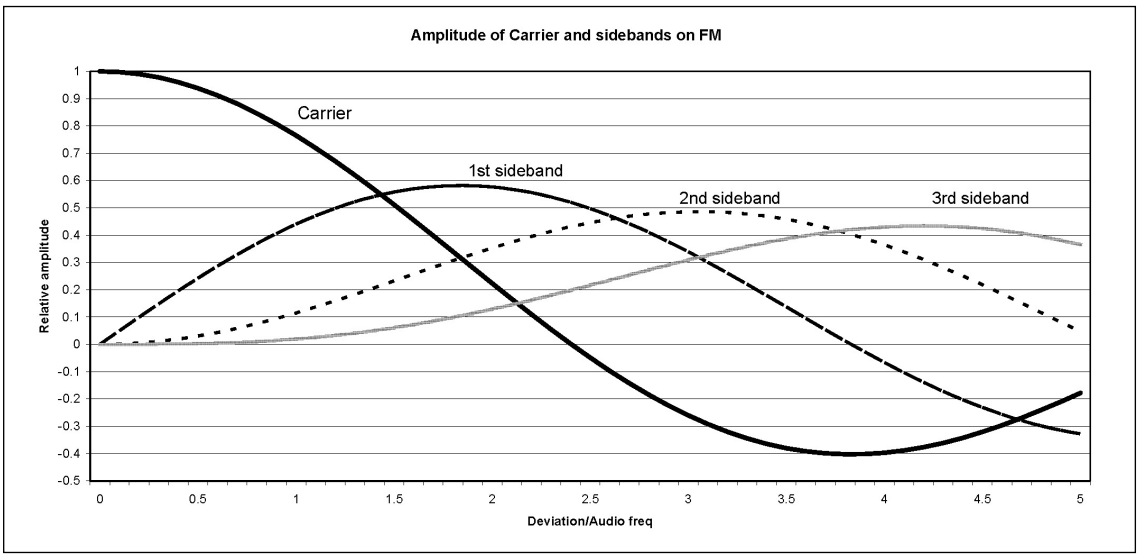


Fig 1.57: The amplitude of the carrier and the sidebands are given by Bessel functions

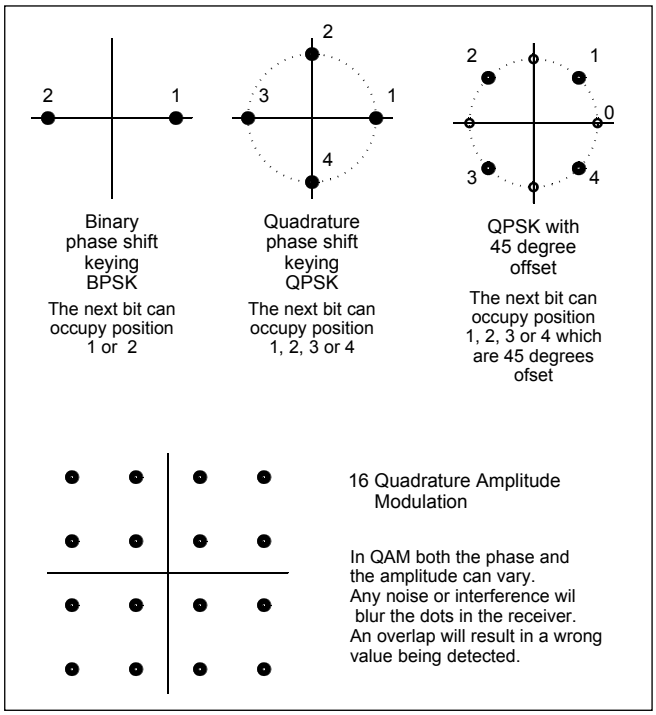


Fig 1.58: Digital modulation schemes often vary the phase of the signal to represent the next transmitted symbol. 16 symbols can transmit four binary bits at once, but use amplitude as well as phase changes

and 1, giving four in total) to be combined into a single character or symbol to be sent. By this means the symbol rate (the number of different symbols per second) is lower, half in this case, than the baud rate, defined as the number of binary bits per second.

