# FOUNDATIONS OF WIRELESS

M. G. SCROGGIE

B.Sc., M.I.E.E.

LONDON · ILIFFE & SONS LTD

| | |
|---|---|
| *First Edition (by A. L. M. Sowerby, M.Sc.)* | 1936 |
| *Second Edition* | 1938 |
| *Third Edition (Revised and enlarged by M. G. Scroggie, B.Sc., M.I.E.E.)* | 1941 |
| *Fourth Edition* | 1944 |
| *Fifth Edition (Entirely rewritten)* | 1951 |
| *Sixth Edition* | 1957 |
| *Seventh Edition (Revised and enlarged)* | 1958 |


*Published for "Wireless World" by Iliffe & Sons, Ltd., Dorset House, Stamford Street, London, S.E.1*
*Made and printed in Great Britain at The Chapel River Press, Andover, Hants*
(*Bks.* 3028)

# Contents

CONTENTS

# Preface

WIRELESS—OR RADIO—has branched out and developed so tremendously that very many books would be needed to describe it all in detail. Besides broadcasting sound and vision, it is used for communication with and between ships, aircraft, trains and cars; for direction-finding and radar (radiolocation), photograph and "facsimile" transmission, telegraph and telephone links, meteorological probing of the upper atmosphere, astronomy, and other things. Very similar techniques are being applied on an increasing scale to industrial control ("automation") and scientific and financial computation. All of these are based on the same foundational principles. The purpose of this book is to start at the beginning and lay these foundations, on which more detailed knowledge can then be built.

At the beginning. . . . If you had to tell somebody about a cricket match you had seen, your description would depend very much on whether or not your hearer was familiar with the jargon of the game. If he wasn't and you assumed he was, he would be puzzled. If it was the other way about, he would be irritated. There is the same dilemma with wireless. It takes much less time to explain it if the reader is familiar with methods of expression, such as symbols, that are taken for granted in scientific discussion but not in ordinary conversation. This book assumes hardly any special knowledge. But if the use of graphs and symbols had to be completely excluded, or else accompanied everywhere by digressions explaining them, it would be very boring for the initiated. So the methods of technical expression are explained separately in a preliminary Initiation. Those already initiated can of course skip it; but it might be as well just to make sure, because this is a most essential foundation.

Then there are the technical terms. They are explained one by one as they occur and their first occurrence is distinguished by printing in italics; but in case any are forgotten they can be looked up at the end of the book; and so can the symbols and abbreviations. These references are there to be used whenever the meaning of anything is not understood.

Most readers find purely abstract principles very dull; it is more stimulating to have in mind some application of those principles. As it would be confusing to have all the applications of radio in mind at once, broadcasting of sound is mentioned most often because it directly affects the largest number of people. But the same principles apply more or less to all the other things. The reader who is interested in the communicating of something other than sound has only to substitute the appropriate word.

*March*, 1958 M. G. SCROGGIE

# Initiation into the Shorthand of Wireless

GLANCING THROUGH THIS BOOK, you can see numerous strange signs and symbols. Most of the diagrams consist of little else, while the occasional appearance of what looks like algebra may create a suspicion that this is a Mathematical Work and therefore quite beyond a beginner.

Yet these devices are not, as might be supposed, for the purpose of making wireless look more learned or difficult. Quite the contrary. Experience has shown them to be the simplest, clearest and most compact ways of conveying the sort of information needed.

The three devices used here are Graphs, Circuit Diagrams, and Algebraic Symbols. The following explanations are only for readers who are not quite used to them.

### ALGEBRAIC SYMBOLS

If a car had travelled 90 miles in 3 hours, we would know that its average speed was 30 miles per hour. How? The mental arithmetic could be written down like this:

$$90 \div 3 = 30$$

That is all right if we are concerned only with that one particular journey. If the same car, or another one, did 7 miles in a quarter of an hour, the arithmetic would have to be:

$$7 \div \tfrac{1}{4} = 28$$

To let anyone know the speed of any car on any journey, it would be more than tedious to have to write out the figures for every possible case. All we need say is, " To find the average speed in miles per hour, divide the number of miles travelled by the number of hours taken ".

What we actually would say would probably be briefer still: " To get the average speed, divide the distance by the time ". Literally that is nonsense, because the only things that can be arithmetically divided or multiplied are numbers. But of course the words " the number of " are (or ought to be!) understood. The other words—" miles per hour ", " miles ", " hours "—the units of measurement, as they are called, may also perhaps be taken for granted in such an easy case. In others, missing them out might lead one badly astray. Suppose the second journey had been specified in the alternative form of 7 miles in 15 minutes. Dividing one by the other would not give the right answer in miles per hour. It would, however, give the right number of miles per minute.

In wireless, as in other branches of physics and engineering, this matter of units is often less obvious, and always has to be kept in mind.

Our instruction, even in its shorter form, could be abbreviated by using mathematical symbols as we did with the numbers:

Average speed = Distance travelled ÷ Time taken.

Here we have a concise statement of general usefulness; note that it applies not only to cars but to railway trains, snails, bullets, the stars in their courses, and everything else in heaven or earth that moves.

Yet even this form of expression becomes tedious when many and complicated statements have to be presented. So for convenience we might write:

$$S = D \div T$$

or alternatively $$S = \frac{D}{T}$$

or, to suit the printer, $$S = D/T$$

**What Letter Symbols Really Mean**

This is the stage at which some people take fright, or become impatient. They say, " How *can* you divide $D$ by $T$? Dividing one letter by another doesn't *mean* anything. You have just said yourself that the only things that can be divided are numbers! " Quite so. It would be absurd to try to divide $D$ by $T$. Those letters are there just to show what to do with the numbers when you know them. $D$, for example, has been used to stand for the number of miles travelled.

The only reason why the letters $S$, $D$ and $T$ were picked for this duty is that they help to remind one of the things they stand for. Except for that there is no reason why the same information should not have been written as

$$x = \frac{y}{z}$$

or even $$\alpha = \frac{\beta}{\gamma}$$

so long as we know what these symbols were intended to mean. As there are only 26 letters in our alphabet, and fewer still in the Greek, it may not be feasible to allocate any one of them permanently, be it $S$ or $x$ or $\alpha$, to mean " average speed in miles per hour ". $x$, in particular, is notoriously capable of meaning absolutely anything. Yet to write out the exact meaning every time would defeat the whole purpose of using the letters. How, then, does one know the meaning?

Well, some meanings have been fixed by international agreement. There is one symbol that means the same thing every time, not only in wireless but in all the sciences—the Greek letter $\pi$ (read as " pie "). It is a particularly good example of abbreviation because it stands for a number that would take eternity to write out in full—the ratio of the circumference of a circle to its diameter, which begins: 3·1415926535 . . . . . (The first three or four decimal places give enough accuracy for most purposes.)

Then there is a much larger group of symbols which have been given meanings that hold good throughout a limited field such as electrical engineering, or one of its subdivisions such as wireless, but are liable to mean something different in, say, astronomy or hydraulics. These have to be learnt by anyone going in for the subject seriously. A list of those that concern us appears on pages 373–374 of this book; but do not try to learn them there. It is much easier to wait till they turn up one by one in the body of the book.

Lastly, there are symbols that one uses for all the things that are not in a standard list. Here we are free to choose our own; but there are some rules it is wise to observe. It is common sense to steer clear as far as possible from symbols that already have established meanings. The important thing is to state the meaning when first using the symbol. It can be assumed to bear this meaning to the end of the particular occasion for which it was attached; after that, the label is taken off and the symbol thrown back into the common stock, ready for use on another occasion, perhaps with a different label—provided it is not likely to be confused with the first.

Sometimes a single symbol might have any of several different meanings, and one has to decide which it bears in that particular connection. The Greek letter $\mu$ (pronounced "mew") is an example that occurs in this book. When the subject is a radio valve, it can be assumed to mean "amplification factor". But if iron cores for transformers are being discussed, $\mu$ should be read as "permeability". And if it comes before an upright letter it is an abbreviation for "micro-", meaning "one millionth of".

### Some Other Uses of Symbols

The last of these three meanings of $\mu$ is a different kind of meaning altogether from those we have been discussing. Until then we had been considering symbols as abbreviations for quantities of certain specified things, such as speed in miles per hour. But there are several other uses to which they are put.

Instead of looking on "$S = D/T$" as an instruction for calculating the speed, we can regard it as a statement showing the relationship to one another of the three quantities, speed, time and distance. Such a statement, always employing the "equals" sign, is called by mathematicians an *equation*. From this point of view $S$ is no more important than $T$ or $D$, and it is merely incidental that the equation was written in such a form as to give instructions for finding $S$ rather than for finding either of the other two. We are entitled to apply the usual rules of arithmetic in order to put the statement into whatever form may be most convenient when we come to put in the numbers for which the letters stand.

For example, we might want to be able to calculate the time taken on a journey, knowing the distance and average speed. We can divide or multiply both "sides" of an equation by any number (known, or temporarily represented by a letter) without upsetting their equality. If we multiply both sides of $S = D/T$ by $T$ we get $ST = D$ ($ST$ being the recognized abbreviation for $S \times T$). Dividing both sides of this new form of the equation by $S$ we get $T = D/S$. Our equation is now in the form of an instruction to divide the number of miles by the speed in miles per hour (e.g., 120 miles at 24 m.p.h. takes 5 hours).

When you see books on wireless (or any other technical subject) with pages covered almost entirely with mathematical symbols, you can take it that instead of explaining in words how their conclusions are reached, the authors are doing it more compactly in symbols. It is because such pages are concentrated essence, rather than that the meanings of the symbols themselves are hard to learn, that makes them difficult. The procedure is to express the known or assumed facts in the form of equations, and then to combine or manipulate these equations according to the established rules in order to draw some useful or interesting conclusions from those facts.

This book, being an elementary one, explains everything in words, and only uses symbols for expressing the important facts or conclusions in concise form.

## Abbreviations

Another use of symbols is for abbreviation pure and simple. We have already used one without explanation—m.p.h.—because it is well known that this means " miles per hour ". The mile-per-hour is a unit of speed. " Lb " is a familiar abbreviation for a unit of weight. The unit of electrical pressure is the volt, denoted by the abbreviation V. Sometimes it is necessary to specify very small voltages, such as 5 millionths of a volt. That could be written 0·000005 V. But a more convenient abbreviation is 5 μV, read as " 5 microvolts ". The list on pages 373–374 gives all the abbreviations commonly used in wireless.

Still another use for letters is to point out details on a diagram. $R$ stands for electrical resistance, but one has to guess whether it is intended to mean resistance in general, as a property of conductors, or the numerical value of resistance in an equation, or the particular resistance marked $R$ in a diagram. Often it may combine these meanings, being understood to mean " the numerical value of the resistance marked $R$ in Fig. So-and-so ".

Attaching the right meanings to symbols probably sounds dreadfully difficult and confusing. So do the rules of a new game. The only way to defeat the difficulties is to start playing the game.

But before starting to read the book, here are a few more hints about symbols.

## How Numbers are Used

One way of making the limited stock of letters go farther is to use different kinds of type. It has become a standard practice to distinguish symbols for physical quantities by italic (sloping) letters, leaving the roman (upright) ones for abbreviations; for example, " *V* " denotes an unspecified amount of electrical potential difference, reckoned in volts, for which the standard abbreviation is " V ".

Another way of making letters go farther is to number them. To prevent the numbers from being treated as separate things they are written small near the foot of the letter (" subscript "). For example, if we want to refer to several different resistances we can mark them $R_1$, $R_2$, $R_3$, etc. Sometimes a modification of a thing denoted by one symbol is distinguished by a tick or dash; $A$ might stand for the amplification of a receiver when used normally, and $A'$ when modified in some way.

But on no account must numbers be used " superscript " for this purpose, because that already has a standard meaning. $5^2$ (read as " 5 squared ") means $5 \times 5$; $5^3$ (" 5 cubed ") means $5 \times 5 \times 5$; $5^4$ (" 5 to the 4th power ") means $5 \times 5 \times 5 \times 5$; $x^2$ means the number represented by $x$ multiplied by the same number.

$x^{-2}$ seems nonsensical according to the rule just illustrated. It has been agreed, however, to make it mean $1/x^2$ (called the *reciprocal* of $x$ squared). The point of this appears most clearly when these superscripts, called *indices* (singular: *index*), are applied to the number 10. $10^4 = 10{,}000$, $10^3 = 1{,}000$, $10^2 = 100$, $10^1 = 10$, $10^0 = 1$, $10^{-1} = 0{\cdot}1$, $10^{-2} = 0{\cdot}01$, and so on. The rule is that the power of 10 indicates the number of places the decimal point has to be away from 1; with a positive index it is on the right; negative, on the left. Advantage is taken of this to abbreviate very large or very small numbers. It is easier to see that $0{\cdot}0000000026$ is 2·6 thousand-millionths if it is written $2{\cdot}6 \times 10^{-9}$. Likewise $2{\cdot}6 \times 10^{12}$ is briefer and clearer than 2,600,000,000,000.

$x^{\frac{1}{2}}$ also needs explanation. It is read as " the square root of $x$ ", and is often denoted by $\sqrt{x}$. It signifies the number which, when multiplied by the same number, gives $x$. In symbols $(\sqrt{x})^2 = x$.

Note that $10^6 \times 10^6 = 10^{12}$, and that $10^{12} \times 10^{-3} = 10^{12}/10^3 = 10^9$. In short, multiply by adding indices, and divide by subtracting them. This idea is very important in connection with decibels (p. 146).

## Graphs

Most of us when we were in the growing stage used to stand bolt upright against the edge of a door to have our height marked up. The succession of marks did not convey very much when reviewed afterwards unless the dates were marked too. Even then one had to look closely to read the dates, and the progress of growth was difficult to visualize. Nor would it have been a great help to have presented the information in the form of a table with two columns—Height and Date.

But, disregarding certain technical difficulties, imagine that the growing boy had been attached to a conveyor belt which moved

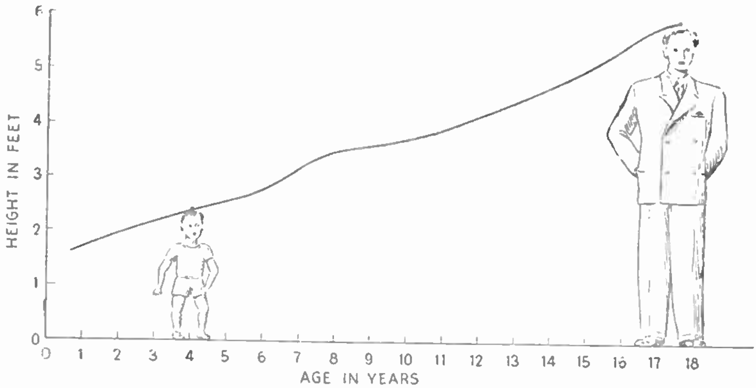



**Fig. 0.1**—Simple (but inconvenient) system for automatic graph plotting of human growth. The equipment consists of a conveyor belt moving at the rate of one foot per year, a pencil, and a white wall

him horizontally along a wall at a steady rate of, say, one foot per year, and that a pencil fixed to the top of his head had been tracing a line on the wall (Fig. 0.1). If he had not been growing at all, this line would, of course, be straight and horizontal. If he had been growing at a uniform rate, it would be straight but sloping upwards. A variable rate of growth would be shown by a line of varying slope.

In this way the progress over a period of years could be visualized by a glance at the wall.

## Scales

To make the information more definite, a mark could have been made along a horizontal line each New Year's Day and the number of the year written against it. In more technical terms this would be a time scale. The advantage of making the belt move at uniform speed is that times intermediate between those actually marked can be identified by measuring off a proportionate *distance*. If one foot represented one year, the height at, say, the end of May in any year could be found by noting the height of the pencil line 5 inches beyond the mark indicating the start of that year.

Similarly a scale of height could be marked anywhere in a vertical direction. It happens in this case that height would be represented by an equal height. But if the graph were to be reproduced in a book, although height would still represent height it would have to do so on a reduced scale, say half an inch to a foot.

To guide the eye from any point on the information line to the two scales, it is usual to plot graphs on paper printed with horizontal and vertical lines so close together that a pair of them is sure to come near enough to the selected point for any little less or more to be judged. The position of the scales is not a vital matter, but unless there is a reason for doing otherwise they are marked along the lines where the other quantity is zero. For instance, the time scale would be placed where height is nil; i.e., along the foot of the wall. If, however, the height scale were erected at the start of the year 1 A.D. there would be nearly half a mile of blank wall between it and the pencil line. To avoid such inconvenience we can use a "false zero", making the scale start at or slightly below the first figure in which we are interested. A sensible way of doing it in this example would be to reckon from the time the boy was born, as in Fig. 0.1. The point that is zero on both scales is called the *origin.*

False zeros are sometimes used in a slightly shady manner to give a wrong impression. Fig. 0.2*a* shows the sort of graph that might appear in a company-promoting prospectus. The word "Profit" would, of course, be in big type, and the figures indistinctly, so that the curve would seem to indicate a sensational growth in profits. Plotted without a false zero, as in Fig. 0.2*b*, it looks much

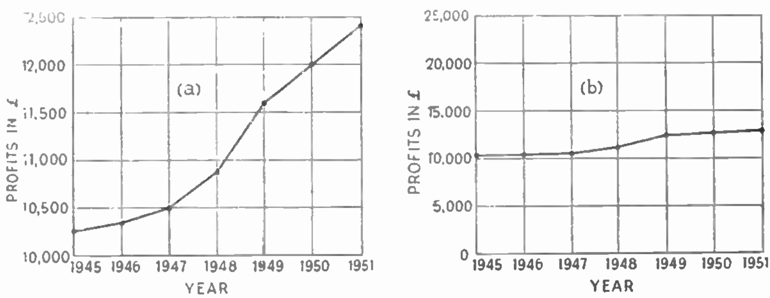



Fig. 0.2—(*a*) Typical financial graph arranged to make the maximum impression consistent with strict truthfulness. (*b*) The same information presented in a slightly different manner

less impressive. Provided that it is clearly admitted, however, a false zero is useful for enabling the significant part of the scale to be expanded and so read more precisely.

Fig. 0.2, by the way, is not a true graph of the sort a mathematician would have anything to do with. Profits are declared annually in lumps, and the lines joining the dots that mark each year's result have no meaning whatever but are there merely to guide the eye from one to another. In a true graph a continuous variation is shown, and the dots are so close together as to form a line. This line is technically a *curve*, whether it is what is commonly understood as a curve or is as straight as the proverbial bee-line.

## What a "Curve" Signifies

Every point on the curve in Fig. 0.1 represents the height of the boy at a certain definite time (or, put in another way, the time at which the boy reached a certain height). Since it is possible to distinguish a great number of points along even a small graph, and each point is equivalent to saying "When the time was $T$ years, the height was $H$ feet ", a graph is not only a very clear way of presenting information, but a very economical one.

Here we have our old friend $T$ again, meaning time, but now in years instead of hours. It is being used to stand for the number of years, as measured along the time scale, represented by *any* point on the curve. Since no particular point is specified, we cannot tell how many feet the corresponding height may be, so have to denote it by $H$. But directly $T$ is specified by a definite number, the number or *value* of $H$ can be found from the curve. And vice versa.

In its earlier role (p. 12) $T$ took part in an equation with two other quantities, denoted by $S$ and $D$. The equation expressed the relationship between these three quantities. We now see that a graph is a method of showing the relationship between two quantities. It is particularly useful for quantities whose relationship is too complicated or irregular to express as an equation—the height and age of a boy, for instance.

## Three-Dimensional Graphs

Can a graph deal with three quantities? There are two ways in which it can, neither of them entirely satisfactory.

One method is to make a three-dimensional graph by drawing a third scale at right angles to the other two. Looking at the corner of a room, one can imagine one scale along the foot of one wall,

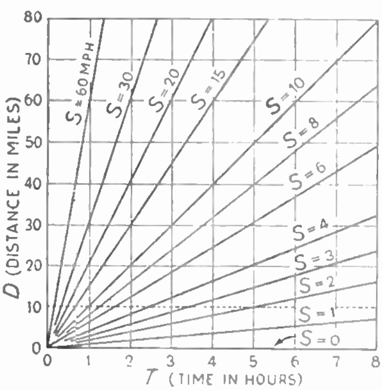



**Fig. 0.3—One way of graphing the relationship between speed, time and distance of a journey**

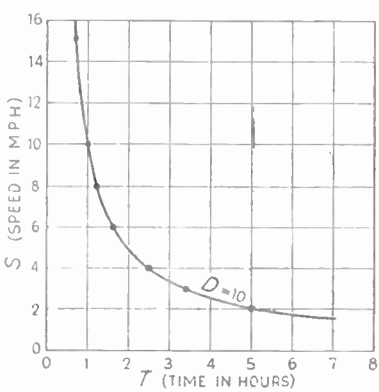



**Fig. 0.4—A graph of speed against time for any given distance can be derived from Fig. 0.3**

another at the foot of the other wall, and the third upwards along the intersection between the two walls. The " curve " would take the form of a surface in the space inside—and generally also outside—the room. There are obvious difficulties in this.

The other method is to draw a number of cross-sections of the three-dimensional graph on a two-dimensional graph. In other words, the three variable quantities are reduced to two by assuming some numerical value for the third. We can then draw a curve showing the relationship between the two. A different numerical value is then assumed for the third, giving (in general) a different graph. And so on.

Let us take again as an example $D = ST$. If any numerical value is assigned to the speed, $S$, we can easily plot a graph connecting $D$ and $T$, as in Fig. 0.3. When the speed is zero, the distance remains zero for all time, so the " $S = 0$ " curve is a straight line coinciding with the time scale. When $S$ is fixed at 1 m.p.h., $D$ and $T$ are always numerically equal, giving a line (" $S = 1$ ") sloping up with a 1-in-1 gradient (measured by the $T$ and $D$ scales). The " $S = 2$ " line slopes twice as steeply; the " $S = 3$ " line three times as steeply, etc.

If a fair number of $S$ curves are available, it is possible to replot the graph to show, say, $S$ against $T$ for fixed values of $D$. A horizontal line drawn from any selected value on the $D$ scale in Fig. 0.3 (such as $D = 10$) cuts each $S$ curve at one point, from which the $T$ corresponding to that $S$ can be read off and plotted, as in Fig. 0.4. Each point plotted in Fig. 0.4 corresponds to an $S$ curve in Fig. 0.3, and when joined up they give a new curve (really curved this time!). This process has been demonstrated for only one value of $D$, but of course it could be repeated to show the time/speed relationship for other distances. Fig. 0.4 gives numerical expression to the well-known fact that the greater the speed the shorter is the time to travel a specified distance.

Graphs of this triple kind are important in many branches of wireless; for example, with valves, whose relationships cannot be accurately expressed as equations.

## Significance of Slope

Even from simple two-dimensional graphs like Fig. 0.1 it may be possible to extract information that is worth replotting separately. It is clear that a third quantity is involved—*rate of growth*. This is not an independent variable; it is determined by the other two. For it is nothing else than rate-of-change-of-height. As we saw at the start, when the curve slopes upward steeply, growth is rapid; when the curve flattens out, it means that growth has ceased; and if the curve started to slope downwards—a negative gradient—it would indicate negative growth. So it would be possible to plot growth in, say, inches per year, against age in years. This idea of rate-of-change, indicated on a graph by slope, is the essence of no less a subject than the differential calculus, which is much easier than it is often made out to be. If you are anxious to get a clearer view of any electrical subject it would be well worth while to read at least the first few chapters of an elementary book on the differential calculus, such as S. P. Thompson's " Calculus Made Easy ".

Fig. 0.5—(Left) Comparison between uniform or " linear " scale (*a*), and logarithmic scale (*b*)

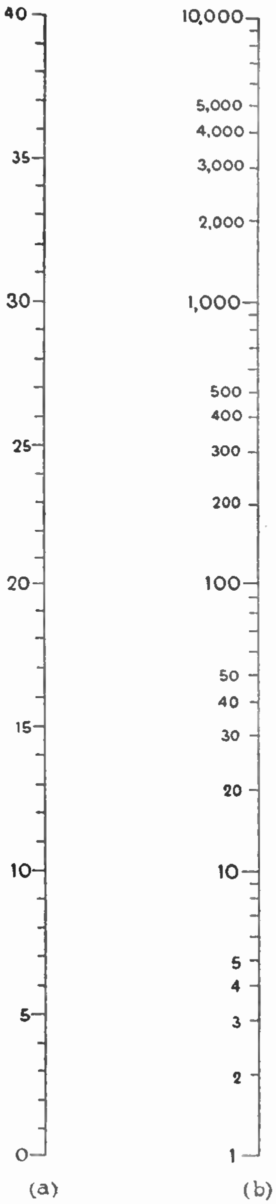
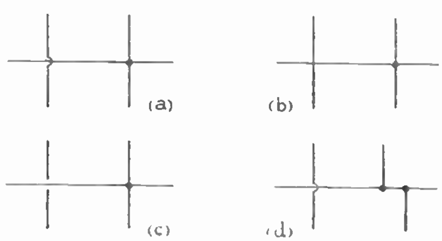



Fig. 0.6—(Below) Alternative methods of distinguishing crossing from connected wires in circuit diagrams



## Non-Uniform Scales

Finally, although uniform scales for graphs are much the easiest to read, there may quite often be reasons which justify some other sort of scale. The most important, and the only one we need consider here, is the *logarithmic* scale. In this, the numbers are so spaced out that a given distance measured anywhere along the scale represents, not a certain *addition*, but a certain *ratio* or *multiplication*. Musicians use such a scale (whether they know it or not); their intervals correspond to ratios. For example, raising a musical note by an octave means doubling the frequency of vibration, no matter whereabouts on the scale it occurs. Users of slide-rules are familiar with the same sort of scale. Readers who are neither musicians nor slide-rule pushers can see the difference by looking at Fig. 0.5, where *a* is an ordinary uniform (or *linear*) scale and *b* is logarithmic. If we start with *a* and note any two scale readings 1½ inches apart we find that they always *differ* by 10. Applying the same test to *b* we find that the difference may be large or small, depending on whether the inch and a half is near the top or bottom. But the larger reading is always 10 times the smaller.

One advantage of this is that it enables very large and very small readings to be shown clearly on the same graph. Another is that with some quantities (such as musical pitch) the ratio is more significant than the numerical interval. Fig. 16.11 is an example of logarithmic scales.

## Circuit Diagrams

It is easier to identify the politicians depicted in our daily papers in the cartoons than in the news photos. There is a sense in which the distorted versions of those gentlemen presented by the caricaturist are more like them than they are themselves. The distinguishing features are picked out and reduced to conventional forms that can be recognized at a glance.

A photograph of a complicated machine shows just a maze of wheels and levers, from which even an experienced engineer might derive little information. But a set of blue-prints, having little relation to the original in appearance, would enable him, if necessary, to reproduce such a machine.

There should be no further need to argue why circuit diagrams are more useful than pictures of radio sets.

In a circuit diagram each component or item that has a significant electrical effect is represented by a conventional symbol, and the wires connecting them up are shown as lines. Fortunately, except for a few minor differences or variations, these symbols are recognizably the same all over the world. Most of them indicate their functions so clearly that one could guess what they mean. However, they are introduced as required in this book, and the whole lot are lined up for reference on p. 375.

## Alternative Methods

The only matter about which there is a sharp difference of opinion is the way in which crossing wires should be distinguished from connected wires. Fig. 0.6*a* shows the method preferred by the author; an unconnected crossing wire is indicated by a little half-loop. In an alternative method *b*, which is general in America, crossing wires are assumed to be unconnected unless marked by a blob. There is obviously a risk that either the blob may be omitted where needed or may form unintentionally (especially when the lines are drawn in ink) where not wanted. Even when correctly drawn the difference is harder to see. A third method *c* is sometimes used; it is neater and easier to draw than *a* and clearer than *b*. The blob is not essential in *a* but it makes the distinction doubly sure. To make triply sure it is recommended that connecting crossings should always be staggered, as at *d*. Of all the methods, *b* is by far the worst. It is only shown here because unfortunately it is often used.

Another variation that may be worth mentioning is that some people omit to draw a ring round the symbols that represent the " innards " of a valve or transistor. In this book they are drawn thicker than the connecting wires, and make the valves (which are usually key components) stand out distinctly. See Fig. 19.21, for example.

## The Importance of Layout

But there is a good deal more in the drawing of circuit diagrams than just showing all the right symbols and connections. If the letters composing a message were written at random all over the paper, nobody would bother to read it even if the correct sequence were shown by a maze of connecting lines—unless it was a clue to valuable buried treasure. And a circuit diagram, though perfectly accurate, baffles even a radio engineer if it is laid out in an unfamiliar manner. The layout has by now become largely standardized, though again there are some national differences. It is too soon in the book to deal with layout in detail, but here are some general principles for later reference.

The first is that the diagram should be arranged for reading from left to right. In a receiver, for example, the aërial should be on the left and the final item, the loudspeaker, on the right.

Next, there is generally an " earthed " or " low-potential " connection. This should be drawn as a thicker line along the foot of the diagram. The highest potential connections should be at the top. This helps one to visualize the potential distribution.

Long runs of closely parallel wires are difficult to follow and should be avoided.

But, as with written symbols, it is easier to get accustomed to circuit diagrams with practice. An effort has been made in this book to exemplify sound practice in diagram drawing.

## Where Circuit Diagrams can Mislead

There is one warning which it is prudent to heed in connection with circuit diagrams, especially when going on to more advanced work. They are an amazingly convenient and well-conceived aid to thought, but it is possible to allow one's thoughts to be moulded too rigidly by them. The diagram takes for granted that all electrical properties are available in separate lumps, like chemicals in bottles, and that one can make up a circuit like a prescription. Instead of which they are more like the smells from the said chemicals when the stoppers have been left open—each one strongest near its own bottle, but pervading the surrounding space and mixing inextricably with the others. This is particularly true at very high frequencies, so that the diagram must not always be taken as showing the whole of the picture. Experience tells how far a circuit diagram can be trusted.

CHAPTER 1

# A General View

## WHAT WIRELESS DOES

PEOPLE WHO REMARK on the wonders of wireless seldom seem to consider the fact that all normal persons can broadcast speech and song merely by using their voices. We can instantly communicate our thoughts to others, without wires or any other visible lines of communication and without even any sending or receiving apparatus outside of ourselves. If anything is wonderful, that is. Wireless is merely a device for increasing the range.

There is a fable about a dispute among the birds as to which could fly the highest. The claims of the wren were derided until, when the great competition took place and the eagle was proudly outflying the rest, the wren took off from his back and established a new altitude record.

It has been known for centuries that communication over vast distances of space is *possible*—it happens every time we look at the stars and detect light coming from them. If there were intelligent beings on the moon they could send messages to us by means of searchlights. We know, then, that a long-range medium of communication exists. Wireless simply uses this medium for the carrying of sound and vision.

Simply?

Well, there actually is quite a lot to learn about it. Before getting to grips with the several kinds of subject matter that must be included in even an elementary book it will be helpful to take a general look at the whole to see why each is necessary and where it fits in.

## THE NATURE OF SOUND WAVES

Wireless, as we have just observed, is essentially a means of extending natural communication beyond its limits of range. It plays the part of the swift eagle in carrying the wren of human speech to unaccustomed distances. So first of all let us see what is required for " natural " communication.

There are three essential parts: the *speaker*, who generates sound; the *hearer* who receives it; and the *medium*, which connects one with the other.

The last of these is the key to the other two. By suspending a source of sound in a container and extracting the air (Fig. 1.1) one can demonstrate that sound cannot travel without air (or some other physical substance) all the way between sender and receiver.

Anyone who has watched a cricket match will recall that the smack of bat against ball is heard a moment after they are seen to

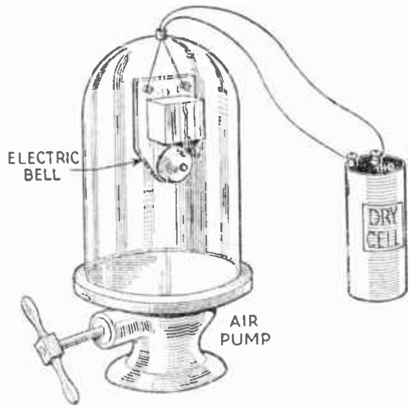



**Fig. 1.1—Experiment to demonstrate that sound waves cannot cross empty space. As the air is removed from the glass vessel, the sound of the bell fades away**

meet; the sound of the impact has taken an appreciable time to travel from the pitch to the viewpoint. If the pitch were 1,100 feet away, the time delay would be one second. The same speed of travel is found to hold good over longer or shorter distances. Knowing the damage that air does when it travels at even one-tenth of that speed (hurricane force) we conclude that sound does not consist in the air itself leaping out in all directions.

When a stone is thrown into a pond it produces ripples, but the ripples do not consist of the water which was directly struck by the stone travelling outwards until it reaches the banks. If the ripples pass a cork floating on the surface they make it bob up and down; they do not carry it along with them.

What does travel, visibly across the surface of the pond, and invisibly through the air, is a *wave*, or more often a succession of waves. While ripples or waves on the surface of water do have much in common with waves in general, they are not of quite the same type as sound waves *through* air or any other medium (including water). So let us forget about the ripples now they have served their purpose, and consider a number of boys standing in a queue. If the boy at the rear is given a sharp push he will bump into the boy in front of him, and he into the next boy; and so the push may be passed on right to the front. The push has travelled along the line although the boys—representing the medium of communication—have all been standing still except for a small temporary displacement.

In a similar manner the cricket ball gives the bat a sudden push; the bat pushes the air next to it; that air gives the air around it a push; and so a wave of compression travels outwards in all directions like an expanding bubble. Yet all that any particular bit of air does is to move a small fraction of an inch away from the cause of the disturbance and back again. As the original impulse is spread over an ever-expanding wave-front the extent of this tiny air movement becomes less and less and the sound fainter and fainter.

## Characteristics of Sound Waves

The generation of sound is only incidental to the functioning of cricket bats, but the human vocal organs make up a generator capable of emitting a great variety of sounds at will. What surprises

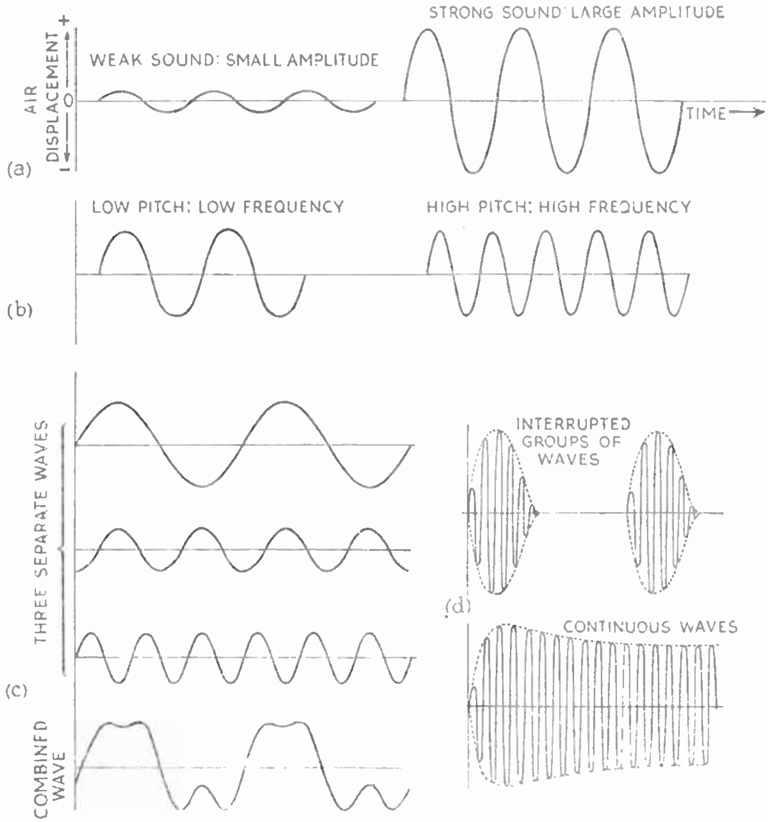



**Fig. 1.2—Graphical representation of the four basic ways in which sound (or any other) waves can differ. (*a*) amplitude; (*b*) frequency; (*c*) waveform; and (*d*) envelope**

most people, when the nature of sound waves is explained to them, is how such a variety of sound can be communicated by anything so simple as to-and-fro tremors in the air. Is it really possible that the subtle inflexions of the voice by which an individual can be identified, and all the endless variety of music, can be represented by nothing more than that?

Examining the matter more closely we find that this infinite variety of sounds is due to four basic characteristics, which can all be illustrated on a piano. They are:

(a) AMPLITUDE. A single key can be struck either gently or hard, giving a soft or loud sound. The harder it is struck, the more violently the piano string vibrates and the greater the tremor in the air, or, to use technical terms, the greater the amplitude of the sound wave. Difference in amplitude is shown graphically in Fig. 1.2*a*. (Note that to-and-fro displacement of the air is represented here by up-and-down displacement of the curve, because horizontal displacement is usually used for time.)

(b) FREQUENCY. Now strike a key nearer the right-hand end of the piano. The resulting sound is distinguishable from the previous one by its higher pitch. The piano string is tighter and lighter, so vibrates more rapidly and generates a greater number of sound waves per second. In other words, the waves have a higher *frequency*, which is represented in Fig. 1.2*b*.

(c) WAVEFORM. If now several keys are struck simultaneously to give a chord, they blend into one sound which is richer than any of the single notes. The first three waves in Fig. 1.2*c* represent three such notes singly; when the displacements due to these separate notes are added together they give the more complicated waveform shown on the last line. Actually any one piano note itself has a complex waveform, which enables it to be distinguished from notes of the same pitch played by other instruments.

(d) ENVELOPE. A fourth difference* in sound character can be illustrated by first "pecking" at a key, getting brief (staccato) notes, and then pressing it firmly down and keeping it there, producing a sustained (legato) note. This can be shown as in Fig. 1.2*d*. The dotted lines, called envelopes, drawn around the wave peaks, have different shapes. Another example of this type of difference is the contrast between the flicker of conversational sound and the steadiness of a long organ note.

Every voice or musical instrument, and in fact everything that can be heard, is something that moves; its movements cause the air to vibrate and radiate waves; and all the possible differences in the sounds are due to combinations of the four basic differences.

The details of human and other sound generators are outside the scope of this book, but the sound waves themselves are the things that have to be carried by long-range communication systems such as wireless. Even if something other than sound is to be carried (pictures, for instance) the study of sound waves is not wasted, because the basic characteristics illustrated in Fig. 1.2 apply to all waves—including wireless waves.

The first (amplitude) is simple and easy to understand. The third and fourth (waveform and envelope) lead us into complications that are too involved for this preliminary survey, and will have to be gone into later. The second (frequency) is very important indeed and not too involved to take at this stage.

* Mathematically it is covered by Waveform, so may not always be listed separately.

## Frequency

Let us consider the matter as exemplified by the vibrating string or wire shown in Fig. 1.3. The upper and lower limits of its vibration are indicated (rather exaggeratedly, perhaps) by the dotted lines. If a point on the string, such as A, were arranged to record its movements on a strip of paper moving rapidly from left to right, it would trace out a wavy line which would be its displacement/time graph, like those in Fig. 1.2. Each complete up-and-down-and-back-again movement is called a *cycle*. In Fig. 1.4, where several cycles are shown, one has been picked out in heavy line. The time it takes, marked $T$, is known as its *period*.

We have already noted that the rate at which the vibrations take place is called the *frequency*, and determines the pitch of the sound produced. It is generally reckoned in cycles per second (abbreviated

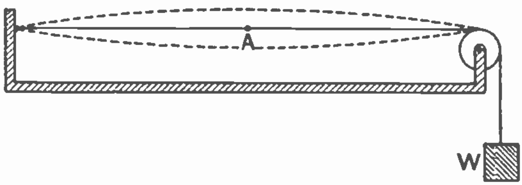



**Fig. 1.3—A simple sound generator, consisting of a string held taut by a weight. When plucked it vibrates as indicated by the dotted lines**

c/s) and its symbol is $f$. Middle C in music has a frequency of 261 c/s. The full range* of a piano is 27 to 3,516 c/s. The frequencies of audible sounds—about 20 to 20,000 c/s for people with normal hearing—are distinguished as *audio frequencies*.

You may have noticed in passing that $T$ and $f$ are closely related. If the period of each cycle is one-hundredth of a second, then the frequency is obviously one hundred cycles per second. Putting it in general terms, $f = 1/T$. So if the time scale of a waveform graph is given, it is quite easy to work out the frequency. In Fig. 1.4 the time scale shows that $T$ is 0·0005 sec, so the frequency must be 2,000 c/s.

## Wavelength

Next, consider the air waves set up by the vibrating string. Fig. 1.5 shows an end view, with point A vibrating up and down. It pushes the air alternately up and down, and these displacements travel outwards from A. Places where the air has been temporarily moved a little farther from A than usual are indicated by full lines, and places where it is nearer by dotted lines; these lines should be imagined as expanding outwards at a speed of about 1,100 feet per

* In wireless the word " range " means distance that can be covered, so to avoid confusion a range of frequencies is more often referred to as a *band*.

second. So if, for example, the string is vibrating at a frequency of 20 c/s, at one second from the start the first air wave will be 1,100 feet away, and there will be 20 complete waves spread over that distance (Fig. 1.6). The length of each wave is therefore one-twentieth of 1,100 feet, which is 55 feet.

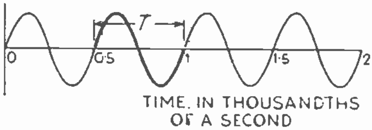



**Fig. 1.4—The extent of one cycle in a series of waves is shown in heavy line. The time occupied by its generation, or by its passing a given point, is marked *T***

Now it is a fact that sound waves of practically all frequencies and amplitudes travel through the air at the same speed, so the higher the frequency the shorter the wavelength. This relationship can be expressed by the equation

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

where the Greek letter $\lambda$ (pronounced " lambda ") stands for wavelength, $v$ for the velocity of the waves, and $f$ for the frequency. If $v$ is given in feet per second and $f$ in cycles per second, then $\lambda$ will be in feet. The letter $v$ has been put here instead of 1,100, because the exact velocity of sound waves in air depends on its temperature; moreover the equation applies to waves in water, wood, rock, or any other substance, if the $v$ appropriate to the substance is filled in.

It should be noted that it is the frequency of the wave which affects the pitch, and that the wavelength is a secondary matter depending on the speed of the wave. That this is so can be shown

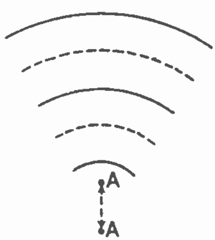



**Fig. 1.5—End view of the vibrating string at A in Fig. 1.3. As it moves up and down over the distance AA it sends out air waves which carry some of its energy of vibration to the listener's ear**

by sending a sound of the same frequency through water, in which the velocity is 4,700 feet per second; the wavelength is therefore more than four times as great as in air, but the pitch, as judged by the ear, remains the same as that of the shorter air wave.

## The Sender

The device shown in Fig. 1.3 is a very simple sound transmitter or sender. Its function is to generate sound waves by vibrating and stirring up the surrounding air. In some instruments, such as violins, the stirring-up part of the business is made more effectual by attaching the vibrating parts to surfaces which increase the amount of air disturbed.

The pitch of the note can be controlled in two ways. One is to vary the weight of the string, which is conveniently done by varying the length that is free to vibrate, as a violinist does with his fingers. The other is to vary its tightness, as a violinist does when tuning, or can be done in Fig. 1.3 by altering the weight W. To lower the

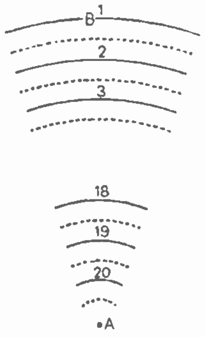



**Fig. 1.6—Twenty successive waves from the string shown in Figs. 1.3 and 1.5**

pitch, the string is made heavier or slacker. As this is not a book on sound there is no need to go into this further, but there will be a lot to say (especially in Chapter 8) on the electrical equivalents of these two things, which control the frequency of wireless waves.

## The Receiver

What happens at the receiving end? Reviving our memory of the stone in the pond, we recall that the ripples made corks and other small flotsam bob up and down. Similarly, air waves when they strike an object try to make it vibrate with the same characteristics as their own. That is how we can hear sounds through a door or partition; the door is made to vibrate, and its vibrations set up a new lot of air waves on our side of it. When they reach our ears they strike the ear-drums; these vibrate and stimulate a very remarkable piece of receiving mechanism, which sorts out the sounds and conveys its findings via a multiple nerve to the brain.

## Electrical Communication by Wire

To extend the very limited range of sound-wave communication it was necessary to find a carrier. Nothing served this purpose very well until the discovery of electricity. At first electricity could only be controlled rather crudely, by switching the current on and off; so spoken messages had to be translated into a code of signals before they could be sent, and then translated back into words at the receiving end. A simple electric telegraph consisted of a battery and a switch or " key " at the sending end, some device for detecting the current at the receiving end, and a wire between to carry the current.

Most of the current detectors made use of the discovery that when some of the electric wire was coiled round a piece of iron the iron became a magnet so long as the current flowed, and would attract other pieces of iron placed near it. If the neighbouring iron was in the form of a flexible diaphragm held close to the ear, its movements towards and away from the iron-cored coil when current was switched on and off caused audible clicks even when the current was very weak.

To transmit speech and music, however, it was necessary to make the diaphragm vibrate in the same complicated way as the sound waves at the sending end. Since the amount of attraction varied with the strength of current, this could be done if the current could be made to vary in the same way as the distant sound waves. In the end this problem was solved by quite a simple device—the carbon microphone, which, with only details improved, is still the type used in modern telephones.

In its simplest form, then, a telephone consists of the equipment shown in Fig. 1.7. Although this invention extended the range of speech from yards to miles, every mile of line weakened the electric currents and set a limit to clear communication. So it was not

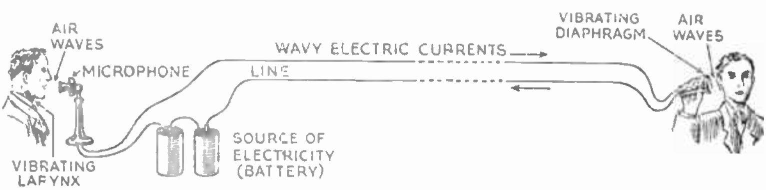



**Fig. 1.7—A very simple form of one-way electric telephone, showing how the characteristics of the original air waves are duplicated in electric currents, which travel farther and quicker, and which are then transformed back into air waves to make them perceptible by the listener**

until the invention of the electronic valve, making it possible to amplify the complicated current variations, that telephoning could be done over hundreds and even thousands of miles.

## Electric Waves

Wire or line telephones and telegraphs were and still are a tremendous aid to communication. But something else was needed for speaking to ships and aeroplanes. And even where the intervening wire is practicable it is sometimes very inconvenient.

At this stage we can profitably think again of the process of unaided voice communication. When you talk to another person you do not have to transfer the vibrations of your vocal organs directly to the ear-drum of your hearer by means of a rod or other " line ". What you do is to stir up waves in the air, and these spread out in all directions, shaking anything they strike, including ear-drums.

"Is it possible", experimenters might have asked, "to stir up *electric* waves by any means, and if so would they travel over greater distances than sound waves?" Gradually scientists supplied the answer. Clerk Maxwell showed mathematically that electric waves were theoretically possible, and indeed that in all probability light, which could easily be detected after travelling vast distances, was an example of such waves. But they were waves of unimaginably high frequency, about $5 \times 10^{14}$ c/s. A few years later Hertz actually produced and detected electric waves of much lower (but still very high) frequency, and found that they shared with sound the useful ability to pass through things that are opaque to light.

Unlike sound, however, they are not carried by the air. They travel equally well (if anything, better) where there is no air or any other material substance present. If the experiment of Fig. 1.1 is repeated with a source of electric waves instead of sound waves, extracting the air makes no appreciable difference. We know, of course, that light and heat waves travel to us from the sun across 93,000,000 miles of empty space. What does carry them is a debatable question. It was named ether (or aether, to avoid confusion with the anaesthetic liquid ether), but experiments which ought to have given results if there had been any such thing just didn't. So scientists deny its existence. But it is very hard for lesser minds to visualize how electric waves, whose behaviour corresponds in so many respects with waves through a material medium, can be propagated by nothingness.

There is no doubt about their speed, however. Light waves travel through space at about 186,282 miles, or nearly 300,000,000 metres, per second. Other sorts of electric waves, such as X-rays, ultra-violet and wireless waves, travel at the same speed, and differ only in frequency. Their length is generally measured in metres; so filling in the appropriate value of $v$ in the formula on p. 19 we have

$$\lambda = \frac{300,000,000}{f}$$

If, for example, the frequency is 1,000,000 c/s, the wavelength is 300 metres.

## Why High Frequencies are Necessary

Gradually it was discovered that electric waves are capable of travelling almost any distance, even round the curvature of the earth to the antipodes.

Their enormous speed was another qualification for the duty of carrying messages; the longest journey in the world takes less than a tenth of a second. But attempts to stir up electric waves of the same frequencies as sound waves were not very successful. To see why this is so we may find it helpful to consider again how sound waves are stirred up.

If you try to radiate air waves by waving your hand you will fail, because the highest frequency you can manage is only a few c/s, and at that slow rate the air has time to rush round from side to side of your hand instead of piling up and giving the surrounding air a push. If you could wave your hand at the speed of an insect's (or even a humming bird's) wing, then the air would have insufficient time to equalize the pressure and would be alternately compressed and rarefied on each side of your hand and so would generate sound waves. Alternatively, if your hand were as big as the side of a house, it could stir up air waves even if waved only a few times per second, because the air would have too far to go from one side to the other every time. The frequency of these air waves would be too low to be heard, it is true; but they could be detected by the rattling of windows and doors all over the neighbourhood.

The same principle applies to electric waves. But because of the vastly greater speed with which they can rush round, it is necessary for the aerial—which corresponds to the waving hand—to be correspondingly vaster. To radiate electric waves at the lower audible frequencies the aerial would have to be miles high, so one might just as well (and more conveniently) use it on the ground level as a telephone line.

Even if there had not been this difficulty, another would have arisen directly large numbers of senders had started to radiate waves in the audible frequency band. There would have been no way of picking out the one wanted; it would have been like a babel of giants.

That might have looked like the end of any prospects of wireless telephony, but fortunately human ingenuity was not to be beaten. The solution arose by way of the simpler problem of wireless telegraphy, so let us follow that way.

## Wireless Telegraphy

Hertz discovered how to stir up electric waves; we will not bother about exactly how, because his method is obsolete. Their frequency was what we would now call ultra-high; in the region of hundreds of millions of c/s, so their length was only a few centimetres. Such frequencies are used nowadays for radar. He also found a method of detecting them over distances of a few yards. With the more powerful senders and more sensitive receivers developed later, ranges rapidly increased, until in 1901 Marconi actually signalled across the Atlantic. The various sorts of detectors that were invented from time to time worked by causing an electric current to flow in a local receiving circuit when electric waves impinged on the receiving aerial. Human senses are unable to respond to these " wireless " waves directly—their frequency is far too high for the ear and far too low for the eye—but the electric currents resulting from their detection can be used to produce audible or visible effects.

At this stage we have the wireless counterpart of the simple telegraph. Its diagram, Fig. 1.8, more or less explains itself. The sending key turns on the sender, which generates a rapid succession of electric waves, radiated by an aerial. When these reach the receiver they operate the detector and cause an electric current to

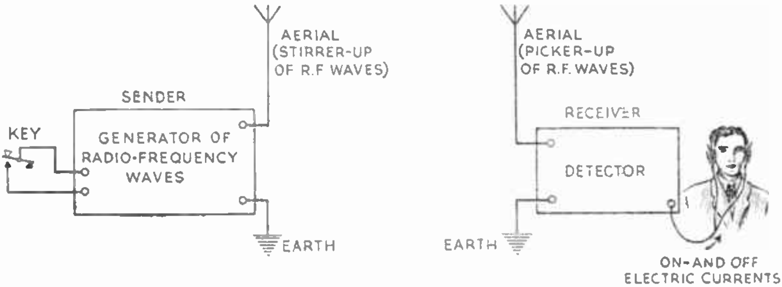



**Fig. 1.8—Elementary wireless telegraph. Currents flow in the receiving headphones and cause sound whenever the sender is radiating waves**

flow, which can make a noise in a telephone earpiece. By working the sending key according to the Morse code, messages can be transmitted. That is roughly the basis of wireless telegraphs to this day, though some installations have been elaborated almost out of recognition, and print the messages on paper as fast as the most expert typist.

## Tuning

One important point to note is that the current coming from the detector does not depend on the frequency of the waves, within wide limits. (The frequency of the current is actually the frequency with which the waves are turned on and off at the sender.) So there is a wide choice of wave frequency. In Chapter 15 we shall consider how the choice is made; but in the meantime it will be sufficient to remember that if the frequency is low, approaching audibility, the aerials have to be immense; while if the frequency is much higher than those used by Hertz it is difficult to generate them powerfully, and they are easily obstructed, like light waves. The useful limits of these *radio* frequencies are about 15,000 to 50,000,000,000 c/s.

The importance of having many frequencies to choose from was soon apparent, when it was found to be possible—and highly advantageous—to make the receiver respond to the frequency of the sender and reject all others. This invention of *tuning* was essential to effective wireless communication, and is the subject of a large part of this book. So it may be enough just now to point out that it has its analogy in sound. We can adjust the frequency of a sound wave by means of the length and tightness of the sender. A piano contains nearly a hundred strings of various fixed lengths and tightnesses, which we can select by means of the keys. If there is a second piano in the room, with its sustaining pedal held down so that all the strings are free to vibrate, and we strike a loud note on the first piano, the string tuned to the same note in the second can be heard to vibrate. If we had a single adjustable string, we could tune it to respond in this way to any one note of the piano and ignore those of substantially different frequency. In a corresponding manner the electrical equivalents of length (or weight) and tightness are adjusted by the tuning controls of a wireless receiver to select the desired sending station.

## Wireless Telephony

The second important point is that (as one would expect) the strength of the current from the detector increases with the strength of the wireless waves transmitted. So, with Fig. 1.7 in mind, we need no further hint to help us to convert the wireless telegraph of Fig. 1.8 into a wireless telephone or a broadcasting system, by substituting for the Morse key a device for varying the strength of the waves, just as the microphone varied the strength of the line current in Fig. 1.7. Such devices are called *modulators*, and will be referred to in more detail in Chapter 13. The microphone is still needed, because the output of the wireless sender is not controlled directly by the impinging sound waves but by the wavily-varying currents from the microphone.

A practical system needs amplifiers in various places, firstly to amplify the feeble sound-controlled currents from the microphone until they are strong enough to modulate a powerful wave sender; then to amplify the feeble radio-frequency currents stirred up in a distant receiving aerial by the far flung waves; then another to amplify the audio-frequency currents from the detector so that they are strong enough to work a loudspeaker instead of the less convenient headphones. This somewhat elaborated system is indicated —but still only in broadest outline—by the block diagram, Fig. 1.9.

## Recapitulation

At this stage it may be as well to review the results of our survey. We realize that in all systems of communication the signals must, in the end, be detectable by the human senses. Of these, hearing and seeing are the only ones that matter. Both are wave-operated. Our ears respond to air waves between frequency limits of about 20 and 20,000 c/s; our eyes respond to waves within a narrow band of frequency centred on $5 \times 10^{14}$ c/s—waves not carried by air but by space, and which have been proved to be electrical in nature.

Although our ears are remarkably sensitive, air-wave (i.e., sound) communication is very restricted in range. Eyes, it is true, can see over vast distances, but communication is cut off by the slightest

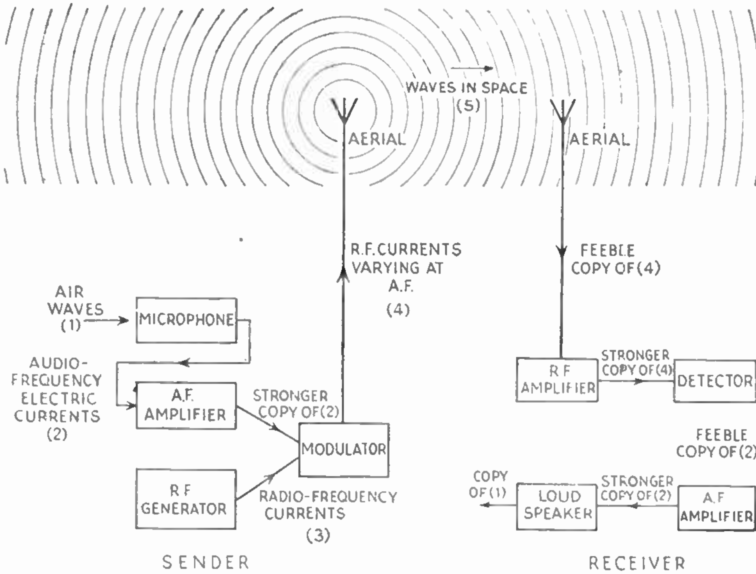



**Fig. 1.9—The wireless telegraph of Fig. 1.8 converted into a wireless telephone by controlling the output of the sender by sound waves instead of just turning it on and off. In practice it is also generally necessary to use amplifiers where shown**

wisp of cloud. If, however, we can devise means for " translating " audible or visible waves into and out of other kinds of waves, we have a far wider choice, and can select those that travel best. It has been found that these are electric waves of lower frequencies than the visible ones, but higher than audible frequencies. The lowest of the radio frequencies require excessively large aerials to radiate them, and the highest are impeded by the air. All travel through space at the same speed of nearly 300,000,000 metres per second, so the wavelength in metres can be calculated by dividing that figure by the frequency in c/s.

The frequency of the sender is arranged by a suitable choice and adjustment of the electrical equivalents of weight and tightness (or, more strictly, slackness) in the tuning of sound generators. In the same way the receiver can be made to respond to one frequency and reject others. There are so many applicants for frequencies—broadcasting authorities, telegraph and telephone authorities, naval, army, and air forces, police, the merchant navies, airlines, and others—that even with such a wide band to choose from it is difficult to find enough for all, especially as only a small part of the whole band is generally suitable for any particular kind of service.

The change-over from the original sound waves to electric waves of higher frequency is done in two stages: first by a microphone into electric currents having the same frequencies as the sound waves; these currents are then used in a modulator to control the radio-frequency waves in such a way that the original sound characteristics can be extracted at the receiver. Here the change-back is done in the same two stages reversed: first by a detector, which yields electric currents having the original sound frequencies; then by ear-phones or loud speakers which use these currents to generate sound waves.

Since the whole thing is an application of electricity, our detailed study must obviously begin with the general principles of electricity. And it will soon be necessary to pay special attention to electricity varying in a wavelike manner at both audio and radio frequencies. This will involve the electrical characteristics that form the basis of tuning. To generate the radio-frequency currents in the sender, as well as to amplify and perform many other services, use is made of various types of electronic valves, so it is necessary to study them in some detail. The process of causing the r.f. currents to stir up waves, and the reverse process at the receiver, demands some knowledge of aerials and radiation. Other essential matters that need elucidation are modulation and detection. Armed with this knowledge we can then see how they are applied in typical senders and receivers. Lastly the special requirements of television and radar will be briefly noted.

CHAPTER 2

# Elementary Electrical Notions

## Electrons

THE EXACT NATURE of electricity is a mystery that may never be fully cleared up, but from what is known it is possible to form a sort of working model or picture which helps us to understand how it produces the results it does, and even to think out how to produce new results. The reason why the very existence of electricity went unnoticed until comparatively recent history was that there are two opposite kinds which exist in equal quantities and (unless separated in some way) cancel one another out. This behaviour reminded the investigating scientists of the use of positive (+) and negative (−) signs in arithmetic: the introduction of either + 1 or − 1 has a definite effect, but + 1 − 1 equals just nothing. So, although at that time hardly anything was known about the composition of the two kinds of electricity, they were called positive and negative. Both positive and negative electricity produce very remarkable effects when they are separate, but a combination of equal quantities shows no signs of electrification.

Further research led to the startling conclusion that all matter consists largely of electricity. All the thousands of different kinds of matter, whether solid, liquid, or gaseous, consist of atoms which contain only a very few different basic components. Of these we need take note of only one kind—particles of negative electricity, called *electrons*. For a simple study it is convenient to imagine each atom as a sort of ultra-microscopic solar system in which a number of electrons are distributed around a central nucleus, somewhat as our earth and the other planets around the sun (Fig. 2.1). The nucleus is generally a more or less composite

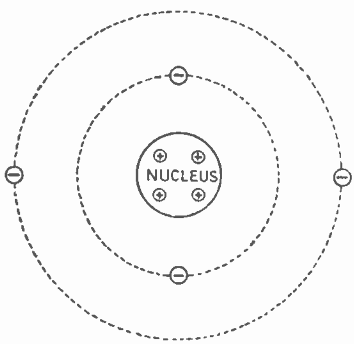



Fig. 2.1—Diagrammatic picture of an atom, showing the nucleus surrounded by a number of electrons

structure; it makes up nearly all the weight of the atom, and the number and type of particles it contains determine which element it is—oxygen, carbon, iron, etc. Recent research has shown how to change some elements into others by breaking up the nucleus under very intense laboratory treatment. For our present purpose, however, the only things we need remember about the nucleus are that it is far heavier than the electrons, and that it has a surplus of positive electricity.

In the normal or unelectrified state of the atom this positive surplus is exactly neutralized by the planetary electrons. But it is a comparatively easy matter to dislodge one or more of these electrons from each atom. One method is by rubbing; for example, if a glass rod is rubbed vigorously with silk some of the electrons belonging to the glass are transferred to the silk. Doing this produces no change in the material itself—the glass is still glass and the silk remains silk. As separate articles they show no obvious evidence of the transfer. But if they are brought close together it is found that they attract one another. And the glass rod can pick up small scraps of paper. It is even possible, in a dry atmosphere, to produce sparks. These curious phenomena gradually fade away, and the materials become normal again.

## Electric Charges and Currents

The silk with its surplus of electrons is said to be negatively charged. The glass has a corresponding deficiency of electrons, so its positive nuclear electricity is incompletely neutralized, and the result is called a positive charge. It is a pity that the people who decided which to call positive did not know anything about electrons, because if they had they would certainly have called electrons positive and so spared us a great deal of confusion. As it is, however, one just has to remember that a *surplus* of electrons is called a *negative* charge.

Any unequal distribution of electrons is a condition of stress. Forcible treatment of some kind is necessary to create a surplus or deficiency anywhere, and if the electrons get a chance they move back to restore the balance; i.e., from negatively to positively charged bodies. This tendency shows itself as an attraction between the bodies themselves. That is what is meant by saying that opposite charges attract.

The space between opposite charges, across which this attraction is exerted, is said to be subject to an *electric field*. The greater the charges, the more intense is the field and the greater the attractive force.

If a number of people are forcibly transferred from town A to town B, the pressure of overcrowding and the intensity of desire to leave B depend, not only on the number of displaced persons, but also on the size of B and possibly on whether there is a town C close by with plenty of accommodation. The electrical pressure also depends on other things than the charge, reckoned in displaced electrons; so it is more convenient to refer to it by another term, namely *difference of potential*, often abbreviated to p.d. We shall consider the exact relationship between charge and p.d. in the next chapter. The thing to remember now is that wherever there is a difference of potential between two points there is a tendency for electrons to move from the point of lower (or negative) potential to that of higher (or positive) potential in order to equalize the distribution.* If they are free to move they will do so; and their movement is what we call an *electric current.*

This is where we find it so unfortunate that the names " positive " and " negative " were allocated before anybody knew about electrons. For it amounted to a guess that the direction of current flow was the opposite to that in which we now know electrons flow. By the time the truth was known this bad guess had become so firmly established that reversing it would have caused worse confusion. Moreover, when the positively charged atoms (or positive *ions*) resulting from the removal of electrons are free to do so they also move, in the direction + to −, though much more slowly owing to their greater mass. So this book follows the usual custom of talking about current flowing from + to − or high to low potential, but it must be remembered that most often this means electrons moving from − to +.

## Conductors and Insulators

Electrons are not always free to move. Except in special circumstances (such as high temperature, which we shall consider in connection with valves) atoms do not allow their electrons to fly off completely on their own, even when they are surplus. The atoms of some substances go so far as to allow frequent exchanges, however, like dancers in a Paul Jones, and in fact such exchanges go on all the time, even in the normal unelectrified state. The directions in which the electrons flit from one atom to another are then completely random, because there is nothing to influence them one way or another.

But suppose the whole piece of substance is pervaded by an electric field. The electrons feel an attraction towards the positive end; so between partnerships they tend to drift that way. The drift is what is known as an electric current, and substances that allow this sort of thing to go on are called *conductors* of electricity. All the metals are more or less good conductors; hence the extensive use of metal wire. Carbon and some liquids are fairly good conductors.

Note that although electrons start to go in at the negative end and electrons start to arrive at the positive end the moment the p.d. is set up, this does not mean that these are the same electrons, which have instantaneously travelled the whole length of the

* Note that + 1 is positive with respect to − 1 or even to 0, but it is negative with respect to + 2; − 1 or 0 are both positive with respect to − 2. There is nothing inconsistent in talking about two oppositely charged bodies one minute, and in the next referring to them both as positive (relative to something else).

conductor. An electron drift starts almost instantaneously throughout miles of wire, but the speed at which they drift is seldom much more than an inch a minute.

Other substances keep their electrons, as it were, on an elastic leash which allows them a little freedom of movement, but never " out of sight " of the atom. If such a substance occupies the space between two places of different potential, the electrons strain at the leash in response to the positive attraction, but a continuous steady drift is impossible. Materials of this kind, called *insulators*, can be used to prevent charges from leaking away, or to form boundaries restricting currents to the desired routes. Dry air, glass, ebonite, rubber, and silk are among the best insulating materials. None is absolutely perfect, however; electrified glass rods, for example, gradually lose their charges.

We shall see later that there is a third class of substances, of such importance that Chapter 10 is devoted to them.

## Electromotive Force

If electrons invariably moved from — to +, as described, they would in time neutralize all the positive charges in the world, and that, for all practical purposes, would be the end of electricity. But fortunately there are certain appliances, such as batteries and dynamos (or generators) which can force electrons to go from + to —, contrary to their natural inclination. In this way they can continuously replenish a surplus of electrons. Suppose A and B in Fig. 2.2 are two insulated metal terminals, and a number of electrons have been transferred from A to B, so that A is at a higher

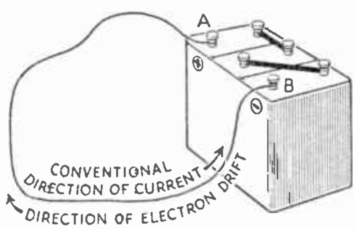



**Fig. 2.2—A very simple electrical circuit, round which the electromotive force of a battery makes a current of electrons flow continuously**

potential than B. If now they are joined by a wire, electrons will drift along it, and, if that were all, the surplus would soon be used up, and the potential difference would disappear. But A and B happen to be the terminals of a battery, and as soon as electrons leave B the battery provides more, while at the same time it withdraws electrons arriving at A. So while electrons are moving from B to A through the wire, the battery keeps them moving from A to B through itself. The battery is, of course, a conductor; but if it were only that it would be an additional path for electrons to go from B to A and dissipate the charge all the quicker. It is remarkable, then, for being able to make electrons move against a p.d. This ability is called *electromotive force* (usually abbreviated to e.m.f.).

Fig. 2.2 shows what is called a *closed circuit*, there being a continuous endless path for the current. This is invariably necessary for a continuous current, because if the circuit were opened anywhere, say by disconnecting the wire from one of the terminals, continuation of the current would cause electrons to pile up at the break, and the potential would increase without limit; which is impossible.

## ELECTRICAL UNITS

The amount or strength of an electric current might reasonably be reckoned as the number of electrons passing any point in the circuit each second, but the electron is so extremely small that such a unit would lead to inconveniently large numbers. For practical purposes it has been agreed to base the unit on the metric system. It is called the *ampere* (or more colloquially the amp), and happens to be about 6,240,000,000,000,000,000 (or $6{\cdot}24 \times 10^{18}$) electrons per second.

As one would expect, the number of amps caused to flow in a given circuit depends on the strength of the electromotive force operating in it. The practical unit of e.m.f. is the *volt*.

Obviously the strength of current depends also on the circuit—whether it is made up of good or bad conductors. This fact is expressed by saying that circuits differ in their electrical *resistance*. The resistance of a circuit or of any part of it can be reckoned in terms of the e.m.f. required to drive a given current through it. For convenience the unit of resistance is made numerically equal to the number of volts required to cause one amp to flow, and to avoid the cumbersome expression " volts per amp " this unit of resistance has been named the *ohm*.

By international agreement the following letter symbols have been allocated to denote these electrical quantities and their units, and will be used from now on:

| *Quantity* | *Symbol for Quantity* | *Unit of Quantity* | *Symbol for Unit* |
|---|---|---|---|
| E.M.F. | $E$ | Volt | V |
| Current | $I$ | Ampere | A |
| Resistance | $R$ | Ohm | Ω |

(Ω is a Greek capital letter, " Omega ".)

These three important quantities are not all independent, for we have just defined resistance in terms of e.m.f. and current. Expressing it in symbols:

$$R = \frac{E}{I}$$

## OHM'S LAW

The question immediately arises: does $R$ depend only on the conductor, or does it depend also on the current or voltage? This was one of the first and most important investigations into electric currents. We can investigate it for ourselves if we have an instrument for measuring current, called an *ammeter*, a battery of identical cells, and a length of wire or other resistive conductor. A closed circuit is formed of the wire, the ammeter, and a varying number of the cells; and the ammeter reading corresponding to each number of cells is noted. The results can then be plotted in the form of a graph (Fig. 2.3). If the voltage of a cell is known, the current can be plotted against voltage instead of merely against number of cells.

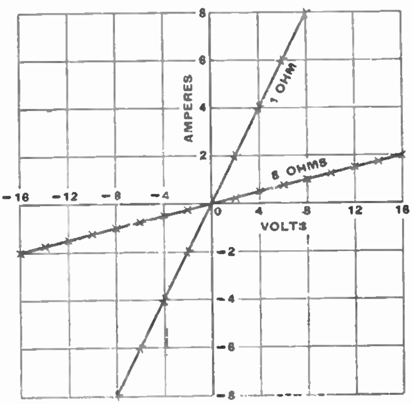



**Fig. 2.3—Graph showing the results of an experiment on the relationship between current in amps and e.m.f. in volts, for two different circuits**

When all the points are joined up by a line it will probably be found that the line is straight and inclined at an angle and passes through the origin (0). If a different resistance is tried, its line will slope at a different angle. Fig. 2.3 shows two possible samples resulting from such an experiment. In this case each cell gave 2 volts, so points were plotted at 2, 4, 6, 8, etc., volts. The points at $-2$, $-4$, etc., volts were obtained by reversing the battery. A current of $-2$ amps means a current of the same strength as $+2$ amps, but in the opposite direction.

Looking at the steep line, we see that an e.m.f. of 2 V caused a current of 2 A, 4 V caused 4 A, and so on. The result of dividing the number of volts by the number of amps is always 1. And this, according to our definition, is the resistance in ohms. So there is no need to specify the current at which the resistance must be measured, or to make a graph; it is only necessary to measure the voltage required to cause any one known current (or the current caused by any one voltage). The amount (or, as one says in technical language, the *value*) of resistance so obtained can be used to find the current at any other voltage. For this purpose it is more convenient to write the relationship $R = E/I$ in the form

$$I = \frac{E}{R} \quad \text{or} \quad E = IR$$

Thus if any two of these quantities are known, the third can be found.

Conductors whose resistance does not vary with the amount of current flowing through them are described as *linear*, because the line representing them on a graph of the Fig. 2.3 type (called their *current/voltage characteristic*) is straight. Although the relationship $I = E/R$ is true for any resistance, its greatest usefulness lies in the fact, discovered by Dr. Ohm, that ordinary conductors are linear.* With this assumption it is known as Ohm's law.

As an example of the use of Ohm's law, we might find, in investigating the value of an unknown resistance, that when it was connected to the terminals of a 100 V battery a current of 0·01 A was driven through it. Using Ohm's law in the form $R = E/I$, we get for the value of the resistance $100/0{\cdot}01 = 10{,}000\ \Omega$. Alternatively, we might know the value of the resistance and find that an old battery, nominally of 120 V, could only drive a current of 0·007 A through it. We could deduce, since $E = I \times R$, that the voltage of the battery had fallen to $10{,}000 \times 0{\cdot}007 = 70$ V. Note that in driving current through a resistance an e.m.f. causes a p.d. of the same voltage between the ends of the resistance. This point is elaborated on page 42.

## Larger and Smaller Units

It is unusual to describe a current as 0·007 ampere, as was done just now; one speaks of " 7 milliamperes ", or, more familiarly, " 7 milliamps ". A milliampere is one-thousandth part of an ampere. Several other prefixes are used; the most frequent are:

| *Prefix* | *Meaning* | *Symbol* |
|---|---|---|
| milli- | one thousandth of | m |
| micro- | one millionth of | $\mu$ |
| micromicro- or pico- | one billionth of (million millionth) | $\mu\mu$ or p |
| kilo- | one thousand | k |
| mega- | one million | M |

($\mu$ is the Greek letter "mu".)

These prefixes can be put in front of any unit; one speaks commonly of milliamps, microamps, megohms, and so on. " Half a megohm " is easier to say than " Five hundred thousand ohms ", just as " ½ MΩ " is quicker to write than " 500,000 Ω ".

It must be remembered, however, that Ohm's law in the forms given above refers to volts, ohms, and amps. If a current of 5 mA is flowing through 15,000 Ω, the voltage across that resistance will

* We shall come across some exceptions, such as valves and rectifiers, from Chapter 9 onwards.

*not* be 75,000 V. But seeing that most of the currents we shall be concerned with are of the milliamp order, it is worth noting that there is no need to convert them to amps if $R$ is expressed in *thousands* of ohms (k$\Omega$). So in the example just given one can get the correct answer, 75 V, by multiplying 5 by 15.

## CIRCUIT DIAGRAMS

At this stage it would be as well to start getting used to circuit diagrams. In a book like this, concerned with general principles rather than with constructional details, what matters about (say) a battery is not its shape nor the design of its label, but its voltage and

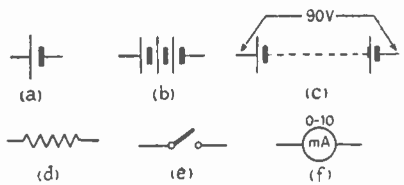



**Fig. 2.4—A first instalment of circuit-diagram symbols, representing (*a*) a cell, (*b*) a battery of three cells, (*c*) a battery of many cells, (*d*) a resistor, (*e*) a switch, and (*f*) a milliammeter**

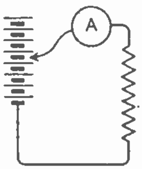



**Fig. 2.5—Circuit diagram of the apparatus used for obtaining Fig. 2.3**

how it is connected in the circuit. So it is a waste of time drawing a picture as in Fig. 2.2. All one needs is the symbol shown as Fig. 2.4*a*, which represents one cell. An accumulator cell gives 2 V, and a dry cell about 1·4 V. The longer stroke represents the + terminal. To get higher voltages, several cells are connected one after the other, as in Fig. 2.4*b*, and if so many are needed that it is tiresome to draw them all one can represent all but the end ones by a dotted line, as in Fig. 2.4*c*, which also shows how to indicate the voltage between two points. Other symbols in Fig. 2.4 are (*d*) a circuit element called a *resistor*, because resistance is its significant feature, (*e*) a switch shown " open " or " off ", and (*f*) a measuring instrument, the type of which can be shown by " mA " for milliammeter, " V " for voltmeter, etc., and the range of measurement marked as here. A simple line represents an electrical connection of negligible resistance, often called a *lead* (rhyming with " feed ").

The circuit diagram for the Ohm's law experiment can therefore be shown as in Fig. 2.5. The arrow denotes a movable connection, used here for including anything from one to eight cells in circuit.

Fig. 2.5 is a simple example of a *series* circuit. Two elements are said to be connected in series when, in tracing out the path of the current, we encounter them one after the other. With steadily flowing currents the electrons do not start piling up locally, so it is clear that the same current flows through both elements. In Fig. 2.5, for instance, the current flowing through the ammeter is the same as that through the resistor and the part of the battery " in circuit ".

Two elements are said to be in *parallel* if they are connected so as to form alternative paths for the current; for example, the resistors in Fig. 2.8*a*. Since a point cannot be at two different potentials at once, it is obvious that the same potential difference exists across both of them (i.e., between their ends or terminals). One element in parallel with another is often described as *shunting* the other.

The same method of connection can sometimes be regarded as either series or parallel, according to circumstances. In Fig. 2.6, $R_1$ and $R_2$ are in series as regards battery $B_1$, and in parallel as regards $B_2$.

Looking at the circuit diagrams of wireless sets—especially for television—one often sees quite complicated networks of resistors. Fortunately Ohm's law can be applied to every part of a complicated system of e.m.fs and resistances as well as to the whole. Beginners often seem reluctant to make use of this fact, being scared by the apparent difficulty of the problem. So it may be as well to see how it works out with more elaborate circuits.

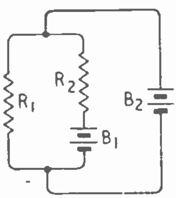



Fig. 2.6—$R_1$ is in series with $R_2$ as regards $B_1$, but in parallel with it as regards $B_2$

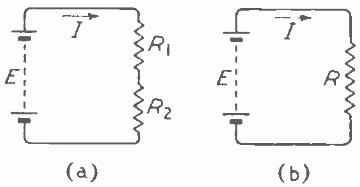



Fig. 2.7—Resistances in series. The circuit *b* is equivalent to the circuit *a*, in the sense that both take the same current from the battery, if $R = R_1 + R_2$

## RESISTANCES IN SERIES AND IN PARALLEL

Complex circuit networks can be tackled by successive stages of finding a single element that is equivalent (as regards the quantity to be found) to two or more. Consider first two resistances $R_1$ and $R_2$, in series with one another and with a source of e.m.f. $E$ (Fig. 2.7*a*). To bring this circuit as a whole within the scope of calculation by Ohm's law we need to find the single resistance $R$ (Fig. 2.7*b*) equivalent to $R_1$ and $R_2$ together.

We know that the current in the circuit is everywhere the same; call it $I$. Then, by Ohm's law, the voltage across $R_1$ is $IR_1$, and that across $R_2$ is $IR_2$. The total voltage across both must therefore be $IR_1 + IR_2$, or $I(R_1 + R_2)$, which must be equal to the voltage $E$.

In the equivalent circuit, $E$ is equal to $IR$, and since, to make the circuits truly equivalent, the currents must be the same in both for the same battery voltage, we see that

$$R = R_1 + R_2.$$

Generalizing from this result, we conclude that: *The total resistance of several resistances in series is equal to the sum of their individual resistances.*

Turning to the parallel-connected resistances of Fig. 2.8*a*, we see that they have the same voltage across them; in this case the e.m.f. of the battery. Each of these resistances will take a current depending on its own resistance and on the e.m.f.—the simplest case of Ohm's law. Calling the currents respectively $I_1$ and $I_2$, we therefore know that $I_1 = E/R_1$ and $I_2 = E/R_2$. The total current drawn is the sum of the two: it is

$$I = \frac{E}{R_1} + \frac{E}{R_2} = E\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$$

In the equivalent circuit of Fig. 2.8*b* the current is $E/R$, which may also be written $E\ (1/R)$. Since, for equivalence between the circuits, the current must be the same for the same battery voltage, we see that

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}$$

Generalizing from this result, we conclude that: *If several resistances are connected in parallel, the sum of the reciprocals of their*

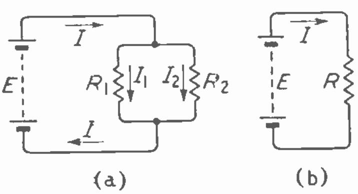



**Fig. 2.8—Resistances in parallel. The circuit *b* is equivalent to the circuit *a* in the sense that both take the current from the battery, if $1/R = 1/R_1 + 1/R_2$**

*individual resistances is equal to the reciprocal of their total resistance.*

If the resistances of Fig. 2.8*a* were 100 Ω and 200 Ω, the single resistance $R$ which, connected in their place, would draw the same current is given by $1/R = 1/100 + 1/200 = 0{\cdot}01 + 0{\cdot}005 = 0{\cdot}015$. Hence, $R = 1/0{\cdot}015 = 66{\cdot}67\ \Omega$. This could be checked by adding together the individual currents through 100 Ω and 200 Ω, and comparing the total with the current taken from the same voltage-source by 66·67 Ω. In both cases the result is 0·015 A per volt of battery.

We can summarize the two rules in symbols:

1. Series Connection: $R = R_1 + R_2 + R_3 + R_4 + \cdots$
2. Parallel Connection: $1/R = 1/R_1 + 1/R_2 + 1/R_3 + \cdots$

For only two resistances in parallel, a more convenient form of the same rule can be obtained by multiplying above and below by $R_1R_2$:

$$R = \frac{1}{1/R_1 + 1/R_2} = \frac{R_1R_2}{R_1 + R_2}$$

## SERIES-PARALLEL COMBINATIONS

How the foregoing rules, derived directly from Ohm's law, can be applied to the calculation of more complex circuits can perhaps

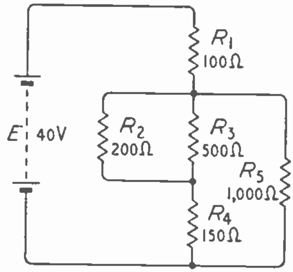



**Fig. 2.9—The current through and voltage across each resistor in this complicated network can be calculated by applying the two simple rules derived from Figs. 2.7 and 2.8**

best be illustrated by a thorough working-out of one fairly elaborate network, Fig. 2.9. We will find the total current flowing, the equivalent resistance of the whole circuit, and the voltage and current of every resistor individually.

The policy is to look for any resistances that are in simple series or parallel. The only two in this example are $R_2$ and $R_3$. Writing $R_{23}$ to symbolize the combined resistance of $R_2$ and $R_3$ taken together, we know that $R_{23} = R_2R_3/(R_2 + R_3) = 200 \times 500/700 = 142{\cdot}8\,\Omega$. This gives us the simplified circuit of Fig. 2.10*a*. If $R_{23}$ and $R_4$ were one resistance, they and $R_5$ in parallel would make another simple case, so we proceed to combine $R_{23}$ and $R_4$ to make $R_{234}$: $R_{234} = R_{23} + R_4 = 142{\cdot}8 + 150 = 292{\cdot}8\,\Omega$. Now we have the circuit of Fig. 2.10*b*. Combining $R_{234}$ and $R_5$, $R_{2345} = 292{\cdot}8 \times 1000/1292{\cdot}8 = 226{\cdot}5\,\Omega$. This brings us within sight of the end; Fig. 2.10*c* shows us that the total resistance of the network now is simply the sum of the two remaining resistances; that is, $R$ of Fig. 2.10*d* is $R_{2345} + R_1 = 226{\cdot}5 + 100 = 326{\cdot}5\,\Omega$.

From the point of view of current drawn from the 40-V source the whole system of Fig. 2.9 is equivalent to a single resistor of this value. The current taken from the battery will therefore be $40/326{\cdot}5 = 0{\cdot}1225$ A $= 122{\cdot}5$ mA.

To find the current through each resistor individually now merely means applying Ohm's law to some of our previous results. Since $R_1$ carries the whole current of 122·5 mA, the potential difference across it will be $100 \times 0{\cdot}1225 = 12{\cdot}25$ V. $R_{2345}$ also carries the whole current (2.10*c*); the p.d. across it will again be the product of resistance and current, in this case $226{\cdot}5 \times 0{\cdot}1225 = 27{\cdot}75$ V. This same voltage also exists, as comparison of the various diagrams will show, across the whole complex system $R_2R_3R_4R_5$ in Fig. 2.9. Across $R_5$ there comes the whole of this voltage; the current through this resistor will therefore be 27·75/1000 A = 27·75 mA.

The same p.d. across $R_{234}$ of Fig. 2.10*b*, or across the system $R_2$ $R_3$ $R_4$ of Fig. 2.9, will drive a current of 27·75/292·8 = 94·75 mA through this branch. The whole of this flows through $R_4$ (2.10*a*), the voltage across which will accordingly be $150 \times 0{\cdot}09475 = 14{\cdot}21$ V. Similarly, the p.d. across $R_{23}$ in Fig. 2.10*a*, or across both $R_2$ and $R_3$

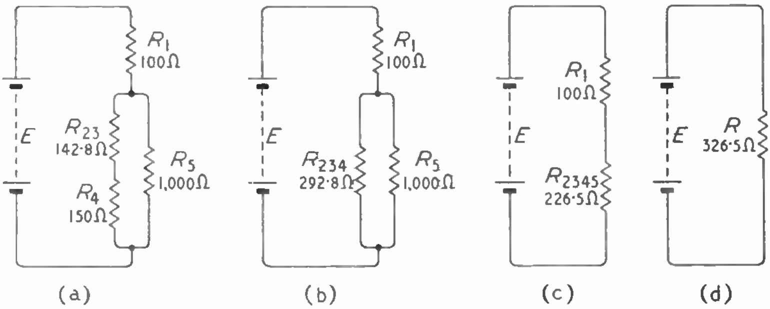



**Fig. 2.10—Successive stages in simplifying the circuit of Fig. 2.9. $R_{23}$ stands for the single resistances equivalent to $R_2$ and $R_3$; and so on. $R$ represents the whole system**

in Fig. 2.9, will be $0{\cdot}09475 \times 142{\cdot}8 = 13{\cdot}54$ V, from which we find that the currents through $R_2$ and $R_3$ will be respectively 13·54/200 and 13·54/500 A, or 67·68 and 27·07 mA, making up the required total of 94·75 mA for this branch.

This completes an analysis of the entire circuit; we can now collect our scattered results in the form of the table below.

| *Resistance (ohms)* | | *Current through it (milliamps)* | *Potential Difference across it (volts)* |
|---|---|---|---|
| $R_1$ | 100 | 122·5 | 12·25 |
| $R_2$ | 200 | 67·68 | 13·54 |
| $R_3$ | 500 | 27·07 | 13·54 |
| $R_4$ | 150 | 94·75 | 14·21 |
| $R_5$ | 1000 | 27·75 | 27·75 |
| $R$ | 326·5 | 122·5 | 40 |

It should be noted that by using suitable resistors any potential intermediate between those given by the terminals of the battery can be obtained. For instance, if the lower and upper ends of the battery in Fig. 2.9 are regarded as 0 and + 40 (they could equally be reckoned as $-40$ and 0, or $-10$ and $+30$, with respect to any selected level of voltage), the potential of the junction between $R_3$ and $R_4$ is 14·21 V. The arrangement is therefore called a *potential divider*, and is often employed for obtaining a desired potential not given directly by the terminals of the source. If a sliding connection is provided on a resistor, to give a continuously variable potential, it is generally known—though not always quite justifiably—as a *potentiometer*.

## Resistance Analysed

So far we have assumed the possibility of almost any value of resistance without inquiring very closely into what determines the resistance of any particular resistor or part of a circuit. We understand that different materials vary widely in the resistance they offer to the flow of electricity, and can guess that with any given material a long piece will offer more resistance than a short piece, and a thin piece than a thick. We have indeed actually proved as much and more; by the rule for resistances in series the resistance

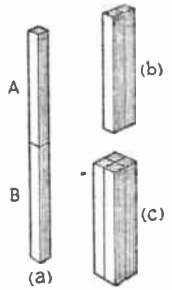



**Fig. 2.11—The rule for resistances in series proves that doubling the length of a uniform conductor doubles its resistance (*a*). By the rule for resistances in parallel, doubling its cross-section area halves its resistance (*b*); and doubling its diameter quarters its resistance (*c*)**

of a uniform wire is exactly proportional to its length—doubling the length is equivalent to adding another equal resistance in series, and so on (Fig. 2.11*a*). Similarly, putting two equal pieces in parallel, which halves the resistance, is equivalent to doubling the thickness; so resistance is proportional to the reciprocal of thickness. Or, in more precise language, to the cross-section area (Fig. 2.11*b*). Altering the shape of the cross-section has no effect on the resistance—with steady currents, at least. Fig. 2.11*c* shows how doubling the *diameter* of a piece of wire divides the resistance by *four*.