Each symbol may be advanced in phase by 45° in addition to the phase change required by the data so that, irrespective of the data, there is always some phase change. This is needed to ensure the receiver keeps time with the transmitter. If this did not happen and the data happened to be a string of 20 zeros with no phase changes, there is a risk that the receiver would lose timing and output either 19, 20 or 21 zeros. It is then out of synchronisation with the transmitter and the output data signal may well be meaningless.

Higher level modulation schemes can combine even more binary digits into a symbol, allowing even higher data rates. Quadrature Amplitude Modulation (QAM) is used professionally but not often in amateur circles. This utilises both amplitude and phase changes to encode the data. Whilst higher data rates are achievable, the difference between each symbol gets progressively less and less as each symbol is used to contain more and more binary digits. **Fig 1.58** shows these systems diagrammatically. In isolation this is all very well but the effects of noise and interference get progressively worse as the complexity of the system increases.

Amateurs have tended to stick to relatively straightforward modulation schemes but have utilised then in novel and clever ways to combat interference or send signals with very low bandwidth, finding a slot in an otherwise crowded band. The data communications chapter has the details.

Shannon limit

We saw above and in the QAM diagram in Fig 1.58 that a single link can use a more sophisticated multi-level modulation scheme to get more binary bits per symbol and thus a faster data (baud) rate for a given symbol rate. That is a higher bits/second/Hz. The downside is that the symbols are closer together, in amplitude, phase or both, so that the link is more susceptible to noise.

In professional planning the aim is to get many links in a given geographical area, maybe all converging on a single node, and the risk then is interference from other links, not natural noise. Higher capacity links need less noise and need to be more widely separated, leading to a trade off.

Amateur practice has been to optimise in a different direction, minimising the bandwidth required, eg PSK31, to fit in between other transmissions, or performing well at levels below the noise floor, eg WSJT.

However that poses the question: how far can we go? Is there a theoretical maximum and how close to it are we? If we are within one decibel of the limit then trying to design for a 3dB improvement is doomed to failure.

We can try to combat bit errors by the use of error correcting codes. Very good codes exist but sending the code bits requires either a higher capacity channel or sacrifices user bit rates. Another trade off.

The limit is given by the Shannon formula:-

$$C = B \log_2\left(1+\frac{S}{N}\right)$$

Where C is the channel capacity in bits/sec (that is binary bits), B is the channel bandwidth in Hz, S is the received signal power in bandwidth B and N is the noise power in the same bandwidth. S/N is the signal to noise ratio in linear units. The signal and noise powers must be expressed as a power, not a voltage.

This formula assumes the noise has a Gaussian distribution, that is it is truly random noise as would be experienced from a hot resistor. If the ambient noise has a different statistical distribution, due typically to a semi-distant transmission or impulsive interference of man-made origin, then the S/N term will require slight modification.

Note that the log term is to the base 2, not the normal base 10. The $\log_{10}$ of 1000 is 3 because $10^3 = 1000$. Similarly $\log_2$ of 16 is 4 because $2^4 = 16$. Finding logs to the base 2 is simply achieved by remembering the definition of a logarithm.

The log of a number, x, is the power to which the base of the log must be raised to equal x.

$$\text{If } \; 2^y = x \qquad \text{then } \; \log_2(x) = y$$

Take logs of both sides ($\log_{10}$)

$$y \log_{10}(2) = \log_{10}(x)$$

$$\text{so } \; y = \frac{\log_{10}(x)}{\log_{10}(2)} \; = \; \frac{\log_{10}(x)}{0\cdot 301}$$

Simply find the log of x (where x=1+S/N) and divide it by 0.3.

Example:

A radio channel has a 3 kHz bandwidth and a signal to noise ratio of 20dB or 100:1

$$C = B \log_2\left(1+\frac{S}{N}\right) = 3000 \log_2\left(1+\frac{100}{1}\right) = 3000 \log_2(101)$$

$$C = 3000 \times 6\cdot 66 \approx 20\, kbit/\sec$$

The formula gives no indication how this may be achieved, that is for the designer; but the designer now knows the maximum bit rate in that bandwidth with that amount of noise (which includes receiver noise) which cannot be exceeded.

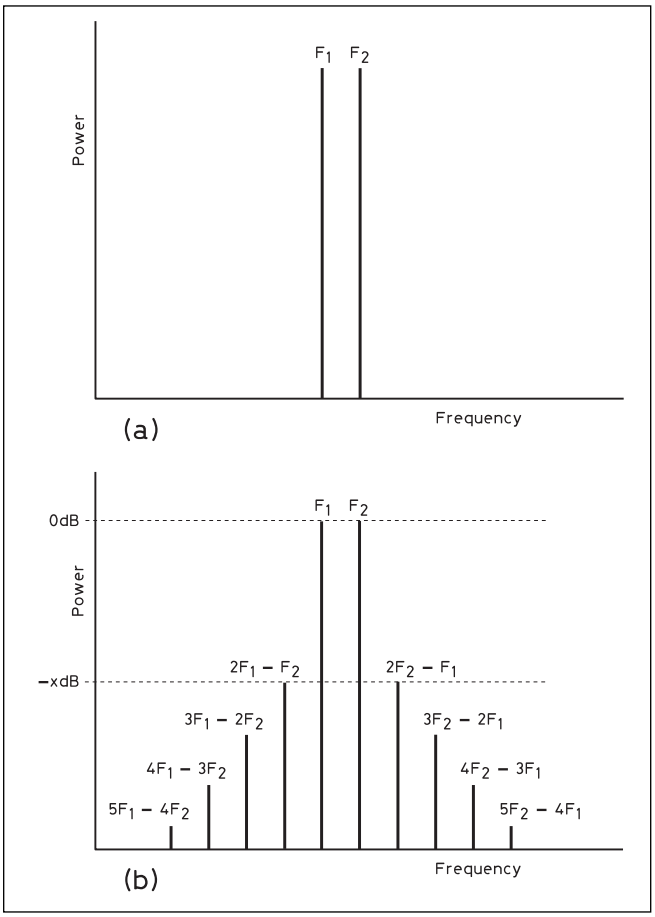


Fig 1.59: Spectrum of two frequencies (a) before and (b) after passing through a non-linear device

Intermodulation

Mixing and modulation often use non-linear devices to perform the required function. Non-linear means that the output is not just a larger (or smaller) but otherwise identical copy of the input. Consider the equation y=4x. The output (y) is simply 4 times the input (x). Plotting the function on graph paper would give a straight line - linear.

Now consider $y=x^2$. Plotting this would give a curve, not a straight line; it is non-linear.

If the input was a sine wave, for example $\sin \omega_c t$, then the output would be of the form:

$$\sin^2 \omega_c t \quad \text{which is} \quad \tfrac{1}{2}(1 - \cos 2\omega_c t)$$

The output will contain a DC component and a sine wave of twice the frequency. This is a perfect frequency doubler.

Similarly if two signals are applied, $\sin \omega_a$ and ω_b, the output will be:

$$(\sin \omega_a + \sin \omega_b)^2$$

which is

$$\sin^2 \omega_a + 2 \sin \omega_a \sin \omega_b + \sin^2 \omega_b$$

We already know the two $\sin^2$ terms are frequency doubling and we also know (from the amplitude modulation section) that sinA x sinB gives us sin (A+B) and sin (A-B), that is the sum and difference frequencies. We now have a mixer.

Unfortunately no device is quite that simple, and amplifiers are not perfectly linear. The errors are usually small enough to neglect but when badly designed, or overdriven, the non-linearities will produce both harmonics of the input frequencies and various sums and differences of those frequencies.

For two sine wave inputs A and B, the output contains frequencies m x A ± n x B where m and n are any integers from 0 upwards. These are called intermodulation products (IMPs). The frequencies 2A, 3A; 2B, 3B etc are far removed from the original frequencies and are usually easily filtered out. However, the frequencies m x A - n x B where m and n differ by 1, are close to the original frequencies.