To compare the resistances of different materials it is necessary to bring them all to the same standard length and cross-section area, which, in what is known as the m.k.s (metre-kilogram-second) system of units, is one metre long by one square metre. This comparative resistance for any material is called its *resistivity* (symbol: $\rho$, pronounced " roe "). Knowing the resistivity (and it can be found in almost any electrical reference book*) one can easily

* In the older books, resistivity is given in ohms per *centimetre* cube, or ohm-cm. Resistivity in metre units is a number 100 times smaller.

calculate the resistance of any wire or piece having uniform section by multiplying by the length in metres ($l$) and dividing by the cross-section area in square metres ($A$). In symbols:

$$R = \frac{\rho l}{A}$$

It is seldom necessary to make such calculations, because there are tables showing the resistance per metre or per yard of all the different gauges of wire, both for copper (usually used for parts of the circuit where resistance should be as low as possible) and for the special alloys intended to have a high resistance.

Resistance varies to some extent with temperature, so the tables show the temperature at which they apply, and also the proportion by which the resistance rises or falls with rise of temperature.

### Conductance

It is often more convenient to work in terms of the *ease* with which a current can be made to flow, rather than the difficulty; in other words, in *conductance* rather than resistance. The symbol for conductance is $G$, and its unit is equal to the number of amps caused to flow by one volt. So $I = EG$ is an alternative form of $I = E/R$; and $G = I/R$. As conductance is thus resistance " up-side down ", its unit is called the mho. It has no official symbol, but ℧ has been suggested.

It is not difficult to see that corresponding to the equation on the previous page for calculating resistance there will be one for conductance:

$$G = \frac{\sigma A}{l}$$

where $\sigma$ (Greek " sigma ") stands for *conductivity* (mhos per metre).

The advantage of conductance appears chiefly in parallel circuits, for which it is easy to see that the formula is $G = G_1 + G_2 + \ldots.$

### Kirchhoff's Laws

We have already taken it as obvious that no one point can be at more than one potential at the same time. (If it is not obvious, try to imagine a point with a difference of potential between itself; then electrons will flow from itself to itself, i.e., a journey of no distance and therefore non-existent.) So if you start at any point on a closed circuit and go right round it, adding up all the voltages on the way, with due regard for + and –, the total is bound to be zero. If it isn't, a mistake has been made somewhere; just as a surveyor must have made a mistake if he started from a certain spot, measuring his ascents as positive and his descents as negative, and his figures told him that when he returned to his starting-point he had made a net ascent or descent.

This simple principle is one of what are known as Kirchhoff's laws, and is a great help in tackling systems of resistances where the methods already described fail because there are no two that can be simply combined. One writes down an equation for each closed loop, by adding together all the voltages and equating the total to zero, and one then solves the resulting simultaneous equations. That can be left for more advanced study, but the same law is a valuable check on any circuit calculations. Try applying it to the several closed loops in Fig. 2.9, such as that formed by E, $R_1$, and $R_5$. If we take the route clockwise we go " uphill " through the battery, becoming more positive by 40 V. Coming down through $R_1$ we move from positive to negative so add $-12{\cdot}25$ V, and, through $R_5$, $-27{\cdot}75$ V, reaching the starting point again. Check: $40 - 12{\cdot}25 - 27{\cdot}75 = 0$.

Taking another clockwise route via $R_3$, $R_5$, and $R_4$, we get $13{\cdot}54 - 27{\cdot}75 + 14{\cdot}21 = 0$. And so on for any closed loop.

The other Kirchhoff law is equally obvious; it states that if you count all currents arriving at any point in a circuit as positive, and those leaving as negative, the total is zero. This can be used as an additional check, or as the basis of an alternative method of solving circuit problems.

## P.D. AND E.M.F.

The fact that both e.m.f. and p.d. are measured in volts may at first be a little confusing. E.m.f. is a cause; p.d. an effect. An e.m.f. can only come from something active, such as a battery or generator. It has the ability to set up a p.d. But a p.d. can exist across a passive element such as a resistor, or even in empty space. The p.d. across a resistor (or, as it is sometimes called, the voltage drop, or just " drop ") is, as we know, equal to its resistance multiplied by the current through it; and the positive end is that at which the current enters (or electrons leave). The positive end of a source of e.m.f., on the other hand, is that at which current would leave if its terminals were joined by a passive conductor. That is because an e.m.f. has the unique ability of being able to drive a current against a p.d.

Note that the positive terminal of a source of e.m.f. is not always the one at which current actually *does* leave; there may be a greater e.m.f. opposing it in the circuit.

It will be worth while to make quite sure that these questions of direction of e.m.f., p.d. and current are clearly understood, or there will be endless difficulty later on. So let us consider the imaginary circuit shown in Fig. 2.12*a*. Assume each cell has an e.m.f. of 2 V.

All five cells in the battery $B_1$ are connected in series and all tend to drive current clockwise, so the total clockwise e.m.f. of $B_1$ is $5 \times 2 = 10$ V, and we can show this as in Fig. 2.12*b*.

$B_2$ seems to be a flat contradiction of Kirchhoff's law. Here we have an e.m.f. of 2 V between two points which must be at the same potential because they are joined by a wire of negligible resistance! The cell is, as we say, *short-circuited*, or " dead-shorted ". Part of the answer to this apparent impossibility is that no real source of e.m.f. can be entirely devoid of resistance. Often it is not enough to matter, so is not shown in diagrams. But where the source is short-circuited it is vital, because it is the only thing to limit the current. To represent a source of e.m.f. completely, then, we must include between its terminals not only a symbol for e.m.f. but also one for resistance. When a source of e.m.f. is open-circuited, so that no current flows, there is no voltage drop in its internal resistance, so the terminal voltage is equal to the e.m.f. When current is drawn the terminal voltage falls, and if the source is dead-shorted the internal drop must be equal to the e.m.f., as shown in Fig. 2.12*b*. The other part of the answer is that when the source has a very low resistance (e.g., a large accumulator

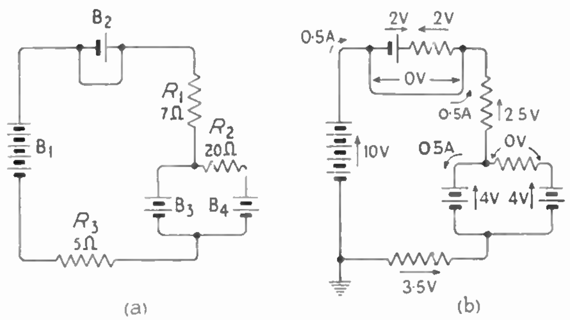



**Fig. 2.12—(*a*) is an imaginary circuit for illustrating the meanings of e.m.f. and p.d. The voltages and currents in different parts of it are marked in (*b*). Arrows alongside voltages show the direction of *rise* of potential**

or generator) such an enormous current flows that even a wire of normally negligible resistance does not reduce the terminal voltage to zero, and the wire is burnt out, and perhaps the source too.

In our example we shall assume that all the resistance is internal, so the external effect of $B_2$ is nil.

If the loop containing $B_3$, $B_4$, and $R_2$ is considered on its own as a series circuit, we see two e.m.fs each of 4 V in opposition, so the net e.m.f. is zero, and no current due to them flows through $R_2$. As regards the main circuit, $B_3$ and $B_4$ are in parallel, and contribute an e.m.f. of $-$ 4 V (because in opposition to $B_1$, which we considered positive). This can be indicated either by " $-$ 4 V " and a clockwise arrow, or by " 4 V " and an anticlockwise arrow as shown.

We see, then, that putting a second battery in parallel does not increase the main circuit e.m.f., but it does reduce the internal resistance by putting two in parallel, so is sometimes done for that reason. In our example the 20 Ω in series with $B_4$ is probably far greater than the internal resistance of $B_3$, so practically all the current will flow through $B_3$, and the effect of $B_4$ will be negligible.

We have, then (ignoring $B_2$), $10 - 4 = 6$ V net e.m.f. in the circuit, and (neglecting the internal resistances of $B_1$ and $B_3$) a total resistance of $7 + 5 = 12\ \Omega$. The current round the circuit is therefore $6/12 = 0{\cdot}5$ A, as marked, and the voltage drops across $R_1$ and $R_3$ are $0{\cdot}5 \times 7 = 3{\cdot}5$ V and $0{\cdot}5 \times 5 = 2{\cdot}5$ V respectively.

To help maintain the distinction between e.m.f. and p.d. it is usual to reserve the symbol $E$ for e.m.f., and to use $V$ to mean p.d.

Note that the potential difference between two points is something quite definite, but the potential of a point (like the height of a hill) is relative, depending on what one takes as the reference level. Heights are usually assumed to be relative to sea level, unless it is obvious that some other " zero " is used (such as street level for the height of a building). The potential of the earth is the corresponding zero for electrical potentials, where no other is implied. To ensure definiteness of potential, apparatus is often connected to earth, as indicated by the symbol at the negative end of $B_1$ in Fig. 2.12*b*.

## Electrical Effects

It is time now to consider what electricity can *do*. One of the most familiar effects of an electric current is heat. It is as if the jostling of the electrons in the conductor caused a certain amount of friction. But whereas it makes the wire in an electric fire red hot, and in an electric light bulb white hot, it seems to have no appreciable effect on the flex and other parts of the circuit—in a reputable installation, anyway. Reasoning of the type we used in connection with Fig. 2.11 shows that the rate at which heat is generated in a conductor containing no e.m.fs is proportional to the product of the current flowing through it and the e.m.f. applied; in symbols, $EI$. Since both $E$ and $I$ are related to the resistance of the conductor, we can substitute $IR$ for $E$ or $E/R$ for $I$, getting $I^2R$ or $E^2/R$ as alternative measures of the heating effect. Thus for a given current the heating is proportional to the resistance. Electric lamps and heaters are designed so that their resistance is far higher than that of the wires connecting them to the generator. $I^2R$ shows that the heating increases very rapidly as the current rises. Advantage is taken of this in fuses, which are short pieces of wire that melt (or " blow ") and cut off the current if, owing to a short-circuit somewhere, it rises dangerously above normal.

Another effect of an electric current is to magnetize the space around it. This effect is particularly marked when the wire is coiled round a piece of iron. As we shall see, it is of far more significance than merely being a handy way of making magnets.

An effect that need not concern us much is the production of chemical changes, especially when the current is passed through watery liquids. A practical example is the charging of accumulators. The use of the word " charging " for this process is unfortunate, for no surplus or deficiency of electrons is accumulated by it.

All these three effects are reversible; that is to say, they can be turned into causes, the effect in each case being an e.m.f. The production of an e.m.f. by heating the junction of two different metals is of only minor importance, but magnetism is the basis of all electrical generating machinery, and chemical changes are very useful on a smaller scale in batteries.

## Instruments for Measuring Electricity

All three effects can be used for indicating the strength of currents; but although hot-wire ammeters are occasionally used in radio senders, the vast majority of current meters are based on the magnetic effect. There are two main kinds: the moving-iron instrument, in which the current coil is fixed, and a small piece of iron with a pointer attached moves to an extent depending on the amperage; and the moving-coil (generally preferred) in which the coil is deflected by a fixed permanent magnet.

Besides these current effects there is the potential effect—the attraction between two bodies at different potential. The force of attraction is seldom enough to be noticeable, but if the p.d. is not too small it can be measured by a delicate instrument. This is the principle of what is called the electrostatic voltmeter, in which a rotatable set of small metal vanes is held apart from an interleaving fixed set by a hairspring. When a voltage is applied between the two sets the resulting attraction turns the moving set through an angle depending on the voltage. This type of instrument, which is a true voltmeter because it depends directly on potential and draws no current, is practical and convenient for full-scale readings from about 1,000 V to 10,000 V, but not for low voltages. (" Full-scale reading " means the reading when the pointer is deflected to the far end of the scale.)

Since $V$ or $E = IR$, voltage can be measured indirectly by measuring the current passed through a known resistance. If this additional conducting path were to draw a substantial current from the source of voltage to be measured, it might give misleading results by lowering that voltage appreciably; so $R$ is made relatively large. The current meter (usually of the moving-coil type) is therefore a milliammeter, or better still a microammeter. For example, a voltmeter for measuring up to 1 V could be made from a milliammeter reading up to 1 mA by adding sufficient resistance to that of the moving coil to bring the total up to 1,000 Ω. The same instrument could be adapted to read up to 10 V (all the scale readings being multiplied by 10) by adding another resistance, called a *multiplier*, of 9,000 Ω, bringing the total to 10,000 Ω. Such an arrangement, shown in Fig. 2.13*a*, can obviously be extended.

The same milliammeter can of course be used for measuring currents up to 1 mA by connecting it in series in the current path; but to avoid increasing the resistance of the circuit more than can be helped the voltage resistances would have to be short-circuited or otherwise cut out. Care must be taken never to connect a current meter directly across a voltage source, for its low resistance might pass sufficient current to destroy it. Higher currents can be measured by diverting all but a known fraction of the current through a by-pass resistance called a *shunt*. If, for example, the resistance of the 0–1 milliammeter is 75 Ω, shunting it with 75/9 = 8·33 Ω would cause nine times as much current to flow through the shunt as passes through the meter, thereby multiplying the range by 10 (Fig. 2.13*b*).

Thus a single moving-coil instrument can be made to cover many ranges of current and voltage measurement, simply by connecting appropriate shunts and multipliers. These are, in effect, known resistances used to reduce the current through the instrument to any

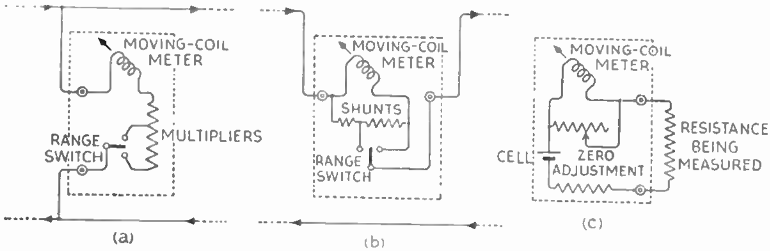



**Fig. 2.13—Showing how a single moving-coil indicating instrument can be used for measuring (*a*) voltage, (*b*) current, (*c*) resistance. The moving coil and pointer attached are shown diagrammatically**

desired fraction. But the operation can be reversed by incorporating a battery sufficient to give full-scale deflection; then, when an *unknown* resistance is connected, either as a shunt or multiplier, according to whether it is small or large, the extent to which the reading is reduced depends on the value of that resistance. It is therefore possible to graduate the instrument in ohms and we have an *ohmmeter* (Fig. 2.13*c*).

## Electrical Power

When everyday words have been adapted for scientific purposes by giving them precise and restricted meanings, they are more likely to be misunderstood than words that were specially invented for use as scientific terms—" electron ", for example. *Power* is a word of the first kind. It is commonly used to mean force, or ability, or authority. In technical language it has only one meaning: rate of doing work. *Work* here is also a technical term, confined to purely physical activity such as lifting weights or forcing things to move against pressure or friction—or potential difference. If by exerting muscular force you lift a 10 lb weight 3 ft against the opposing force of gravity you do 30 foot-pounds of work. And if you take a second to do it, your rate of doing work—your power output—is 30 ft-lbs per second, which is rather more than 1/20th of a horse-power. Correspondingly, if the electromotive force of a battery raises the potential of 10 electrons by 3 volts (i.e., causes them to flow against an opposing p.d. of 3 V) it does 30 electron-volts of work; and if it takes one second to do it the output of power is

30 volt-electrons per second. The electron per second is, as we saw, an extremely small unit of electric current, and it is customary to use one more than $6 \times 10^{18}$ times larger—the ampere. So the natural choice for the electrical unit of power might be called the volt-ampere. Actually, for brevity (and another reason which appears in Chapter 6), it has been given the name *watt*.

We now have the following to add to the table which was given on p. 32:

| *Quantity* | *Symbol for Quantity* | *Unit of Quantity* | *Symbol for Unit* |
|---|---|---|---|
| Electrical power | $P$ | Watt | W |

We also have the equation $P = EI$. So when your 3 V torch battery is lighting a 0·2 A bulb it is working at the rate of $0{\cdot}2 \times 3 = 0{\cdot}6$ W. In this case the battery e.m.f. is occupied in forcing the current through the resistance of the filament in the bulb, and the resulting heat is visible. We have already noted that the rate at which heat is produced in a resistance is proportional to $EI$—the product of the e.m.f. applied and the current flowing—and now we know this to be a measure of the power expended.

Alternative forms of the power relationship can be found as on p. 44:

$$P = I^2R \quad \text{and} \quad P = \frac{E^2}{R}.$$

Thus if any two of the four quantities $E$, $I$, $R$, and $P$ are given, all are known. For example, if a 1,000 Ω resistor is to be used in a circuit where 50 V will be maintained across it, the power dissipated as heat will be $50^2/1{,}000 = 2{\cdot}5$ W. In choosing such a resistor one must take care not only that the resistance is correct but also that it is large enough to get rid of heat at the rate equivalent to 2·5 W without reaching such a high temperature as to damage itself or things near it.

To familiarize oneself with power calculations it would be a good exercise to add another column to the table on p. 39, headed " Power (watts) ", working out the figures for each resistor and checking that they are the same whether derived from $EI$, $I^2R$, or $E^2/R$. A further check is to add up all the wattages for $R_1$ to $R_5$ and see that the total is the same as that for $R$.

It is often necessary to be able to tell whether a certain part of a circuit is acting as a giver (or generator, or source) of power, or as a receiver (or load). The test is whether the current is going through it in the direction of rising potential, in which case power is being generated there, or of falling potential, in which case power is being delivered. In Fig. 2.12 the current is going through $B_1$ from $-$ to $+$, so $B_1$ is a generator, as one would expect of a battery.

But it is going through $B_3$ from + to −, so although in a different circuit $B_3$ could be a generator, in this one it is a load, into which power is being delivered, perhaps for charging it. Through *any* resistance alone, current can only flow from + to −, for a resistance is essentially a user-up of power.

## A Broader View of Resistance

It will be found later on in this book that there are other ways in which electrical power can be " lost " than by heating up the circuit through which the current flows. This leads to a rather broader view of resistance; instead of defining it as $E/R$, resulting from an experiment such as that connected with Fig. 2.3, one defines it as $P/I^2$ (derived from $P = I^2R$), $P$ being the expenditure of power in the part of the circuit concerned, and $I$ the current through it. The snag is that $P$ is often difficult to measure. But as a definition to cover the various sorts of " resistance " encountered in radio it is very useful.

A feature of electrical power expended in resistance—however defined—is that it leaves the circuit for good. Except by some indirect method it is not possible to recover any of it in the form of electrical power. But there are ways in which electrical power can be employed to create a store of energy that can be released again as electrical power. *Energy* here is yet another technical term, meaning the amount of work that such a store can do. It can be reckoned in the same units as work, and so power can be defined alternatively as rate of release of energy. To return to our mechanical analogy, 30 ft-lbs of energy expended by exerting a force of 10 lbs to push a box 3 ft across the floor is all dissipated as heat caused by friction.* But 30 ft-lbs used in raising a 10 lb weight 3 ft is stored as what is called potential energy. If the weight is allowed to descend, it delivers up the 30 ft-lbs of energy, which could be used to drive a grandfather clock for a week. Another way of storing the energy would be to push a heavy truck mounted on ball bearings along rails. In this case very little of the pressure would be needed to overcome the small amount of friction; most of it would have the effect of giving it momentum which would keep it going long after the 3 ft push had finished. A heavy truck in motion is capable of doing work because the force setting it in motion has been used to store energy in it; this kind is called kinetic energy.

In the next chapter we shall begin the story of how electrical power can be stored and released in two ways, corresponding to potential and kinetic energy, and how this makes it possible to tune in to different stations on the radio.

* Electrical units are defined in terms of mechanical forces etc., in metric units. Because of the relationships between metric and British units, a power of 746 watts is equal to 550 ft-lbs per second (1 horse power). It is confirmed by experiment that the amount of heat is the same, whether produced by dissipation of 1 mechanical HP or 746 electrical watts.

CHAPTER 3

# Capacitance

## Charging Currents

Although most of the last chapter was about steady e.m.fs driving currents through resistances, you may remember that the whole thing began with the formation of an electric charge. We likened it to the deportation of unwilling citizens from town A to town B. Transferring a quantity of electrons from one place to another and leaving them there sets up a stress between the two places, which is only relieved when the same quantity of electrons has been allowed to flow back. The total amount of the stress is the p.d., measured in volts. And the place with the surplus of electrons is said to be negatively charged. The original method of charging—by friction—is inconvenient for most purposes and has been generally superseded by other methods. If an e.m.f. of, say, 100 V is used, electrons will flow until the p.d. builds up to 100 V; then the flow will cease because the charging and discharging forces are exactly equal.

It was at this stage that we provided a conductive path between the negatively and positively charged bodies, allowing the electrons to flow back just as fast as the constant e.m.f. drove them, so that a continuous steady current was set up. This state of affairs could be represented as in Fig. 3.1. Our attention then became attached to this continuous current and the circuit through which it flowed. The electron surplus at the negative end—and the deficit at the positive end—seemed to have no bearing on this, and dropped out of the picture. So long as e.m.f. and current are quite steady, the charges can be ignored.

But not during the process of charging. While that is going on there must be more electrons entering B than are leaving it; and vice versa during discharging. Assuming that there are parts of a circuit requiring a substantial number of electrons to charge them to

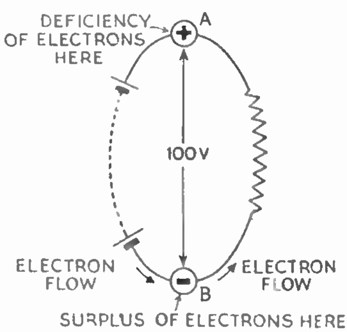



**Fig. 3.1—Before a p.d. can be established between any two parts of a circuit (such as A and B) they must be charged, and when the p.d. is rapidly varied the charging current may be important**

a potential equal to the applied e.m.f., then it is clear that the ordinary circuit principles we have been studying become more complicated whenever the e.m.f. varies. For one thing, it seems as if the current is no longer the same in all parts of a series circuit. The moment the e.m.f. is applied, a charging current starts to flow, over and above any current through conducting paths. If the e.m.f. is increased at any time, a further charging current flows to raise the p.d. correspondingly. Reducing the e.m.f. causes a discharging current. Seeing that both wire and wireless communication involve e.m.fs that are continually varying, it is evidently important to know how far the charging and discharging currents affect matters.

## Capacitance: What It Is

The first thing is to find what decides the number of electrons needed to charge any part of a circuit. The number of people that could be transferred to another town would probably depend on (*a*) the pressure brought to bear on them, (*b*) the size of the town, and (*c*) its remoteness. The number of electrons that are transferred as a charge certainly depends on the electrical pressure applied—the e.m.f. As one would expect, it is proportional to it. If the battery in Fig. 3.1 gave 200 V instead of 100, the surplus of electrons forced into B would be just twice as great. The quantity of electrons required to charge any part of a circuit to 1 V is called its *capacitance*.

We have already found that as a practical unit of electric current one electron-per-second would be ridiculously small, and the number that fits the metric system is rather more than $6 \times 10^{18}$ electrons per second (=1 A). So it is natural to take the same number as the unit of electric charge or *quantity* of electricity (symbol: $Q$). It is therefore equal to the number that passes a given point every second when a current of 1 A is flowing, and the name for it is the *coulomb*. So the obvious unit of capacitance is that which requires 1 coulomb to charge it to 1 volt. It is named the *farad* (abbreviation: F) in honour of Michael Faraday, who contributed so much to electrical science. In symbols, the relationship between electric charge, voltage, and capacitance is

$$Q = VC$$

which may be compared with the relationship between electric current, voltage, and conductance (p. 1). $V$ is used instead of $E$ because fundamentally it is the p.d. that is involved rather than the charging e.m.f. If the conducting path and e.m.f. in Fig. 3.1 were both removed, leaving A and B well insulated, their charge and the p.d. would remain.

As capacitance is equal to the charge required to set up a given potential *difference* between two parts of a circuit or other objects, one has to speak of the capacitance *between* those objects, or of one to another. And because the capacitance between, say, A and B in Fig. 3.1 is in parallel with the resistance it does not really make an exception to the rule that the current is the same throughout a series circuit.

## Capacitance Analysed

We saw that when any object is being charged the current going into it is greater than that coming out. So if Kirchhoff's law about currents is true in all circumstances, the object must be more than just a point; it must have some size. We would expect the capacitance of objects to one another to depend in some way on their size; the question is, what way?

When analysing resistance (p. 40) we found that it depended on the dimensions of the conductor and what it was made of. So, of course, does conductance, though in inverse ratio (p. 41). Capacitance, however, has to do with the space between the conductors; it results from the force of attraction that can be set up in that space, or from what is called the electric field. Suppose A and B in Fig. 3.2 are our oppositely-charged terminals. The directions along which the force is felt are indicated by dotted lines, called *lines of electric force*. These show the paths electrons would take if they were free to move slowly under the force of attraction to +. (But to agree with the convention for direction of electric currents the arrows point in the opposite direction.) To represent a more intense field, lines are drawn more closely together. Each can be imagined to start from a unit of positive charge and end on a unit of negative charge. It must be understood that it is a matter of convenience what size of unit is taken for this purpose; and that the lines are not meant to convey the idea that the field itself is in the form of lines, with no field in the spaces between.

The fact that the space between two small widely-separate objects such as these is not confined to a uniform path makes the capacitance between them a more difficult thing to calculate than the resistance of a wire. But consider two large flat parallel plates, shown edge-on in Fig. 3.3*a*. Except for a slight leakage around the edges, the field is confined to the space directly between the plates, and this space, where the field is uniformly distributed, is a rectangular slab.

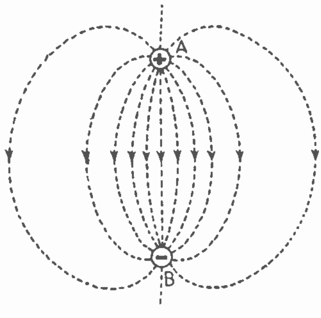



**Fig. 3.2—The dotted lines indicate the direction and relative intensity of electric field between two oppositely-charged terminals, A and B**

Suppose, for the sake of example, the plates are charged to a p.d. of 100 V. Since the field is uniform, the potential changes at a uniform rate between one plate and the other. Reckoning the potential of the lower plate as zero, the potential half way between the plates is therefore + 50 V. A third plate could be inserted here, and we would then have two capacitances in series, each with a p.d. of 50 V. Provided that the thickness of the third plate was

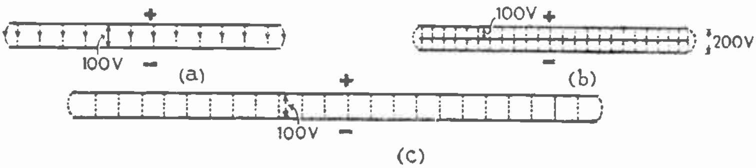



**Fig. 3.3—Consideration of the electric field between closely-spaced parallel plates leads to the relationship between dimensions and capacitance**

negligible, its presence would make no difference to the amount of the charge required to maintain 100 V across the outer plates. Each of the two capacitances in series would therefore have the same charge but only half the voltage. To raise their p.d. to 100 V each, the amount of charge would have to be doubled (Fig. 3.3*b*). In other words (remembering that $C = Q/V$) halving the spacing doubles the capacitance. In general, $C$ is inversely proportional to the distance between the plates (which we can denote by $t$).

Next, imagine a second pair of plates, exactly the same as those in Fig. 3.3*a*. Obviously they would require the same quantity of electricity to charge them to 100 V. If the two pairs were joined together as in Fig. 3.3*c*, they would form one unit having twice the cross-section area and requiring twice the charge to set up a given p.d.; in other words, twice the capacitance. In general, $C$ is directly proportional to the cross-section area of the space between the plates (which we can denote by $A$).

So the capacitance between two plates separated by a rectangular slab of space in which the electric field is uniform depends on the dimensions of the slab in the same way as does the conductance of a piece of material (p. 41):

$$C = \frac{\epsilon A}{t}$$

Here $\epsilon$ ("epsilon") is the "constant", corresponding to conductivity, needed to link the chosen system of units with the experimentally observed facts. Formerly the symbol $\kappa$ was used. One would expect it to be called "capacitivity", but actually the name is *absolute permittivity*. When, as in the m.k.s. system of units (p. 40), $C$ is reckoned in farads, $t$ in metres, and $A$ in square metres, and the space is completely empty, $\epsilon$ turns out to be $8{\cdot}854 \times 10^{-12}$.

This particular value is sometimes called the "electric space constant" and is given the distinctive symbol $\epsilon_0$.

Filling the space with air makes hardly any difference; but if you use solid or liquid materials such as glass, paper, plastics or oil, the capacitance is increased. You may remember that a feature of insulators (p. 31) is that their electrons are unable to drift through the material in large numbers but that they "strain at the leash" under the influence of an electric field. When the field is set up—by applying an e.m.f.—this small shift of the electrons all in one direction is equivalent to a momentary current, just like a charging current. Similarly their return to normal position in the atoms is equivalent to a discharging current. The total charging and discharging currents are therefore greater than when the space is unoccupied. The amount by which the capacitance is multiplied in this way by filling the whole of the space with such a material is called the *relative permittivity* of the material, or often just *permittivity* (symbol: $\epsilon_r$).

Insulating materials in this role are called *dielectrics*, and another name for relative permittivity is *dielectric constant*. For most solids and liquids it lies between 2 and 10, but for special materials it may be as much as several thousands.

Just as the embodiment of resistance is a resistor, a circuit element designed for its capacitance is called a *capacitor*; the older name *condenser* is also used. Usually it consists of two or more parallel plates spaced so closely that most of the field is directly between them, as in Fig. 3.3.

The capacitance of a capacitor thus depends on three factors: area and thickness of the space between the plates, and permittivity of any material there. For convenience, the permittivity specified for materials is the relative value, $\epsilon_r$; and since $\epsilon$ is $\epsilon_r$ times $\epsilon_0$ the formula is usually adapted for $\epsilon_r$ by filling in the value of $\epsilon_0$:

$$C = \frac{8{\cdot}854 \times 10^{-12}\,\epsilon_r\,A}{t}$$

This practice is so common that the little r is usually omitted, "$\epsilon$" being understood to be the relative value.

Supposing, for example, the dielectric had a permittivity of 5 and was 0·1 cm thick, the area needed to provide a capacitance of 1 F would be about 9 square miles. In practice the farad is far too large a unit, so is divided by a million into microfarads (μF), or even by $10^{12}$ into picofarads (pF), sometimes called micro-microfarads (μμF). Adapting the above formula to the most convenient units for practical purposes, we have:

$$C = \frac{\epsilon A}{11{\cdot}3t} \text{ picofarads,}$$

the dimensions being in *centimetres*.

## CAPACITORS

The general symbol for a capacitor in circuit diagrams is Fig. 3.4*a*. The capacitors themselves appear in great variety in radio circuits, according to the required capacitance and other circumstances. Some actually do consist of a single sheet of insulating material such as mica sandwiched between a pair of metal plates, but sometimes there are a number of plates on each side as indicated in Fig. 3.4*b*. The effective area of one dielectric is of course multiplied by their number—in this case four. Fig. 3.4*b* was at one time used in circuit diagrams to indicate a large capacitance.

A very common form of capacitor consists of two long strips of tinfoil separated by waxed paper and rolled up into a compact block. Above one or two microfarads it is often more economical to use aluminium plates in a chemical solution or paste which causes an extremely thin insulating film to form on one of them. Apart from this film-forming, the solution also acts as a conductor between the film and the other plate. These *electrolytic* capacitors

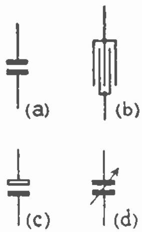



**Fig. 3.4—Standard symbols for (*a*) fixed capacitor, (*c*) the same (electrolytic), in which the white plate is positive, and (*d*) variable capacitor. The symbol (*b*) denoting one of large capacitance is generally obsolete**

must always be connected and used so that the terminal marked + never becomes negative relative to the other. Even when correctly used there is always a small leakage current. To help the wireman, the symbol Fig. 3.4*c*, in which the + plate is distinguished by outlining, is often used for electrolytic types.

Air is seldom used as the dielectric for fixed capacitors, but is almost universal in variable capacitors. These, indicated in diagrams as Fig. 32*d*, consist of two sets of metal vanes which can be progressively interleaved with one another, so increasing the effective area, to obtain any desired capacitance up to the maximum available, seldom more than 500 pF.

Besides the capacitance, an important thing to know about a capacitor is its maximum working voltage. If the intensity of electric field in a dielectric is increased beyond a certain limit the elastic leashes tethering the electrons snap under the strain, and the freed electrons rush uncontrollably to the positive plate, just as if the dielectric were a good conductor. It has, in fact, been broken down or punctured by the excessive voltage. For a given thickness, mica stands an exceptionally high voltage, or, as one says, has a relatively high dielectric strength. Air is less good, but has the advantage that the sparking across resulting from breakdown does it no permanent harm.

The highest voltage that could be safely applied to a variable capacitor would not cause nearly enough attraction between the fixed and moving vanes to make them turn into the maximum capacitance position; but an electrostatic voltmeter (p. 45) can be looked upon as a very small variable capacitor with its moving vanes so delicately suspended that they do move under the attractive force and indicate the voltage.

In our concentration on intentional capacitance between plates, we should not forget that, whether we like it or not, every part of a circuit has capacitance to surrounding parts. When the circuit e.m.fs are varying rapidly, these stray capacitances may be very important.

## Charge and Discharge of a Capacitor

It is instructive to consider the charging process in greater detail. In Fig. 3.5 the capacitor can be charged by moving the switch to A, connecting it across a battery; and discharged by switching to B. $R$ might represent the resistance of the wires and capacitor plates,

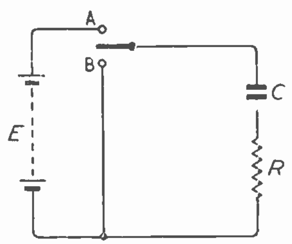



**Fig. 3.5—Circuit used for obtaining the charge and discharge curves shown in Fig. 3.6**

but as that is generally very small it will be easier to discuss the matter if we assume we have added some resistance, enough to bring the total up to, say, 200 Ω. Suppose the capacitance $C$ is 5 μF and the battery e.m.f. 100 V.

At the exact moment of switching to A the capacitor is as yet uncharged, so there can be no voltage across it; the whole of the 100 V e.m.f. of the battery must therefore be occupied in driving current through $R$, and by Ohm's law that current is found to be 0·5 A (extreme left of Fig. 3.6*a*). Reckoning the potential of the negative end of the battery as zero, we have at this stage the positive end of the battery, the switch, the upper plate of the capacitor, and (because the capacitor has no p.d. across it) the lower plate also, all at + 100 V. The lower end of the resistor is zero, so we have 100 V drop across the resistor, as already stated. Note that *immediately* the switch is closed the potential of the lower plate of the capacitor as well as that of the upper plate jumps from zero to + 100. The use of a capacitor to transfer a sudden change of potential to another point, without a conducting path, is very common.

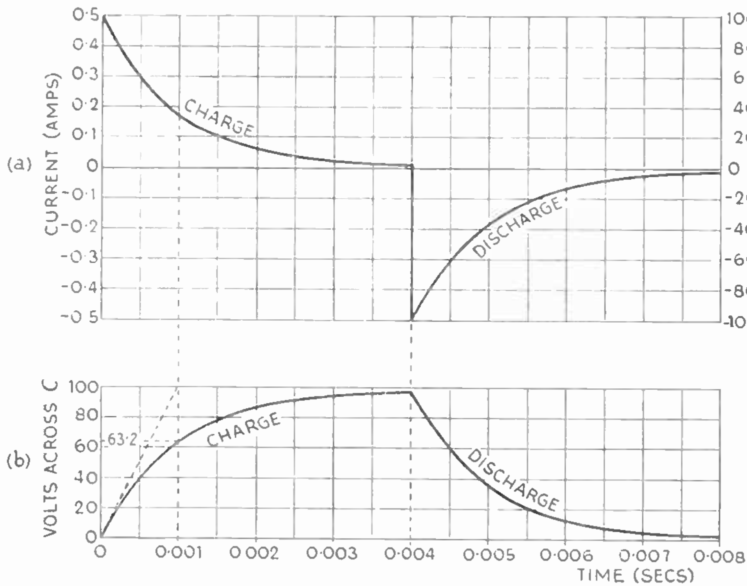



**Fig. 3.6—These curves show the current and voltage in the circuit of Fig. 3.5 during charge and discharge**

We have already seen that the number of coulombs required to charge a capacitance of $C$ farads to $V$ volts is $VC$. In this case, it is $100 \times 5 \times 10^{-6}$, which is 0·0005. As a coulomb is an ampere-second, the present charging rate of 0·5 A, *if maintained*, would fully charge the capacitor in 0·001 sec. The capacitor voltage would rise steadily as shown by the sloping dotted line in Fig. 3.6*b*. Directly it starts to do so, however, there are fewer volts available for driving current through $R$, and so the current becomes less and the capacitor charges more slowly. When it is charged to 50 V, 50 V are left for the resistor, and the charging rate is 0·25 A. When $C$ is charged to 80 V, 20 are left for R, and the current is 0·1 A. And so on, as shown by the curves in Fig. 3.6.

Curves of this type are known as *exponential*, and are characteristic of many growth and decay phenomena. Note that the current curve's downward slope, showing the rate at which the current is decreasing, is at every point proportional to the current itself. It can be shown mathematically that in the time a capacitor would take to charge at the starting rate it actually reaches 63·2% of its full charge. By using the principles we already know we have calculated that this time in the present example is 0·001 sec. Try working it out with letter symbols instead of numbers, so that it covers all charging circuits, and you should find that it is $CR$ secs, regardless of the voltage $E$. $CR$—equal to the time taken to reach

63·2% of final charge—is of course a characteristic of the circuit, and is called its *time constant*.

Theoretically, the capacitor is never completely charged, because the charging depends on there being a difference between the applied e.m.f. and the capacitor voltage, to drive current through the resistance. If the current scale is multiplied by 200 (the value of $R$), the upper curve is a graph of this voltage across $R$. Note that the voltages across $C$ and $R$ at all times up to 0·004 sec add up to 100, just as Kirchhoff says they ought (p. 41).

For practical purposes the capacitor may be as good as fully charged in a very short time. Having allowed 0·004 sec for this to happen in the present case, we move the switch to B. Here the applied voltage is zero, so the voltages across $C$ and $R$ must now add up to zero. We know that the capacitor voltage is practically 100, so the voltage across $R$ must be $-$ 100, and the current $-$ 0·5 A; that is to say, it is flowing in the opposite direction, discharging the capacitor. And so we get the curves in Fig. 3.6 from 0·004 sec onwards.

## Where the Power Goes

The height of the charging curve above the zero line in Fig. 3.6*a* represents the current supplied by the battery. Multiplied by its e.m.f. (100 V) it represents the power. When first switched on, the battery is working at the rate of $0{\cdot}5 \times 100 = 50$ watts, but at the end of 0·004 sec this has fallen off almost to nothing. At first the whole 50 watts goes into the resistance, so all its energy is lost as heat. But the effort of the e.m.f. immediately becomes divided, part being required to drive the current through $R$ and the rest to charge $C$. It is clear that energy is going into the capacitance (p. 47); what happens to it? The answer is that it is stored as an electric field, in a similar way to the storage of mechanical energy in a spring when it is compressed by an applied force. During the charging process as a whole, half the release of energy from the battery is lost as heat in the resistor, and half is stored in the capacitor.

When we come to the discharge we find that the voltage across $R$ and the current through $R$ follow exactly the same curves as during the charge, except that they are both reversed. But multiplying two negatives together gives a positive, so the energy expended in $R$ is the same amount as during charge. The capacitor voltage and current during the discharge are the same as for the resistor, so the amount of energy is equal; in fact, it is the same energy, for what goes into the resistor has to come out of the capacitor. Applying the test described on p. 47, we can check that the capacitor is a source of energy; the current that is going into the resistor at its positive terminal is coming out from the capacitor's positive terminal.

Summing up: during the charging period half the energy supplied by the battery is stored as an electric field in the capacitor and half is dissipated as heat in the resistor; during discharge the half that was stored is given back at the expense of the collapsing field and is dissipated in the resistor.

It may be as well to repeat the warning that the " charging " and " discharging " of accumulator batteries is an entirely different process, accompanied by only minor variations in voltage. The electricity put into the battery is not stored as an electric charge; it causes a reversible chemical change in the plates.

A point to remember is that the p.ds set up across $R$ at the instants of switching to A and B are short-lived; some idea of their duration in any particular case can be found by multiplying $C$ by $R$ in farads and ohms respectively, giving the answer in seconds. These temporary changes are called *transients*.

CHAPTER 4

# Inductance

## Magnets and Electromagnets

If a piece of paper is laid on a straight " bar " magnet, and iron filings are sprinkled on the paper, they are seen to arrange themselves in a pattern something like Fig. 4.1*a*. The lines show the paths along which the attraction of the magnet exerts itself. (Compare the lines of electric force in Fig. 3.2.) As a whole, they map out the *magnetic field*, which is the " sphere of influence ", as it were, of the magnet. The field is most concentrated around two regions, called *poles*,* at the ends of the bar. The lines may be supposed to continue right through the magnet, emerging at the end marked N and returning at S. This direction, indicated by the arrows, is (like the direction of an electric current) purely conventional, and the lines themselves are an imaginary representation of a condition occupying the whole space around the magnet.

The same result can be obtained with a previously quite ordinary and unmagnetized piece of iron by passing an electric current through a wire coiled round the iron—an arrangement known as an *electromagnet*. It is not even necessary to have the iron core (as it is called); the coil alone, so long as it is carrying current, is interlinked with a magnetic field having the same general pattern as that due to a bar

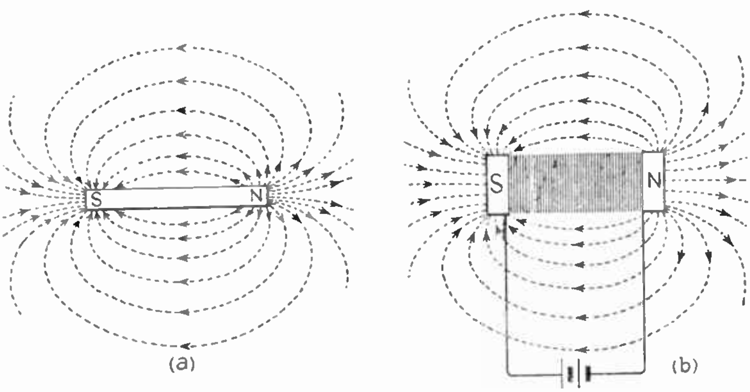



**Fig. 4.1—The dotted lines indicate the direction and distribution of magnetic field around (a) a permanent bar magnet, and (b) a coil of wire carrying current**

*The terminals of a battery or other voltage generator, between which electric field lines are imagined, are sometimes also called poles. The matter of which end of a magnet is N, or which terminal of a battery is +, is called its *polarity*.

magnet, as can be seen by comparing Fig. 4.1*a* and *b*. Without the iron core, however, it is considerably weaker. Finally, if the wire is unwrapped, every inch of it is still found to be surrounded by a magnetic field. Though very much less concentrated than in the coil, it can be demonstrated with filings as in Fig. 4.2, if the current is strong enough.

The results of these and other experiments in electromagnetism are expressed by saying that wherever an electric current flows it surrounds itself with a field which sets up magnetic *flux* (symbol: $\Phi$, pronounced " fie "). This flux and the circuit link one another like

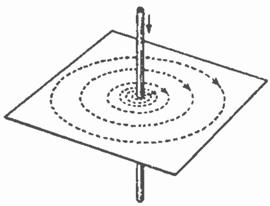


**Fig. 4.2—Conventional direction of magnetic field around a straight wire carrying current**

two adjoining links in a chain. By winding the circuit into a compact coil, the field due to each turn of wire is made to operate in the same space, producing a concentrated magnetic flux (Fig. 4.1*b*). Finally, certain materials, notably iron and its alloys, described as *ferromagnetic*, are found to have a very large multiplying effect on the flux, this property being called the *relative permeability* of the material, denoted by the symbol $\mu$. The permeabilities of most things, such as air, water, wood, brass, etc., are all practically the same and are therefore generally reckoned as 1. Certain iron alloys have permeabilities running into many thousands. Unlike resistivity and permittivity, however, the permeability of ferromagnetic materials is not even approximately constant but varies greatly according to the flux density (i.e., the flux passing through unit area at right angles to its direction).

It is now known that all magnets are really electromagnets. Those that seem not to be (called *permanent magnets*) are magnetized by the movements of the electrons in their own atomic structure. The term " electromagnet " is however understood to refer to one in which the magnetizing currents flow through an external circuit.

### Interacting Magnetic Fields

The " N " and " S " in Fig. 4.1 stand for North-seeking and South-seeking respectively. Everybody knows that a compass needle points to the north. The needle is a magnet, and turns because its own field interacts with the field of the earth, which is another magnet. Put differently, the north magnetic pole of the earth attracts the north-seeking pole of the needle, while the earth's south pole attracts its south-seeking pole. The poles of a magnet are often called, for brevity, just north and south, but strictly, except in referring to the earth itself, this is incorrect. By bringing the

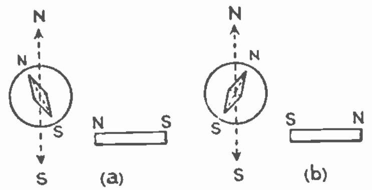



**Fig. 4.3—Deflection of a compass needle by a magnet, proving that like poles repel and unlike poles attract**

two poles of a bar magnet—previously identified by suspending it as a compass—in turn towards a compass needle it is very easy to demonstrate, as in Fig. 4.3, that unlike poles attract one another. This reminds one of the way electric charges behave.

Now suppose we hang a coil in such a way that it is free to turn, as suggested in Fig. 4.4. So long as no current is passing through the coil it will show no tendency to set itself in any particular direction, but if a battery is connected to it the flow of current will transform the coil into a magnet. Like the compass needle, it will then indicate the north, turning itself so that the plane in which the turns of the coil lie is east and west, the axis of the coil pointing north. If now the current is reversed the coil will turn through 180 degrees, showing that what was the N pole of the coil is now, with the current flowing the opposite way, the S.

The earth's field is weak, so the force operating to turn the coil is small. When it is desired to make mechanical use of the magnetic effect of the current in a coil it is better to provide an artificial field of the greatest possible intensity by placing a powerful magnet as close to the coil as possible. This is the basis of all electric motors.

In radio, however, we are more interested in other applications of the same principle, such as the moving-coil measuring instruments described very briefly on p. 45. The coil is shaped somewhat as in Fig. 4.4, and is suspended in the intense field of a permanent magnet. Hairsprings at each end of the coil serve the double purpose of conducting the current to the coil and keeping the coil normally in the position where the pointer attached to it indicates zero. When current flows through it the coil tends to rotate against the restraint of the springs, and the angle through which it moves

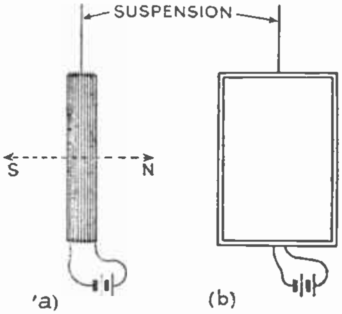



**Fig. 4.4—A coil carrying current, and free to rotate, sets itself with its axis pointing N and S. *a* is a side view and *b* an end view**

is proportional to the magnetic interaction, which in turn is proportional to the value of the current.

## INDUCTION

On the last page but one we noted the fact that the magnetic circuit formed by the imaginary flux lines is linked with the electric circuit that causes them, like two adjoining links in a chain (Fig. 4.5). Now suppose the electric circuit includes no source of e.m.f. and

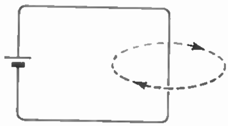


**Fig. 4.5—An electric current and the magnetic flux set up thereby are linked with one another as indicated in this diagram, which shows the relative directions established by convention**

consequently no current, but that some magnetic flux is made to link with it. There are various ways in which this could be done: a permanent magnet or an electromagnet—or any current-carrying circuit—could be moved towards it, or current could be switched on in a stationary circuit close to it. Whatever the method, the result would be the same: so long as the flux linked with the circuit was

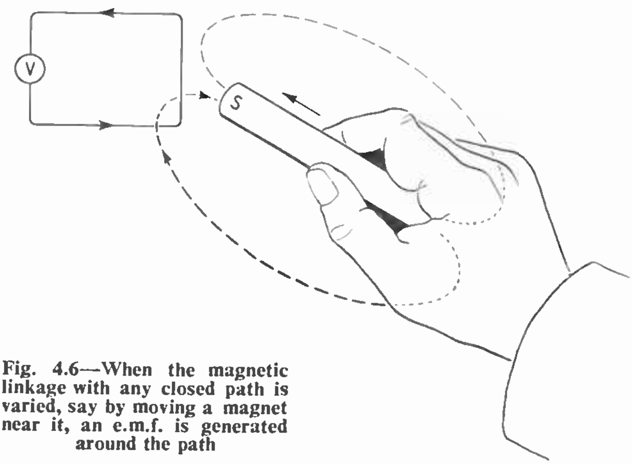



**Fig. 4.6—When the magnetic linkage with any closed path is varied, say by moving a magnet near it, an e.m.f. is generated around the path**

increasing, an e.m.f. would be generated in the circuit. The direction and magnitude of this *induced* e.m.f. can be found by means of a suitable voltmeter, as in Fig. 4.6; the e.m.f. is proportional to the rate at which the amount of flux linked with the circuit is changing. So a constant flux linkage yields no e.m.f. When the flux is removed or decreased, a reverse e.m.f. is induced. If the circuit were closed, the e.m.f. would make a current flow around it.

This result, first discovered by Michael Faraday in 1831, is the basis of all electrical engineering. The electrical generators in power stations are devices for continuously varying the magnetic flux linked with a circuit, usually by making electromagnets rotate past fixed coils. The alternative method of varying the flux linkages—by varying the current in adjacent fixed coils—is adopted in transformers (p. 69) and, as we shall see, in all kinds of radio equipment.

### Self-Inductance

So far we have been thinking of an e.m.f. being induced in a circuit by the varying of magnetic flux due to current in some other circuit. But in Fig. 4.5 we see a dotted loop representing the flux due to current flowing in that circuit itself. Before the current flowed there was no flux. So, directly the current was switched on, the flux linked with the circuit must have been increasing from zero to its present value, and in the process inducing an e.m.f. in its own circuit. This effect is—very naturally—known as *self-induction*. Since Fig. 4.5 can be taken to represent *any* circuit carrying current, and because a flow of current inevitably results in some linked magnetic flux, it follows that an e.m.f. is induced in any circuit in which the current is varying.

The amount of e.m.f. induced in a circuit (or part of a circuit) when the current is varied at some standard rate is called its *self-inductance*, or often just " inductance " (symbol: $L$). To line up with the units we already know, its unit, called the *henry* (abbreviation: H), was so defined that the inductance of a circuit in henries is equal to the number of volts self-induced in it when the current is changing at the rate of 1 ampere per second.

Actually it is the change of *flux* that induces the e.m.f., but since information about the flux is less likely to be available than that about the current, inductance is specified in terms of current. What it depends on, then, is the amount of flux produced by a given current, say 1 A. The more flux, the greater the e.m.f. induced when that current takes one second to grow or die, and so the greater the inductance. The flux produced by a given current depends, as we have seen, on the dimensions and arrangement of the circuit or part of circuit, and on the permeability.

### Inductance Analysed

We had no great difficulty in finding a formula expressing resistance in terms of resistivity and the dimensions of the circuit element concerned, because currents are usually confined to wires or other conductors of uniform cross-section. Capacitance was not so easy, because it depends on the dimensions of space between conductors, and the distribution of the electric field therein is likely to be anything but uniform, but for the particular and practically-important case of a thin space between parallel plates we found a formula for capacitance in terms of permittivity and the dimensions of the space. Inductance is even more difficult, because the magnetic fields of many of the circuit forms used for providing it, known in general as *inductors*, are not even approximately uniform or easily calculated. Another complication is that some of the flux may link with only part of the circuit. In coils, for example, the flux generated by one turn usually links with some of the others but not all.

However, Fig. 4.7 shows a type of inductor in which the coil is linked with an iron core of nearly uniform cross-section area. The permeability of the iron is usually so high that the core carries very nearly all the magnet flux; the remainder—called *leakage* flux—

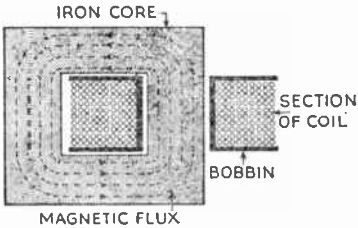



**Fig. 4.7—Showing how a closed iron core links with the turns of the coil and carries nearly all the flux set up by the coil**

that strays from it through the surrounding air or the material of the coil itself, can be neglected. So it is nearly true to assume that all the flux links all the turns, and that it is uniformly distributed throughout the whole core. On that assumption, the inductance of a coil is related to the dimensions in the same way as for conductance and capacitance:

$$L = \frac{\mu A}{l}$$

where $A$ is the cross-section area of the core, $l$ is its length, and $\mu$ is its absolute permeability. As with permittivity, the " absolute " value is the one that gives the answer in the desired units without the assistance of other numbers (" constants "). In m.k.s. units (p. 40), in which $L$ is in henries, $A$ in square metres and $l$ in metres, its value for empty space (sometimes called the " magnetic space constant "), denoted by $\mu_0$, is $1{\cdot}257 \times 10^{-6}$. The same value is practically correct for air and most other substances except the ferromagnetic materials. Although strictly the permeability ($\mu$) is this $\mu_0$ multiplied by the relative permeability $\mu_r$ (p. 60), the values quoted for materials are invariably relative, so the formula is usually adapted for them by incorporating $\mu_0$:

$$L = \frac{1{\cdot}257 \times 10^{-6}\,\mu_r\,A}{l}$$

And, as with permittivity, $\mu_r$ is then usually written " $\mu$ " and called just "permeability". For air, etc., it can be omitted altogether, being practically 1. If the dimensions are in centimetres, $10^{-8}$ must be substituted for $10^{-6}$.

The above equation is for only one turn. If there are $N$ turns the same current flows $N$ times around the area $A$, so the flux (assuming $\mu$ to be constant) is $N$ times as much. Moreover, this $N$-fold flux links the circuit $N$ times, inducing $N$ times as much e.m.f. when it varies as in one turn. So the inductance is $N^2$ times as great:

$$L = \frac{1{\cdot}257 \times 10^{-6}\, \mu_r\, AN^2}{l}$$

## Practical Considerations

In practice this formula is not of much use for calculating the inductance of even iron-cored coils. For one thing, in order to get the core into position around the coil it is necessary to have it in more than one piece, and the joint is never so magnetically perfect as continuous metal. Even a microscopic air gap makes an appreciable difference, because it is equivalent to an iron path perhaps tens of thousands of times as long. Then the permeability of iron depends very much on the flux density and therefore on the current in the coil. So one cannot look it up in a list of materials as one can resistivity or permittivity. It is usually found plotted as a graph against magnetizing force in ampere-turns per unit length of core.

With " air-core " coils, even when all the turns are close together it is hardly correct to assume that all the flux due to each turn links with the whole lot; and the error is of course much greater if the turns are widely distributed as in Fig 4.1*b*. An even greater difficulty is that the flux does not all follow a path of constant area $A$ and equal length $l$.

A general formula for inductance has nevertheless been given, to bring out the basic similarity in form to those for conductance and capacitance, but for practical purposes it is necessary either to measure inductance or calculate it from the various formulae and graphs that have been worked out for coils and straight wires of various shapes and published in radio reference books.

Between a short length of wire and an iron-cored coil with thousands of turns there is an enormously large range of inductance. The iron-cored coil might have an inductance of many henries; the short wire a small fraction of a microhenry. Yet the latter is not necessarily unimportant. The maximum current may be much less than 1 A, but if it comes and goes a thousand million times a second its rate of change may well be millions of amps per second and the induced voltage quite large.

It is, of course, this induced voltage that makes inductance significant. Just how significant will appear in later chapters, but we can make a start now by comparing and contrasting it with the voltage that builds up across a capacitance while it is being charged.

## Growth of Current in Inductive Circuit

Firstly, what is the direction of the induced voltage? There are two possibilities: it might be in such a direction as to assist or to oppose the change of current that caused it. If it were to assist, the current would change more rapidly, inducing a greater voltage, increasing the change of current still more, and there would be no limit to the thing. That is clearly absurd, and in fact the induced voltage always tends to *oppose* the change of current responsible for it.

What happens when current is switched on in an inductive circuit can be studied with the arrangement shown in Fig. 4.8, which should be compared with Fig. 3.5. To make comparison of the

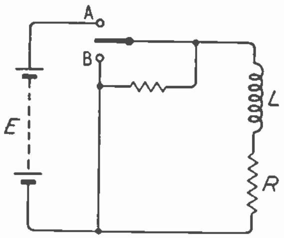



**Fig. 4.8—Circuit used for studying the rise and fall of current in an inductive circuit**

results easy, let us assume that the inductance, for which the curliness marked $L$ is the standard symbol, is 0·2 H, and that $R$ (which includes the resistance of the coil) is 200 Ω, and $E$ is 100 V as before.

With the switch as shown, no current is flowing through $L$ and there is no magnetic field. At the instant of switching to A, the full 100 V is applied across $L$ and $R$. The current cannot instantly leap to 0·5 A, the amount predicted by Ohm's law, for that would mean increasing the current at an infinite rate, which would induce an infinite voltage opposing it. So the current must rise gradually, and at the exact moment of closing the circuit it is zero. There is therefore no voltage drop across $R$, and the battery e.m.f. is opposed solely by the inductive e.m.f., which must be 100 V (Fig. 4.9*a*). That enables us to work out the rate at which the current will grow. If $L$ were 1 H, it would have to grow at 100 A per sec to induce 100 V. As it is only 0·2 H it must grow at 500 A per sec.

In the graph (Fig. 4.9*b*) corresponding to Fig. 3.6*b* for the capacitive circuit, the dotted line represents a steady current growth of 500 A per sec. If it kept this up, it would reach the full 0·5 A in 0·001 sec. But directly the current starts to flow it causes a voltage drop across $R$. And as the applied voltage remains 100, the induced voltage must diminish. The only way this can happen is for the current to grow less rapidly. By the time it has reached 0·25 A there are 50 V across $R$, therefore only 50 across $L$, so the current must be growing at half the rate. The full line shows how it grows; here is another exponential curve and, as in the capacitive circuit, it theoretically never quite finishes. In the time that the current would take to reach its full Ohm's law value if it kept up its starting pace, it actually reaches 63·2% of it. This time is, as before, called the time constant of the circuit. A little elementary algebra based on the foregoing shows it to be $L/R$ seconds.

The voltage across $L$ is also shown. Compared with Fig. 3.6, current and voltage have changed places. The voltage across $R$ is of course proportional to the current and therefore its curve is the same shape as Fig. 4.9*b*. Added to the upper, it must equal 100 V as long as that is the voltage applied.

When the current has reached practically its full value, we flick the switch instantaneously across to B. The low resistance across the contacts is merely to prevent the circuit from being completely interrupted in this process. At the moment the switch is operated, the magnetic field is still in existence. It can only cease when the current stops, and the moment it begins to do so an e.m.f. is induced which tends to keep it going. At first the full 0·5 A is going, which requires 100 V to drive it through $R$; and as the battery is no longer in circuit this voltage must come from $L$, by the current falling at the required rate: 500 A per sec. As the current wanes, so does the voltage across $R$, and so must the induced voltage, and therefore the current dies away more slowly, as shown in the continuation of Fig. 4.9. In $L/R$ (= 0·001) secs it has been reduced by the inevitable 63·2%.

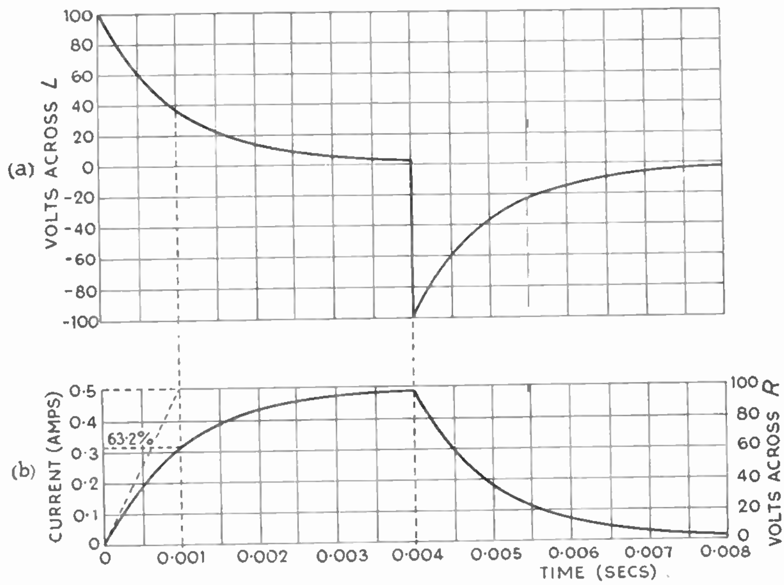



**Fig. 4.9—Curves showing the current and voltage in the circuit of Fig. 4.8 when the switch is moved first to A and then to B**

## Power During Growth

In reckoning how the power in the circuit varies during these operations, we note firstly that the input from the battery (calculated by multiplying the current at each instant by its 100 V) instead of starting off at 50 W and then tailing off almost to nothing, as in the capacitive example, starts at nothing and works up to nearly 50 W. If there had been no *C* in Fig. 3.5 there would have been no flow of power at any time; the effect of *C* was to make possible a temporary flow, during which half the energy was dissipated in *R* and half stored in *C*. In Fig. 4.8, on the other hand, absence of *L* would have meant power going into *R* at the full 50-watt rate all the time; the effect of *L* was to withhold part of that expenditure from *R* for a short time. Not all this energy withheld was a sheer saving to the battery, represented by the temporarily sub-normal current; part of it—actually a half—went into *L*, for the current had to be forced into motion through it against the voltage induced by the growing magnetic field. The energy was in fact stored in the field, in a similar way to the storage of mechanical energy when force is applied to set a heavy vehicle in motion.

Just as the vehicle gives up its stored energy when it is brought to a standstill, the result being the heating of the brakes, so the magnetic field in *L* gave up its energy when the current was brought to a standstill, the result being the heating of *R*.

## More Comparison and Contrast

In the previous chapter we saw that in addition to the conducted currents, reckoned by dividing voltage by resistance, there is an extra current whenever the voltage is varying. This current is made up of movements of charges caused by the electric field adjusting itself to the new voltage. It is proportional not only to the rate at which the voltage is changing but also to the capacitance (intentional or otherwise) of the part of the circuit considered. In this chapter we have seen that in addition to the resistive voltage, reckoned by multiplying current by resistance, there is an extra voltage whenever the current is varying. This voltage is caused by the magnetic field adjusting itself to the new current. It is proportional not only to the rate at which the current is changing but also to the inductance (intentional or otherwise) of the part of circuit considered.

## Mutual Inductance

If a coil with the same number of turns as that in Fig. 4.8 were wound so closely around it that practically all the magnetic flux due to the original *primary winding* embraced also this *secondary winding*, then an equal voltage would at all times be induced in the secondary. Generally it is impracticable to wind so closely that the whole flux links or couples with both coils, and the secondary voltage is less, according to the looseness of magnetic coupling.

But by giving the secondary more turns than the primary, it is possible to obtain a greater voltage in the secondary than in the primary. Remember that this voltage depends on the *varying* of the primary current. A device of this kind, for stepping voltages up or down, or for inducing voltages into circuits without any direct

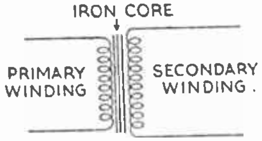



**Fig. 4.10—How a transformer is represented in circuit diagrams**

connection, is named (rather inappropriately) a *transformer*. It is represented in circuit diagrams by two or more coils drawn closely together. An iron core is shown by one or more straight lines drawn between them (Fig. 4.10).

The effect that one coil can exert, through its magnetic field, on another, is called *mutual* inductance (symbol : $M$), and like self-inductance is measured in henries. The definition is similar too: the mutual inductance in henries between two coils is equal to the number of volts induced in one of them when the current in the other is changing at the rate of 1 A per second.

Transformers are an immensely useful application of the inductive effect, and in fact are the main reason why alternating current (a.c.) has almost superseded direct current (d.c.) for electricity supply. So we shall now proceed to study a.c.

CHAPTER 5

# Alternating Currents

## Frequencies of Alternating Current

In our first chapter we saw that speaking and music are conveyed from one person to another by sound waves, which consist of rapid vibrations or alternations of air pressure. And that to transmit them over longer distances by telephone it is necessary for electric currents to copy these alternations. And further, that for transmitting them across space, by wireless, it is necessary to use electric currents alternating still more rapidly. We have now just noted the fact that the great advantage of being able to step voltages up and down as required is obtainable only when the electricity supply is continually varying, which is most conveniently done by arranging for it to be alternating. Public electricity supplies which light and heat our houses and provide the power that works our wireless sets (instead of the more expensive and troublesome batteries) are therefore mostly of the alternating-current kind.

The only essential difference between all these alternating currents is the number of double-reversals they make per second; in a word, the frequency. We have already gone fairly fully into the matter of frequency (p. 18), so there is no need to repeat it; but it may be worth recalling the main divisions of frequency.

There is no hard-and-fast dividing line between one lot of frequencies and another; but those below 100 cycles per second are used for power (the standard in Britain is 50 c/s); those between about 20 and 20,000 c/s are audible, and therefore are classed as audio frequencies (a.f.); while those above about 15,000 are more or less suitable for carrying signals across space, and are known as radio frequencies (r.f.). Certain of these, such as 525,000 to 1,605,000, are allocated for broadcasting.

The useful band of r.f. is subdivided as shown in the table below.

| *Class* | *Abbreviation* | *Frequencies* | *Wavelengths* |
|---|---|---|---|
| Very low | v.l.f. | Below 30 kc/s | Above 10,000 metres |
| Low | l.f. | 30–300 kc/s | 1,000–10,000 metres |
| Medium | m.f. | 300–3,000 kc/s | 100–1,000 metres |
| High | h.f. | 3–30 Mc/s | 10–100 metres |
| Very high | v.h.f. | 30–300 Mc/s | 1–10 metres |
| Ultra high | u.h.f. | 300–3,000 Mc/s | 10–100 cm |
| Super high | s.h.f. | 3,000–30,000 Mc/s | 1–10 cm |
| Extremely high | e.h.f. | Above 30,000 Mc/s | Below 1 cm |

All those above about 1,000 Mc/s (wavelengths shorter than about 1 ft.) are often lumped together as *microwave* frequencies.

What has been said about currents applies to voltages, too; it requires an alternating voltage to drive an alternating current.

The last two chapters have shown that when currents and voltages are varying there are two circuit effects—capacitance and inductance—that have to be taken into account as well as resistance. We shall therefore be quite right in expecting a.c. circuits to be a good deal more complicated and interesting than d.c. Before going on to consider these circuit effects there are one or two things to clear up about a.c. itself.

## The Sine Wave

It is easy to see that an alternating current might reverse abruptly or it might do so gradually and smoothly. In Fig. 3.6*a* we have an example of abrupt reversal. There is, in fact, an endless variety of ways in which current or voltage can vary with time, and when graphs are drawn showing how they do so we get a corresponding variety of wave shapes. The steepness of the wave at any point along the time scale shows the rapidity with which the current is changing at that time.

Fortunately for the study of alternating currents, all wave shapes, however complicated, can be regarded as built up of waves of one fundamental shape, called the *sine wave* (adjective: sinusoidal). Fig. 1.2*c* is an example of this. Note the smooth, regular alternations of the component sine waves, like the swinging of a pendulum. Waves of even such a spiky appearance as those in Fig. 3.6 can be analysed into a number of sine waves of different frequencies.

Most circuits used in wireless and other communications consist basically of sources of alternating e.m.f. (hereinafter called *generators*) connected to combinations of resistance, inductance and/or capacitance (often called *loads*). With practice one can reduce even very complicated-looking circuits to standard generator-and-load combinations.

We have just seen that although there is no end to the variety of waveforms that generators can produce, they are all combinations of simple sine waves. So it is enough to consider the sine-wave generator. Sources of more complicated waveforms can be imitated by combinations of sine-wave generators. Theoretically this dodge is always available, but in practice (as, for example, with square waves) there are sometimes easier alternatives. Unless the contrary is stated, it is to be assumed from now on that a generator means a sine-wave generator.

## Circuit with Resistance only

The simplest kind of a.c. circuit is the one that can be represented as a generator supplying current to a purely resistive load (Fig. 5.1). We have already found (p. 34) how to calculate the current any given e.m.f. will drive through a resistance. Applying this method to an alternating e.m.f. involves nothing new. Suppose, for example, that the voltage at the crest of each wave is 200, and that its frequency is 50 c/s. Fig. 5.2 shows a graph of this e.m.f. during rather more than one cycle. And suppose that $R$ is 200 Ω. Then

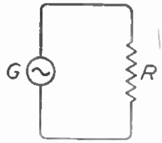



**Fig. 5.1—Circuit consisting of an a.c. generator (G) feeding a purely resistive load (*R*)**

we can calculate the current at any point in the cycle by dividing the voltage at that point by 200, in accordance with Ohm's law. At the crest it is, of course, 200/200 = 1 A; half-way up the wave it is 0·5 A. Plotting a number of such points we get the current wave, shown dotted. It is identical in shape, though different in vertical scale.

## R.M.S. Values

How is an electricity supply that behaves in this fashion to be rated? Can one fairly describe it as a 200-volt supply, seeing that the actual voltage is changing all the time and is sometimes zero? Or should one take the average voltage?

The answer that has been agreed upon is based on comparison with d.c. supplies. It is obviously very convenient for a lamp or fire intended for a 200-V d.c. system to be equally suited to a.c. mains of the same nominal voltage. This condition will be fulfilled if the average *power* taken by the lamp or fire is the same with both

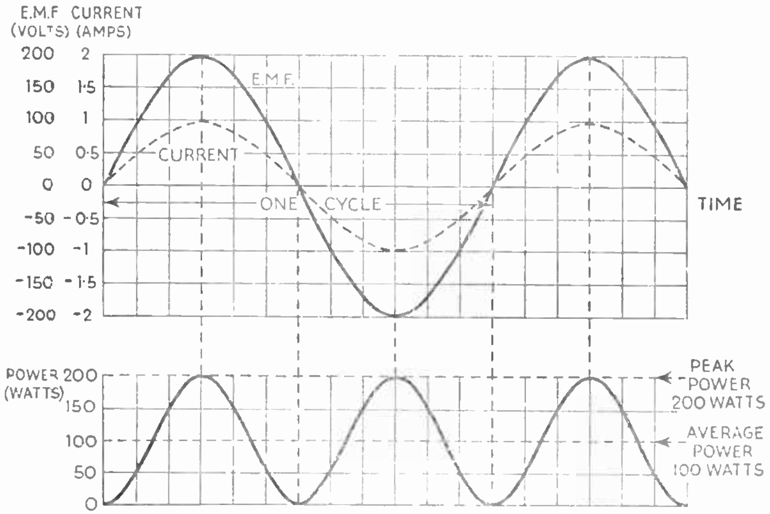



**Fig. 5.2—Time graphs of e.m.f., current and power in the circuit Fig. 5.1**

types of supply, for then the filament will reach the same temperature and the light or heat will be the same in both cases.

We know (p. 47) that the power in watts is equal to the voltage multiplied by the amperage. Performing this calculation for a sufficient number of points in the alternating cycle in Fig. 5.2, we get a curve showing how the power varies. To avoid confusion it has been drawn below the $E$ and $I$ curves. A significant feature of this power curve is that although during half of every cycle the voltage and current are negative, their negative half-cycles always exactly coincide, so that even during these half-cycles the power is (by the ordinary rules of arithmetic) positive. This mathematical convention agrees with and represents the physical fact that although both current and voltage reverse their direction the flow of power is always in the same direction, namely, out of the generator (p. 47).

Another thing one can see by looking at the power curve is that its average height is half the height of the crests—which, incidentally, occur 100 times per second when the frequency of the voltage and current are 50 c/s. The crest height is $200 \times 1 = 200$ W; so the average power must be 100 W. The next step is to find what direct voltage would be needed to dissipate 100 W in 200 $\Omega$; the answer is on p. 47: $P = E^2/R$, from which $E^2 = PR$, which in this case is $100 \times 200 = 20{,}000$. So $E = \sqrt{20{,}000} = 141$ approximately.

Judged on the basis of equal power, then, an a.c. supply with a maximum voltage of 200 is equivalent to a d.c. supply at 141 V.

Generalizing this calculation, we find that when an alternating voltage is adjusted to deliver the same power to a given resistance as a direct voltage, its crest voltage—called its *peak* value—is $\sqrt{2}$ times the direct voltage, or about 1·414 times as great. Put the other way round, its nominal or equivalent or effective voltage—called its *r.m.s.* (root-mean-square) value—is $1/\sqrt{2}$ or about 0·707 times its peak value. Since the resistance is the same in both cases, the same ratio exists between r.m.s. and peak values of the current. What is called a 230 V a.c. supply therefore alternates between + and $-\ \sqrt{2} \times 230 = 325$ V peak.

The r.m.s. value is *not* the same as the average voltage or current (which is actually 0·637 times the peak value). If you have followed the argument carefully you will see that this is because we have been comparing the a.c. and d.c. on a power basis, and power is proportional to voltage or current *squared.*

It is important to remember that the figures given above apply only to the sine waveform. With a square wave, for example, it is obvious that peak, r.m.s., and average (or mean) values are all the same.

There is one other recognized "value"—the instantaneous value, changing all the time in an a.c. system. It is the quantity graphed in Fig. 5.2.

As regards symbols, plain capital letters such as $E$, $V$ and $I$ are generally understood to mean r.m.s. values unless the contrary is obvious. Instantaneous values, when it is necessary to distinguish them, are denoted by small letters, such as $i$; peak values by $I_{max}$; and average values by $I_{ave}$.

Using r.m.s. values for voltage and current we can forget the rapid variations in instantaneous voltage, and, *so long as our circuits are purely resistive*, carry out all a.c. calculations according to the rules discussed in Chapter 2.

## A.C. Meters

If alternating current is passed through a moving-coil meter (p. 61) the coil is pushed first one way and then the other, because the current is reversing in a steady magnetic field. The most one is likely to see is a slight vibration about the zero mark. Certainly it will not indicate anything like the r.m.s. value of the current.

If, however, the direction of magnetic field is reversed at the same times as the current, the double reversal makes the force act in the same direction as before, and the series of pushes will cause the pointer to take up a position that will indicate the value of current. The obvious way to obtain this reversing magnetic field is to replace the permanent magnet by a coil and pass through it the current being measured. When the current is small, the field also is very weak, and the deflection too small to be read; so this principle is seldom used, except in *wattmeters*, in which the main current is passed through one coil, and the other—the *volt coil*—is connected across the supply.

A more usual type is that in which there is only one coil, which is fixed. Inside are two pieces of iron, one fixed and the other free to move against a hairspring. When either d.c. or a.c. is passed through the coil, both irons are magnetized with the same polarity, and so repel one another, to a distance depending on the strength of current. These *moving-iron* meters are useful when there is plenty of power to spare in the circuit for working them, but they tend to use up too much in low-power circuits.

Another method is to make use of the heating effect of the current. When a junction of two different metals is heated, a small unidirectional e.m.f. is generated, which can be measured by a moving coil meter. Instruments of this kind, called *thermojunction* or *thermocouple* types, if calibrated on d.c. will obviously read r.m.s. values of a.c. regardless of waveform. They are particularly useful for much higher frequencies than can be measured with instruments in which the current to be measured has to pass through a coil.

The electrostatic instrument (p. 45) can be used for alternating voltages and responds to r.m.s. values.

But perhaps the most popular method of all is to convert the alternating current into direct by means of a *rectifier*—a device that allows current to pass through it in one direction only—so that it can be measured with an ordinary moving-coil meter. The great advantage of this is that by adding a rectifier a multi-range d.c. instrument can be used on a.c. too. Most of the general-purpose test meters used in radio are of this kind. Because the moving-coil instrument measures the average current, which in general is not the same as the r.m.s. current, the instrument is so arranged as to take account of the factor necessary to convert one to the other. This, as we have seen, is approximately 0·707/0·638 (= 1·11) for sine waves. It is different for most other waveforms, so the rectifier type of meter reads them incorrectly.

## Phase

Looking again at the current graph in Fig. 5.2 we see that it not only has the same shape as the e.m.f. that causes it but it is also exactly in step with it. Consideration of Ohm's law proves that this must be so in any purely resistive circuit. The technical word for being in step is being *in phase*. The idea of phase is very important, so we had better make sure we understand it.

Suppose we have two alternating generators, A and B. The exact voltages they give will not matter, but to make it easier to distinguish their graphs we shall suppose A gives about double the

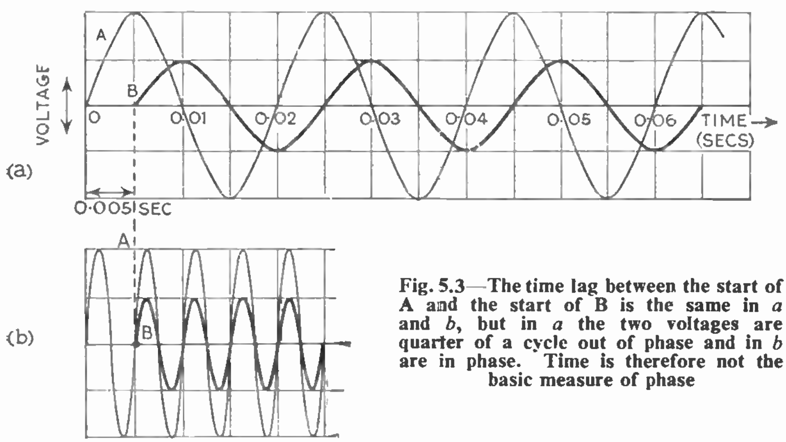



**Fig. 5.3—The time lag between the start of A and the start of B is the same in *a* and *b*, but in *a* the two voltages are quarter of a cycle out of phase and in *b* are in phase. Time is therefore not the basic measure of phase**

voltage of B. If we start drawing the voltage graph of A just as it begins a cycle the result will be something like waveform A in Fig. 5.3*a*. Next, consider generator B, which has the same frequency, but is out of step with A. It starts each of its cycles, say, quarter of a cycle later than A, as shown by waveform B. This fact is expressed by saying that voltage B lags voltage A by a phase difference of quarter of a cycle. Seeing that it is a *time* graph, it might seem more natural to say that B lagged A by a phase difference of 0·005 sec. The incorrectness of doing so is shown in Fig. 5.3*b*, where the frequency of A and B is four times as great. The time lag is the same as before, but the two voltages are now *in phase*. So although phase is usually very closely related to time, it is not advisable to think of it as time. The proper basis of reckoning phase is in fractions of a cycle.

## VECTOR DIAGRAMS

The main advantage of a time graph of alternating voltage, current, etc., is that it shows the waveform. But on those very many occasions when the form of the wave is not in question (because we have agreed to stick to sine waves) it is a great waste of effort to draw a number of beautifully exact waves merely in order to show the relative phases. Having examined Figs. 1.2, 5.2 and 5.3, we ought by now to be able to take the sine waveform for granted, and be prepared to accept a simpler method of indicating phases.

In Fig. 5.4, imagine the line OP to be pivoted at O and to be capable of rotating steadily in the direction shown (anticlockwise). (The arrowhead at P is there to show which end of OP is moving,


**Fig. 5.4—Starting position of an alternative method of representing sinusoidal variation**

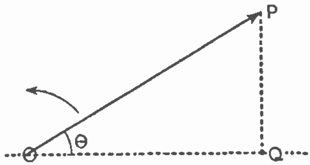



**Fig. 5.5—The Fig. 5.4 vector after having turned through an angle, $\theta$**

not the direction in which P is moving.) Then the height of the end P above a horizontal line through O varies in exactly the same way as a sine wave. Assuming OP starts, as shown in Fig. 5.4, at " 3 o'clock ", P is actually on the horizontal line, so its height above it is, of course, nil. That represents the starting point of a sine-wave cycle. As OP rotates, its height above the line increases at first rapidly and then more slowly, until, after a quarter of a revolution, it is at right angles to its original position. That represents the first quarter of a sine-wave cycle. During the third and fourth quarter of the revolution, the height of P is negative, corresponding to the same quarters of a cycle.

Readers with any knowledge of trigonometry will know that the ratio of the height of P to its distance from O (i.e., PQ/PO) is called the sine of the angle that OP has turned through (Fig. 5.5); hence the name given to the waveform we have been considering and which is none other than a time graph of the sine of the angle $\theta$ (abbreviated " sin $\theta$ ") when $\theta$ (pronounced " theta ") is increasing at a steady rate.

We now have a much more easily drawn diagram for representing sine waves. The length of a line such as OP represents the peak value of the voltage, etc., and the angle it makes with the " 3 o'clock " position represents its phase. The line itself is called a *vector*. The angle it turns through in one whole revolution (which, of course, brings it to the same position as at the start) is 360°, and that corresponds to one whole cycle of the voltage. So a common way of specifying a phase is in angular degrees. Quarter of a cycle is 90°, and so on. Usually even a waveform diagram such as Fig. 5.3 is marked in degrees.

Besides the common angular measure, in which one whole revolution is divided into 360 degrees, there is the mathematical angular measure in which it is divided into $2\pi$ *radians*. Seeing that $\pi$ is a very odd number (p. 2) this may seem an extremely odd way of doing things. It arises because during one revolution P travels

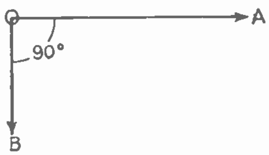



**Fig. 5.6—Vector diagram corresponding to Fig. 5.3*a***

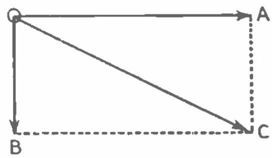



**Fig. 5.7—OC is the vector representing the resultant of OA and OB**

$2\pi$ times its distance from O. Even this may not seem a good enough reason, but the significance of it will appear very soon in the next chapter.

As an example of a vector diagram, Fig. 5.6 is the equivalent of Fig. 5.3*a*. In the position shown it is equivalent to it at the beginning or end of each cycle of A, but since both vectors rotate at the same rate the 90° phase difference shown by it between A and B is the same at any stage of any cycle. If the voltages represented had different frequencies, their vectors would rotate at different speeds, so the phase difference between them would be continually changing.

## Adding Alternating Voltages

A great advantage of the vector diagram is the way it simplifies adding and subtracting alternating voltages or currents that are not in phase. To add voltages A and B in Fig. 5.3*a*, for example, one would have to add the heights of the A and B waves at close intervals of time and plot them as a third wave. This would be a very laborious way of finding how the peak value and phase of the combined voltage (or *resultant*, as it is called) compared with those of its component voltages. In the vector diagram it is quickly done by dotting in from A a line equal in length and direction to OB (Fig. 5.7), to arrive at point C. Then OC is the vector representing voltage A plus voltage B. Alternatively, a line can be drawn from B equal to OA. Note that the result of adding two voltages that are not in phase is less than the result of adding them by ordinary arithmetic, and its phase is somewhere between those of the components. Since it is representable by a vector, it is sinusoidal, like its components.

If you mean to study a.c. seriously you should read a good elementary book on vectors and also learn the corresponding algebraical method of calculation, using the symbol $j$ to distinguish vertical from horizontal distances.*

## Direction Signs

Fig. 5.8*a* shows a simple resistive circuit at the instant when the alternating e.m.f. is at its maximum positive clockwise, as indicated by the + and − sign at the generator and the $E$ vector in Fig. 5.8*b*. We know that at this precise moment the current is also at its positive maximum (p. 73). But in which direction? The generally agreed convention (p. 47) is that it is regarded as flowing through a resistance from + to −; that is, in the direction indicated by the arrow in Fig. 5.8*a*. This, of course, is the same as the direction in

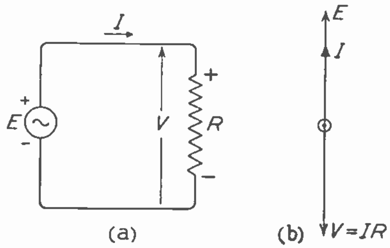



**Fig. 5.8—(*a*) A.c. circuit at the instant when the e.m.f. is at its positive maximum, and (*b*) the corresponding vector diagram**

which the e.m.f. $E$ is regarded as driving it, so we can draw a current vector $I$ in phase with $E$ in Fig. 5.8*b*. As we may be drawing more complicated vector diagrams before long, it will be a convenient convention to distinguish current vectors by blacked-in arrow heads.

The p.d. across the resistance is equal to that across the generator terminals, but its direction is anticlockwise. We can therefore draw a p.d. vector $V$, equal and opposite to $E$. This is merely the vectorial way of expressing the Kirchhoff law we studied in connection with d.c. circuits (p. 41), by which the sum of all the voltages completely round a circuit must be equal to zero.

If there were several resistances in the circuit, each would have across it an alternating p.d. which could be represented by its own vector; and these separate vectors, all in the same direction, would add up to one p.d. equal and opposite to the e.m.f.

In general, it is of no importance which particular instant in the alternating cycle one selects for drawing a vector diagram. The relative positions of the vectors, which are what matters, are the same throughout. (It is, in fact, more usual for the " reference " vector—the first one drawn—to be placed so that it points at the conventional zero angle, " 3 o'clock ".)

That being so, it might be supposed that the direction and polarity markings in Fig. 5.8*a* are unnecessary—and indeed meaningless,

* For example, *Basic Mathematics for Radio and Electronics*, by F. M. Colebrook and J. W. Head. Published for *Wireless World* by Iliffe & Sons, Ltd., 17s. 6d.

seeing that the direction is continually reversing. And certainly many circuit diagrams for which vector diagrams are drawn are devoid of such markings. But the reason for the omission is not that they are inappropriate to a.c. circuits, but because the drawer of the diagrams assumes that everyone would choose the same relative directions as he.

It is likewise sometimes assumed that everyone invited to a social event will know what type of dress will be worn; but just as absence of precise instructions on this point sometimes leads to embarrassment, so with diagrams. It is better to include direction markings, even when they might seem unnecessary, than to leave room for wrong guesses. Even in such a simple circuit as Fig. 5.8*a* the correctness of the vector diagram *b* depends on the direction of *I* being as shown at *a*, which in turn depends on a mere optional custom—the custom of preferring directions that lead to positive values.

To understand what this means, suppose the *I* arrow had been drawn pointing the other way. This would have been inconvenient, and almost certainly confusing for beginners, but it would not necessarily have been wrong. It would have been inconvenient, because it would have meant that whenever *E* was positive in the direction marked, *I* would have been negative. It would not have been wrong, because a negative current flowing one way is exactly the same thing as a positive current flowing the other way. Because this choice of direction for the *I* arrow in *a* would imply negative values relative to *E*, the *I* vector in *b* would have had to be drawn in the reverse direction. This is the important point: that the correct directions of the vectors depend entirely on the relative directions assumed in the circuit. Note, *relative* directions. Reversing *all* the directions (which is what happens when the current alternates!) makes no difference.

In Fig. 5.8 it is pretty obvious which is the more convenient choice of the two, but in more complicated circuits it may not be at all obvious at the start which choice is going to give positive values. Moreover, we shall soon be learning that during the positive half-cycle of e.m.f. it is the exception rather than the rule for the current to be either positive or negative all that time. And even when it is, there might be several e.m.fs in opposing directions—or several branch currents. So although in a circuit with only one e.m.f. and current it is the custom to assume the same direction for both, it is not safe to leave out the signs in more complicated circuits. So long as circuit and vector diagrams agree, there is no question of right or wrong; we can choose any direction we like, so long as we make our choice clear.

# Capacitance in A.C. Circuits

## Current Flow in a Capacitive Circuit

The next type of basic circuit to consider is the one shown in Fig. 6.1. If the generator gave an unvarying e.m.f. no current could flow, because there is a complete break in the circuit. The most that could happen would be a momentary current when switching on, as shown in Fig. 3.6.

Some idea of what is likely to happen when an alternating e.m.f. is applied can be obtained by an experiment similar to Fig. 3.5 but with an inverted battery in the B path, as shown in Fig. 6.2*a*. Moving the switch alternately to A and B at equal time intervals will then provide an alternating voltage of square waveform (Fig. 6.2*b*).

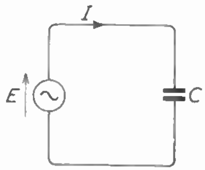



**Fig. 6.1—Circuit consisting of an a.c. generator with a purely capacitive load**

Whenever the switch is moved to A there is a momentary charging current in one direction, and moving it to B causes a similar current in the reverse direction.

We know that for a given voltage the quantity of electricity transferred at each movement of the switch is proportional to the capacitance (p. 50). And it is obvious that the more rapidly the switch is moved to and fro (that is to say, the higher the frequency of the alternating e.m.f.) the more often this quantity of electricity will surge to and fro in the circuit (that is to say, the greater the quantity of electricity that will move in the circuit per second, or, in other words, the greater the current).

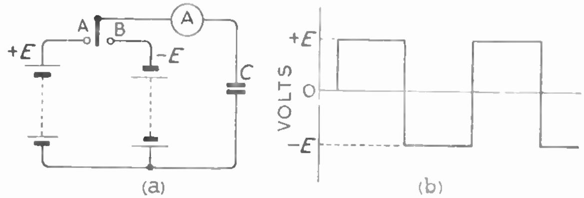
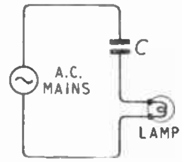



**Fig. 6.2—If the generator in Fig. 6.1 consisted of the periodically switched batteries as shown at *a*, the waveform would be as at *b*, and the current (assuming a certain amount of resistance) could be calculated as discussed in connection with Fig. 3.6**

Thus, although the circuit has no conductance, so that no continuous d.c. can flow, an alternating e.m.f. causes an alternating current, flowing not *through* but *in and out* of the capacitor. And the amount of current due to a given voltage will depend on two things—the capacitance and the frequency. The greater they are, the greater the current.

The truth of the first statement—that an alternating e.m.f. causes a current to flow in a circuit blocked to d.c. by a capacitance in

**Fig. 6.3—Demonstration of current without conduction**




series—can easily be demonstrated by bridging the open contacts of an electric light switch by a capacitor; say 2 μF for a 40 W lamp (Fig. 6.3). The lamp will light and stay alight as long as the capacitor is connected. But its brilliance will be below normal.

## Capacitive Current Waveform

Let us now consider the action of Fig. 6.1 in greater detail, by drawing the graph of instantaneous e.m.f. (Fig. 6.4) exactly as for the resistive circuit, in which we assumed a sine waveform. But unlike the resistive case there is no Ohm's law to guide us in plotting the current waveform. We have, however, a rather similar relationship (p. 50):

$$V = Q/C$$

where $V$ is the p.d. across $C$ and $Q$ is the charge in $C$. This is true at every instant, so we can rewrite it $v = q/C$, to show that we mean instantaneous values. In Fig. 6.1, $v$ is always equal in magnitude to $e$, the instantaneous e.m.f. So we can say with confidence that at the moments when $e$ is zero the capacitor is completely uncharged, and at all other moments the charge is exactly proportional to $e$.

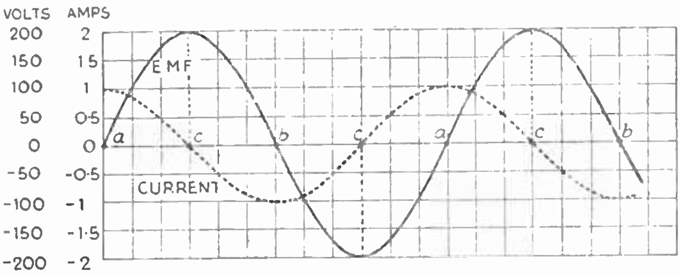



**Fig. 6.4—E.m.f. and current diagram for Fig. 6.1, drawn to correspond with Fig. 5.2**

If we knew the right scale, the voltage wave in Fig. 6.4 would do also as a charge curve. But we are not so much interested in the charge as in the current; that is to say, the rate at which charging takes place. At points marked *a*, although the voltage and charge are zero they are growing faster than at any other stage in the cycle. So we may expect the current to be greater than at any other times. At points marked *b*, the charge is decreasing as fast as it was growing at *a*, so the current is the same in magnitude but opposite in direction and sign. At points marked *c*, the charge reaches its maximum, but just for an instant it is neither growing nor waning; its rate of increase or decrease is zero, so at that instant the current must be zero. At intermediate points the relative strength of current can be estimated from the steepness of the voltage (and charge) curve. Joining up all the points gives a current curve shaped like the dotted line.

If this job is done carefully it is clear that the shape of the current curve is also sinusoidal, as in the resistive circuit, but out of phase, being quarter of a cycle (or 90°) ahead of the voltage.

On seeing this, students are sometimes puzzled and ask how the current *can* be ahead of the voltage. "How does the current know what the voltage is going to be, quarter of a cycle later?" This difficulty arises only when it supposed, quite wrongly, that the current maximum at *a* is caused by the voltage maximum quarter of a cycle later, at *c*; whereas it is actually caused by the fact that at *a* the voltage is increasing at its maximum rate.

### The "Ohm's Law" for Capacitance

What we want to know is how to foretell the actual magnitude of the current, given the necessary circuit data. Since, as we know (p. 50), 1 volt established between the terminals of a 1-farad capacitor causes a charge of 1 coulomb to enter it, we can see that if the voltage were increasing at the rate of 1 volt per second the charge would be increasing at 1 coulomb per second. But we also know that 1 coulomb per second is a current of 1 amp; so we can express our basic relationship, $Q = VC$, in the form

$$\text{Amps} = \text{Volts-per-second} \times \text{Farads}.$$

The problem is how to calculate the volts-per-second in an alternating e.m.f., especially as it is varying all the time. All we need do, however, is find it at the point in the cycle where it is greatest—at *a* in Fig. 6.4. That will give the peak value of current, from which all its other values follow (p. 73).

Going back to Fig. 5.5, you may remember that one revolution of P about O represents one cycle of alternating voltage, the fixed length OP represents the peak value of the voltage, and the length of PQ (which, of course, is varying all the time) represents the instantaneous voltage. When P is on the starting line, as in Fig. 5.4, the length of PQ is zero, but at that moment it is increasing at the rate at which P is revolving around O. Now the distance travelled by P during one cycle is $2\pi$ times the length of OP. And if the frequency is 50 c/s, it does this distance 50 times per second. Its rate is therefore $2\pi \times 50$ times OP. More generally, if $f$ stands for the frequency in cycles per second, the rate at which P moves round O is $2\pi f$ times OP. And as OP represents $E_{max}$, we can say

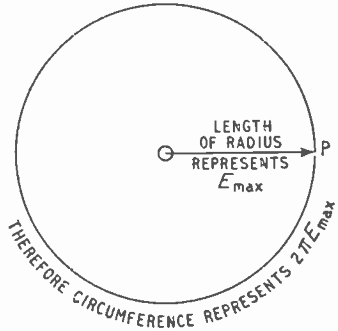



**Fig. 6.5—From this diagram it can be inferred that the maximum rate at which $E$ increases is $2\pi f$ $E_{max}$ volts per second**

that P's motion represents $2\pi f E_{max}$ volts per second (Fig. 6.5). But we have just seen that it also represents the maximum rate at which PQ, the instantaneous voltage, is growing. So, fitting this information into our equation, we have

$$I_{max} = 2\pi f E_{max} C$$

where $I$ is in amps, $E$ in volts, and $C$ in farads. Since $I_{max}$ and $E_{max}$ are both $\sqrt{2}$ times $I$ and $E$ (the r.m.s. values) respectively, it is equally true to say

$$I = 2\pi f EC$$

And if we rearrange it like this :

$$\frac{E}{I} = \frac{1}{2\pi f C}$$

and then take another look at one of the forms of expressing Ohm's law:

$$\frac{E}{I} = R$$

we see they are the same except that $1/2\pi fC$ takes the place of $R$ in the role of limiting the current in the circuit. So, although $1/2\pi fC$ is quite a different thing physically from resistance, for purposes of calculation it is of the same kind, and it is convenient to reckon it in ohms. To distinguish it from resistance it is named *reactance*. To avoid having to repeat the rather cumbersome expression $1/2\pi fC$, the special symbol $X$ has been allotted. The "Ohm's law" for the Fig. 6.1 circuit can therefore be written simply as

$$\frac{E}{I} = X \quad \text{or } I = \frac{E}{X} \quad \text{or } E = IX$$

For example, the reactance of a 2 μF capacitor to a 50 c/s supply is $1/(2\pi \times 50 \times 2 \times 10^{-6}) = 1{,}590\ \Omega$, and if the voltage across it were 230 V the current would be $230/1590 = 0{\cdot}145$ A.

## Capacitances in Parallel and in Series

The argument that led us to conclude that the capacitance between parallel plates is proportional to the area across which they face one another (p. 52) leads also to the conclusion that capacitances in parallel add up just like resistances in series.

We can arrive at the same conclusion by simple algebraic reasoning based on the behaviour of capacitors to alternating voltage. Fig. 6.6 represents an a.c. generator connected to two capacitors in parallel. The separate currents in them are respectively $E \times 2\pi f C_1$

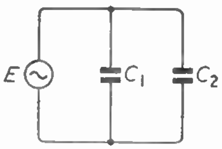



**Fig. 6.6—In this circuit the capacitance of $C_1$ and $C_2$ together is equal to $C_1 + C_2$**

and $E \times 2\pi f C_2$. The total current is thus $E \times 2\pi f(C_1 + C_2)$, which is equal to the current that would be taken by a single capacitance equal to the sum of the separate capacitances of $C_1$ and $C_2$.

We also saw (p. 52) that if a single capacitor is divided by a plate placed midway between its two plates, the capacitance between the middle plate and either of the others is twice that of the original

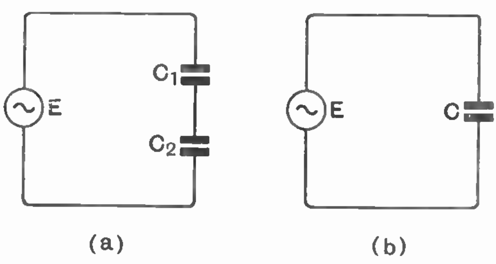



**Fig. 6.7 — If the one capacitor $C$ in $b$ is to take the same current as the two in series at $a$, its capacitance must be equal to $1/(1/C_1 + 1/C_2)$**

capacitor. In other words, the capacitance of two equal capacitors in series is half that of each of them. Let us now consider the more general case of any two capacitances in series.

If $X_1$ and $X_2$ are respectively the reactances of $C_1$ and $C_2$ in Fig. 6.7, their combined reactance $X$ is $(X_1 + X_2)$, as in the case of resistances in series. By first writing down the equation $X = X_1 + X_2$, and then replacing each $X$ by its known value, of form $1/2\pi f C$, we deduce that

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2}$$

That is, the sum of the reciprocals of the separate capacitances is equal to the reciprocal of the total capacitance.

So the rule for capacitances in series is identical with that for resistances or reactances in parallel. From the way it was derived it is evidently not limited to two capacitances only, but can be applied to any number. It implies that if capacitors are connected in series, the capacitance of the combination is always less than that of the smallest.

Applied to two capacitances only, the rule can, as with resistances (p. 38), be put in the more easily workable form

$$C = \frac{C_1C_2}{C_1 + C_2}$$

## Power in a Capacitive Circuit

Fig. 6.4 was drawn to match Fig. 5.2 as regards peak voltage and current. The only difference is the current's 90° phase lead. Let us now complete the picture by calculating the wattage at a sufficient number of points to draw the power curve (Fig. 6.8). Owing to the phase difference between $E$ and $I$, there are periods in the cycle when they are of opposite sign, giving a negative power. In fact, the power is alternately positive and negative, just like $E$ and $I$ except for having twice the frequency. The net power, taken over a whole cycle, is therefore zero. Since we are applying the sign " + " to positive current going through the generator towards positive potential, positive power means power going out of the generator, and negative power is power returned to the generator by the load (p. 47).

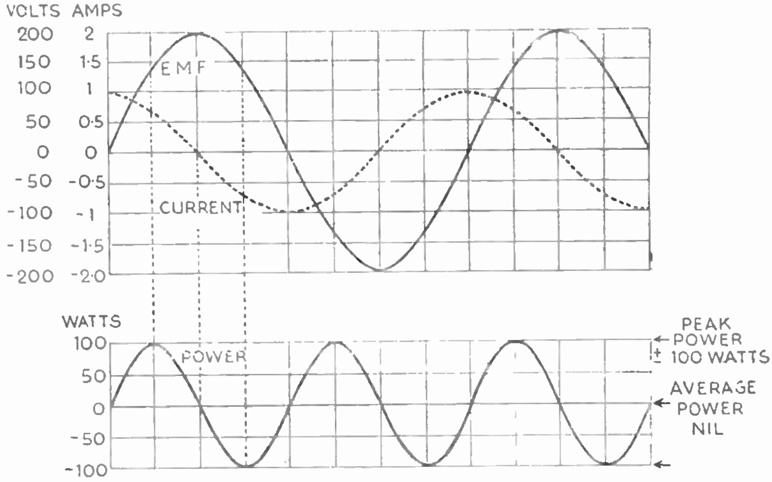



**Fig. 6.8—The current and voltage graphs for a capacitance load (Fig. 6.4) are here repeated, with a power graph derived from it plotted below**

The power curve in Fig. 6.8 represents the fact that although a pure capacitance draws current from an alternating generator it does not permanently draw any power—it only borrows some twice per cycle, repaying it with a promptitude that might be commended to human borrowers.

A demonstration of this fact is given by the capacitor in series with the lamp (Fig. 6.3). The lamp soon warms up, showing that electrical power is being dissipated in it, but although the capacitor is carrying the same current it remains quite cold.

## Capacitance and Resistance in Series

Fig. 6.9 is an example of the next type of circuit to be considered: $C$ and $R$ in series. We know how to calculate the relationship (magnitude and phase) between voltage and current for each of these separately. When they are in series the same current will flow through both, and this current cannot at one and the same time be in phase with the generator voltage and 90° ahead of it. But it

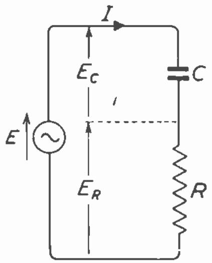



**Fig. 6.9—For finding the current, *I*, caused by *E* to flow in a circuit comprising *C* and *R* in series, it is helpful to split *E* up into the two parts needed to drive *I* through *C* and *R* separately**

is quite possible—and in fact essential—for it to be in phase with the voltage across the resistance and 90° ahead of the voltage across the capacitance. These two voltages, which must obviously be 90° out of phase with one another, must add up to equal the generator voltage. We have already added out-of-phase voltages by two different methods (Fig. 5.3 and Fig. 5.7), so that part of it should not be difficult. Because the current, and not the voltage, is common to both circuit elements, it will be easier to reverse the procedure we have adopted until now, and, starting with a current, find the e.m.f. required to drive it. When we have in that way discovered the key to this kind of circuit, we can easily use it to calculate the current resulting from a given e.m.f.

For brevity, let $E_C$ and $E_R$ denote the e.m.fs needed for $C$ and $R$ (Fig. 6.9). We could, of course, represent $E_C$ and $E_R$ by drawing waveform graphs, and, by adding them, plot the graph of $E$, the total generator e.m.f., and compare it for phase and magnitude with the current graph. But seeing we have taken the trouble to learn the much quicker vector method of arriving at the same result, we might as well use it. (If you find the vector method unconvincing, there is nothing to stop you drawing the waveforms.)

The first thing, then, is to draw a vector of any convenient length to represent the current, $I$ (Fig. 6.10). Since we are concerned only with *relative* phases, it can be drawn in any direction. The $E_R$ vector must be drawn in the same direction, to represent the fact that the current driven through a resistance is in phase with the

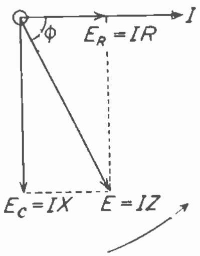



**Fig. 6.10—Vector diagram relating to Fig. 6.9. $E_R$ is in phase with $I$, but $E_C$ lags $I$ by 90°. $E$ is the vector sum of $E_R$ and $E_C$, and lags $I$ by the angle $\phi$**

e.m.f. Its length represents the voltage of the e.m.f. ($=IR$) to any convenient scale. The length of the $E_C$ vector represents $E_C$ ($=IX$ or $I/2\pi fC$) to the same scale, and since we know the current is 90° ahead of it we must draw it 90° behind the current vector. The vector representing $E$ in magnitude and phase is then easily found as explained in connection with Fig. 5.7.

### IMPEDANCE

A new name is needed to refer to the current-limiting properties of this circuit as a whole. *Resistance* is appropriate to $R$ and *reactance* to $C$; a combination of them is called *impedance* (symbol: $Z$). Resistance and reactance themselves are special cases of impedance. So we have still another relationship in the same form as Ohm's law; one that covers the other two, namely:

$$I = \frac{E}{Z}$$

Although $Z$ combines $R$ and $X$, it is not true to say $Z = R + X$. Suppose that in Fig. 6.9 $C$ is the 2 μF capacitor which we have already calculated has a reactance of 1,590 Ω at 50 c/s, and that $R$ is a round 1,000 Ω. $I$, shall we say, is 0·1 A. Then $E_R$ must be 100 V and $E_C$ 159 V. If the vector diagram is drawn accurately to scale it will show $E$ to be 188 V, so that $Z$ must be 1,880 Ω, which is much less than 1,590 + 1,000. It should be obvious from Fig. 6.10 that the three voltage vectors could, by suitable choice of scale, be used to represent the respective impedances. Then, since these three vectors placed together make a right-angled triangle, the celebrated Theorem of Pythagoras tells us that $Z^2 = R^2 + X^2$, or

$$Z = \sqrt{(R^2 + X^2)}.$$

So we have the alternative of calculating $Z$ by arithmetic instead of drawing a vector diagram or a waveform diagram accurately to scale. Since the voltages are proportional to the impedances,

$$E = \sqrt{(E_R^2 + E_C^2)}.$$

The angle by which $I$ leads $E$ in Fig. 6.10 is marked $\phi$, which is the usual symbol for a phase difference. It is the Greek small letter "phi". (We have already used the capital, $\Phi$, to mean flux.) The trigonometrists will see that $\tan\phi = X/R$.

Given the values of $R$ and $C$, and the frequency, we now have three alternative methods of finding their series impedance and the phase difference between current and voltage; also, given either current or voltage, of finding the other. Although the method of adding out-of-phase voltages by drawing their waveforms takes longest to do, it does perhaps bring out most clearly the reason why the total voltage is less than the sum of the two separate voltages—namely, that their peak values occur at different times.

From what has already been said about power in a.c. circuits it should be clear that in Fig. 6.9, for example, the wattage dissipated is given by multiplying $I$ by $E_R$ (*not* by $E$), so that only part of the volt-amps supplied by the generator ($EI$) represents power permanently delivered.

## Capacitance and Resistance in Parallel

In solving this type of circuit, Fig. 6.11, we revert to the practice of starting with the voltage $E$, because that is common to both. What we do, in fact, is exactly the same as for the series circuit except that currents and voltages change places. We now have two

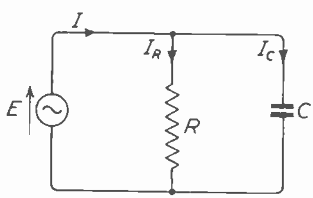



**Fig. 6.11—The branch currents in this circuit, $I_R$ and $I_C$, are 90° out of phase with one another**

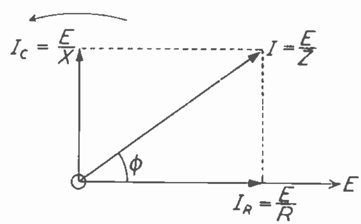



**Fig. 6.12—Vector diagram relating to Fig. 6.11**

currents, $I_R$ and $I_C$, one in phase with $E$ and the other 90° ahead of it; and by adopting the same methods we find that $I=\sqrt{(I_R^2+I_C^2)}$. Fig. 6.12 is a typical vector diagram, which should need no further explanation.

Calculating the impedance is slightly more complicated, however, because the currents are *inversely* proportional to the impedances through which they flow. So $1/Z=\sqrt{(1/R^2+1/X^2)}$, or

$$Z = \frac{1}{\sqrt{\frac{1}{R^2}+\frac{1}{X^2}}}$$

which can be simplified to

$$Z = \frac{RX}{\sqrt{R^2 + X^2}}$$

Compare the formula for two resistances in parallel, p. 38.

It should be noted that this addition of squares applies only to circuits in which the two currents or voltages are exactly 90° out of phase, as when one element is a pure resistance and the other a pure reactance. Combining series impedances which are themselves

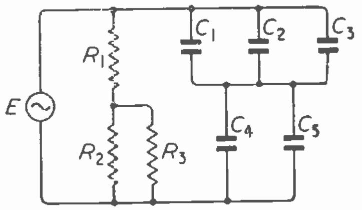



**Fig. 6.13—Example of a complex circuit that can be reduced to Fig. 6.11 by applying the rules for resistances and capacitances in series and in parallel**

combinations of resistance and reactance in parallel, or vice versa, is a rather more advanced problem than will be considered here. Such a circuit as Fig. 6.13, however, can be tackled by rules already given. $R_2$ and $R_3$ are reduced to a single equivalent resistance (p. 44) which is then added to $R_1$. $C_1$, $C_2$ and $C_3$ are just added together, and so are $C_4$ and $C_5$; the two resulting capacitances in series are reduced to one (p. 85). The circuit has now boiled down to Fig. 6.11.

CHAPTER 7

# Inductance in A.C. Circuits

## Current Flow in an Inductive Circuit

ALTHOUGH THE PHENOMENON of magnetism, which gives rise to inductance, differs in many ways from electrostatics, which gives rise to capacitance, it is helpful to draw a very close parallel or analogy between them. Inductance and capacitance are, in fact, like opposite partners; and this chapter will in many ways be a repetition of the last one, but with a few basic situations reversed.

We have already noticed some striking similarities, as well as differences, between capacitance and inductance (p. 68). One thing that could be seen by comparing them in Figs. 3.6 and 4.9 was an exchange of roles between voltage and current.

The simple experiment of Fig. 6.2 showed that, in a circuit consisting of a square-wave alternating-voltage generator in series with a capacitor, the current increased with frequency, beginning with zero current at zero frequency. This was confirmed by examining in more detail a circuit (Fig. 6.1) with a sine-wave generator, in which the current was found to be exactly proportional to frequency.

Comparing Fig. 4.9 with Fig. 3.6, we can expect the opposite to apply to inductance. At zero frequency it has no restrictive influence on the current, which is limited only by the resistance of the circuit. But when a high-frequency alternating e.m.f. is applied, the current has very little time to grow (at a rate depending on the inductance) before the second half-cycle is giving it the " about turn ". The gradualness with which the current rises, you will remember, is due to the magnetic field created by the current, which generates an e.m.f. opposing the e.m.f. that is driving the current. It is rather like the gradualness with which a heavy truck gains speed when you push it. If you shake it rapidly to and fro it will hardly move at all.

## Inductive Current Waveform

For finding exactly how much current a given alternating voltage will drive through a given inductance (Fig. 7.1) we have the basic relationship (p. 63):

$$\text{Volts} = \text{Amps-per-second} \times \text{Henries}$$

It looks as if it is going to be difficult to find what we want directly from this; so let us tackle it in reverse, beginning with a sinusoidal current and seeing what voltage is required. The dotted curve in Fig. 7.2 represents such a current during rather more than one cycle. At the start of a cycle (*a*), the current is zero, but is increasing faster than at any other phase. So the amps-per-second is at its maximum, and so therefore must be the voltage. After half a cycle (*b*) the current is again zero and changing at the same rate, but this time it is *decreasing*, so the voltage is a maximum *negative*. Halfway between (*c*) the current is at its maximum, but for an instant it is neither increasing nor decreasing, so the voltage must be zero. And so on.

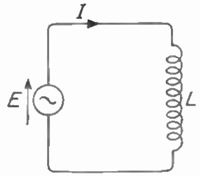



**Fig. 7.1—Circuit consisting of an a.c. generator with a purely inductive load**

Completing the voltage curve in the same way as we did the current curve in Fig. 6.4, we find that it also is sinusoidal. This is very fortunate, for it allows us to say that if the sinusoidal voltage represented by this curve were applied to an inductance, the current curve would represent the resulting current, which would be sinusoidal.

## The "Ohm's Law" for Inductance

We have already found (p. 83) that when an alternating e.m.f. $E$, of frequency $f$, is sinusoidal, its r.m.s. volts-per-second is $2\pi fE$. The same method applies equally to current, so the r.m.s. amps-per-second is $2\pi fI$. Fitting this fact into our basic principle we get

$$E = 2\pi fIL$$

where $L$ is the inductance in henries.

What we set out to find was the current due to a given e.m.f., so the appropriate form of the above equation is

$$I = \frac{E}{2\pi fL}$$

Again, this is in the same form as Ohm's law, $2\pi fL$ taking the place of $R$. This $2\pi fL$ can therefore also be reckoned in ohms, and, like $1/2\pi fC$, it is called reactance and denoted by $X$. Whenever it is necessary to distinguish between inductive reactance and

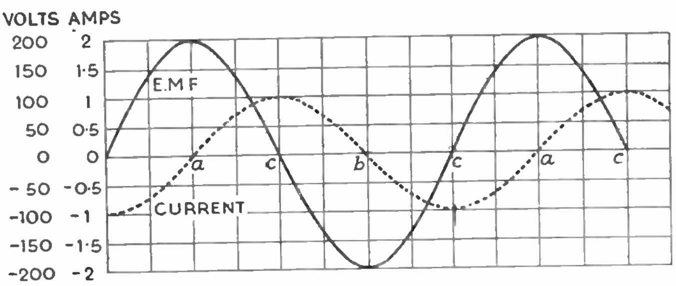



**Fig. 7.2—E.m.f. and current graphs for Fig. 7.1, drawn to correspond with Figs. 5.2 and 6.4**

capacitive reactance they are denoted by $X_L$ and $X_C$ respectively. $X$ is the general symbol for reactance.

A question that may come to mind at this point is: can $X_L$ and $X_C$ in the same circuit be added in the same simple way as resistances? The answer is so important that it will be reserved for the next chapter.

Note in the meantime that $X_L$ is proportional to $f$, whereas $X_C$ is proportional to $1/f$; this expresses in exact terms what our early experiments had led us to expect about the opposite ways in which frequency affected the current due to a given voltage.

An example may help to clinch the matter. What is the reactance of a 2 H inductor at 50 c/s? $X = 2\pi fL = 2\pi \times 50 \times 2 = 628\ \Omega$. So if the voltage is 230, the current will be $230/628 = 0{\cdot}366$ A.

The multiple of frequency, $2\pi f$, occurs so often in electrical calculations that it has been given a special symbol, the Greek letter $\omega$ (omega). (We have already met capital omega, $\Omega$, on p. 32.)

### Inductances in Series and in Parallel

The previous section has shown that the reactive impedance o an inductance is directly proportional to the inductance—oppositely to that of capacitance (which is proportional to $1/C$), but exactly like resistance. So inductances can be combined in the same way as resistances. The total effect of two in series, $L_1$ and $L_2$, is equal to that of a single inductance equal to $L_1 + L_2$. And inductances in parallel follow the reciprocal law: $1/L = 1/L_1 + 1/L_2$, or

$$L = \frac{L_1 L_2}{L_1 + L_2}$$

These principles can easily be verified by adding the reactances when they are in series, and the currents when the inductances are in parallel.

The above rules are subject to one important condition, however: that the inductors are so placed that their mutual inductance (p. 68) or magnetic coupling is negligible. If mutual inductance, $M$, does exist between two coils having separate self-inductances $L_1$ and $L_2$, the total self-inductance of the two in series is $L_1 + L_2 + 2M$ or $L_1 + L_2 - 2M$, according to whether the coils are placed so that their separate magnetic fields add to or subtract from one another. The reason $M$ is doubled is that each coil is affected by the other, so in the combination the same effect occurs twice.

The corresponding elaboration of the formula for two inductances in parallel gives the result

$$L = \frac{(L_1 \pm M)(L_2 \pm M)}{L_1 + L_2 \pm 2M}$$

where "$\pm$" signifies "$+$ or $-$".

Complications arise also when the fields of two capacitors interact, but with the usual forms of construction such interaction is seldom enough to trouble about.

## Power in an Inductive Circuit

Comparing Figs. 6.4 and 7.2 we see one difference. Whereas the capacitive current leads its terminal voltage by quarter of a cycle (90°), the inductive current lags by that amount. If you want something to do you might care to calculate the instantaneous power at intervals throughout the cycle in Fig. 7.2 and draw a power curve, as in Fig. 6.8. But the same result can be achieved with less trouble simply by reversing the time scale in Fig. 6.8, making it read from right to left. So except for relative phases the conclusions regarding power in a pure inductance are exactly the same as for a capacitance—the net power taken over a whole cycle is zero, because during half the time the generator is expending power in building up a magnetic field, and during the other half the energy thus built up is returning power to the generator. So in the example we took, with 230 V passing 0·366 A through 2 H, 230 × 0·366 does not represent watts dissipated in the circuit (as it would with resistance) but half that number of volt-amps tossed to and fro between generator and inductor at the rate of 100 times a second.

If you try an experiment similar to Fig. 6.3, but using a coil of several henries in place of the capacitor, you may find that it does become perceptibly warm. This is because the coil is not purely inductive. Whereas the resistance of the plates and connections

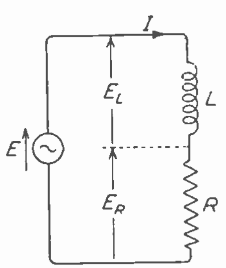



**Fig. 7.3—This circuit may be compared with Fig. 6.9**

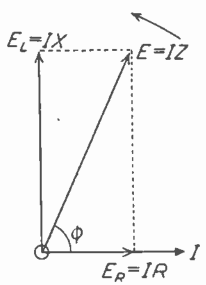



**Fig. 7.4—Vector diagram relating to Fig. 7.3**

of a capacitor is usually negligible compared with its reactance, the resistance of the wire in a coil is generally very appreciable. Although the resistance and inductance of a coil can no more be separate than the body and soul of a live person, it is allowable to consider them theoretically as separate items in series with one another (Fig. 7.3).

## Inductance and Resistance in Series

*R* in Fig. 7.3 represents the resistance of the inductor whose inductance is represented by *L*, plus any other resistance there may be in series. This circuit can be tackled in exactly the same manner as Fig. 6.9. $E_L$ is the e.m.f. required to drive a sinusoidal current $I$ through $L$, and is exactly equal to the back e.m.f. generated by the alternating magnetic flux linking the turns of the coil. $E_R$ is, as before, the e.m.f. required to drive the same current through $R$. $E_L$ and $E_R$ must be 90° out of phase with one another, because

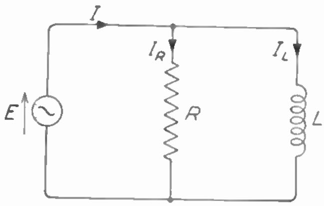



**Fig. 7.5—As in Fig. 6.11, the branch currents are 90° out of phase with one another, the only difference being that the reactive current $I_L$ is 90° later than $I_R$**

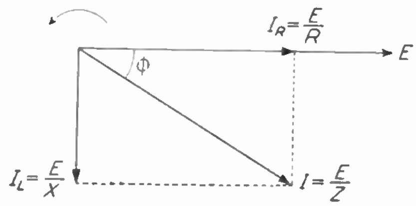



**Fig. 7.6—Vector diagram relating to Fig. 7.5**

whereas $I$ is in phase with $E_R$ it lags 90° behind $E_L$. So $E_L$ and $E_R$ cannot be added straightforwardly to give $E$, but by the same method as we found for Fig. 6.9, namely $E=\sqrt{(E_R^2 + E_L^2)}$. The vector diagram is obviously the same except for the current lagging $E_L$ instead of leading $E_C$; compare Fig. 7.4 with Fig. 6.10.

In the same way, too, the impedance of the circuit is $Z = \sqrt{(R^2 + X^2)}$.

## Inductance and Resistance in Parallel

This circuit, Fig. 7.5, corresponds with Fig. 6.11; and its impedance is found in the same way, by combining the branch currents, remembering that this time the reactive current *lags* by 90°. The vector diagram, Fig. 7.6, is therefore with this exception the same as Fig. 6.12. So we have for the total current $I = \sqrt{(I_R^2 + I_L^2)}$; and as the impedances are inversely proportional to the currents, $1/Z = \sqrt{(1/R^2 + 1/X^2)}$, giving as before

$$Z = \frac{RX}{\sqrt{R^2 + X^2}}$$

These calculations give only the magnitudes of the impedances, without regard for the amount of the phase angle $\phi$, or whether it is leading or lagging. So for a complete specification of an impedance it is necessary to add this information.

The more advanced books extend the methods and scope of circuit calculation enormously; but before going on to them one should have thoroughly grasped the contents of these last two chapters. Some of the most important results can be summarized like this:

| | | | | Current |
|---|---|---|---|---|
| Impedance, $Z$ ohms | Resistance, $R$ ohms | | | In phase with e.m f. |
| | Reactance, $X$ ohms | Capacitive, $X_C = 1/2\pi fC$ | Leads | e.m.f. by 90° |
| | | Inductive, $X_L = 2\pi fL$ | Lags | |

| | | |
|---|---|---|
| Impedance of $R$ and $X$ | in series | $Z = \sqrt{(R^2 + X^2)}$ |
| | in parallel | $Z = \dfrac{RX}{\sqrt{(R^2 + X^2)}}$ |

But the most important thing is not to memorize results but to understand the methods used for obtaining them, so as to be able to tackle new problems.

## Transformers

On p. 69 we had a brief introduction to the transformer, which is an appliance very widely used for radio and electrical purposes. The main features of its action can be grasped by considering the simple arrangement shown in Fig. 7.7*a*. Such complications as resistance of the windings are conveniently assumed

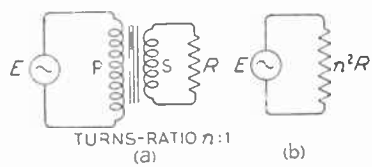



**Fig. 7.7—(*a*) Iron-cored transformer drawing current from the generator $E$ and delivering it, at a voltage equal to $E/n$, to the load $R$. From the point of view of the generator, diagram *b* represents an equivalent circuit (assuming a perfect trausformer)**

to be absent. An iron core is used, not only to increase the amount of magnetic flux and hence the e.m.f. generated by a given current, but also to ensure that practically all of it links both primary and secondary windings (p. 64). It will be seen later that transformers used for high-frequency a.c. often have no iron core, and only part of the field due to one winding links with the other; in other words, the coupling is loose. It is much simpler, to begin with, to assume that the coupling is 100%.

Firstly, consider what happens when the resistance $R$ is unconnected. There being no resistance in the primary circuit either, the generator e.m.f. $E$ is opposed solely by the back voltage caused by the varying of the magnetic flux. The situation is exactly as described with reference to Fig. 7.1—a purely inductive load. The vector diagram is as Fig. 7.8*a*, where $I_m$ denotes the current. As we have seen (p. 93) it lags the applied voltage $E$ by 90° and sets up an alternating magnetic flux $\Phi$ in phase with itself. It is this flux that induces the e.m.f. $E_p$ which, in order to prevent any further growth of current, must automatically be equal and opposite to $E$. Provision of the right amount of flux calls for a certain number of ampere-turns; with an iron core that number is far smaller than it would be without. If the number of turns is large the current is small. It is called the *magnetizing current.*

How about the secondary coil? Because $R$ is disconnected, the circuit is broken and no current can flow; but, as we have assumed that the whole of the sinusoidally varying magnetic flux associated with the primary links with the secondary too, an e.m.f. is generated

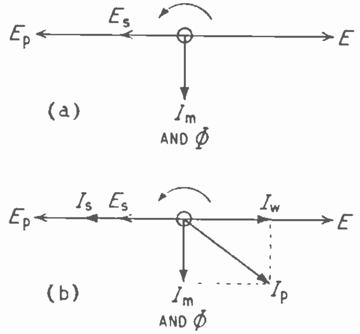



**Fig. 7.8—Simplified vector diagrams for a transformer; (*a*) unloaded, and (*b*) loaded**

in the secondary. The e.m.f. generated by a given change of flux linked with a coil is proportional to the number of turns. Since the e.m.f. generated in the primary is $E_p$, and the secondary has one $n$th as many turns, the secondary voltage $E_s = E_p/n = - E/n$.

For example, suppose P is connected to a 200-V supply. The magnetizing current rises until sufficient alternating magnetic flux is developed to generate an opposing 200 V. Suppose 100 amp-turns are needed to do this. Then if P has 2,000 turns, the magnetizing current is 100/2,000 = 0·05 A. If the transformer were required to supply current at the voltage for heating the valves in a radio receiver, which is 6·3 V, then $n$ would have to be 200/6·3 = 32, and S would have 2,000/32 = 63 turns. In practice it is quite usual to have several secondary windings delivering different voltages for various purposes, all energized by a single primary.

## The Primary Load Current

Now suppose $R$ to be connected, and its value to be 2·6 Ω. The secondary voltage, 6·3, causes a current of 6·3/2·6 = 2·4 A, which thus introduces a magnetizing force of 2·4 × 63 = 151 amp-turns. As the secondary current is flowing in a resistive circuit, it must be in phase with the secondary e.m.f. $E_s$, so at least as regards phase is represented in Fig. 7.8*b* by $I_s$. The original 100 amp-turns due to $I_m$ are now augmented by 151 due to $I_s$, 90° out of phase. The total is obviously wrong, both in magnitude and phase, for generating the e.m.f. $E_p$ needed to balance $E$ exactly. The only thing that can happen to end this impossible situation is for sufficient additional current to flow in the primary to provide 151 amp-turns in exactly opposite phase to those due to the secondary, neutralizing them. If you are guaranteed a net income of £100 a year, and your expenses in a certain year are £151, the matter can only be put right by paying an additional £151 into your account to neutralize the outgoings.

The extra primary current—in this case 151/2,000 = 0·0755 A—is represented in Fig. 7.8*b* by $I_w$, exactly opposite in phase to $I_s$, and is called the *load current*, because it is the current needed to make good the results of connecting the secondary load, $R$. Even if one did not know how a transformer worked, one would expect it to be in phase with $E$, for the secondary is delivering $6{\cdot}3 \times 2{\cdot}4 = 15{\cdot}1$ watts of power, and this can only come from the supply connected to the primary. 15·1 W at the supply voltage 200, is $15{\cdot}1/200 = 0{\cdot}0755$ A, which agrees with what we have already found in another way.

$I_p$ is the total primary current; in our example it would be $\sqrt{(0{\cdot}05^2 + 0{\cdot}0755^2)} = 0{\cdot}9$ A.

## Transformer Losses

In practice the magnetizing current is not quite wattless, because it has to flow through the primary winding, which inevitably has some resistance. Moreover, the alternating flux in the core generates a certain amount of heat in it, which has to come from somewhere, and in accordance with our wider definition of resistance (p. 48) can be considered as equivalent to some extra resistance in the primary. The primary resistance, augmented in this way, causes $I_m$ to lag $E$ by less than 90°.

Nor is the power from the supply transferred to the load without the transformer taking its rake-off. The primary and secondary load currents have to contend with the resistance of the windings, so there is bound to be some loss of voltage when $R$ is connected. The output wattage is therefore somewhat less than the input, the difference being dissipated as heat in the transformer.

The foregoing example is typical of a small " mains " transformer. Although the same principles apply, the emphasis is rather different in a.f. and r.f. types, which will be considered later.

## Impedance Transformation

We already know that (neglecting the losses) the secondary voltage given by a transformer with a $n:1$ turns ratio is $E/n$. You will probably have realized by now that the secondary current is $n$ times the primary load current. For one thing, it must be so to make the input and output wattages equal. For another, it must be so to make the primary load current exactly neutralize the secondary current by providing an equal number of amp-turns.

Now if the primary voltage ($E$) is $n$ times as great as the secondary voltage, and the primary load current is one-$n$th, connecting $R$ to the secondary draws the same load current from the supply as connecting a resistance equal to $n^2R$ directly to the supply. So, except for magnetizing current and losses, the transformer and its load $R$ can be replaced by a resistance $n^2R$ (Fig. 7.7*b*), without making any difference from the generator's point of view. This principle applies not only to resistances but to any type of impedance.

One of the uses of a transformer is for making a load of a certain impedance, say $Z_s$, equivalent to some other impedance, say $Z_p$. So it is often necessary to find the required turns ratio. Since $Z_p = n^2 Z_s$, it follows that $n$ must be $\sqrt{(Z_p/Z_s)}$.

Another use is for insulating the load from the generator. That is why 1 : 1 transformers are occasionally seen. But if it is not necessary to insulate the secondary winding from the primary, there is no need for two separate coils. The winding having the smaller number of turns can be abolished, and the connections tapped across the same number of turns forming part of the other winding. This device is called an *auto-transformer*.

CHAPTER 8

# The Tuned Circuit

## Inductance and Capacitance in Series

In the previous chapter we have seen what happens in a circuit containing capacitance *or* inductance (with or without resistance), and the question naturally arises: what about circuits containing both? We know that two or more reactances of either the inductive or capacitive kind in series can be combined just like resistances, by adding; and either sort of reactance can be combined with resistance by the more complicated square-root process. The outcome of combining reactances of opposite kinds is so fundamental to wireless that it needs a chapter to itself.

We shall begin by considering the simplest possible series circuit, Fig. 8.1. The method is exactly the same as for reactance and resistance in series; that is to say, since the current is common to

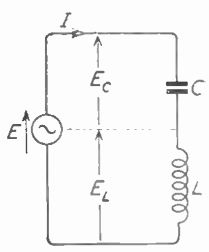



**Fig. 8.1—Circuit consisting of an a.c. generator in series with capacitance and inductance**

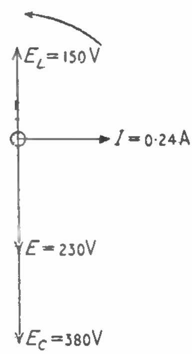



**Fig. 8.2—Vector diagram relating to Fig. 8.1. when $E$ = 230 V, $X_c$ = 1,590 Ω and $X$ = 628 Ω**

both it is easiest to start from it and work backwards to find the e.m.f. needed to drive it. We already know how to find the separate e.m.fs for *C* and *L*; Figs. 6.4 and 6.10 are respectively the waveform and vector diagrams for *C*, and Figs 7.2 and 7.4 for *L*. It is therefore an easy matter to combine them in a diagram to represent the Fig. 8.1 situation. With either type of diagram it is clear that the e.m.fs are exactly in opposition; 180° out of phase, as one would say. This can only mean that one e.m.f. must be *subtracted* from the other to give the necessary driving e.m.f. And since reactance is equal to e.m.f. divided by current (the same in both cases) it follows that the total reactance of the circuit is also equal to one of the separate reactances *less* the other.

Which e.m.f. or reactance must be subtracted from which to give the total? There is no particular reason for favouring either, but the agreed convention is to call $X_L$ positive and $X_C$ negative.

For example, suppose $L$ in Fig. 8.1 is 2 H, $C$ is 2$\mu$F, and E is 230 V at 50 c/s. We have already calculated the reactances of 2 H and 2$\mu$F at 50 c/s (p. 92 and p. 84) and found them to be 628 $\Omega$ and 1,590 $\Omega$. So the total reactance must be $628 - 1{,}590 = -962\ \Omega$, which at 50 c/s is the reactance of a 3·3 $\mu$F capacitor. The generator, then, would not notice any difference if a 3·3 $\mu$F capacitor were substituted for the circuit consisting of 2 $\mu$F in series with 2 H. The magnitude and phase of the current would be the same in both cases; namely, 230/962 = 0·24 A, leading by 90°.

To make sure, let us check it by calculating the voltage needed to cause this current. $E_C$ is equal to $IX_C = 0{\cdot}24 \times 1{,}590 = 380$ V, and its vector must be drawn so that $I$ leads it by 90° (Fig. 8.2). $E_L$ is $IX_L = 0{\cdot}24 \times 628 = 150$ V, drawn so that $I$ lags it by 90°. The resultant of $E_C$ and $E_L$, represented in Fig. 8.2 by the net effect of a distance downwards representing 380 V and one upwards of 150 V, is clearly 230 V.

The fact that the voltage across the capacitor is greater than the total supplied may be difficult to believe at first, but it is nothing to what we shall see soon!

## $L$, $C$ and $R$ all in Series

Bringing $R$ into the circuit introduces no new problem, because we have just found that $L$ and $C$ can always be replaced (for purposes of the calculation) by either $L$ *or* $C$ of suitable value, and the method of combining this with $R$ was covered in the previous chapter. Elaborating the equation given therein (p. 87) to cover the new information, we have

$$Z = \sqrt{(X_L - X_C)^2 + R^2}$$

Faced with a circuit like Fig. 8.3, we might begin by combining $R_1$ with $X_C$ and $R_2$ with $X_L$, afterwards combining the two results. But since neither of these pairs would be either a pure resistance or a

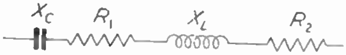



**Fig. 8.3—Capacitance, inductance and resistance in series**

pure reactance, we should have no immediate knowledge of the relative phases of the voltages across them. The final stage of the process would therefore be outside the range of the methods we have discussed. We can get round the difficulty by first subtracting $X_C$ from $X_L$ to find the total reactance of the circuit, then finding the total resistance by adding $R_1$ to $R_2$, and finally working out the impedance as for any other simple combination of reactance and resistance. The fact that neither the two reactances nor the two resistances are neighbours in the circuit does not matter, for the same current flows through all in series.

## The Series Tuned Circuit

We have already seen that the reactance of a capacitor falls and that of an inductor rises (p. 90) as the frequency of the current supplied to them is increased. It is therefore going to be interesting to study the behaviour of a circuit such as Fig. 8.4 over a range of frequencies. For the values given on the diagram, which are typical of practical broadcast receiver circuits, the reactances of coil and capacitor for all frequencies up to 1,800 kc/s are plotted as curves in Fig. 8.5. The significant feature of this diagram

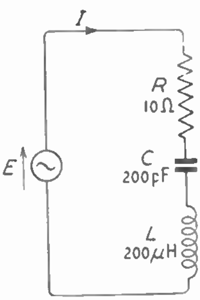



**Fig. 8.4—The way the reactances in this circuit vary with the frequency of *E*, graphed in Fig. 8.5, leads to interesting conclusions regarding this type of circuit in general**

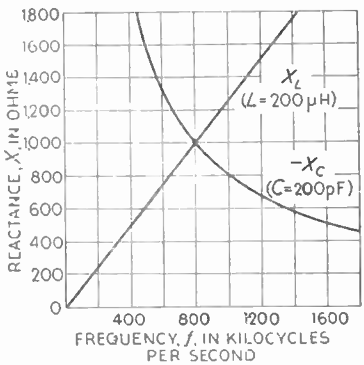



**Fig. 8.5—Reactances of the coil and capacitor in Fig. 8.4 plotted against frequency**

is that at one particular frequency, about 800 kc/s, $L$ and $C$ have equal reactances, each amounting to about 1,000 Ω. So the total reactance, being the difference between the two separate reactances, is zero. Put another way, the voltage developed across the one is equal to the voltage across the other; and since they are, as always, in opposition, the two voltages cancel out exactly. The circuit would therefore be unaltered, so far as concerns its behaviour *as a whole* to a voltage of this particular frequency, by completely removing from it both $L$ and $C$. This, leaving only $R$, would result in the flow of a current equal to $E/R$.

Let us assume a voltage not unlikely in broadcast reception, and see what happens when $E = 5$ mV. The current at 800 kc/s is then $5/10 = 0{\cdot}5$ mA, and this current flows, not through $R$ only, but through $L$ and $C$ as well. Each of these has a reactance of 1,000 Ω at this frequency; the potential difference across each of them is therefore $0{\cdot}5 \times 1{,}000 = 500$ mV, which is just one hundred times the voltage $E$ of the generator to which the flow of current is due.

That so small a voltage should give rise to two such large voltages elsewhere in the circuit is one of the queer paradoxes of alternating currents that make wireless possible. If the foregoing paragraphs have not made clear the possibility of the apparent absurdity, try making a waveform or vector diagram for the circuit of Fig. 8.4 at 800 kc/s; it will show that the presence of two large voltages in opposition is inevitable.

## Magnification

In the particular case we have discussed, the voltage across the coil (or across the capacitor) is one hundred times that of the generator. This ratio is called the *magnification* of the circuit.

We have just worked out the magnification for a particular circuit; now let us try to obtain a formula for any circuit. The magnification is equal to the voltage across the coil divided by that from the generator, which is the same as the voltage across $R$. If $I$ is the current flowing through both, then the voltage across $L$ is $IX_L$, and that across $R$ is $IR$. So the magnification is equal to $IX_L/IR$, which is $X_L/R$ or (because in the circumstances considered $X_C = X_L$) $X_C/R$. At any given frequency, it depends solely on $L/R$, the ratio of the inductance of the coil to the resistance of the circuit.

To obtain high magnification of a received signal (for which the generator of Fig. 8.4 stands), it is thus desirable to keep the resistance of the circuit as low as possible.

The symbol generally used to denote this voltage magnification is $Q$ (not to be confused with $Q$ denoting quantity of electricity).*

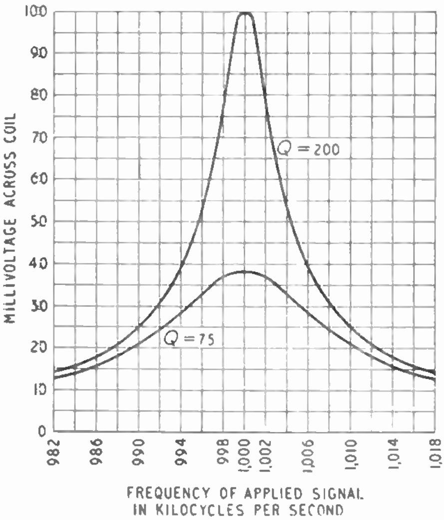
## Resonance Curves

At other frequencies the impedance of the circuit is greater, because in addition to $R$ there is some net reactance. At 1,250 kc/s, for example, the individual reactances are 1,570 and 636 Ω (see Fig. 8.5), leaving a total reactance of 934 Ω. Compared with this, the resistance is negligible, so that the current, for the same driving voltage of 5 mV, will be 5/934 mA, or roughly 5 μA. This is approximately one hundredth of the current at 800 kc/s. Passing through the reactance of $L$ (1,570 Ω) it gives rise to a voltage across it of $1{,}570 \times 5 = 7{,}850$ μV = 7·85 mV, which is only about one sixty-fourth as much as at 800 kc/s.

By extending this calculation to a number of different frequencies we could plot the current in the circuit, or the voltage developed across the coil, against frequency. This has been done for two circuits in Fig. 8.6. The only difference between the circuits is that

* In any real tuned circuit the coil is not pure inductance as shown in Fig. 8.4, but contains the whole or part of $R$ and also has some stray capacitance in parallel with it. Consequently the magnification obtained in practice may not be exactly equal to $Q$ (as calculated by $X_L/R$); but the difference is usually unimportant.

**Fig. 8.6—Showing how the voltage developed across the coil in a tuned circuit varies with frequency. Curves are plotted for $L=180\ \mu H$, $C=141$ pF, $E$ (injected voltage) $= 0{\cdot}5$ V, and $R = 15\ \Omega$ for the $Q = 75$ circuit and $5{\cdot}66\ \Omega$ for the $Q = 200$ circuit**




in one the resistance is 15 Ω, giving $Q = 75$, and in the other it is 5·66 Ω, giving $Q = 200$. In both, the values of $L$ and $C$ are such that their reactances are equal at 1,000 kc/s. These curves illustrate one thing we already know—that the *response* (voltage developed across the coil) when the reactances are equal is proportional to $Q$. They also show how it falls off at frequencies on each side, due to unbalanced reactance. At frequencies well off 1,000 kc/s the reactance is so much larger than the resistance that the difference between the two circuits is insignificant. The shapes of these curves show that the response of a circuit of the Fig. 8.4 type is far greater to voltages of one particular frequency—in this case 1,000 kc/s—than to voltages of substantially different frequencies. The circuit is said to be *tuned* to, or to *resonate* to, 1,000 kc/s; and the curves are called *resonance* curves. This electrical resonance is very closely analogous to acoustical resonance: the way in which hollow spaces or pipes magnify sound of a particular pitch.

The principle on which a receiver is tuned is now beginning to be evident: by adjusting the values of $L$ or $C$ in a circuit such as that under discussion one can make it resonate to any desired frequency. Any signal voltages received from the aerial at that frequency will have preferential amplification; and the desired transmission, distinguished from the rest by its frequency, will be received to the comparative exclusion of the others.

## Selectivity

This ability to pick out signals of one frequency from all others is called *selectivity*. It is an even more valuable feature of tuned circuits than magnification. There are alternative methods (which we shall consider later) of magnifying incoming signal voltages, but by themselves they would be useless, because they fail to distinguish between the desired programme and others.

The way the curves in Fig. 8.6 have been plotted focuses attention on how the value of $Q$ affects the response at resonance. The comparative selectivity is more easily seen, however, if the curves are

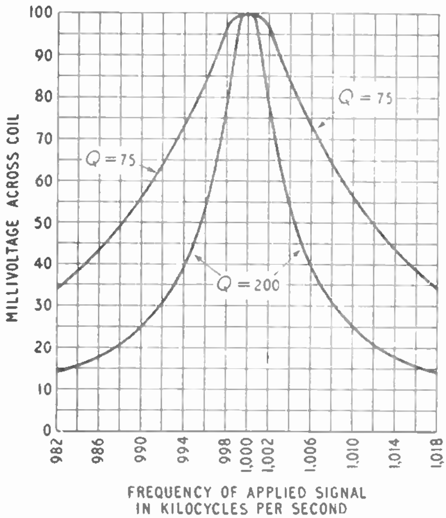



**Fig. 8.7—In this diagram the curve " $Q$ = 200 " is the same as in Fig. 8.6, but, to enable the selectivity of the $Q$=75 circuit to be more easily compared, the voltage injected into it has been raised from 0·5 mV to 1·33 mV, so as to make the output voltage at resonance equal to that across the $Q$ = 200 circuit**

plotted to give the same voltage at resonance, as has been done in Fig. 8.7. It must be borne in mind, of course, that this implies raising the input to the low-$Q$ circuit from 0·5 V to 1·33 V. Since the maximum voltage across both the coils is now 100 mV, the scale figures can also be read as percentage of maximum. Fig. 8.7 brings out more clearly than Fig. 8.6 the superior selectivity of the high-$Q$ circuit. For a given response to a desired station working on 1,000 kc/s, the 200-$Q$ response to 990 kc/s voltages is less than half that of the 75-$Q$ circuit. In general, the rejection of frequencies well off resonance is nearly proportional to $Q$.

Broadcasting stations in this waveband work at intervals of 9 or 10 kc/s, and although a $Q$ of 200 is rather better than the average, Fig. 8.6 shows that the response at 10 kc/s off resonance is still as much as one quarter that at resonance. This amount of selectivity is quite insufficient to keep out a possible interfering station, especially if it is stronger than the wanted one. So nearly all receivers include several tuned circuits. A second one of the same $Q$ would reduce the quarter to one sixteenth; a third would reduce it to a quarter of one sixteenth—one sixty-fourth—and so on.

## Frequency of Resonance

It is obviously important to be able to calculate the frequency at which a circuit containing known $L$ and $C$ resonates, or (performing the same process in reverse) to calculate the $L$ and $C$ required to tune to a given frequency. The required equation follows easily from the fact that resonance takes place at the frequency which makes the reactance of the coil equal that of the capacitor:

$$2\pi f_r L = \frac{1}{2\pi f_r C}$$

where the symbol $f_r$ is used to denote the frequency of resonance. Rearranging this, we get

$$f_r^2 = \frac{1}{(2\pi)^2 LC}, \qquad \text{so } f_r = \frac{1}{2\pi\sqrt{LC}}$$

If $L$ and $C$ are respectively in henries and farads, $f_r$ will be in cycles per second; if $L$ and $C$ are in henries and microfarads, $f_r$ will be in kc/s. But perhaps the most convenient units for radio purposes are $\mu$H and pF, and when the value of $\pi$ has been filled in the result is

$$f_r \text{ (in kc/s)} = \frac{159{,}200}{\sqrt{LC}} \quad \text{or} \quad f_r \text{ (in Mc/s)} = \frac{159 \cdot 2}{\sqrt{LC}}$$

If the answer is preferred in terms of wavelength, we make use of the relationship $f = 3 \times 10^8/\lambda$ (p. 22) to give

$$\lambda_r = 1 \cdot 885 \sqrt{LC} \qquad (\lambda_r \text{ in metres; } L \text{ in } \mu\text{H; } C \text{ in pF})$$

One thing to note is that $f_r$ and $\lambda_r$ depend on $L$ multiplied by $C$; so in theory a coil of any inductance can be tuned to any frequency by using the appropriate capacitance. In practice the capacitance cannot be reduced indefinitely, and there are disadvantages in making it very large. Most variable tuning capacitors for the medium-frequency band give a range of about 30 to 530 pF. The capacitance of the wiring and circuit components necessarily connected in parallel may add another 30 pF. This gives a ratio of maximum to minimum of $560/60 = 9 \cdot 35$. But the square root sign in the above equations means that if the inductance is kept fixed the ratio of maximum to minimum $f_r$ (or $\lambda_r$) is $\sqrt{9 \cdot 35}$, or $3 \cdot 06$. Any band of frequencies with this range of maximum to minimum can be covered with such a capacitor, the actual frequencies in the band being dependent on the inductance chosen for the coil.

Suppose we wished to tune from 1,605 kc/s to 1,605/3·06 or 525 kc/s, corresponding to the range of wavelengths 187 to 572 metres. For the highest frequency or lowest wavelength the capacitance will have its minimum value of 60 pF; by filling in this value for $C$ and 1·605 Mc/s for $f_r$, we have $1{\cdot}605 = 159{\cdot}2/\sqrt{60L}$, from which $L = 164\ \mu$H.*

If we calculate the value of $L$ necessary to give 0·525 Mc/s with a capacitance of 560 pF, it will be the same.

In the same way the inductance needed to cover the short-wave band 9·8 to 30 Mc/s (30·6 to 10 metres) can be calculated, the result being 0·471 $\mu$H.

Notice how large and clumsy numbers are avoided by a suitable choice of units, but of course it is essential to use the equation having the appropriate numerical " constant ". If in any doubt it is best to go back to first principles ($2\pi f_r L = 1/2\pi f_r C$) and use henries, farads, and c/s.

## $L$ AND $C$ IN PARALLEL

If the input voltage, instead of being applied in series with $L$ and $C$ as in Fig. 8.1, is in parallel (Fig. 8.8), the result is a system which is in many ways similar, but in which voltage and current have changed

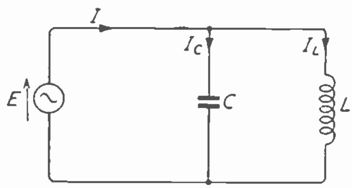



**Fig. 8.8—Parallel resonant circuit, which may be compared with the series resonant circuit, Fig. 8.1**

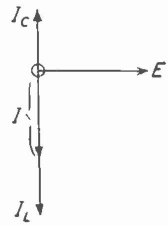



**Fig. 8.9—Vector diagram relating to Fig. 8.8**

places. (We noticed a similar relationship when, on p. 68, we compared inductance with capacitance.) In elucidating the series circuit we started with the current, because that was common to both $L$ and $C$, and found the voltages across each and hence across the whole. In Fig. 8.8 the currents through $L$ and $C$ are not necessarily equal, but the voltages across them obviously must be, both being equal to $E$. So we start the vector diagram (Fig. 8.9) by drawing an $E$ vector.

With no resistance, the current in the $L$ branch will be determined by the reactance $2\pi fL$ of the coil; it will be $E/2\pi fL$. In the $C$ branch, it will similarly be $E/(1/2\pi fC) = E.2\pi fC$. We know, too, that $I_L$ will lag $E$ by 90°, and $I_C$ will lead $E$ by 90°; so the

* For this reverse process it saves time to adapt the formula, thus:

$$f_r = 159{\cdot}2/\sqrt{LC} \quad (\mu\text{H, pF, Mc/s}).$$
$$\therefore f_r^2 = 159{\cdot}2^2/LC$$
$$\therefore L = 25{,}350/f_r^2C$$

The corresponding adaptation for calculating $C$ is

$$C = 25{,}350/f_r^2L$$

currents must be in exactly opposite phase, as were the voltages in Fig. 8.2. The net current taken from the generator will, therefore, be the simple difference between the two individual currents.

Suppose now that the frequency is such as to make the reactances equal. The currents will therefore be equal, so that the difference between them will be zero. We then have a circuit with two parallel branches, both with currents flowing in them, and yet the current supplied by the generator is nil!

It must be admitted that such a situation is impossible, the reason being that no practical circuits are entirely devoid of resistance. But if the resistance is small, when the two currents are equal the system does behave approximately as just described. It is a significant fact that the frequency which makes them equal is the same as that which in a series circuit would make $L$ and $C$ resonate.

## The Effect of Resistance

When the resistance of a circuit such as Fig. 8.8 has been reduced to a minimum, the resistance in the $C$ branch is likely to be negligible compared with that in the $L$ branch. Assuming this to be so, let us consider the effect of resistance in the $L$ branch (Fig. 8.10), representing it by the symbol $r$ as a reminder that it is small compared with the reactance. We have already studied this branch on its own (Fig. 7.3) and noted that one effect of the resistance is to cause the phase angle, by which the current lags the total applied voltage, to be less than 90°; the greater the resistance the less the angle (Fig. 7.4). Applying the same principle, we modify Fig. 8.9 as shown in Fig. 8.11*a*. (If you have any difficulty in seeing that the principle is the same, look at Fig. 7.4 with the right-hand edge of the page at the bottom, so that the two diagrams are in phase with one

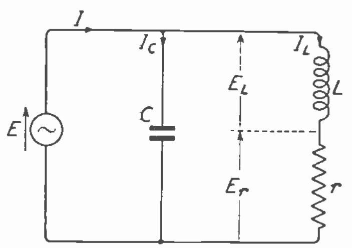



**Fig. 8.10—Parallel resonant circuit with resistance in the inductive branch, which may be compared with Fig. 8.4**

another.) The angle between $E$ and $I_L$ is now less than 90°, but since we are assuming that the resistance in the $C$ branch is negligible we must draw the $I_C$ vector at right angles to $E$. We are also assuming that $I_C = I_L$, but as they are now not exactly opposite in phase the result of combining them is a small current $I$, being the current supplied by the generator. To avoid confusing Fig. 8.11*a* this combining process has been transferred to Fig. 8.11*b*.

One detail that should be noticed about Fig. 8.11*b* is that although $I$ is nearly in phase with $E$ it is not exactly so. If $I_C = I_L$, $I$ must be exactly midway between them in phase, whereas $E$ is 90° from $I_C$ and less than 90° from $I_L$. It is easy to see that $I$ could be brought exactly into phase with $E$ by a slight reduction in frequency, which would increase $I_L$ and reduce $I_C$. But if (as we are supposing) $r$ is

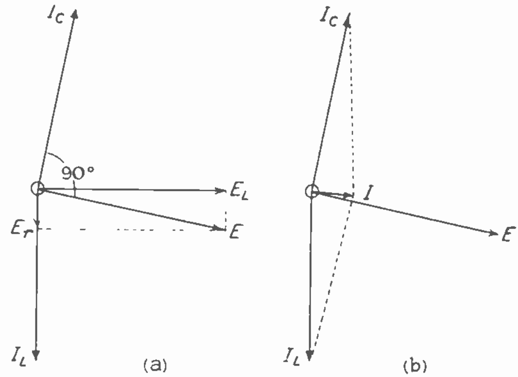



**Fig. 8.11—(*a*) Modification of Fig. 8.9 to allow for the presence of resistance as in Fig. 8.10**

much smaller than $X_L$ there is no need to worry about this, because the difference in phase between $E$ and $I$ will then be negligible. In Fig. 8.11 we have had to make $r$ nearly a quarter of $X_L$ to show any difference at all, whereas in practice $r$ is more likely to be 10–50 times smaller still. And of course that will make $I$ very small indeed compared with $I_L$ or $I_C$.

What all this amounts to is that under the conditions named (inductive reactance of the coil much greater than its resistance, and frequency such as to make the currents in the two branches equal) the current that has to be supplied by the generator is very small and practically in phase with its e.m.f. In other words, it is identical with the current that would flow if the two reactive branches were replaced by one consisting of a high resistance.

## Dynamic Resistance

It is a matter of considerable practical interest to know how high this resistance is (let us call it $R$), and how it is related to the values of the real circuit: $L$, $C$ and $r$. We have already seen that if $r = 0$ the current $I$ is zero, so in that case $R$ is infinitely large. Fig. 8.11 shows that as $r$ is increased the voltage $E_r$ increases, bringing $I_L$ more nearly into phase with $E$ and therefore with $I_C$, so that $I$ increases, representing a decrease in $R$. Whereas when $r$ is much less than $X_L$ the value of $R$ is much greater than $X_L$, you can find by drawing the vector diagram that when $r$ is about equal to $X_L$

then $R$ will be not very much more than $X_L$. So at least we can say that the smaller we make $r$ in Fig. 8.10, the greater is the resistance which the circuit as a whole presents to the generator.

It can be shown mathematically that, so long as $r$ is much less than $X_L$, the resistance $R$, to which the parallel circuit *as a whole* is equivalent at a certain frequency, is practically equal to $X_L^2/r$. The proof by algebra is given in standard textbooks, but geometry-minded readers should not find it difficult to prove from the vector diagram, Fig. 8.11.

To take an example, suppose $C$ and $L$ are the same as in Fig. 8.4, and $r$ is equal to 10 Ω (Fig. 8.12*a*). At 800 kc/s the reactances are each 1,000 Ω (Fig. 8.5). So $R$ is $1{,}000^2/10 = 100{,}000$ Ω (Fig. 8.12*b*). If $L$, $C$ and $r$ were hidden in one box and $R$ in another, each with a

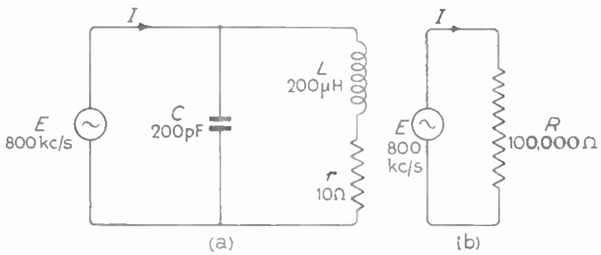



**Fig. 8.12—The circuit composed of $C$, $L$ and $r$ having the values marked ($a$) can *at 800 kc/s* be replaced by a high resistance ($b$), so far as the phase and magnitude of the current taken by the generator is concerned**

pair of terminals for connecting the generator, it would be impossible to tell, by measuring the amount and phase of the current taken, which box was which. Both would appear to be resistances of 100,000 Ω, *at that particular frequency*. We shall very soon consider what happens at other frequencies, but in the meantime it should be noted that the apparent or equivalent resistance, which we have been denoting by $R$, has the special name *dynamic resistance*.

## Parallel Resonance

Suppose $E$ in Fig. 8.12 is 1 V. Then under the conditions shown the current $I$ must be 1/100,000 A, or 10 μA. Since the reactances are each 1,000 Ω, and the impedance of the $L$ branch is not appreciably increased by the presence of $r$, the branch currents are each 1,000 μA. But let us now change the frequency to 1,000 kc/s. Fig. 8.5 shows that this makes $X_C$ and $X_L$ respectively 800 Ω and 1,250 Ω. The current $I_C$ taken by $C$ is then 1/800 A = 1,250 μA and $I_L$ is 1/1,250 A = 800 μA. Since $r$ is small, these two currents are so nearly in opposite phase that the total current is practically equal to the difference between them, 1,250 − 800 = 450 μA. This balance is on the capacitive side, so at 1,000 kc/s the two branches can be replaced by a capacitance having a reactance = $E/I$ = 1/0·00045 = 2,222 Ω, or 112·5 pF. The higher the frequency the greater is the current, and the larger this apparent capacitance; at very high frequencies it is practically 200 pF—as one would expect, because the reactance of $L$ becomes so high that hardly any current can flow in the $L$ branch.

Using the corresponding line of argument to explore the frequencies below 1,000 kc/s, we find that the current becomes larger

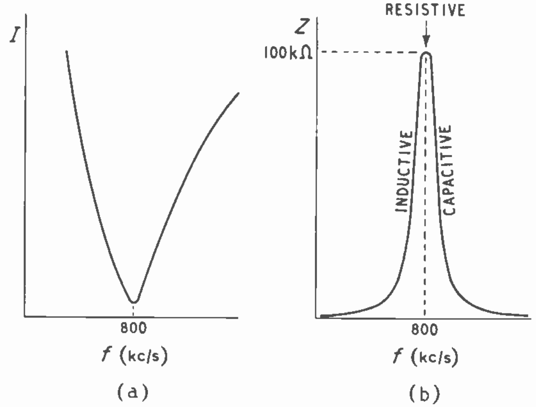



**Fig. 8.13 — Plotting the current and impedance of the Fig. 8.12a circuit as a whole at various frequencies gives the shapes shown here: (*a*) current, (*b*) impedance**

and larger, and that its phase is very nearly the same as if $C$ were removed, leaving an inductance rather greater than $L$ but not much greater at very low frequencies.

Plotting the current over a range of frequency, we get the result shown in Fig. 8.13*a*, which looks like a series-circuit resonance curve turned upside down.

The graph of impedance against frequency (Fig. 8.13*b*) is still more clearly a resonance curve.

## The Frequency of Parallel Resonance

Before we complicated matters by bringing in $r$ we noticed that the frequency which would make $L$ and $C$ resonate in a series circuit had the effect in a parallel circuit of reducing $I$ to zero and making the impedance infinitely great. In these circumstances the frequency of parallel resonance is obviously the same as that of series resonance. But actual circuits always contain resistance, which raises some awkward questions about the frequency of resonance. In a series circuit it is the frequency that makes the two reactances equal. But in a parallel circuit this does not make the branch impedances exactly equal unless both contain the same amount of resistance, which is rarely the case. The two branch currents are not likely to be exactly equal, therefore. And even if they were, Fig. 8.11 shows that it would not bring the total current $I$ exactly into phase with $E$. So are we to go on defining $f_r$ as the frequency that makes $X_L = X_C$, or as the frequency which brings $I$

into phase with $E$ so that the circuit behaves as a resistance $R$, or as the frequency at which the current $I$ is least?

Fortunately it happens that in normal radio practice this question is mere hair-splitting; the distinctions only become appreciable when $r$ is a large-sized fraction of $X_L$; in other words, when $Q$ $(=X_L/r)$ is abnormally small—say less than 5. *On the understanding that such unusual conditions are excluded, therefore*, the following relationships are so nearly true that the inaccuracy is of no practical importance. The frequency of parallel resonance ($f_r$) is the same as that for series resonance, and it makes the two reactances $X_L$ and $X_C$ equal. The impedance at resonance ($R$) is resistive, and equal to $X_L^2/r$ and also to $X_C^2/r$. And, because $X_L = X_C$ at resonance,

$$R = \frac{X_L X_C}{r} = \frac{2\pi f_r L}{2\pi f_r C r} = \frac{L}{Cr}$$

from which
$$r = \frac{L}{CR}$$

In all these, $r$ can be regarded as the whole resistance in series with $L$ and $C$, irrespective of how it is distributed between them. Since $Q = X/r$,

$$R = QX, \quad r = \frac{X}{Q}, \quad \text{and} \quad \frac{R}{Q} = \frac{Q}{r}.$$

All these relationships assume resonance.

## Series and Parallel Resonance Compared

The resonance curve in Fig. 8.13*b* showing $R = 100\ \text{k}\Omega$ applies to the circuit in Fig. 8.12*a*, in which $Q = X_L/r = 2\pi f L/r = 2\pi \times 800{,}000 \times 0{\cdot}0002/10 = 100$. If $r$ were reduced to 5 Ω, $Q$ would be 200, and $R$ would be 200 kΩ. But reducing $r$ would have hardly any effect on the circuit at frequencies well off resonance; so the main result would be to sharpen the resonance peak. Increasing $r$ would flatten the peak. This behaviour is the same as with series resonance, except that a peak of impedance takes the place of a peak of voltage across the coil or of current from the generator. Current from the generator to the parallel circuit reaches a minimum at resonance (Fig. 8.13*a*), as does impedance of the series circuit.

At resonance a series circuit is (except for $r$) a perfect short-circuit to the generator, so is called an *acceptor circuit*. A parallel circuit at resonance is (except for $R$, which is an inversion of $r$) a perfect open-circuit to the generator, so is called a *rejector circuit*. One should not be misled by these names into supposing that there is some inherent difference between the circuits; the only difference is in the way they are used. To a generator connected in one of the branches, Fig. 8.12*a* is an acceptor circuit at the same time as it is a rejector circuit to the generator shown.

If the generator in a circuit such as Fig. 8.12*a* were to deliver a constant current instead of a constant e.m.f., the voltage developed across it would be proportional to its impedance, and so would vary with frequency in the manner shown in Fig. 8.13*b*. So the result is the same, whether a tuned circuit is used in series with a source of voltage or in parallel with a source of current. In later chapters we shall come across examples of both.

## THE RESISTANCE OF THE COIL

In the sense that it cannot be measured by ordinary direct-current methods—by finding what current passes through on connecting it across a 2 V cell, for example—it is fair to describe $R$ as a fictitious resistance. Yet it can quite readily be measured by any method suitable for measuring resistances at the frequency to which the circuit is tuned; in fact, those methods by themselves would not disclose whether the thing being measured was the dynamic resistance of a tuned circuit or the resistance of a resistor.

If it comes to that, even $r$ is not the resistance that would be indicated by any d.c. method. That is to say, it is not merely the resistance of the wire used to make the coil. Although no other cause of resistance may be apparent, the value of $r$ measured at high frequency is always greater, and may be many times greater, than the d.c. value.

One possible reason for this has been mentioned in connection with transformers (p. 97). If a solid iron core were used, it would have currents generated in it just like any secondary winding: the laws of electromagnetism make no distinction. These currents passing through the resistance of the iron represent so much loss of energy, and, as we have seen, that necessitates some e.m.f. in phase with the primary current, just as if there were some extra resistance in the coil to be overcome. To stop these *eddy currents*, as they are called, iron cores are usually made up of thin sheets arranged so as to break up circuits along the lines of the induced e.m.f. At radio frequencies even this is not good enough, and if a magnetic core is used at all it is either iron in the form of fine dust bound together by an insulator, or a non-conducting magnetic material called a *ferrite*.

Another source of loss in magnetic cores, called *hysteresis*, is a lag in magnetic response, which shifts the phase of the primary current so that it is partly in phase with the applied e.m.f., and therefore represents resistance, or energy lost (p. 48).

The warming-up of cores in which a substantial number of watts are being lost in these ways is very noticeable.

Although an iron core introduces the equivalent of resistance into the circuit, its use is worth while if it enables a larger amount of resistance to be removed as a result of fewer turns of wire being needed to give the required inductance. Another reason for using iron cores, especially in r.f. coils, is to enable the inductance to be varied, by moving the core in and out.

Even the metal composing the wire itself has e.m.fs induced in it in such a way as to be equivalent to an increase in resistance. Their distribution is such as to confine the current increasingly to the surface of the wire as the frequency is raised. This *skin effect* occurs even in a straight wire; but when wound into a coil each turn lies in the magnetic field of other turns, and the resistance is further increased; so much so that using a thicker gauge of wire sometimes actually increases the r.f. resistance of a coil.

## Dielectric Losses

In the circuits we have considered until now, $r$ has been shown exclusively in the inductive branch. While it is true that the resistance of the capacitor plates and connections is usually negligible (except perhaps at very high frequencies), insulating material coming within the alternating electric field introduces resistance rather as an iron core does in the coil. It is as if the elasticity of the " leashes " (p. 59) were accompanied by a certain amount of friction, for the extra circuit current resulting from the electron movements is not purely capacitive in phase. The result is the same as if resistance were added to the capacitance.

Such materials are described as poor dielectrics, and obviously would not be chosen for interleaving the capacitor plates. Air is an almost perfect dielectric, but even in an air capacitor a part of the field traverses solid material, especially when it is set to minimum capacitance. Apart from that, valve holders, valve bases, terminal blocks, wiring, and other parts of the circuit—including the tuning coil—have " stray " capacitances; and if inefficient dielectrics are used for insulation they will be equivalent to extra resistance.

## R.F. Resistance

All these causes of loss in high-frequency circuits come within the general definition of resistance (p. 48), as (electrical power dissipated) ÷ (current-squared). In circuit diagrams and calculations it is convenient to bring it all together, with the ordinary wire resistance, in a single symbol such as in Fig. 8.12*a*, in conjunction with a perfectly pure inductance and capacitance, $L$ and $C$. But we have seen that at resonance the whole tuned circuit can also be represented as a resistance $R$ (Fig. 8.12*b*). We also know that a perfect $L$ and $C$ at resonance have an infinite impedance, so could be shown connected in parallel with $R$ as in Fig. 8.14; and as this tuned circuit would carry the large circulating currents $I_L$ and $I_C$ it would represent the actual circuit more closely. Furthermore, it represents the circuit not only at resonance but (except in so far as $R$ varies somewhat with frequency) at frequencies on each side of resonance.

More generally, for every resistance and reactance in series one can calculate the values of another resistance and reactance which, connected in parallel, at the same frequency, are equivalent. And if (as we are assuming) the reactance is much greater than the resistance, it has very nearly the same value in both cases. The values of series and parallel resistance, $r$ and $R$ respectively, are connected by the same approximate formula $R = X^2/r$, which we have hitherto regarded as applying only at the frequency of resonance (p. 109). This ability to reckon resistance as either in parallel or in

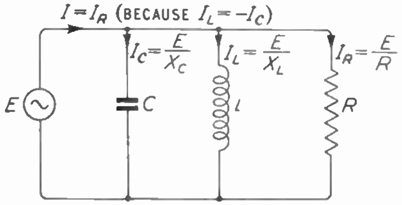



**Fig. 8.14—An alternative way of representing a circuit such as Fig. 8.12 (a). If $r$ is small compared with $X$, $R = X^2/r$**

series is very useful. It may be, for example, that in an actual circuit there is resistance connected both ways. Then for simplifying calculations it can be represented by a single resistance, either series or parallel.

An example, worked out with the help of relationships explained in this chapter, may make this clearer. A certain coil has, say, an inductance of 1·2 mH, and at 300 kc/s its r.f. resistance, reckoned as if it were in series with a perfect inductance, is 25 Ω (Fig. 8.15*a*). At that frequency the reactance, $X_L$, is $2\pi f L = 2\pi \times 300 \times 1{\cdot}2 =$ 2,262 Ω. So $Q_L = 2{,}262/25 = 90{\cdot}6$. ($Q_L$ is the $Q$ of the coil, defined as $X_L/r_L$; it is also the magnification of a resonant circuit consisting of this coil and a perfect capacitor.) The same coil can be represented as a perfect inductance of very nearly 1·2 mH in

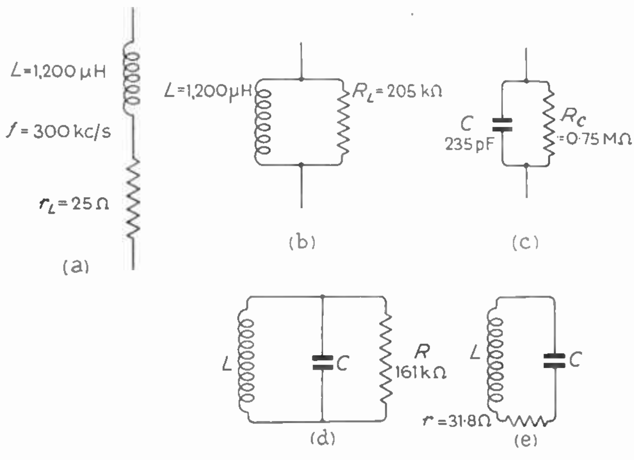



**Fig. 8.15—Example to illustrate series and parallel equivalents. *b* is equivalent to *a*; and *d* and *e* are alternative ways of expressing the result of combining *c* with *b***

parallel with a resistance $R_L = X_L^2/r_L = 2{,}262^2/25 = 205{,}000\ \Omega$ or 205 k$\Omega$ (Fig. 8.15*b*). If $r_L$ remained the same at 200 kc/s, the value of $R_L$ would be $(2\pi \times 200 \times 1{\cdot}2)^2/25 = 91$ k$\Omega$; so when the equivalent parallel value has been found for one frequency it must not be taken as holding good at other frequencies. As a matter of fact, neither $r_L$ nor $R_L$ is likely to remain constant, but in most practical coils $Q$ is fairly constant over the useful range of frequency. So it is likely that at 200 kc/s $R_L$ would be somewhere around 135 k$\Omega$, and $r_L$ would be about 17 $\Omega$.

The capacitance required to tune the coil to 300 kc/s, or 0·3 Mc/s, is (p. 106, footnote) $C = 25{,}350/f_r^2L = 25{,}350/(0{\cdot}3^2 \times 1{,}200) =$ 235 pF. This, of course, includes the tuning capacitor itself plus the stray capacitance of the wiring and of the coil. The whole of this is equivalent to a perfect 235 pF in parallel with a resistance of, shall we say, 0·75 M$\Omega$ (Fig. 8.15*c*). The reactance, $X_C$, is bound to be the same as $X_L$ at resonance; but to check it we can work it out from $1/2\pi fC = 1/(2\pi \times 0{\cdot}3 \times 0{\cdot}000235) = 10^6/(2\pi \times 0{\cdot}3 \times 235) =$ 2,262 $\Omega$. $Q_C$ is therefore $R_C/X_C = 750{,}000/2{,}262 = 332$. (This figure is the magnification of a resonant circuit consisting of this capacitor and a perfect coil, but as real coils are far from perfect it is more useful as a measure of the " goodness " of the capacitor.)

If the two components are united as a tuned circuit, the total loss can be expressed as a parallel resistance, $R$, by reckoning $R_L$ and $R_C$ in parallel: $R_LR_C/(R_L + R_C) = 0{\cdot}205 \times 0{\cdot}75/0{\cdot}955 = 0{\cdot}161$ M$\Omega$ (Fig. 8.15*d*). Although this is less than $R_L$, it indicates a greater loss. The $Q$ of the tuned circuit as a whole is $R/X = 161{,}000/2{,}262 = 71$. Incidentally, it may be seen that this is $Q_LQ_C/(Q_L + Q_C)$. The total loss of the circuit can also be expressed as a series resistance, $r = X^2/R = 2{,}262^2/161{,}000 = 31{\cdot}8\ \Omega$ (Fig. 8.15*e*). Calculating $Q$ from this, $X/r = 2{,}262/31{\cdot}8$, gives 71 as before.

Before going on, the reader would do well to continue calculations of this kind with various assumed values, taking note of any general conclusions that arise. For instance, a circuit to tune to the same frequency could be made with less inductance; assuming the same $Q$, would the dynamic resistance be lower or higher? Would an added resistance in series with the lower-$L$ coil have less or more effect on its $Q$? And what about a resistance in parallel?

CHAPTER 9

# Valves: The Simpler Types

## EMITTED ELECTRONS

SO FAR WE have considered an electric current as a stream of electrons along a conductor (p. 30). The conductor is needed to provide a supply of loose electrons, ready to be set in motion by an e.m.f. An e.m.f. applied to an insulator causes no appreciable current because the number of loose electrons in it is negligible.

To control the current it is necessary to move the conductor, as is done in a switch or rheostat (a variable resistor). This is all right occasionally, but quite out of the question when (as in radio) we want to vary currents millions of times a second. The difficulty is the massiveness of the conductor; it is mechanically impracticable to move it at such a rate.

This difficulty can be overcome by releasing electrons, which are inconceivably light, from the relatively heavy metal or whatever conductor is used. Although in constant agitation in the metal, they have not enough energy at ordinary temperature to escape beyond its surface. But if the metal is heated they " boil off " from it. A source of free electrons, such as this, is called a *cathode.* To prevent the electronic current from being hindered by the surrounding air, the space in which it is to flow is enclosed in a glass bulb and as much as possible of the air pumped out, giving a vacuum. We are now well on the way to manufacturing a *thermionic valve*, or, as it is called in America, a vacuum tube.

Cathodes are of two main types: *directly heated* and *indirectly heated.* The first, more usually known as a *filament*, consists of a fine wire heated by passing a current through it. To minimize the current needed, the wire is usually coated with a special material that emits electrons freely at the lowest possible temperature. Valves with this type of cathode are used chiefly in battery-driven sets.

The indirectly heated cathode is a very narrow tube, usually of nickel, coated with the emitting material and heated by a separate filament, called the *heater*, threaded through it. Since the cathode is insulated from the heater, three connections are necessary compared with the two that suffice when the filament serves also as the source of electrons, unless two are joined together internally.

In all essentials the two types of cathode work in the same way; in dealing with valves we shall therefore omit the heater or filament circuit altogether after the first few diagrams, indicating the cathode by a single connection. The operation of a valve depends upon the emission from the cathode; the means by which the cathode is heated to obtain this emission has little significance except in connection with the design of a complete receiver (p. 357).

The free electrons do not tend to move in any particular direction unless urged by an electric field. In the absence of such a field they accumulate in the space around the cathode, enclosing it in an electronic cloud. Because this cloud consists of electrons, it is a negative charge, known as the *space charge*, which repels new arrivals back again to the cathode, so preventing any further accumulation of electrons in the vacuous space.

## THE DIODE VALVE

To overcome the stoppage caused by the negative space charge it is necessary to apply a positive potential. This is introduced into the valve by means of a metal plate called the *anode*. The simplest type of valve—the diode—contains only two *electrodes*: cathode and anode.