This can happen in an SSB power amplifier where two (or more) audio tones are mixed up to RF and then amplified. The unwanted intermodulation products may well be inside the pass-band of the amplifier or close to it such that they are not well attenuated by the filter. This, then represents distortion of the signal, or, worse, interference to adjacent channels. **Fig 1.59** illustrates this.

The value of m+n is known as the 'order' of the intermodulation product. It will be noticed that it is only the odd orders that result in frequencies close to the wanted signals. It is also the case that the lower order products (ie 3rd) will have a greater amplitude than higher order products. The power in these relative to the fundamentals, xdB in **Fig 1.59(b)**, should be at least 30dB down and preferably more.

Intermodulation can also occur in other systems. In receivers it can occur in the 'front-end' if this is operating at a non-linear level (due for example by being overloaded). For instance, in a crowded band such as 7MHz with high-powered broadcast signals close to (and even in!) the amateur band, intermodulation can occur between them. It should be noted that these signals may be outside the bandwidth of the later stages of the radio receiver but inside the band covered by the RF front end. The IMPs occur in the overloaded front end (normally in the mixer as this is a non-linear stage), producing extra signals which may well be close to, or on top of, the wanted signal.

This degrades the receiver performance if the products are within the pass-band of the receiver. The use of high-gain RF amplifiers in receivers can be a major source of this problem and may be countered by reducing the gain by using an input attenuator.

Alternatively, only just enough gain should be used to overcome the noise introduced by the mixer. For frequencies up to 14MHz, it is not necessary to have any pre-mixer gain provided the mixer is a good low-noise design (see the chapter on HF receivers).

Intermodulation can also occur in unlikely places such as badly constructed or corroded joints in metalwork. This is the so-called rusty-bolt effect and can result in intermodulation products which are widely separated from the original transmitter frequency and which cause interference with neighbours' equipment (see the chapter on electromagnetic compatibility).

ANALOGUE TO DIGITAL CONVERSION

Frequently it is desirable to handle an analogue signal by digital means. Digital signals have discrete values so random noise, an ever present problem, must be sufficiently strong to exceed half the difference between the discrete values before there is risk of error. Furthermore, with digital systems, it is possible to have parity bits and other more complex error detection and correction systems. In many cases these will allow the RF signal to be successfully recovered even though it is well below the noise floor. Later chapters in this book discuss such systems.

Functions, such as filtering and demodulation can be performed as mathematical processes, often with better results than achievable by analogue circuitry. The wanted signal often

had identifiable characteristics, whereas the noise (AWGN, Additive White Gaussian Noise) appears as random fluctuations. That allows the noise to be substantially suppressed, improving the perceived signal to noise ratio. Similarly, multipath reception can be either allowed for or its effects minimised. All this has the umbrella term DSP, Digital Signal Processing.

However, before any of this can be carried out the analogue signal must be digitised by sampling it and representing the samples as digital values.

Digital Sampling

Two decisions must be made when sampling an analogue signal. How accurately must we represent the amplitude of the signal, and how often must we do it?

Quantisation distortion

Analogue signals are continuously variable and may adopt any level from zero to a set maximum, and may be either positive or negative. A binary word of 8 bits, eg 10010011, can have 2^8 different values, including zero. That is 256 different levels. The nearest of those 256 levels is chosen to represent the analogue signal.

There will be a small error, known as the quantisation error, as a result of digitising the signal. The reason for the error is that there are only 256 different levels to choose from, rather than the infinite number needed to give an exact fit to the analogue signal. These errors can be regarded as a real signal plus quantisation 'noise', giving rise to the concept of a signal to quantisation noise ratio.

Fig 1.60 shows part of an analogue signal and the errors in selecting the best fit digital representation. The vertical samples are restricted to one of eight levels each side of zero. A four bit binary word is being used, giving 16 values with 0000 representing the -8 level, 1000 representing the signal 0 point, and 1111 representing the +7 level.