When the anode is made positive relative to the cathode, it attracts electrons from the space charge, causing its repulsion to diminish so that more electrons come forward from the cathode. In this way current can flow through the valve and round a circuit such as in Fig. 9.1. But if the battery is reversed, so that the anode is more negative than the cathode, the electrons are repelled towards their source, and no current flows. The valve will therefore permit

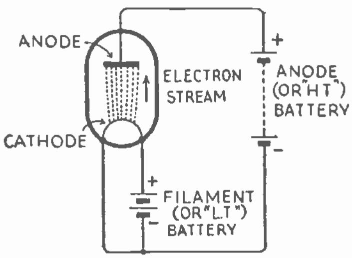



**Fig. 9.1—Electron flow from cathode to anode in a directly-heated diode valve**

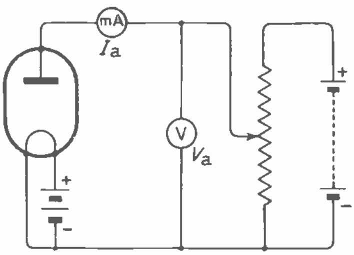



**Fig. 9.2—Circuit for finding the relationship between the anode voltage ($V_a$) and the anode current ($I_a$) in a diode**

current to flow through it in one direction only, like the air valve attached to a tyre, and this is why it was so named. It is one particular type of rectifier (p. 359). The anode battery or other supply is often called the h.t. (signifying " high tension ") and the filament or heater source the l.t. (" low tension ").

If the anode of a diode is slowly made more and more positive with respect to the cathode, as, for example, by moving the slider of the potentiometer in Fig. 9.2 upwards, the attraction of the anode for the electrons is slowly augmented and the current increases. To each value of anode voltage $V_a$ there corresponds some value of anode current $I_a$, and if each pair of readings is recorded on a graph a curve like that of Fig. 9.3 (called a valve $I_a/V_a$ *characteristic curve*) is obtained.

The shape of the curve shows that at low voltages the anode collects few electrons, being unable to overcome the repelling effect of the space-charge. The greater the positive anode voltage the greater the negative space charge it is able to neutralize; that is to say, the greater the number of electrons that can be on their way between cathode and anode; in other words, the greater the anode current. By the time the point C is reached the voltage is so high that electrons are reaching the anode practically as fast as the cathode can emit

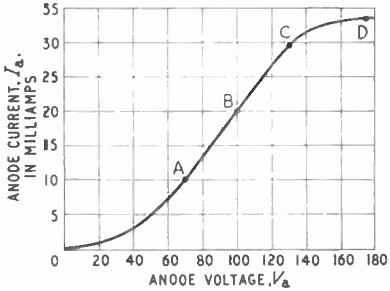



**Fig. 9.3—Characteristic curve of a diode**

them; a further rise in voltage collects only a few more strays, the current remaining almost constant from C to D and beyond. This condition is called *saturation*. Because the saturation current is limited by the emission of the cathode, and that depends on the temperature of the cathode, it is often called a *temperature-limited* current, to distinguish it from the *space-charge limited* current lower down the curve. In most modern valves—unlike the one whose characteristic is shown in Fig. 9.3—the saturation current is high above the working range.

At B an anode voltage of 100 V drives through the valve a current of 20 mA; the valve could therefore be replaced by a resistance of $100/20 = 5$ kΩ without altering the current flowing at this voltage. This value of resistance is therefore the equivalent d.c. resistance of the valve at this point. The curve shows that although the valve is in this sense equivalent to a resistor, it does not conform to Ohm's law; its resistance depends on the voltage applied. To drive 10 mA, for example, needs 70 V; $V/I = 70/10 = 7$ kΩ.

This is because the valve, considered as a conductor, is quite different from the material conductors with which Ohm experimented. Its so-called resistance—really the action of the space charge—is like true resistance in that it restricts the current that flows when a given voltage is applied, but it does not cause the current to be exactly proportional to the voltage as it was in Fig. 2.3 In other words, it is non-linear.

## ANODE A.C. RESISTANCE

One can, however, reckon the resistance of a valve in another way. The portion of the curve round about B is nearly linear (i.e., straight). Within it, an increase or decrease of 30 anode volts causes an increase or decrease in anode current of 10 mA. The resistance over this region of the curve would therefore appear to be 30/10 = 3 kΩ. This resistance is effective towards current *variations* within the range A to C; if, for example, a steady anode voltage of 100 V were applied (point B) and then an alternating voltage of peak value 20 V were superposed on this, the resulting alternating current through the valve would be 6·7 mA peak. Based on this, the resistance, as before, comes out to 20/6·7 = 3 kΩ. The figure derived in this way is the resistance offered to an alternating voltage superposed on a steady anode voltage; it is therefore called the *anode a.c. resistance* of the valve. Its importance in electronics is so great that it has had the special symbol $r_a$ allotted to it. It is also, but not so precisely, called the "impedance" of the valve. The number of steps up in current for one step up in voltage is the *slope* of the curve. But $r_a$ is the number of voltage steps for one current step, so it is 1 divided by the slope, or, as it is called, the reciprocal of the slope, at the particular point selected on it. As can be seen from Fig. 9.3, a steep slope means a low anode a.c. resistance.

The equivalent d.c. resistance of a valve is a quantity seldom used or mentioned; it was introduced here only for the sake of calling attention to the non-linearity of the valve's resistance.

## The Triode Valve

As we shall see in Chapter 16, the one-way traffic of the diode makes it very useful for converting a.c. to d.c.—*rectifying*. But if a wire grating is inserted between cathode and anode, so that, in order to reach the anode, electrons have to pass through it, a much fuller control of the electron current becomes possible. It enables us, in effect, to vary the space-charge, which is the only thing (short of saturation) affecting the amount of current that a given anode voltage causes to flow.

If the potential of this new electrode, the *grid*, is made positive with respect to the cathode it will assist the anode to neutralize the space charge, so increasing the anode current; and, being nearer

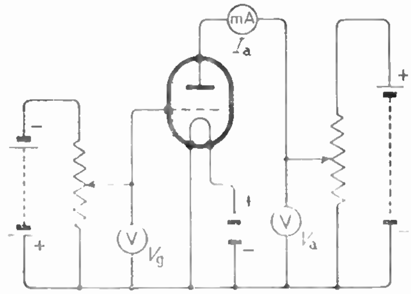



Fig. 9.4—Circuit for taking characteristic curves of a triode

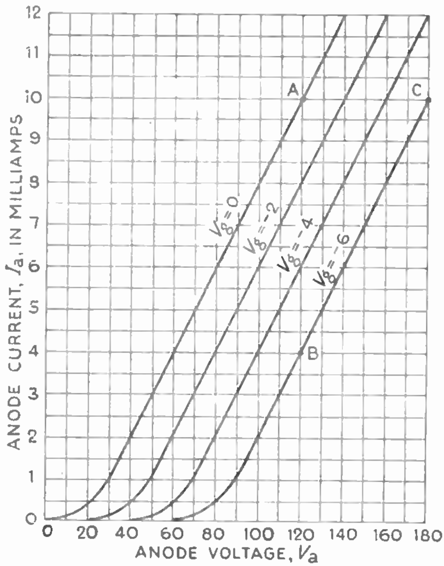



**Fig. 9.5—Characteristic curves of triode valve, at different grid voltages. (These curves are somewhat idealized, being straighter and more parallel than in actual practice)**

to the cathode, one grid volt will be more effective than one anode volt. If, on the other hand, it is made negative it will assist the space charge in repelling electrons back towards the cathode.

Fig. 9.4 shows the rather more complicated apparatus needed to take characteristic curves of this three-electrode valve, or *triode*. And in Fig. 9.5 are some such curves, showing the effects of both anode and grid voltages. Each of these curves was taken with the fixed grid voltage indicated alongside. Notice that this voltage, like all others relating to a valve, is reckoned from the cathode as zero. If, therefore, the cathode of a valve is made two volts positive with respect to earth, while the grid is connected to earth, it is correct to describe the grid as "two volts negative", the words "with respect to the cathode" being understood. In a directly heated valve, voltages are reckoned from the negative end of the filament.

## Amplification Factor

One would expect that if the grid were made neither positive nor negative the triode would be much the same as if the grid were not there; in other words, as if it were a diode. This guess is confirmed by the curve in Fig. 9.5 marked "$V_g = 0$", for it might easily be a diode curve. Except for a progressive shift towards the right as the grid is made more negative, the others are almost identical. This means that while a negative grid voltage reduces the anode current in the way described, this reduction can be counterbalanced by a suitable increase in anode voltage. In the valve for which curves are shown, an anode current of 10 mA can be produced by an anode voltage of 120 if the grid is held at zero potential. This is indicated by the point A. If the grid is now made 6 V negative the current drops to 4 mA (point B), but can be brought up again to its original value by increasing the anode voltage to 180 V (point C).

Looked at another way, the distance A to C represents 60 V on the $V_a$ scale and $-6$ V on the $V_g$ scale, with no change in $I_a$. So we can say that a change of 6 V at the grid can be compensated for by a change of 60 V, or ten times as much, at the anode. For reasons that will soon appear, this ratio of 10 to 1 is called the *amplification factor* of the valve. It is yet another thing to be denoted by $\mu$.

As with the diode, the anode resistance of the valve can be read off from the curves. All four curves of Fig. 9.5 will give the same value over their upper portions, since they all have the same slope; over the lower parts, where the steepness varies from point to point, a whole range of values for the anode resistance exists. Over the linear portions of the curves this resistance is 10 k$\Omega$, as can be seen from the fact that the anode voltage must change by 10 to alter the anode current by 1 mA.

When reckoning the $r_a$ of a triode, it is most important that the change in anode voltage should not be accompanied by any change in grid voltage.

## Mutual Conductance

We have seen that 1 V at the grid of this particular triode is equivalent to 10 V at the anode; an alternative way of changing the anode current by 1 mA is therefore to change the grid potential by 1 V. This fact also can be read from the curves, by observing (for example) that at $V_a = 100$ the anode current for $V_g = 0$ and $V_g = -2$ is 8 and 6 mA respectively.

The response of the anode current of a valve to changes in grid voltage is a useful measure of the control that the grid exercises over the electron stream through the valve. It is called the *mutual conductance* (symbol: $g_m$), and in basic units would be expressed in amps per volt, or mhos (p. 41); but because valve currents usually amount only to milliamps it is nearly always in mA/V, which could also be called millimhos. Americans often use a smaller unit still, the $\mu$A/V or micromho.

The value of $g_m$ can, as we have seen, be derived from an $I_a/V_a$ diagram such as Fig. 9.5, provided that there are curves for at least two different grid voltages, by noting the change in $I_a$ corresponding to a change in $V_g$. And, of course, the pair of readings marking out this change must both be at the same anode voltage. An alternative form of characteristic curve is obtained by plotting $I_a$ against $V_g$, taking care to keep $V_a$ adjusted to a fixed value for each curve, as in Fig. 9.6; a single curve of this kind is obviously sufficient to indicate $g_m$. For instance, BC represents the increase

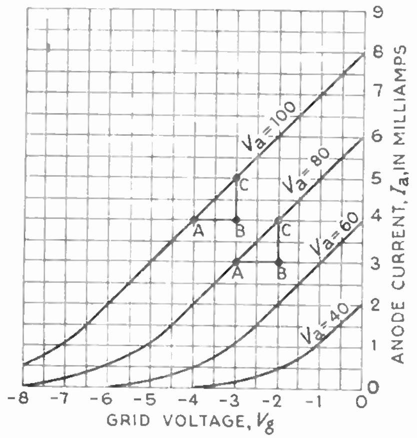



Fig. 9.6—Four samples of anode-current/grid-voltage characteristic curve, each taken at a different anode voltage

in $I_a$ caused by an increase in $V_g$ represented by AB; so $g_m$ is given by the ratio BC/AB—in this case 1 mA/V. Since this ratio also defines the slope of the curve, it is quite common to refer to the mutual conductance as the *slope* of the valve. But it must be clearly understood that it is only the slope of this particular kind of valve characteristic curve. The slope of the $I_a/V_a$ curve (Fig. 9.5) is obviously different, and, in fact, is the anode conductance, $1/r_a$. It, too, can be measured in mA/V, the volts being anode volts instead of grid volts. But we know that, in its effects on $I_a$, one grid volt is equivalent to $\mu$ anode volts. So the mutual conductance must be $\mu$ times the anode conductance. In symbols:

$$g_m = \mu \times \frac{1}{r_a}, \quad \text{or} \quad g_m = \frac{\mu}{r_a}, \quad \text{or} \quad \mu = g_m r_a, \quad \text{or} \quad r_a = \frac{\mu}{g_m}.$$

If $r_a$ is in ohms, $g_m$ will be in A/V; and if $g_m$ is in mA/V, $r_a$ will be in kΩ.

We see, then, that if any two of these valve *parameters* are known, the third follows. It must not be forgotten that their values depend on the part of the valve curve selected; that is to say, on the steady voltages applied to the electrodes. The more nearly linear the curves, the more nearly constant the parameters. At the bends, they vary enormously, especially $r_a$ and $g_m$; $\mu$ is the most nearly constant.

## Alternating Voltage at the Grid

In Fig. 9.7 we have an $I_a/V_g$ curve for a typical triode. As the slope of the curve shows, its mutual conductance is about $3\frac{1}{2}$ to 4 mA/V for anode currents greater than about 4 mA, but less for lower currents. Suppose that, as suggested in the inset to that figure, we apply a small alternating e.m.f., $v_g$, to the grid of the valve, what will the anode current do? If the batteries supplying anode and grid give 200 and $-2\frac{1}{2}$ V respectively, the anode current will set itself at about $5\frac{3}{4}$ mA: point A on the curve.

If $v_g$ has a peak value of 0·5 V, the total voltage on the grid will swing between $-2$ and $-3$ V, alternate half-cycles adding to or subtracting from the steady voltage $E_g$. The anode current will swing correspondingly with the changes in grid voltage, the points B and C marking the limits of the swing of both. The current, swinging between $7\frac{1}{2}$ and 4 mA, is reduced by $1\frac{3}{4}$ mA on the negative half-cycle and increased by the same amount on the positive one. The whole is therefore equivalent to the original steady current with an alternating current of $1\frac{3}{4}$ mA peak superposed on it.

There are two ways in which this a.c. in the anode circuit can be usefully employed. It can be used for producing an alternating voltage, by passing it through an impedance. If the voltage so obtained

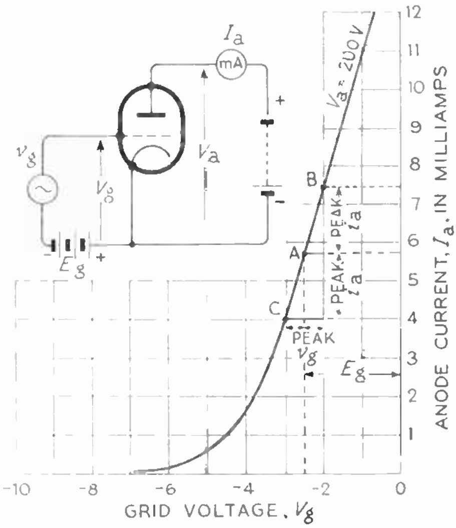



**Fig. 9.7—The result of applying an alternating e.m.f. ($v_g$) to the grid of a triode can be worked out from the $I_a/V_g$ curve, as here. The fixed starting voltage, $E_g$, in this case is $-2{\cdot}5$ V**

is larger than the alternating grid voltage that caused it, we have a voltage amplifier. Alternatively, if the alternating current is strong enough it can be used to operate a loudspeaker or other device, in which case the valve is described as a power amplifier. It is this ability to amplify that has made modern radio and the many kindred developments comprised in electronics possible. The subject is so important that several chapters will be devoted to it, beginning with Chapter 11.

## Grid Bias

You may have been wondering why the grid of the valve has never been shown as positive. The reason is that a positive grid attracts electrons to itself. It is, in effect, an anode. The objection to this is not so much that it robs the official anode of electrons—for up to a point it continues to increase the flow there as well as to itself—but because current flowing in the grid circuit calls for expenditure of power there. The full implications of this will become clearer later on, but even at this stage it may be appreciated that the existence of grid current would deprive the valve of one of its most attractive features—its ability to release or control power without demanding any.

To prevent the flow of grid current during use, a fixed negative voltage, known as *grid bias*, is applied. The amount of bias required is equal to—or preferably a volt or so more than—the positive peak of the signal input voltage which the valve is expected to accept. The reason for making the bias greater than the signal peak is that some electrons are emitted from the cathode with sufficient energy to force a landing on the grid against a negative bias of anything up to about one volt.

CHAPTER 10

# Semi-Conduction: Transistors

## Varieties of Conduction

During the preceding chapters we have come across two ways in which electricity can be conducted from place to place—or not conducted, as the case may be. Almost our first notion of electricity (p. 30) was that substances can be classified into conductors and insulators according as their atoms have or have not electrons so loosely attached that the electric field produced by an e.m.f. can make them drift through the material in a particular direction. Later (p. 116) we found that by means of heat it is possible to detach electrons from matter and use them as an electric current across the otherwise empty space in a vacuum tube or valve. The reason for doing this is that it makes the current much easier to control, especially at high speed. But because their positive partners are left behind, the cloud of electrons in the space is an unneutralized negative charge which tends to repel those following, so that a relatively high voltage is needed to keep up the flow, making the valve appear to have a high resistance. Later on (p. 328) we shall learn that this can be overcome, at some sacrifice of controllability, by admitting a small amount of matter in the form of low-pressure gas, the atoms of which are split up into positive and negative charges by the moving electrons. Such conduction has features in common with both metallic and vacuum kinds.

Not all solid substances can be sorted into the two classes first mentioned—good conductors and good insulators. Some are in between, conducting perhaps a million times less than copper, but billions of times more than rubber. These are called *semi-conductors*. In general, as regards the way they conduct and their consequent electrical behaviour, they are more complicated than metallic conductors.

## Germanium Crystals

The semi-conductor of greatest practical interest at the present time is germanium, but what is to be said about it applies in principle also to silicon. Germanium is a comparatively rare element, one of a small class that forms itself into crystals by the process of every atom linking up with four other atoms. This crystal structure is of course in three dimensions, but can be represented diagrammatically in two as in Fig. 10.1. Each of the circles represents one atom. It has been found that the links between are due to four electrons which each atom carries outermost. Now germanium, being No. 32 in the list of elements, carries altogether 32 electrons around its nucleus (p. 28), so the nucleus has a positive electrical charge of 32 electron units. But all except four of the electrons are closely bound to the nucleus, so for our present purpose we shall regard them as lumped together with it to form

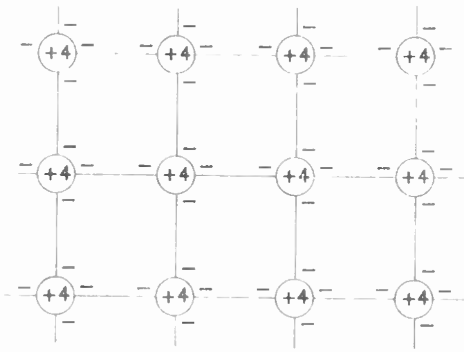



**Fig. 10.1—Diagrammatic representation of the four-fold linkage of germanium atoms to form a perfect crystal. The "valency" electrons responsible are shown as four minus signs**

the main part of the atom, having a net positive charge of four units to balance the four linking electrons. These particular electrons, by the way, are the ones responsible for chemical combination with other elements, so are known by the chemists' term as *valency* electrons.

## Holes

A crystal of this kind, if perfect, would have no electrons free to roam around and would therefore be a perfect insulator. But for various reasons it never is. One reason is that at ordinary room temperatures there is enough heat energy to shake out a valency electron here and there, leaving behind a place with an electron short. Now "an electron short" is, in effect, a positive charge; and the curious and interesting thing is that this shortage can wander about and cause conduction, almost like the loose electron. Although the exact way in which this happens is difficult to understand, the general effect can be pictured by imagining a row of patients sitting in a doctor's waiting room (Fig. 10.2). The patient in seat No. 1 has just been called to the consulting room, leaving that seat vacant. The patient in No. 2 moves forward into No. 1 so as to be ready; No. 3 moves into No. 2; and so on, until the situation is as at *b*. The empty seat has "drifted" the whole length of the room, although no patient has moved more than a small fraction of that distance. The doctor's call has not only "attracted" a patient but has also "repelled" a vacant seat. It is more convenient to look

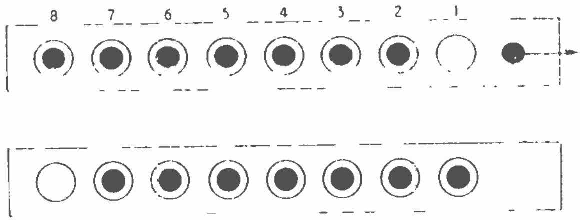



**Fig. 10.2—Plan view of a doctor's waiting room seated for eight: (*a*) while a patient is moving into the consulting room, and (*b*) a few moments later**

at the matter in this way than to regard all the separate movements of patients.

In semi-conductor electronics the " vacant seat " is termed a *hole*. For practical purposes a hole can be regarded as equivalent to a mobile electron, except for its being positive instead of negative. When a meter reads 1 mA it could be due to $6{\cdot}24 \times 10^{15}$ electrons per second moving through it one way, or an equal number of holes the other way—or a mixture of both. Conduction caused by heat disturbance is therefore two-fold; each electron released creates a hole, and when an e.m.f. is applied the current made up of electrons moving towards the positive terminal is augmented by holes moving towards the negative terminal. Even so, the combined number falls far short of the free electrons in a good conductor, so the result is only semi-conduction.

It must be emphasized that strictly speaking a hole does not move, any more than a chair moved through the consulting room. But the net result is the same as if it did, and it is easier to take account of the apparent movement of a hole than of the enormous number of very short electron movements that actually cause it.

## Intrinsic Conduction

A more comprehensive model of a semi-conductor could be made by parading a large number of small balls on a flat table in rank-and-file formation to represent the crystalline structure of Fig. 10.1. They could be held in place by resting in shallow depressions in the table. A difference of potential would be represented by tilting the table very slightly, the upper end being " negative ". Such a tilt would cause no movement of balls (electrons), except perhaps a slight lurch to the side of each cup nearest the " positive " end, representing a displacement or capacitive current (p. 53). This is like a crystal at a very low temperature, when it is an insulator.

The effect of raising the temperature could be shown by making the table vibrate with increasing vigour. At a certain intensity a ball here and there would be shaken loose, and on a perfectly level table (no e.m.f.) would wander around at random. In doing so it would sooner or later fall into a hole—not necessarily the one from which it came. There would therefore be some random "movement" of holes. At any given degree of vibration (temperature) a balance would eventually be reached between the rate of shaking loose and recombination, but the greater the vibration the greater the number of free electrons and holes at any one time, and the greater the drift of balls towards the " positive " end of the table when tilted. One result of this movement would be a tendency for more holes to be reoccupied at the lower end than the upper; in effect, there would be a movement of holes up the table towards the negative end.

To represent a sustained e.m.f. it would be necessary for the balls that reached the bottom of the table to be conveyed by the source of the e.m.f. to the top end, where their filling of the holes would be equivalent to a withdrawal of the holes there.

If you have been able to visualize this model sufficiently you should by now have gathered that at very low temperatures a germanium crystal is an insulator, but that as the temperature rises to ordinary room level and higher it increasingly conducts. In other words its resistance falls, in contrast to that of metals, which rises with temperature.

This sort of conduction is called *intrinsic*, because it is an inescapable property of the material itself.

A similar effect is caused by light; a fact that is turned to advantage in the use of semi-conductors as light detectors and measurers.

Unless a substance crystallizes under ideal conditions, it tends to be a jumble of small crystals, at the boundaries of which there are numerous breaks in the regularity of the lattice; this is yet another cause of conduction.

## Effects of Impurities

So far we have assumed that our crystal, however it may be disturbed by heat, light, or its original formation, is at least all of one material—germanium. In practice however it is never perfectly pure. Now if perchance the impurities include any element having five valency electrons (such as phosphorus, arsenic or antimony) its atoms will be misfits. When they take their places in the crystal lattice—as they readily do—one electron in each atom will be at a loose end, more or less free to wander and therefore to conduct. Such an incident is represented in Fig. 10.3. Intruding atoms that yield spare electrons are called *donors*. It might be supposed that because each surplus electron leaves behind it an equal positive charge it thereby creates a hole, but it should be understood that the technical term " hole " applies only to a *mobile* positive charge. In this case the positive charge is in the nucleus, and, as all four valency electrons needed for the lattice are in fact present, there is no vacancy into which a strolling electron can drop. Therefore the positive charges are fixed, and conduction is by electrons only.

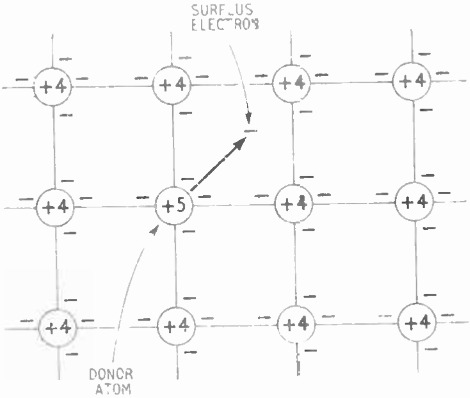



**Fig. 10.3—Showing how the perfect crystal (Fig. 10.1) is modified when an impurity atom with five valency electrons occupies a place**

Because electrons are *n*egative charges, germanium with do*n*or impurity is classified as *n-type*.

An incredibly small amount of impurity is enough to raise the conductivity appreciably; something of the order of one part in a thousand million ($10^9$)—far less than can be detected by the most sensitive chemical analysis.

Some other elements, such as aluminium, gallium and indium, have three valency electrons per atom; and when any of these substances is present as an impurity the situation is the opposite of that just described. The vacancy in one of the four links constitutes a hole. This is liable to borrow an electron from some other atom, giving the effect of movement by holes. The impurity atom in this case is an *acceptor*. Note that when the hole has migrated from the acceptor—in other words, an electron from elsewhere has filled up or cancelled the hole in the acceptor atom—that atom constitutes a negative charge, because its main body has only three

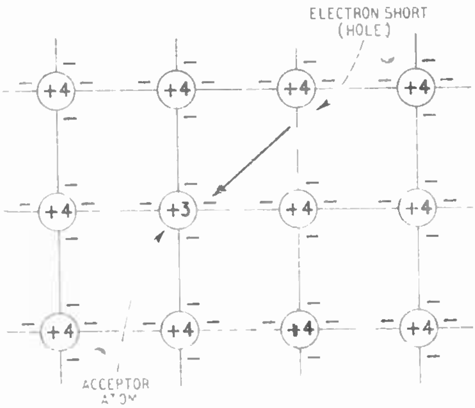



**Fig. 10.4—Showing how the perfect crystal is modified when an impurity atom with three valency electrons occupies a place**

net positive charges to offset what are now four valency electrons; but such negative charges are fixed and take no direct part in current flow. Because conduction is by holes only, which are *positive* charges, germanium with acce*p*tor impurity is called *p-type*.

So far as impurity conduction is concerned, we see that the germanium itself plays only a passive part, as a framework for the broken impurity atoms. In the diagrams that follow, it will therefore be omitted in order to focus attention on those relatively very few atoms. Fig. 10.5 shows the standard symbols used for this purpose. The mobile charges—electrons and holes—are represented by simple minus and plus signs respectively; and the

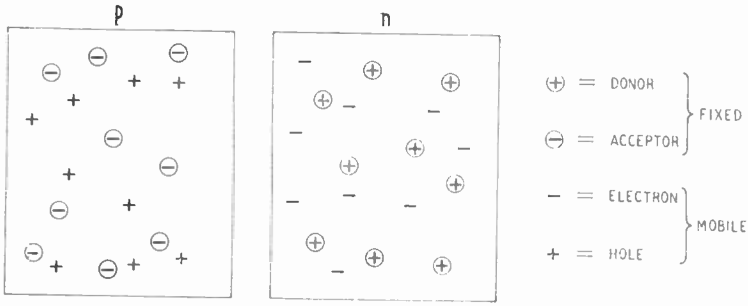



**Fig. 10.5—Simplified representation of pieces of impure germanium, showing only the electric charges of broken impurity atoms. In *p*-type material the mobile charges are predominantly positive (holes), and in *n*-type material negative (electrons)**

fixed charges—donor and acceptor atoms—are distinguished by circles around them. Although these atoms are all members of the crystalline lattice, they are such a small minority that the pattern can no longer be discerned.

One might ask what happens when both donor and acceptor impurities are present at the same time. The answer is that they tend to cancel one another out, and when present in equal quantities the germanium is much as if it had no impurity at all. But although this *compensation* as it is called is used for converting *n*-type into *p*-type (or vice versa) by adding more than enough opposite impurity to neutralize the impurity already present, the balance would be too fine to enable heavily contaminated germanium to be made apparently pure by the same method.

## P-N Junctions

We can have, then, *p*-type material, in which conduction is by positive holes; *n*-type, in which it is by negative electrons; and what is sometimes called *i*-type (for " intrinsic "), in which it is by both in equal numbers. None of these by itself is particularly useful; it is when more than one kind are in contact that things begin to happen. Just bringing two pieces together is not good enough, however; the combination has to be in one piece, divided into two or more regions having different impurities. Various manufacturing methods have been devised for achieving this.

Fig. 10.6*a* shows in standard diagram such a combination, forming what is known as a *p-n* junction. Even when no electric field is created by an e.m.f., mobile charges tend to move by what is called *diffusion* into a uniform distribution. Some of the holes in the *p* region therefore move across the junction into the relatively holeless territory beyond, and electrons move oppositely. Both these migrations are an electric current from *p* to *n*, so *n* begins to charge positively and *p* negatively. A potential difference thus grows between *p* and *n* of such polarity as to discourage further migration, until it is just sufficient to bring it to a halt. This steady balance is indicated in Fig. 10.6*a* by the slight redistribution near the junction, which produces an excess of negative charge on the *p* side and of positive charge on the *n* side. An alternative way of indicating this difference of potential is shown at *b*. The difference amounts to a few tenths of a volt.

Next, consider what happens when an external e.m.f. is applied. If the polarity is as shown at *c* it attracts the free electrons to itself, leaving more donor charges in the *n* region unneutralized; similarly the holes in the *p* region are swept to the left. When this process

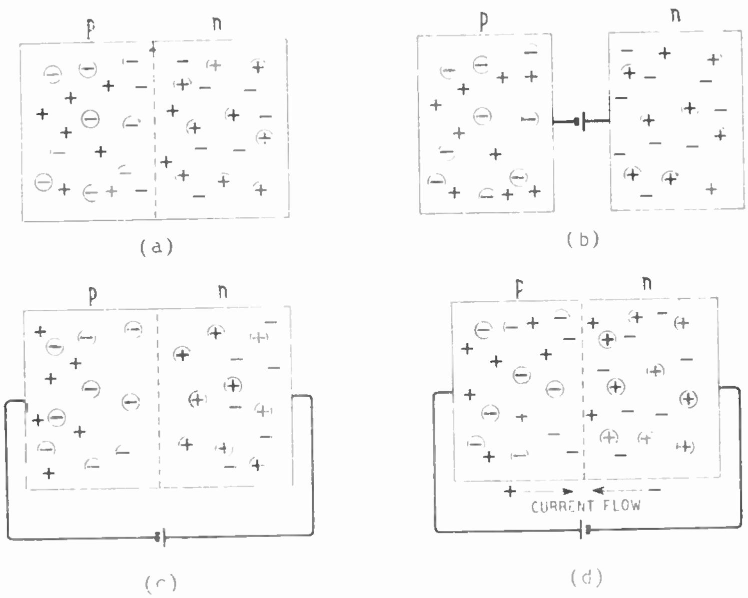



Fig. 10.6—Some mobile charges stray across an undefended frontier (*a*) causing an excess of negative on the *p* side and positive on the *n* side; this can be represented symbolically as at *b*. An external e.m.f. connected + to *n* (*c*) increases this p.d. against itself, sealing the circuit against current flow; but with + to *p* (*d*) it reduces or even removes the p.d. and current flows freely

has gone far enough for the p.d. between the regions to build up to that of the battery, another no-current balance is established. In other words, to an e.m.f. of this polarity the junction behaves as a non-conductor. Fig. 10.6*b* indicates this quickly in a general sort of way by the fact that the imaginary internal battery is in opposition to the external one, which is described as a *reverse* e.m.f. What it fails to show is that the voltage of the internal one rises to equal that of the external. In fact, the junction behaves to a reverse e.m.f. rather like a capacitor (p. 50).

Suppose however that the external e.m.f. is applied the other way round, as at *d*. The direction of its field makes electrons move to the left and holes to the right; but this time the flow continues because the external e.m.f. is of such a polarity as to be able to replenish the supply of both. So although the flowing of holes and electrons through one another's territory causes some electrons to fill holes, cancelling both, the numbers do not diminish.

In relation to Fig. 10.6*b* we see that the internal " battery " assists the external, which means that the external is connected the right way for current flow—the *forward* direction. The internal p.d. is however reduced by the holes flowing from the left and electrons from the right.

## The Germanium Diode

In short, a *p-n* junction is a rectifier. The foregoing explanation of why it rectifies may seem complicated in comparison with the action of the vacuum diode (p. 117), so let us look at them side by side, as in Fig. 10.7. Here diodes of both vacuum and *p-n* junction types are connected both ways across a battery. At *a* we are reminded that current can consist either of negative charges flowing to + or positive charges to −. Or, of course, both at once. When a vacuum diode is connected as at *b*, current can flow, because the heated cathode is an emitter of electrons, which are negative charges. But when reversed, as at *c*, there is no current, because the anode does not emit electrons. And neither electrode emits positive charges. In a *p-n* diode connected *p* to + as at *d*, current flows,

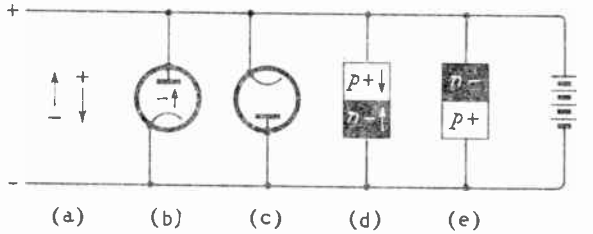



**Fig. 10.7—Showing how in diodes the presence or absence of current depends on the presence or absence of the right kind of current-carriers**

because the $n$ region emits electrons, and because the $p$ region emits holes, which are positive charges. When reversed ($e$) both regions emit the wrong charges, so there is no current.

The germanium diode rectifier has some important advantages over the vacuum type: it needs no cathode heating; it is less fragile; and the emitted electrons do not form a space charge tending to oppose the anode voltage, because their charge is neutralized in the crystal by the fixed positive charges, so less voltage is lost in the diode when current flows and as a rectifier it is more efficient.

On the other hand there are one or two complications. So far we have ignored the intrinsic conduction which, as we have seen, results from the releasing of electrons and holes in equal numbers throughout the material, so does not depend on the direction of the e.m.f. To the extent that it is present, therefore, it reduces the effectiveness of a semi-conductor rectifier by allowing current to

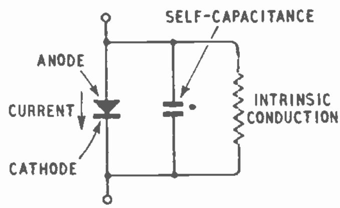



**Fig. 10.8—Two imperfections of a semi-conductor rectifier can be represented by circuit components in parallel with an ideal rectifier**

flow the " wrong " way. It is as if the rectifier were shunted by a conductor. One cause—light—can easily be excluded by an opaque covering. But the amount of heat at ordinary temperatures is sufficient for the reverse (or leakage) current of a germanium diode to be appreciable, and it approximately doubles with every 9° C rise in temperature. This is one of the most serious limitations of germanium. In this respect silicon is an attractive alternative because its intrinsic conductivity is about a thousand times less.

Then the capacitor action mentioned above is another limitation, for it allows current to flow the wrong way in proportion to frequency. So the junction type—and especially those with sufficiently large junction areas to pass heavy current without serious rise in temperature—are unsuitable for high frequencies.

A junction rectifier therefore behaves somewhat as if it were composed as shown in Fig. 10.8, where the symbol on the left is the general symbol for a semi-conductor rectifier.

The miniature germanium rectifiers or diodes used so extensively in television receivers and elsewhere are mainly of the point-contact type, consisting of a small piece of $n$-type crystal with a pointed springy wire impinging on it, enclosed in a glass bulb (Fig. 10.9). During manufacture a pulse of current is passed, which has the effect of giving $p$-type characteristics to the germanium immediately surrounding the point. Because of the small area of the boundary so formed, the capacitance is less than in the junction type, so these diodes are effective at much higher frequencies than the junction type. A wide variety of characteristics are obtainable by varying the impurity of the germanium and other details, but the general form is as shown in Fig. 10.10. When the reverse voltage is increased a point is reached where the current increases

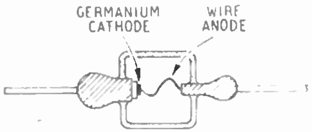



**Fig. 10.9—Construction of a point-contact germanium rectifier**

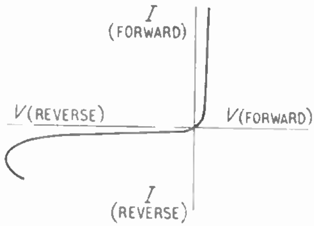



**Fig. 10.10—Typical characteristic curve of a germanium rectifier**

rapidly, even if the voltage is then reduced. This point is called the *turnover*, and with the purer kinds of germanium may be as high as 200 V. The less pure kinds are better for high frequencies—some are effective beyond 1,000 Mc/s—but the working voltages are much lower.

## The Junction Transistor

Remembering how greatly the usefulness of the thermionic valve was increased by adding a third electrode to the original diode, we can hardly be surprised that attempts were made to do the same with the semi-conductor diode. These efforts were successful, and the result is known as a *transistor*. It has already displaced the thermionic valve in some applications, and is finding its own new applications.

The first transistors were point-contact, but for purposes of explanation it is easier to begin with the junction type. In its usual form this is a sandwich of *n*-type germanium with a very thin

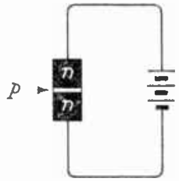



**Fig. 10.11—A transistor triode consists of two semi-conductor diodes back-to-back. When it is connected like this, very little current can pass either way**

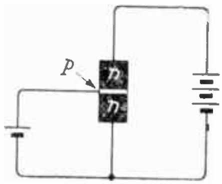



**Fig. 10.12 — Applying bias of the right polarity to the middle layer causes increased current between the outers**

filling of *p*-type, shown diagrammatically in Fig. 10.11, and called an *n-p-n* transistor; or alternatively the opposite, *p-n-p*. Either way, it adds up to a pair of rectifiers back to back, so whichever direction an e.m.f. is applied from end to end one of the junctions is in reverse, and apart from a normally small leakage current due to intrinsic conduction no current will flow.

This is not very useful, so the next step is to give the *p* layer a small positive bias (Fig. 10.12). This causes current to flow, which is what one would expect of a rectifier biased in the forward direction. What one might not expect, however, is that most of the current flows from end to end through the original part of the circuit, only a small part emerging from the *p* layer.

How is it that current can now flow freely in the "wrong" direction through the upper rectifier? The answer comes if we recall, with the aid of Fig. 10.7 if necessary, why it is that current cannot normally flow this way through a *p-n* junction—because the only mobile charges present in substantial numbers in a *p* region are holes, which are repelled rather than attracted by the positive potential of the upper *n* region. But the lower *n* region has vast numbers of mobile electrons, and making the *p* layer positive attracts them into it. In other words, the lower *n* region emits electrons into what was an electron "vacuum". This *n* region is therefore called the *emitter*, and it corresponds to the cathode of a vacuum valve. The electrons thus brought into the *p* layer are attracted by the positive upper *n* region, which is called the *collector*. It corresponds to the anode of a valve. The amount of collector current depends on the amount of bias on the *p* layer, which is therefore analogous to a valve grid, and is called the *base*. Unlike the other two names, this seems strangely inappropriate. The reason for it, which will appear later, is historical rather than functional.

If the grid of a valve were made positive like the base, electrons would be attracted to it as well as to the anode, causing grid current. And so it is in the transistor. But by making the base extremely thin—about a thousandth of an inch or even less—so that electrons entering it are only that distance from the collector, it has been found possible to ensure that something like 95–99 per cent of the emitter current reaches the collector. This fraction, slightly less than 1, is distinguished by the symbol $\alpha$ (Greek "alpha"). The collector current, $I_c$, is therefore 20–100 times the base current, $I_b$, needed to make it flow.

So, although in some ways a transistor is strikingly analogous to a triode valve of the vacuum kind, there are important differences. The valve is a rectifier connected to a supply voltage (h.t.) in the right direction to make current flow through it. This current can be either increased or decreased by applying grid bias, according to whether it is positive or negative; but to avoid the flow of grid current it is usual to make it negative, so that in practice the anode current is controlled downwards. The transistor, on the contrary, faces the supply with a rectifier in reverse, so except for a small leakage the normal state is no current. Biasing the base negative would merely add another reversed rectifier to the score, so to obtain any useful result it is necessary for the bias to be positive; and although under these conditions base current is inevitable it enables a much larger collector current to be created and controlled.

As with the junction diode, the transistor does not contain an unneutralized space charge to impede current flow, so only a few volts are needed at the collector—certainly not enough to be dignified by the description "h.t." The base bias is small, too; usually only a fraction of a volt.

## Characteristic Curves

Let us go into the matter further, using the same test circuit as for the triode valve (Fig. 9.4) except for the absence of cathode heating, the addition of a milliammeter for base current, and lower supply voltages, both of the same polarity relative to emitter; see Fig. 10.13. With this set-up one could obtain data for plotting a set of $I_c/V_c$ characteristic curves at a number of fixed base voltages, similarly to what was done with the valve to produce curves of the Fig. 9.5 type. If we did we would find that instead of the even spacing of the valve curves there

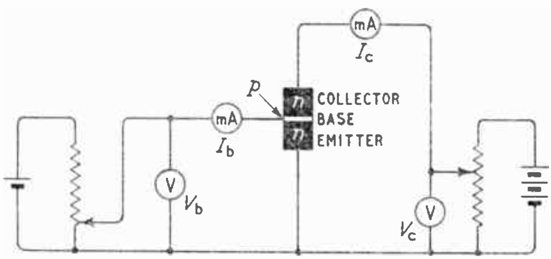



**Fig. 10.13—Arrangement, similar to Fig. 9.4, for obtaining characteristic curves of a transistor**

would be crowding together at first, with progressive thinning out. This would show that the changes in $I_c$ are not in constant proportion to the changes in $V_b$ causing them. But by plotting the $I_c/V_c$ curves at equal intervals of fixed base *current*, as in Fig. 10.14, we find they are spaced with even greater regularity than the valve curves.

This can be seen still more clearly by plotting $I_c$ against $I_b$ at a fixed $V_c$, as in Fig. 10.15, which is strikingly linear.

Everything so far has referred to an *n-p-n* transistor, because that is the type most easily compared with the triode valve. To convert the *n-p-n* information to *p-n-p*, merely substitute "—" for "+" everywhere, and "holes" for "electrons".

## Transistor Parameters

Looking at Fig. 10.14 we see at once, from the very slight slope of the curves over nearly all their extent, that the collector a.c. conductance is smaller than that of a typical triode valve;

that is to say, its resistance is larger than such a triode's $r_a$. For reasons that will appear later (p. 341), this resistance is not called $r_c$. At $I_b = 0$, the collector is working under reversed diode conditions, so there is only intrinsic conduction. The electrons responsible for it are so few that less than one tenth of a volt $V_c$ is sufficient to cause saturation. (Contrast *forward* saturation of a vacuum valve, p. 118.) This $I_c/V_c$ curve therefore hardly rises above the zero line, and being almost horizontal it indicates a very high collector a.c. resistance indeed. Ideally it is infinitely large. As $I_b$ is raised, the resistance gradually falls, but even at 0·3 mA is still high by comparison with vacuum triodes. The almost constant small $I_c$ at $I_b = 0$ is given the special symbol $I_{c0}$, and has only nuisance significance.

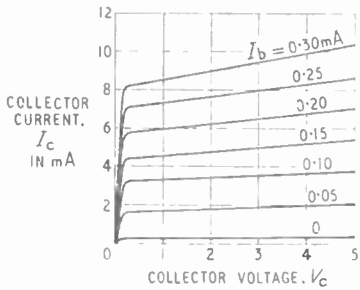



Fig. 10.14—Collector current/voltage characteristic curves of an *n-p-n* junction transistor

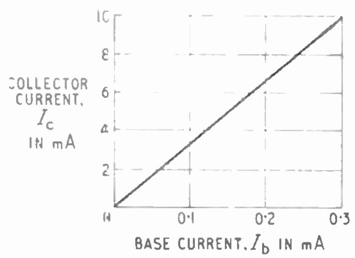



Fig. 10.15 — Collector-current/base-current characteristic of transistor connected as in Fig. 10.13

The mutual conductance, which is equal to the slope of the $I_c/V_b$ curve (compare p. 121), is not usually referred to, because this slope varies so greatly, which is why we did not bother to plot it. For the same reason the voltage amplification factor is also left out of account. Instead, attention is focused on the slope of the $I_c/I_b$ " curve ", which Fig. 10.15 shows to be remarkably constant. For obvious reasons it is called the current amplification factor. We have already noted that the ratio of collector current to emitter current ($I_c/I_e$) is denoted by $\alpha$, which is a little less than 1. Now $I_e$ is $I_c$ and $I_b$ combined:

$$I_e = I_c + I_b$$

So $$\alpha = \frac{I_c}{I_c + I_b}$$

$$\alpha I_c + \alpha I_b = I_c$$

$$\alpha I_b = (1-\alpha) I_c$$

So the current amplification $$\frac{I_c}{I_b} = \frac{\alpha}{1-\alpha}$$

Various symbols are used for it, such as $\alpha'$, $\beta$ (" beta "), $\alpha_{cb}$ and $\alpha_e$. The commonest is $\alpha'$, but we shall see later that there is much in favour of $\alpha_e$.

For broadly comparing one triode valve's characteristics with another's it is sufficient to know any two of the three parameters we have studied: $r_a$, $\mu$ and $g_m$. The third can always be very simply calculated from the other two. For a transistor, because there is base voltage *and* current, it is necessary to know at least four, selected from a bewilderingly large list of possible parameters. Unfortunately at the present time there is no general agreement as to which four.

This problem will be gone into in Chapter 22. In the meantime it will be sufficient to remember that the current amplification factor, $\alpha_e$, usually comes within the range 20 to 100, and that typical values of collector-emitter and base-emitter a.c. resistance are 30 kΩ and 1 kΩ respectively. $I_{c0}$ should of course be as small as possible; 25 μA at room temperature is typical. The fact that the transistor is *current* controlled, through quite a low input resistance, is the most notable difference between it and the vacuum valve.

## Point-Contact Transistors

The original type of transistor, first announced in 1948, was a modification of the germanium diode shown in Fig. 10.9, made by applying a second point contact very close to the first. One of these contacts was named the emitter, the other the collector, and the germanium electrode the base—a name more appropriate to this type than for the functionally equivalent part of the junction type. It has since been discovered that the process of forming the transistor by applying current pulses via the point contacts converts the germanium closely surrounding the points into *p*-type, so a point-contact transistor can be classed as *p-n-p*. Like the *p-n-p* junction type, it requires a negative collector voltage and a positive emitter bias. Incidentally, it is helpful to remember that in *all* transistors the supply voltages should bias the emitter in the forward direction and the collector in the reverse direction.

The reason for placing the point contacts close together—usually a few thousandths of an inch—is similar to that for using a thin base layer in the junction type: putting it crudely, to enable holes from the emitter to slip across easily to the collector instead of emerging at the base terminal. In fact, in the point type of transistor a given current change in the emitter changes the collector current by a larger amount than itself; in other words, $\alpha$ instead of being slightly less than 1 is usually about 2 or more. So instead of the $V_b$ source having to supply a small base current, a comparatively large base current (from the collector) flows into the source against its e.m.f. The reason for this is rather obscure, and investigation has shown that the point transistor is really a good deal more complicated than a simple *p-n-p* system. The large $\alpha$ has far-reaching effects on performance and use; in particular, point transistors connected in Fig. 10.12 are unstable, and unless special precautions are taken in using them they are liable to pass excessive current and burn out. This tendency is increased by the working parts being so small that they can safely handle only very small amounts of power. They are also more noisy than the junction type. For these and other reasons they are now rarely used.

## Transistor Symbols

The transistor symbol shown in Fig. 10.16*a* or *b* obviously dates from the time when only point types were in existence; nevertheless it is still commonly used to denote also *p-n-p* junction transistors.

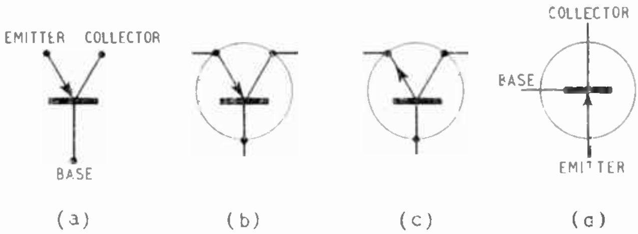



**Fig. 10.16—(*a* and *b*) Usual symbol for transistor (point-contact or *p-n-p* junction types); the "envelope" is optional. The *n-p-n* type is distinguished by a reversed arrow head (*c*). Symbol *d* is more appropriate to represent *p-n-p* junction types**

The *n-p-n* type is distinguished by reversing the arrow head, as at *c*. Because it is confusing to represent a junction transistor by a symbol which strongly suggests something quite different, various more suitable alternatives have been suggested. Fig. 10.16*d* is a particularly convenient example, suitable for modern circuits, and is used in this book to replace the one in Fig. 10.13, which, though suggestive of the construction, is awkward to draw by hand.

# Amplification

## A Vicious Circle

CHAPTER 9 ENDED with the disclosure that an alternating voltage applied to the grid of a valve is reproduced in the anode circuit as an alternating current. Then in the last chapter we saw that a current fed to the base of a transistor is reproduced in the collector circuit as a much larger current.

Simply having an alternating current flowing in the anode or collector circuit through the battery or other source is no help to anyone. To be of use the a.c. must be used in something, and that something is bound to offer some impedance to the current. Consequently the voltage across the something—which is officially termed the *load*—is bound to vary; and, since the battery maintains a steady voltage, that variation must be passed on to the valve.

This complicates matters, because the mutual conductance $g_m$ (or the $I_a/V_g$ graph) which tells us how much the anode current $I_a$ will be changed by a given change in $V_g$ is based on the assumption

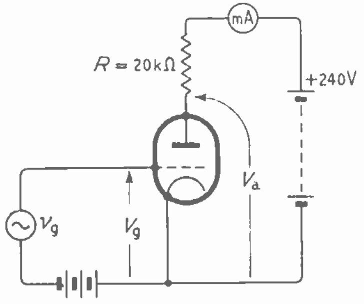



**Fig. 11.1—Simple voltage-amplifier circuit**

that meanwhile $V_a$ does not change. The anode resistance $r_a$ (or the $I_a/V_a$ graph) tells us how much $I_a$ depends on $V_a$, but the awkward thing in this case is that in order to find how much $V_a$ varies we have to know how much $I_a$ varies, and to find that we have to know how $V_a$ varies.

Take for example Fig. 11.1, which is the same circuit as in Fig. 9.7 with the addition of an anode load. To avoid frequency complications let us make it a simple resistance, $R$. It would be easy to calculate the output alternating current if $R$ were not there; it would simply be equal to the alternating grid voltage multiplied by the anode current per grid volt—in brief, $v_g g_m$. But the a.c. through $R$ makes $V_a$ alternate, so our $g_m$ no longer holds good. If we knew the amount of the a.c. we could calculate how much $V_a$ alternates, but we do not.

## The Load Line

One way of breaking this vicious circle is by means of a set of $I_a/V_a$ curves. To avoid having to look back at Fig. 9.5, here in Fig. 11.2 is another set, more typical of an actual valve. Let us suppose that the anode supply voltage is 240, and $R$ is 20 kΩ. We can then at once say what $V_a$ will be when any given anode current is flowing. When $I_a = 0$, the voltage drop (which, of course, is equal to $I_aR$) is also zero, and therefore the full battery voltage reaches the anode; $V_a = 240$ V. When $I_a = 1$ mA, $I_aR$ is 20 V, so $V_a = 240-20=220$ V. When $I_a = 2$ mA, $V_a = 200$ V. And so on. Plotting all these points on Fig. 11.2, we get the straight line marked "$R = 20$ kΩ". The significance of this line is that it marks out the only combinations of $I_a$ and $V_a$ that are possible when the resistance $R$ is 20 kΩ. Since $R$ is a load resistance, the line is called a load line. It would be equally easy to draw a load line for $R = 10$ kΩ or any other value.

At the same time, the valve curves mark the only possible combinations of $I_a$ and $V_a$, when that particular valve and $V_g$ are used. So when that valve and that resistance are used together the only possible combinations of $I_a$ and $V_a$ are those marked by points located both on the load line and on the appropriate valve curves; for example, point A, which indicates $I_a = 3\frac{1}{2}$ mA at $V_a = 170$ V. A current of $3\frac{1}{2}$ mA flowing through 20 kΩ drops $3\frac{1}{2} \times 20 = 70$ V, so the total voltage to be supplied is $170 + 70 = 240$ V, which is correct. Since the valve curve on which A falls is marked "$V_g = -2{\cdot}5$ V", that is the only grid voltage that will satisfy all the conditions mentioned.

## Voltage Amplification

Suppose now we alter the grid voltage to $-1{\cdot}5$ V. The working point (as it is called) must move to B, because that is the only point

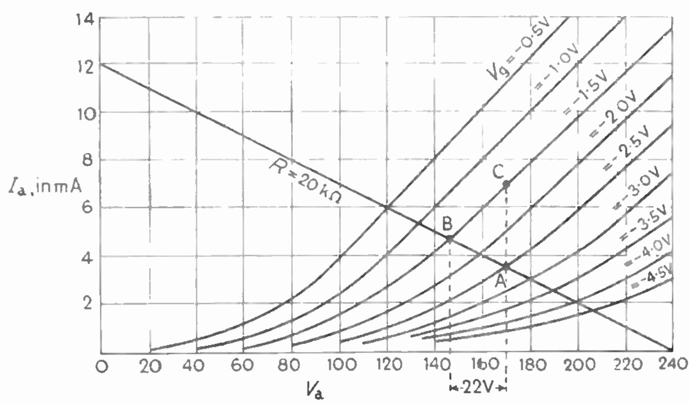



**Fig. 11.2—The problem of finding the amplification of the valve in Fig. 11.1 is here worked out on an $I_a/V_a$ curve sheet**

on both the "$V_g = -1{\cdot}5$ V" curve and the load line. The anode current rises from 3·5 mA to 4·6 mA—an increase of 1·1 mA. The voltage drop due to $R$ therefore increases by $1{\cdot}1 \times 20 = 22$ V. So the voltage at the anode ($V_a$) falls by that amount.

Note that a grid voltage change equal to 1 V has caused an anode voltage change of 22 V, so we have achieved a voltage multiplication —called amplification or *gain*—of 22. If the anode had been connected direct to a 170 V battery there could of course have been no change in $V_a$ at all, and the working point would have moved to C, representing an $I_a$ increase of 3·3 mA. The reason why the increase with $R$ in circuit was only 1·1 mA was the drop of 22 V in $V_a$, which partly offset the rise in $V_g$.

Another thing to note is that making the grid less negative caused the anode to become less positive. So an amplifier of this kind reverses the sign of the signal being amplified.

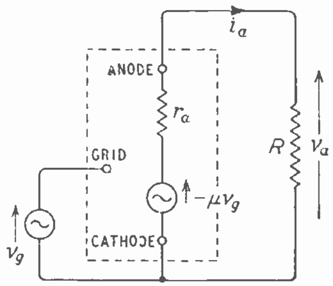
## The "Valve Equivalent Generator"

It would be very convenient to be able to calculate the voltage amplification when a set of curves was not available and only the valve parameters, $\mu$ and $r_a$, were known. To understand how this can be done, let us go back to Fig. 9.4 with its controls for varying anode and grid voltages. First, leaving the $V_g$ control untouched, let us work the $V_a$ control. The result, indicated on the milliammeter, is a variation in anode current. The same current variation for a given variation in $V_a$ would be obtained if an ordinary resistance equal to $r_a$ were substituted for the valve (p. 119). Considering only the variations, and ignoring the initial $V_a$ and $I_a$ needed to make the valve work over an approximately linear part of its characteristics, we can say that from the viewpoint of the anode voltage supply the valve looks like a resistance $r_a$.

Now keep $V_a$ steady and vary $V_g$. To the surprise of the $V_a$ supply (which does not understand valves!), $I_a$ again starts varying. If the $V_a$ supply could think, it would deduce that one (or both) of two things was happening: either the resistance $r_a$ was varying, or the valve contained a source of varying e.m.f. To help it to decide between these, we could vary $V_a$ slightly—enough to check the value of $r_a$ by noting the resulting change in $I_a$—at various settings of the $V_g$ control. Provided that we took care to keep within the most linear working condition, the value of $r_a$ measured in this way would be at least approximately the same at all settings of $V_g$. So that leaves only the internal e.m.f. theory in the running. We (who do understand valves) know that varying $V_g$ has $\mu$ times as much effect on $I_a$ as varying a voltage directly in the anode circuit (namely, $V_a$).

So we can now draw a diagram, Fig. 11.3 (which should be compared with Fig. 11.1) to show what the valve looks like from the point of view of the anode circuit. Its behaviour can be accounted for by supposing that it contains a source of e.m.f., $\mu v_g$, in series with a resistance, $r_a$. We have already come across examples of substituting (on paper or in the imagination) something which, within limits, behaves in the same way as the real circuit, but is easier for calculation. One of them was a dynamic resistance

Fig. 11.3—The "equivalent generator" circuit of a valve, which can be substituted for Fig. 11.1 for purposes of calculating the performance of the valve




in place of a parallel resonant circuit (Fig. 8.12). And now this trick of the *valve equivalent generator*, which is one of the most important and useful of all.

But "within limits" must be remembered. In this case there are two limits: (1) for a valve to behave like an ohmic resistance it must be working at anode and grid voltages corresponding to linear parts of its characteristic curves, and they are never *perfectly* linear, so at best $r_a$ is no more than a good approximation to an ohmic resistance; and (2) the initial values of $V_a$, $V_g$, and $I_a$ are left out of account—they are merely incidental conditions necessary to achieve (1)—and variations or *signal* voltages and currents within the linear region are considered on their own. For example, if $I_a$ is alternating between 3 mA and 5 mA, then the valve equivalent generator regards only $i_a$, the anode signal current, peak value 1 mA, and ignores the initial 4 mA on which it is based.

It is failure to separate in one's mind these two things (the supply or feed current and the signal current) that leads to confusion about the next idea—the minus sign in front of $\mu v_g$ in Fig. 11.3. Some people argue that, because a positive movement in $v_g$ (say from $-2{\cdot}5$ V to $-1{\cdot}5$ V) *increases* $I_a$, the imaginary signal voltage $\mu v_g$ must be positive. And then they are stuck to account for $v_a$, the amplified signal voltage, being negative as we found on p. 125. But if they realized that the feed voltage and current have nothing to do with Fig. 11.3 they would avoid this dilemma. For it just happens that the feed current flows anticlockwise round the circuit; but as we are ignoring this there is nothing to stop us from adhering to the established convention of reckoning voltages with respect to the cathode. If the imaginary internal generator voltage were $+\mu v_g$, a positive signal $v_g$ on the grid would make the anode go positive, which is contrary to the facts. To put this right by turning the generator upside down and reckoning the cathode voltage with respect to the anode is contrary to established convention. The logical course is to reckon the generator voltage as $-\mu v_g$.

## Calculating Amplification

Now let us apply this valve equivalent generator technique to the problem we set out to solve—finding the amplification of a valve. The signal current $i_a$ in Fig. 11.3 is (by Ohm's law) the signal e.m.f. divided by the total circuit resistance: $-\mu v_g/(R + r_a)$. The signal output voltage, $v_a$, is caused by this current flowing through $R$, so is $-\mu v_g R/(R + r_a)$. The voltage amplification (which is often denoted by $A$) is $v_a/v_g$, so

$$A = \frac{-\mu R}{R + r_a}$$

This minus sign agrees with our finding on p. 142. Let us apply this to the case we considered with the help of Fig. 11.2. Using the methods already described (p. 119 and p. 121) we find that in the region of point A the valve in question has a $r_a$ of about 14·5 kΩ and $\mu$ about 39. Substituting these, and $R = 20$ kΩ, in the formula: $A = -39 \times 20/(20 + 14{\cdot}5) = -22{\cdot}6$, which agrees pretty well with the figure obtained graphically. Since it is generally understood that $v_a$ is in opposite phase to $v_g$, the minus is often omitted.

## The Effect of Load on Amplification

The amplification formula that we have just used shows at once that making $R = 0$ would make $A = 0$. It also shows that making $R$ so large that $r_a$ was negligible in comparison with it would make $A$ almost equal to $\mu$, which is the reason for calling this parameter the amplification factor. What happens at intermediate values of $R$ can best be seen by using the formula to plot a graph of $A$ against $R$—the full-line curve in Fig. 11.4.* Lest this suggest that the best policy is to use the largest possible $R$, it should be mentioned that in order to keep $r_a$ from increasing too much it is necessary to apply sufficient anode voltage to maintain the feed current. If $R$ were raised from 20 kΩ to 100 kΩ in order to raise $A$ from 22 to 34, an extra $80 \times 3{\cdot}5 = 280$ V would be needed to keep at the same 3·5 mA working point. Any gain above about two-thirds $\mu$ is obtained at an uneconomical cost in supply voltage.

All these facts can be deduced also from Fig. 11.2. For a higher $R$ than 20 kΩ the load line would be more nearly horizontal, and to keep it passing through the same working point A it would have to pivot around A so that its left-hand end would touch the voltage scale at a higher point than 240 V. At the same time the anode current would be kept more nearly constant as $V_g$ varied, so the resulting variation in $V_a$ would necessarily be more nearly equal to $\mu v_g$. And vice versa for lower values of $R$.

## The Maximum-Power Law

For some purposes (such as working loudspeakers) we are not interested in the *voltage* output so much as the *power* output. This,

* Logarithmic scales are used (p. 11) in order to cover wide ranges of values without squeezing most of them into a corner. See also p. 148.

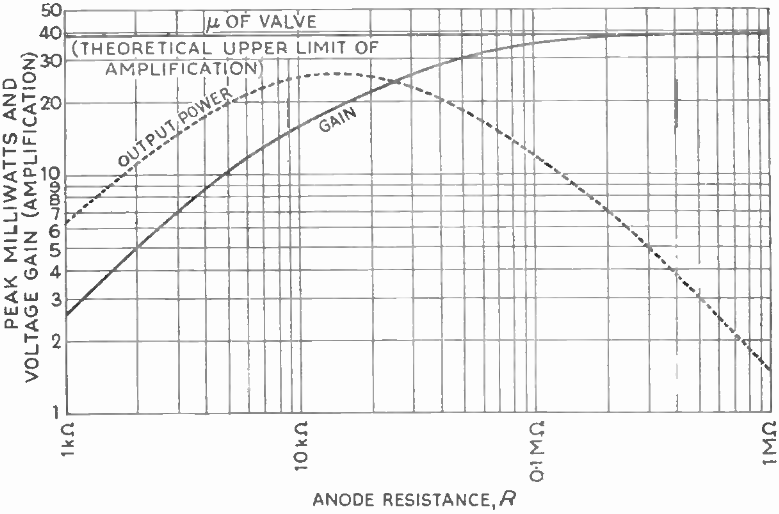



**Fig. 11.4—The full line shows how the voltage amplification depends on the value of load resistance $R$. It is calculated for a valve with $\mu = 39$ and $r_a = 14{\cdot}5$ kΩ. The dotted line shows the peak milliwatts delivered to $R$ for a grid input of 1 V peak.**

of course, is equal to $i_a v_a$ or $i_a^2 R$ (p. 47), and can therefore be calculated by filling in the value of $i_a$ we found overleaf; namely $-\mu v_g/(R + r_a)$. Denoting the power by $P$, we therefore have

$$P = \left(\frac{\mu v_g}{R + r_a}\right)^2 R.$$

A graph of this expression, for $v_g = 1$ V, and $\mu = 39$ and $r_a = 14{\cdot}5$ as before, is shown dotted in Fig. 11.4. The interesting thing about it is that it has a maximum value when $R$ is somewhere between 10 kΩ and 20 kΩ. The only thing in the circuit that seems to give any clue to this is $r_a$, 14·5 kΩ. Could the maximum power result when $R$ is made equal to $r_a$?

It could, and does; as can be proved mathematically. This fact is not confined to this particular valve, or even to valves in general, but (since, you remember, it was based on Fig. 11.3) it applies to all circuits which consist basically of a generator having internal resistance and working into a load resistance. Making the load resistance equal to the generator resistance is called *load matching*.

It does not follow that it is always desirable to make the load resistance equal to the generator resistance. Attempting to do so with a power-station generator would cause so much current to flow that it would be disastrous! But it is true, and can be confirmed by experiment, that the greatest power output *for a given signal voltage applied to the grid* is obtained when the load resistance is made equal to the valve resistance $r_a$. Again, the usual assumptions about linear characteristics apply.

## Transistors and Valves

The transistor seems to have fallen out by the way; can the same procedure be applied for calculating its amplification, or finding the most suitable load resistance? In principle, yes; but its equivalent generator and the calculations are considerably more complicated. So much so that beginners might find it discouraging to study them in detail at this stage, and perhaps confusing to have a mixture of valves and transistors in the various kinds of circuits to be considered in the chapters that follow. Although transistor circuits often look remarkably like the corresponding valve circuits, this similarity is somewhat deceptive. Since all the circuits can be considered in conjunction with valves, whose operation in them is on the whole easier to follow, more detailed study of transistors will be postponed to Chapter 22.

Advance notice is also given that use of the cathode of a valve as the electrode common to input (grid) and output (anode) is not the only arrangement possible; there are two others, using grid (p. 276) or anode (p. 315) as the common electrode. The same applies to the three electrodes of a transistor (p. 341).

## Varieties of Amplification

On p. 144 we reckoned voltage amplification, $A$, as the number of signal volts received out of the valve per signal volt put in. The same principle can be adopted with current amplification, such as that obtainable from a transistor. In both cases one must remember, of course, that the amplification actually given by a valve or transistor in a circuit is not the same thing as its voltage or current amplification *factor*, $\mu$ or $\alpha_e$. The actual amplification would only equal these factors if the load resistance of the valve were infinitely great; or, for the transistor, zero; and usually neither of these conditions is practical.

Account can also be taken of power amplification, the meaning of which is obvious. The only question that might be asked is how it applies to a valve, seeing that care is taken to avoid expenditure of any input power (p. 124). On that basis, provided any power at all was delivered by the valve, its power amplification would be infinitely great! It is true that for this reason the power amplification of a valve is often really meaningless and would not be quoted, but we shall see that in some circumstances, especially at very high frequencies (p. 241), a valve requires appreciable input power even when its grid is negatively biased.

The kind of amplification meant is sometimes indicated by a subscript, thus: $A_V$, $A_I$, $A_P$.

## Decibels

Suppose a 0·5 V signal is put into an amplifier which raises it to 12 V. This 12 V is then put into another amplifier which raises it to 180 V. Which amplifier amplifies most; the one that adds 11·5 V to the signal, or the one that adds 168 V? If we went by these figures we would have to say the second, because it added so many more volts to the signal than the first. In spite of this, most people would say the first, because its $A$ is 24 compared with 15 for the second. This instinct would be well-founded, because we do not judge by the amount by which a sound is increased in strength, but by the *ratio* in which it is increased. Doubling the power into a loudspeaker produces much the same impression of amplification whether it be from 50 to 100 mW or from 500 to 1,000.

This is like the way the ear responds to the pitch of sounds. While it is pleased by combinations of sounds whose frequencies have certain fixed ratios to one another, it can find no significance at all in any particular additive frequency intervals, say an increase of 250 c/s. Musical intervals are all ratios, and the significant ratios have special names, such as the octave, which is a doubling of frequency wherever on the frequency scale it may come. The octave is subdivided into tones and semitones.

The logical thing is to adopt the same principle with the *strength* of sound or electrical signals, and reckon amplification—or indeed any change in power—in ratios. Instead of the 2 to 1 ratio, or octave, a power ratio of 10 to 1 was chosen as the unit. The advantage of this choice is that the number in such units is simply equal to the common logarithm of the power ratio. It was named the bel (abbreviation: B) after A. Graham Bell, the telephone inventor. So an amplifier that multiplies the power by 10 can be said to have a gain of 1 bel. A further power multiplication by 10, giving a total of 100, adds another bel, making 2 bels.

The bel, being rather a large ratio for practical purposes, is divided into 10 decibels (dB). One tenth of a 10 to 1 ratio is not at all the same as one tenth of 10; it is 1·259 to 1. If you have any doubts, try multiplying 1 by 1·259 ten times.

So a gain of 1 dB is a 25·9% increase in power (which, incidentally, is about the least that is easily noticeable in a sound programme). The voltage or current increase giving a 25·9% increase in power is 12·2% *provided that the resistance is the same for both voltages (or currents) in the ratio*, because power is proportional to the square of voltage or current (p. 47). The fact that a voltage or current decibel is a smaller ratio than a power decibel does not mean that there are two different sizes of decibels, like avoirdupois and troy ounces; it arises because the ratio of volts or amps is smaller than the ratio of power in the same situation. Decibels are fundamentally *power* ratios.

If 1 is multiplied by 1·259 three times, the result is almost 2, so a doubling of power is a gain of 3 dB. Doubling voltage or current quadruples the power, so is a gain of 6 dB. A table connecting decibels with power and voltage (or current) ratios is printed on p. 376. Since the number of bels is $\log_{10} P_2/P_1$, or $2 \log_{10} V_2/V_1$ (where $P_2/P_1$ is the power ratio and $V_2/V_1$ the voltage ratio) the number of decibels is 10 $\log_{10} P_2/P_1$ or 20 $\log_{10} V_2/V_1$. If Fig. 0·5*a* is a scale of dB, then *b* is the same scale in power ratios. The dB scale, being linear, is obviously much easier to interpolate (i.e., to subdivide between the marked divisions). Fig. 11.4 would be improved if the logarithmic voltage-ratio scale were replaced or supplemented by a uniformly divided scale with its 0 at 1 and its 20 at 10, which would indicate dB.

An advantage of working in dB is that instead of having to multiply successive changes together to give the net result, as with the *A* figures, they are simply added. A voltage amplification ($A_V$) of 24 is + 27·6 dB, and $A_V = 15$ is +23·5 dB, so the total ($A_V = 360$) is + 51·1 db. The + indicates amplification or gain; attenuation or loss shows as a –. For instance, if the output voltage of some circuit were half the input, so that the power was one quarter—i.e., a power " gain " of 0·25—the 1 to 4 power ratio would be denoted by the same number of dB as 4 to 1, but with the opposite sign: –6 dB. Zero dB, equivalent to $A = 1$, means " no change ".

Another advantage is that when gain is graphed against something else, as for example in Fig. 11.4, the shape of the curve is unaffected by a gain or loss of an equal number of dB all over. The same curve plotted to an *A* scale would be distorted.

An important thing to realize is that decibels are not units of power, still less of voltage. So it is nonsense to say that the output from a valve is 20 dB, unless one has also specified a reference level, such as 0 dB = 50 mW. Then 20 dB, being 100 times the power, would mean 5,000 mW.

CHAPTER 12

# Oscillation

## Generating an Alternating Current

In the last few chapters we have considered alternating currents in a variety of circuits, but have taken the generator of such currents for granted, representing it in diagrams by a conventional symbol. The only sort of a.c. generator actually shown (Fig. 6.2) is incapable of working at the very high frequencies necessary in radio, and produces a square waveform which is generally less desirable than the " ideal " sine shape (Fig. 5.2).

The alternating currents of power frequency—usually 50 c/s in Great Britain and many other countries—are generated by rotating machinery which moves conductors through magnetic fields produced by electromagnets (p. 62). Such a method is quite impracticable for frequencies of millions per second. This is where the curious properties of inductance and capacitance come to the rescue.

## The Oscillatory Circuit

Fig. 12.1 shows a simple circuit comprising an inductor and a capacitor in parallel, connected between the anode of a valve and its positive source. The valve is not essential in the first stage of the experiment, which could be performed with a battery and switch, but it will be needed later on. With the grid switch as shown, there is no negative bias, and we can assume that the anode current flowing through the coil L is fairly large. It is a steady current, and therefore any voltage drop across L is due to the resistance of the coil only. As none is shown, we assume that it is too small to take into account, and therefore there is no appreciable voltage across L and the capacitor is completely uncharged.

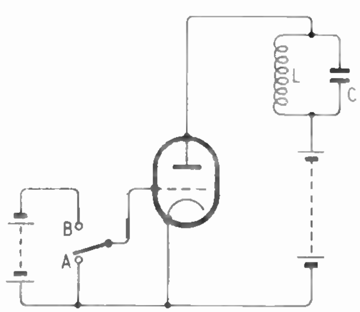



**Fig. 12.1—Valve-operated device for setting up oscillations in the circuit LC**

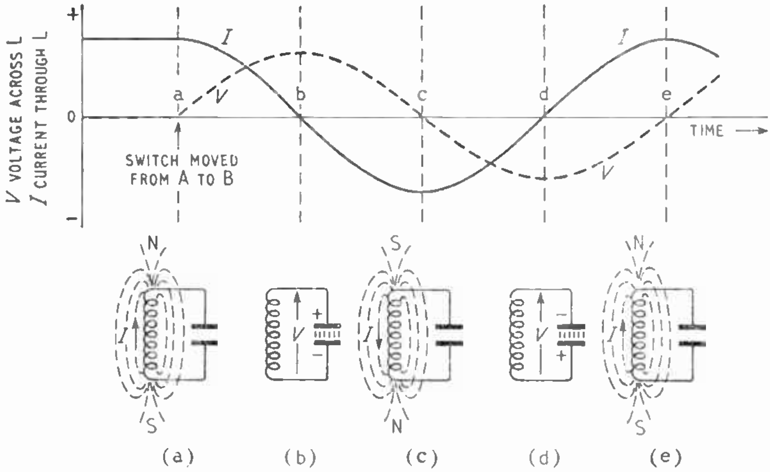



**Fig. 12.2—Sequence of events in Fig. 12.1 after the switch has been moved to B. Stages *a* to *e* cover one complete cycle of oscillation**

The current through L has set up a magnetic flux which by now has reached a steady state, and represents a certain amount of energy stored up, depending on the strength of current and on the inductance *L*. This condition is represented in Fig. 12.2 by *a*, so far as the circuit LC is concerned. The conventional direction of current (opposite to electron flow) is indicated.

Now suppose the switch is moved from A to B, putting such a large negative bias on the grid that anode current is completely cut off. If the current in L were cut off instantaneously, an infinitely high voltage would be self-induced across it (p. 63), but this is impossible owing to C. What actually happens is that directly the current through L starts to decrease, the collapse of magnetic flux induces an e.m.f. in such a direction as to oppose the decrease, that is to say, it tends to make the current continue to flow. It can no longer flow through the valve, but there is nothing to stop it from flowing into the capacitor C, charging it up. As it becomes charged, the voltage across it rises (p. 50), demanding an increased charging voltage from L. The only way in which the voltage can rise is for the current through it to fall at an increasing rate, as shown in Fig. 12.2 between *a* and *b*.

After a while the current is bound to drop to zero, by which time C has become charged to a certain voltage. This state is shown by the curves of *I* and *V* and by the circuit sketch at the point *b*. The magnetic field has completely disappeared, its energy having been transferred to C as an electric field.

The voltage across C must be balanced by an equal voltage across L, which can result only from the current through L continuing to fall at the same rate; that means it must become negative, reversing its direction. The capacitor must therefore be discharging through L. Directly it begins to do so it inevitably loses volts, just as a punctured tyre loses pressure. So the rate of change of current (which causes the voltage across L) gradually slackens until, when C is completely discharged and there is zero voltage, the current is momentarily steady at a large negative value (point *c*). The whole of the energy stored in the electric field between the plates of C has been returned to L, but in the opposite polarity. Assuming no loss in the to-and-fro movement of current (such as would be caused by resistance in the circuit), it has now reached a negative value equal to the positive value with which it started.

This current now starts to charge C, and the whole process just described is repeated (but in the opposite polarity) through *d* to *e*. This second reversal of polarity has brought us to exactly the same condition as at the starting point *a*. And so the whole thing starts all over again and continues to do so indefinitely, the original store of energy drawn from the battery being forever exchanged between coil and capacitor; and the current will never cease oscillating.

It can be shown mathematically that voltage and current vary sinusoidally.

## Frequency of Oscillation

From the moment the valve current was switched off in Fig. 12.1 the current in L was bound to be always equal to the current in C (shown by the curve $I$ in Fig. 12.2) because, of course, it was the same current flowing through both. Obviously, too, the voltage across them must always be equal. It follows that the reactances of $L$ and $C$ must be equal, that is to say $2\pi f L = 1/2\pi f C$. Rearranging this we get $f_0 = 1/2\pi\sqrt{LC}$, where $f_0$ denotes frequency of oscillation. This is the same as the formula we have already had (p. 105) for the frequency of resonance. It means that if energy is once imparted to an oscillatory circuit (i.e., L and C joined together as shown) it goes on oscillating at a frequency called the natural frequency of oscillation, which depends entirely on the values of inductance and capacitance. By making these very small—using only a few small turns of wire and small widely-spaced plates—very high frequencies can be generated, running if necessary into hundreds of millions per second.

## Damping

The non-stop oscillator just described is as imaginary as perpetual motion. It is impossible to construct an oscillatory circuit entirely without resistance of any kind. Even if it were possible it would be merely a curiosity, without much practical value as a generator, because it would be unable to yield more than the limited amount of energy with which it was originally furnished. After that was exhausted it would stop.

The inevitable resistance of all circuits—using the term resistance in its widest sense (p. 48)—dissipates a proportion of the original energy during each cycle, so each is less than its predecessor, as in Fig. 12.3. This effect of resistance is called *damping*; a highly damped oscillatory circuit is one in which the oscillations die out quickly. If the total effective series resistance $r$ is equal to $2\sqrt{L/C}$

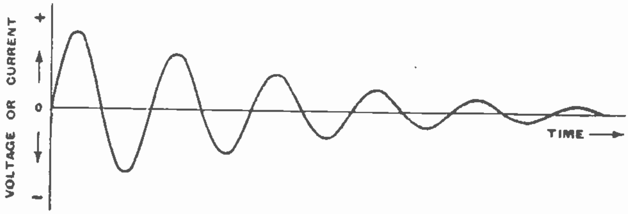



**Fig. 12.3.—The first few cycles of a train of damped oscillations**

the circuit is said to be critically damped; with less resistance the circuit is oscillatory, with more it is non-oscillatory and the voltage or current does not reverse after the initial kick due to an impulse but tends towards the shape shown in Figs. 3.6 and 4.9.

## The Valve-Maintained Oscillator

What we want is a method of keeping an oscillator going, by supplying it periodically with energy to make good what has been usefully drawn from it as well as what has unavoidably been lost in its own resistance. We have plenty of energy available from batteries, etc., in a continuous d.c. form. The problem is to apply it at the right moments.

The corresponding mechanical problem is solved in a clock. The pendulum is an oscillatory system which on its own soon comes to a standstill owing to frictional resistance. The driving energy of the mainspring is " d.c." and would be of no use if applied direct to the pendulum. The problem is solved by the escapement, which is a device for switching on the pressure from the spring once in each oscillation, the " switch " being controlled by the pendulum itself.

Turning back to Fig. 12.1 we have the oscillatory circuit LC which is set into oscillation by moving the switch from A to B at the stage marked *a* in Fig. 12.2. By the time marked *e* a certain amount of the energy would have been lost in resistance, but by operating the switch from B to A at point *c* the current will be built up to its full amplitude, at the expense of the anode battery, by the time *e* is reached.

What is needed, then, is something (an actual switch is not practicable for high frequencies) to make the grid more negative from *a* to *c* and more positive from *c* to *e*, and so on for each cycle.

At the frequency of oscillation the circuit LC is effectively a high resistance (p. 109). So we have all the essentials of a resistance-coupled amplifier (p. 140). Because the valve amplifies, quite a small part of the oscillating anode voltage, if applied to the grid, would release sufficient energy from the anode battery to cancel losses. From *a* to *c* the anode is more positive than its average, and from *c* to *e* more negative, which is just the opposite to what we have just seen to be necessary at the grid. We have already noticed (p. 142) that in a stage of valve amplification the anode voltage is opposite in polarity to the grid voltage causing it.

Suppose, as in Fig. 12.4*a*, a small generator of the required frequency is available to apply an alternating voltage $v_g$ to the grid. LC at that frequency being equivalent to a resistance, an amplified voltage $v_a$, *A* times as large, appears across it (Fig. 12.4*b*). If now a fraction $1/A$ of this output voltage is turned upside down and

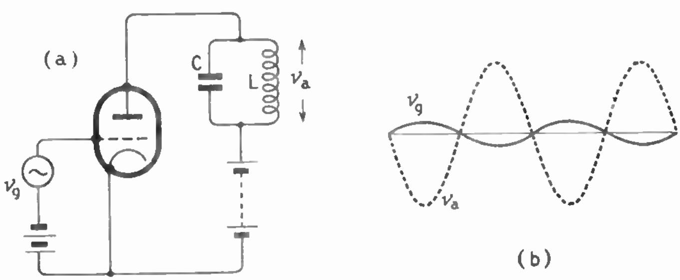



**Fig. 12.4—Diagram *b* shows the relative amplitude and phase of grid and anode voltages when their frequency is the same as the natural frequency of the circuit LC**

applied to the grid there is no need for the generator; the amplifier provides its own input and has become a generator of sustained oscillations, or, briefly, a valve oscillator.

## Valve Oscillator Circuits

The most obvious practical method of doing what has just been described is to connect a coil in the grid circuit and couple it inductively to L until it has induced in it sufficient voltage to "oscillate" (Fig. 12.5). The arrow indicates a variable coupling, obtained by varying the distance between the coils. The grid coil—called a *reaction* (or, more precisely, *retroaction*) coil—can be connected either way round in the circuit; one of these ways gives the reversal of sign necessary to maintain oscillation. The reaction process is sometimes called positive feedback. When the coupling is sufficiently close and in the right direction for oscillation it is not necessary to take any trouble to start it going; the ceaseless random movements of electrons in conductors (p. 30) are enough to start a tiny oscillation that very quickly builds up.

If the grid and anode are connected to opposite ends of the oscillatory coil, and some point in between is connected through the anode battery to the cathode and is thereby kept at a fixed potential, the alternating grid voltage is always opposite to the alternating anode voltage, and can be adjusted in relative magnitude by choosing the tapping point on L (Fig. 12.6). Two extra components are required, because if the grid were connected straight to the coil it would receive the full positive voltage of the anode battery which would upset things badly (p. 124). $C_g$ is used as a *blocking capacitor* to stop this (p. 55), and R is used to convey a suitable grid bias. The capacitance of $C_g$ is sufficient for its react-

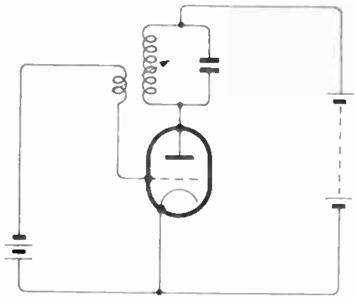


**Fig. 12.5—The reaction-coil valve oscillator circuit**

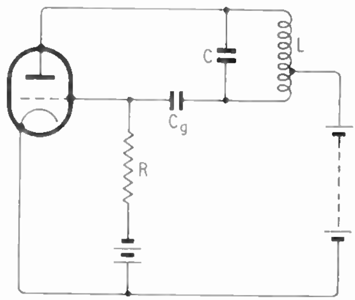



**Fig. 12.6—The series-fed Hartley circuit, showing parallel-fed grid bias using a grid leak R and capacitor $C_g$**

ance to offer little obstruction at the oscillatory frequency compared with the resistance of R, which is made large to prevent the alternating voltage from being short-circuited through the bias source. This circuit, known as the Hartley, is a very effective oscillator, useful when the losses are so great that it might be difficult to obtain sufficiently close coupling to get Fig. 12.5 to oscillate. Fig. 12.6 shows the *series-fed* variety, so called because the anode

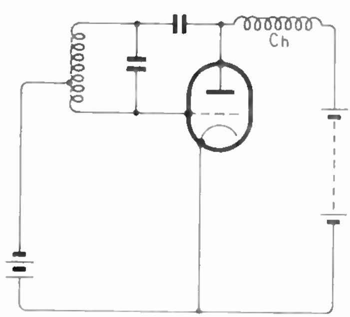



**Fig. 12.7— Parallel-fed Hartley circuit with the anode current supplied through a choke Ch. The grid bias is series-fed**

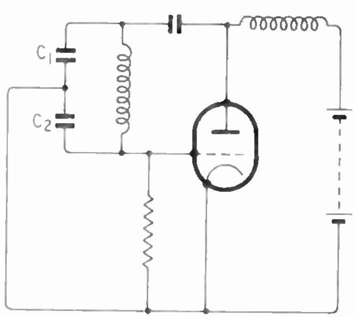



**Fig. 12.8—Parallel-fed Colpitts circuit**

feed current flows through the oscillatory circuit. The grid bias is parallel-fed through R, sometimes called a *grid leak*.

In the *parallel-fed* Hartley (Fig. 12.7) the anode is connected in this way, but to avoid loss of anode supply volts an inductor Ch, known as a choke coil, is used instead of a resistance. The grid is shown series-fed, but it is more often parallel-fed to allow the coil to be connected to the cathode, which is usually earthed. Moreover, for reasons to be explained in Chapter 16, the combination of $C_g$, R and the valve generates its own grid bias and the battery can be left out.

An alternative intermediate-potential point can be obtained on the oscillatory circuit by "tapping" the capacitor instead of the coil, in the Colpitts circuit, of which one variety is shown in Fig. 12.8. In this the capacitance of $C_1$ and $C_2$ in series is made equal to that of the one capacitor shown in the preceding circuits. The Colpitts is particularly favoured for very high frequency oscillators. $C_2$ then sometimes consists merely of the small capacitor formed by the grid and cathode of the valve itself.

Another method of obtaining the necessary reversal between anode and grid is to take advantage of the fact, shown in Fig. 8.2, that when a current flows through a coil and capacitor in series the voltages across them (neglecting resistance) are 180° out of phase (i.e., in opposite directions). This coil and capacitor, $L_1$ and $C_2$ in Fig. 12.9, are proportioned so that the reactance of $L_1$, and hence

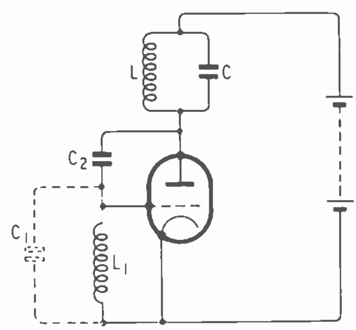



**Fig. 12.9—Tuned-anode tuned-grid (T.A.T.G.) oscillator circuit. The phase-reversing part of the circuit is $C_2$ (normally the valve itself) and $L_1$**

the voltage across it, is much less than that of $C_2$, so that the voltage across it is less and in opposite phase to that across $C_2$ and $L_1$ combined, which voltage is the same as that across the oscillatory circuit (battery voltages disregarded). In practice $C_2$ is usually the capacitance between the anode and grid themselves, while the inductive reactance between grid and cathode is adjusted by $C_1$. The name of this circuit, "tuned-anode tuned-grid", then seems reasonable; but it should not be taken to mean that grid and anode circuits are both tuned to the frequency of oscillation. If the combination $L_1$ $C_1$ were so tuned, it would be equivalent to a resistance, and not an inductive reactance as required; while LC must also be slightly off tune so that it has sufficient reactance to shift the phase slightly to make up for the fact that the phase shift in $C_2L_1C_1$, due to their resistance, is less than 180°.

The oscillator circuits shown in Figs. 12.5 to 12.9 look considerably different from one another at first glance, and can be modified to look even more different without affecting their basic similarity, which is that they are all valve amplifiers in which the grid *drive* or *excitation* is obtained by taking a part of the resulting anode alternating voltage and applying it in the opposite polarity, instead of going to some independent source for it.

## Amplitude of Oscillation

You might ask: what happens if the valve feeds back to its grid a greater voltage than is needed to maintain oscillation? The oscillation will grow in amplitude; the voltage at the grid will increase; that will be amplified by the valve, increasing the output at the anode; and so on. What stops it growing like a snowball?

Part of the answer is on p. 124, where it was pointed out that if the grid is allowed to become positive with respect to the cathode, it causes current to flow in that circuit and absorb power. When the amplitude of oscillation has grown to the point at which this loss of power due to grid current balances the surplus fed-back power, it will remain steady at that level.

If the grid bias is increased in an effort to obtain a greater amplitude of oscillation before grid current steps in to restrict it, another limiting influence is felt.

Look at Fig. 12.10 showing some typical valve characteristic curves of the same type as those in Fig. 11.2. Suppose the anode voltage is 300 V and the grid bias $-5$ V. Then any greater amplitude (i.e., peak value) than about 5 V causes grid current. Suppose also, for example, that the anode oscillatory circuit is equivalent at its resonant frequency to a resistance $R$ of 8 kΩ (dynamic resistance) represented by the load line shown.* Then if the grid voltage has an amplitude of 5 V, the anode voltage swings from 210 to 380 V; i.e., it has an amplitude of 85 V. If one-seventeenth of this (5 V) is fed back to the grid in the correct phase, the conditions for oscillation are fulfilled. Any tendency for a slight overcoupling to the grid causes it to go positive, grid current flows, and by transformer action (p. 96) this is equivalent to a resistance load in parallel with $R$, so reducing the anode voltage amplitude and hence the voltage fed back to the grid. And so a brake is put on any tendency for the amplitude of oscillation to increase indefinitely.

Now suppose the grid bias is increased, say to $-10$ V. The voltage amplitude at the grid can increase to about 10 V before grid current flows. Drawing the appropriate load line in Fig. 12.10, centred on $V_g = -10$, we see that the negative half-cycle of grid voltage cuts off the anode current altogether, so that it is distorted.

* Note that as the 8 kΩ applies only to a.c. of the resonant frequency, and not to d.c., the voltage actually reaching the anode is practically the same as that of the supply.

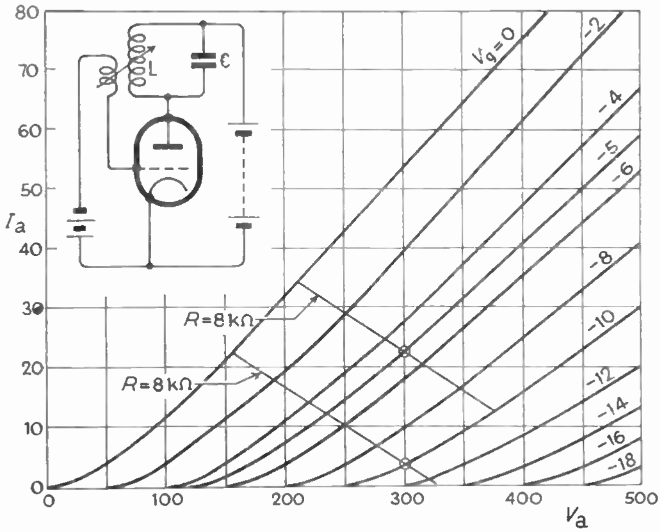



**Fig. 12.10—The oscillatory circuit LC is equivalent at resonance to a high resistance *R* and the voltage across it can be derived from a load line on the valve characteristic curves. From this follows the proportion to be fed back to the grid to sustain oscillation**

The whole conception of dynamic resistance was based on a sinusoidal current (p. 71), and the simple method of calculation employed above breaks down. It should be fairly clear, however, that if during a substantial part of each cycle of oscillation the resistance of the valve, $r_a$, is abnormally high, or even infinity, due to the oscillatory voltage reaching to and beyond the " bottom bend " of the characteristic curves, the average amplification of the valve is reduced, and the tendency to oscillate also reduced. So here is another factor limiting the amplitude of oscillation.

## Distortion of Oscillation

The previous section shows that if a valve oscillates fiercely the positive half-cycles of grid voltage cause grid current to flow, which imposes a heavy load on those parts of the cycle, and they are consequently clipped at the peaks. The corresponding anode-current peaks are therefore similarly distorted. The negative half-cycles may cause anode current to be cut off, so they are clipped too. The result in either or both cases is a distorted wave, which for some purposes is undesirable. Obviously, to obtain as nearly as possible a pure sine wave from a valve oscillator it is desirable so to adjust the back-coupling that a very slight clipping is sufficient to stop the amplitude of oscillation from growing any more.

It should be clear from Fig. 12.10 that the greater the dynamic resistance, *R*, of the anode circuit (load line more horizontal), the greater the amplitude that is possible before either grid current or bottom-bend limiting set in, and the less, in proportion, is a given amount of clipping at the peaks.

We have seen (p. 111) that the dynamic resistance is equal to $QX$, where $X$ is the reactance of either L or C at resonance. So to obtain a pure waveform the tuned circuit should have a high $Q$. This, incidentally, tends to increase the amplification, so that the amount of feedback needed to maintain oscillation is comparatively small.

There is a more important effect of high $Q$ in minimizing distortion. Fig. 8.11 showed how the smaller the series resistance, $r$, is in relation to the reactance $X$, the smaller is the external current fed to a parallel resonant circuit in relation to the current circulating round it. And we have seen that the ratio $X/r$ is $Q$. The same thing can be seen in another way by considering the circuit, Fig. 12.1, with which we started to explain oscillation. Here we started a sinusoidal oscillation by means of a square waveform produced by switching grid bias on and off. If the circuit has a high $Q$, the oscillatory current will build up to a value many times greater—to be precise, $Q$ times greater—than the valve anode current of that frequency. The valve anode current therefore is such a small proportion of the whole that it does not matter very much whether it is distorted or not.

For example, suppose $Q$ is 100, and $R$ is 20 kΩ. Then, as $R = QX$, $X$ is 200 Ω. The impedance to alternating current of the resonant frequency from the valve is thus 100 times greater than the impedance of the coil or of the capacitor. As the voltage across them is the same, the valve current is only 1% of the oscillatory current circulating internally between coil and capacitor, and any distortion it may have hardly counts. The oscillatory current itself is always sinusoidal in normal circumstances.

To obtain as nearly as possible a pure sine waveform from an oscillator, then, use a high-$Q$ circuit and the least back-coupling that will do.

## Stability of Frequency

The frequency of an oscillator, as we have seen, is adjusted by varying the inductance and/or capacitance forming the oscillatory circuit. For most purposes it is desirable that when the adjustment has been made the frequency shall remain perfectly constant. This means that the inductance and capacitance (and, to a less extent, the resistance) shall remain constant. For this purpose inductance includes not only that which is intentionally provided by the tuning coil, but also the inductance of its leads and the effect of any coils or pieces of metal (equivalent to single short-circuited turns) inductively coupled to it. Similarly capacitance includes that of the coil, the wiring, valve electrodes, and valve and coil holders. Resistance is varied in many ways, such as a change in the supply voltages of the oscillator valve, causing $r_a$ to alter.

Inductance and capacitance depend mainly on dimensions, which expand with rising temperature. If the tuning components are shut up in a box along with the oscillator valve, the temperature may rise considerably, and the frequency will drift for some time after switching on. Frequency stability therefore is aided by keeping the valve, and any resistors which develop heat, well away from the tuning components; by arranging effective ventilation; and by designing components so that expansion in one dimension offsets the effect of that in another. Fixed capacitors are obtainable employing a special ceramic dielectric material whose variations with temperature oppose those of other capacitors in the circuit; this reminds one of temperature compensation in watches and clocks.

Transistor oscillators are dealt with in Chapter 22.

CHAPTER 13

# The Sender

## Essentials of a Sender

The last few chapters provide the material for tackling in greater detail what was indicated in barest outline in Fig. 1.9. Considering the sending or transmitting system first, we have as the essentials for, say, broadcasting sound:

(*a*) A radio-frequency generator.
(*b*) A microphone, which may be regarded as an audio-frequency generator.
(*c*) A modulator, combining the products of (*a*) and (*b*).
(*d*) An aerial, radiating this combination.

In a television sender the camera takes the place of the microphone, and it generates a considerably wider band of frequencies, called *video* frequencies.

## The R.F. Generator

In the valve oscillator we have a r.f. generator capable of working at any frequency up to many millions of cycles per second. It is easily controlled, both in frequency and output. For short-range communication, or occasionally long ranges in special circumstances, ordinary receiving valves can be used, or even transistors. But to broadcast over a large area, or consistently over a long range, it is necessary to generate a large amount of power.

The difference between a receiving valve generating a few milliwatts (or at the most a few watts) of r.f. power, and a sending valve generating 100 kilowatts, is in the construction rather than the basic principles; and as this book deals with principles there is no need to go into the subject of high-power valves as such.

The higher the power the sender is required to produce, the more important it is that the percentage wasted shall be small. If, in generating 10 kW of r.f. power, 30 kW is wasted, the total power to be supplied is 40 kW, and the efficiency (ratio of useful power to the total power employed) is 25%. By increasing the efficiency to 75% it is possible to obtain 30 kW of r.f. for the same expenditure, or, alternatively, to reduce the power employed to 13·3 kW for the same output. In the latter case it would be possible to reduce the size and cost of the valves and other equipment as well.

Referring back to Fig. 12.10, suppose the upper load line, based on a grid bias of − 5 V and an anode voltage of 300, represents the voltage swing of an oscillator, working under practically distortionless or so-called " Class A " conditions. The average anode voltage applied is 300, and the average current 23 mA, so the input power is 6·9 W.

The useful output is obtained by multiplying r.m.s. alternating voltage by r.m.s. current. The swing is 170 V; therefore peak value $170/2 = 85$ V; r.m.s. value $85/\sqrt{2} = 60$ V. Current swing 22 mA; r.m.s. value 7·7 mA; power 0·462 W. This example points the way to a short cut for working out the a.c. power:

$$\text{Watts} = (\text{Voltage swing} \times \text{Current swing})/8.$$

The efficiency is therefore $0{\cdot}462/6{\cdot}9 = 0{\cdot}067$, or 6·7%. This is very poor indeed, because under the conditions illustrated there is a considerable anode current flowing all the time and a considerable voltage between anode and cathode of the valve. The power represented by this is dissipated as heat at the anode, and in the larger valves it is difficult to carry away all this heat. Many normally run bright red hot.

### High-Efficiency Oscillators: "Class B"

Obviously the aim should be to pass as little current through the valve as possible, consistent with keeping the oscillation going; and to pass that current only at times when the voltage across the valve is low.

As it happens, we started our study of oscillators on the right lines, because in Fig. 12.2 the valve was totally cut off by grid bias during half the cycle, from *a* to *c*; and was allowed to pass current only during the half-cycle from *c* to *e* when the voltage across the tuned circuit was in opposition to that of the supply, and therefore the anode voltage low. That is just what we want, and is obtained by biasing the valve until it just fails to pass current (i.e., the working point is at the "bottom bend").

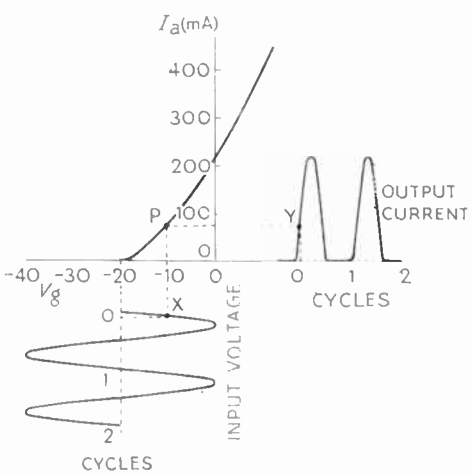



**Fig. 13.1 — Illustrating "Class B" operation with the aid of a valve anode-current/grid-voltage characteristic curve**

These conditions are shown in Fig. 13.1, representing a valve with a negative bias of 20 V and the feedback adjusted to give a peak grid voltage of the same amount, shown by the sine waves below the grid voltage scale.* The resulting anode current has half of each cycle suppressed, shown by the waveform on the right; but, as we have seen (p. 58), this current is normally very small in comparison with the sinusoidal oscillatory current.

## "Class C"

The process can be carried farther, and the efficiency further improved, by increasing the negative bias to well beyond cut-off. The grid swing has to be increased too, and the conditions, referred to as "Class C", are illustrated in Fig. 13.2. Note that the period during which anode current flows is less than half of each cycle, and

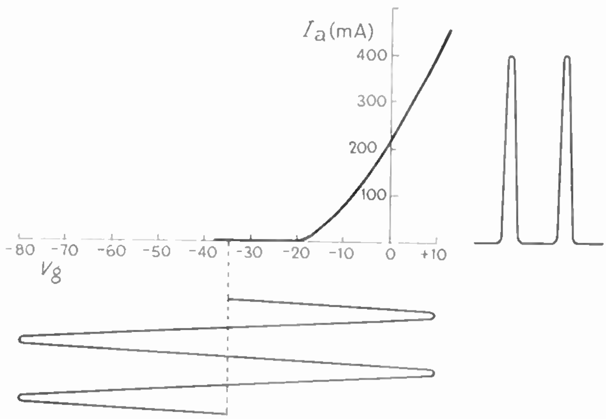



**Fig. 13.2—" Class C " operation. Comparing this with Fig. 13.1, we see that the working grid bias (—35 V in this example) is considerably more than enough to cut off anode current, which flows during only part of the positive half-cycles of signal**

is confined near the point *d* in Fig. 12.2, at which the anode voltage is a minimum, and therefore the power wasted in the valve is also very small.

Note too that it is advantageous to drive the grid positive, because, although there is a certain loss due to grid current, it occurs during only a small portion of each cycle, whereas the power developed in

* If the method of deriving the output waveform from the characteristic curve is unfamiliar it should be noted, as it will be used a good deal from now on. The output waveform is traced out point by point over the cycle; for example, at one-twelfth of a cycle (30°) from the start the input signal voltage is indicated by the point X. The corresponding point on the valve curve, P, is vertically above it, and indicates the anode current, which is then transferred horizontally to the left to give a point (Y) on the output waveform above 1/12th cycle on the output cycles scale. Points at other stages in the cycle are obtained similarly.

the anode circuit is much increased. Normally the valve is operated so that the positive grid voltage is nearly equal to the minimum anode voltage. Working efficiencies of the order of 75% are practicable, but beyond a certain point the power begins to drop off steeply. Moreover there may be difficulty in providing the very large grid swing.

Under Class C conditions the oscillator has either to be given a start or else begun with reduced bias, because full working bias cuts off anode current and prevents feedback. If worked as a self-oscillator (instead of merely a driven amplifier) the latter method is used. Chapter 16 will show a method of generating grid bias proportional to the amplitude of oscillation; an ideal arrangement for Class C, because it ensures the most powerful starting conditions.

## Constant Frequency

The need for constancy of frequency (p. 158) applies with particular force to radio senders. To avoid confusion it is internationally agreed that their frequencies must keep within very narrow limits; and for special purposes—such as working more than one sender on the same frequency—it is now common practice to keep within less than one cycle in a million.

In early days, the tuned oscillatory circuit (sometimes called a *tank circuit*, because of its capability for storing oscillatory energy) was coupled straight to the aerial, either inductively or by a direct tapping, as in Fig. 13.3. Aerials will be discussed in detail in Chapter 15; in the meantime they can be considered as opened-out capacitors (or sometimes inductors) necessary for effectively radiating waves over distances. They have both capacitance and inductance, which depend on their height above ground and other objects; so on a windy day are likely to vary erratically. Now, the frequency of an oscillator arranged as in Fig. 13.3 depends on the inductance and capacitance of the tuned circuit, of which the aerial forms part. The frequency stability of such a system is not nearly good enough for present-day needs.

Any circuits or components that can affect the frequency should be rigid and compact structures—compact, so that their electric or

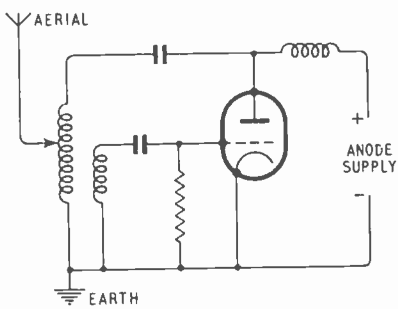



**Fig. 13.3—Direct coupling of aerial to the tuned circuit of a self-oscillating sender. The other end is usually connected to earth**

magnetic fields do not spread out where persons or things might move and upset the field distribution and hence the capacitance or inductance. We shall see (p. 240) that it is possible to prevent such variations by shutting up the components in a metal box or *screen*. But, of course, that is no good for radiating.

## The Master-Oscillator Power-Amplifier System

The way out of this dilemma is to separate the function of oscillating at the required frequency from that of feeding r.f. power to the aerial, instead of trying to make one valve do both. The oscillation frequency is generated by a valve oscillator in which all practicable precautions are taken to avoid frequency drift, and in which no attempt is made to obtain an extremely high power efficiency. This *master oscillator* is coupled, by some means that do not seriously affect the frequency stability, to one or more stages of amplification, the last of which is coupled to the aerial. Being designed to feed as much power to it as possible, this stage is called the *power amplifier*. The whole is therefore called a master-oscillator power-amplifier system; for short, M.O.P.A.

## Crystal Control

To obtain extreme frequency stability even in a M.O.P.A. requires very good design and considerable expense. Where it is not necessary to cover a range of frequency, but is enough to be able to work on a few " spot " frequencies, a way of guaranteeing correct frequency to within very narrow limits quite simply and cheaply is to exploit the remarkable properties of certain crystals, notably quartz. Such crystals are capable of vibrating mechanically at very high frequencies and in doing so develop an alternating e.m.f. between two opposite faces. Conversely, an alternating e.m.f. causes them to vibrate.

The subject is a large and specialized one, but for most purposes the crystal can be regarded as equivalent to a tank circuit of remarkably high inductance and $Q$. Unlike the coil-and-capacitor tank circuit, there is practically nothing about it that can be altered by fair wear and tear. There is a small change of frequency with temperature, but where requirements are stringent the crystal is kept in a temperature-controlled box.

The crystal is mounted between two flat metal plates, generally with a small air gap intervening. Fig. 13.4 shows two varieties of crystal oscillator circuit. Imagining the crystal to be a tuned circuit, we see *a* to be a T.A.T.G. circuit (compare Fig. 12.9), and *b* is a Colpitts circuit (compare Fig. 12.8). In *a* the tuning of the anode circuit is adjusted to a point at which oscillation takes place.

## Telegraph Senders: Keying

We have now reached the point—in theory at any rate—at which we can send out into space a continuous stream of waves of constant

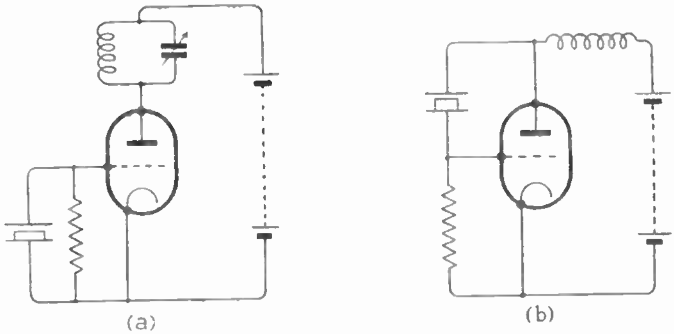



Fig. 13.4—Two types of crystal-controlled oscillator circuit

frequency and (if desired) high power. This alone, however, is not enough to convey messages, still less speech or music. The simplest way of sending messages is to break up the stream of waves into short and long bursts to represent the dots and dashes of the Morse code.

To do this it is necessary to connect a Morse key—which is simply a form of switch that makes a contact when pressed and breaks it when let go—in such a way that it starts and stops the radiation from the sender.

It might appear that almost any part of the circuit would do. In practice there are quite a number of things that have to be considered. The key should not be asked to break large amounts of power, because the resulting sparking would soon burn the contacts away. It cannot break a very high voltage, which would just spark across; and there is safety to consider, too. Therefore the anode circuit is practically ruled out, and so is the aerial circuit. The filament circuit is no good, because of the time taken to heat and cool. And wherever the keying is done it must not affect the frequency of oscillation, so it is better not to keep stopping and starting the master oscillator.

A common method of suppressing radiation is to apply a large negative bias to a low-power r.f. amplifier valve by disconnecting the cathode. Another method is to keep the radiation going all the time and use the key to shift its frequency.

## Radiotelephony and Broadcasting: Modulation

To convey speech and music or vision, something more elaborate than simple interruption of the transmission is required. It must be, as it were, moulded into the shape of the sounds. This process is called *modulation*, and the raw material that is modulated is called the *carrier wave*.* Keying is an extreme and special case of modula-

* Strictly, one should say " succession of waves " instead of " wave "; but that is taken as understood.

tion, in which the strength of the carrier wave is made to vary suddenly between zero and full power.

Suppose, however, that it is desired to transmit a pure note of 1 kc/s, available as an electric current derived from the microphone before which the note is being sounded. This current will have the form of a sine wave, as shown in Fig. 13.5*a*.

For the reasons given in Chapter 1, waves of such a low frequency cannot be radiated effectively through space. We therefore choose a r.f. carrier wave, say 1,000 kc/s. As 1,000 of its cycles occur in the time of every one of the a.f. cycles, they cannot all be drawn to the same scale, so the cycles in Fig. 13.5*b* are representative of the much greater number that have to be imagined. What we want is something which, like *b*, is entirely radio-frequency (so that it can be radiated), and yet carries the exact imprint of *a* (so that the a.f. waveform can be extracted at the receiving end). One way of obtaining it is to vary the amplitude of *b* in exact proportion to *a*, increasing the amplitude when *a* is positive and decreasing it when *a* is negative, the result being such as *c*. As we can see, this modulated wave is still exclusively r.f. and so can be radiated in its entirety. But although it contains no a.f. current, the tips of the waves trace out the shape of the a.f. waveform, as indicated by the dotted line, called the *envelope* of the r.f. wave train.

For simplicity we have considered a "programme" no more exciting than a tuning note; but the same principle applies to the

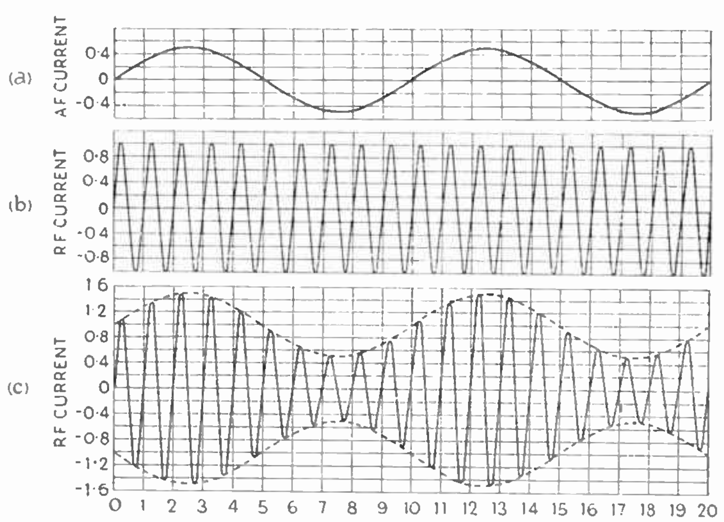



**Fig. 13.5—Curve *a* represents two cycles of audio-frequency current, such as might be flowing in a microphone circuit; *b* represents the much larger number of radio-frequency cycles generated by the sender. When these have been amplitude-modulated by *a*, the result is as shown at *c***

far more complicated waveforms of music and speech, or of light and shade in a televised scene—their complexity can be faithfully represented by the envelope of a modulated carrier-wave.

## Depth of Modulation

Since receivers are designed so that an unmodulated carrier wave (between items in the programme) causes no sound in the loud-speaker, it is evident that a graph such as Fig. 13.5*c* represents a note of some definite loudness, the loudness depending on the amount by which the r.f. peaks rise and fall above and below their mean value. The amount of this rise and fall, relative to the normal amplitude of the carrier wave, is the *depth of modulation.*

For distortionless transmission the increase and decrease in carrier amplitude, corresponding to positive and negative half-cycles of the modulating voltage, must be equal. The limit to this occurs

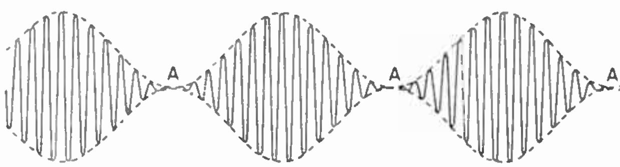



**Fig. 13.6—Carrier-wave modulated to a depth of 100%. At its minima (points A) the r.f. current just drops to zero**

when the negative half-cycles reduce the carrier-wave amplitude exactly to zero, as shown in Fig. 13.6. At its maximum it will then rise to double its steady value. Any attempt to make the maximum higher than this will result in the carrier actually ceasing for an appreciable period at each minimum; over this interval the envelope of the carrier-amplitude can no longer represent the modulating voltage, and there will be distortion.

When the carrier has its maximum swing, from zero to double its mean value, it is said to be modulated to a depth of 100%. In general, half the variation in amplitude, maximum to minimum, expressed as a percentage of the mean amplitude, is taken as the measure of modulation depth. Thus, when a 100-V carrier wave is made to vary between 50 V and 150 V, the depth of modulation is 50%.

In transmitting a musical programme, variations in loudness of the received music are produced by variations in modulation depth, these producing corresponding changes in the amount of a.f. output from the receiver.

## Methods of Modulation

Without going into details, we may note that one basic method of amplitude modulation is to vary the anode voltage of a r.f. oscillator or amplifier valve. Fig. 13.7 shows an outline of this method applied to an oscillator.

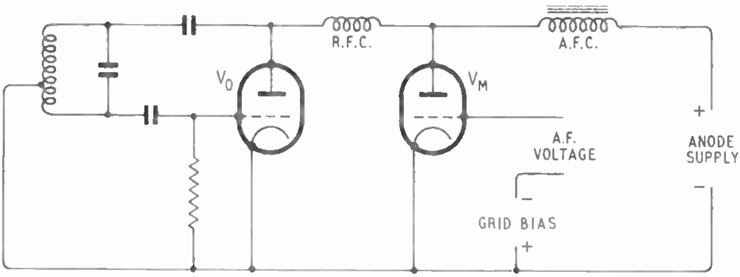



**Fig. 13.7—Anode modulation, sometimes called "choke control" modulation. RFC is a r.f. choke for keeping r.f. current out of the modulator circuit**

$V_O$ is the oscillator valve, shown working in a parallel-fed Hartley circuit (p. 155). If it were fed direct from the anode supply it would of course generate r.f. power of constant (i.e., unmodulated) amplitude. But the anode current to both valves is supplied through an inductor AFC (a.f. choke), which has a large impedance to all audio frequencies. So when a.f. voltage is supplied to the grid of $V_M$, that valve operates as an amplifier (p. 141) reproducing the a.f. voltage on a larger scale at its anode. This a.f. voltage alternately raises and lowers the anode voltage supplied to the oscillator valve $V_O$, modulating the oscillation.

Unfortunately the frequency of oscillation depends slightly on the anode voltage and consequently varies during the cycle of modulation. This rules the method out from modern practice, in which the frequency of the carrier wave must be kept constant within extremely narrow limits. It is therefore necessary to apply the method to a r.f. amplifier stage in a M.O.P.A. system (p. 164).

If it is applied to the final r.f. stage, another difficulty arises. Suppose for example that the anode of this stage is normally fed at 5,000 V. To cause 100% modulation it is necessary for the modulator alternately to reduce this to the verge of zero and raise it to 10,000 V. In doing this it has to control the full anode current of the r.f. valve, so the amount of power required by the modulator is of the same order as for the r.f. power amplifier. This is quite serious in a high-power sender and necessitates large and expensive modulator valves as well as a considerable addition to the power consumption.

Applying the modulating voltage to the grid of an oscillator or r.f. amplifier valve might seem to be a simpler and much more economical proceeding; but although grid modulation has been applied successfully it is less straightforward than might appear, and generally results in more distortion than does anode modulation.

There are a number of other methods of modulation, some of which are of practical value, especially in low-power senders. But the anode method, of which there are many variations, is the one generally favoured where freedom from distortion is important.

### Frequency Modulation

Instead of keeping the frequency constant and varying the amplitude, one can do vice versa. Fig. 13.8 shows the difference; *a* represents a small sample—two cycles—of the a.f. programme, as in Fig. 13.5, and *b* is the corresponding sample of r.f. carrier wave after it has been amplitude-modulated by *a*. If, instead, it were frequency-modulated it would be as at *c*, in which the amplitude remains constant but the frequency of the waves increases and decreases as the a.f. voltage rises and falls. In fact, with a suitable vertical scale of cycles per second, *a* would represent the frequency of the carrier wave.

The chief advantage of frequency modulation ("f.m.") is that under certain conditions reception is less likely to be disturbed by interference or noise. Since with f.m. the sender can radiate at its full power all the time, it is more likely to outdo interference than

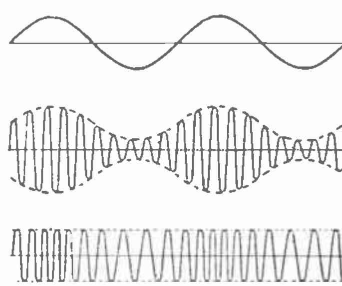



**Fig. 13.8—Curve (*a*) represents 2 cycles of audio frequency used to modulate a r.f. carrier wave. If amplitude modulation is used, the result is represented by (*b*); if frequency modulation, by (*c*)**

with amplitude modulation ("a.m."), which has to use a carrier wave of half maximum voltage and current and therefore only one quarter maximum power. At the receiving end, variations in amplitude caused by noise, etc., coming in along with the wanted signals can be cut off without any risk of distorting the modulation.

Another contribution to noise reduction results from the fact that modulation can be made to swing the carrier-wave frequency up and down over a relatively wide range or "band" as it is called. Exactly how this advantage follows is too involved to explain at the present stage, but since the change of frequency imparted by the modulator is proportional to the a.f. voltage it is obvious that the greater the change the greater the audible result at the receiving end.

In a.m., 100% modulation varies the amplitude between zero and double. If a f.m. station varied its frequency between zero and double it would (apart from other impracticabilities!) leave no frequencies clear for any other stations, so some arbitrary limit must be imposed, to be regarded as 100% modulation. This limit is called the *deviation*, and for sound broadcasting is commonly 75 kc/s above and below the frequency of the unmodulated carrier wave. Stations working on the "medium" broadcast frequencies (525–1,605 kc/s) are spaced only 9 or 10 kc/s apart, and if they used f.m. they would spread right across one another. For this and other reasons it is necessary for f.m. stations to use much higher frequencies, usually in the "very high frequency" (v.h.f.) band, 30 to 300 Mc/s. Although the principle of f.m. is simple, the actual methods of modulation adopted in practical senders are too complex to describe here.

A silent background to reception is especially desirable in broadcasting, and for this reason f.m. is much used for sound programmes, with or without accompanying vision. It is also widely used in radio telegraphy and still-picture transmission.

## Microphones

So far we have learnt something about two of the boxes at the sending end of Fig. 1.9: the r.f. generator and the modulator.

A device by which sound, consisting of air waves, is made to set up electric currents or voltages of identical waveform, which in turn are used to control the modulator, is one that most people handle more or less frequently, in telephones. It is the *microphone*; and the type used in telephones, but not for radio broadcasting, is essentially the same as the original invention by Hughes in 1878. The sounds are directed on to a circular diaphragm, to the centre

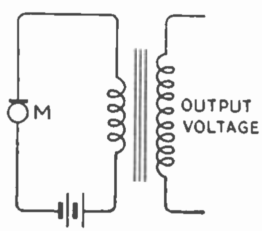



**Fig. 13.9—How a carbon microphone (M) is connected to give an a.f. voltage**

of which is attached a carbon button. Beyond it is a fixed carbon button, and the space between is filled with carbon granules. So between the terminals, which are connected to the diaphragm and fixed button, there are many not very firm contacts between pieces of carbon. Their resistance depends on the pressure, which is varied by the air waves striking the diaphragm.

A carbon microphone is, in fact, a resistance that varies to correspond with the sound waves, and so the current delivered by the battery in Fig. 13.9 varies in the same way. Since this current passes through the primary of a transformer, a varying voltage is produced at the secondary terminals. Sound waves of normal intensity are very feeble, and even with a step-up transformer the result is generally only a few volts at most, so for modulating a powerful sender several stages of amplification may be necessary.

While such a microphone can be designed to give a reasonably large output over the band of frequencies corresponding to the essentials of intelligible speech, it does so by virtue of a certain amount of mechanical resonance, and many of the frequencies needed to give full naturalness to speech and music are not fairly represented. If the resonance is reduced in order to give more uniform response to all the audio frequencies, then the output becomes less, and more amplification is needed. Compare electrical resonance, as in Fig. 8.6. The carbon microphone is also, on account of its loose contacts, noisy. So various other systems have been tried; but generally speaking it is true that the higher the quality—i.e., the more faithfully the electrical output represents the sound—the lower the output.

The electrostatic microphone consists of a diaphragm, usually metal, with a metal plate very close behind. When vibrated by sound, the capacitance varies, and the resulting charging and discharging currents from a battery, passed through a high resistance, set up a.f. voltages. Its disadvantages are very low output and high impedance.

The crystal or piezo-electric microphone depends on the same properties as are applied in crystal control (p. 164); the varying pressures directly give rise to a.f. voltages. Crystals of Rochelle salt are particularly effective.

The type most commonly used for broadcasting and the better class of " public address " is the electromagnetic. There are many varieties: in some an iron diaphragm varies the magnetic flux through a coil, so generating voltages; in others the diaphragm bears a moving coil in which the voltages are generated by flux from a stationary magnet; while in the most favoured types a very light metal ribbon is used instead of the coil. Electromagnetic and crystal microphones generate e.m.fs directly, so need no battery; but electromagnetic types need a transformer to step them up.

The counterpart of the microphone in a vision system is the camera, which is described on p. 333.

## Coupling to Aerial

The last main component is the aerial. This is the subject of Chapter 15; but one aspect that we can consider at this point is the way in which the sender is connected to the aerial, because the power efficiency is greatly affected thereby.

We saw (p. 142) that a valve can be regarded as a generator having internal resistance ($r_a$), and that the maximum output power for a given alternating voltage at the grid is obtained by making the load resistance equal to the internal resistance. In a sender, the valve in question is the output or " last-stage " valve, and the load is the dynamic resistance of the tuning circuit combined with the aerial. Generally it is practicable to keep the losses in the tuning circuit relatively small, so we shall regard the aerial as the load. Assuming for the moment that it can be regarded as a resistance, should our aim be to make it equal to $r_a$? If so, the efficiency—which means that fraction of the total power generated in the valve which is delivered to the aerial—is only 50%, because half the power is wasted in the valve. But the real limiting factor is not the grid e.m.f.—which can readily be increased—but the amount of heat the valve can dissipate (p. 161). If, for example, the rated anode dissipation of a valve is 100 W, and the efficiency of power transfer to the aerial is 50%, then it can put up to 100 W into the aerial. But if, by increasing the load resistance *and* the input voltage, the efficiency is raised to 66⅔%, the maximum output will be double the dissipation, namely, 200 W. For various reasons, the efficiency cannot be raised indefinitely in this way, so the valve manufacturer specifies the *optimum load resistance* for each type of valve.

Leaving detailed discussion of aerials as loads to Chapter 15, we can assume in the meantime that their impedance varies greatly with frequency. If by some stroke of luck the aerial impedance at the working frequency were to equal the optimum load resistance for the transmitting valve we could connect the aerial and earth terminals to opposite ends of the anode tuning coil. In general, however, the effective impedance between these terminals will not be equal to the optimum for the valve. A matching device is necessary.

On p. 97 we saw that a transformer is just such a device, because by a suitable choice of turns ratio any resistance can be made

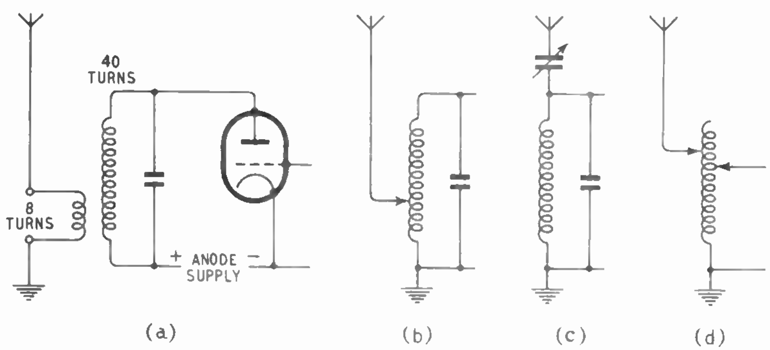



**Fig. 13.10**—Some of many varieties of method in coupling aerial to sender. *a* is the inductive coupling or r.f. transformer method, worked out in the text for a particular example; *b* is the aerial tapping or auto-transformer; *c* is the series capacitor method; and at *d* the valve is tapped down to reduce the effective load resistance

equivalent to any other resistance. The rule was that to match $R_s$ to $R_p$ the transformer should have $\sqrt{(R_p/R_s)}$ primary turns for every secondary turn. That is the same thing as $\sqrt{(R_s/R_p)}$ secondary turns for every primary turn. For example, if our tuning coil had 40 turns, and the optimum load for the valve was 3,000 Ω, and at the working frequency the aerial was equivalent to a resistance between aerial and earth terminals of 120 Ω, then it should be coupled to the tuning coil by a secondary winding of $40 \times \sqrt{120/3{,}000} = 8$ turns, as in Fig. 13.10*a*.

Alternatively, an auto-transformer (p. 98) may be used, *b*. There are other ways of matching, such as *c*. Like method *a*, it insulates the aerial from the anode coil, which may be at anode potential. Method *d* is used when the optimum load is *less* than the aerial impedance.

Although it is easy to work out the right number of turns according to this simple theory (and rather less easy to calculate the right value of capacitance in Fig. 13.10*c*) there are complications that make it a more formidable task to calculate the best practical design. For example, we assumed that *all* the magnetic flux linked both primary and secondary coils, whereas this is far from being so in r.f. transformers. In practice, the right tapping or coupling is calculated as nearly as possible, and provision made for it to be varied until the best results are obtained. Fig. 11.4 shows that it is necessary to depart quite a lot from the optimum before there is a serious loss in output.

Another complication is that an aerial nearly always has reactance as well as resistance, which affects the tuning and makes it necessary to use less or more tuning capacitance; less, if the reactance effectively in parallel with the coil is capacitive, and greater if it is inductive.

CHAPTER 14

# Transmission Lines

## Feeders

IN DEALING WITH circuit make-up we have for the most part assumed that resistance, inductance and capacitance are concentrated separately in particular components, such as resistors, inductors and capacitors. True, any actual inductor (for example) has some of all three of these features, but it can be imitated fairly accurately by an imaginary circuit consisting of separate $L$, $C$ and $R$ (p. 114). In the next chapter, however, we shall turn attention to aerials, in which $L$, $C$ and $R$ are all mixed up and distributed over a considerable length. The calculation of aerials is therefore generally much more difficult than anything we have attempted, especially as they can take so many different forms.

But one example of distributed-circuit impedance does lend itself to rather simpler treatment, and makes a good introduction to the subject, as well as being quite important from a practical point of view. It is the *transmission line*. The most familiar example is the cable connecting the television aerial on the roof to the set

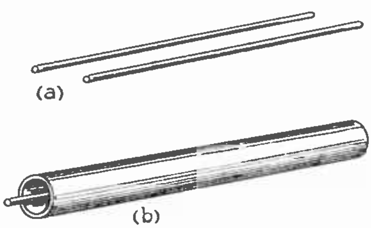



**Fig. 14.1—The two main types of r.f. transmission line or feeder cable: (*a*) parallel-wire, and (*b*) coaxial**

indoors. For reasons to be explained later (p. 332) television uses signals of very high frequency, which are best received on aerials of a definite length (p. 193). It is desirable that the wires connecting aerial to receiver should not themselves radiate or respond to radiation, nor should they weaken the v.h.f. signals. The corresponding problem at the sending end is even more important, because a large amount of power has to be conveyed from the sender on the ground to an aerial hundreds of feet up in the air. A transmission line for such purposes is usually called a *feeder*.

To reduce radiation and pick-up to a minimum, its two conductors must be run very close together, so that effects on or from one cancel out those of the other. One form is the parallel-wire feeder (Fig. 14.1*a*); still better is the coaxial feeder (*b*), in which one of the conductors totally encloses the other.

## Electrical Equivalent of a Line

With either type, and especially the coaxial, the closeness of the two leads causes a high capacitance between them; and at very high frequencies that means a low impedance, which, it might be supposed, would more or less short-circuit a long feeder, so that very little of the power put in at one end would reach the other. But it must be remembered that every inch of the feeder not only has a certain amount of capacitance in parallel, but also, by virtue of the magnetic flux set up by any current flowing through it (Fig. 4.2), a certain amount of inductance in series. Electrically,

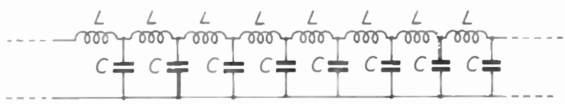



**Fig. 14.2—Approximate electrical analysis of a short length of transmission line**

therefore, a piece of transmission line can be represented with standard circuit symbols as in Fig. 14.2, in which $L$ and $C$ are respectively the inductance and capacitance of an extremely short length, and so are extremely small quantities. Each inductance is, of course, contributed to by both wires, but it makes no difference to our argument if for simplicity the symbol for the total is shown in one wire; it is in series either way. As we shall only be considering lines that are uniform throughout their length, every $L$ is accompanied by the same amount of $C$; in other words, the ratio $L/C$ is constant.

Having analysed the line in this way, let us consider the load at the receiving end; assuming first that its impedance is a simple

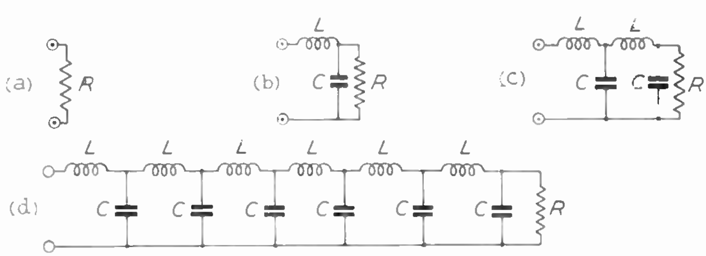



**Fig. 14.3—Synthesis of transmission line from basic circuit elements**

resistance, $R$ (Fig. 14.3$a$). If we were to measure the voltage across its terminals and the amperage going into it, the ratio of the two readings would indicate the value $R$ in ohms.

Next, suppose this load to be elaborated by adding a small inductance in series and a small capacitance in parallel, as at $b$. Provided that the capacitance really is small, its reactance, $X_C$, will be much greater than $R$. By an adaptation of the method used on p. 114 a relatively large reactance in parallel with a resistance can, with negligible error, be replaced by an equivalent circuit consisting of the same resistance in series with a relatively small reactance. Calling this equivalent series capacitive reactance $X_C'$, we have $X_C'/R = R/X_C$, or $X_C' = R^2/X_C$. And if we make this $X_C'$ equal to $X_L$, it will cancel out the effect of $L$ (p. 101), so that the whole of Fig. 14.3*b* will give the same readings on our measuring instruments as $R$ alone. $X_L$ is, of course, $2\pi fL$, and $X_C$ is $1/2\pi fC$; so the condition for $X_C'$ and $X_L$ cancelling one another out is

$$2\pi fL = R^2 \times 2\pi fC$$

$$\text{or,} \quad L = R^2C$$

$$\text{or,} \quad R = \sqrt{\frac{L}{C}}$$

Notice that frequency does not come into this at all, except that if it is very high then $L$ and $C$ may have to be very small indeed in order to fulfil the condition that $X_C$ is much larger than $R$.

Assuming, then, that in our case $\sqrt{L/C}$ does happen to be equal to $R$, so that $R$ in Fig. 14.3*a* can be replaced by the combination $LCR$ in *b*, it will still make no difference to the instrument readings if $R$ in *b* is in turn replaced by an identical $LCR$ combination, as at *c*. $R$ in this load can be replaced by another $LCR$ unit, and so on indefinitely (*d*). Every time, meters connected at the terminals will show the same readings, just as if the load were $R$ only.

The smaller $L$ and $C$ are, the more exactly the above argument is true at all practical frequencies. This is very interesting, because by making each $L$ and $C$ smaller and smaller and increasing their number we can approximate as closely as we like to the electrical equivalent of a parallel or coaxial transmission line in which inductance and capacitance are uniformly distributed along its length. In such a line the ratio of inductance to capacitance is the same for a foot or a mile as for each one of the infinitesimal " units ", so we reach the conclusion that provided the far end of the line is terminated by a resistance $R$ equal to $\sqrt{L/C}$ ohms the length of the line makes no difference; to the signal-source or generator it is just the same as if the load $R$ were connected directly to its terminals.

That, of course, is exactly what we want as a feeder for a load which has to be located at a distance from the generator. But there is admittedly one difference between a real feeder and the synthetic one shown in Fig. 14.3*d*; its conductors are bound to have a certain amount of resistance. Assuming that this resistance is uniformly distributed, every 100 feet of line will absorb a certain percentage of the power put into it; for example, if the power put into a 100-foot length has to be 25·9% greater than that coming out at the far end, the loss is said to be 1 dB per 100 feet (p. 147); and we then know that a 20-foot length of the same line will cause a

0·2-dB loss. Obviously, a line made of a very thin wire, of comparatively high resistance, will cause greater loss than one made of low-resistance wire. So, too, will a pair separated by poor insulating material (p. 113). Although the loss due to resistance is an important property of a line, it is generally allowable to neglect the resistance itself in calculations such as that at the top of this page.

## Characteristic Resistance

The next thing to consider is how to make the feeder fit the load, so as to fulfil the necessary condition, $\sqrt{L/C} = R$. $\sqrt{L/C}$ is obviously a characteristic of the line itself, and to instruments connected to one end the line appears to be a resistance; so for any particular line it is called its *characteristic resistance*, denoted by $R_0$.*

If you feel that the resistance in question really belongs to the terminating load and not the line, then you should remember that the load resistance can always be replaced by another length of line, so that ultimately $R_0$ can be defined as the input resistance of an infinitely long line. Alternatively (and more practically) it is a resistance equal to whatever load resistance can be fed through any length of the line without making any difference to the generator.

In any case, $L$ and $C$ depend entirely on the spacing and diameters of the wires or tubes of which the line consists, and on the permittivity and permeability of the spacing materials. So far as possible, to avoid losses, feeders are air-spaced. The closer the spacing, the higher is $C$ (as one would expect) and the smaller is $L$ (because the magnetic field set up by the current in one wire is nearly cancelled by the field of the returning current in the other). So a closely-spaced line has a small $R_0$, suitable for feeding low-resistance loads. Other things being equal, one would expect a coaxial line to have a greater $C$, and therefore lower $R_0$, than a parallel-wire line. Formulae have been worked out for calculating $R_0$; for a parallel-wire line (Fig. 14.4*a*) it is practically 276 $\log_{10}(2D/d)$ (so long as $D$ is at least 4 or 5 times $d$), and for a coaxial line (Fig. 14.4*b*) 138 $\log_{10}(D/d)$. Graphs such as these can be used to find the correct dimensions for a feeder to fit a load of any resistance, within the limits of practicable construction. It has been found that the feeder dimensions causing least loss make $R_0$ equal to about 600 Ω for parallel-wire and 80 Ω for coaxial types; but a reasonably efficient feeder of practical dimensions can be made to have an $R_0$ of anything from, say 200 to 650 and 40 to 140 ohms respectively. We shall see later what to do if the load resistance cannot be matched by any available feeder.

In the meantime, what if the load is not a pure resistance? Whatever it is, it can be reckoned as a resistance in parallel with a

* If loss resistance is taken into account the expression is slightly more complicated than $\sqrt{L/C}$, and includes reactance; so the more comprehensive and strictly accurate term is characteristic (or surge) impedance, $Z_0$. With reasonably low-loss lines there is not much difference.

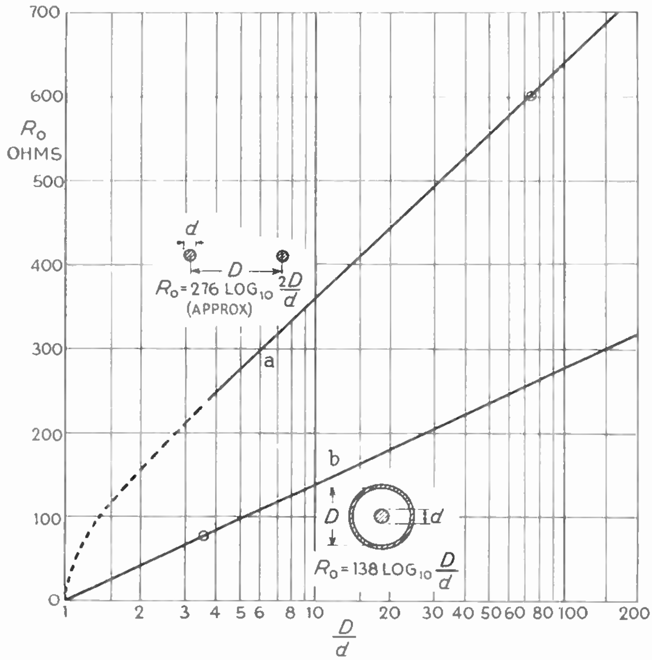



**Fig. 14.4—Graphs showing characteristic resistance ($R_0$) of ($a$) parallel-wire and ($b$) coaxial transmission lines in terms of dimensions**

reactance. And this reactance can always be neutralized or tuned out by connecting in parallel with it an equal reactance of the opposite kind (p. 106). So that little problem is soon solved.

## Waves along a Line

Although interposing a loss-free line (having the correct $R_0$) between a generator and its load does not affect the voltage and current at the load terminals—they are still the same as at the generator—it does affect their phase. The mere fact that the voltage across $R$ in Fig. 14.3$b$ is equal to the voltage across the terminals makes that inevitable, as one can soon see by drawing the vector diagram. There is a small phase lag due to $L$. In a line, the phase lags adds up steadily as one moves along it. If we consider what happens in a line when the generator starts generating (Fig. 14.5) we can see that what the gradual phase lag along it really means is that the current takes time to travel from generator to load.

During the first half-cycle (positive, say) current starts flowing into $C_1$, charging it up. The inductance $L_1$ prevents the voltage across $C_1$ rising to its maximum until a little later than the generator voltage maximum. The inertia of $L_2$ to the growth of current through it allows the charge to build up in $C_1$ a little before $C_2$, and so on. Gradually the charge builds up on each bit of the line in turn. Meanwhile, the generator has gone on to the negative half-cycle, and this follows its predecessor down the line. At its first negative maximum, the voltage distribution will be as shown at *b*. With a little imagination one can picture the voltage wave flashing down the line, rather like the wave motion of a long rope waggled up and down at one end.

Our theory showed that wherever the line is connected to measuring instruments it appears as a resistance, which means that the current is everywhere in phase with the voltage, so the wave diagram also represents power flowing along the line.

The speed of the wave clearly depends on its capacitance and inductance; the larger they are the longer the time each bit of line takes to charge and the longer the time for current to build up in it. We have already noted that $C$ is increased by spacing the two conductors more closely, but that the $L$ is reduced thereby. As it happens, the two effects exactly cancel one another out, so the speed of the wave is unaffected; the only result is that the ratio $L/C$—and hence $R_0$—is decreased. The only way of increasing $C$ without reducing $L$ is to increase the permittivity in the space between the conductors (p. 53); similarly $L$ is affected by the permeability there (p. 64). It has been discovered that the speed of the waves is equal to $1/\sqrt{\epsilon\mu}$. The lowest possible permittivity and permeability are in empty space (though air is practically the same); in m.k.s. units they are $8{\cdot}854 \times 10^{-12}$ and $1{\cdot}257 \times 10^{-6}$ respectively, so the speed of the waves works out at $1/\sqrt{(8{\cdot}854 \times 1{\cdot}257 \times 10^{-18})}$ = very nearly 300 million metres per second, or 186,240 miles per second.

There have to be solid supports for the conductors, and sometimes the space between is completely filled. In practice the permeability of such material would not differ appreciably from that of empty

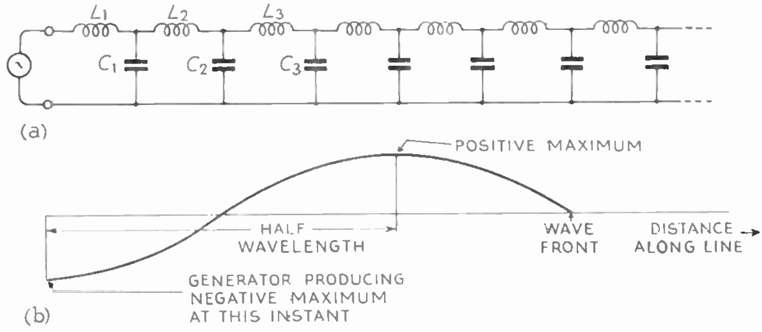



**Fig. 14.5—(*a*) Electrical representation of the generator end of a transmission line, and (*b*) instantaneous voltage (and current) diagram $\frac{3}{4}$ of a cycle from the start of generating a sine wave**

space, but the permittivity would be higher. Consequently the speed of the waves would be lower; for example, if the relative permittivity were 4 the speed would be reduced to a half. But it would still be pretty high! So the time needed to reach the far end of any actual line is bound to be very short—a small fraction of a second. But however short, there must be some interval between the power first going into the line from the generator and its coming out of the line into the load. During this interval the generator is not in touch with the load at all; the current pushed into the line by a given generator voltage is determined by $R_0$ alone, no matter what may be connected at the far end; which is further evidence that $R_0$ is a characteristic of the line and not of the load.

During this brief interval, when the generator does not " know " the resistance of the load it will soon be required to feed, the $R_0$ of the line controls the rate of power flow tentatively. In accordance with the usual law (p. 148), the power going into the line during this transient state is a maximum if $R_0$ is equal to $r$, the generator resistance.

Neglecting line loss, the voltage and current reaching the load end will be the same as at the generator. So if the load turns out to be a resistance equal to $R_0$, it will satisfy Ohm's law, and the whole of the power will be absorbed just as fast as it arrives. For example, if a piece of line having an $R_0$ of 500 Ω is connected to a 1,000-V generator having an $r$ or 500 Ω, the terminal voltage is bound to be 500 V and the current 1 A until the wave reaches the far end, no matter what may be there. If there is a load resistance of 500 Ω, the r.m.s. current will, of course, continue at 1 A everywhere.

## Wave Reflection

But suppose that the load resistance is, say, 2,000 Ω. According to Ohm's law it is impossible for 500 V applied across 2,000 Ω to cause a current of 1 A to flow. Yet 1 A is arriving. What does it do? Part of it, having nowhere to go, starts back for home. More scientifically, it is *reflected* by the mismatch. The reflected current, travelling in the opposite direction, can be regarded as opposite in phase to that arriving, giving a total which is less than 1 A. The comparatively high resistance causes the voltage across it to build up above 500; this increase can be regarded as a reflected voltage driving the reflected current. If half the current were reflected, leaving 0·5 A to go into the load resistance, the voltage would be increased by a half, making it 750. A voltage of 750 and current 0·5 A would fit a 1,500 Ω load, but not 2,000 Ω, so the reflected proportion has to be greater—actually 60%, giving 800 V and 0·4 A at the load.

We now have 1 A, driven by 500 V, travelling from generator to load, and 0·6 A, driven by 300 V, returning to the generator. (The ratio of the reflected voltage to the reflected current must, of course, equal $R_0$.) The combination of these two at the terminals of the load gives, as we have seen, $1 - 0{\cdot}6 = 0{\cdot}4$ A, at $500 + 300 =$ 800 V. But at other points on the line we have to take account of the phase lag. At a distance from the load end equal to quarter of a wavelength ($\lambda/4$) the arriving and returning waves differ in phase by half a wavelength ($\lambda/2$) or 180° as compared with their relative phases at the load (because a return journey has to be made over the $\lambda/4$ distance). So at this point the current is $1 + 0{\cdot}6 = 1{\cdot}6$ A at $500 - 300 = 200$ V. At a point $\lambda/8$ from the load end the phase separation is $\lambda/4$ or 90°, giving 1·16 A at 582 V. At intervals of $\lambda/2$, the two waves come into step again.

## Standing Waves

Calculating the current and voltage point by point in this way and plotting them, we get the curves in Fig. 14.6. It is important to realize that these are not, as it were, flashlight photographs of

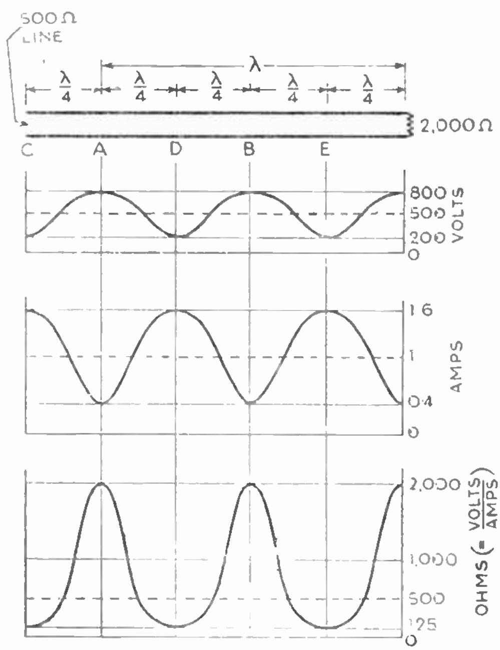



**Fig. 14.6 — Diagrams of voltage, current, and (as a result) impedance at the load end of an unmatched transmission line, showing standing waves**

the waves travelling along the line; these are r.m.s. values set up continuously at the points shown, and would be indicated by meters connected in or across the lines at those points (assuming the meters did not appreciably affect the $R_0$ of the line where connected). Because this wavelike distribution of current and voltage, resulting from the addition of the arriving and reflected waves travelling in opposite directions, is stationary, it is called a *standing wave.* For comparison, the uniform distribution of current and voltage when the load resistance is equal to $R_0$ is shown dotted.

The ratio of maximum to minimum current or voltage is called the *standing wave ratio*; in our example it is 800/200 (or 1·6/0·4) = 4.

In due course the reflected wave reaches the generator. It is in this indirect way that the load makes itself felt by the generator. If the 2,000-Ω load had been directly connected to the generator terminals, the current would have been 1,000/(500 + 2,000) = 0·4 A, and the terminal voltage 0·4 × 2,000 = 800 V, and the power 800 × 0·4 = 320 W. This, as we have seen, is exactly what the load at the end of the line is actually getting. But the power that originally went out from the generator, being determined by $R_0$, was 500 × 1 = 500 W. The reflected power is 300 × 0·6 = 180 W, so the net outgoing power is 500 − 180 = 320 W, just as it would have been with the load directly connected.

## LINE IMPEDANCE VARIATIONS

So the power adjustment is (in this case) quite simple. But the current and voltage situation at the generator is complicated by the time lag, and is not necessarily the same as at the load. Unless the length of the line is an exact multiple of $\lambda/2$, the phase relationships are different. Suppose, for example, it is an odd multiple of $\lambda/4$; say, $5\lambda/4$, as in Fig. 14.6. Then the current and voltage at the generator end will be 1·6 A and 200 V respectively. That makes 320 W all right—but compare the power loss in the generator. 1·6 A flowing through $r$ = 500 Ω is $1{\cdot}6^2 \times 500$ = 1,280 W, whereas at an 800-V point the current would be only 0·4 A and the loss in the generator only $0{\cdot}4^2 \times 500$ = 80 W! So when there are standing waves, the exact length of the line (in wavelengths) is obviously very important. If there are no standing waves, it does not matter if the line is a little longer or shorter or the wavelength is altered slightly; and that is one very good reason for matching the load to the line. The usual way of making sure that the load is right is by running a voltmeter or other indicator along the line and seeing that the reading is the same everywhere, except perhaps for a slight gradual change due to line loss.

Connecting the generator to a point on the line where the voltage and current are 200 V and 1·6 A respectively is equivalent to connecting it to a load of 200/1·6 = 125 Ω. The impedance of the line at all points can easily be derived from the voltage and current curves, as at the foot of Fig. 14.6. This curve indicates the impedance measured at any point when all the line to the left of that point is removed. The impedance depends, of course, not only on the distance along the line to the load but also on $R_0$ and the load impedance. From a consideration of the travelling waves it can be seen that at $\lambda/4$ intervals the phases of the currents and voltages are exactly the same as at the load, or exactly opposite; so that if the load is resistive the input resistance of the line is also resistive. At all other points the phases are such as to be equivalent to introducing reactance.

In our example we made the generator resistance, $r$, equal to $R_0$. If it were not, the situation would be more complicated still, because the reflected wave would not be completely absorbed by the generator, so part would be reflected back to the load, and so on, the reflected power being smaller on each successive journey, until finally becoming negligible. The standing waves would be the resultant of all these travelling waves. Even quite a long line settles down to a steady state in a fraction of a second, but it is a state that is generally undesirable because it means that much of the power is being dissipated in the line instead of being delivered to the load.

It should be noted that mismatching at the generator end does not affect the standing-wave ratio, but does affect the values of current and voltage attained.

Another result of making $r = R_0$ is that any reflection from the load or elsewhere inevitably impairs the generator-to-line matching. That is so even if the generator is connected to a point where the impedance curve coincides with the 500-Ω line in Fig. 14.6, because then the impedance is reactive. But if $r$ were, say, 2,000 Ω, so that it would be mismatched to the 500-Ω line during the brief moment following the start, there would be a chance that when the reflected wave arrived it would actually improve the matching, even to making it perfect (e.g., if connected at A or B in Fig. 14.6); but on the other hand it might make it still worse (if connected at C, D, or E).

It should be remembered that " perfect " matching is that which enables maximum power to be transferred; but in practice, as we have seen in another connection (p. 172) there may be good reasons for deliberately mismatching at the generator. To take figures we have already had, it may be considered better to deliver 320 W with a loss of 80 W than the maximum (500 W) with a loss of 500 W.

## The Quarter-Wave Transformer

An interesting result of the principles exemplified in Fig. 14.6 is that a generator of one impedance can be perfectly matched to a load of another by suitably choosing the points of connection. For instance, a generator with an internal resistance of 125 Ω would be matched to the 2,000 Ω load if connected at C, D or E. The line then behaves as a 1 : 4 transformer. Points can be selected giving (in this case) any ratio between 1 : 4 and 4 : 1, but except at the lettered points there is reactance to be tuned out.

It is not necessary to use the whole of a long line as a matching transformer; in fact, owing to the standing waves by which it operates it is generally undesirable to do so. We can see from Fig. 14.6 that the maximum ratio of transformation, combined with non-reactive impedance at both ends, is given by a section of line only quarter of a wavelength long. We also see that the mismatch ratio to the line is the same at both ends; in this particular case it is 1 : 4 (125 Ω to 500 Ω at the generator end and 500 Ω to 2,000 Ω at the load end), which, incidentally, is equal to the voltage ratio of the whole transformer. In general terms, if $R_1$ denotes the input resistance of the line (with load connected), $R_1$ is to $R_0$ as $R_0$ is to $R$; which means $R_1/R_0 = R_0/R$, or $R_0 = \sqrt{R_1R}$.

This formula enables us to find the characteristic resistance of the quarter-wave line needed to match two unequal impedances, $R_1$ and $R$. Suppose we wished to connect a 70-Ω load to a 280-Ω

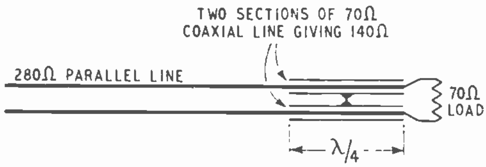



**Fig. 14.7—Example of a quarter-wave length of line being used as a matching transformer**

parallel-wire feeder without reflection at the junction. They could be matched by interposing a section of line $\lambda/4$ in length, spaced to give a $R_0$ of $\sqrt{(70 \times 280)} = 140\ \Omega$. Parallel wires would have to be excessively close (Fig. 14.7*a*), but the problem could be solved by using a length of 70-Ω coaxial cable for each limb and joining the metal sheaths together as in Fig. 14.7, putting the 70-Ω impedances in series across the ends of the 280-Ω line.

## Fully Resonant Lines

Going now to extremes of mismatch, it is of interest to inquire what happens when the " load " resistance is either infinite or zero; in other words, when the end of the line is open-circuited or short-circuited. Take the open circuit first. If this were done to our original 500-Ω example, the current at the end would obviously be nil, and the voltage would rise to 1,000—double its amount across the matched load. This condition would be duplicated by the standing waves at points A and B in Fig. 14.6; while at C, D, and E the voltage would be nil and the current 2 A. The impedance curve would consequently fluctuate between zero and infinity.

With a short-circuited line, there could be no volts across the end, but the current would be 2 A; in fact, exactly as at E with the open line. A shorted line, then, is the same as an open line shifted quarter of a wavelength along. Reflection in both cases is complete, because there is no load resistance to absorb any of the power.

If the generator resistance is very large or very small, *nearly* all the reflected wave will itself be reflected back, and so on, so that if the line is of such a length that the voltage and current maximum points coincide with every reflection, the voltages and currents will build up to high values at those maximum points—dangerously high with a powerful sender. This reminds us of the behaviour of a high-$Q$ resonant circuit (p. 102). The maximum current or voltage points are called *antinodes*, and the points where there is *no* current or voltage are *nodes*.

When the length of a short-circuited or open-circuited line is a whole number of quarter-wavelengths, the input impedance is approximately zero or infinity. An *odd* number of quarter-wavelengths gives opposites at the ends—infinite resistance if the other end is shorted, and vice versa. An *even* number of quarter-wavelengths gives the same at each end.

In between, as there is now no load resistance, the impedance is a pure reactance. At each side of a node or antinode there are opposite reactances—inductive and capacitive. If it is a current node, the reactance at a short distance each side is very large; if a voltage node, very low. Fig. 14.8 shows how it varies. It is clear from this that a short length of line—less than quarter of a wavelength—can be used to provide any value of inductance or capaci-

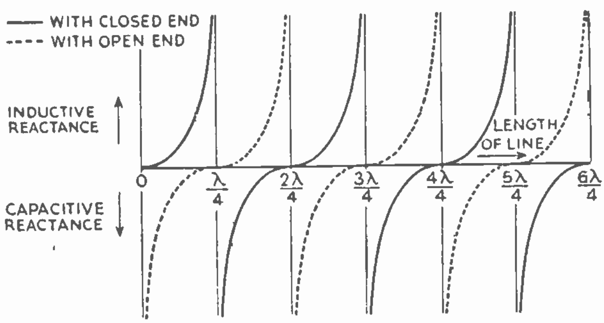



**Fig. 14.8—Showing how the reactances of short-circuited and open-circuited lines vary with their length**

tance. For very short wavelengths, this form is generally more convenient than the usual inductor or capacitor, and, under the name of a *stub* is often used for balancing out undesired reactance.

## Lines as Tuned Circuits

A quarter-wave line shorted at one end is, as we have just seen, a high—almost infinite—resistance. But only at the frequency that makes it quarter of a wavelength (or an odd multiple of $\lambda/4$). This resistance is closely analogous to the dynamic resistance of a parallel resonant circuit; and in fact a line can be used as a tuned circuit. At wavelengths less than about 2 metres it is generally more efficient and easily constructed than the conventional sort.

The lower the loss resistance of the wire, the higher the dynamic resistance across the open ends; and to match lower impedances all that is necessary is to tap it down, just as if it were any other sort of tuned circuit.

Fig. 14.9 is an example of a conventional tuning circuit, such as might be used in a receiver, and *b* is the coaxial equivalent. A coaxial feeder is also used to connect the aerial, and being normally about 80Ω it is tapped low down, near the earthed end. The

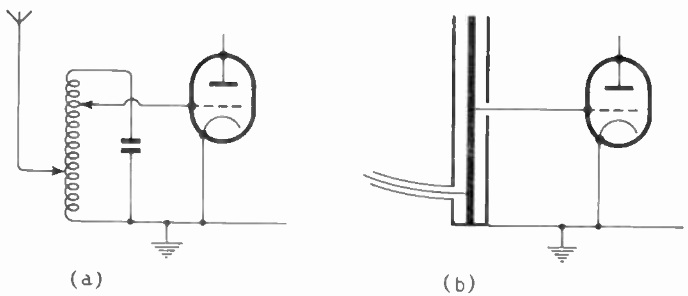



**Fig. 14.9—(*a*) Conventional aerial input circuit using "lumped" components, and (*b*) distributed equivalent using a quarter-wave coaxial line**

impedance at the top end is normally many thousands of ohms, and may be too high for the input of a valve, which is quite low at very high frequencies.

At still higher frequencies, with waves only a few centimetres long, the inner conductor becomes unnecessary, and power can be transmitted along hollow pipes (called *waveguides*) and tuned by cavities. But that is a subject in itself.

CHAPTER 15

# Radiation and Aerials

### BRIDGING SPACE

IN THE FIRST CHAPTER of this book the processes of radio communication were very briefly traced from start to finish; since then we have considered the sender in some detail. Before going on to the receiver we should know something about how the space in between is bridged.

We have just seen how electrical power can be transmitted from one place to another at the not inconsiderable speed of nearly 300,000 kilometres (186,240 miles) per second, in the form of waves along a line consisting of two closely spaced conductors. A negro who was asked to explain how it was possible to signal by telegraph from New York to New Orleans replied by pointing out that when you tread on the tail of a dog the bark comes out at its head. A telegraph was like a very long dog, with the tail in New York and the head in New Orleans. " But what about wireless? " he was asked. " Same thing, 'cept the dawg am imagin'ry ". This reply, though ingenious, is not perhaps altogether satisfying. How is it possible to transmit the waves without any conductors to carry the currents to and fro?

The significant thing about the section on how waves are transmitted along a line (p. 178) is that conduction was hardly mentioned; the important properties of a line are its distributed capacitance and inductance. In other words, transmission of waves along a line is primarily a matter of oscillating electric and magnetic fields, located in the space between and around the two conductors. These conductors are not required to carry current from one end to the other, any more than waves on a pond carry the water from one side to the other; all that the electrons in the conductors do is oscillate a very short distance locally.

### THE QUARTER-WAVE RESONATOR AGAIN

Near the end of the same chapter we found that a line only quarter of a wavelength long and open at the end, as in Fig. 15.1*a*, is equivalent from the generator's point of view to a series resonant circuit (*b*); that is to say, the only impedance the generator faces is the incidental resistance of the circuit. We saw on p. 102 that if this resistance is small a very little generator voltage can cause a very large current to flow and build up large voltages across C and L. This remarkable behaviour of C and L is due to the electric field concentrated mainly between the plates of C and the magnetic field concentrated mainly inside and close around the turns of L. The main difference between *a* and *b* is that in *a* the inductance and capacitance—and therefore the fields—are mixed up together and uniformly distributed along the length. The distance of the dotted lines from the conductors in Fig. 15.1 indicates the distribution of current and voltage along them (p. 181). The dotted lines can be regarded as marking the limits of the alternations of current and voltage, just as the dotted lines in Fig. 1.3 marked the limits of vibration of a stretched string.

An essential feature of a transmission line, even such a short one as this, is that the distance between the conductors is small compared with a wavelength—or even quarter of a wavelength. The reason

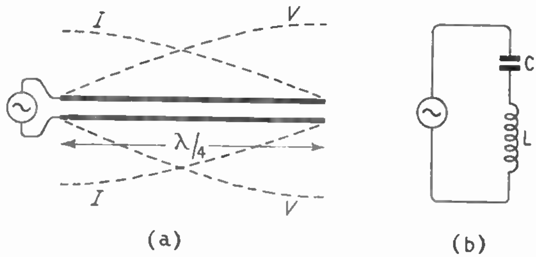



**Fig. 15.1—(*a*) Above—A quarter-wave length line, showing the distribution of current and voltage, and (*b*) its circuit equivalent**

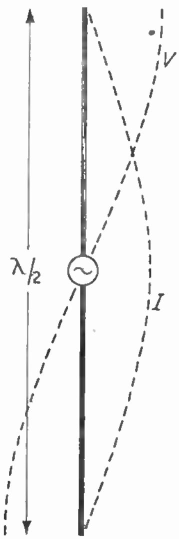



**Fig. 15.2—Left—The quarter-wave line of Fig. 15.1 opened fully out**

is that the explanation of its action is based on neglecting the time taken for the fields to establish themselves when a current flows or a p.d. is created. Provided nearly all of each field is close to the conductors, this assumption is fair enough. But when a current is switched on, the resulting magnetic field cannot come into existence instantaneously everywhere; just as a current takes time to travel along the line, so its effect in the form of a magnetic field takes time to spread out. And when the current is cut off, the effect of the remoter parts of the magnetic field collapsing takes time to get back to the conductor. At points in the field one eighth of a wavelength from the conductor of the current, the combined delay is quarter of a wavelength, or quarter of a cycle, which upsets the whole idea of self-inductance. In fact, this phase delay between a change of current and the self-induced e.m.f. calls for a driving e.m.f. in phase with the current, just as if there were a resistance in the circuit—so far as that piece of field is concerned.

Suppose that in order to study this we open out the two conductors of our quarter-wave parallel line to the fullest possible extent, as in Fig. 15.2. Corresponding to our lumped $L$ and $C$ approximation to the original parallel line we could draw Fig. 15.3. The opening out greatly increases the area through which the magnetic field can pass, so increasing $L$. The lines of electric field are on the average

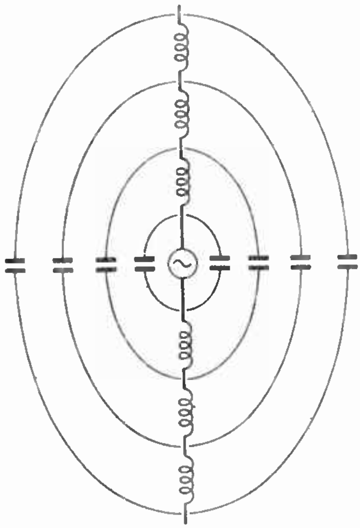


Fig. 15.3—Approximate representation of the distributed inductance and capacitance in the half-wavelength system of Fig. 15.2

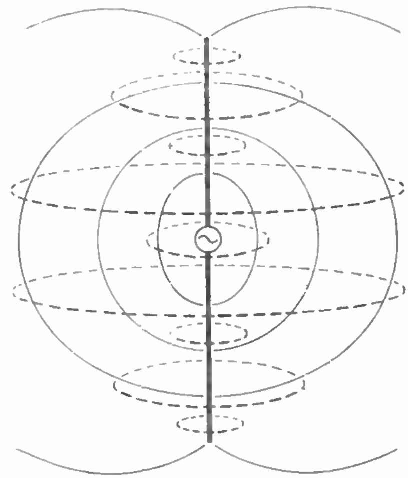


Fig. 15.4—The distributed inductance and capacitance in Fig. 15.2 are better indicated by drawing a few lines of force. The full lines are electric and the broken lines magnetic

much longer, so $C$ is decreased. Again, these two effects almost exactly cancel out, so the line is still nearly resonant to the same generator frequency. It can be brought exactly into tune by shortening it a very few per cent, which will be assumed to be done even though (as in Fig. 15.2) the total length is for simplicity marked $\lambda/2$. So the distribution of current and voltage along the conductors is much the same as before. Looking at Fig. 15.3 we would say that the current was greatest at the middle because it has to charge all the capacitance in parallel, whereas near the ends there is hardly any to charge.

Fig. 15.4 shows some typical electric and magnetic lines of force. At any one radius from the generator they would trace out a sort of terrestrial globe, on which the magnetic lines were the parallels of latitude and the electric lines the meridians of longitude. This is helpful as a picture, but we must remember once again that field lines are as imaginary as lines of latitude and longitude.

## A Rope Trick

It is clear that quite an appreciable portion of each field is now far enough away to be subject to a phase delay in its reaction on the conductors. To the generator the effect is as if some resistance had been inserted in the circuit. If so, that means the generator is now supplying energy. Where does it go?

When we were young we must all have taken hold of one end of a rope and waved it up and down. If this is done slowly, nothing much happens except that the part of the rope nearest us moves up and down in time with our hand, as in Fig. 15.5*a*. But if we shake it rapidly, the rope forms itself into waves which run along it to the far end (*b*). This is quite an instructive analogy. The up-and-down movement of the hand represents the alternation of

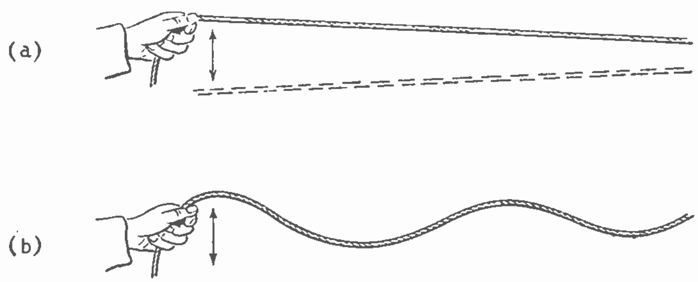



**Fig. 15.5—Well-known results of waving a rope up and down, (*a*) slowly, (*b*) quickly**

charge in the $\lambda/2$ rod caused by the generator. When the current is upwards it charges the upper half positively and the lower half negatively, resulting in an electric field downwards between the two. Half a cycle later the pattern is reversed. The up-and-down displacement of the rope from its middle position represents this alternating electric field. The displacement travels outward from the hand as each bit of rope passes it on to the next. If the time of each cycle of hand movement is long compared with the time for the movement to pass along the rope, all the rope for at least quite a distance appears to move as a whole, " in phase ". Because the electric field has to spread out in all directions, and not just in one direction like the rope, its amplitude falls off much more rapidly with distance from the source than does that of the rope.

When the hand is moved so fast that the rope displacement has not gone far before the hand reverses, the displacements become progressively out of phase with one another, and this is why travelling waves are formed, as in Fig. 15.5*b*. The same applies to the alternations of electric field; they form a wavelike pattern which travels outwards, just like the electric field between the conductors of a long line in that direction, except for a falling off in amplitude because it is not going out only along one radius but all around the " equator ".

## Electromagnetic Waves

A very fair question at this stage would be: what keeps it going? The rope wave would not keep going unless the rope had not only displacement but *inertia*. The fact that we have already noted a resemblance between inertia and inductance (p. 68) may remind us that we seem to have forgotten about the magnetic field But of course that too is spreading out, and in fact is essential to the wave motion. When we did the experiments with iron filings around a wire carrying current (p. 60) we said that the cause of the magnetic field was the electric current in the wire. But we know that an electric current consists of electric charges in motion, and charges are surrounded by an electric field. So the electric current, which generates the magnetic field, is really a moving electric field. We know, too, that a moving magnetic field generates an e.m.f. We tend to think, perhaps, of an e.m.f. only in a circuit, but it is really an electric field, whether there happen to be any circuits or conductors about or not. Putting all this concisely:

Moving electric field causes magnetic field
Moving magnetic field causes electric field

Without going into the rather advanced mathematics of the thing, one might guess the possibility of the two fields in motion keeping one another going. Such a guess would be a good one, for they can and do. The speed they have to keep up in order to be mutually sustaining is the same as when they are guided by the transmission line. The combination of fast-moving fields is known as an *electromagnetic wave*. The union of the two fields is so complete that each depends entirely on the other and would disappear if it were destroyed. The speed of the wave is the same as that of light—a significant fact, which led to the discovery that light waves are the same as those used for radio communication, except for their much higher frequency.

## RADIATION

An important thing to note about the rope trick is that the rope itself does not travel away from us—only a wavy pattern. But there is another thing: some of the energy we put into it travels to the far end, for it is capable of setting a support there into vibration. Electromagnetic waves are similar in this respect too; they convey energy away from the source. The outward movement of the waves from the source is called *radiation*. All variations in current cause radiation, but when the field energy is mainly localized, say between the conductors of a transmission line or the plates of a capacitor, the amount is much less than when it is spread out as in Fig. 15.4. Radiation also increases with the rate of variation; the 50 c/s variations of ordinary a.c. power systems cause negligible radiation, but 50 Mc/s is quite a different matter. So although we have spoken of the waveforms of low-frequency a.c., this should not be taken as suggesting that they are waves in the sense used in this chapter—electromagnetic waves.

We saw in connection with the lumped tuned circuit of Fig. 12.2 that the electric and magnetic fields, responsible for the capacitance and inductance, are quarter of a cycle out of phase with each other. Even in the most spread-out circuit, such as our half-wavelength rod, the majority of the energy of the surrounding fields is able to return to the circuit twice each cycle, to be manifested there as its distributed capacitance and inductance. This majority portion of each field, mainly near the source, is similarly out of phase with the other. It is distinguished by the term *induction field*, and it falls off inversely as the square of the distance, which means that most of it is quite close to the source.

In order to convey energy right away from the source, however, there have to be components of fields in phase with one another, and it is this portion, called the *radiation field*, in which we are interested now. It falls off inversely as the distance, which means that although it starts as a minority it spreads much more, and far away from the source it is very much in the majority.

## POLARIZATION

Fig. 15.4 shows that the electric and magnetic fields are at right angles to one another and to the direction in which they radiate outwards as electromagnetic waves. Imagine we are looking

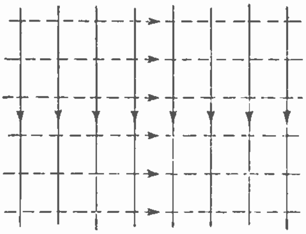


**Fig. 15.6—Relative direction of electric (full line) and magnetic (broken line) fields in an electromagnetic wave advancing towards the viewer**

towards the vertical source and the oncoming waves. Then the electric field alternates up and down and the magnetic field side to side, as shown in Fig. 15.6. This pattern can be imagined as due to a transmission line consisting of parallel strips top and bottom. At the instant depicted, the potential is positive at the top, so the electric field is downwards from it. This p.d. would be driving current towards us through the top conductor and away from us through the bottom one. The " corkscrew rule " (Fig. 4.2) indicates a clockwise magnetic field round the bottom current and anticlockwise round the top one; both combine to give a magnetic field from left to right, as shown.

If for any reason we prefer the directions of the two fields to be interchanged—electric horizontal and magnetic vertical—that can easily be arranged, by turning the radiator horizontal. Similarly for any other angle. The direction of the field in an electromagnetic wave is known as its *polarization*. Since there are two fields, mutually at right angles, there would be ambiguity if it had not been agreed to specify the polarization of a wave as that of its *electric* field. This custom has the advantage that the angle of polarization is the same as that of the radiator.

Because radiation starts off with a certain polarization it does not follow that it will arrive with the same. If it has travelled far, and especially if on its way it has been reflected from earth or sky, it is almost certain to have become disarranged. Another thing is that polarization is not necessarily confined to a single angle. After reflection, radiation often has both vertical and horizontal components, and the maximum may be at some odd angle. If it is equal at all angles, the wave is said to be circularly polarized, because at any point it consists of fields of constant strength rotating continuously at the rate of one revolution per cycle. Sometimes for special purposes the radiator is designed to make the waves circularly polarized from the start.

## AERIALS

Having considered the process of radiation, we are due to turn attention to the radiator. If intended for that purpose it is called an *aerial* or antenna. The form we developed from a quarter-wavelength transmission line, known as a half-wave *dipole*, is perhaps the most important both from the theoretical and (since the arrival of television) practical point of view.

We have already noted that this straight metal wire or rod is a tuned circuit, and that its sound-wave analogue is a stretched cord as in Fig. 1.3. The fundamental frequency at which it resonates is the one that makes half a wavelength equal to (or, to be precise, a few per cent greater than) the length of the rod.

There are certain higher frequencies at which the same length of aerial can be set into vibration at harmonic frequencies; but because of complications due to opposing currents in different parts, this condition is seldom used. It is not essential for an aerial to be in resonance at all, but as the radiation is proportional to the current, which is a maximum at resonance, aerials are practically always worked in that condition.

Just as with the tuned circuits in Chapter 8, the originator of the current has been shown as a generator in series; actually at the centre of the dipole. But in practice a tuned circuit is often "excited" by an external field, as for example a magnetic field in inductive coupling. The same applies to our metal rod. If it is exposed to electromagnetic waves, the magnetic field pattern sweeping across it induces an e.m.f. in it. This statement must not be taken to mean that if there were no metal there would be no e.m.f. It is the movement of the magnetic field that induces the electric field, as we have seen; and that electric field constitutes an alternating e.m.f. in space.

It is unimportant whether the e.m.f. acting on the aerial is credited to the electric field or the magnetic field—they are simply two different descriptions of a single action—but it is usually convenient to work in terms of the electric component of the waves. Near a radiator it is reckoned in volts per metre, or microvolts per metre if far off. If a vertical wire two metres long is exposed to vertically polarized radiation of 100 $\mu$V per metre, then an e.m.f. of 200 $\mu$V will be induced in it. But if either the wire or the radiation is horizontal, no e.m.f. is induced. In general, e.m.f. is proportional to the cosine of the angle between the wire and the direction of polarization.

It is in this way that an aerial is used at the receiving end. The only difference in principle between a sending and a receiving aerial is that one is excited by a r.f. generator connected to it, and the other by the electromagnetic waves created by the former. The factors that make for most effective radiation in any particular direction are identical with those for the strongest reception from that direction, so whatever is said about aerials as radiators can readily be adapted from the reception standpoint.

## Radiation Resistance

We have noted that the dipole aerial, like the quarter-wave line from which we developed it, has inductance and capacitance. What about resistance? There is of course the resistance of the metal of which it is made, together with the various sources of loss resistance described on p. 113. But we have seen that the waves radiated from a sending aerial carry away energy from it, otherwise they would not be able to make currents flow in receiving aerials. It is convenient to express any departure of energy from a circuit in terms of resistance (p. 48). But whereas the other kinds of aerial resistance are wasteful, this kind, called *radiation resistance*, measures the ability of an aerial to do its job.

The radiation resistance of an ordinary tuning circuit is generally very small compared with its other resistance (radiation, after all, is not its job!); at the other extreme the loss resistance of a dipole, made perhaps of copper tubing, is small compared with its radiation resistance, which at resonance is about 70$\Omega$ if it is measured at the centre where the current is greatest. If therefore a current of, say, 2A can be generated there, the power radiated, calculated in the usual way as $I^2R$, is $2^2 \times 70 = 280$ W. If the radiation resistance were measured at some other point along the dipole it would be greater, but this would not mean greater radiation, because the current there would be less.

A receiving aerial too has radiation resistance, and some of the energy imparted to it by the waves is re-radiated—a fact of great importance in the design of receiving aerials, as we shall see.

## Directional Characteristics

We based our study of radiation from a dipole on the production of waves along a rope. But of course there is no reason for electromagnetic radiation to be confined along a single line like this; it goes out equally in all directions around the "equator". So the wave pattern in this plane can be likened to the ripples created on the surface of a pond by dropping in a stone. A skeleton of it could be produced by having ropes radiating from one's hand like the spokes of a horizontal wheel. But what about other angles, upwards and downwards?