To reduce the error, more binary bits are needed. 16 bits will give 65536 levels to choose from. That is much more accurate, but at the expense of needing more bits to represent the signal. That will translate into more DSP processing power required.

Assuming the quantisation noise has an amplitude uniformly distributed between plus or minus one half of the distance between the least significant bits, then the signal to noise ratio is given by

$$\mathrm{SNR} = 20\mathrm{Log}(2^{Q}) \quad \mathrm{dB}$$

where Q is the number of binary bits *(For explanation of dB see later in this chapter).*

For sine wave signals and assuming the full range of the converter is used, this formula can be simplified to

$$\mathrm{SNR} = 1 \cdot 76 + 6\mathrm{Q} \quad \mathrm{dB}$$

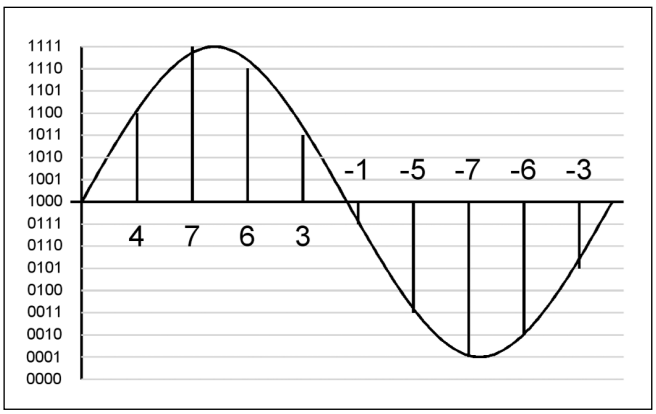


Fig 1.60: Digitising the amplitude results in quantisation errors

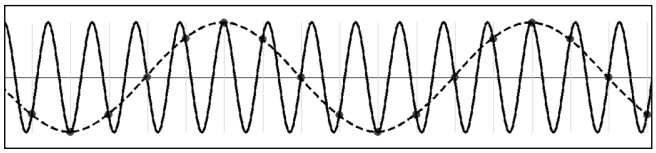


Fig 1.61: Under-sampling the signal

Sampling rate

Clearly we need to sample sufficiently often to capture all the details in the signal; but how fast is that?

Fig 1.61 shows a sine wave, the solid line, sampled 17 times in 15 cycles, that is just over once per cycle. The sample points are shown as dots. The sampled points suggest the signal was the dashed waveform. The sampling is too infrequent to capture all the detail in the signal.

Fig 1.62 shows a signal being sampled very frequently. The exact shape of the signal is captured well. This results in a lot more data than needed and extra work for the digital processing system. Fewer samples would capture this signal perfectly well enough to reproduce it properly.

Finding the best sampling rate

Fig 1.63 shows 10ms of time divided up into a sample taken every half millisecond, 21 samples if we include time zero.

The solid waveform is a sine wave signal at 900Hz; there are nine cycles in 10ms. The 0.5ms sampling points are shown by the black dots. Also shown, dashed, is an 1100Hz signal. It will be noticed that the results of the sampling of that signal are identical to the 900Hz one. The sampling is failing to distinguish between a 900Hz signal and an 1100Hz signal. Which one is correct? It can be shown that it is necessary to sample at least as fast as twice the highest frequency signal present. The proof is quite complex and the reader is referred to texts on the Nyquist Sampling Theorem for further information [6, 7, 8].