Is it too much to imagine oneself in outer space, where ropes have no weight, at the centre of great numbers of them stretching in every direction? Those immediately above and below receive

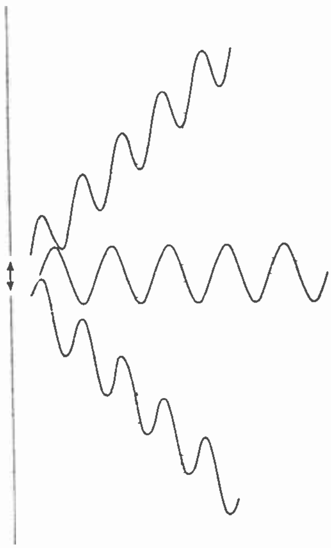


**Fig. 15.7—If the rope trick of Fig. 15.5 (*b*) were performed with ropes radiating in all directions, the results would look something like this**

no wavy pattern; only lengthwise stretching and relaxing. At intermediate angles there would be a combination of the two effects, as illustrated by the few samples in Fig. 15.7. By filling in as many others as possible, we can get an idea of the whole pattern of radiation: most strongly around the equator, and none at all along the axis. The pattern would be still more completely represented by the effect on a vast light elastic jelly of making a point at its centre vibrate up and down. The relative strength of radiation in different directions is however usually represented by what is called a *polar diagram*. This is a curve drawn around a centre, such that the distance from the centre to the curve in any direction represents the radiation in that direction. Ideally, a polar diagram would be a three-dimensional surface, but on paper one usually has to be content with the curves, which are sections of the surface at particular planes.

Fig. 15.8*a* is the horizontal polar diagram of a vertical dipole. showing that radiation (or reception) is equal in all horizontal directions. At *b* is the vertical polar diagram, showing how the radiation varies from maximum at right angles to the dipole to nil end-on. So if a dipole is placed horizontally it can be used to cut out undesired reception from two opposite directions. As regards reception, these diagrams assume, of course, that the radiation arriving has the correct polarization. A horizontal

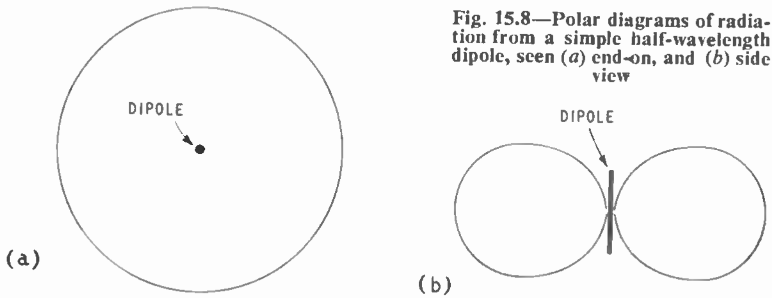



Fig. 15.8—Polar diagrams of radiation from a simple half-wavelength dipole, seen (*a*) end-on, and (*b*) side view

dipole does not respond to vertically polarized waves from any direction.

## Reflectors and Directors

Any metal rod is a tuned circuit which will have maximum currents set going in it by waves twice its own length. These currents radiate waves, just as if they were caused by a directly connected generator. If such a rod is placed close to a radiating dipole, the re-radiation from it is strong enough to modify the primary radiation very considerably, subtracting from it in some directions and adding to it in others. When the parasitic aerial (as it is called) is about the same length as the driven one, and quarter of a wavelength away from it, the phase of the re-radiation is such as to reverse much of the radiation beyond the rod, which is therefore called a *reflector*—see Fig. 15.9. The same principle applies to a receiving aerial, and reflectors are frequently used; not so much to strengthen

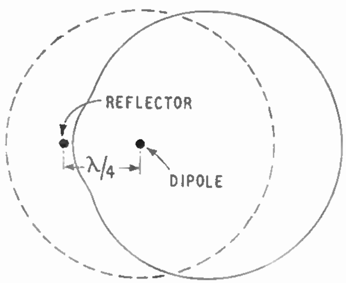



Fig. 15.9—Showing how the equal-all-round radiation of Fig. 15.8 (*a*) (repeated here dotted) is modified by an unconnected dipole placed quarter of a wavelength away from the first, and of such a length as to have a 45° phase angle

the wanted signal (though that is usually helpful) but to cut out interference from the oppositite direction.

The effect just described depends on the phase relationship between the primary and secondary radiation. When considering tuned circuits we may have noticed (from Fig. 8.11, for example) that a small departure from exact tuning causes a large phase change between voltage and current. It is the same with dipoles; a small increase or decrease in length alters the phase and therefore the polar diagrams. The phase can also be varied by the spacing between the rods. If the parasitic rod is shorter than the driven one, and only about $0{\cdot}1\lambda$ away from it, the effect is opposite to that shown in Fig. 15.9, so such a rod is called a *director*. The use of reflectors or directors affects the resistance of the driven dipole as a load, usually lowering it.

The one-way tendency can be increased by using both a reflector and a director, one on each side; or even several directors, as in the

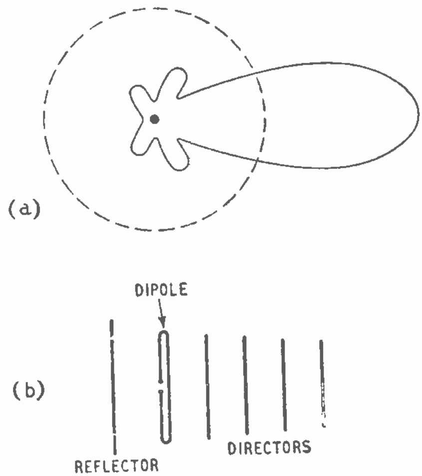



**Fig. 15.10—(*a*) Polar diagram of an aerial system consisting of two parallel units each made up of a dipole, one reflector and four directors (*b*). The diagram for a single dipole is again shown dotted for comparison**

array shown in Fig. 15.10*b*, which is called a Yagi aerial. Its polar diagram (*a*) shows a very considerable advantage over that of the single dipole.

But this is only the beginning of what can be done. The number of driven (or receiver-connected) dipoles can be increased, by placing them side by side, or in stacks one above the other, or both, and they can also be provided with parasitic elements. The underlying principle of all is the same: to arrange them so that the separate radiations add up in phase in the desired directions and cancel out in the undesired directions. It is necessary to consider not only the distribution horizontally but also the angle of elevation above the horizon. By using a sufficient number of dipoles, suitably arranged, the radiation can be concentrated into a narrow beam in any desired direction.

## Aerial Gain

Concentrating the radiation in the direction or directions where it will be most useful increases the apparent power of the sender. The number of times the intensity of radiated power in any particular direction is greater than it would be if the same amount of power were radiated equally in all directions is called the *power gain* (or just " gain ") of the aerial in that direction. For example, if the radiated power were concentrated uniformly into one tenth of the whole surrounding sphere the gain would be 10. Within that beam, the radiation would be the same as if ten times the power were being radiated equally in all directions. The effective power of a sender is therefore equal to the actual radiated power multiplied by the aerial gain; this figure is known as the *effective radiated power* (e.r.p.).

In broadcasting, the aim is to increase the e.r.p. by directing as much of the radiation as possible horizontally, rather than upwards where it would be wasted. The apparent power of a sender, so far as any particular receiver is concerned, can of course be further increased by using a receiving aerial with gain in the direction of the sender.

A simple dipole, by favouring the equatorial as compared with polar directions, has a gain of 1·64, and the gains of other aerials are sometimes quoted relative to this rather than to the theoretical spherical radiator.

## Choice of Frequency

A feature of the dipole aerial and its derivations which we have considered so far is that their dimensions are strictly related to the wavelength and therefore to the frequency (p. 22). What decides the choice of frequency: a convenient aerial length, or something else?

The shorter the wavelength and higher the frequency, the smaller and cheaper the aerial and the more practicable it is to direct its radiation and increase its gain. For example, as long ago as 1931 a radio link was established across the English Channel on a wavelength of 17 centimetres, a half-wave aerial for which is only about 3½ inches long! Why, then, erect at vast expense aerials hundreds of feet high?

The main reason is that such waves behave very much like the still shorter waves of light, in not being able to reach anywhere beyond the direct line of " view " from the sending aerial. So either the aerials have to be put at the tops of high towers (in which case their cheapness disappears) or the range is restricted to a few miles. Wavelengths shorter than 10 metres are used for television, not because the shortness of wavelength in itself has any particular merit for that purpose, but because it corresponds to a very high frequency, which, as we shall see in Chapter 18, is necessary for a carrier wave that has to carry modulation frequencies of up to several Mc/s. They are also used for short-distance communication such as police cars, radiotelephone links between islands and mainland, and other specialized short-range purposes.

## Influence of the Atmosphere

At various heights between 60 and 300 miles above the earth, where the atmosphere is very rarified, there exist conditions which cause wireless waves to bend, rather as light waves bend when they pass from air into water. The very short (high-frequency) waves we have just been considering normally pass right through these layers and off into space, as shown in Fig. 15.11*a*. Where they are close to the ground, they are rapidly absorbed, as indicated by the shortness of the ground-level arrow. So their effective range is limited.

As the frequency is reduced below about 30 Mc/s—the exact dividing line depends on time of day and year, sun-spot activity,

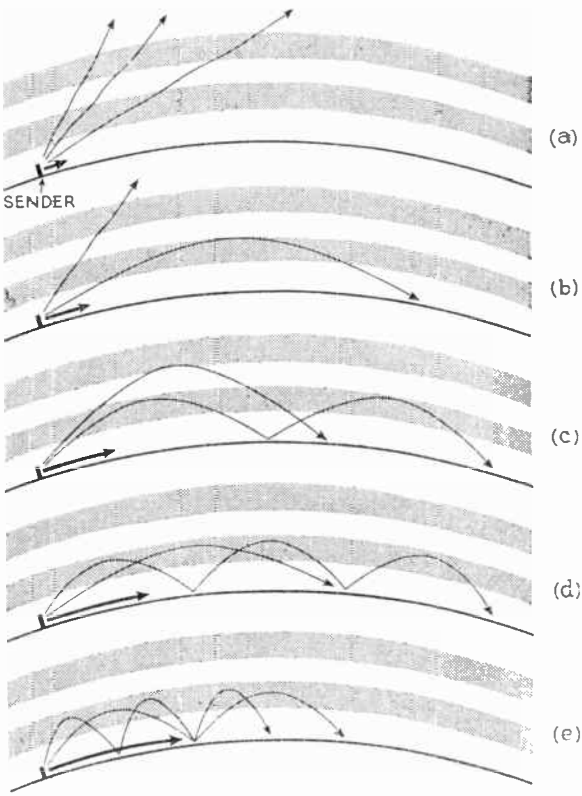



**Fig. 15.11—Showing (not to scale) the relative ranges of ground wave and reflected wave from very high frequencies (*a*) to very low (*e*)**

and other conditions—the range of the wave-front travelling along the surface of the earth (and therefore termed the ground wave) increases slowly, while the sky wave is bent so much that it returns to earth at a very great distance, generally several thousands of miles (Fig. 15.11*b*). Between the maximum range of the ground wave and the minimum range of the reflected wave there is an extensive gap, called the skip distance, which no appreciable radiation reaches.

As the frequency is further reduced this gap narrows, and earth reflections may cause the journey to be done in several hops (*c*). Since the distances at which the sky waves return to earth vary according to time and other conditions as mentioned, it is rather a complicated business deciding which frequency to adopt at any given moment to reach a certain distance. But a vast amount of data has been accumulated on this and a judicious choice enables reliable communication to be maintained at nearly all times. As waves usually arrive at the receiver by more than one path simultaneously, and tend to interfere with one another, fading and distortion are general unless elaborate methods are adopted for sorting them out. At a certain frequency, of the order of 2 Mc/s, the ground wave and reflected wave begin to overlap at night, while during daylight the reflected wave is more or less absent. Over the ranges at which there is overlap the two waves tend to interfere and cause fading and distortion, as they do with more than one reflected wave.

Finally, the range of the ground wave increases and becomes less affected by daylight or darkness, so that frequencies of 50–20 kc/s have a range of thousands of miles and are not at the mercy of various effects that make long-distance short-wave communication unreliable. For this reason they were originally selected as the only wavelengths suitable for long ranges, but now only a small fraction of long distance communication is borne by very long waves. The disadvantages of the latter are (1) the enormous size of aerial needed to radiate them; (2) the low efficiency of radiation even with a large and costly aerial system; (3) the higher power needed to cover long ranges, due largely to (4) the great intensity of atmospherics—interference caused by thunderstorms and other atmospheric electrical phenomena; and (5) the very limited number of stations that can work without interfering with one another, because the waveband is so narrow in terms of frequency—which is what matters; see Chapter 18.

## Earthed Aerials

We see, then, that although for convenience of dipole aerial size one would choose frequencies of a few hundred Mc/s, these are not suitable for all purposes. For distances of more than about 50 miles it is usually necessary to use longer waves (lower frequencies). Until now we have assumed that the aerial is suspended well clear of the earth, but this is not convenient with a dipole for wavelengths of many metres, and the length of the aerial itself may be unwieldy.

Marconi used the ground itself as the lower half of a vertical dipole; the remaining half sticking up out of it is therefore often called a quarter-wave Marconi aerial. The current and voltage distribution are shown in Fig. 15.12, which is the same as the top half of Fig. 15.2.

For the earth to be a perfect substitute for the lower half of a dipole it must be a perfect conductor, which of course it never is. Sea water is a better approximation. To overcome the loss due to

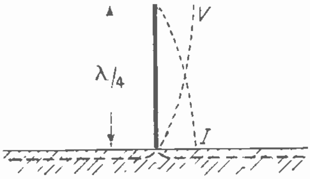



**Fig. 15.12—The earth can be used as one half of a radiator, leaving a quarter-wavelength wire above earth**

imperfectly conducting earth, it is a common practice to connect the lower end of the aerial to a radial system of copper wires. An insulated set of wires stretched just above the ground is known as a *counterpoise*. Alternatively, to avoid tripping up people or animals, the wires are often buried just below the surface, as shown dotted in Fig. 15.12.

## Feeding the Aerial

One advantage of the Marconi aerial is that it places the middle of the dipole at ground level, which is helpful for coupling it to the sender or receiver. Even so, it is not always convenient to have the output circuit of a sender close enough to the foot of the aerial to couple it directly, and it is seldom convenient to have either sender or receiver at the centre of a suspended dipole, so there is a need for some means of linking the aerial with the ground equipment. The previous chapter showed how the problem can be conveniently and efficiently solved by means of a transmission line designed to match the impedance of the aerial. The object of matching, we recall, is to avoid reflections, which prevent the maximum transfer of power to or from the aerial and cause excessive line currents and voltages at a sender, or signal echo effects at a receiver.

We have noted that an ordinary centre-fed dipole has a radiation resistance of about 70 Ω. At resonance, which is the normal working condition, inductance and capacitance cancel one another out. Other kinds of resistance are usually negligible, so the dipole can be treated as a 70 Ω load. The simplest scheme is to use a line having a characteristic resistance, $R_0$, of about 70 Ω, as in Fig. 15.13. Here the dipole can be regarded as a radiating extension of the line. At the other end, the line should be matched to the sender output circuit (p. 171).

Fig. 14.4 shows that 70 Ω is an awkward $R_0$ for a parallel-wire line, for the wires have to be almost touching; it is also a long way from the lowest-loss value. On the other hand it is about optimum for coaxial cable. Moreover the coaxial type radiates less and picks up less interference. But it has the disadvantage of being unsymmetrical; that is to say, unbalanced with respect to earth. The outer conductor is in fact itself normally earthed, so although

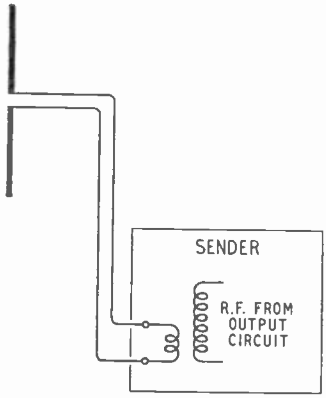



**Fig. 15.13 (left)—A simple dipole connected to a sender (or receiver) by means of a matched-impedance transmission line**

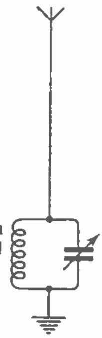


**Fig. 15.14—A common method of tuning an aerial to a desired wavelength**

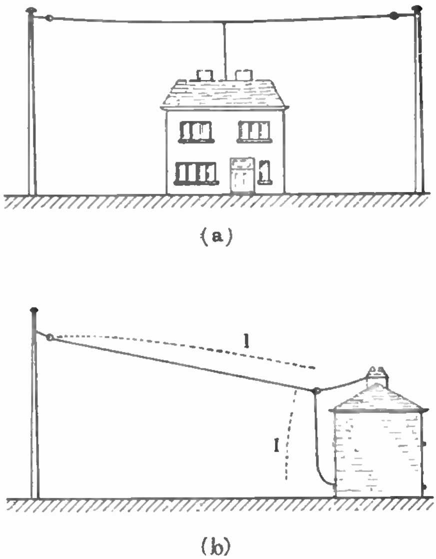
the line can easily be coupled to the ground equipment (which is usually earthed) it is not strictly correct to connect it to a dipole, which is symmetrical. Where maximum efficiency is important, as in a sender, a balance-to-unbalance transformer is used; for receiving, the unbalance is usually tolerated. Alternatively, parallel-wire line of relatively high $R_0$ can be matched to the dipole by means of a transformer such as the one shown in Fig. 14.7. This is only one of a great variety of matching and balancing devices.

## Tuning

A dipole aerial is self-tuned, by virtue of its chosen length, to a particular frequency. The thinner the wire, the sharper is its resonance peak. But an aerial is often required to work over a very wide range of frequency, especially for receiving. Instead of varying the length of the aerial wire, it is usual to augment its distributed inductance and capacitance by means of variable capacitors or inductors, as for example in Fig. 15.14. These tuning components reduce the resonant length of the aerial for a given frequency, which may be very desirable; for example, even a quarter-wave self-tuned aerial for 30 kc/s would have to be 8,000 feet high! It is quite usual for a considerable proportion of the reactance of an aerial system, especially for medium and long waves, to be in concentrated form; but it must be realized that this is at the expense of radiation or reception.

Where height is limited, by restrictions imposed by aviation or by the resources of the owner, it is possible to increase radiation

**Fig. 15.15 — To increase the radiation from a vertical aerial without increasing its height, a horizontal top is commonly added, forming (*a*) the T or (*b*) the inverted L type. Note the more uniform current distribution in the vertical part (*b*)**




moderately by adding a horizontal portion, as in the familiar T or inverted L aerials, Fig. 15.15. The effect of the horizontal extension is to localize the bulk of the capacitance in itself so that the current in the vertical part, instead of tailing off to zero, remains at nearly the maximum and therefore radiates more. The top portion radiates too, but with a different polarization, so the addition due to it may not be very noticeable in a vertical receiving aerial.

## EFFECTIVE HEIGHT

In general, the greater the radiation resistance of an aerial the more effective it will be for receiving as well as for sending. But a more useful piece of information about an aerial when it is used for receiving is what is called its *effective height*. It is the height which, when multiplied by the electric field strength, gives the e.m.f. generated in it. For example, if the wave from a sender has a strength of 2 mV per metre at a receiving station, and the effective height of the aerial is 5 metres, the signal voltage generated in the aerial will be 10 mV. The effective height is less than the physical height. It can be calculated for a few simple shapes, such as a vertical wire rising from flat open ground, but usually has to be measured, or roughly estimated by experience.

## MICROWAVE AERIALS

What happens if, instead of making the wavelength longer than is most handy for dipoles, it is made much shorter—say a few centimetres? The radiation from a single dipole is of course limited by its size, and to bring the e.r.p. (p. 198) up to a useful figure a relatively large reflector is needed. Instead of an enormous number of small dipoles, it is made of continuous metal. The currents induced in this by the driven dipole cause reflection, but in a more controllable pattern, for the metal can be formed into a parabolic reflector, like a large headlight or bowl fire. In this way the radiation or reception can be confined very closely in the required direction.

Just as at these frequencies the two-conductor transmission line gives place to a waveguide (p. 186), the dipole gives place to a horn, which is an expanding end to the waveguide, arranged to direct the radio power on to the reflector.

Such frequencies and aerials are commonly used for radar, where narrow beaming is essential, and for point-to-point transmission of television signals.

## Inductor Aerials

In all the aerials considered, the circuit between the terminals is completed by the capacitance from one arm to the other or to earth. Although they have inductance, they can be regarded roughly as opened-out capacitors. It is possible instead to use an opened-out inductor, in which of course there is a wire path all the way. This type is known as a frame or loop aerial, and because its loss resistance is usually larger than the radiation resistance it is seldom used as a radiator, but mainly for receiving, especially for portable sets and direction finding.

We noted that reception by an open aerial is usually calculated from the strength of the electric field, although the same answer would be given by the magnetic field. The opposite applies to

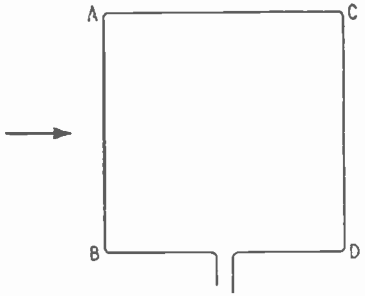



**Fig. 15.16—The effectiveness of a receiving aerial consisting of a loop or coil of wire depends on the direction from which the waves are arriving**

loop aerials. Taking as example a single square turn of wire as in Fig. 15.16, one can see that waves arriving broadside on, whatever their polarization, induce no net e.m.f. in the aerial. A vertical electric field would cause e.m.fs in the portions AB and CD, but these would be exactly in opposition around the loop. It is easier, however, to see that the magnetic field, being in the same plane as the loop, would not link with it at all, and that is sufficient to guarantee no signal pick-up.

From any direction in the plane of the paper—say that indicated by the arrow—the magnetic field is at right angles to the loop, so links it to the maximum extent, provided the width is not more than half a wavelength. Although the e.m.fs from A to B and C to D are equal (and most of the time opposite, if the loop is small compared with a wavelength) they are out of phase, so there is a net e.m.f. around the loop.

At intermediate angles the results are intermediate, so the polar diagram is very like that of a dipole at right angles to the loop.

The size of a loop aerial in a portable broadcast receiver is, at

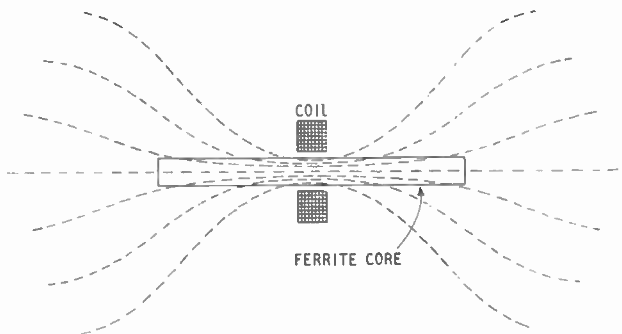



**Fig. 15.17—The concentrating effect of a suitable magnetic core on the magnetic field of the waves greatly increases the effectiveness of a small coil as an aerial**

most, very small compared with the wavelengths to be received, so the signal e.m.f. obtained by it is relatively weak. If, in order to embrace as much field flux as possible, aerial wire is wound around the set just inside its outer covering, it also embraces a good deal of assorted material, much of which causes considerable losses of the kind referred to on p. 113. If, however, the aerial winding is provided with a high-permeability core, the magnetic flux over a wide area tends to pass through the core, as shown in Fig. 15.17, just as electric current favours a low-resistance path. In this way a coil less than one inch in diameter is equivalent to a very much larger one without the core. It is essential, of course, for the core to be of low-loss material; the non-conducting magnetic materials known as ferrites are used.

CHAPTER 16

# Detection

## THE NEED FOR DETECTION

TURNING NOW to the reception end in Fig. 1.9, we see a number of boxes representing sections of a receiver. Although all of these sections are included in most receivers, they are not all absolutely essential. So we begin with the detector, because it is the one thing indispensable, for reasons now to be explained.

At the receiving aerial, the modulated carrier wave generates e.m.fs whose waveforms exactly match those of the currents in the sending aerial. The frequency of those currents, as we know (p. 22), is so high as to be inaudible. It has been chosen with a view to covering the desired range most economically (as discussed in the previous chapter), and bears no relationship to the frequencies of the sounds being transmitted. In order to reproduce those sounds, it is necessary to produce currents having the frequencies and waveforms of the sounds. For example, given the amplitude-modulated carrier wave shown in Fig. 13.5*c* (or 16.1*a*), the problem is to obtain from it Fig. 13.5*a*. The device for doing so is the *detector*, or, as it is often called, the demodulator.

For receiving frequency-modulated signals (p. 169) it is necessary also to include means for converting the frequency variations into the amplitude variations on which the detector works. This device —the *discriminator*—is considered at the end of this chapter.

(a)

(b)

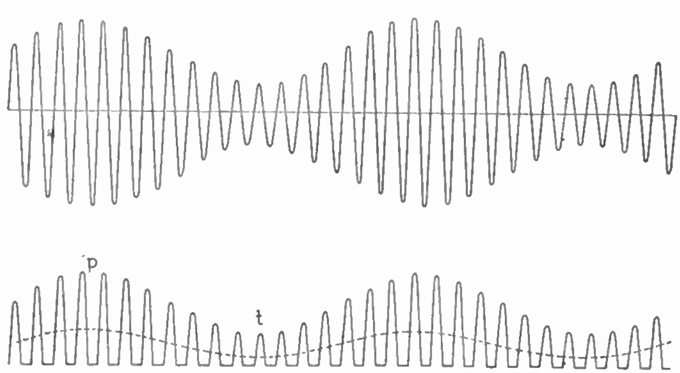



**Fig. 16.1—(*a*) Sample of an amplitude-modulated carrier wave, similar to Fig. 13.5 *c*. The average value of each cycle is zero; but if one half of each is eliminated (*b*) the average value of the other halves, shown dotted, fluctuates in the same manner as the modulating waveform (e.g., Fig. 13.5 *a*)**

## The Detector

The simplest method of detection is to eliminate half of every cycle, giving the result shown in Fig. 16.1*b*. At first glance, this might not seem to differ fundamentally from *a*, for the original frequency appears to be still there, though at half-strength. It is quite true that the original frequency is there, and means have to be provided for getting rid of it. The fundamental difference is that whereas the average value of each cycle in *a* is zero (because the negative half cancels the positive half), in *b* it is proportional to the amplitude (at that moment) of the modulated carrier wave, which in turn is proportional to the instantaneous value of the a.f. modulating signal. The dotted line in Fig. 16.1*b* indicates the average level of the current, and it can be seen to vary at the same frequency and with the same waveform as Fig. 13.5*a*. There is admittedly some loss in amplitude, but we shall see that the types of detector in common use manage to avoid most of this loss and give an output almost equal to the peak values.

The process of suppressing half of each cycle, converting alternations of current into a series of pulses of unidirectional current, is known by the general term *rectification*. The detector usually consists of a rectifier adapted to the particular purpose now in view, and may be taken to include means for extracting the desired a.f. from a mixture like Fig. 16.1*b*. It should be noted that this rectified signal contains not only r.f. and a.f. components, but also a unidirectional component (d.c.). That is, the desired a.f. current does not alternate about zero, as in Fig. 13.5*a*, but about some steady current—positive or negative, depending on which set of half-cycles was eliminated by the rectifier. Although this d.c. is unnecessary and in fact undesirable for reproducing the original sounds, it has its uses as a by-product (p. 288).

## Rectifiers

A perfect rectifier would be one that had no resistance to current flowing in one direction, and an infinite resistance to current in the opposite direction. It would, in fact, be equivalent to a switch, completing the circuit during all the positive half-cycles in Fig. 16.1*a*, thus enabling them to be passed completely, as in *b*; and interrupting it during all the negative half-cycles, suppressing them completely, as also shown in *b*. For very low-frequency alternating currents such as 50 c/s it is possible to construct such a switch (known as a vibrator rectifier; p. 364), but any device involving mechanical motion is impracticable at radio frequencies.

The two main types of electronic rectifier have already been considered: thermionic valves in Chapter 9, and semi-conductor rectifiers in Chapter 10. In the early days of radio, point-contact semi-conductor rectifiers using various minerals were much used under the name of crystal detectors. Between the world wars these were almost entirely superseded by vacuum valves, but were revived and improved during the second world war as the silicon diodes used for radar. More recently, germanium point-contact diodes have been competing with vacuum diodes as detectors for domestic and other receivers.

In considering resistance we began with the current/voltage graphs of ordinary conductors (Fig. 2.3), and noted that because such graphs were straight lines the conductors described by them were called linear, and that since linear resistances were covered by the very simple relationship known as Ohm's law there was really no need to spend any more time drawing their graphs. Then we came to valves, and again drew current/voltage graphs, such as Fig. 9.3, and found that they were curved, indicating non-linear characteristics, not conforming to Ohm's law (except in the modified and limited sense of Fig. 11.3, and then only approximately). All non-linear conductors are capable of some degree of rectification, and all rectifiers are necessarily non-linear in the above sense; that is, their resistance depends on the amount of voltage or current, instead of being constant as in Ohm's law.

Here now in Fig. 16.2 are some more current/voltage characteristics, the lines COD and AOB representing (as can be calculated from the current and voltage scales) 500 Ω and 2,000 Ω respectively.

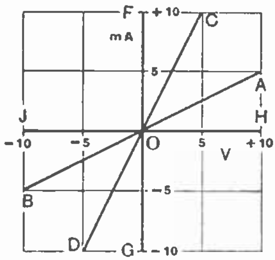
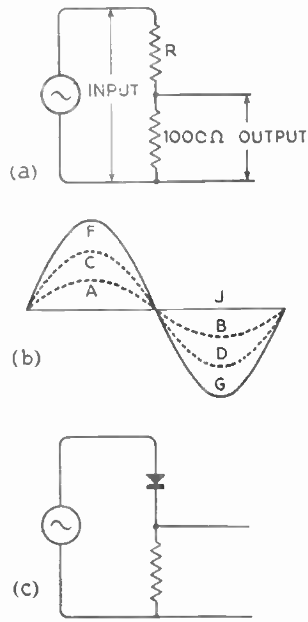



**Fig. 16.2—Characteristic curves of various linear resistances. The line COD is seen to represent 500 Ω. If the resistance to negative voltages is different from that to positive, as shown for example by the line COB, the result is a partial rectifier**

As we noticed before, the steeper the slope the lower the resistance. To go to extremes, the line FOG represents zero resistance, and HOJ infinite resistance; so FOJ is the graph of a perfect rectifier. COB is an example of a partial rectifier—a resistance of 500 Ω to positive currents and 2,000 Ω to negative currents. Now apply an alternating voltage to a circuit consisting of any of the foregoing imaginary resistances in series with a fixed resistance of 1,000 Ω (Fig. 16.3*a*). The voltage across the 1,000 Ω—call it the output voltage—is equal to the applied voltage only if *R* is zero. Both voltages are then indicated by the full line in Fig. 16.3*b*. If *R* were 500 Ω (COD), the output would be represented by the dotted line CD, and if 2,000 Ω by the dotted line AB.

If now a rectifier is substituted for R, as shown in Fig. 16.3*c*, the positive and negative output voltages are unequal. For example, the perfect rectifier (FOJ) gives the output FJ in *b*; and the partial rectifier gives CB. The average voltage, taken over each whole

**Fig. 16.3—The results of applying a sinusoidal voltage to a circuit (*a*) consisting of any of the resistances represented by Fig. 16.2 (denoted by R) in series with 1,000 Ω are indicated by *b*, which is lettered to correspond with the various characteristics in Fig. 16.2. When R is a rectifier it is represented as in *c***




cycle, is, of course, nil when R is not a rectifier. Using the perfect rectifier, the average of the positive half is 64% of the maximum (p. 73), while the negative is, of course, nil; so the average rectified current taken over a whole cycle is 32% of the peak value. Using the rectifier with the graph COB, the negative half-cycle (B in Fig. 16.3*b*) cancels out half of the positive half (C), which is two-thirds of the input; so the rectified result is less than 11% of the peak input.

## Linearity of Rectification

Whether complete or partial, however, the degree of rectification from rectifiers having characteristics like those in Fig. 16.2 would be independent of the amount of voltage, provided, of course, that the

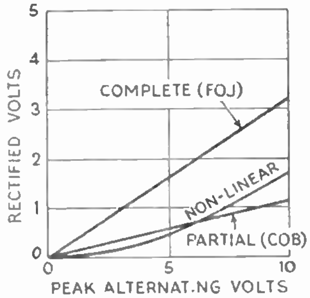



**Fig. 16.4—The rectified outputs obtained by applying alternating voltages to various rectifiers are shown in these characteristic curves, lettered to correspond with Figs. 16.2 and 16.3**

rectifier was arranged to work exactly at the sharp corner in its characteristic, whether that happened to correspond with zero voltage, as in Fig. 16.2, or not. So if we were to draw a graph showing the net or average or rectified voltage output against peak alternating voltage input to a rectifier of this kind we would get a straight line, such as the two examples in Fig. 16.4, corresponding to characteristics FOJ and COB in Fig. 16.2. These are therefore called linear rectifiers, it being understood that the term is in reference to a graph of this kind and not to one of the Fig. 16.2 kind.

It should be clear from Fig. 16.1*b* that it is necessary for the rectifier to be linear if the waveform of the modulation is not to be distorted. It is therefore unfortunate that in reality there is no such thing as a perfectly linear rectifier. The best one can do is to approach very close to perfection by intelligent use of the rectifiers available. The weak feature of all of them is the more or less gradual transition from high to low resistance, exemplified by the typical " bottom bend " in Fig. 9.3, which causes weak signals to be rectified less completely than strong ones. Such a characteristic is represented in Fig. 16.4 by the curved line, and is explained in greater detail later (p. 220).

## Rectifier Resistance

There were two reasons for the poor results of rectifier COB in Fig. 16.2 compared with FOJ. One was its *forward resistance* (500 Ω), which absorbed one-third of the voltage, leaving only two-thirds for the 1,000 Ω load. The other was its finite *backward resistance* (2,000 Ω) which allowed current to pass during the negative half-cycles, so neutralizing another one-third of the voltage. With such a rectifier it would clearly be important to choose the load resistance carefully in order to lose as little as possible in these ways. As it happens, 1,000 Ω does give the largest output voltage in this case, the general formula being $\sqrt{R_F R_B}$, where $R_F$ and $R_B$ are respectively the forward and backward resistances of the rectifier.

As compared with crystal diodes, the vacuum diode has one strong point in its favour. Although it may not be perfect in the forward direction, for it always has some resistance, it is practically perfect in the reverse direction; that is to say, it passes negligible backward current, even when quite large voltages are applied. So even if, owing to a high forward resistance, the forward current were very small, the output voltage could, theoretically at least, be made to approach that given by a perfect rectifier, merely by choosing a sufficiently high load resistance.

Although, as we shall soon see, there are limits to the resistance that can be used in practice, the loss of rectification can usually be made insignificantly small. In fact, by a very simple elaboration of Fig. 16.3*c*—a capacitor in parallel with the load resistance—it is possible to approach a rectification efficiency of 100% instead of the

32% that is the maximum without it, and so do nearly three times better than our "perfect" rectifier. This point deserves closer consideration.

## Action of Reservoir Capacitor

The modified rectifier circuit is now basically as Fig. 16.5. Suppose, in order to get the utmost rectified voltage, we made the load resistance *R* infinitely great. The capacitor C would then be in series with the diode, with no resistor across it. To simplify

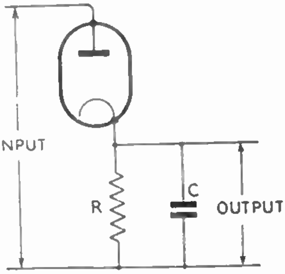



**Fig. 16.5—Diode rectifier with the load resistor (R) shunted by a reservoir capacitor, C**

consideration we shall apply a square wave instead of a sine wave; amplitude 100 V (Fig. 16.6*a*). We shall also assume that its frequency is 500 kc/s, so that each *half* cycle occupies exactly one millionth of a second (1 μsec). At the start C is uncharged, and therefore has no voltage across it. It can acquire a charge only by current flowing through the rectifier, which offers a certain amount of resistance, and so when the first positive half-cycle arrives its 100 V at first appears wholly across that resistance, as shown by the full line in Fig. 16.6*b*.

The current driven by this voltage starts charging the capacitor, whose voltage thereby starts to rise towards 100 V, as shown by the dotted line. It is clear that the voltages across rectifier and capacitor, being in series, must add up to give 100 V so long as that is the applied voltage; and Fig. 16.6*b* shows this to be so. The greater the resistance of the rectifier and the capacitance of the capacitor, the slower the rate of charge, just as with a very large balloon

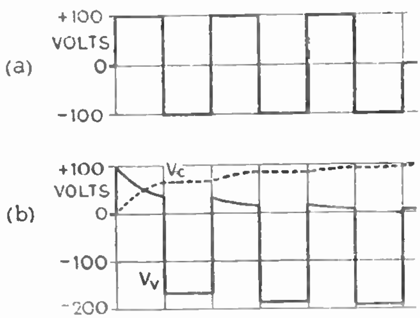



**Fig. 16.6—Analysis of application of a square alternating voltage (*a*) to the circuit Fig. 16.5. The dotted line in *b* represents the voltage across C; the full line, that across the diode**

inflated through a very narrow tube. As we noted in connection with Fig. 3.6, which showed another example of the same thing, the capacitance in microfarads multiplied by the resistance in megohms is the time constant—the number of seconds required for the capacitor voltage $V_C$ to reach 63% of the applied voltage. Suppose that the rectifier resistance (assumed constant) is 0·01 MΩ and the capacitance 0·0001 μF. Then the time constant is 0·000001 second or 1 μsec. In this case that happens to be the time occupied by one half-cycle of the 500 kc/s applied voltage. So at the end of the first positive half-cycle $V_C$ is 63, while $V_V$ has dropped to $100 - 63 = 37$. Then comes the negative half-cycle. The diode ceases to conduct, and while the capacitor therefore cannot charge up any more it likewise has no conducting path through which to discharge, and so remains at 63 V until the second positive half-cycle. Meanwhile, $V_C$ (63 V) plus $V_V$ must now be equal to the new applied voltage, $-100$. The voltage across the rectifier must therefore be $-163$.

The net voltage applied to the capacitor when the second positive half-cycle arrives is $100 - 63 = 37$ V, so at the end of this half-cycle the voltage across the capacitor will increase by 63% of 37, or 23 V, which, added to the 63 it already possessed, makes 86. The charge thus gradually approaches the peak signal voltage in successive cycles, while the average voltage across the rectifier falls by the same amount, as shown in Fig. 16.6*b*. The rectified output voltage therefore approaches 100% of the peak applied r.f. voltage. A similar result is obtained (but more slowly) with a sine wave signal.

Because C maintains the voltage during the half-cycles when the supply is cut off, it is called a *reservoir capacitor*.

## Choice of Component Values

This, of course, is excellent so far as an unmodulated carrier is concerned. When it is modulated, however, the amplitude alternately increases and decreases. The increases build up the capacitor voltage still further; but the decreases are powerless to reduce it, for the capacitor has nothing to discharge through. To enable $V_C$ to follow the modulation it is necessary to provide such a path, which may be in parallel with either the capacitor or the diode, so long as in the latter case there is a conducting path through the input to complete the discharge route. The resistance of the path should be low enough to discharge the capacitor C at least as fast as the modulation is causing the carrier amplitude to decline. The speed of decline increases with both depth and frequency of modulation, and calculating the required resistance properly is rather involved, but a rough idea can be obtained as follows.

Suppose, for example, the highest modulation frequency is 10,000 c/s. Then the time elapsing between a peak of modulation and a trough ($p$ to $t$ in Fig. 16.1*b*) is half a cycle, or 1/20,000th sec (50 μsec). If C had to discharge at least 63% of the difference in carrier amplitude in that time, the time constant would have to be not more than 50 μsec.

Of the two things determining this time constant, the value of the resistance is to some extent already decided on other grounds, for it must be much higher than the diode forward resistance, or there will be loss of detector efficiency. Incidentally, there will also be damping of the tuned circuit from which the r.f. voltage was derived. In practice one may reckon, as a fair approximation, that the damping effect of a diode detector of the Fig. 16.5 type with a load resistance $R$ is equivalent to connecting a resistance $R/2$ directly across the tuned circuit.* For these two reasons $R$ for a vacuum diode is seldom made much less than 0·5 MΩ. The value of $C$ is thereby fixed. If $R$ is 0·5 MΩ and $CR = 0{\cdot}00005$, then $C$ must be 0·0001 μF (or 100 pF), which in fact is quite a usual figure.

Note that at modulation frequencies as high as 10,000 c/s these values for $C$ and $R$ would result in quite an appreciable loss, for we were reckoning on $V_0$ following the modulation to the extent of only 63%. So from this point of view it would be better for $CR$ to be still smaller.

Is there any limit to reducing $C$? The answer can be obtained by reconsidering Fig. 16.6 on the assumption that the time constant is very short. Instead of $V_C$ remaining level during the negative half-cycles, as shown there, or declining slightly as it would with a moderate time constant, it would fall rapidly. So the reservoir action would largely fail; each positive half-cycle would have to start recharging $C$ almost from scratch, and there would be a loss of detection efficiency—at all modulation frequencies. The correct choice of $C$ is most difficult when the carrier-wave frequency is not many times greater than the top modulation frequency. If equal treatment of all modulation frequencies is reckoned more important than detection efficiency over all frequencies, then $C$ should be kept low; and vice versa.

At the much higher frequencies used for television and f.m. broadcasting, there is no difficulty in meeting requirements. The usual detector is a crystal diode, with values of both $C$ and $R$ considerably less than those arrived at for the lower frequency valve detector. In some circumstances the backward resistance of a crystal diode is a sufficient discharge path for C, but because it is so non-linear it is usually supplemented by R.

## The Diode Detector in Action

As the action of the deceptively simple-looking circuit in Fig. 16.5 is not very easy to grasp, and is also most important, it will be

* This is quite easily proved. Denote the resistance which, if connected across the tuned circuit in place of the diode and its load resistor R and reservoir C, would be equivalent to them, by $R'$. Then assuming the resistance of the diode is negligible in comparison with $R$, C will charge up to the peak voltage $\sqrt{2}E$, $E$ being the r.m.s. voltage developed across the tuned circuit. As R will be the only component dissipating power, the power dissipated (equal to the square of the voltage across R) divided by $R$ is $2E^2/R$. But from the way in which we have defined $R'$, this wattage must also equal $E^2/R'$. Therefore $R' = R/2$.

worth spending some more time on it. Fig. 16.7 is a diagram which, compared with Fig. 16.6, is speeded up so as to show some modulation, and uses proper sine waves for the carrier. Also, the voltage $V_C$ across the capacitor—which, of course, is the output voltage—is drawn together with the r.f. input voltage $V_I$ at *a*. Notice that whenever $V_I$ is more positive than $V_O$ the diode conducts and C charges rapidly, because the time constant is $CR_V$, $R_V$ (the forward resistance of the valve) being much less than $R$. Directly

(a)

(b)

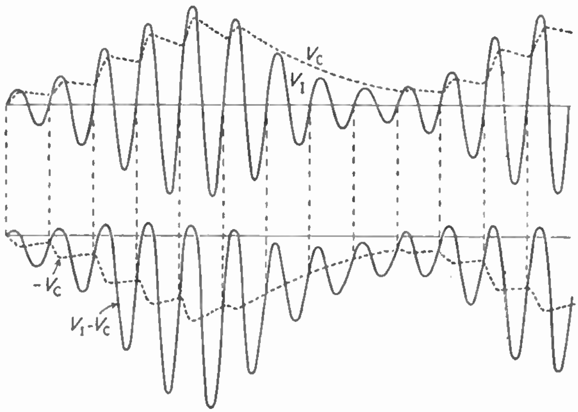



**Fig. 16.7—In diagram *a*, $V_I$ represents a modulated r.f. input voltage and $V_C$ the resulting output voltage. In diagram *b*, the full line shows $V_I - V_C$; i.e., the voltage across the diode**

$V_I$ drops below $V_C$, C discharges through the only path available (R) with the much longer time constant $CR$. It is too long in this case, because $V_C$ loses contact with $V_I$ altogether during the troughs, and its waveform is obviously a distorted version of the modulation envelope. But that is because for the sake of clearness the carrier frequency shown is excessively low for the modulation frequency.

Note too that in addition to the desired frequency, $V_C$ has also a saw-tooth ripple at the carrier frequency, but much less of it than in Fig. 16.1*b*. Not only has C improved the rectification efficiency but it has also got rid of most of the unwanted r.f.

The third component making up $V_C$ is a steady positive voltage (zero frequency), represented by the average height of the dotted line above the zero line. The voltage across the valve must, of course, be equal to $V_I$ minus $V_C$ (Fig. 16.7*b*). So it also contains the z.f. and a.f. voltages, and therefore could be used alternatively as the output voltage, but with the disadvantage of having nearly the full r.f. voltage along with it.

## Varieties of Diode Detector Circuit

We have just seen that a diode detector circuit arranged as in Fig. 16.5 gives an output consisting of the desired a.f., plus a positive z.f., plus a much reduced r.f. This arrangement, repeated in Fig. 16.8*a*, is the most usual with crystal diodes, but with thermionic diodes is open to the objection that if a.c. is used for heating the cathode an appreciable fraction of its voltage may appear across R via the stray capacitance between heater and cathode. And in some receivers there is a need for a *negative* voltage proportional to the carrier amplitude. It can be obtained if the circuit is rearranged as in Fig. 16.8*b*, and at the same time the possible heater trouble is removed. This circuit is the one generally used when the input is tuned to a fixed frequency. But if it is necessary to use a variable tuning capacitor, it is often inconvenient to have both its terminals at varying potential. So if *a* is ruled out, *c* is an alternative, in which the output is drawn from across the diode and so contains most of the original r.f. If, for reasons that will appear in Chapter 17, the tuned circuit is at a high positive potential, circuit *c* would apply this potential via R to the diode, taking it away from its effective rectifying point. To obviate this, R must be connected

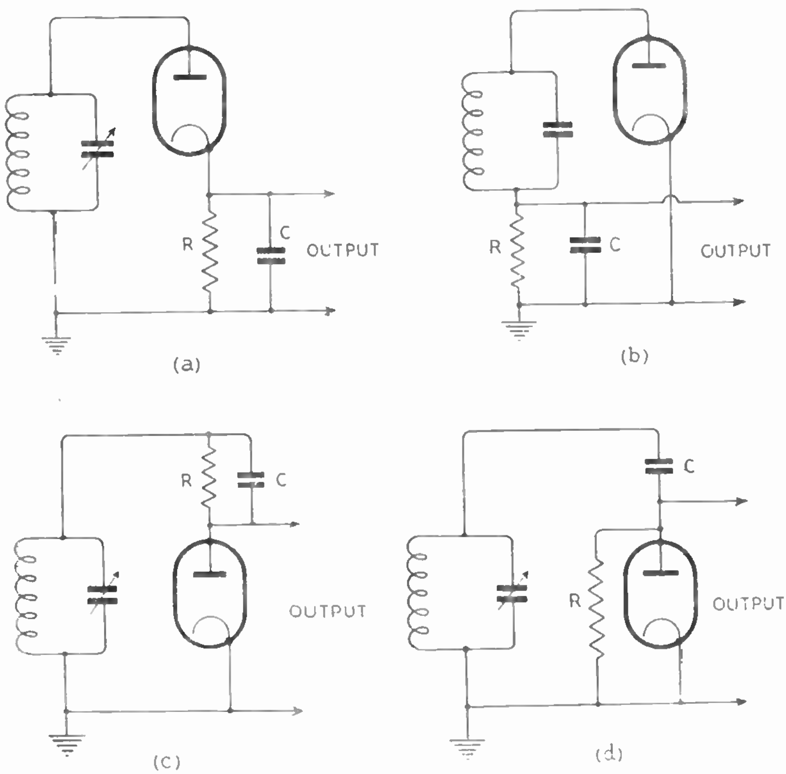


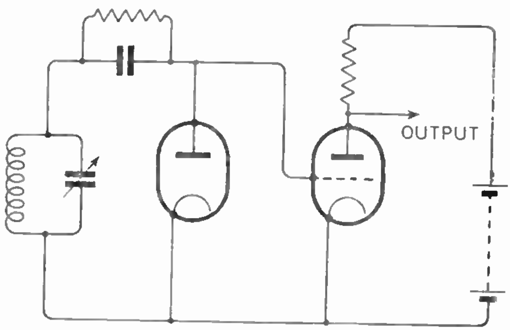

Fig. 16.8—The principal varieties of diode detector circuit

to the cathode side, as at *d*. Unlike the three other arrangements, in this one R is not by-passed by C, so it carries the r.f. a.c. as well as the d.c. and consequently loads the tuned circuit rather more heavily. The effective resistance is in fact approximately $R/3$ instead of $R/2$. As in *c*, the output contains a large proportion of r.f., so *a* and *b* are used wherever possible.

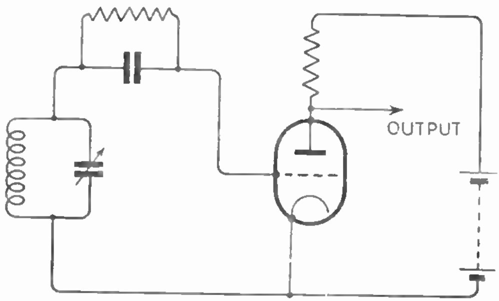
## THE GRID DETECTOR

If any of the foregoing circuits were connected directly to a loudspeaker or a pair of headphones, the comparatively low resistance of such devices would completely upset the carefully-chosen detector circuit values, even if the amount of power available were sufficient to give useful results. So in practice a diode detector is followed by one or more stages of amplification.

If we connect one of these last two detectors, say Fig. 16.8*c*, straight to an amplifier, as shown in Fig. 16.9*a*, we find that the anode-cathode path of the diode is in parallel with the grid-cathode path of the triode. Since both the anode of the diode and the grid




(a)




(b)

**Fig. 16.9—If a diode detector (Fig. 16.8*c* variety) is connected straight to a triode amplifier (*a*), the grid-cathode path of the triode renders a separate diode unnecessary; this is what happens in the grid type of detector (*b*)**

of the triode consist simply of an electrode close to an emitting cathode, there would seem to be no need to have them both. Experiment confirms this supposition; there is no significant change in the performance of the system if the diode is removed from its socket.

The simplified result, Fig. 16.9*b*, is the *grid detector*, in which the grid and cathode, acting as a diode, rectify just as in Fig. 16.8*c*. The resulting a.f. voltages at the grid then control the electron stream through the triode and so produce an amplified voltage at the anode in the way described on p. 142.

In practice this apparently neat idea does not usually work out very satisfactorily. As we have already noticed, the output of the type of diode circuit employed contains not only the desired a.f. but also a negative z.f. and a large amount of r.f. The negative z.f. may be useful, as a source of the necessary bias voltage for the triode (p. 123). But the r.f. is a nuisance, because it limits the amplitude of signal that can be handled without distortion to about half of what it could be if the r.f. were not there. Moreover, the amplified r.f. is liable to work round to the input through stray couplings and may even cause oscillation.

We shall soon see that restricting the detector input to not more than a volt or two of r.f. tends to introduce distortion in another way, and also limits the usefulness of the detector for auxiliary duties (p. 288). If, to escape this dilemma, the anode voltage of the triode is increased, it is found that the anode current runs up to an alarming figure, and even then may not satisfy the requirements in modern receivers, which often call for a rectified output of 20 V or more.

The apparent economy of this circuit therefore proves in most cases to be illusory. With the separate diode there is no restriction on the r.f. voltage, and the residue left in the output can be removed from the a.f. before it is passed on to the next stage.

## FILTERS

Devices for separating currents of different frequencies are known as *filters*, and are a vast subject in themselves. Those used in detectors are usually of the very simplest types, however. All filters depend on the fact that inductive and capacitive impedances vary with frequency. The simple circuits of Figs. 6.9, 7.3 and 8.1 can all be used as filters.

Take Fig. 6.9: C and R in series. At very low frequencies, C offers a high impedance compared with R, and nearly all the voltage applied across them both appears across C. As the frequency rises the impedance of C falls, until, at a very high frequency, most of the voltage appears across R. The dividing frequency, at which the voltages across C and R are equal to one another (and to the input voltage divided by $\sqrt{2}$), is given by $X_C = 1/2\pi fC = R$, so $f = 1/2\pi CR$.

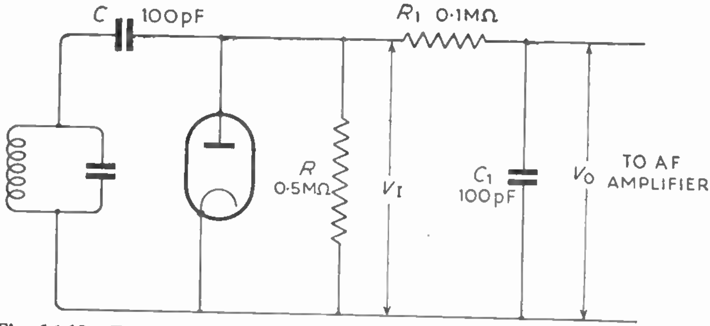



**Fig. 16.10**—To prevent the r.f. voltages left over after rectification being passed on to the a.f. amplifier, a simple filter, $R_1$ and $C_1$, is commonly used. The component values shown are typical

Suppose the lowest carrier frequency to be handled by a detector is 150 kc/s, and the highest modulation frequency is 10 kc/s. One might make the dividing frequency 15 kc/s. Knowing $R$ we can then find $C$, or vice versa. If the input voltage is to be derived straight from a diode detector, the impedance of the filter ought not to be too low, or $C$ too high. Let us try 100 pF. Then, using c/s, μF and MΩ units, $R = 1/(2\pi \times 15{,}000 \times 0{\cdot}0001)$ = about 0·1 MΩ.

Fig. 16.10 shows a diode circuit (type Fig. 16.8*d*) with this filter incorporated, as $R_1$ and $C_1$. R, of course, is the load resistor. Having calculated the reactance of $C_1$ at various frequencies, one can easily work out the ratio of output voltage to input voltage, $V_O/V_I$, and plot it as a graph, Fig. 16.11. This shows clearly how a filter with the selected component values would cause negligible loss of modulation frequencies but would reduce the residual r.f. considerably. Admittedly this graph does not present an exact

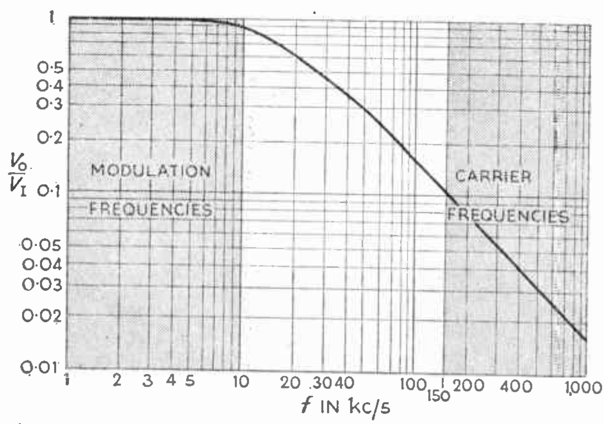



**Fig. 16.11**—The filtering action of $R_1$ $C_1$ in Fig. 16.10 is here shown as a graph of $V_O/V_I$ against frequency

picture, because the filter is a shunt circuit to the load resistance and would affect the rectification somewhat; but with such high-impedance filter components it is broadly true.

The possibility of using the negative voltage resulting from rectification as a bias for the following amplifier valve has been mentioned. It is not generally considered good to have a bias voltage that varies with the amplitude of r.f. reaching the detector, and a more usual policy is to exclude this voltage and apply a suitable bias to the amplifier in some other way. A filter to cut out a frequency lower than the desired frequencies—zero-frequency in this case—can be similar to the previous one, but with C and R interchanged. C in series definitely prohibits z.f., so is often called a blocking capacitor; but it must have sufficiently large

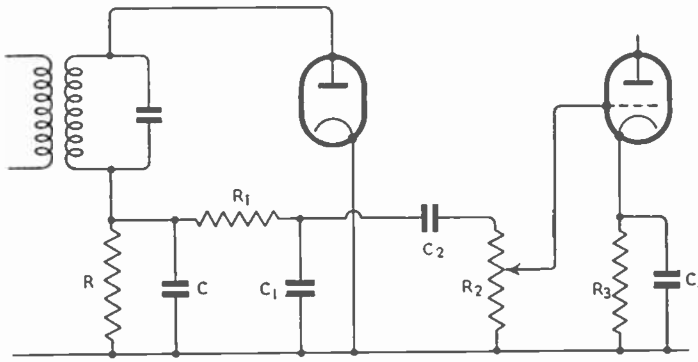



**Fig. 16.12—Typical detector circuit, in which R is the load resistor, C the reservoir, $R_1C_1$ the r.f. filter, $R_2C_2$ the z.f. filter, $R_2$ the volume control, $R_3$ the biasing resistance, and $C_3$ its by-pass**

capacitance to pass the lowest modulation frequency. If the associated resistance is 1 MΩ, 0·01 μF or larger is usually satisfactory.

## Complete Detector Circuits

This resistance, incidentally, is commonly used for volume control in the manner shown in Fig. 16.12, which is a typical detector circuit, developed from Fig. 16.8*b*. $R_1$ and $C_1$, as before, are the r.f. filter, and $R_2C_2$ the z.f. filter, $R_2$ also serving as volume control. This is the best place for it, because, as we shall see, distortion is minimized by having the maximum possible signal amplitude at the diode detector and the minimum going into the amplifier. The purpose of $R_3$ and $C_3$ is to provide bias for the amplifier valve, as explained on p. 269.

Although for clearness separate valves are shown, in practice the diode is generally incorporated with the triode, making use of the same cathode. To prevent the bias due to $R_3$ from affecting the diode action, the lower end of R is joined to the common cathode instead of to the common negative line.

An alternative arrangement, which dispenses with $R_3$ and $C_3$ and is commonly used in a.m. broadcast receivers, is shown in Fig. 16.13. The volume control $R_2$ serves also as the load resistance, so no R appears; and the grid bias is provided $C_4$ and $R_4$ acting in the same way towards the a.f. signal as C and R in Fig. 16.8*d* to the

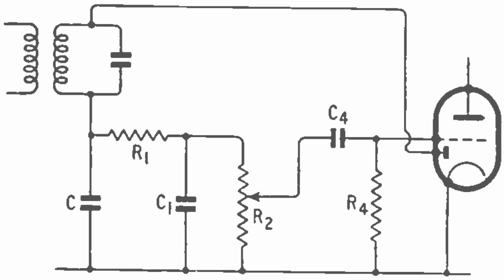



**Fig. 16.13—Alternative arrangement in which the negative grid bias for the a.f. amplifier is developed across a grid leak ($R_4$) of exceptionally high resistance**

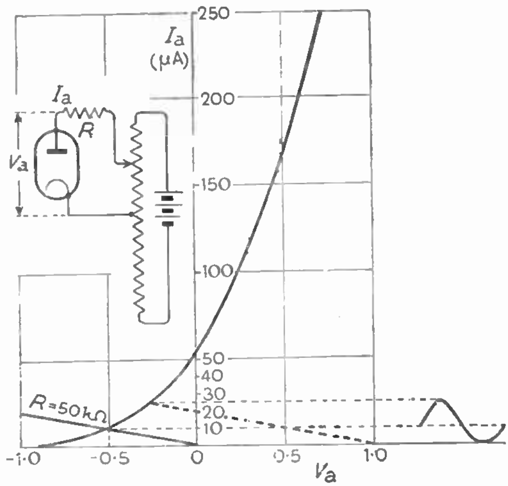
r.f. signal, as shown in Fig. 16.7*b*. To keep the variations in bias voltage well below the audible limit, the time constant $C_4R_4$ is made large; and to keep grid current very small, $R_4$ is large, usually 10 MΩ or even more. A common value for $C_4$ is 0·01 μF.

## Detector Characteristics

In discussing rectifiers we noted that they had to be non-linear conductors in order to rectify, but that if the non-linearity consisted of a sharp corner between two straight portions (e.g., COB in Fig. 16.2), the rectifier itself could be described as linear (e.g., COB in Fig. 16.4). And that no real rectifiers have perfectly sharp bends, so all are more or less non-linear. Now that we have seen the details of complete detector circuits we shall return to this point and consider how to ensure that the non-linearity is less rather than more.

The various valve characteristic curves in Chapter 9 show that anode current begins gradually at first, and then the curve straightens out considerably. But this does not happen until the anode current is several milliamps. If we were to use a detector load resistance so low as to pass several milliamps, it would damp the tuned circuit excessively (p. 213). So for detector purposes the only part that concerns us is the microamp region, for which we need an enlarged view of the bottom bend (Fig. 16.14). Anode current ($I_a$) normally begins at an anode voltage ($V_a$) somewhere between 0 and − 1·5 V. So if the anode were connected direct to the cathode with no load resistance, a current of about 50 μA (in this particular example) would flow.

**Fig. 16.14—The action of a detector can be examined in detail with the help of the load-line technique**




When a load resistance is inserted, the drop in voltage across it due to current flowing through it makes the anode more negative. To find how much more, we adopt the load-line method explained in Chapter 11. For clearness let us take the rather low load of 50 kΩ, connected straight to cathode (zero volts). That means the load line must be drawn from the origin of the graph, 0. When so drawn, it cuts the valve curve at $-$ 0·5 V, 10 μA, these being respectively $V_a$ and $I_a$ under the conditions assumed.

Now suppose a small alternating signal voltage is applied in series with R and the diode. Its effect can be represented on the diagram by making the load line vibrate horizontally to right and left of point 0 to the extent of the signal amplitude. The point of intersection with the valve curve will move up and down, indicating an alternating anode current. If first we suppose that the signal voltage is very small, say 0·01 V peak, then the tiny portion of curve actually used is practically straight, and there will be hardly any rectification. $I_a$ will alternate slightly above and below 10 μA, the negative half-cycle almost completely neutralizing the positive half. Small-signal detection is therefore very inefficient.

If the valve curve started up from zero at a sudden angle, even if only a slight one, we could improve matters by bringing our working point to it, either by applying a negative bias or by using a much higher load resistance. But magnifying the bottom part of the curve fails to reveal any sharp corner, so all we would do in our attempt to suppress the negative half-cycles would be to suppress the positive halves as well!

Now imagine that the signal voltage is increased to, say, 1 V peak. Moving the load line up to + 1 V to represent the positive half-cycle brings the anode-current peak up to about 25μA, a net peak value of + 15 μA; whereas the negative half-cycle cannot be more than − 10 μA at most. Averaged over the whole cycle, the anode current (shown at the extreme right) shows a small net positive value. In other words, there is some rectification, though still not very much. To show the results with very large signal voltages it would be necessary to use an $I_a/V_a$ diagram on such a scale that the gradual bend would look like a sharp corner, and rectification would approach the ideal.

Drawing a curve of rectified current, or rectified voltage due to this current passing through R, against signal voltage, we get a curve such as that in Fig. 16.15, showing poor detection of weak signals, improving with increase in signal.

## Detector Distortion

Consider now what happens when the signal is amplitude-modulated. The signal strength periodically waxes and wanes at the modulation frequency. The curve at the foot of Fig. 16.15 represents (on the r.f. voltage scale) one cycle of sinusoidal modula-

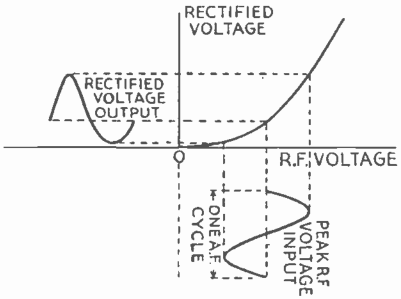



**Fig. 16.15—Graph of one-way or z.f. voltage derived from a rectifier is plotted against r.f. input voltage. It enables the distortion of the modulation to be ascertained for any r.f. amplitude and depth of modulation. The input curve here represents 50% modulation**

tion to a depth of 50%. It traces the positive tips of all the many r.f. cycles during one a.f. cycle. The negative modulation half-cycle reduces the rectified voltage less than the positive half-cycle increases it, so the rectified waveform is distorted as shown to the left of the diagram. From the shape of the curve one can see that there would be some distortion at any depth of modulation, but the deeper the modulation the greater the distortion.

Further consideration of Fig. 16.15 will show that if the carrier wave is considerably stronger, enough to bring it on to the straight, then there will be no distortion so long as the depth of modulation is not enough to bring the signal strength down on to the bend. Modulation to a depth of 100% is bound to be distorted because it brings the signal strength at the negative peaks down to zero. The greater the carrier-wave voltage, the greater the depth of modulation before distortion begins, and the less the distortion at 100% modulation. That is one of the reasons why it is a good thing for the receiver to be designed so that the signal voltage at the detector is as large as possible, and at least not less than several volts.

In discussing Fig. 16.14, we assumed a load resistance with no reservoir capacitance. The effect of including the reservoir is to introduce a negative bias voltage proportional to the rectified current, pushing the initial working point lower down the curve. But since with most valves the degree of curvature anywhere on the bottom bend is very much the same, the general results are substantially as described above for the unshunted load resistor.

A great advantage of the diode detector over other types is that correct operation is mainly a matter of arranging that the input signal voltage is not too small. There is, however, a rather serious possible cause of distortion which is unaffected by signal strength. In the typical circuit shown in Fig. 16.12 the load resistance, so far as d.c. is concerned, is $R$. But a.c. at modulation frequency has another path in parallel, via $C_2$ to $R_2$. Since the impedance of $C_2$ is made relatively small, the a.c. load resistance is effectively $R$ in parallel with $R_1 + R_2$, and therefore lower than the d.c. load resistance. This causes the current through the diode to swing more widely, above and below the mid position, when the r.f. signal is modulated.

With 100% modulation this cannot happen at the negative peaks, because the diode current cannot become less than zero. So the negative peaks are clipped, distorting them. To avoid distortion of deep modulation it is advisable to make $R_1 + R_2$ much greater than $R$. This condition is admirably fulfilled by the Fig. 16.13 circuit.

## The Anode-Bend Detector

We have given most attention to the diode detector because it is by far the commonest. The closely-related grid detector has also come under notice. There is one other type important enough to look at here.

We have seen that the $I_a/V_g$ characteristics of a triode (Fig. 9.6) are very similar in shape to the $I_a/V_a$ characteristics of a diode (Fig. 9.3). Both are one-way current devices and therefore rectifiers. When using a triode for amplifying, we adjust the negative grid bias to place the working point on the straightest part of the curve. But by increasing the bias so as nearly to cut off the anode current, we can use the valve as a detector, as in Fig. 16.16, where the components are marked to correspond with their equivalents in Fig. 16.12. This *anode-bend* detector has the advantage over the diode type that the input circuit is non-conducting—so long as the peaks of input voltage do not run into grid current. The power in the load resistance is supplied by the anode battery, instead of being taken from the tuned input circuit.

On the debit side, however, the bottom bend in the characteristic curve of a triode is generally more gradual than that of a diode, and the disadvantages described in the last two sections are consequently greater. Moreover, the remedy applicable to the diode—making the input amplitude so large that the bend is relatively unimportant—

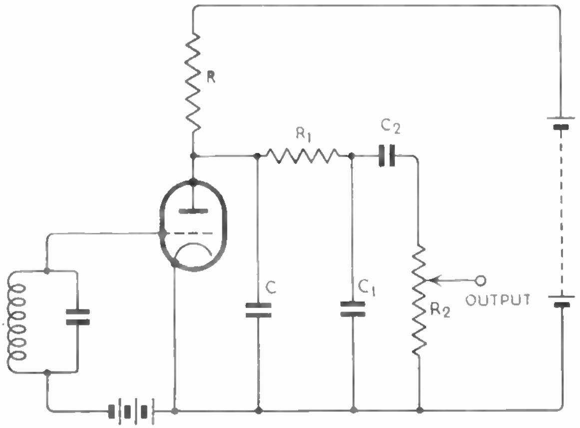



**Fig. 16.16—Anode-bend detector circuit, which should be compared with Fig. 16.12**

does not work with the triode, because grid current would flow at the positive peaks, tending to rectify them in opposition to the negative rectification at the bend, as well as causing serious distortion. For this and other reasons anode-bend detection is now seldom used except for special purposes.

## F.M. Detectors

The performance of all the detectors considered so far depends hardly at all on the frequency of the carrier wave between very wide limits. Consequently they treat a frequency-modulated carrier wave (p. 169) as if it were unmodulated. In order to receive f.m. signals it is necessary either to convert the variations of frequency into corresponding variations of amplitude so that an ordinary detector can be used, or else use a different kind of detector altogether.

The first method can be employed in a very simple manner with an ordinary a.m. receiver by adjusting the tuning so that instead of the incoming carrier-wave frequency coinciding with the peak of the resonance curve (Fig. 8.6) it comes in on one of the slopes. As its frequency rises and falls by modulation, the amplitude of response rises and falls. But because the slope of the resonance curve is not quite linear, the amplitude modulation is not a perfect copy of the frequency modulation, so there is some distortion. More serious still, the potential benefits of f.m. are not achieved, because the receiver is open to amplitude modulation caused by interference. And both selectivity and efficiency are poor.

Receivers designed for f.m. reception therefore include means for rejecting amplitude modulation and responding only to frequency modulation. The first of these functions is performed by a *limiter* and the second by a *discriminator*. In one of the most popular types of f.m. detector—the ratio detector—these functions are combined. However, for a first approach it will be easier to consider a system in which they are separate.

In connection with oscillators we saw (p. 156) that if the amplitude exceeds a certain amount the valve cuts off the positive or negative peaks or both. A limiter is usually a valve so arranged that it does this at anything above quite a small signal amplitude, so that however much greater the incoming signal may be it is cut down or limited to the same size. Variations in amplitude are therefore almost completely removed, and with them much of the effects of interference.

Fig. 16.17*a* shows the essentials of the Foster-Seeley or phase discriminator. It consists of two Fig. 16.8*a* type a.m. detectors back to back, both receiving the incoming r.f. carrier-wave voltage $V_p$ from a tuned circuit $L_p$ $C_p$, and each receiving in addition half the voltage from the tuned circuit $L_s$ $C_s$. $L_p$ and $L_s$ are the

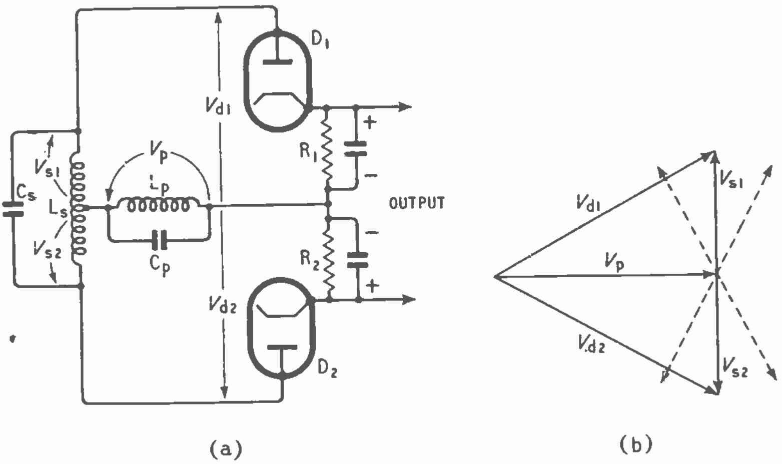



**Fig. 16.17—(*a*) Circuit of phase discriminator for f.m. detection, arranged to bring out the principle : $L_p$ and $L_s$ are the primary and secondary windings of the input r.f. transformer; (*b*) vectors of voltages marked, for carrier wave frequency. Dotted lines indicate the effect of f.m. deviation**

primary and secondary windings of a r.f. transformer, and because both windings are exactly tuned to the carrier wave and are coupled only very loosely the voltages across them are 90° out of phase, as indicated by the vectors $V_p$, $V_{s1}$ and $V_{s2}$ in Fig. 16.17*b*. The total r.f. input to diode $D_1$ is therefore $V_{d1}$, and to diode $D_2$ it is $V_{d2}$. These are obviously equal, so the rectified voltages set up across the load resistors $R_1$ and $R_2$ will also be equal. But as they are connected in opposite polarity, the net output is nil.

When the carrier wave is frequency-modulated it swings off the resonance peak and, as explained on p. 109, the impedance of the tuned circuit changes from resistance to inductance on one side and capacitance on the other. The phase swings correspondingly, as indicated by the dotted lines in Fig. 16.17*b*, changing the one-to-one ratio of $V_{d_1}$ to $V_{d_2}$ to something alternately greater and less. The value of $V_{d_1} - V_{d_2}$ therefore becomes alternately positive and negative at the frequency of modulation, and the modulating signal is reproduced at the output.

Practical phase-discriminator circuits differ in detail from Fig. 16.17*a*—for instance, $L_p$ is parallel-fed like the tuned circuit in Fig. 12.7 and for a similar reason—but the main principle is the same. The ratio detector circuit is not unlike the phase discriminator, but it combines the limiter function in rather a subtle way. Both types commonly make use of crystal diodes in place of the vacuum type shown. There are also several entirely different but comparatively seldom used types of f.m. discriminator.

# Radio-Frequency Amplification

## The Need for R.F. Amplification

In our study of detectors we noted that the effectiveness of all types deteriorated very rapidly as the r.f. voltage was reduced much below about 1 V. Signals picked up by the receiving aerial are seldom more than a few millivolts, and are often only microvolts; so they would be far too weak, even when magnified to some extent by the tuned circuit, to operate a detector effectively—to say nothing of distortion.

In practice, then, it is almost invariably necessary to amplify the signals before they reach the detector; that is to say, while they are still at *radio* frequency. For this reason, no time has been spent in considering details of means for coupling the receiver aerial to the detector.

One might suppose that the desired result could be obtained by putting a resistance-coupled amplifier stage, as described on p. 140,

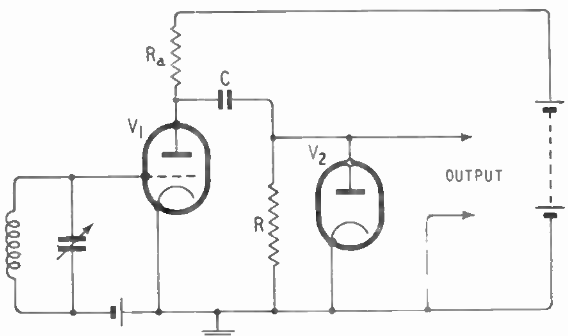



Fig. 17.1—This resistance-coupled amplifier is of little or no value for r.f amplification

between the tuned circuit which for the time being we are regarding as the source of r.f. signal voltage, and the detector, as in Fig. 17.1. In valve circuits it is usually convenient for as much of the circuit as possible to be at the negative or earthed end; and because the anode of the amplifying valve is at a comparatively high positive potential we choose variety *d* from the selection of diode detectors in Fig. 16.8. Here, C serves not only as the reservoir capacitor to be charged by the one-way current through $V_2$, but also as a blocking capacitor (p. 154). The symbol $R_a$ denotes the external anode load resistance for $V_1$, the capital letter being used to distinguish it from the valve's internal anode resistance $r_a$ (p. 119). Grid bias for $V_1$

is for simplicity shown as coming from a battery; in practice other methods are more often used.

From what was said in Chapter 11 one would say that the signal voltage developed across the tuned circuit would be reproduced on a considerably enlarged scale across $R_a$, and the resulting r.f. voltage at the anode would be good meat for the detector. But if one were to try this circuit the result would be disappointing. Reception might even be weaker than it was without the amplifier, especially at television or f.m. frequencies. Yet it looks all right according to the book.

## Stray Capacitance

This is where we begin to make acquaintance with the complications that arise when we have to deal with real circuits instead of ideal circuits on paper. This particular circuit has an important " invisible component ": the stray capacitance in parallel with $R_a$. This includes the capacitance to earth of the anode of $V_1$, the anode-connected part of its valve holder, the capacitor C as a whole, and the anode of $V_2$, with associated wiring. All this may not look as if it were across $R_a$, but so far as signal currents are concerned the battery is just a low resistance and in effect $R_a$ is connected straight from anode to earth. Even with careful arrangement, the total is likely to be at least 16 pF, which at 1 Mc/s is 10 kΩ. At 50 Mc/s, it is only 200 Ω! No matter how high a resistance may be chosen for $R_a$, the effective coupling impedance cannot exceed these figures. A few calculations as explained in Chapter 11 would show that with any available triodes the gain could be no more than moderate at medium frequencies, more or less non-existent on Band I television frequencies (41–68 Mc/s), and actually a loss on higher bands. But there is worse to come.

## Miller Effect

Another invisible component is the capacitance from anode to grid, shown in Fig. 17.2 as $C_{ag}$. With the minimum of associated wiring, etc., it would probably amount to about 4 pF. The stray capacitance from anode to earth which we have just discussed is also shown, as $C_{ae}$. $R_a$ can be regarded as reduced slightly so as

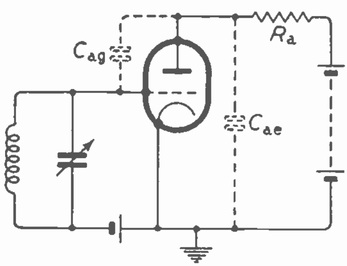



**Fig. 17.2—The reasons for the lack of success with Fig. 17.1 are the capacitances from anode to grid and earth, shown here dotted**

to incorporate the input resistance of the detector. So altogether the working conditions of $V_1$ in Fig. 17.1 are fairly represented.

Since the amplifying action of the valve produces a signal voltage at the anode, a small signal current flows through $C_{ag}$ and the grid circuit to the cathode of the valve. In flowing through the components in the grid circuit, this current develops across them a voltage, and this voltage may have any phase relationship with the voltage already present due to the signal. If it were in phase with the original signal voltage the two would simply add, and the original voltage would be increased. If it were 180° out of phase, on the other hand, this new voltage would be in opposition to that

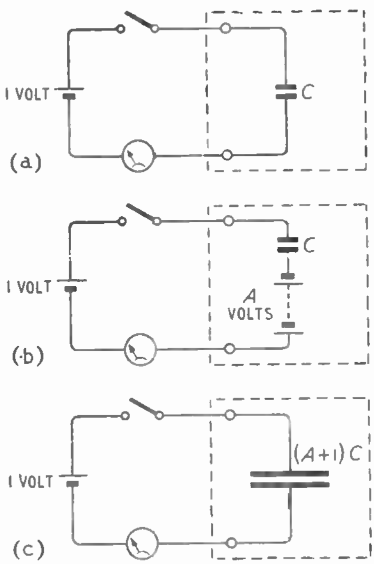



**Fig. 17.3—Showing how the introduction of an extra voltage, such as that fed back from anode to grid in a valve, can be equivalent to a multiplied capacitance**

already there, and the energy fed through $C_{ag}$ would tend to damp out and reduce the signal voltage. In a third case the voltage fed back from the anode might be 90° out of phase with that already present, and would therefore produce the same effect on the grid signal as a reactance.

In general there will be a combination of this with one of the other two.

The phase in any particular circuit will depend on the nature of the impedance in the anode circuit, which we shall call $Z_a$. Let us first neglect $C_{ae}$ and consider the case in which $Z_a$ is a resistance, because we have already seen (p. 142) that the signal voltage at the anode of a resistance-coupled amplifier is opposite in polarity to that applied to the grid, and $A$ times as great (where $A$ denotes the stage amplification). Therefore every signal volt applied to the grid causes $A + 1$ volts between anode and grid. Consider Fig. 17.3*a*, in which C is a capacitor hidden in a box, in series with a special sort of meter that measures the charge passing into the capacitor when the switch is closed. As capacitance is equal to the charge required to raise the p.d. across it to 1 V (p. 50), we have here a means of measuring $C$. But suppose (*b*) that a demon in the box connects an $A$-volt battery in series with C whenever we connect our 1-volt battery. The charging voltage is $A + 1$ volts, and the meter indicates a charge $A + 1$ times as large as it would without the demoniacal activity. Knowing nothing of this, we conclude that the box contains a capacitor having a capacitance $A + 1$ times as large as $C$. So far as external electrical tests can tell, the boxes *b* and *c* are identical.

Thus the effect of the grid-to-anode capacitance $C_{ag}$ is to make our resistance-coupled amplifier behave as if a capacitance $A + 1$ times as large as $C_{ag}$ were connected across the tuned circuit. As $A$ might be 50 or more, this apparent capacitance could be well over 100 pF.

If $Z_a$ is an inductance, the phase of the voltage fed back is such that it reinforces and increases the signal voltage on the grid.

If $Z_a$ is capacitive, the fed-back voltage tends to oppose the signal voltage applied to the grid. The result is the same as if a conducting path for signals were connected across the input to the valve. This is the effect we have expressly tried to avoid (p. 124) by using negative grid bias.

These feedback effects via $C_{ag}$ are named after their discoverer, J. M. Miller. How do they concern our would-be amplifying circuit represented in Fig. 17.2?

At very high frequencies, $Z_a$ is mainly capacitive, because $C_{ae}$ has such a low impedance that nearly all the current goes that way rather than via $R_a$. So Miller effect is mainly of the kind that causes loss of signal voltage. The loss will not be great, because the lowness of $Z_a$ ensures that $A$ will be very small. But that is not much comfort!

At frequencies of the order of 1 Mc/s, used for medium-wave sound broadcasting, the reactance of $C_{ae}$ and the resistance $R_a$ are likely to be of the same order of magnitude, in which case both capacitive and conductive paths will be thrown across the input to the valve. Owing to $Z_a$ being relatively large, a useful amplification might be possible were it not that the greater it is the more it will destroy itself by Miller effect.

With this type of amplifier, then, we seem to be in a cleft stick; whatever the frequency within the wide r.f. range, the amplification of the valve is brought to little or nothing.

## The Tuned-Anode Circuit

Let us tackle the first of the twin trouble-makers—$C_{ae}$. Chapter 8 showed us how to tune out capacitance by means of inductance in series or parallel. In this case obviously it should be in parallel, because what we want is a high impedance. We saw that in

Fig. 17.1 the stray capacitance $C_{ae}$ was in parallel with $R_a$, so if as in Fig. 17.4 we substitute for $R_a$ the correct amount of inductance it will neutralize or tune out $C_{ae}$. An advantage of this position is that $L_a$ serves to conduct the anode current, and does so with far less voltage loss than $R_a$. The inductance to provide a specified neutralizing reactance depends on the frequency; as does

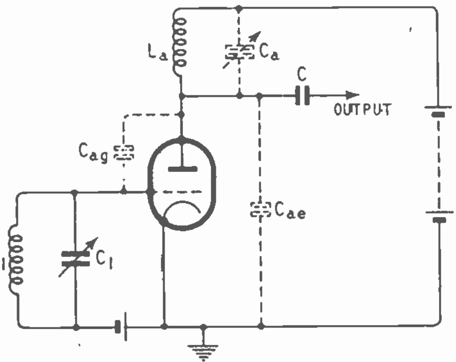



**Fig. 17.4—The effect of $C_{ae}$ in Fig. 17.2 can be overcome by substituting a tuned coupling for the resistance, but there is still trouble due to $C_{ag}$**

also the reactance to be neutralized, but in the opposite way. In practice a tuning inductance is often varied by moving an iron-dust or ferrite core in and out, termed *permeability tuning*. An alternative is to vary the reactance to be neutralized, augmenting $C_{ae}$ by means of a variable capacitor $C_a$ across $L_a$, as shown dotted.

When this anode coupling impedance has been tuned to the frequency of the desired signal it behaves as a high resistance—the dynamic resistance (p. 108). For that particular frequency we have therefore achieved the equivalent of the unrealizable pure resistance-coupled arrangement of Fig. 17.1, with two advantages added. Of these, the minor advantage is avoiding the loss of anode supply voltage in $R_a$. The major advantage is that full amplification is conferred only on signals of the desired frequency; in other words, the selectivity is improved, as will be discussed at length in the next chapter.

The gain given by a tuned-anode stage is calculated by the formula $A = \mu R/(R + r_a)$ given on p. 144 for a resistance-coupled stage, $R$ being interpreted as the dynamic resistance of the tuned circuit.

We have found a remedy for the stray capacitance $C_{ae}$; what now about $C_{ag}$?

## Instability

As long as the anode circuit in Fig. 17.4 is tuned exactly to the signal frequency it is purely resistive, so Miller effect neither assists nor damps down the signal voltage at the grid but throws in capacitance there, merely reducing the amount of $C_1$ needed for tuning to that frequency. If the applied frequency (or alternatively the capacitance $C_a$) is increased, more current will flow through $C_a$ (and of course $C_{ae}$) than through $L_a$, so that the anode circuit as a whole becomes capacitive. The fed-back voltage will then, as we have seen, tend to damp down the signal.

If, on the other hand, the signal frequency (or alternatively $C_a$) is reduced, the majority of current will flow through $L_a$; the anode circuit will be inductive, and the fed-back voltage will augment the signal. That might strike one as a good thing. But the augmented signal is immediately amplified by the valve, more is fed back, and the amplitude increases quite out of control and independently of the original signal that started it. Our amplifier circuit has turned out to be an oscillator circuit; none other, in fact, than the t.a.t.g. oscillator circuit of Fig. 12.9.

That does not necessarily mean that it will actually oscillate. It will not do so unless the voltage fed back via $C_{ag}$ is at least as great as that causing the anode voltage, and is in phase with it. If the amplifier were amplifying to a worth-while extent, the risk would be so great as to amount to a certainty. The better the amplifier, the more likely it is to oscillate. For instance, if $A$ were 50, only one-fiftieth of the anode voltage finding its way back to the grid would be sufficient to cause oscillation.

The effect of continuous oscillation in a receiver is a rushing noise and loud whistles and squeaks as it is tuned. A set which is liable to oscillate is said to be *unstable*.

Seeing that this undesirable condition was brought about by slightly mistuning the anode circuit, one might suppose that the remedy was simply to tune it exactly. But although that would bring the anode circuit into tune with the grid circuit and with the desired signal, at a slightly lower frequency the conditions for oscillation would exist, and any small signal at that frequency—or even the random movements of electrons in circuits—would assuredly start oscillation.

## Neutralization

One method of making the amplifier stable is to feed back a second voltage, equal to the first but opposite in phase. Fig. 17.5 shows how the part of Fig. 17.4 concerned can be modified to do this: by tapping $L_a$ at its centre so that the tuned circuit is symmetrical each side of the supply connection (which, so far as signals are concerned, is an " earth " or no-potential point), and by providing between the top end of $L_a$ and the grid a small capacitor, $C_n$. The signal voltage across the tuned circuit see-saws about the centre-tap, so that the top end is always at the opposite signal potential to the bottom. If $C_n$, called the *neutralizing capacitor*, is adjusted so that its capacitance is equal to $C_{ag}$, the net feedback is reduced to zero or at most something near it.

When triodes were the only amplifying valves available, neutralization was the only effective way of making r.f. amplification possible.

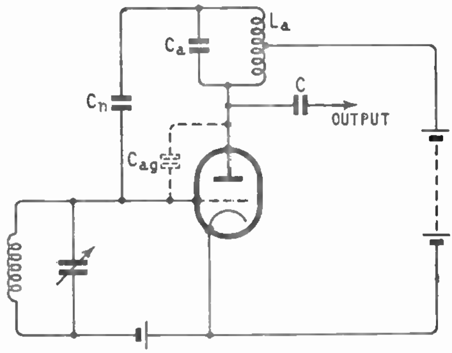



**Fig. 17.5—One way of overcoming the effect of a signal fed back via $C_{ag}$ is to feed back an equal and opposite signal via $C_n$**

As transistors are still in the triode stage of development, neutralization is used in r.f. transistor amplifiers (p. 352). It is still used to some extent in the r.f. amplifying stages of fixed-frequency senders. But in receivers for tuning over a wide range of frequencies it is difficult to obtain a single setting of $C_n$ for them all, and a more satisfactory solution was found which will now be considered.

## Screening

The capacitance between any two objects can be eliminated by placing between them as a screen an earthed metal sheet of sufficient size. The operation of such a screen can be understood by considering Fig. 17.6*a*, in which a stray capacitance is represented by two plates A and B separated from one another by an air-space. The a.c. e.m.f. $E$ will therefore drive a current round the circuit Earth—$E$—A—B—$Z_1$—Earth. Across $Z_1$, which is an impedance of some kind between B and earth, the current will develop a potential difference, and this p.d. will be the voltage appearing on B as a result of current passing through the capacitance AB.