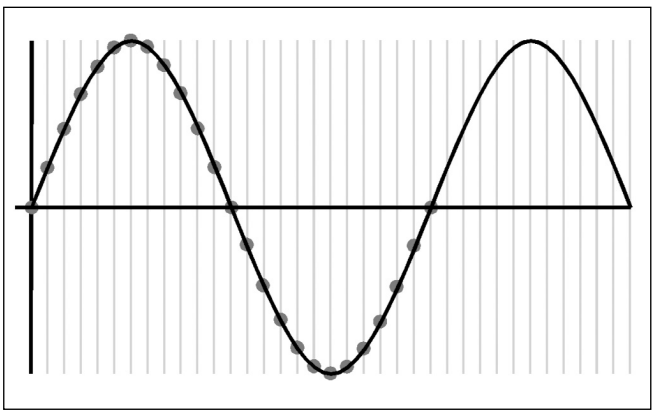


Fig 1.62: Over-sampling the signal results in a lot of data

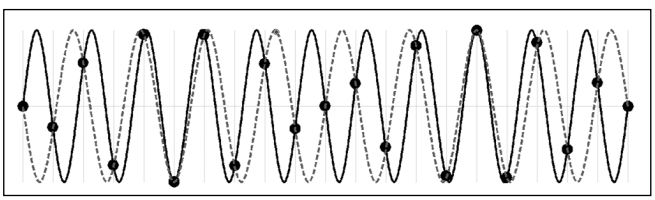


Fig 1.63: Sampling at the optimum rate with an unwanted higher frequency present

Aliasing and anti-aliasing filter

In the example above, it would be necessary to have a good filter in front of the analogue to digital conversion process, to remove all frequencies above 1000Hz. If a small trace of an 1100Hz signal were allowed through the input filter, it would be processed as if it were a 900Hz signal. It is quite likely that it would interfere with the wanted signals. The 1100Hz, unwanted, signal is called an *alias* and the effect of these higher frequency signals appearing as lower frequency ones inside the wanted frequency range, is called *aliasing*. Since the input filter cannot be perfect, in practice it is necessary to sample rather more often than twice the highest frequency present.

Digital to Analogue Conversion

Having been processed, the digital signal must be converted back into analogue form. The output of a digital to analogue convertor (DAC or D2A) will be a series of steps both in amplitude and time. This is shown in **Fig 1.64**.

The output of a DAC is shown as a solid black line. It is based on the samples of the dotted sine wave. At each conversion point the DAC changes its output to match the present value of that sine wave.

The DAC output is squared off into one of the 31 discrete levels with level 16 (binary 1000) being taken as the zero level allowing the signal to have both positive and negative values. Negative values are signified by binary numbers below 1000, and positive values by binary numbers above 1000.

The dashed sine wave shows the sinewave representing the average value of the DAC output.

Two errors are noticed: Firstly, the output is delayed with respect to the digital representation. This is in addition to any delays occasioned by any mathematical processing in the DSP. Secondly, the square waveform is not a good representation of the smooth sinewave.

The timing error is of no consequence unless the original signal can be heard. On devices used to replay an earlier recording, this is irrelevant. Where the processing introduces extended delays, some interruption in smooth voice communication can be observed.

The fine black line shows the difference between the averaged sinewave and the actual output. The output can be regarded as the wanted sinewave and an additional signal represented by the fine black line. That is noise introduced by the digital to analogue conversion process.

Inspection of the noise waveform shows it is of much higher frequency. Consequently it can be removed by filtering. Passing the DAC output through a filter will remove all the higher frequency components and leave just the dashed waveform (further delayed by around half a sample period). The filter has smoothed out the rough edges of the DAC output.

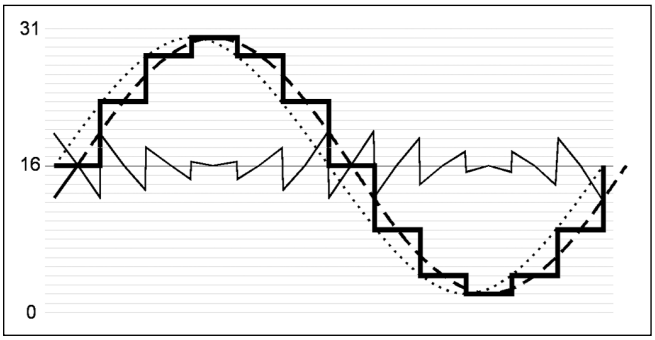


Fig 1.64: The DAC output of a sinewave

In practice a 5-bit DAC is too coarse for practical use. A minimum of 8 bits is required and music systems are likely to utilise 14 or 16 bit DACs.

These devices are much more common than might be realised. CD and MP3 players all store the material, usually music, in digital form, which requires conversion to analogue. Mobile phones are now fully digital and require both A to D and D to A conversion.

DECIBELS

Consider the numbers:

100 = 10^2
1000 = 10^3
10000 = 10^4

The logarithm of a number is defined as the power to which the base of the logarithm (in this case 10) must be raised to equal the number. So the logarithm (log for short) of 1000 is 3, because 10^3 is 1000. This is written as Log_{10} (1000) = 3. However it is usually understood that base 10 is used so the expression is often simplified to Log1000 = 3.

It should also be noted that 10^3 (1,000) x 10^4 (10,000) is 10,000,000 or 10^7. In multiplying the two numbers the powers have been added.

It is possible to consider the Logs of numbers other than powers of 10. For example, the Log of 2 is 0·3010 and the Log of 50 is 1·6990. If these two Logs are added, we get 0·3010+1·6990 =2. We know the Log of 100 is 2, so the anti-Log of 2 is 100. So adding the Log of 2 and the Log of 50 gives us an answer equal to the Log of 100.

Adding the logs of two numbers and then finding the anti-log of the result gives us an answer equal to the product of the two original numbers.

Today this seems an incredibly awkward way of working out that 2 x 50=100. However, computers and calculators are relatively recent devices and Logs were a simple way of multiplying rather more cumbersome numbers.

The aim of the above is really to set out a principle. Multiplying and dividing can be achieved by adding and subtracting the Logs of the numbers concerned.

Consider the route from a microphone to the distant loudspeaker. A microphone amplifier provides some gain, the modulator may provide a gain or a loss, the RF power amplifier will provide gain, the feeder to the antenna will cause a loss and the antenna may well have some gain.

The radio path will have losses associated with distance, trees and building clutter, hills and weather conditions.

The receive antenna will provide gain, the feeder a loss and then the many stages in the receiver will have their own gains and losses.

These gains and losses will not be simple round numbers and they all need multiplying or dividing.

It is normal in radio and line transmission engineering to express all these gains and losses in terms of units proportional to the logarithm of the gain (positive) or loss (negative). In this way gains and losses in amplifiers, networks and transmission paths can be added and subtracted instead of having to be multiplied together.

The unit is known as the decibel or dB. It is defined as:

$$\text{Gain (or loss)} = 10\text{Log}_{10}\left(\frac{P_{out}}{P_{in}}\right)$$

It is always a power ratio, not a current or voltage ratio. If current or voltage ratios were used, a voltage step-up transformer

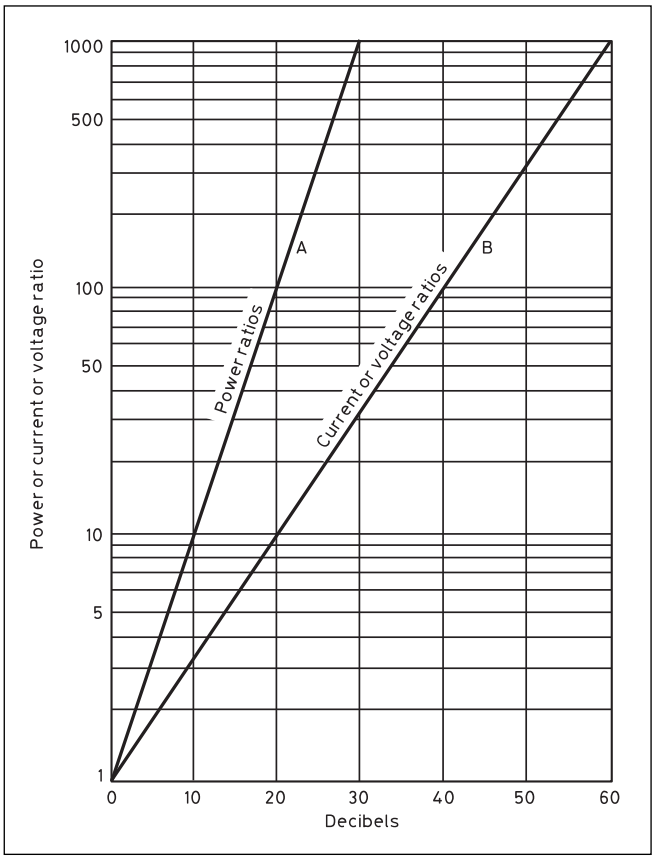


Fig 1.65: Relating decibels to power or current or voltage ratio

would appear to have a gain, which is not the case. The power out of a transformer is never greater than the power in.