At *b* a third plate S, larger than either of the two original plates, is inserted between them in such a way that no lines of force pass directly from A to B. We now have no direct capacitance between A and B, but we have instead two capacitances, AS and SB, in series.

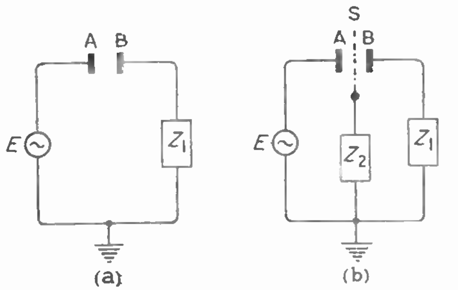



**Fig. 17.6—Illustrating the theory of screening**

If an impedance $Z_2$ is connected between S and earth the current round the circuit Earth—$E$—A—S—$Z_2$—Earth will develop a p.d. across $Z_2$. Since $Z_2$ is also included in the right-hand circuit the voltage across it will drive a current round the circuit Earth—$Z_2$—S—B—$Z_1$—Earth, and this will give rise to a potential on B. So far, S has not screened A from B; there remains an effective capacitance between them which, if $Z_2$ is infinitely large, amounts to the capacitance equivalent to that of AS and SB in series. If S is thin this is practically equal to the original direct capacitance between the two plates.

Now imagine $Z_2$ to be short-circuited. Current will flow round the first circuit, but since there is now no impedance common to both there will be no driving voltage to produce a current in the latter. No matter what alternating voltages are applied to A, none will appear on B, even though large currents may flow via S to earth. The effective capacitance between A and B has therefore been reduced to zero, and B is completely screened from A.

It is very important to note that S is only effective as a screen if it is definitely connected to earth either by a direct wire or through an impedance which is negligibly small.

### The Screened Valve

This is the principle used in reducing the anode-to-grid capacitance of a valve. A screen is put between anode and grid within the bulb, while capacitance between the wires running to grid and anode is avoided by extending the screening down to and below the base, or in some valves by taking one or other of them out through the top of the bulb.

Clearly, a solid metal sheet, while providing perfect screening, would cut off the electron flow from cathode to anode. The screen actually used therefore consists either of metal gauze or a wire grid similar to the first or control grid. The spaces between the wires, while allowing electrons to pass freely, affect the screening remarkably little. In an unscreened valve, $C_{ag}$ is usually of the order of 1 to 4 pF; with a screen as described, this is commonly reduced to less than 0·01 pF, and may even be less than 0·002 pF.

If the screen were earthed by connecting it straight to cathode, it would cut off the attraction of the positive anode; electrons near the grid would then not be drawn onwards, and the anode current would fall practically to zero. But since, as Fig. 17.6 shows, the requirements of screening can be met by making $Z_2$ negligibly small, we can connect a large capacitance between the screen grid and cathode of the valve, and then supply a suitable positive potential to the screen grid.

The portion of the valve comprising cathode, control grid, and screen grid is a triode, and could be used as one, the screen grid acting as an anode. The current through the valve is practically unaffected by whatever voltage may be connected to the real anode, because its electric field is screened off. But if, as is usual, the screen grid is given a potential of the order of + 100 V, the electrons from the cathode are accelerated towards it so rapidly that unless they chance to hit one of its wires they go right through. What happens to them then depends on the potential of the anode. If it is zero or less, practically all the electrons are attracted back to the screen grid, just as a ball thrown up into the air reverses and comes back. But if the anode is substantially positive it attracts and retains those electrons that come near it. Since the number that do so depends mainly on the potential of the screen grid and very little on the potential of the anode, increasing the positive anode voltage beyond a certain amount has very little effect on the anode current.

Thus, the total current through the valve (at a given control-grid voltage) is decided mainly by the screen-grid potential, but the way it is shared between screen-grid and anode depends also on the anode potential.

## Characteristics of a Screened Valve

This can be seen more clearly by looking at Fig. 17.7, which shows anode current plotted against anode voltage in an early type of screened *tetrode* (as a four-electrode valve is called). It would be as well at this point to take note of the standard system of referring to the electrodes, and their voltages and currents. In valves having, like this one, more than one grid, the grids are numbered, beginning with the one nearest the cathode. So the control-grid is called $g_1$ and the screen grid $g_2$. $V_{g2}$ and $I_{g2}$ mean respectively the voltage of the screen grid and the current collected by it; all voltages, as usual, being reckoned relative to the cathode, which is

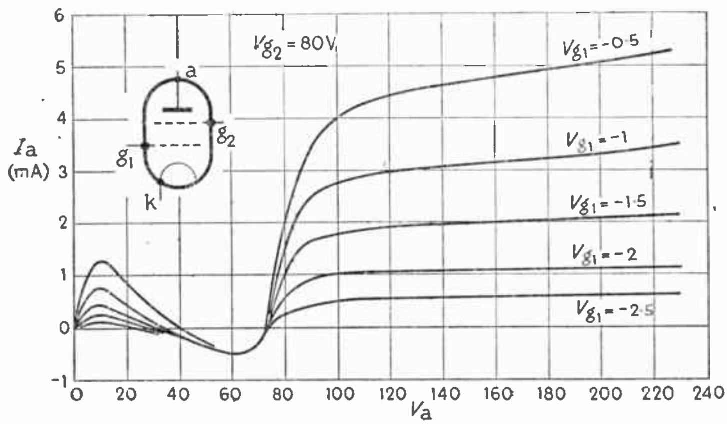



**Fig. 17.7—Characteristic curves of a screened tetrode of the original type, which has now gone out of use**

denoted by k. The standard symbols for the various kinds of valves are shown on p. 375.

Looking now at the curves in Fig. 17.7, we see that the anode current is controlled by $V_{g1}$, just as in a triode. But the shape of each curve is quite different from the corresponding triode curves. Concentrate for the moment on the right-hand portions, where $V_a$ is appreciably greater than $V_{g2}$ (80 V in this case). The anode current increases only very slightly with increasing anode voltage. Another way of putting this is to say that $r_a$ is far higher than in a comparable triode. For the curve $V_{g1} = -1{\cdot}5$, the change of $I_a$ by 0·2 mA for a 100-V change in $V_a$ indicates $r_a = 100/0{\cdot}2 = 500$ kΩ.

But a 1-V change in $V_{g1}$ in this region alters $I_a$ by about 2 mA, so $g_m$ is 2 mA/V. Deriving $\mu$ in the usual way (p. 122), we have $\mu = 500 \times 2 = 1{,}000$. So the introduction of the screen grid not only eliminates most of $C_{ag}$ but also has quite a startling effect on the amplification factor and a.c. resistance. Besides being both much higher, they vary considerably more with the electrode voltages than in a triode, where they are determined mainly by the dimensions and spacing of the electrodes. The mutual conductance, however, is quite normal, since the influence of the control grid is almost unaffected by the division in the structure beyond it.

Although the right-hand portions of the curves in Fig. 17.7 agree with expectations, the left-hand portions certainly do not. What is the explanation of the strange hollow?

## Secondary Emission

If the only effect of raising the anode voltage, beginning at zero, were to rob $g_2$ of more and more electrons, $I_a/V_a$ curves would take forms such as that shown dotted in Fig. 17.8. But when electrons are made to bombard electrodes, they often dislodge electrons already there. This effect is called *secondary emission*. Once liberated, these secondary electrons are free to be attracted by the most positively charged object in the neighbourhood, just as if they had been emitted in the regular way from a cathode. Secondary emission is not noticeable with a triode, because the only electrode to be bombarded violently enough is the anode, and, as it is the most positive electrode around, the secondary electrons all fall back on to it. But the situation is different in a tetrode.

As the anode voltage is raised from zero, the curve of $I_a$ at first follows the dotted line (A). All the secondary emission from $g_2$ returns to it, and $V_a$ is not high enough for secondary emission from the anode. At about point B, anode secondary emission begins; and as $g_2$ is so much more positive it collects it all. So secondary emission is a dead loss to the anode; hence the falling off in anode current. Increasing $V_a$ has, as we know, little effect on the number of electrons arriving, but it increases the emission and hence the number leaving. So we have the paradoxical situation that increasing $V_a$ reduces $I_a$. This is represented by the negative slope between B and C, and by the statement that $r_a$ (over this range of $V_a$) is negative. Such a statement about a real ohmic resistance would, of course, be nonsense, but as we have seen (p. 142), $r_a$ is just a convenient pretence.

Some types of electrode surface emit more than one secondary electron, on the average, for each primary electron, resulting in a negative anode current such as that indicated near C.

Beyond this point, as $V_a$ approaches and then exceeds $V_{g2}$,

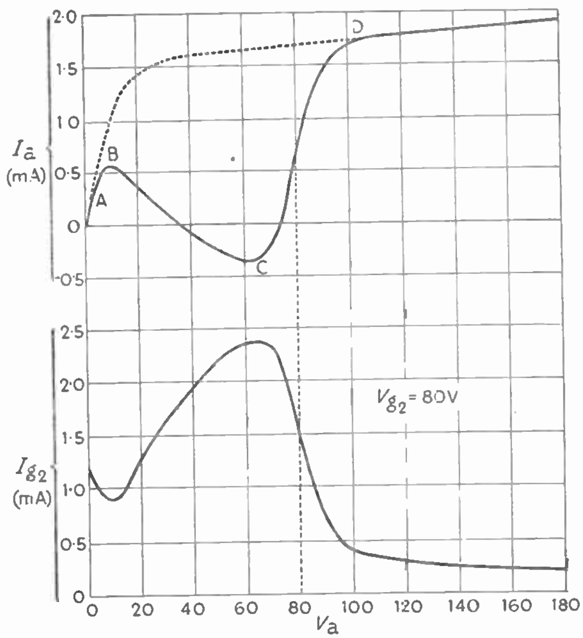



**Fig. 17.8—The dotted line indicates the shape the anode-current curve would take but for secondary emission, which in tetrodes causes the curious departure BC. The corresponding variation in screen-grid current is graphed below**

increasing numbers of secondary electrons return to the anode, until finally none escape to $g_2$ and the valve curve rejoins the dotted line at D. Above this point, secondary emission has no effect on the net anode current.

The shape of the $I_{g2}$ curve, shown below, is like the $I_a$ curve upside down. If $I_a$ and $I_{g2}$ are added together, they come to nearly the same at all anode voltages, confirming the statement that the total " space " current in the valve is scarcely affected by $V_a$. $I_{g2}$, which is wasted, is made as small as possible by careful design of the valve electrodes.

## THE SCREENED PENTODE

The secondary emission effects just described make a big hole in the useful range of anode voltage, for they render the valve unsuitable for amplification unless $V_a$ is at all times substantially greater than $V_{g2}$. Not only is there some inconvenience in having to supply the two electrodes at different voltages, but if the valve is intended for handling a considerable signal voltage one must take care that the negative half-cycles do not bring the anode potential on to the steep slope, where the sudden change in $r_a$ would cause distortion.

Several methods have been found for eliminating the secondary-emission " kink ", giving characteristic curves like the dotted line in Fig. 17.8. The most used of these is to insert between the screen grid and anode a third grid, called (for obvious reasons) the *suppressor grid.* Such a valve, having five electrodes, is called a *pentode.* The suppressor grid, $g_3$, is normally connected to cathode, so is at zero volts, and therefore prevents any secondary electrons from crossing between anode and $g_2$. The faster-moving primary electrons have sufficient momentum to carry them past $g_3$ into the zone of anode attraction.

In many types of pentode, $g_3$ is brought out to a separate connection instead of being connected internally to the cathode, and is thereby made available for certain special uses.

The addition of $g_3$, kept at constant zero potential, tends to reduce still further the influence of the anode voltage on the space current; in other words, the pentode has a higher $r_a$, and consequently a higher $\mu$, than a corresponding tetrode. Since, in a r.f. amplifier, $r_a$ acts as a shunt across the anode tuned circuit, the higher it is the less the damping.

Pentodes thus have several advantages over the type of tetrode described, which is now obsolete: a possible slight improvement in selectivity due to reduced damping; the availability for signal handling of the whole range of anode voltage down to about 20 V; and the simplification of being able to supply $V_a$ from the same source as $V_{g2}$.

## AMPLIFICATION USING SCREENED VALVES

The r.f. amplifier circuit of Fig. 17.4, which we found to be unworkable in practice owing to feedback through the grid-anode capacitance of the triode being sufficient to make the amplifier oscillate, can be converted into a practical amplifier simply by substituting a screened valve for the triode and providing for $g_2$ to be maintained at a voltage recommended by the manufacturer. Some types are designed for $V_{g2}$ to be the same as $V_a$, so that $g_2$ merely needs to be joined to the anode " h.t." supply; others work better with $g_2$ at a lower voltage, which is usually supplied through a resistor as shown in Fig. 17.9. The value of $I_{g2}$ at the specified $V_{g2}$ is stated by the manufacturer, so the calculation of $R_S$ is a simple exercise in Ohm's law. In either case a capacitor $C_S$, having not more than a few ohms reactance at the frequency to be amplified, should be connected as directly as possible between $g_2$ and cathode or earth, to keep the potential of $g_2$ steady.

We have already noted that the voltage gain given by a tuned-anode stage of amplification can be calculated by the same formula as for a resistance coupling, $A = \mu R/(R + r_a)$, $R$ being the dynamic resistance of the tuned circuit, including the input resistance

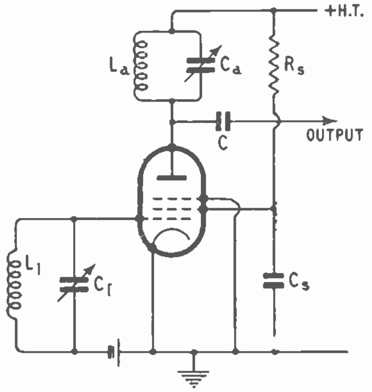



**Fig. 17.9—Modification of Fig. 17.4 to convert it into a workable amplifier circuit by using a screened valve**

of the next stage and any other things that are in parallel. But we have also noted that the $\mu$ and $r_a$ of pentodes vary considerably with the conditions of use, and are both many times higher than for triodes. Typical values are: $\mu = 2{,}000$ and $r_a = 1\ \text{M}\Omega$. The value of $R$ is rarely as much as one-tenth of $r_a$, and is often far less than that; so, taking into account the fact that the value of $r_a$ is unlikely to be known within 10%, it is quite a good enough approximation to neglect $R$ in comparison with $r_a$ and say $A \simeq \mu R/r_a$. But $\mu/r_a$ is $g_m$, so we can simplify the matter still more and say

$$A \simeq g_m R$$

This is much simpler than the original formula: two rather variable valve parameters have been replaced by one ($g_m$) which, for a given valve, is relatively constant. The most convenient units are: $g_m$ in mA/V and $R$ in k$\Omega$.

We see then that provided $r_a$ is many times greater than $R$ (as it almost invariably is in practice) the conditions for a high voltage gain with a screened valve are simply that we choose a valve with a high " slope ", working into a load having high dynamic resistance.

At very high frequencies (30 to 300 Mc/s) tuning circuits with dynamic resistances as high as those used for medium-frequency broadcasting are not practicable. Even if they were, there would be no point in using them; for various reasons they would find themselves shunted by comparatively low impedances. For example, the modulation frequencies used in television run up to several Mc/s; at these frequencies the stray capacitance across the detector load resistance offers such a low-impedance path that the load resistance cannot be more than a few kΩ if the load impedance as a whole is to be reasonably level at all frequencies. If, on the other hand, our r.f. amplifier stage were to be used to supply a second such stage, it would find that owing to the appreciable time (relative to one cycle of this v.h.f.) taken by electrons to cross the space inside the valve its input resistance would be quite low. This effect (called *transit-time effect*) is similar in its results to Miller effect with a capacitive anode circuit. One way or another, then, the overall dynamic resistance $R$ is inevitably low, so to keep up the stage gain in television and other v.h.f. receivers it has been necessary to develop valves with very high $g_m$. Values of 5–12 mA/V are usual in valves designed for television, compared with 1·5–3 mA/V for the lower frequencies.

At these lower radio frequencies it is possible to obtain a very high overall $R$, say 100 kΩ. It might seem that by using this in conjunction with a high-slope pentode a gain of 500 or more could be obtained from a single stage. But it must be remembered that if the signal voltage at the anode is 500 times as great as the grid, it needs only 1/500th of it to work back to the grid, in phase with the signal there, to cause continuous oscillation. Even although the grid-anode capacitance in a screened valve is so small, it would usually be enough. To ensure that the amplifier is always stable, a stage gain much higher than 150 is seldom attempted at radio frequencies.

## External Screening

The achievements of the valve designer in nearly abolishing $C_{ag}$ can easily be nullified by injudicious construction of the rest of the amplifier. If the anode and grid leads from the valve were placed close together for several inches, the capacitance between them would be enough to make any reasonably effective r.f. amplifier unstable. So it is usual to lay them in opposite directions and to screen one or both of them by an earthed metal covering or partition.

Having taken precautions against the amplifier behaving as a t.a.t.g. oscillator, one must also consider it as a possible reaction-coil oscillator (p. 153). There will, of course, be no intentional magnetic (or inductive) coupling, but accidental coupling is difficult to avoid when two tuning coils are mounted in the same unit. We have seen how a very minute capacitive coupling is enough to make a r.f. amplifier unstable, so it is not difficult to imagine that a very small stray induction has the same effect. Most radio apparatus is required to be compact, which means that the coils cannot be spaced far enough apart to avoid risk of instability. The larger the coils, the greater the risk, for two reasons: the magnetic field spreads out more, and the dynamic resistance tends to be higher.

If a coil is enclosed in a metal box or can, constructed so as to give a continuous low-resistance path parallel to the coil winding, the metal is in effect a transformer secondary winding with a single short-circuited turn. The currents induced in it by the tuning coil set up a magnetic field which opposes that due to the coil, so the resultant field outside this screen is greatly reduced.

If the coil is surrounded closely by the screen, its own self-inductance is largely neutralized, which means that more turns must be used to tune to a given frequency, and the r.f. resistance is greater. So it is desirable for the diameter of the screening can to be not less than twice that of the coil; and even then some allowance must be made in the design of the coil for reduction of inductance by the screen. The need for this is lessened by using an iron or ferrite core, which prevents much of the magnetic field from spreading outside.

Comparing magnetic (or inductive) screening with electric (or capacitive) screening, note that it necessitates low-resistance paths parallel to the coil winding but it need not be earthed, whereas capacitive screening need not provide closed circuits but must be kept at some constant potential. Coils need both kinds of screening; the capacitance between them would be enough to by-pass the internal valve screening, quite apart from inductive coupling. Comprehensive screening is provided by enclosing them completely in metal cans in contact with the earthed part of the circuit.

## The Aerial Coupling

We have already considered the problem of coupling a sender to its aerial (p. 171); now we have to couple the receiving aerial to the receiver. The same basic principles apply, and the various methods shown in Fig. 13.10 are used, but the aims are rather different. The main object with the sender, you may recall, was to get the maximum power into the aerial; and the purpose of the coupling was to transform the actual impedance of the aerial into the optimum load for the output valve.

At the receiving end we have the aerial picking up a small signal voltage and therefore acting as the generator. It also has a certain impedance. The load consists of the input tuning circuit seen in the circuit diagrams of r.f. amplifiers in this chapter, together with the input impedance of the valve. There is the valve's input capacitance ($C_{gk}$), usually 3–8 pF. We can assume that the flow of grid current is prevented by negative bias. But we have seen how Miller effect can create the equivalent of capacitance and positive or negative conductance across the input to the valve; this extra capacitance further alters the setting of the tuning capacitor needed to establish resonance, and the conductance decreases or increases the dynamic resistance of the resonant tuned circuit. Using a screened valve greatly reduces Miller effect but does not abolish it entirely. At low and medium r.f. (say, not more than a few Mc/s) the input impedance is not usually low enough for it to have a controlling influence on the tuning circuit. At very high frequencies, however, there is not only the likelihood of greater Miller effect (because the residual stray capacitance from the anode represents a lower impedance) but there is the transit-time effect just mentioned, and a similar effect due to cathode lead inductance, which combine to make the valve look to the tuning circuit like a comparatively low resistance.

The impedance of the aerial depends on the type used. There are the three main classes considered in Chapter 15: (1) the dipole; (2) the earthed aerial; and (3) the loop or coil. The purpose of the coupling from aerial to input tuned circuit is to convey as much of the desired signal as possible without unduly upsetting the tuning and selectivity. We know (p. 145) that the greatest power is transferred to a load from a generator when the load resistance is, from the generator's point of view, equal to the generator's resistance; which is the same as when the generator resistance is, from the load's point of view, equal to the load resistance. This is on the assumption that neither load nor generator has reactance. If it has, then the matching process includes tuning it out by an equal quantity of the opposite kind.

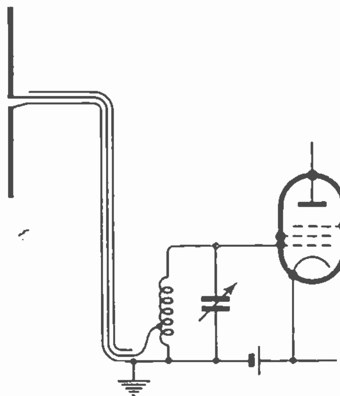


Fig. 17.10—Although it does not provide for balancing to earth, this simple method of connecting a dipole aerial to the receiver by means of a coaxial feeder is often used

The problem is simplified if the impedance of the aerial to be used is known. For v.h.f. (television and f.m. broadcasting, for example) the aerial is usually some kind of dipole connected via a cable having a characteristic resistance of about 70 Ω. The band of frequencies is usually not so wide that the input impedance of tuned circuit and valve varies enormously, and an average figure should be ascertainable; say 7 kΩ. In this example the resistance step-up is 1 : 100, so a 1 : 10 turns ratio should give correct matching (p. 98). Fig. 17.10 shows the simplest way in which this might be accomplished. In practice there are many variations of method: a separate primary winding may be used (especially if the cable is balanced about earth as centre); tuning may be done by moving the core of the coil instead of by capacitor; and a different kind of valve connection altogether (p. 276) may be adopted.

The characteristics of the aerial are also known if it is an inductor built into the receiver. The tuning inductor and the aerial are then usually one and the same, so no question of coupling arises. In receivers for accurate direction finding, however, the matter is less simple and is in fact a highly specialized subject.

The most awkward problem is when the characteristics of the aerial to be provided for are unknown, as in broadcast receivers for frequencies lower than v.h.f., and " communication " receivers to work over a very wide range of frequency. The aerial may be anything from a few feet of wire indoors to a large outdoor erection. As an impedance, it is roughly equivalent to a capacitance of about 20–1,000 pF in series with a resistance of perhaps 50–5,000 Ω—somewhat indefinite!

Supposing, for example, it was 200 pF and 400 Ω, the aerial could be represented as a signal generator connected through them to the receiver aerial and earth terminals, as in Fig. 17.11. At a frequency of 1 Mc/s the tuning circuit might well have the values shown, in which the 100 kΩ represents its dynamic resistance including the effect of the valve. The question is how to connect the aerial.

Ideally, the capacitive reactance would be cancelled by putting an equal inductive resistance in series, and the aerial resistance would be matched by a 1 : $\sqrt{(100/0{\cdot}4)}$ step-up between an aerial

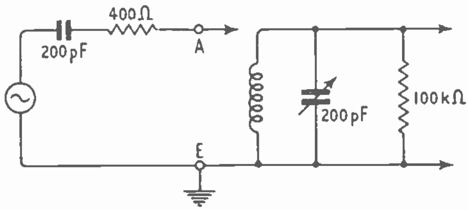



**Fig. 17.11—A example of how connecting an aerial directly across the tuning circuit may have very bad effects. The components to the left of the A and E terminals are the circuit equivalent of the aerial**

primary winding and the tuning coil. Both the series inductance and the ratio would have to be adjusted for every frequency, which is out of the question. In practice the aerial reactance has to be tolerated.

We noted (p. 114) that for any reactance and resistance in series one can calculate a reactance and resistance in parallel, to which they are equivalent. At 1 Mc/s, 200 pF and 400 Ω in series are equivalent to 160 pF and 2 kΩ. If, therefore, the aerial were connected straight to the top end of the tuning circuit, its dynamic resistance would be brought down from 100 kΩ to 100 kΩ and 2 kΩ in parallel: just under 2 kΩ. So the $Q$ of the circuit (which is a measure of its signal magnification and selectivity) would be reduced to one fiftieth! And the 200 pF tuning capacitance would have to be reduced to 40 pF to re-establish resonance. There are always other circuits in the receiver to be tuned simultaneously, and for convenience all the tuning capacitors are controlled by a single knob and cannot be adjusted separately to allow for the effect of the aerial, which varies with frequency. So the capacitance change due to connecting the aerial must be kept quite small. Direct connection in this typical case therefore can be ruled out completely on this ground alone, to say nothing of the loss of signal and selectivity.

By tapping the aerial lower down the coil, or to a primary winding with fewer turns, the equivalent aerial resistance across the whole coil is raised, increasing the overall $Q$; and the capacitance is reduced. The signal voltage reaches a maximum at a certain ratio, then falls off; but selectivity and tuning disturbance continue to improve. It is better to under-couple rather than over-couple; although signal is lost either way, over-coupling is bad for selectivity and tuning, whereas under-coupling is good for both. In practice, under-coupling is essential, because coupling that is optimum for signal strength disturbs the tuning far too much.

The larger the aerial, the greater the signal it brings, and the greater its capacitance, so the looser the coupling should be. A single degree of coupling must be very much of a compromise, unlikely to give anything like best results with extremely large or small aerials. So often two or more alternative aerial terminals are provided.

The system of connection just discussed—a step-up transformer or auto-transformer—has the advantage that the signal loss due to

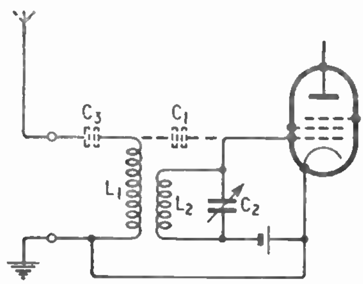



Fig. 17.12—Methods of adapting an aerial to the receiver

tapping low enough to avoid excessive de-tuning is partly compensated by the voltage step-up. In another system, Fig. 17.12, the primary ($L_1$) of a r.f. transformer is designed so that with its own stray capacitance it resonates at a frequency just below the lowest in the range of the receiver. The larger the aerial, the more the resonant frequency of this winding departs from that of the secondary and so the less it disturbs it. The greater signal loss is compensated by the more that the larger aerial brings in. Signal transfer naturally tends to be poor at the highest frequencies, at which $L_2 C_2$ is most out of tune with the aerial, so a small capacitor $C_1$ is sometimes provided to help things along. To limit to a known amount the capacitance that can be put into the circuit by an aerial, a capacitor $C_3$ is often connected in series with either the only or an alternative aerial connection.

## Two Stages of Amplification

For receiving distant stations, the gain—and the selectivity—of a single stage of amplification is insufficient. Each stage that is

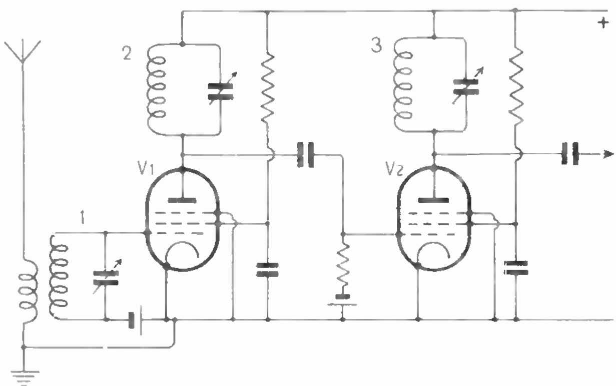



**Fig. 17.13—Two-stage r.f. amplifier, using tuned-anode couplings**

added accentuates the problem of keeping the amplifier stable. Fig. 17.13 shows how two tuned-anode r.f. stages would be connected one after the other, or *in cascade* as it is called. Examining it, we see that tuned circuit No. 2, besides tuning the anode of $V_1$, also tunes the grid of $V_2$, for it is connected through a capacitance of negligible impedance to the grid, and through the h.t. source—or more usually through a by-pass capacitor not shown—to the cathode. In its role as grid tuner for $V_2$, it is liable to have positive feedback from circuit No. 3, increasing its dynamic resistance above normal. This in turn may increase the risk of feedback to No. 1 leading to instability.

Because of the greatly increased total gain, the direct stray capacitance from output to input needed to render the amplifier unstable is correspondingly smaller; and so is the stray magnetic coupling. A high standard of screening is therefore needed.

As we shall see in the next chapter, greater selectivity is needed than can be provided by single tuned circuits between stages, so separately-tuned transformer coils are almost invariably used in place of them. This incidentally somewhat eases the stability problem. But because a different technique is usually adopted for obtaining a large amount of amplification and selectivity before the detector, it is not worth while studying the details of multi-stage r.f. amplifiers just now.

CHAPTER 18

# Selectivity

## SELECTIVITY AND *Q*

THE REASON GIVEN for using r.f. amplification was the need for *sensitivity*; that is to say, smallness of signal input sufficient to yield the required output from the receiver. A set with high sensitivity may therefore be expected to have a good range of reception. There have also been several incidental references to *selectivity*, which means ability to exclude unwanted signals. The need for this was mentioned as far back as p. 103, when the usefulness of tuned circuits was explained. Without it, sensitivity would be a positive embarrassment.

We should have gathered by now that both sensitivity and selectivity are improved by increasing the $Q$ of a tuned circuit. Graphs such as Fig. 8.6 show how response at the frequency to which a circuit is tuned is increased in direct proportion to $Q$, whereas the response at other frequencies is less affected; in fact, at widely different frequencies it hardly changes at all. The problem of improving reception seems therefore to resolve itself into a matter of increasing $Q$. As we saw in Chapter 8, the $Q$ of a circuit is the ratio of its reactance (either inductive or capacitive) to its series resistance, which is the same thing as the ratio of its dynamic resistance to its reactance:

$$Q = \frac{X}{r} = \frac{R}{X}$$

Now the amount of reactance is decided by the requirements of tuning, so attention must be confined to reducing the series resistance, $r$. This will have the effect of increasing the dynamic resistance, $R$. We noted (p. 112) that $r$ is a sort of hold-all figure including everything that causes energy loss from the tuned circuit; not only the resistance of the coil of wire, but loss in the solid core (if any), in adjacent metal such as screens, and in dielectrics associated with the circuit. These can all be reduced by careful choice of materials and careful design, but size and cost limit what can be done, so the result is a $Q$ seldom very much better than about 200.

## USE OF FEEDBACK

In Chapter 12 we saw how the loss that damps out oscillations in tuned circuits can be completely neutralized by energy fed to it at the right moment in each cycle by a valve, so that oscillations once started continue indefinitely. This would not do in a receiver, because we want the strength of oscillation in the tuning circuit to depend all the time on what is coming from the distant sender.

But if the coupling from the anode circuit of the valve to the grid is reduced, the amount of energy fed back will *partly* neutralize the losses. Since we are already using the symbol $r$ to include not only resistance in its strictest sense but all other causes of loss, there is no reason why we should not also make it take account of a source of energy *gain*. In this sense, then, positive feedback reduces $r$, which is precisely what we want.

We can put this to the test, using our original valve oscillator circuit (Fig. 12.5), by increasing the distance between the two coils just sufficiently to make the continuous oscillation stop. The tuned anode circuit will then behave exactly as if its resistance were abnormally low; any small voltage injected in series with it at its resonant frequency is magnified much more than if the valve were not working. The same result is obtained if the tuned circuit and the reaction coil are interchanged; in this way a grid tuning circuit likewise can be given a sharper resonance.

In Fig. 18.1 are plotted some of the resonance curves that might be obtained, using different amounts of "reaction" or positive feedback. The "$Q = 40$" curve represents a rather poor circuit with no feedback, and the others with progressively increased amounts. To raise the $Q$ from 40 to 8,000 necessitates very careful

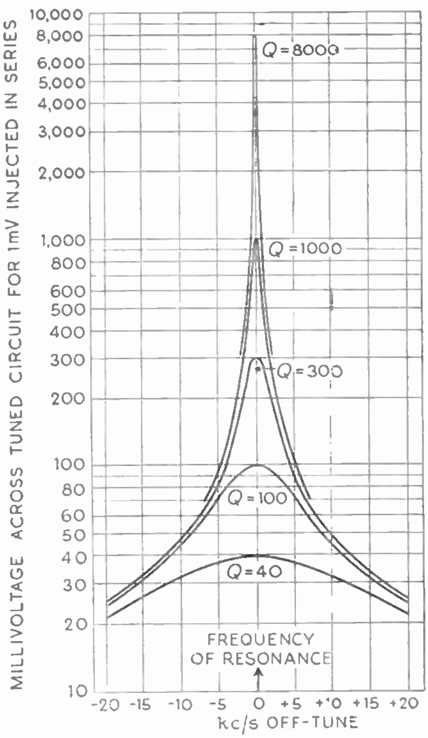



**Fig. 18.1—Showing how r.f. voltage across a tuned circuit varies with frequency at different *Q* values, obtained by adjusting positive feedback**

adjustment of the coupling, as it is then very near the self-oscillation point, and a slight change in supply voltage—h.t. or l.t.—would either carry it past this point or in the opposite direction where $Q$ would be substantially less. In other words, the adjustment of coupling is extremely critical. In practice it would be difficult to maintain the $Q$ close to such a high value for long, and one would be wise to be content with rather less. Even so, the benefit that can be obtained by feedback is so great that it could not be shown on a graph like those that have been used hitherto; Fig. 18.1 has therefore been drawn to a logarithmic voltage scale, so compared with Fig. 8.6 the increase is much greater than it looks.

Before screened valves were available to provide effective r.f. amplification, much use was made of controllable feedback. In the effort to obtain the utmost from it the oscillation point was often crossed, and the resulting oscillation—radiated from the receiving aerial—caused serious interference with neighbouring receivers. Provided that the control was always kept on the right side of this point, however, it might be thought that the nearer it was to it the better, for both sensitivity and selectivity would be at their best.

If only a simple carrier wave were being received, that would be true, but quite useless. It is the modulation of the carrier that conveys the desired information (using that word in its widest sense). If the carrier is frequency-modulated (p. 169) it is obvious that a definite band of frequencies, with the carrier frequency at its centre, must be received to an equal degree. The need for covering a band of frequencies when the carrier is amplitude-modulated is less obvious, but it none the less exists.

## The Theory of Sidebands

Strictly speaking, it is only an exactly recurrent phenomenon that can be said to possess a definite frequency. The continuous change in amplitude of a carrier wave during amplitude modulation makes the r.f. cycle of the modulated wave non-recurrent, so that in acquiring its amplitude variations it has lost its constancy of frequency.

A mathematical analysis shows that if a carrier of $f_1$ c/s is modulated at a frequency $f_2$ c/s the modulated wave is exactly the same as the result of adding together three separate waves of frequencies $f_1$, $(f_1 + f_2)$, and $(f_1 - f_2)$. It is not easy to perform the analysis of the modulated wave into its three components by a graphical process, but the corresponding synthesis, adding together three separate waves, demands little more than patience.

Fig. 8.2 shows at *a*, *b*, and *c* three separate sine-wave trains, there being 35, 30, and 25 complete cycles, respectively, in the width of the diagram. By adding the heights of these curves above and below their own zero lines point by point, the composite waveform at *d* is obtained. It has 30 peaks of varying amplitude, and the amplitude rises and falls five times in the period of time represented. If this is a thousandth part of a second, curve *d* represents what we have come to know as a 30 kc/s carrier amplitude-modulated at 5 kc/s. Fig. 18.2 shows it to be identical with three constant-amplitude wave-trains having frequencies of 30, 30 + 5, and 30 − 5 kc/s.

Thus a carrier modulated at a single frequency is equivalent to three simultaneous signals: the unmodulated carrier itself and two other steady frequencies spaced from the carrier on each side by an amount equal to the frequency of modulation. In a

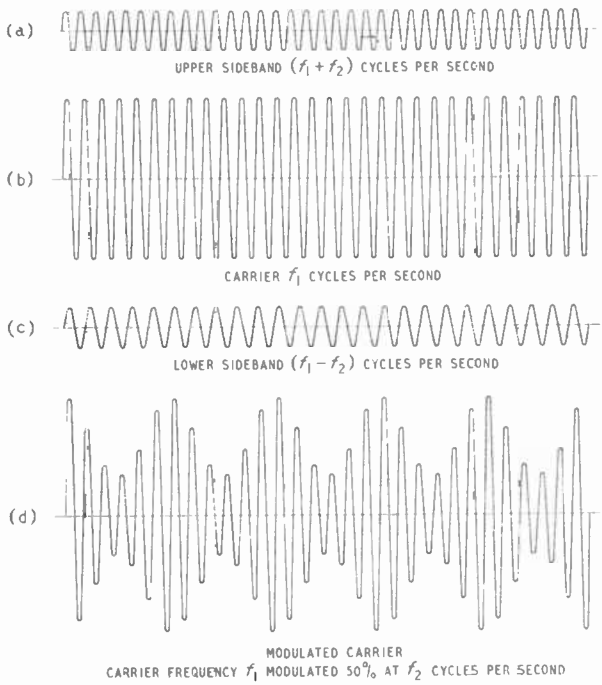



**Fig. 18.2—Showing that a modulated carrier wave (*d*) is identical with three unmodulated waves (*a*, *b* and *c*) added together**

musical programme, in which a number of modulation frequencies are simultaneously present, the carrier is surrounded by a whole family of extra frequencies. Those representing the lowest musical notes are close to the carrier on each side, those bringing the middle notes are farther out, and the highest notes are the farthest removed from the carrier frequency. This spectrum of associated frequencies on each side of the carrier is called a *sideband*, and as a result of their presence a musical programme, nominally transmitted on a frequency of, say, 1,000 kc/s, spreads over a band of frequencies extending from about 990 to 1,010 kc/s.

The same facts can be illustrated by a vector diagram. In such a diagram, as we know (p. 76), a single unmodulated carrier-wave of frequency $f$ is represented by a vector such as OP in Fig. 5.5, pivoted at one end (O) and rotating $f$ times per second. The length of the vector represents (to some convenient scale) the peak voltage or current, and the vertical " projection " PQ represents the instantaneous value, which of course becomes alternately positive and negative once in each cycle.

If this were related to a waveform diagram such as Fig. 18.2, it should be clear that amplitude modulation would be represented by a gradual alternate lengthening and shortening of the vector as

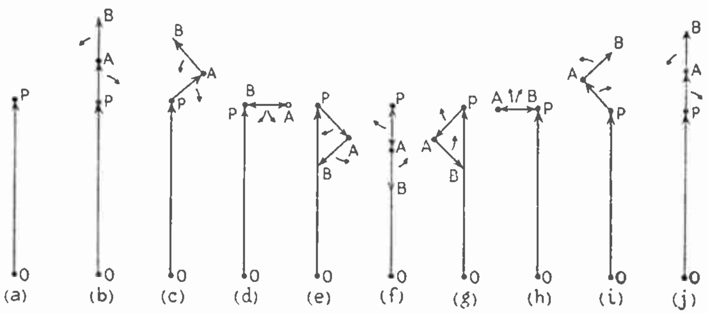



**Fig. 18.3—Progressive series of vector diagrams, showing how adding a pair of oppositely-rotating vectors (PA and AB) alternately lengthens and shortens the original vector (OP) without affecting its phase**

it rotated. Thus, modulation of a 600 kc/s carrier to a depth of 50% at a frequency of 1 kc/s would be represented by a vector revolving 600,000 times per second, and at the same time increasing and decreasing in length by 50% once during each 600 revolutions.

The attempt to visualize this rapid rotation makes one quite dizzy, so once we have granted the correctness of the vector representation we may forget about the rotation and focus our entire attention on how the alternate growing and dwindling can be brought about by vectorial means.

OP in Fig. 18.3*a*, then, is the original unmodulated-carrier vector, now stationary. Add to it two other vectors, each one quarter the length of OP. If they are both in the same direction as OP, they increase its length by 50%; if in the opposite direction they decrease it by 50%. The intermediate stages, corresponding to sinusoidal modulation, are represented by the two small vectors rotating *in opposite directions* at modulation frequency. Various stages in one cycle of modulation are shown at Fig. 18.3*b–j*, the total length of all three vectors added together being OB in every case. Notice that in spite of the small vectors going through complete 360° cycles of phase relative to OP, the triple combination OB is always exactly in phase with OP. It therefore gives the true lengthening and shortening effect we wanted.

Having agreed about this, we can now set OP rotating at 600,000 r.p.s. again. In order to maintain the cycle of changes shown in Fig. 18.3 the little vector AB will therefore have to rotate at 600,000 r.p.s. plus once in every 600; that is, 601,000: while PA must rotate at 600,000 less once in every 600; 599,000.

Translated back into electrical terms, this confirms the statement that amplitude-modulating a carrier-wave of $f_1$ c/s at a frequency of $f_2$ c/s creates side-frequencies, $f_1 + f_2$ and $f_1 - f_2$. And Fig. 18.3 also shows that the amplitude of each of these side-waves relative to that of the carrier is *half* the modulation depth (p. 167).

## Over-Sharp Tuning

We now have a direct relationship between the selectivity of a tuned circuit and its ability to receive the highest notes likely to be present as modulation on the carrier. If the resonance curve of the circuit is not substantially flat over a central portion wide enough to include the whole of the required sidebands, high notes will be attenuated—they will be quite literally tuned out owing to over-selectivity. In the curve for $Q = 8{,}000$, in Fig. 18.1, the sidebands corresponding to a modulation frequency of 5 kc/s are shown as being magnified only about 1·3% as much as the central carrier frequency. Lower notes are more fully preserved; higher notes even more greatly attenuated. The result is " woolly " and more or less unintelligible speech, and " boomy " music. For a tuned circuit in which $Q = 100$, however, 5 kc/s notes are passed at 70% of the carrier amplitude.

With frequency modulation there is the same need to preserve sidebands, but the result of failing to do so is different. In broadcasting, the deviation (p. 169) is usually 75 kc/s each side of the carrier frequency; this represents 100% modulation, and for perfect reproduction it is necessary for the receiver to respond equally over a total band of at least 150 kc/s.* If the band accepted is less than this, then reception of less modulation depth may be satisfactory but deep modulation is distorted, almost regardless of whether the audible notes are high or low. The " almost " must be said, because the maximum width of the sidebands depends not only on the deviation but also in a complicated way on the modulation frequency. When the deviation is so much greater than the highest modulation frequency as is 75 kc/s, the additional bandwidth due to even the highest modulation frequency is negligible. But if the deviation were, say, 1 kc/s, this would not mean that a 10 kc/s note could be transmitted with sidebands only up to 1 kc/s: they would actually extend to about 11 kc/s.

It is clear from these considerations that high selectivity is not the unmixed blessing it appeared at first. This discovery that the sharpness of resonance that can be employed is strictly limited if

* Although the limiter (p. 224) helps to iron out inequalities in response, it could not cope satisfactorily with severe "sideband cutting".

the desired modulation (by frequency or by amplitude) is not to be partly lost and thereby distorted, immediately raises more problems. The restriction on selectivity restricts at the same time the amount of signal amplification by circuit resonance alone, making it necessary to rely for high sensitivity on valve amplification. But we saw in the last chapter that r.f. signals cannot be amplified effectively unless the coupling circuits are tuned. Will not the number of tuned circuits required re-introduce the same loss of sidebands that we are trying to avoid by banning a single very sharp circuit? In any case, how can the need for level response over the band of desired frequencies be reconciled with effective selectivity? We shall proceed to tackle both of these questions, but to appreciate what is involved in the second of them let us consider how the frequencies of different senders are spaced.

## Channel Separation

The "spread" of the sidebands on each side of the carrier wave is determined (with a.m.) by the frequency or (with f.m.) by the depth of the modulation, and not at all by the frequency of the carrier wave. In broadcasting sound, the modulation frequencies are the audible frequencies, say 30 to 15,000 c/s; so

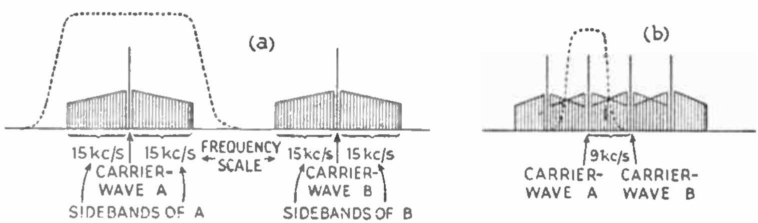



**Fig. 18.4—At (*a*) is shown how broadcasting stations' carrier waves ought to be spaced on the frequency scale in order to permit full reception of each without interference from others. How they actually are spaced (*b*), shows that interference is almost inevitable if full reception of sidebands is attempted**

with a.m. the sidebands might spread up to 15 kc/s on each side of the carrier-wave frequency. Ideally, then, the carrier waves of different broadcasting stations should be spaced at intervals of at least 30 kc/s, plus a margin for selectivity, as suggested in Fig. 18.4*a*, where the dotted line is an ideal overall receiver response curve, embracing the whole of transmission A and completely rejecting transmission B.

Unfortunately the pressure of national demands for broadcasting "channels" has squeezed this ideal severely; the standard interval in Europe on the low and medium frequencies is only 9 kc/s. Obviously this faces receivers with an impossible task, for it brings 9 kc/s modulation of one station on to the same frequency as the carrier wave of the next and the 9 kc/s modulation of the next but one. Therefore few, if any, broadcasting stations on these frequencies actually radiate sidebands up to 15 kc/s; 8 kc/s is a more practical maximum. There is therefore some loss at the start. And even what is transmitted cannot be fully reproduced without interference, unless the transmission is so much stronger than those on adjacent frequencies as to drown them. So receivers for other than local-station reception must be designed to cut off at about 4·5 kc/s, sacrificing quality of reproduction in the interests of selectivity (Fig. 18.4*b*), and even then liable to some interference from adjacent-channel sidebands if they are strong enough. At very high frequencies there is less overcrowding; so sound transmitted on those frequencies (such as f.m. and television sound) is not subject to the same degradation of quality.

An important point is that the spacing of carrier waves is necessarily on a frequency basis, not a wavelength basis. If a station transmits on 200 kc/s, the two nearest in frequency will be on 191 and 209 kc/s. The wavelengths corresponding to these frequencies (p. 22) are 1,571, 1,500 and 1,435 metres, so the average wavelength spacing is 68 metres. Three consecutive carrier-wave frequencies higher in the scale are, according to the same 9 kc/s spacing, 1,502, 1,511 and 1,520 kc/s. The corresponding wavelengths are 199·7, 198·6, and 197·3 metres, so the wavelength spacing here is only 1·2 metres.

Whether we are concerned with programme separation or loss of sidebands, then, the basis of reckoning is "frequency off-tune". To distinguish this particular frequency from frequency in general, we shall denote it by $f'$.

## A Universal Resonance Curve

When we first considered resonance curves we found that they could be more easily compared for selectivity if the sensitivity element was removed by adjusting the input to each circuit so as to bring all their resonance peaks to the same level, as was done in turning Fig. 8.6 into Fig. 8.7. The important quantity is then not how much the curves rise above some arbitrary zero level but how much they descend below their common peak level at a given frequency off tune. So it will be more logical and helpful to make our response scales begin at peak level and work downwards. The appropriate measure of response is a decibel scale (p. 146) in which all the numbers—since they represent reductions relative to "in-tune"—are negative. For the sake of readers who are not yet used to thinking in decibels, alternative scales will show the corresponding number of times by which the response at a given $f'$ is less than at resonance. For instance, instead of saying that the level of the "$Q = 8{,}000$" curve in Fig. 18.1 at 10 kc/s off tune is 0·0067 relative to that at resonance we will say it is 1/0·0067=150 "times down" at that frequency. For convenience we shall call the numbers on this scale "$S$" values (since they concern selectivity). Because the decibel scale is uniformly divided, the corresponding $S$ scale must be logarithmic.

We have seen that sharpness of resonance, which can be reckoned as the value of $S$ at a given $f'$, depends very largely on $Q$. Does it depend on anything else? If we were to plot resonance curves for a number of tuned circuits having the same $Q$ we would find that they would not be identical unless they resonated at the same frequency. Further experiment would show that to make the $S/f'$ curves identical it is necessary for their $Q$s to be exactly proportional to their frequencies of resonance. For example, if one circuit resonated at 500 kc/s and had a $Q$ of 100, and another resonated at 1,000 kc/s it would have exactly the same $S/f'$ curve if its $Q$ was 200. So if in every case we were to divide $Q$ by the

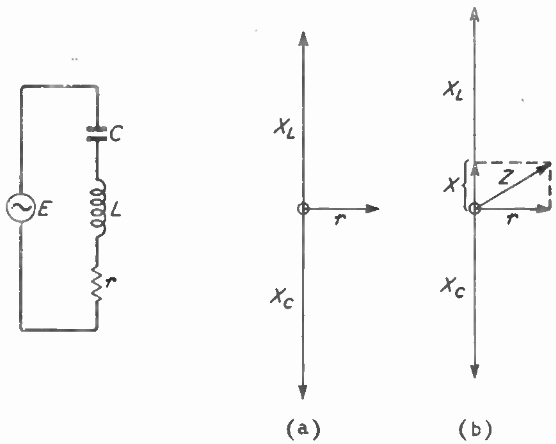



**Fig. 18.5**—Vector diagrams of a tuned circuit, (*a*) exactly at resonance, and (*b*) off resonance, showing how the total impedance increases from $r$ to $z$

resonant frequency, $f_r$, we would arrive at a factor defining the sharpness of resonance as seen on a graph of $S$ plotted against $f'$. Since in a simple tuned circuit $Q = 2\pi f_r L/r$ (p. 102), this factor simplifies to $2\pi L/r$. We can further simplify it by knocking out the $2\pi$ because that is just a constant number. So the selectivity factor is just $L/r$, the ratio of tuning-coil inductance to series r.f. resistance.

That this is so can be seen more definitely by drawing the vector diagram for the impedance of a tuned circuit. In Fig. 18.5*a*, $X_L$ and $X_C$, the inductive and capacitive reactances, are equal, so cancel one another out, leaving $r$ as the sole impedance; this represents the state of resonance. Now suppose the frequency is increased by $f'$ c/s, as shown in Fig. 18.5*b*. $X_L$ increases in direct proportion, from $2\pi f_r L$ to $2\pi(f_r + f')L$; so the increase is $2\pi f'L$. Provided that $f'$ is small compared with $f_r$, the reduction in $X_C$ is very nearly the same. So the total reactance, $X$, is practically equal to $4\pi f'L$. The diagram shows that it is the ratio of this to $r$ that decides how much the current through the circuit falls off when the frequency is detuned by $f'$ c/s. Again, $4\pi$ is a constant number. If, therefore, we plot $S$ against $f'L/r$ we get a resonance curve that holds good for all tuning circuits of this kind, regardless of the values of $L$, $C$ and $r$. When we know $L$ and $r$ for any particular circuit we can fill these values in, and their ratio adjusts the $f'L/r$ scale to fit that circuit and enable us to read off $S$ corresponding to any $f'$.

This will be clearer when we actually have the curve to refer to, but before constructing it we might as well incorporate two other helpful features. It has no doubt been noticed that resonance curves appear to be symmetrical. If we plotted them far enough each side of resonance we would find that they are not perfectly so, but since we are assuming $f'$ is considerably smaller than $f_r$ we are justified in neglecting this discrepancy and plotting only one half of the curve, regarding $f'$ as either $+$ or $-$.

The other idea is to make the horizontal scale logarithmic as well as the vertical, because this gives greatest space to the part of the curve nearest resonance. It also allows us to continue the lower part of the curve as a straight line.

Fig. 18.6 is the result of all this cogitation—a resonance curve applicable to *any* simple tuning circuit, and adapted for convenient use. If $f'L/r$ is divided by $L/r$ it becomes simply $f'$, so all we have to do to convert the horizontal scale to an $f'$ scale is to divide all the numbers by $L/r$ (or, what is the same thing, by $Q/2\pi f_r$), taking care always to use the units specified. For instance; if $L$ is 150 $\mu$H and $r$ is 24 $\Omega$, $L/r$ is 6·2, so 75 on the $f'L/r$ scale is 12 kc/s off-tune, the corresponding $S$ being nearly 1·4, or roughly 3 dB.

The curve can be used to find the resistance of a coil to give a certain selectivity, say $-20$ dB at 40 kc/s off-tune. The curve shows that at $-20$ dB $f'L/r$ is about 800; so at $f' = 40$, $L/r$ is 20. If $L$ is 5,000 $\mu$H, then $r$ should be 250 $\Omega$. The loss of 7·5 kc/s side frequency due to such a coil is found by looking up $S$ for $f'L/r = 7{\cdot}5 \times 20 = 150$, namely about 7 dB.

It will have become abundantly clear by now that the sort of response curve that is really needed is one which remains level at $S = 1$ as far as the highest sideband frequency required, and then cuts steeply off to reject other transmissions. But what we have found is that the shape of the resonance curve of a single tuned circuit, drawn on a diagram of the most appropriate kind, is fixed by Nature and cannot be altered. All we can do is slide the curve horizontally along the $f'$ scale. Once we have decided on the response at any one off-tune frequency, the response at all others is thereby decided too, whether we like it or not.

## More Than One Tuned Circuit

Let us now consider what happens when additional tuned circuits are used. If we can assume that these circuits do not react on one another, then it is simple. If the grid circuit of a r.f. amplifier

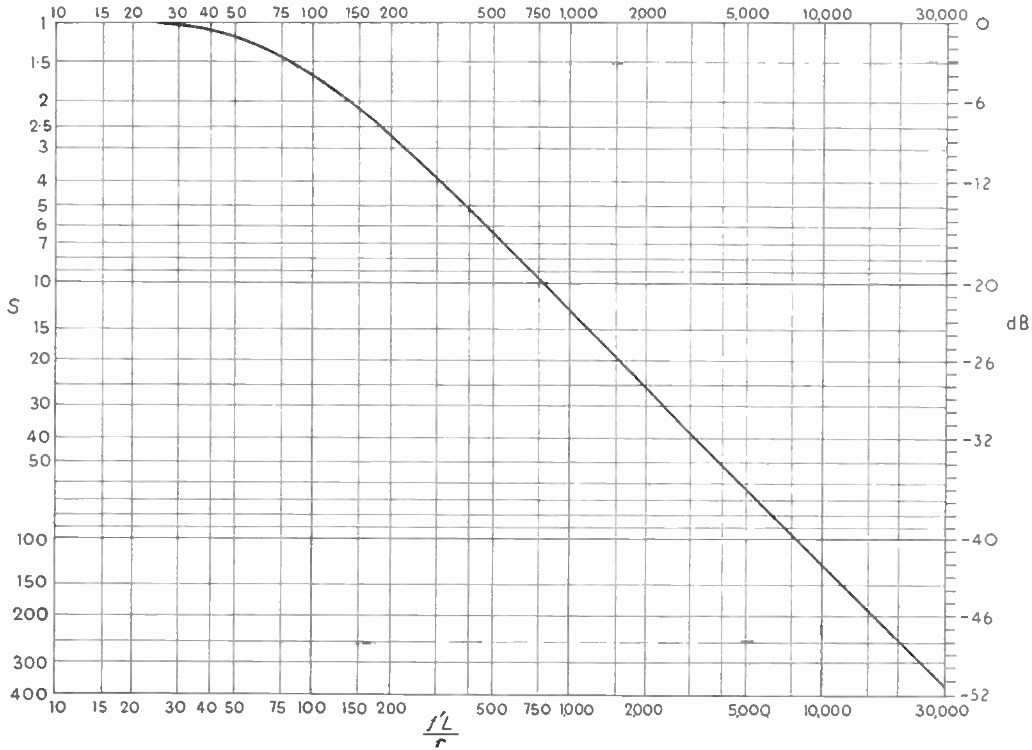



**Fig. 18.6—Generalized resonance curve for any single tuned circuit. To make it refer to any particular circuit, divide the numbers on the horizontal scale by $L/r$ (or by $Q/2\pi f_r$) to convert them to a scale of $f'$. The units are: $L$ in $\mu$H ; $r$ in $\Omega$; $f'$ in kc/s; $f_r$ in Mc/s**

reduces signals at a certain off-tune frequency 3 times (relative to the response at resonance), and the anode circuit is somewhat more lightly damped and reduces them 4 times, then obviously the total relative reduction is 12 times. If there is a second r.f. stage, and its anode circuit gives a further reduction of 3 times, the total is 36. In other words, this 2-stage amplifier amplifies this certain frequency 36 times less than it amplifies the frequency to which all three circuits are tuned. In decibels, the three circuits cause reductions of 9·5, 12 and 9·5 dB; the total is found by simple addition: 31 dB.

If the problem is simplified by assuming all the circuits have the same $L/r$, then the curve in Fig. 18.6 is made to apply to a pair of circuits by squaring $S$; to three circuits by cubing it; and so on. Or by multiplying the dB scale by 2, 3, etc.

Comparison between one and several tuned circuits can be made in different ways. First of all we shall compare them on the basis of equal gain, in order to dispose finally of the question of feedback versus r.f. amplification. Taking some typical figures, suppose that the r.f. amplifier is a single stage giving a gain of 50 times, with input and output tuned circuits for which $L/r = 10$. Squaring the $S$ values in Fig. 18.6, we get curve $a$ in Fig. 18.7. To obtain equal gain by means of feedback, it would be necessary to apply enough of it to multiply $Q$ by 50. This would multiply $L/r$ by the same factor. So curve $b$ is drawn for a single tuned circuit with $L/r = 500$. We see that the modulation frequency corresponding to a 5 kc/s note is reduced 1·4 times by the r.f. amplifier; that is to say, it retains about 70% of its relative strength. But the receiver with feedback reduces it 32 times, leaving only about 3%. True, it gives excellent selectivity; but it would take all the life out of music and make speech difficult to follow. At 9 kc/s off tune the r.f. amplifier is very inselective. But, significantly, at the more remote frequencies its selectivity is beginning to overtake its rival.

The excessive sharpness of $b$ in Fig. 18.7 was a result of having to make the tuned circuit provide as much gain as a stage of amplification. But now let us disregard gain and compare one circuit with several on a basis of equal selectivity. Suppose the intention is to reduce signal voltages at two channels off-tune (18 kc/s) to one-hundredth (−40 dB) relative to the on-tune voltage. Fig. 18.6 shows that the required $L/r$ for a single circuit would be $8{,}000/18 = 445$; nearly as sharp as $b$ in Fig. 18.7. With two circuits, $S$ for each would be − 20 dB, so $L/r = 800/18 = 44·5$. With four circuits, $S$ would be − 10 dB, and $L/r = 13$. With six, $S$ would be $-6\frac{2}{3}$ dB, and $L/r = 8·4$. The overall response at other frequencies can be derived from Fig. 18.6 by taking $S^2$ or twice the number of dB for two circuits, and so on. Plotting the results for one and six circuits on one graph, we get Fig. 18.8, which shows very clearly that a single tuned circuit, whether its extreme sharpness is necessitated by the requirements of gain or of selectivity, can fulfil these requirements only at the cost of drastic cutting of the wanted sidebands, and that a large number of relatively flat circuits is much to be preferred. Extensions

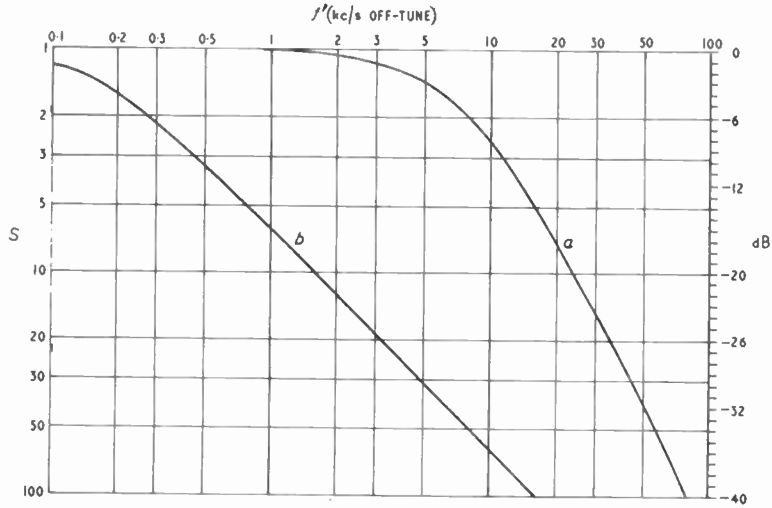



**Fig. 18.7**—Comparison on an equal-gain basis of a stage of amplification with two tuned circuits (curve *a*) with a single tuned circuit given a high magnification by positive feedback (*b*)

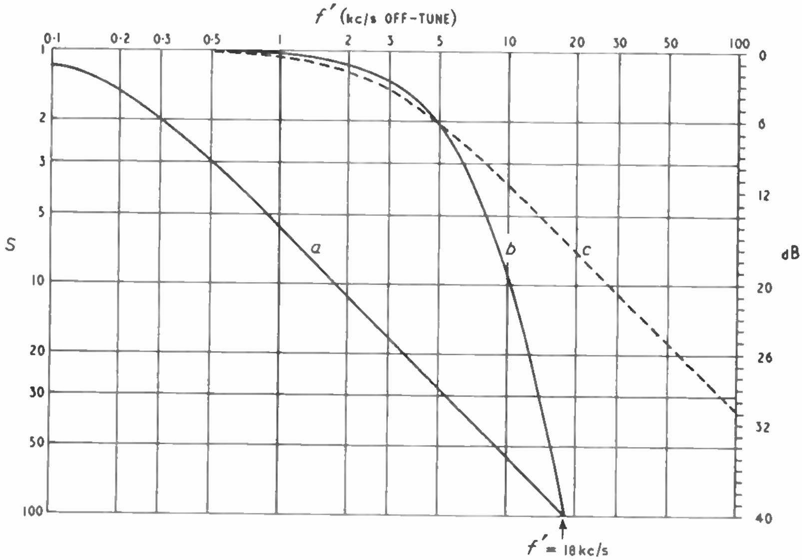



**Fig. 18.8**—Response curves of one (*a*) and six (*b*) tuned circuits on a basis of equal selectivity at 18 kc/s, showing how the six preserve sideband frequencies much better. Curve *c* refers to a single tuned circuit with its resonance flattened to give equal response at 5 kc/s off-tune

of the curves show, too, that the selectivity at frequencies beyond the chosen reference point of 18 kc/s goes on increasing much more rapidly with many circuits than with few.

Lastly, we can compare one circuit with six on a basis of equal retention of sidebands. Fig. 18.8 shows that with six circuits the loss at 5 kc/s is just over 6 dB. The effect of flattening the single tuned circuit until its loss at this point is the same is represented by sliding curve *a* along until it coincides with *b* there (position *c*). This shows how poor the single-circuit selectivity would then be, in spite of the quite severe sideband cut.

## Tuning Difficulties

It is simple enough to talk about choosing $L/r$ ratios to provide the response curves we want, and not usually very hard to achieve them at any one frequency. The difficulty is in maintaining the desired performance when the frequency is varied. When, as is usual, a coil is tuned by a variable capacitor, $L$ of course remains the same over the whole band, but $r$ is much less at the lowest frequency (maximum capacitance) than at the highest. So the selectivity varies accordingly. Even if some way out of this varying-selectivity trouble were found, it would only be to run into another. For if $L/r$ is kept constant over a band of frequency, the dynamic resistance $R$—which as we know is equal to $L/Cr$ (p. 111)—must be inversely proportional to the tuning capacitance, and therefore directly proportional to the *square* of the frequency. This means very much less amplification at the low frequency end of the band than at the high.

Tuning by varying $L$ instead of $C$ could theoretically avoid both these difficulties at once, if $r$ were made to vary in proportion, for then $L/r$ and $L/Cr$ would both be constant. But this is easier said than done.

Summing up the results of this chapter, we find it does much to blight the encouraging prospect of r.f. amplification opened up in the previous one. Using even six tuned circuits, the overall response curve is far from ideal, and seems unlikely to yield sufficient selectivity *and* good quality of reproduction, at any rate in the medium- and long-wave broadcasting frequencies. And whatever balance of selectivity and gain is chosen, with capacitance tuning it is bound to vary widely over the tuning range. Inductive tuning, though probably better in this respect, offers greater difficulties. And with any method of tuning, the problem of keeping, say, six separate circuits in tune over wide ranges of frequency is unattractive from the designer's point of view. So, too, is the problem of keeping stable the number of amplifying stages required.

The same problems arise in varying degree in the tuning of television and other v.h.f. receivers. As we shall see, the television sidebands to be carried are vastly wider than for sound.

So it is not surprising that attempts to produce a really practical high-performance receiver on the lines of variable-frequency r.f. amplification have failed. The generally-accepted solution is the subject of the next chapter.

# The Superheterodyne Receiver

## A Difficult Problem Solved

If we consider the difficulties listed at the end of the previous chapter we find that they reduce to the one basic difficulty of having to tune over a wide range of frequency. Designing a receiver for a fixed frequency is comparatively easy. The variation of gain and selectivity with frequency does not, of course, arise at all. Nor do the mechanical problems of making one knob control a number of tuning circuits simultaneously; they can all be adjusted by skilled operators in the factory and fixed at that. So it would not be unreasonable to suggest having six or even more tuned circuits, thereby obtaining a better compromise between the conflicting requirements of selectivity and quality. In fact, as we shall see, it is practicable to use slightly more complicated tuning circuits, having response curves better suited to the purpose than the one shown in Fig. 18.6. Seeing, however, that the market for receivers capable of working on one frequency only is negligible, is it not just rather tantalizing to imagine how easy it would be to design them?

Fortunately, a combination of the advantages of fixed-tuned selectivity with the ability to receive over a wide range of frequency is not so impossible as it might seem. It is actually accomplished in the great majority of present-day receivers, under the name of supersonic heterodyne—usually abbreviated to superheterodyne or just "superhet". A device called a frequency-changer shifts the carrier wave of any desired station, together with its sidebands, to whatever fixed frequency has been chosen for the selective circuits. This frequency is called the *intermediate frequency* (i.f.), and the following choices are typical:

| *For receiving—* | *I.F.* |
|---|---|
| Sound broadcasts (a.m.) on low and medium frequencies | 470 kc/s |
| Sound broadcasts (f.m.) on very high frequencies | 10·7 Mc/s |
| Television broadcasts vision on very high frequencies | 34·65 Mc/s |
| Television broadcasts sound on very high frequencies | 38·15 Mc/s |

The frequencies of all incoming signals are shifted by the frequency changer to the same extent. So if, for example, the original carrier-wave frequency of the wanted programme was 865 kc/s and it has been shifted to 470 kc/s, adjacent-channel carrier waves at 856 and 874 kc/s would be passed to the i.f. amplifier as 461 and 479 kc/s.

After the desired signal has been amplified at its new frequency and at the same time separated from those on neighbouring frequencies, it is passed into the detector in the usual way, and the remainder of the receiver is in no way different from a " straight " set.

## THE FREQUENCY-CHANGER

The most important thing to grasp in this chapter is the principle on which the frequency-changer works. All else is mere detail.

Suppose a sample of the carrier wave coming from the desired station during one hundred-thousandth of a second to be represented by Fig. 19.1*a*. As it contains 10 cycles, the frequency must be

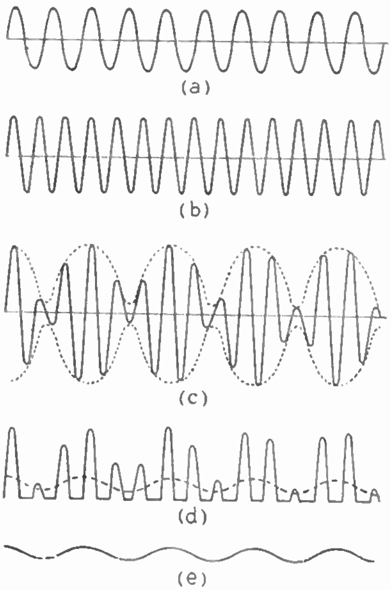



**Fig. 19.1—Showing the action of a frequency-changer**

1,000 kc/s. Similarly *b* represents a sample of a continuous wave-train having a frequency of 1,450 kc/s, which is being generated by a small oscillator in the receiver. Now add the two together. Since the local oscillation alternately falls into and out of step with the incoming signal, it alternately reinforces and weakens it, with the result shown at *c*: waves varying in amplitude at a frequency of 450 kc/s. This part of the process is often referred to as *b beating* or *heterodyning* with *a*, and 450 kc/s in this case would be called the *beat frequency*. Being above audibility, it is described as *supersonic*. The same kind of thing takes place audibly when two musical notes of nearly the same pitch are sounded together.

Notice particularly that what is happening at 450 c/s is only amplitude variation (or modulation) of waves of higher frequency.

*No signal of* 450 *kc/s is present*, for each rise above the centre line is neutralized by an equal and opposite fall below the line. (Fig. 18.2*d* is another example of the same thing, except that it is made up of waves of three frequencies, all higher than the frequency at which its amplitude is varying.) The average result, when applied to a circuit tuned to 450 kc/s, is practically nil. We have been up against this kind of thing already (see Fig. 16.1) and the solution is the same now as then—rectify it. After eliminating the negative half-cycles (*d*) the smoothed-out or averaged result of all the positive half-cycles (shown dotted) is a signal at 450 kc/s; that is to say, the difference between the frequencies of *a* and *b*. The mixture of 450 kc/s and the original frequencies is then passed to the i.f. circuits, which reject all except the 450 kc/s, shown now alone at *e*.

So far, the result is only a 450 kc/s carrier wave of constant amplitude, corresponding to the 1,000 kc/s carrier wave received from the aerial. But if this 1,000 kc/s wave has sidebands, due to modulation at the sender, these sidebands when added to the 1,450 kc/s oscillation and rectified give rise to frequencies which are

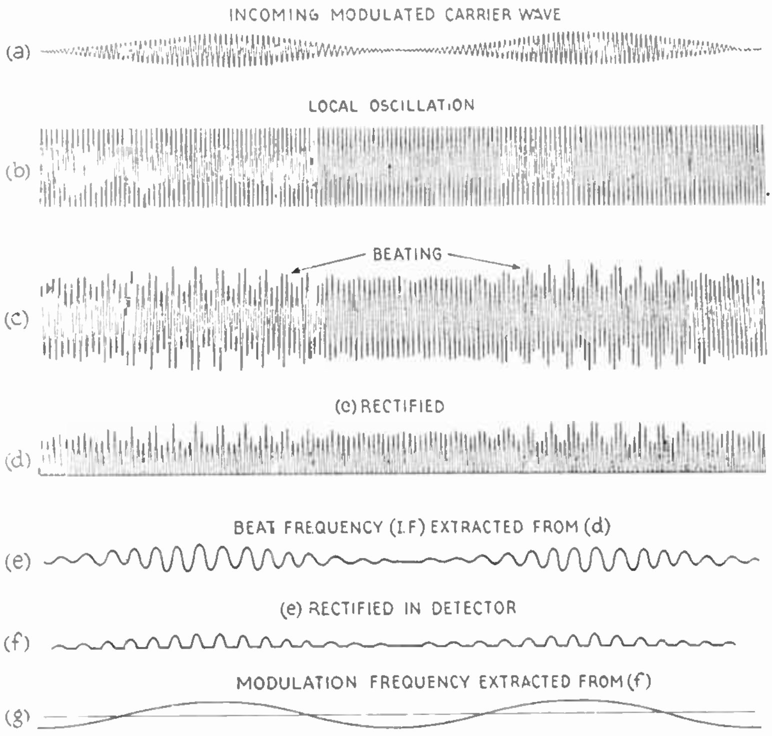



**Fig. 19.2—In this diagram the sequence of Fig. 19.1 is elaborated to show what happens when the incoming carrier wave is modulated. The additional lines, (*f*) and (*g*), show the extraction of the modulation frequency**

sidebands to 450 kc/s. So far as the i.f. amplifier and the rest of the receiver are concerned, therefore, the position is the same as if they were a " straight " set receiving a programme on 450 kc/s.

Any readers who, in spite of Fig. 18.3, have difficulty in seeing amplitude modulation as sidebands, may find Fig. 19.2 helpful. This shows a rather longer sample of the carrier wave (*a*), the individual cycles being so numerous that they are hardly distinguishable. And this time it is fully modulated; the sample is sufficient to show two whole cycles of modulation frequency. The local oscillation, which is certain to be much stronger, and of course is unmodulated, is shown at *b*. When *a* is added to it (*c*), nothing happens to it at the troughs of modulation, where *a* is zero, but elsewhere the cycles of *a* alternately strengthen and weaken those of *b*, causing its amplitude to vary at the beat frequency, in the manner shown. The amplitude of beating is proportional to the amplitude of *a*, and the frequency of beating is equal to the difference between the frequencies of *a* and *b*. The next picture, *d*, shows the result of rectifying *c*. We now have, in addition to other frequencies, an actual signal at the beat frequency (or i.f.), shown at *e* after it has been separated from the rest by the i.f. tuned circuits. The result of rectifying *e* in the detector is *f*; and after the remains of the i.f. have been removed by the detector filter the final result (*g*) has the same frequency and waveform as that used to modulate the original carrier wave.

Fig. 19.3 shows one way in which the process just described can be carried out. The incoming signal and the local oscillation are added together by connecting their sources in series with one

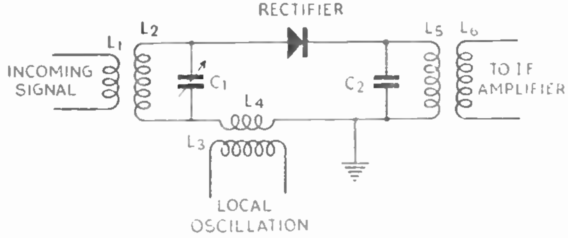



**Fig. 19.3—Simple type of frequency-changer, in which the incoming signal and the local oscillation are added together by connecting their sources in series, and then rectified**

another, and the two together are fed into a rectifier—some kind of diode. $C_1$ is used for tuning $L_2$ to the signal, so as to develop the greatest possible voltage at that frequency. There is, of course, no difficulty in getting an ample voltage from the local oscillator. $C_2$ serves the double purpose of tuning the transformer primary ($L_5$) to the intermediate frequency, and of providing a short-circuit for the two input frequencies, which are not wanted there.

## FREQUENCY-CHANGERS AS MODULATORS

If the rectifier in Fig. 19.3 had been omitted, or replaced by an ordinary linear resistor, no new signal frequency would have been created. In linear circuits, currents of different frequencies can flow simultaneously without affecting or being affected by the others. In non-linear circuits that is not so. A rectifier is an extreme kind of non-linear resistance. Fig. 19.4 shows the current/voltage characteristic curve of a milder kind. In diagram *a* a

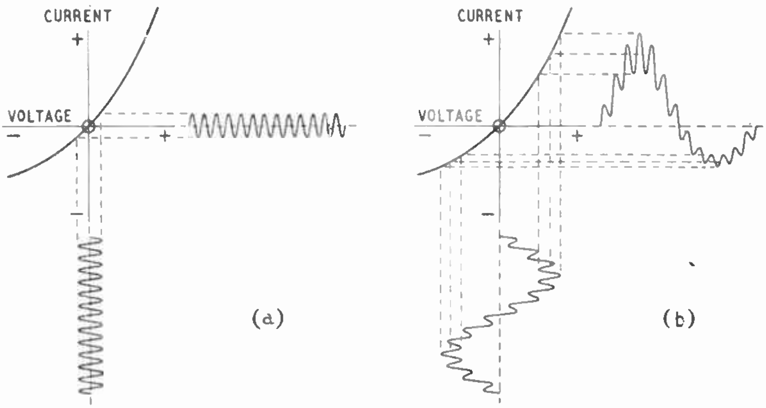



**Fig. 19.4—Showing the effect of adding a strong signal to a weak signal through a non-linear device such as a valve. The amplitude of the weak signal is varied (or modulated) by the strong**

single small alternating voltage—frequency, say $f_1$—is applied, and the resulting current waveform is derived and shown to the right of the graph. It is slightly distorted owing to the non-linearity, but each of its cycles is the same as all the others. Diagram *b* shows the result of applying the same voltage plus another, which has been chosen to have a much larger amplitude and lower frequency ($f_2$) to make it easily distinguishable. Its positive half-cycles carry the $f_1$ signal up the curve to where the resistance is less, so the $f_1$ current cycles are larger than in *a*. The $f_2$ negative half-cycles carry the $f_1$ signal to where the resistance is greater, so the $f_1$ current cycles are smaller than in *a*. The $f_2$ signal has left its mark on the $f_1$ signal, by amplitude-modulating it. The depth of modulation obviously depends on the degree of non-linearity of the circuit.

We already know that the process of amplitude modulation creates signals having frequencies different from both the modulating and modulated frequencies; to be precise, these frequencies are equal to the sum and difference of $f_1$ and $f_2$—see Fig. 18.2 again. These are not necessarily the only new frequencies created—that depends on the kind of non-linearity—but we shall confine our attention to them.

In Fig. 19.1 the difference frequency is clearly visible, especially at *e*, after it has been extracted from the mixture. The fact that in *d* there is a frequency equal to that of *b plus* that of *a* is not at all obvious, but its existence can be shown mathematically, or by actual experiment—picking it out from the rest by tuning a sharply resonant circuit to it.

The purpose of the local oscillator can now be seen as a means of varying the conductance of a part of a circuit carrying the incoming signal, amplitude-modulating it and thereby creating signals of at least two new frequencies.

Either of these could be chosen as the i.f. in a superhet, but the lower or difference frequency is the one usually favoured, because it is easier to get the desired amplification and selectivity at a relatively low frequency.

## Types of Frequency-Changer

A frequency-changer therefore consists of two parts: an oscillator, and a device whereby the oscillation modulates or varies the amplitude of the incoming signal. This device is commonly called the mixer. If the word " mix " is understood in the sense a chemist (or even an audio engineer) would give it—just adding the two things together—then it is a very misleading word to use. Simply adding voltages or currents of different frequencies together creates no new frequencies, just as merely mixing oxygen and hydrogen together creates no new chemical substance. So from this point of view " combiner " would be a better name for the second part of a frequency-changer. It can also be fairly termed the modulator.

The oscillator presents no great difficulty; almost any of the circuits shown in Chapter 12 could be used. It is, of course, necessary to provide for varying the frequency over a suitable range, in order to make the difference between it and the frequency of the wanted station always equal to the fixed i.f.

The simplest type of modulator or mixer is the plain non-linear element—a diode—already seen in Fig. 19.3. Frequency-changers essentially of this type are used in radar and other receivers working on frequencies higher than about 300 Mc/s, for reasons which will appear shortly. But they have certain disadvantages which have resulted in their having been superseded for most other purposes. For one thing, adjustment of $C_1$ is liable to affect the frequency of the oscillator. The local oscillations on their part are liable to find their way via $L_2$ and $L_1$ to the aerial, causing interference with other receivers. In short, the signal and oscillator circuits interact, elsewhere than in the proper direction—through the rectifier. And the rectifier does not amplify; on the contrary, it and its output circuit tend to damp the signal presented to it by the tuned circuit $L_2C_1$. So at frequencies not so high as to cause counter-difficulties, valves with three or more electrodes are preferred.

Such valves can be made to present a very high impedance to the input circuit, and so take little current from it. And while doing the job of modulator they can also, if desired, be made to do the job of oscillator. Fig. 19.5 shows a typical frequency-changer circuit for a f.m. v.h.f. receiver. $L_1C_1$ is the tuned circuit via which the incoming signal is obtained. $L_4$ is tuned to the required oscillator frequency by means of $C_2$, and is made to oscillate by feedback via the reaction coil $L_3$. The rather complicated system of capacitors $C_3C_4C_5$ is designed to provide a point (the junction of $C_3$ and $C_4$) of zero oscillating potential for connecting the signal

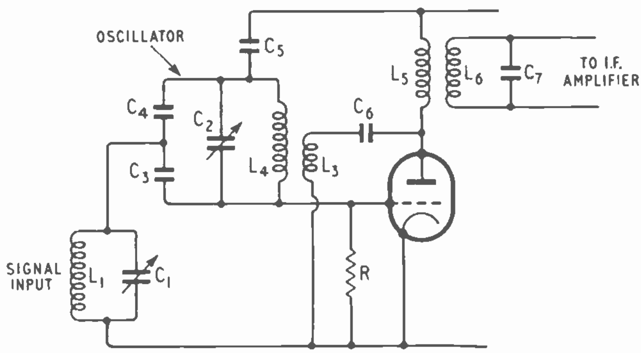



**Fig. 19.5—Typical additive frequency-changer, in which a single valve serves as both modulator and generator of the local oscillations**

input, so that oscillations do not pass into the input circuit $L_1C_1$ and thence to the aerial. $L_5$ is tuned to the relatively low i.f. by $C_6$; $L_3$ has a very low impedance at that frequency, so hardly counts. $L_5$, on the other hand, has a very high impedance at the oscillator frequency, so prevents oscillations leaking away there. The grid of the valve receives a considerable proportion of the signal voltage developed across $L_1$, plus a much larger voltage at the oscillator frequency.

This frequency-changer is certainly the most economical in valves, and is satisfactory over the comparatively small tuning ratio needed for f.m. broadcasts (about 87–100 Mc/s); but making a valve do two things at once does not always pay, especially under more difficult conditions. For television covering Band I (41–68 Mc/s) it is usual to have separate valves, though they may for convenience be enclosed in the same " bottle ": a triode as oscillator, and a pentode as modulator. The object of using a pentode is to avoid interaction between the i.f. and signal-frequency circuits, which are usually not so very widely different in frequency as they are in a f.m. receiver. The use of separate valves gives the designer freedom to adjust the coupling between the oscillator and modulator to give the latter the oscillation amplitude that suits it best.

Because both oscillation and signals are added together before being applied to the control grid of any of these frequency-changers, there is bound to be a certain amount of coupling between their two circuits. Generally speaking, the smaller the difference in frequency between them—i.e., the lower the i.f.—the more difficult it is to get satisfactory results. The more nearly one circuit is in tune with another, the more the resistance and reactance of both are altered by coupling them. This effect is usually called " pulling ". So when the i.f. is low, as it usually is in receivers for frequencies lower than about 30 Mc/s, the frequency-changer is preferably of a kind in which the oscillator controls the mutual conductance of the modulator through a different electrode from the one where the signal comes in. A simple example is the hexode, Fig. 19.6*a*, in which the signal is introduced as usual at the control grid, $g_1$, but the oscillator is connected to a modulating grid, $g_3$. This $g_3$ is screened from $g_1$

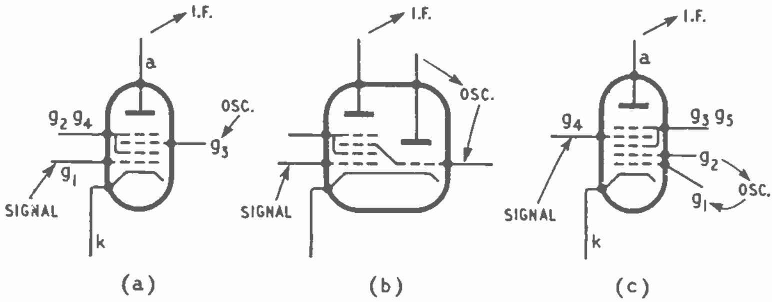



**Fig. 19.6—Some frequency-changer valves: (*a*) hexode; (*b*) triode-hexode; (*c*) heptode**

and anode by grids $g_2$ and $g_4$, which are connected to a fixed-potential point just like the screen grid in a tetrode or pentode. The oscillator triode is often enclosed in the same bulb, forming a triode-hexode (*b*). Pentode-like characteristics can be imparted by inserting a suppressor grid between $g_4$ and anode; the valve is then a triode-heptode.

A heptode of this kind can be used with a separate triode, but there is another kind of heptode (sometimes called a pentagrid) which needs no separate oscillator valve. As shown in Fig. 19.6*c*, the first two grids act as grid and anode of a triode oscillator, modulating the electron stream before it reaches the signal control grid, $g_4$. This type of frequency-changer is less good than the triode-hexode or -heptode, at least at frequencies over about 10 Mc/s, for two reasons: the mutual conductance of the oscillator section may not be sufficient for easy oscillation; and there is more interaction between oscillator and signal sections. But it is still a popular and economical choice for portable receivers and others mainly for low and medium frequencies.

## Conversion Conductance

The effectiveness of an ordinary amplifying or oscillating type of valve is, as we know, expressed as the mutual conductance, in milliamps of signal current in the anode circuit per volt of signal applied at the grid. In a frequency-changer valve, however, the anode current at signal-frequency is a waste product, to be rejected as soon as possible; what we are interested in is the milliamps of i.f. current per volt of signal-frequency. This is known as the *conversion conductance* of the valve, denoted by $g_c$. Obviously it does not depend on the valve alone but also on the oscillation voltage. If this is only a volt or two, it does not vary the mutual conductance over anything like its full range; consequently the signal current is modulated to only a fraction of its full amplitude, and the i.f. current is less than it need be. The conversion conductance reaches its maximum when the local oscillation is enough to bring the mutual conductance to zero during its negative half-cycles and to maximum during as much as possible of its positive half-cycles. The positive part of Fig. 19.2*c* is therefore reproduced with the full $g_m$ of the valve, while the negative part is completely suppressed; the net result is the same as with a perfect rectifier, Fig. 19.2*d*. The incoming signal, however weak, is swept right round the bend by the oscillation. That explains why a superhet is sensitive to the feeblest signals, which, if applied straight to the detector, would make no perceptible impression (p. 221).

In Fig. 19.7*a*, a small sample of oscillation plus signal (such as Fig. 19.2*c* or 19.1*c*) is shown in relationship to the $V_g$ scale of an $I_a/V_g$ valve curve. (The fact that the oscillation and signal may actually be applied to different grids does not affect the essential

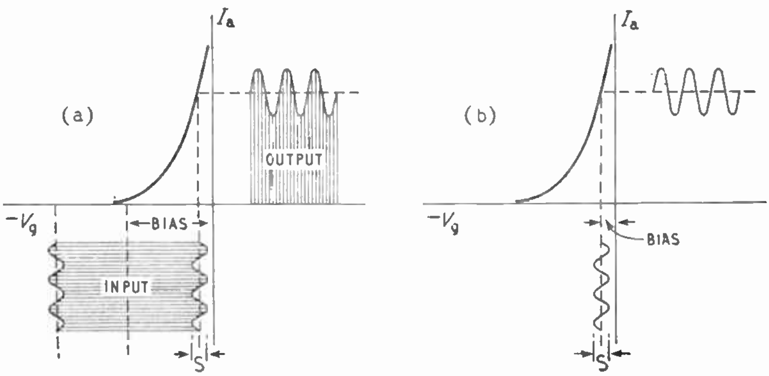



**Fig. 19.7—Diagram showing how conversion conductance compares with mutual conductance**

principle.) Since the i.f. fluctuations are due to the signal alternately strengthening and weakening the local oscillation, their amplitude is, as we have seen, equal to that of the signal. Neglecting the effect of the load in the anode circuit, as we generally can with these types of valve, corresponding fluctuations in anode-current half-cycles are practically equal to those that would be caused by an i.f. voltage of the same amplitude as the signal voltage amplified straightforwardly, as indicated at *b*. If the i.f. current amplitude in *a* were the same thing as the amplitude of the i.f. envelope, this would mean that the conversion conductance was practically equal to the mutual conductance. But, as is shown more clearly in Fig. 19.1*d*, the fluctuation in *mean* value of the half-cycles is $1/\pi$ or only about 32% of the peaks (p. 209). So $g_c$ is only about one-third of $g_m$ under ideal conditions, and in practice is between one-third and one-quarter.