However, provided and only provided, the currents or voltages are being measured in the same resistance (so a voltage increase does signify a power increase) then it is possible to use the definition:

$$\text{Gain (or loss)} = 20\text{Log}_{10}\left(\frac{V_{out}}{V_{in}}\right) \quad \text{or} \quad 20\text{Log}_{10}\frac{I_{out}}{I_{in}}$$

The change to 20x follows from the formula $P=V^2/R$ or I^2R. If a number is squared (ie multiplied by itself) then the Log will be doubled (added to itself).

Line A in **Fig 1.65** can be used for determining the number of decibels corresponding to a given power ratio and vice versa. Similarly line B, shows the relationship between decibels and voltage or current ratios.

multiplier	construction	dB construction	dB value
2	-	-	3
4	2 x 2	3+3	6
8	2 x 2 x 2	3+3+3	9
16	2 x 2 x 2 x 2	3+3+3+3	12
10	-	-	10
100	10 x 10	10+10	20
20	2 x 10	3+10	13
40	2 x 2 x 10	3+3+10	16
5	10 ÷ 2	10-3	7
50	10 x 10 ÷ 2	10+10-3	17
25	10 x 10 ÷ 2 ÷ 2	10+10-3-3	14
1·26	-	-	1
1·58	-	-	2

Table 1.5: Construction of dB values

The use of Log tables (or the Log function on a calculator) can often be avoided by remembering two values and understanding how decibels apply in practice.

It was stated above that Log2 = 0·3010, so a doubling of power is 3·01dB, equally a halving is -3·01dB. The ·01 is usually omitted so doubling is simply assumed to be 3dB.

Quadrupling is double double and, like Logs, decibels are added, giving 6dB. The other one to remember is 10 times is 10dB. **Table 1.5** can be constructed from just those two basic values.

It can also be remembered that, with the same resistance (or impedance), the ratios in column 1 applied to voltage or current will double the dB value.

REFERENCES AND FURTHER READING

[1] *Tables of Physical and Chemical Constants*, Kaye and Laby, 14th edition, Longman, London, 1972, pp102-105
[2] *Electronic and Radio Engineering*, Terman, 4th edn, McGraw Hill
[3] *Advance! The Full Licence Manual*, Alan Betts, 1st edn, RSGB
[4] *The RSGB Guide to EMC,* Robin Page-Jones, 2nd edn, RSGB
[5] *Four Figure mathematical tables*, Castle, Macmillan
[6] *Data Communications and Networks*, R L Brewster, pp 25 and 44
[7] *Telecommunications*, J Brown and E V D Glazier, pp 47
[8] http://en.wikipedia.org/wiki/Nyquist%E2%80%93Shannon_sampling_theorem

About the Author

Alan Betts *B.Sc. C.Eng MIEE MIfL, was first licensed as G8TKV in 1978 having been active in electronics construction since 1966. He started as a tutor of radio amateurs in 1980 and has seen several hundred students obtain their licences.He still conducts amateur radio licence training courses and is currently the RCF Chief Examiner. Alan is one of the presenters on the STELAR teacher's course, which trains teachers to obtain an amateur radio licence and introduces amateur radio into schools. He was the IEE representative to the City & Guilds Advisory Group on amateur radio and a professional radio engineer at the Radiocommunications Agency. Alan has also written* Foundation Licence – Now! *and* Advance – The Full Licence Manual*, both books published by the Radio Society of Great Britain (RSGB). He has also contributed to a number of other books published by the RSGB and written for the Society's journal* RadCom*.*