In a r.f. tetrode or pentode the valve resistance is so much higher than any normal load resistance ($R$) that the voltage gain is approximately equal to $g_m R$ (p. 239). The corresponding quantity in a frequency-changer, the *conversion gain*, is equal to $g_c R$ and is the ratio of output i.f. voltage to input r.f. voltage. Even though it gives only about a quarter of the gain of a comparable r.f. amplifier, the frequency-changer makes a very welcome contribution to the gain of a receiver, in addition to its main function.

## Noise

Gain is not everything, however. If there is noise or interference along with the signal, more gain increases the noise too. In these days there is no real problem in getting all the gain one wants, so the limit is fixed by the noise that remains after as much as possible has been eliminated. The word " noise " is used to refer to the electrical disturbances which, in a receiver for reproducing sound, cause audible noise; but the same disturbances cause equally undesirable results when the end-product is not sound—they mar the picture in television, for example. Some kinds of interference and noise enter the receiver from outside. These include unwanted signals, which can be tuned out provided they are not on the same frequency as the wanted signal. Then noise is created by many kinds of electrical appliances. Anything that causes abrupt changes of current can do so, even when the actual amount of current is not large; it is the rate of change that counts. Such interference is not usually confined to a narrow band of frequency; it is of the nature of sudden shocks which can set any tuned circuit into damped oscillation at its own natural frequency. Lightning flashes are caused by abrupt changes of very large currents indeed, so cause extensive interference, especially with receivers tuned to long waves.

But when everything that comes in has been tuned out or suppressed, there is an unavoidable residue of electrical noise generated in the receiver itself. An electric current is not a continuous flow; it is made up of individual electrons, and even when there is no net current in any one direction the electrons in, say, a resistor are in a continuous state of agitation due to heat. Their random movements, like the aimless jostling of a crowd which, as a whole, remains stationary, are equivalent to tiny random currents, and they give rise to small voltages across the resistor. Quite a moderate gain, such as can be obtained with three or four stages, is enough to enable them to be heard as a rushing or hissing sound, or seen on the television screen as an animated graininess.

The electron agitation increases with temperature, and can only be quelled altogether by reducing the temperature to absolute zero ($-273^\circ$ C) which is hardly a practical proposition. It is known variously as *thermal agitation*, *circuit noise*, and *Johnson effect*. Although these voltages occur in a completely random fashion, their r.m.s. value over a period of time is practically constant; rather as the occurrence of individual deaths in a country fluctuates widely but the death-rate taken over a period of months changes very little from year to year. Since the frequency of the jostlings is completely random, the power represented by it is distributed uniformly over the whole frequency band. So if the amplifier is selective, accepting only part of the frequency band, the noise is reduced. The actual formula is

$$E = 2\sqrt{kTRf_w}$$

where $E$ is the r.m.s. value of noise voltage, $k$ is what is called Boltzmann's constant and is equal to $1{\cdot}37 \times 10^{-23}$, $T$ is the temperature in degrees Kelvin (Centigrade $+ 273^\circ$), $R$ is the resistance in ohms, and $f_w$ is the frequency bandwidth in c/s.

For example, suppose the input resistance of an amplifier is 0·1 MΩ, and it is tuned to accept a band 20 kc/s wide. Room temperature is usually about 290° K. So the r.m.s. noise voltage would be $2\sqrt{1{\cdot}37 \times 10^{-23} \times 290 \times 10^5 \times 2 \times 10^4} = 5{\cdot}6\ \mu$V. If the total voltage gain were, say, one million, this noise would be very disturbing. Note that the 0·1 MΩ might be the dynamic resistance of a tuned circuit; but if so the noise voltage across it would be less than that across a 0·1 MΩ resistor because its resistance is less than 0·1 MΩ at all except the resonant frequency.

A somewhat similar source of noise, usually called *shot effect*, is due to the anode current in a valve being made up of individual electrons. It is not so easily calculated as thermal noise, because it depends on the design of the valve, space charge, and other things; but like thermal noise it is proportional to $\sqrt{f_w}$. It also tends to be proportional to $\sqrt{I_a}$, but the situation is complicated when there are other current-carrying electrodes such as screening grids. These cause an increase, called *partition noise*. Also a frequency-changer introduces considerably more noise than an ordinary amplifying stage.

These facts help to explain the choices of frequency-changer for receivers for different frequency bands. At low and medium frequencies, useful gain is limited by incoming noise and interference, so there is no objection to using valves such as heptodes, with separate oscillator and signal inputs, which are the noisiest. The higher the frequency, the less the incoming noise, and in general the weaker the signals, so the greater the need to study receiver noise. So we have the pentode, which has some partition noise though not much, or even the triode, which has none, but is least good in other respects.

## Ganging the Oscillator

We have seen that the intermediate frequency is equal to the difference between the signal frequency and the oscillator frequency. With an i.f. of 470 kc/s, for example, the oscillator must be tuned to a frequency either 470 kc/s greater or 470 kc/s less than the signal.

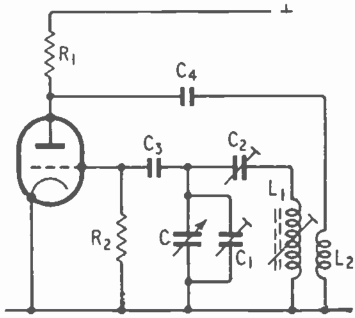
Suppose the set is to tune from 525 to 1,605 kc/s. Then, if higher in frequency, the oscillator must run from (525 + 470) to (1,605 + 470), i.e., from 995 to 2,075 kc/s. If, on the other hand, the oscillator is lower in frequency than the signal, it must run from (525 − 470) to (1,605 − 470), or 55 to 1,135 kc/s. The former range gives 2·09 as the ratio between highest and lowest frequency; the latter 20·6. Since even the signal-circuit range of 3·06 is often quite difficult to achieve, owing to the irreducible capacitance likely to be present in a finished set, the oscillator range from 995 to 2,075 kc/s is the only practical choice.

The next problem is to control both signal-frequency circuit tuning and oscillator frequency with one knob. Because there must be a constant frequency difference between the two, it hardly seems that a ganged variable capacitor—consisting, in effect, of two identical capacitors on the same spindle—would do. As we have just seen, the ratio of maximum to minimum capacitance is quite different. This difference, however, can easily be adjusted by putting a fixed capacitance either in parallel with the oscillator capacitor to increase the minimum capacitance, or in series to reduce the maximum. The ratio of maximum to minimum having been corrected in either of these ways, correct choice of inductance for the oscillator coil will ensure that it tunes to the right frequency at the two ends of the tuning-scale.

In the middle, however, it will be widely out, but in opposite directions in the two cases. It is found that a judicious combination of the two methods, using a small parallel capacitor to increase the minimum a little, and a large series capacitor to decrease the maximum a little, will produce almost perfect "tracking" over the whole wave-band.

The resulting circuit is shown in Fig. 19.8. Here C is a section of an ordinary gang capacitor; at every dial reading it has the same capacitance as its companion sections tuning the signal-frequency circuits. Its minimum capacitance is increased by the *trimmer* $C_1$, and its maximum (so far as the oscillator tuning coil $L_1$ is concerned) is reduced by the *padder* $C_2$; note the special symbol for $C_1$ and $C_2$, indicating that they are *pre-set* to the capacitances giving best tracking. $L_1$ is also pre-set by shifting its iron-dust core, denoted by the dotted lines. The reaction coil $L_2$ is

**Fig. 19.8—Arrangement of trimming capacitor ($C_1$) and padding capacitor ($C_2$) in a local oscillator, to enable its frequency to be correctly controlled by a tuning section (C) ganged to the section that tunes the signal circuit**




parallel-fed via $C_4$. $R_2$ and $C_3$ are the usual grid leak and capacitor for making the oscillation generate its own negative bias (p. 217).

Various switching arrangements are in use for extending this system to more than one waveband.

## Whistles

A super-heterodyne is liable to certain types of interference from which an ordinary set is free. The most noticeable result of these is a whistle, which changes in pitch as the tuning control is rotated, as when using a " straight " receiver in an oscillating condition, but generally less severe. There are many possible causes, some of which are quite hard to trace.

The best-known is *second-channel* or *image* interference. In the preceding section we have seen that it is usual for the oscillator to be adjusted to a frequency higher than that of the incoming signal. But if, while reception is being obtained in this way, another signal comes in on a frequency higher than that of the oscillator by an equal amount, an intermediate frequency is produced by it as well. If the frequency difference is not *exactly* the same, but differs by perhaps 1 kc/s, then two i.f. signals are produced, differing by 1 kc/s, and the detector combines them to give a continuous note of 1 kc/s.

An example will make this clear; especially if for the i.f. we choose a round number, 100 kc/s. Suppose the station desired works on a frequency of 950 kc/s. When tuned to it, the oscillator is 100 kc/s higher, 1,050 kc/s, and yields a difference signal of 100 kc/s, to which the i.f. amplifier responds. So far all is well. But if a signal of 1,149 kc/s is also able to reach the frequency-changer it will combine with the oscillation to produce a difference signal of $1,149 - 1,050 = 99$ kc/s. This is too near 100 kc/s for the i.f. amplifier to reject, so both are amplified together by it and are presented to the detector, which produces a difference frequency, 1 kc/s, heard as a high-pitched whistle. Slightly altering the tuning control alters the pitch of the note, as shown in the table:

| *Set tuned to:* | *Oscillator at:* | *I.F. signal due to* 950 *kc/s signal (wanted)* | *I.F. signal due to* 1,149 *kc/s signal (interfering)* | *Difference (pitch of whistle)* |
|---|---|---|---|---|
| *kc/s* | *kc/s* | *kc/s* | *kc/s* | *kc/s* |
| 945 | 1,045 | 95 | 104 | 9 |
| 946 | 1,046 | 96 | 103 | 7 |
| 947 | 1,047 | 97 | 102 | 5 |
| 948 | 1,048 | 98 | 101 | 3 |
| 949 | 1,049 | 99 | 100 | 1 |
| 949½ | 1,049½ | 99½ | 99½ | 0 |
| 950 | 1,050 | 100 | 99 | 1 |
| 951 | 1,051 | 101 | 98 | 3 |
| 952 | 1,052 | 102 | 97 | 5 |
| 953 | 1,053 | 103 | 96 | 7 |
| 954 | 1,054 | 104 | 95 | 9 |

**An example of how second-channel interference is produced**

Since both stations give rise to carriers falling within the band to which the i.f. amplifier must be tuned to receive one of them, this part of the set can give no protection against interference of this sort. So the superhet principle does not allow one to do away entirely with selective circuits tuned to the original signal frequency. Such circuits are collectively called the *preselector*.

Another form of interference, much more serious if present, but fortunately easy to guard against, is that due to a station operating within the i.f. band itself. Clearly, if it is able to penetrate as far as the i.f. tuning circuits it is amplified by them and causes a whistle on *every* station received. Again, a good preselector looks after this: but the 525 kc/s end of the medium frequency band may be dangerously close to a 470 kc/s i.f. If so, a simple rejector circuit tuned to the i.f. and placed in series with the aerial will do the trick. It is known as an i.f. trap.

The foregoing interferences are due to unwanted stations. But it is possible for the wanted station to interfere with itself! When its carrier arrives at the detector it is, of course, always at intermediate frequency. The detector, being a distorter, inevitably gives rise to harmonics of this frequency (p. 294); that is to say, currents of twice, thrice, etc., times the frequency. If, therefore, these harmonics are picked up at the aerial end of the receiver, and the frequency of one of then happens to be nearly the same as that of the station being received, the two combine to produce a whistle. It is easy to locate such a defect by tuning the set to two or three times the i.f. The cure is to by-pass all supersonic frequencies that appear at the detector output, preventing them from straying into other parts of the wiring and thence to the preselector circuits or aerial.

Oscillator harmonics are bound to be present, too; and are a possible cause of whistles. Suppose the receiver is tuned to 200 kc/s and the i.f. is 470. Then the oscillator frequency is 670, and that of its second harmonic is 1,340 kc/s. If now a powerful local station were to be operating within a few kc/s of 870, its carrier would combine with the harmonic to give a product beating at an audio frequency with the 470 kc/s signal derived from the 200 kc/s carrier. Such interference is likely to be perceptible only when the receiver combines poor preselector selectivity and excessive oscillator harmonics.

Excluding most of these varieties of interference is fairly easy so long as there are no overwhelmingly strong signals. But if one lives under the shadow of a sender it is liable to cause whistles in a number of ways. Besides those already mentioned, an unwanted carrier strong enough to force its way as far as the frequency-changer may usurp the function of the local oscillator and introduce signals of intermediate frequency by combining with other unwanted carriers. Any two stations transmitting on frequencies whose sum or difference is nearly equal to the i.f. may cause interference of this kind. The designer guards against the danger by a suitable choice of i.f., provided that the frequencies of all pairs of stations that might be "local" are known. And, of course, preselector selectivity comes to the rescue once more.

The foregoing list of possible causes of interference is by no means exhaustive; but it is only in exceptional situations that whistles are conspicuous in a receiver of good design.

## The Preselector

As we have just seen, some selectivity is needed between the aerial and the frequency-changer. It is not called upon to deal with adjacent-channel interference, for we saw in the previous chapter that no reasonably small number of tuned circuits is capable of doing so without severely cutting sidebands. Leaving that task mainly to the i.f. amplifier, the preselector has to deal with second-channel interference, which is separated from the wanted channel by twice the i.f. At one time, an i.f. of about 110 kc/s was usual in broadcast receivers, and to exclude strong second-channel interference at least two preselector tuned circuits were necessary. The demand for cheap receivers (and therefore the fewest possible variable tuned circuits) led to the i.f. being raised to about 470 kc/s, giving a separation of nearly 1 Mc/s from the image frequency. Consequently, the majority of broadcast receivers now contain only a single preselector circuit. Except for such possible details as an i.f. trap, the circuit of a superhet of this kind, as far as the frequency-changer valve, is the same as the corresponding part of a "straight" circuit. It is reasonably satisfactory on the low and medium frequencies. At the higher frequencies of short waves, however, even 2 × 470 kc/s is only a few per cent off-tune, and it is usual to find that most short-wave stations can be received at two positions of the tuning knob. Receivers designed exclusively for them—and still more for television and f.m.—have suitably higher i.fs, as tabulated on p. 261.

On p. 270 we noted that the range of receivers for very high frequencies is usually limited by noise generated in the receiver itself, and that the greater the number of current-carrying electrodes the noisier the valve, the frequency changer being much the worst. Seeing that the first stage in the receiver—the one followed by the greatest amplification—is the one that chiefly decides the signal-to-noise ratio, it is clearly undesirable for the first stage in a v.h.f. receiver to be the frequency-changer. So f.m. and television receivers almost invariably start with a stage of amplification at the original signal frequency. This r.f. or preselector stage confers the additional benefits of reducing (1) oscillator radiation from the aerial and (2) the kinds of interference considered in the last section.

## Common-Grid Configuration

Although a pentode can be used as the preselector valve as described in Chapter 17, the lower noise contribution of the triode makes it an attractive candidate if only the feedback through its inter-electrode capacitance can be overcome. It can be, if its one and only grid is used as a screening grid by earthing it. The input is still connected between it and the cathode (Fig. 19.9) so it follows that the cathode cannot be earthed but is the " live " or high-

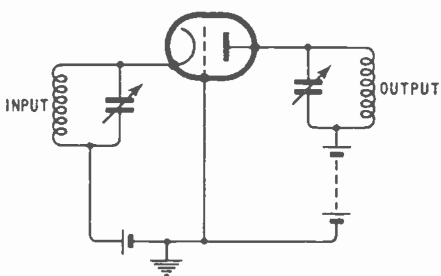



**Fig. 19.9—Common-grid amplifier circuit**

potential input electrode. And because the anode current must get back to the cathode it also follows that the output current must flow through the input circuit. When an input signal makes the cathode more positive, that is the same thing as grid more negative, which reduces anode current. That makes anode more positive, so this type of amplifier circuit does *not* invert the polarity of the signal. Passing through the input circuit impedance, the reduced anode current reduces the voltage drop, tending to make the cathode *less* positive. It therefore opposes the input signal, so in contradistinction to positive feedback this effect is called negative feedback. The gain is therefore considerably less than when the valve is used in its more usual manner, but even if it were only two or three times it would be worth while, by improving the signal/noise ratio. More gain can always be obtained following the frequency changer. Another disadvantage is that the feedback makes the resistance of the valve between its input terminals appear quite low; compare Miller effect (p. 228). But because the dynamic resistances of v.h.f. tuning circuits are inevitably rather low they are less sensitive to this than they would be at lower frequencies.

This arrangement, for obvious reasons, is called an *earthed-grid* stage; and because the grid is the electrode common to both input and output circuits the valve is said to be in common-grid configuration.

### The Cascode

To overcome its two disadvantages, an earthed-grid triode is often combined with an earthed-cathode triode as in Fig. 19.10, to form a *cascode* amplifying circuit. Here the earthed-grid valve

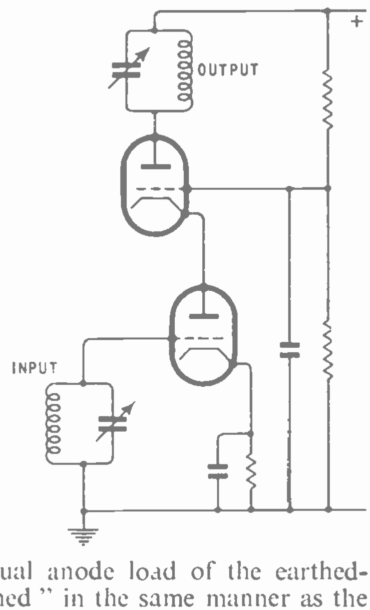



**Fig. 19.10—Cascode amplifier circuit**

directly takes the place of the usual anode load of the earthed-cathode valve, so its grid is " earthed " in the same manner as the screen grid of a pentode (p. 239). In fact, the cascode as a whole resembles a pentode in its general performance as well as in this detail, but without the higher noise of the pentode.

One might well wonder how, if the first triode is working as a conventional common-cathode stage, it does not become unworkable on account of Miller effect, especially at such high frequencies. The answer lies in the low impedance of the second valve. Because of this, the first triode gives little or no voltage gain, so feedback is insufficient to cause trouble; but it does pull its weight by presenting a very much higher input impedance, thereby enabling a much greater signal voltage to be developed across the input tuned circuit.

## The Task of the I.F. Amplifier

Having dealt with the preselector and frequency-changer stages, we can get back to the subject of selectivity, which is chiefly the concern of the i.f. amplifier. The two great advantages obtained by frequency-changing are the fixed tuning and the fact that whatever the signal frequencies may be the amplification can be carried out at a frequency of one's choice. The fixed tuning means that the selectivity and gain are the same at all signal frequencies, and nearer the ideal than the best that would be practicable with variable tuning. Choice of frequency for amplification means that if necessary a higher gain can be obtained.

Fig. 18.8 showed us, however, that the combined effort of even six tuned circuits may fall quite a long way short of giving sufficient adjacent-channel selectivity together with well-preserved sidebands. The real limitation is the basic resonance curve of the simple tuned circuit, Fig. 18.6, because it has the wrong shape—quite different from the flat-topped, steep-sided ideal. And to have a valve in between every pair of tuned circuits is inconvenient and expensive, and gives unusably high gain. If you remember (p. 255), the reason for separating the tuned circuits by valves (apart from the desire for amplification) was to prevent the resonance curves of adjacent circuits from being distorted by the reactive coupling needed to transfer the signal from one to the other. Introducing such coupling makes it much more difficult to predict what sort of resonance curve will be obtained with any particular set of circuits. But although such reluctance to couple the circuits is natural enough when approaching the subject for the first time, it would be foolish to avoid it in practical design, because it can be shown (by a rather greater mathematical effort than is appropriate in a book of this kind) that such coupling not only saves valves but also gives a better-shaped resonance curve.

## Critically-Coupled Tuning Circuits

There are many ways of coupling two tuned circuits; Fig. 19.11 shows some of them. The feature common to all is a certain amount of shared (or mutual) reactance. In type *a* it is capacitance, $C_m$; in *b*, self-inductance, $L_m$; and in *c*, mutual inductance, $M$. Fortunately for calculation, the result does not depend (at any one frequency) on the kind of mutual reactance but only on its amount; so we can denote it simply by $x$. As a matter of fact, method *c* is nearly always used, because no extra component is needed and the coupling can be adjusted simply by varying the spacing of the two coils. Just as $L/r$ is the factor determining

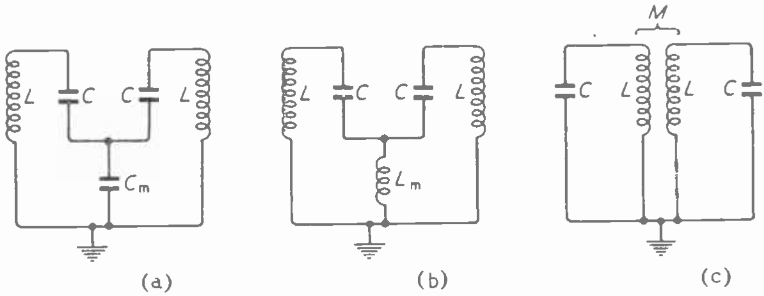



**Fig. 19.11—Three methods of coupling two tuned circuits: (*a*) by capacitance common to both, $C_M$; (*b*) by common inductance, $L_M$; and (*c*) by mutual inductance, *M***

the selectivity of a single tuned circuit, $x/r$ determines the modifying effect of the coupling on the combined resonance curves of two similar circuits.

This result can be shown most clearly in the form of the overall resonance curve for two circuits, (*a*) with no coupling, other than through a valve, and (*b*) with reactance coupling. The question at once arises: with how much? The shape of the curve is different for every value of $x$. Since the primary purpose of the coupling is to transfer the signal from one circuit to the other, a natural coupling to try for a start is the amount that transfers it most effectively. This is not, as might at first be supposed, the greatest possible coupling. It is actually (with low-loss circuits) quite small, being simply $x = r$.

For this value, called the *critical coupling*, the curve is Fig. 19.12*b*. The curve applies to a pair in which the two circuits have the same $L/r$: if there is any difference, an approximation can be made by taking the mean $L/r$. Curve *a*, for uncoupled circuits, is the same as Fig. 18.6 but with all the $S$ figures squared. Although the difference in shape is not enormous, it is quite clear that the curve for the coupled circuits is considerably flatter-topped. But it is also less selective. The advantage gained by the coupling is more obvious if the resistance of the coupled circuits is reduced to half that of the uncoupled circuits, as in Fig. 19.12*c*. If, for example, the $L/r$ of each of the uncoupled circuits represented by *a* were 10, the $L/r$ for each of the coupled circuits represented by *c* would be 20, and the off-tune or sideband frequency treated equally by both combinations (represented by the intersection of *a* and *c*) would be 6 kc/s. Although the coupled circuits would thus retain sideband frequencies up to 6 kc/s better than the uncoupled circuits they would also give substantially more selectivity.

## Effects of Varying the Coupling

Even so, an adjacent-channel selectivity of less than 8 is hardly an adequate protection against interference equal in strength to

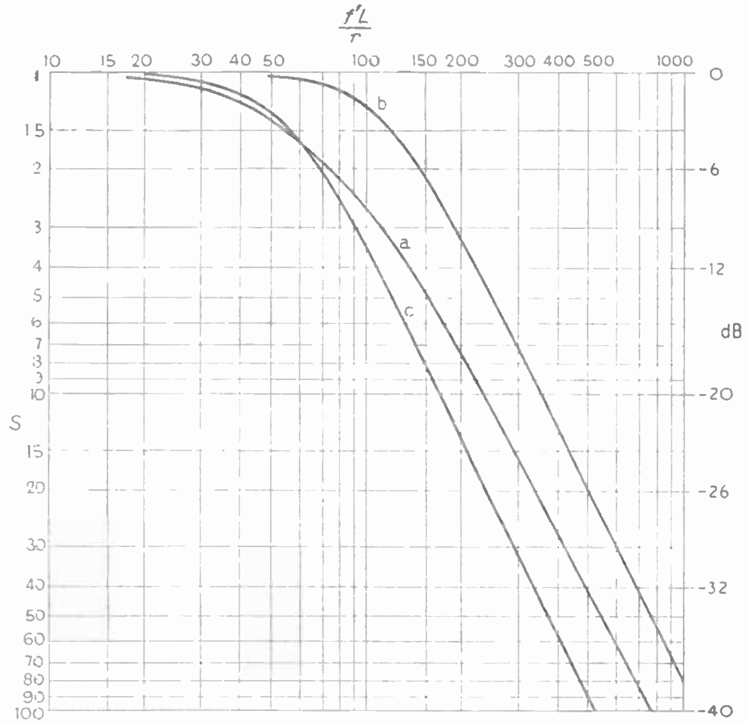
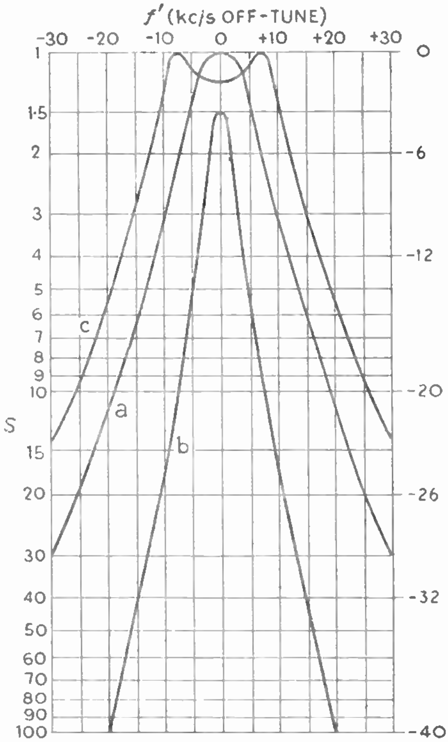



Fig. 19.12—Comparison between the overall resonance curve for a pair of similar tuned circuits, (*a*) uncoupled, and (*b*) "critically" coupled. Curve *c* is the same as *b* but shifted to the left to represent a halving of resistance

the wanted signal, still less against relatively strong interference. We may rightly ask whether any other degree of coupling than critical would give a still better shape; for, if so, any loss in signal transference caused thereby would be well worth while, as it could easily be made up by a little extra amplification. Fig. 19.13 shows what happens when the coupling between two tuned circuits is varied. The curves have been plotted to a linear frequency scale both sides of resonance, to show the kind of shapes seen on an oscilloscope screen (p. 327).

Reducing the coupling below critical, by making $x$ less than $r$, not only reduces the signal voltage across the second circuit but also sharpens the peak, so is wholly undesirable. Increasing the coupling splits the peak into two, each giving the same voltage as with the critical coupling at resonance; and the greater the coupling the farther they are apart in frequency. Here at last we have a means by which the selectivity can be increased by reducing the damping of the circuits, while maintaining sideband response at a

Fig. 19.13—Changes in resonance curve of a pair of coupled circuits caused by varying the coupling: curve *a* is for critical coupling; *b* is for less than critical; and *c* for more




maximum. Coupled pairs of circuits are one type of what are known, for obvious reasons, as *band-pass filters*. There is a limit, however, because if one goes too far with the selectivity one gets a result something like Fig. 19.14, in which a particular sideband

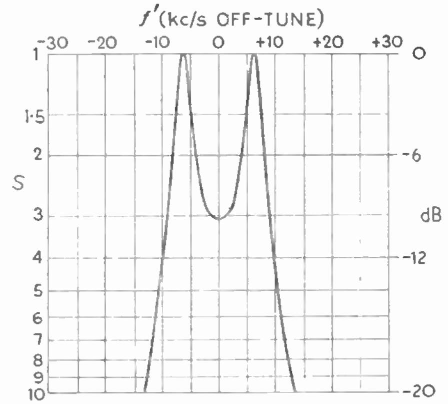



Fig. 19.14—The result of trying to maintain sidebands when using excessively selective coils in a bandpass filter—the "rabbit's ears" produced by over-coupling

frequency is accentuated excessively. Symmetrical twin peaks are in any case difficult to achieve in practice and still more difficult to retain over a long period, because the coupling adjustments are so extremely critical. Moreover, they make the set difficult to tune, because one tends to tune it to the frequency that gives the loudest sound, which in this case is not the carrier frequency.

Perhaps the best compromise is to have one pair of circuits critically coupled, and sufficiently selective to cause appreciable

(a)

(b)

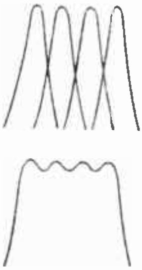


**Fig. 19.15—" Stagger-tuning," in which the separate tuned circuits are set to " peak " at slightly different frequencies (as shown at (*a*) ), so that the combined result (*b*) has a fairly flat top and sharp cut-off**

sideband cutting; and the other pair slightly over-coupled, to give sufficient peak-splitting to compensate for the loss due to the first pair. The ultimate cut-off will then be very sharp and give excellent selectivity.

Owing to the difficulty of " lining up " the circuits (i.e., pre-setting the tuning) with sufficient accuracy to fulfil this plan, an alternative sometimes adopted, especially for the very wide television sidebands, is *stagger* tuning, in which a number of single or double circuits are peak-tuned to slightly different frequencies, as in Fig. 19.15.

## A Typical I.F. Amplifier

Most a.m. broadcast sound receivers have one stage of i.f. amplification, with one pair of coupled circuits between frequency-changer and i.f. valve and another pair between i.f. valve and detector. Television receivers have to retain sidebands up to 2·5 Mc/s or more, so a much higher i.f. must be used, and the circuits have to be flatly tuned; so much so that damping resistors are usually connected across them. So their dynamic resistance, and therefore the gain obtainable per stage, must be small, making it necessary to use several stages.

To illustrate the orders of magnitude used in a.m. broadcast sound receivers, let us work out a simple example, in which the i.f. is 470 kc/s and both pairs of circuits are critically coupled. Fig. 19.12 can therefore be used to find the $L/r$ needed to meet a given specification. Suppose we allow an overall loss of 6 dB at 4 kc/s or 3 dB per pair. The corresponding $f'L/r$ is 112, and dividing this by 4 we get $L/r=28$. Since the resonant frequency is 470 kc/s, $Q=2\pi f_r L/r=2\pi \times 0{\cdot}470 \times 28 = 83$. (The fact that $L$ is in *micro*henries is counterbalanced by expressing $f_r$ in *mega*cycles per second.) Various combinations of $L$ and $C$ can be used to tune to 470 kc/s; and as the dynamic resistance is equal to $QX_L$ (p. 111), the gain is greater the larger the value of $L$. The greatest would be obtained with no added capacitance, $C$ consisting merely of the valve and circuit capacitances. Tuning could then be done by varying $L$, by screwing a plug of solidified iron dust in or out of the coil. But the tuning would be very sensitive to small variations in capacitance, due, say, to change of valve; so whether they are variable for tuning or not, capacitors are usually connected in parallel with the coils to bring the total capacitance up to some such value as 200pF. Assuming this figure, $L$ must be 573 $\mu$H (p. 106). The dynamic resistance of each circuit separately is $QX_L = 83 \times 2\pi \times 0{\cdot}470 \times 573 = 140{,}000\ \Omega$. As regards dynamic resistance and output voltage, a critically coupled pair is equivalent to a single circuit with r.f. resistance equal to $2r$—they are both halved. Supposing then that in our example the mutual conductance of the i.f. valve is 2 mA/V, the voltage gain of the stage is approximately $2 \times 140/2 = 140$ times, between the grid of the i.f. valve and the detector. It is assumed, of course that the input resistance of the detector has been taken into account in the figure for $L/r$; in practice it might not be easy to achieve an overall figure of 28 for the secondary.

Taking 0·7 mA/V as a likely value for the conversion conductance, the gain due to the frequency-changer working into a similar coupled pair would be $0{\cdot}7 \times 140/2 = 49$, making the total from grid of frequency-changer to detector $49 \times 140 = 6{,}860$. To deliver 10 V of i.f. to the detector, the r.f. input to the frequency-changer would have to be $10/6{\cdot}86 = 1{\cdot}46$ mV. Owing to the magnification of the preselector circuit, the input from the aerial, even without an r.f. amplifier, would normally be considerably less.

## Control of Gain

Most receivers have to be provided with a considerable amount of pre-detector gain in order to be capable of supplying the detector with an adequate signal voltage even when tuned to the weakest transmission within range. The question then arises as to what happens when the same receiver is tuned to a powerful local station. We have just considered an example in which 1·46 mV at the input to the frequency-changer provided the detector with 10 V. A local station might well give the frequency-changer 100 mV, in which case the same amount of gain would present the detector with 685 V. In practice this would not occur, because with the usual h.t. supply of not more than about 250 V the i.f. valve would be quite unable to develop anything like such a high voltage. It would be driven to current cut-off and there would be severe distortion. Even if it could, and the detector were able to stand such a high voltage, it would be most undesirable, because the full control of volume from zero to full would be compressed into the first 1% or so of the volume-control scale; the rest would cause excessive loudness (or picture brightness) and distortion.

What is really wanted is control of the pre-detector gain so that no matter what the strength of the incoming transmission the detector is supplied with the same maximum voltage—enough to provide full volume and a little over when the volume control is turned full on.

One possible method is to control the signal passing from the aerial to the receiver by means of a variable resistor or potential divider. That would mean the pre-detector stages would always be working at their full gain, so would always deliver their maximum noise to the detector. While perhaps tolerable when receiving the

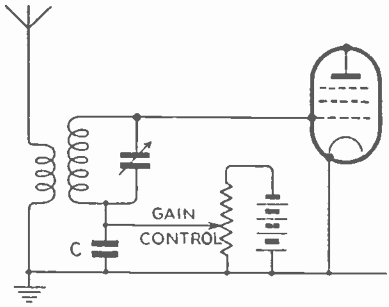



**Fig. 19.16—Method of controlling the amplification (or gain) of a valve by adjusting its grid bias**

most distant stations, this should if possible be avoided when receiving a near one. If on the other hand the gain of the amplifier itself is reduced to control surplus incoming signal strength, then noise is reduced accordingly and the signal/noise ratio improved.

Since the gain of a stage is approximately equal to $g_m R$, we have the choice of varying either or both of these factors. The dynamic resistance, $R$, is a very awkward thing to control by a knob; but $g_m$, which is the slope of the valve's $I_a/V_g$ curve, is quite easy. The curves for pentodes are like those for triodes (e.g. Fig. 9.7) in sloping less and less steeply as the negative grid bias is increased. All we need is a source of negative voltage with a potential divider to apply any desired proportion of it to the grid, in series with the signal source, as in Fig. 19.16. If there are several valves in the amplifier all their grids can be controlled simultaneously by this means. C is used to provide a low-impedance earth connection for the lower end of the tuned circuit.

## Distortion in the R.F. (and I.F.) Amplifier

When receiving the weakest signals, this gain control would be turned up so as to apply only the least bias necessary to prevent the flow of grid current (p. 124). In these circumstances the working point is on a very nearly linear part of the valve characteristic, and as the signal voltage is very small the risk of distortion can be ignored.

But now suppose that a local station is tuned in. To keep the signal voltage to the detector down to the same level as for the weak signal, it is necessary to reduce the gain, by applying more bias. This brings the working point down on the foot of the curve; at the same time the input signal voltage is much greater, so the signal cycles are considerably distorted—partly rectified, in fact.

To make sure that we do not take too pessimistic a view of this situation it may be as well to pause a moment to think what we mean by distortion here. Remember, it is the a.f. waveform, with which the r.f. carrier was modulated at the sender, that we are anxious to preserve. So long as this is done, it matters nothing what happens to the waveforms of the individual cycles of r.f. They could all be changed from sine waves into square waves, or any other shapes, but so long as they retained the same relative amplitudes they would still define the original a.f. envelope (p. 166). A perfect rectifier cuts off the negative half of every cycle, yet this drastic r.f. distortion introduces no distortion of the modulation. As we saw in connection with Fig. 16.15, it is the imperfect " curved " rectifier that distorts the modulation, by altering the relative proportions of the r.f. cycles; but unless the modulation is fairly deep it is surprising how little it is distorted by even a considerable curvature of the amplifying valve's characteristics.

## Cross-Modulation

Distortion of the modulation waveform, however, is not the only thing that can happen as a result of non-linearity in the r.f. (in-

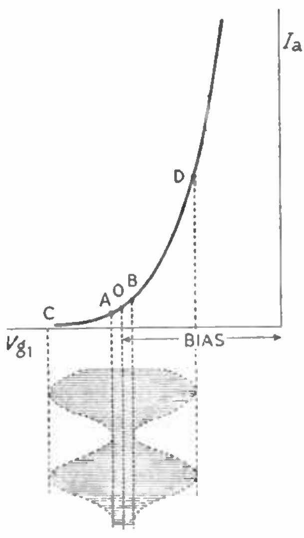



**Fig. 19.17—Showing how the average slope of the valve curve varies during each modulation cycle of a strong signal**

cluding i.f.) amplifier. There is what is known as *cross-modulation*, which makes its appearance in the misleading guise of lack of selectivity. Suppose a set is being used to receive a station 50 kc/s away from the local. We may be quite right in assuming that the overall selectivity of the several tuned circuits is enough to reduce the local station to inaudibility when they are all tuned 50 kc/s away from it. But the grid of the first valve may be protected from it by only one tuned circuit and therefore receive quite a large voltage along with that of the weaker station to which the set is tuned.

What happens can be examined in the same way as for the frequency changer, which is another example of a non-linear valve handling more than one signal at a time. In Fig. 19.17 the $I_a/V_{g1}$ curve is shown with a sample of the local station's modulated carrier wave below. The working point, set by the gain control bias, is marked 0.

Now consider what is happening to the wanted signal. It is being swung up and down the curve by the local station's carrier wave; between A and B during the troughs of local station modulation, and between C and D during the crests. Owing to the curvature, the amplification it receives varies in a manner determined by the local-station modulation.

On the average it is greater between C and D than between A and B. So the amplification, and consequently the output amplitude, of the wanted carrier wave is varied at the modulation frequency of the local station. In other words, the wanted carrier wave is to some extent modulated by the modulation of the unwanted station. So even although the unwanted carrier wave may be removed entirely by subsequent selectivity, it has left its mark indelibly on the wanted carrier, which, with its two-programme modulation, passes through the receiver, and both programmes are heard together.

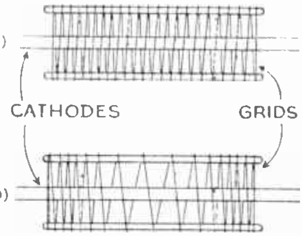
If the transmission to which the set was tuned were switched off, the programme of the interfering station would disappear with it, proving that the interference was due to cross-modulation and not simply to insufficient selectivity.

To prevent cross-modulation it is necessary to reduce the curvature of the valve characteristic, if possible without sacrificing the ability to vary the gain.

## Variable-Mu Valves

The fault of the ordinary type of valve is that the slope of its $I_a/V_{g1}$ curve changes very rapidly near the foot and much less so higher up. For the purpose of gain control this not only has the disadvantage of concentrating the curvature at the end that is used for local-station reception and thereby causing distortion of those programmes for which it is least tolerable, but it also tends to unevenness of control. What is needed is a valve whose $I_a/V_{g1}$ curve increases its slope steadily; if anything, less rapidly towards its foot.

Fig. 19.18—Comparison between the grid of an ordinary valve ($a$) and a variable-mu valve ($b$)




This result is achieved very simply by winding the control grid with variable spacing as shown in Fig. 19.18*b* instead of evenly as in ordinary valves (*a*). So when the negative bias is gradually increased, current from the ends of the cathode is shut down first, while electrons can still pass through the centre part of the grid where there are relatively wide gaps between the negative wires; and it takes perhaps 30 or 40 volts to cut the current off everywhere. The difference in characteristics is shown in Fig. 19.19 where curve *a* refers to a steep-slope pentode with sharp cut-off, and *b* to a comparable type having non-uniform grid spacing. It is known as a *variable-mu* valve; the name can be taken to mean that the mutual conductance is variable in a manner well suited to gain-control. How far this is successful can be better judged by plotting $g_m$ in place of $I_a$, as in Fig. 19.20, which refers to the same valve and conditions as Fig. 19.19*b*. A logarithmic scale is used for $g_m$, not only for the sake of enabling a wide range of values to be shown clearly, but also because the ideal is for the gain to increase in the same *ratio* for equal changes in $V_{g1}$, and this ideal is represented on such a graph by a straight line. Fig. 19.20 therefore indicates a remarkably close approach to the ideal.

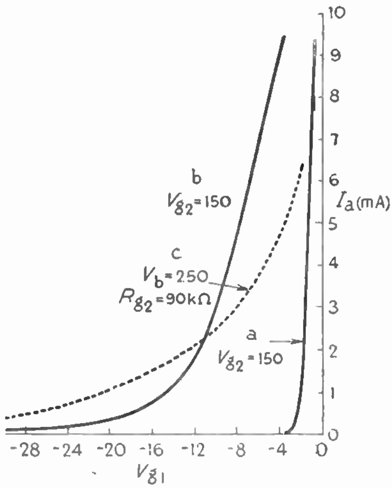



Fig. 19.19—$I_a/V_{g1}$ curves of ($a$) an ordinary screened pentode, ($b$) a variable-mu type, and ($c$) a variable-mu valve with $g_2$ fed through a resistor so that $V_{g2}$ rises as $-V_{g1}$ is increased

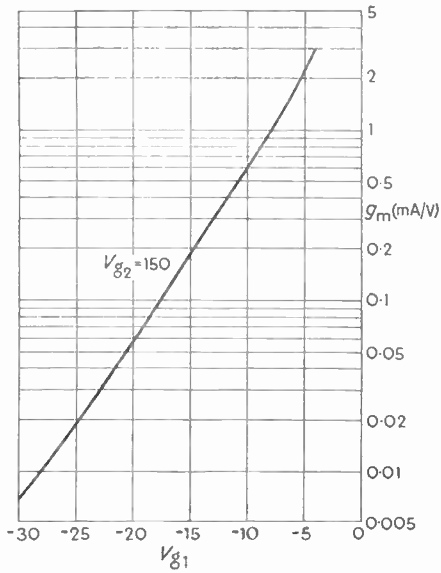



**Fig. 19.20—Mutual conductance plotted against $V_{g1}$ for the valve represented by *b* in Fig. 19.19**

It does show a slight tendency for $g_m$ to increase too fast at the upper end, however, and this tendency is undesirable if only because it goes along with a rapid increase in $I_a$, which may cause difficulties in the supply of current. It can very easily be avoided by feeding $g_2$ through a suitable resistor ($R_s$ in Fig. 17.9), which has the effect of reducing $V_{g2}$ as the negative bias is reduced. Increasing the bias reduces $I_{g2}$ as well as $I_a$ and so reduces the voltage drop in $R_s$, and $V_{g2}$ rises, tending to prolong the range of control. This effect is illustrated by curve *c* in Fig. 19.19, where instead of feeding $g_2$ direct from 150 V it is fed from 250 V through 90 kΩ. By suitable choice of resistance and supply voltage, the gain-control characteristic can be varied by the designer to meet requirements.

## Automatic Gain Control

In a perfect receiver the adjustment of r.f. gain to compensate for the different strengths of incoming signals would take place automatically. All would then (for equal depth of modulation) be reproduced at the same loudness, determined by the listener's setting of the volume control. The ordinary listener does not want to be bothered with the purely technical task of ensuring that the r.f. gain is always such that the various stages of the receiver are working under the best available conditions.

Elsewhere (p. 215) we noted that there was a reason why diode detectors were generally connected in such a way that the steady voltage produced by the rectified carrier-wave was negative. That reason now emerges, for a negative voltage increasing with signal

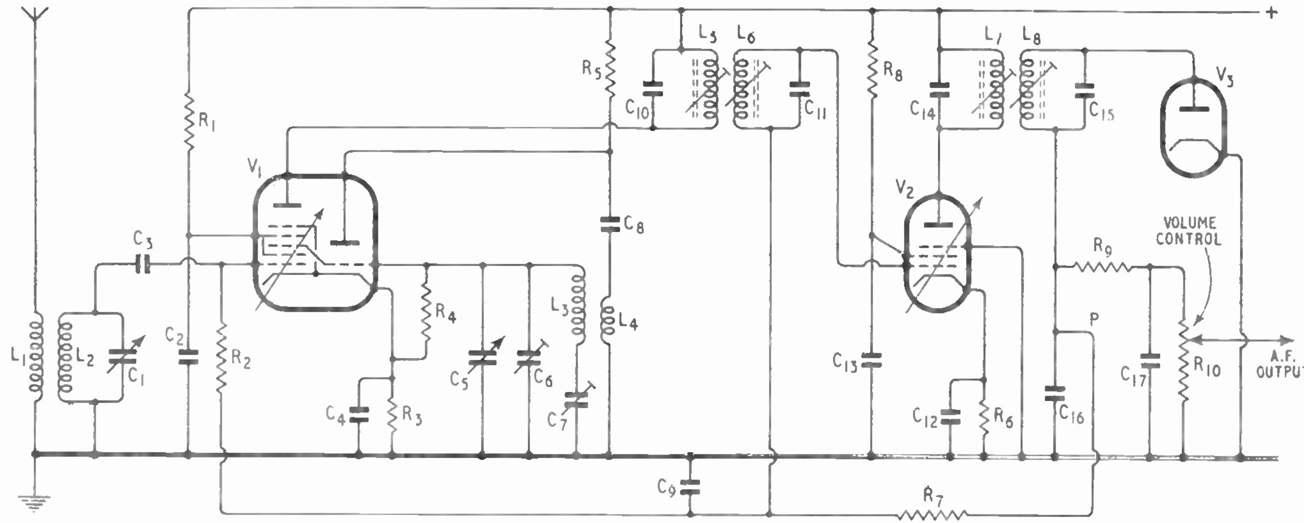



**Fig. 19.21—Circuit diagram of a typical superheterodyne receiver from aerial to volume control, including frequency-changer, i.f. amplifier and detector stages, with automatic gain control**

strength can be used as a gain-reducing bias—a device known as *automatic gain control* or a.g.c. (The name automatic volume control, a.v.c., still sometimes used, is clearly less appropriate.)

The signal level at the detector output must, in order to vary the gain, itself vary to some extent; so this system of a.g.c. is by no means perfect. But by various elaborations it can be made very nearly perfect. One of the commonest devices is to withhold the a.g.c. bias until the signal has attained a working level. Details of such elaborations can be found in books on receiver design.

Fig. 19.21, besides showing how simple a.g.c. is applied, will help us to review what we have learnt about receivers so far, for it is a circuit diagram of those parts of a typical receiver that have been considered in the last four chapters. It has been simplified by removing all switching and tuning circuits for other frequency bands, and though it may still look rather complicated it is just a gathering together of the separate sub-circuits we have already had. Before reading further it would be a good thing to attempt to identify these and explain to oneself the purpose of every component.

$V_1$ is a triode-heptode frequency-changer valve, and $V_2$ the i.f. amplifier. The arrow drawn through each is to indicate that they are both variable-mu valves and therefore suitable for operating on by a.g.c. bias. (Frequency-changers of the additive or single-input types cannot be included in the a.g.c. system, for the variable bias would upset the oscillator.) During reception, the potential of point P is affected in the Fig. 16.7*b* manner, yielding a mixture of r.f., a.f., and negative z.f. The r.f. is filtered out by $C_{16}$ $R_9$ $C_{17}$, and both r.f. and a.f. by $R_7$ $C_9$ so that only the negative z.f. voltage reaches the grids as a.g.c. bias.

The vision section of a television receiver cannot be controlled so simply, because the vision signal has no constant mean carrier amplitude; it varies with the average light and shade in the picture, and one does not want the differences between light and dark scenes to be ironed out. It is therefore necessary to select certain parts of the signal which are not reproduced on the screen and to use them to generate a.g.c. bias.

CHAPTER 20

# Audio-Frequency Amplification

## The Purpose of A.F. Stages

In all the parts of receivers discussed until now the duty of each valve has been to supply the drive to the next valve; that is to say, the signal voltage to its grid. Because valves are voltage-operated devices—the current in the grid circuit normally being so small that the power is almost negligible—the output of each preceding valve has been considered in terms of voltage. Transistors on the other hand are current operated. But when we come to the last stage in the receiver—the output stage—we have to take into account the ultimate purpose of the receiver. In those used for receiving sound, it is usually to work a loudspeaker. To stir up a sufficiently loud sound, appreciable power is needed; that is to say, both voltage and current. The output valve or transistor has therefore to be of such a type, and so operated, that sufficient audio-frequency power is available, without excessive distortion.

The output stage could be connected direct to the detector. This arrangement is sometimes adopted, but more often there is at least one stage of a.f. amplification between the detector and the output stage. In view of the statement that the diode detector works best when the signal voltage it is required to handle is greatest, such additional amplification might seem illogical. Where it is used there are usually one or both of two reasons. Firstly, the amount of a.f. power required may sometimes call for a greater signal voltage at the grid of the output valve than the r.f. stage before the detector can supply without distortion. (" R.f." in this chapter must be understood to include " i.f." where applicable.) It must be remembered that each a.f. volt from the detector requires more than one r.f. volt into the detector, because there are losses in the detector and its filter, and because the average depth of modulation is much less than 100%. Secondly, if the receiver is required to be very sensitive the necessary amount of amplification may be difficult to obtain at a single frequency or even when divided between r.f. and i.f. It might amount to a voltage gain of 10,000,000 times, which would necessitate extraordinary screening to avoid all risk of oscillation (p. 245). The problem is greatly eased by dividing this total gain between several frequencies; for example, 50 at r.f., 4,000 at i.f. (including the frequency-changer) and 50 at a.f.

Then, of course, very many a.f. amplifiers are needed for non-radio purposes; for example, deaf aids, sound reinforcement, disk and tape recording and reproduction, in which the microphone or pick-up generates far too little signal power to be used direct.

## Distortion

A.f. amplification is the same in principle as r.f. amplification, but the emphasis is rather different. The problem of selectivity is removed, but the problem of distortion is greatly increased. So before going further into the methods adopted in the a.f. end of a receiver, we had better consider distortion in general.

There are two main kinds. The first is called amplitude/frequency distortion (often just "frequency distortion"), because it means that the gain or loss of the system varies with frequency, so that the audio frequencies present in the original sounds are not reproduced at the correct relative strengths. One example that we have already discussed is excessive selectivity in the tuned circuits, causing the high audio frequencies to be reduced relative to the low.

The other kind of distortion is non-linearity. It causes the instantaneous amplitudes to be reproduced out of proportion so that the waveform is distorted. In general it is more unpleasant for the listener than amplitude/frequency distortion.

## Frequency Distortion

Information about the frequency distortion in any piece of equipment can be presented very clearly and comprehensively by means of a graph in which gain (or loss) is plotted against frequency. As perception of pitch (corresponding to audible frequency) is on a logarithmic basis the frequency scale is almost invariably logarithmic, usually running from 20 to 20,000 c/s in three 1 : 10 ranges. In this kind of graph we are more interested in the ratios of different gains, amplitudes, voltages or powers than their actual amounts,

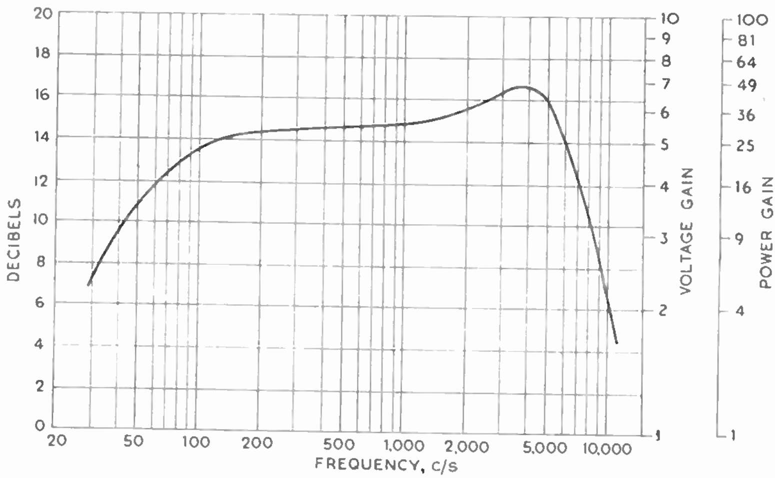



**Fig. 20.1—A typical a.f. gain/frequency curve, with logarithmic scales in both directions**

so they too are usually plotted logarithmically. As we saw in Chapter 11, if the units themselves are logarithmic—decibels—a linear scale can be used, as in Fig. 20.1, which shows a typical amplitude/frequency characteristic curve.

As a guide to interpreting frequency characteristics, it is worth noting that 1 dB is about the smallest change in general level of sound that is perceptible. If the loss or gain is confined to a small part of the audible frequency range, several dB may go unnoticed. So far as audible distortion is concerned, therefore, there is no sense in trying to " iron out " irregularities of a fraction of a decibel. A peak of one or two decibels, though unnoticeable as frequency distortion, may however cause non-linearity distortion by overloading the amplifier. An average fall of 10 dB over all frequencies above, say, 1,000 c/s would be easily noticeable as muffled and indistinct reproduction. A rise of the same amount, or a falling off below 1,000 c/s, would be heard as thin shrill sound.

A cheerful side to the subject is that frequency distortion in one part of the apparatus can usually be compensated elsewhere. For example, if in the pursuit of selectivity some falling-off in top notes has resulted, it can be put right in the a.f. amplifier by deliberately introducing distortion of the opposite kind, re-emphasizing the weakened parts. But such methods should not be relied on more than is really necessary, because the more compensation is needed the greater the risk that slight changes in either it or the original distortion will be noticeable.

Arrangements whereby the listener can adjust the frequency characteristic to suit his taste are called *tone controls*.

## Non-Linearity Distortion

Non-linearity (p. 34) can be expressed as a graph of output against input. A diagonal straight line passing through the origin, as in Fig. 2.3, represents perfect linearity, and means that any given percentage increase or decrease of input changes the output in exactly the same proportion. Valves are examples of non-linear devices, as shown by the curvature in their characteristics (Fig. 16.14, etc.). We have come up against this in connection with r.f. amplifiers (p. 284) and detectors (p. 222), and noticed that distortion is avoided if the action of the valve is confined to those parts of its characteristics that are practically straight. The bad effects of unavoidable curvature can sometimes be reduced by " swamping " the non-linear resistance of the valve with added linear resistance.

Judging from an output/input graph, the most obvious effect of non-linearity might seem to be that the volume of sound produced would not be in exact proportion to the original. It is true that reducing the signal strength at the detector to one-tenth is likely to reduce the a.f. output to far less than one-tenth, because of the " bottom bend " (p. 221). But that is about the least objectionable of the effects of non-linearity. What matters much more is the distortion of waveform. This has not bothered us very much until now, because we have been concerned mainly with the amplification of carrier waves, and the form of their individual cycles is unimportant (p. 285). Now that we have the a.f. waves, corresponding to those of the original sound waves, we must preserve their form with the utmost care. Otherwise the sound will be harsh and we shall not be able to distinguish one musical instrument from another (p. 17). If the waveform is sinusoidal, the sound has a round pure tone, like that of a flute blown softly. Fig. 1.2*c* shows samples of three sinusoidal waves, differing in frequency, and therefore heard with different pitch, but all having the same purity of tone. The fourth waveform is the result of combining the first three, and would be heard as a less smooth tone. Now if the first sample is an original waveform, as broadcast, and the fourth is the result emerging from the loudspeaker, distorted by non-linearity somewhere, then the reproduced sound is not a true copy of the original. The thing to note is that the reproduction contains frequencies (namely, those of the second and third samples) not present in the original.

## Generation of Harmonics

This point is so important that it is worth examining more closely. Fig. 20.2 is an example of a non-linear valve curve, and Fig. 20.3*a* is one cycle of a sine wave applied to it. The scale of grid voltage at its side enables us to plot the resulting wave of anode current (Fig. 20.3*b*). Obviously it is distorted; the positive half-cycle is somewhat pointed, and the negative half is flattened. At *c* this distorted wave is repeated for close comparison with a pure sine wave (shown dotted) having the same peak-to-peak voltage. The

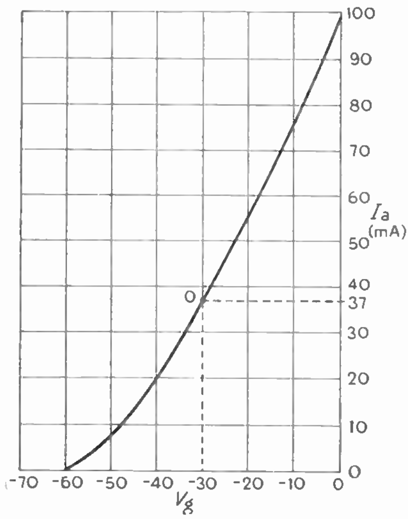



**Fig. 20.2**—Typical characteristic curve of a triode, showing non-linearity. If the working point is O (as a result of applying a grid bias of −30 V), a 30 V positive half-cycle of signal on the grid would increase the anode current by 62 mA, but an equal negative half-cycle would decrease it by only 37 mA; see Fig. 20.3

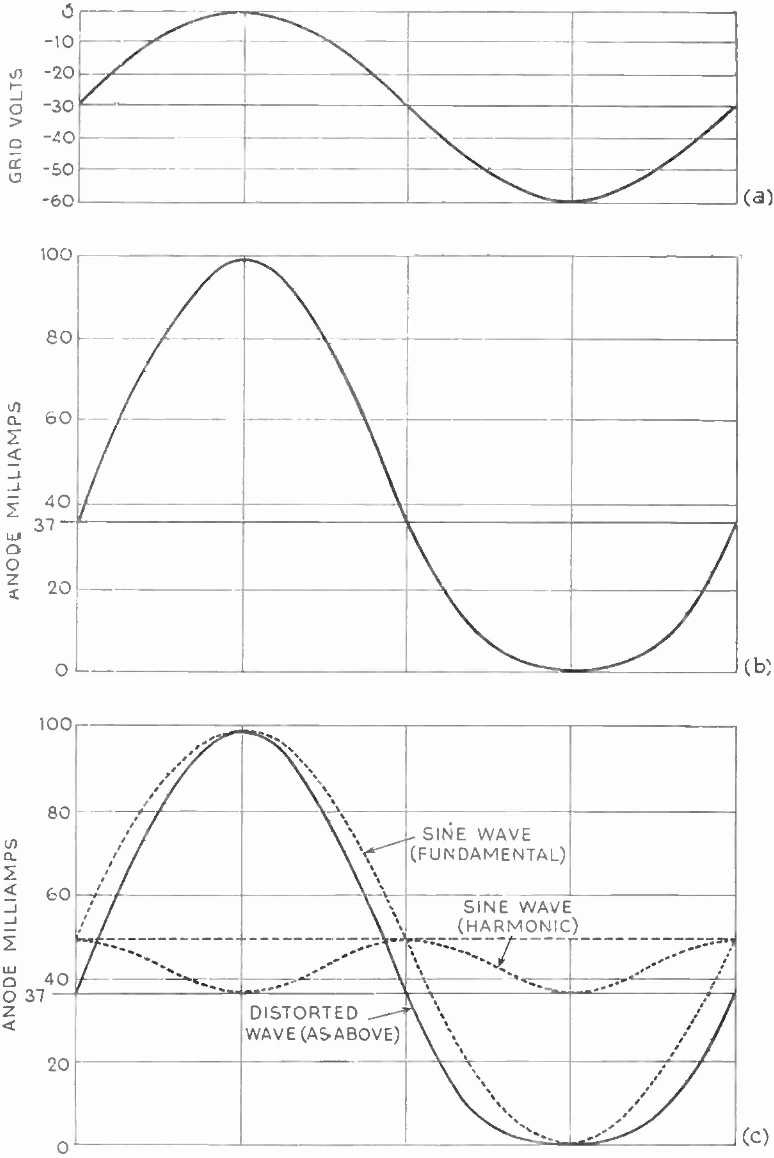



Fig. 20.3—Showing that the distorted output (*b*) resulting from applying the input (*a*) to the valve whose characteristics are given in Fig. 20.2 is equivalent to (*c*) the combination of an undistorted wave with its second harmonic

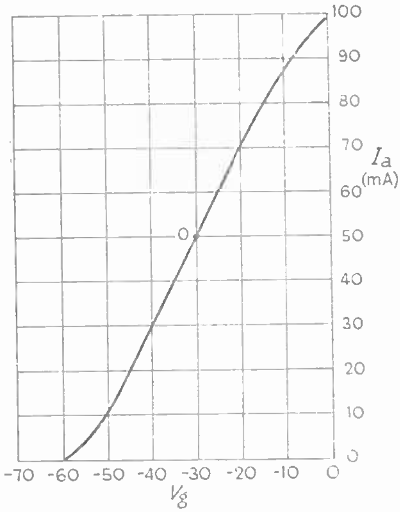



**Fig. 20.4—Non-linearity of this kind does not cause second-harmonic distortion, for the half-cycles of anode current resulting from equal half-cycles of grid voltage are also equal; but it does introduce third-harmonic distortion**

difference between the two (i.e., the result of distortion) is also plotted as a dotted line, and turns out to be a sine wave of twice the frequency. This double-frequency wave is called the *second harmonic* of the original wave, which is the first harmonic, more commonly called the *fundamental*. The peak-to-peak amplitude, or *swing*, of the second harmonic in this example is about 13 V, compared with about 100 V fundamental. The amplifier is therefore causing about 13% second-harmonic distortion.

Second-harmonic distortion is characteristic of the type of non-linearity shown in Fig. 20.2, in which the slope continually increases or decreases. The steeper slope in one direction gives that half-cycle a bigger peak, and the smaller slope in the other direction blunts the other half-cycle. Non-linearity of the kind shown in Fig. 20.4, working from point 0, would blunt both half-cycles, and is found to introduce a frequency three times that of the original—the third harmonic. Usually the characteristic curves of valves are less simple; one generally finds both second and third harmonic, accompanied by progressively weaker fourth, fifth, sixth, etc.

The ear tolerates a fairly large percentage of second-harmonic distortion, less of third harmonic, and so on. A fraction of 1% of the higher harmonics such as the eleventh introduces a noticeable harshness; the reason being that a second harmonic differs by an exact octave from the original tone, whereas the high harmonics form discords.

## Intermodulation

Although for simplicity the distortion of a pure sine wave has been shown, the more complicated waves corresponding to musical instruments and voices are similarly distorted. A most important principle, associated with the name of Fourier, is that all repetitive waveforms, no matter how complicated, can be analysed into pure sine waves, made up of a fundamental and its harmonics, which are all exact multiples of the fundamental frequency. So when one is dealing with waveforms other than sinusoidal it is usual to perform this analysis, and then tackle them separately on the basis of simple sine-wave theory; like the man in the fable who found it easier to break a bundle of faggots by undoing the string and attacking them one by one. Each harmonic current and voltage can be calculated as if it were the only one in the circuit—*provided that the circuit is linear*. But just now we are studying non-linear conditions, in which it is not true to say that two currents flowing at the same time have no effect on one another. If it were, then the frequency-changer would not work. We found (p. 265) that when two voltages of different frequency are applied together to a non-linear device such as a rectifier the resulting current contains not only the original frequencies but also two other frequencies, equal to the sum and difference of the original two. That is a very useful result in a frequency-changer or in a modulator, but extremely undesirable in an a.f. amplifier. It means that all the frequencies present in the original sound—and there may be a great many of them when a full orchestra is playing—interact or *intermodulate* to produce new frequencies; and, unlike the lower harmonics, these frequencies are generally discordant. Even such a harmonious sound as the common chord (C, E, G) is marred by the introduction of discordant tones approximating to D and F. When non-linearity distortion is severe, these intermodulation or combination tones make the reproduction blurred, tinny, harsh, rattling, and generally unpleasant. And once they have been introduced it is impracticable to remove them.

## Allowable Limits of Non-Linearity

Whenever one attempts to get the greatest possible power from a valve, one comes up against non-linearity. It is, in fact, what limits the power obtainable. So naturally the problem is most acute in the output stage. If the signal amplitude is reduced, so that it operates over the most nearly linear parts of the valve characteristics, the distortion can be made very small, but then the output will be uneconomically small in relation to the power being put into the valve. It is not a question of adjusting matters until distortion is altogether banished, because that could be done only by reducing the output to nil. It is necessary to decide how much distortion is tolerable, and then adjust for maximum output without exceeding that limit.

Fixing such a limit is no simple matter, because it should (in reproduction of music, etc.) be based on the resulting unpleasantness judged by the listener. Different listeners have different ideas about this, and their ideas vary according to the nature of the programme. And, as has just been pointed out, the amount of distortion that can be heard depends on the order of the harmonic (second, third, etc.) or intermodulation tones, as well as on the percentage present.

Various schemes for specifying the degree of distortion in terms that can be measured have been proposed, but the nearer they approach to a fair judgment the more complicated they are; so the admittedly unsatisfactory basis of percentage harmonic content is still commonly used. Preferably the percentage of each harmonic is specified, but more often they are all lumped together as " total harmonic distortion ".

We shall see later how the percentage harmonic distortion can be found, at least approximately, from the valve curves.

## Phase Distortion

The main result of passing a signal through any sort of filter, even if it is only a resistance and a reactance in series, is to alter the amplitudes of different frequencies to different extents (p. 217). If it discriminates appreciably between frequencies within the a.f. band it introduces frequency distortion, which, of course, may be intentional or otherwise. It also shifts the phases to a differing extent. If the signal is a pure sine wave, the phase shift has no effect on the waveform; but few people spend their time listening to continuous pure sine waves.

To take a very slightly more complex example, a pure wave is combined with a third harmonic, shown separately in Fig. 20.5*a*. The waveform of this combination, obtained by adding these two together, tends towards squareness, as shown at *b*. Suppose now that after passing through various circuits the relative phases of the component waves have been altered, as at *c*. Adding these together, we see the resulting waveform, *d*, which is quite different from *b*. Yet oddly enough the ear, which is so sensitive to some changes of waveform (such as those caused by the introduction of new frequencies), seems quite incapable of detecting the difference between *d* and *b*, or indeed any differences due to phase shifts in continuous waves.

One has to be much more careful about phases when dealing with television signals, however, because the result is judged by the eye, to which form is of paramount importance. Phase distortion such as that shown in Fig. 20.5 would be very objectionable.

Generally, it can be said that if care is taken to avoid phase distortion, frequency distortion will take care of itself.

## Distortion in Resistance-Coupled Amplifiers

Coming now to apply this general information about distortion to actual amplifiers, we begin with a.f. voltage amplification, of which the typical example is the stage between detector and output valve in a receiver. Almost invariably resistance coupling is used, and we have already (in Chapter 11) discussed this method fairly fully, at least as regards stage gain and choice of load resistance. The question of distortion now arises.

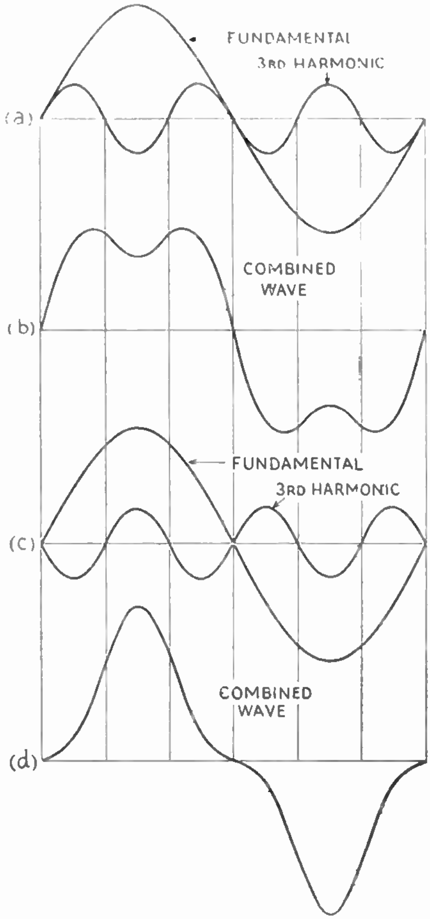



**Fig. 20.5**—Showing that the waveform resulting from adding together two component waves is greatly altered by shifting the relative phase of the components

In most sets the voltage needed to drive the output valve fully is not more than about 10 V, so the a.f. amplifier does not have to work very hard, and there is seldom any difficulty in keeping non-linearity within tolerable limits. This kind of distortion will therefore be considered in connection with the output stage where it is the main problem. We shall concentrate now on frequency distortion.

When (in Chapter 17) we tried resistance-coupled amplification for radio frequencies, we ran into two troubles, due to the stray capacitances from anode to grid and to earth respectively (Fig. 17.2). The lower the frequency the less the current that can pass via a given capacitance, so we would expect both of these troubles to be far less at audio frequencies. In combination, however, they could be serious; for we saw that when the anode load of a valve is resistive the effect due to the anode-to-grid capacitance $C_{ag}$ (Miller effect) is to add an input capacitance to the valve equal to $(A + 1)\,C_{ag}$, where $A$ is the voltage gain. This adds to the anode-to-earth capacitance $C_{ae}$ of the previous stage (if any) and with a triode's $C_{ag}$ might easily make it many times greater—sufficient to be troublesome where $C_{ae}$ alone would not be.

For this reason, and perhaps even more for getting a larger gain, pentodes are almost always used instead of triodes for a.f. amplifiers. Their Miller effect is quite negligible at a.f., so only $C_{ae}$ need be

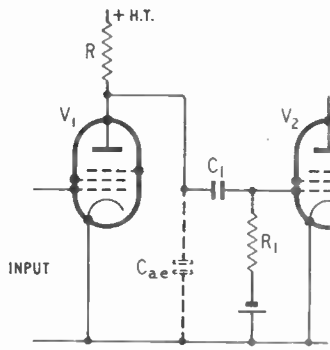



**Fig. 20.6—Simple form of resistance-coupled amplifier. $C_{ae}$ denotes unavoidable stray capacitance**

considered. Fig. 20.6 shows a complete stage, in which $C_{ae}$ includes all the capacitance effectively in parallel with the coupling resistance between the valves; the input capacitance of $V_2$, for instance. Suppose we take 30 pF as an outside figure for $C_{ae}$; with a little care it could be made smaller.

The table below shows its reactance at various frequencies, calculated from $X_c = 1/2\pi fC$.

| $f$, *in c/s* | $X_c$ *in k*$\Omega$ |
|---|---|
| 100 | 53,000 |
| 1,000 | 5,300 |
| 10,000 | 530 |
| 100,000 | 53 |

At 100 c/s its reactance, 53 M$\Omega$, has a quite negligible effect. Even at 1,000 c/s it is not likely to do more than slightly shift the phase. But if, with the object of getting a high stage gain, we were to use a fairly high value of $R$, there might be an appreciable loss at 10 kc/s, due to a reduction in the effective coupling impedance.

We know (p. 144) that the stage gain $A$ is $\mu R/(r_a + R)$, and that $\mu = g_m r_a$ (p. 122). Combining these we get

$$A = \frac{g_m r_a R}{r_a + R}$$

that is to say, $A$ is proportional to the resistance of $R$ in parallel with $r_a$ (p. 38). But if this resistance combination is shunted by an equal reactance, the total impedance is reduced to $1/\sqrt{2}$ or 70·7% of what it was (p. 88), and the gain is reduced accordingly. So it follows that high notes of frequency $f$ are reduced to 70·7% as compared with low frequencies—a 3-dB loss—when $1/2\pi fC = Rr_a/(R + r_a)$; that is to say, when the reactance of the stray capacitance is equal to the anode resistance of the valve in parallel with the load resistance.

On the assumption that the $r_a$ of the pentode is 1·5 MΩ and $R$ with $R_1$ in parallel is 0·5 MΩ, Fig. 20.7 is the frequency characteristic curve (compare Fig. 16.11). Any reduction in reactance (increase in $C_{ae}$) or increase in $R$ or $r_a$ leads to greater proportionate loss of high notes. The curve is not altered in shape thereby; it is merely shifted to the left. It will be clear that where a high capacitance is inevitable (as in long screened lines to a distant amplifier), or the frequency is very high, choosing a valve of low anode resistance, with a coupling resistance of low value, helps to keep the loss of higher frequencies within reasonable bounds. A better method for special purposes is described on p. 315.

In television, video-frequency (v.f.) amplifiers are required to cover all frequencies from zero to several Mc/s; and it is necessary to use very low coupling resistance, and to compensate as much as possible for the reduction in gain on that account by using valves having especially high slope.

The grid capacitor and leak, $C_1$ and $R_1$ in Fig. 20.6, form a potential divider across the source of amplified voltage (anode of valve to earth; remembering that " + H.T." counts as earth

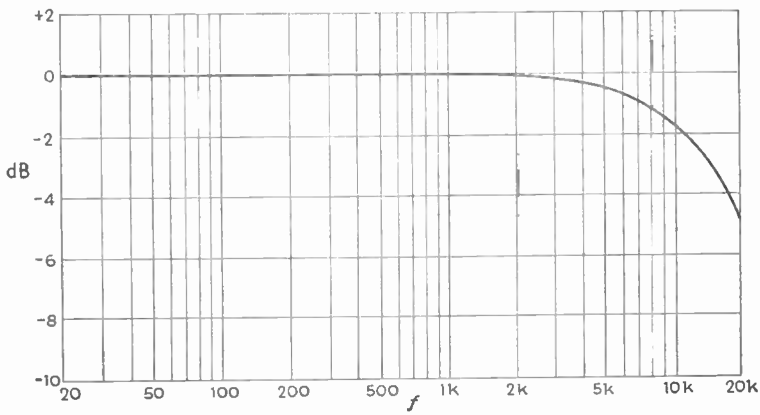



**Fig. 20.7—Frequency curve of the amplifier of Fig. 20.6 when $r_a$ of $V_1$ is 1·5 MΩ, $R$ and $R_1$ in parallel is 0·5 MΩ, and $C_{ae}$ is 30 pF, showing falling-off (in decibels) at the top audio frequencies due to $C_{ae}$**

so far as signals are concerned). Only the signal voltage appearing across $R_1$ reaches the grid of the next valve. This time it is the lowest frequencies that are reduced, by $C_1$ acting as an impedance in series. (Compare the filter $R_2C_2$ in Fig. 16.12.) The shape of the frequency characteristic is exactly the same as that in Fig. 20.7, except for being reversed, left to right. And again the position of the 70·7% point (= − 3 dB) can be found, by putting $R_1 = 1/2\pi f C_1$. With p. 56 in mind, another way of stating the same condition is to say that the time constant, $C_1R_1$, is $1/2\pi f$—roughly one-sixth of the time period of one cycle of the voltage.

Since the − 3 dB point can be regarded as marking the frequency where the loss begins to be noticeable, this " reactance = resistance " criterion is a very useful one to remember for both low- and high-note loss.

A usual combination is $C_1 = 0{\cdot}01\ \mu F$, $R_1 = 0{\cdot}5\ M\Omega$, with which the − 3 dB frequency is about 32 c/s.

## Transformer Coupling

Resistance coupling is cheap and compact, and by suitable choice of components can easily be made to have negligible frequency distortion over the useful a.f. band; but the mean anode voltage is inevitably less than that of the h.t. source because of the drop in the resistance $R$, and this reduces the signal output obtainable within the limits of reasonable non-linearity. A higher output

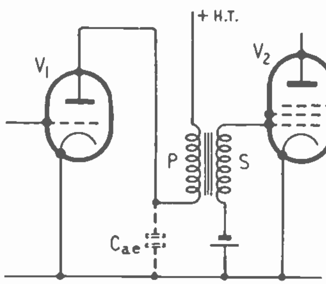



**Fig. 20.8 — Simple transformer-coupled amplifier circuit**

voltage can be obtained by transformer coupling (Fig. 20.8), not only because the resistance of its primary winding is relatively low, but also because a voltage step-up is possible.

We can regard this primary (P) as taking the place of R in Fig. 20.6. Unlike R, its impedance depends very greatly on frequency; in fact, it is really a rejector circuit (p. 111), its inductance being " tuned " by the stray capacitance $C_{ae}$. This fact might seem utterly incompatible with reasonably uniform amplification over the whole a.f. band; and certainly if a pentode were used, so that gain was nearly proportional to the coupling impedance (p. 239), the prospect for avoiding frequency distortion would be very poor. But the full-line curve in Fig. 11.4 shows that when the coupling resistance is much greater than the valve resistance it can vary over quite a wide range without much effect on gain. The same applies to coupling impedance in general. So by using a low-$r_a$ valve—a triode—and taking care to keep the impedance of the coupling relatively high at all the frequencies to be amplified, frequency distortion need not be excessive.

At the top end, $C_{ae}$ is again the limiting factor; and unfortunately the self-capacitance of a transformer winding is likely to be much greater than that of a resistor. Moreover, the capacitance of the secondary winding is equivalent to $n^2$ times as much across the primary, where $n$ is the step-up ratio (p. 97). The more turns that are used on the coils to keep up the impedance at low frequencies, the greater $C_{ae}$. So skilled design is needed in order to achieve a satisfactory frequency characteristic. A transformer is inevitably larger and much more expensive than a suitable resistor, and the benefit of a voltage step-up is almost if not entirely lost by having to use a low-$\mu$ valve. For these and other reasons it is now more usual to use a second resistance-coupled stage in the relatively few cases where more gain is required than a single stage can provide.

In transistor a.f. amplifiers (p. 348) there is some scope for coupling transformers, because of the need to match the output impedance of one to the much lower input impedance of the next, thereby greatly increasing the current amplification.

## The Output Stage

As a load, a loudspeaker can be regarded (very approximately) as a low resistance, usually between 1 and 20 Ω, with a certain amount of inductance in series. If it were connected directly in the anode circuit of even the lowest-resistance valve, the system (considered as an equivalent-generator circuit, Fig. 11.3) would be very inefficient, because nearly all the power would be wasted in the resistance of the valve. Sometimes a special high-resistance speaker is connected direct to a low-resistance amplifier; but except for this the speaker impedance is matched to the valve

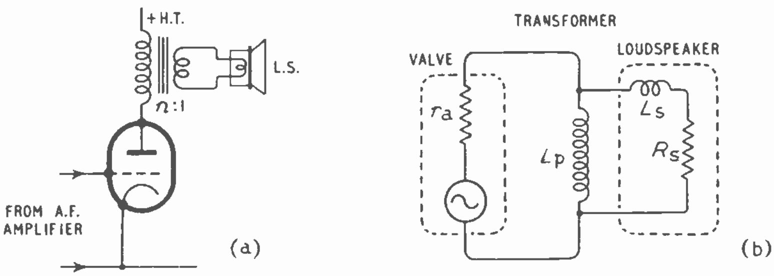



**Fig. 20.9—(*a*) Circuit of output stage in its simplest form. Electrically this can be approximately represented by the generator circuit (*b*)**

(p. 98) by means of a transformer—the output transformer. We can therefore assume that the load can be made to have any resistance to suit the output valve. The primary of the transformer introduces a parallel inductance, so the output stage as a whole, connected as in Fig. 20.9*a*, is roughly equivalent to Fig. 20.9*b*, in which the resistance and inductance $L_s$ and $R_s$ are $n^2$ times the actual resistance and inductance of the loudspeaker.

Next to giving it the correct ratio, the most important thing about the transformer is to make sure that the inductance of its primary ($L_p$) is high enough not to by-pass the signal current at the lower frequencies. On the principle we have already noted for intervalve reactances (p. 302), it is usual to make the reactance of $L_p$ equal to $R_s$ at the lowest wanted frequency; i.e., $2\pi f L_p = R_s$.

The effect of $L_s$ is negligible at low frequencies, and at high frequencies depends somewhat on $r_a$, but in general it causes high-note loss.

The main problem, however, is to find the best value for $R_s$ (which really means the best transformer ratio) for any given valve. On p. 144 we found theoretically that, for a given fixed signal voltage on the grid, the greatest power was obtained in the load by making the load resistance equal to the generator resistance. But that was on the assumption that the valve was a perfect linear generator, which it certainly is not. In practice this general law has to be modified to allow for the bends and curves of the valve characteristics. The usual way of doing this is by making use of the actual valve characteristic diagram, fitting a load line to it, as explained on p. 141. To avoid awkward complications it is assumed that the frequency is neither so low nor so high that $L_p$ or $L_s$ need be taken into account, and therefore that the load is a pure resistance and so can be represented on the diagram as a straight line.

There is one modification to be made to the method as originally

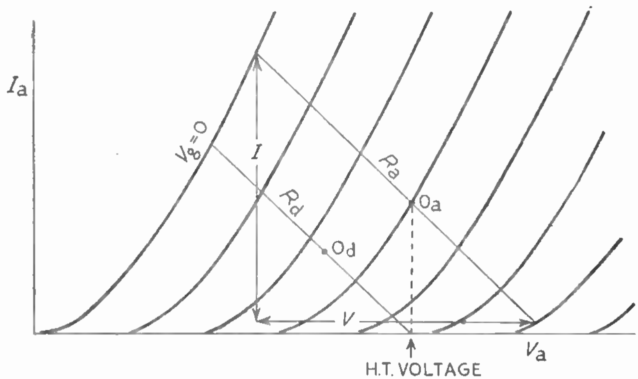



**Fig. 20.10—$R_d$ is a typical load line for a direct resistance load. $R_a$ represents a load having negligible resistance to d.c. and the same resistance to a.c. as $R_d$**

explained. Then we were dealing with a resistance load directly in series with the valve, so that the steady feed current as well as the alternating signal caused a voltage drop, bringing the mean voltage at the anode well below that of the source. Now we are feeding the valve through a transformer winding, having a resistance so low that for approximate purposes it can be neglected, and the a.c. resistance is what is transferred into it from the secondary by transformer action. This difference in conditions is represented by making the initial or working point start at a value of $V_a$ equal to the h.t. source voltage, as shown in Fig. 20.10. Here $R_d$ is the load line for a directly-connected load, with one end pivoted at the h.t. voltage, as in Fig. 11.2. The working point, $O_d$, is normally about half-way along the line, reckoned in terms of grid voltage, so that the cycles of input signal voltage swing the grid up to $V_g = 0$ and down to anode-current cut-off. A purely a.c. load, such as a resistance connected by a perfect transformer, is represented by a load line such as $R_a$; and it is easy to see how avoiding the d.c. voltage drop enables a much larger output voltage and current to be obtained from a given valve and h.t. voltage. The peak-to-peak signal voltage across the load is marked $V$, and the corresponding current $I$; and we found (p. 161) that the power output is $VI/8$. The problem is so to place the load line that the maximum power is obtained, provided that:

(1) the working point is not beyond the maximum rated anode voltage and current;

(2) the top end does not run into grid current;

(3) the bottom end does not more than cut off anode current;

(4) the specified non-linearity is not exceeded.

## Optimum Load Resistance

The first three conditions are easy to observe: No. 4 is the trouble. The usual procedure is to mark the working point $O_a$ at the maximum allowable (or available) $V_a$ and $I_a$. The necessary grid bias is indicated at once if $O_a$ happens to fall on one of the curves; otherwise it must be estimated from the nearest curves on each side. Then the load line is rotated about $O_a$ until equal grid voltages measured from $O_a$ up and down the line reach the start of grid current at the top end (roughly $V_g = 0$, but actually a volt or so negative) and slightly above $I_a = 0$ at the other.

Then comes the tricky job of deciding whether the result is within the tolerable limits of distortion. With typical triode characteristics, the positive peaks of output current turn out larger than the negative, as we saw when looking at Fig. 20.3. Assuming, as we generally can with such characteristics, that the distortion is predominantly second-harmonic, we can see from a closer look at Fig. 20.3*c* (repeated here as Fig. 20.11) that the amount by which the distortion increases the positive half-cycle is equal to twice the peak

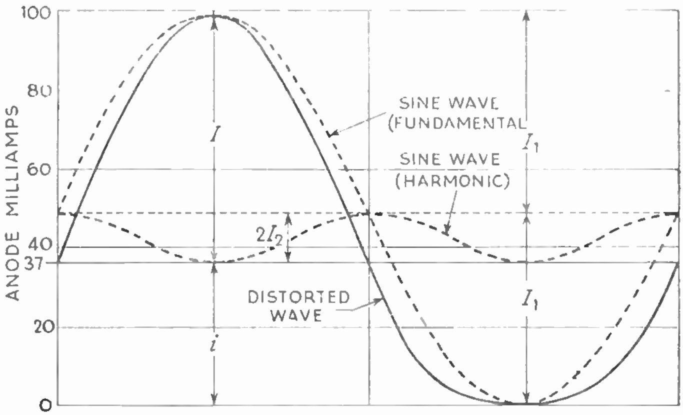



**Fig. 20.11—Fig. 20.3 *c* again, with symbols added for convenience in deriving a formula for percentage second-harmonic distortion. $I_1$ is the peak value of the fundamental current, and $I_2$ of the harmonic; $I$ is the peak current of the larger half-cycle of the distorted current, and $i$ the smaller**

current of the harmonic, and that the negative half-cycle is reduced by the same amount. In the symbols of Fig. 20.11,

$$I = I_1 + 2\,I_2$$

$$\text{and } i = I_1 - 2\,I_2$$

Subtracting the second from the first,

$$I - i = 4\,I_2$$

And adding them,

$$I + i = 2\,I_1$$

Dividing $I_2$ by $I_1$ thus obtained, we have:

$$\frac{I_2}{I_1} = \frac{I - i}{2(I + i)}$$

Multiplied by 100, this is the percentage second harmonic. This result leads to a simple method of finding the percentage harmonic distortion from the valve curve sheet. An example will help to make it all clearer.

Suppose the valve characteristics are as in Fig. 20.12 and that the maximum rated anode voltage and current are 250 V and 37 mA. The working point (O) is marked at this rating, and at once shows the necessary grid bias to be $-$ 30 V. In the absence of any guidance in the matter, let us try to find the best load resistance. To meet condition 2, the peak signal voltage should be + 30 V or slightly less; and to meet condition 3 the equal negative peak should arrive at or slightly above zero $I_a$. We rotate a straight load line about O until it complies with these conditions, as it does in the position BOC. The slope of this line shows it to represent a load resistance of just over 2 kΩ. It is, in fact, the load line from which Fig. 20.2 was derived.* But we saw that the output waveform resulting from an input of 30 V peak was very noticeably distorted, and calculated that the second-harmonic distortion was about 13%. We can now do this more easily direct from Fig. 20.12. $I$, the positive peak of anode current, represented by the height of B above O, is $99 - 37 = 62$ mA; and $i$, the negative peak, represented by the height of O above C, is $37 - 0 = 37$ mA. So the percentage second-harmonic distortion, $100\ (I - i)/2(I + i)$, is $50 \times 25/99 = 12{\cdot}6$.

If we want to restrict the distortion to 5%, then we put $50(I - i)/(I + i) = 5$, which by a little algebra yields $I/i = 11/9$. So it is

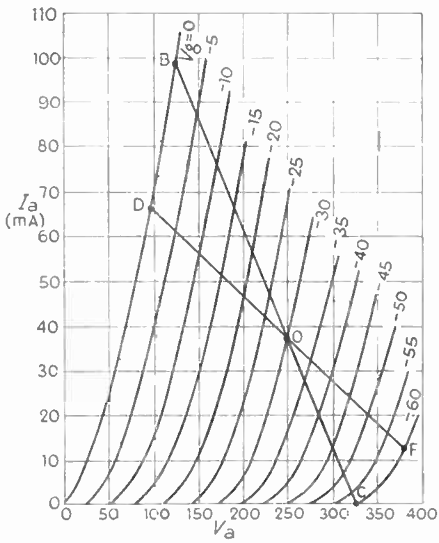



**Fig. 20.12—Two trial load lines drawn on a curve sheet relating to a triode output valve. With load line BOC the distortion is excessive (see Fig. 20.3, which was derived from this). The higher resistance represented by DOF is better**

necessary to make sure that the amplitudes of the output half-cycles (for equal input half-cycles) are in a ratio not exceeding 11 to 9. With the load resistance represented by BOC, the greatest input within this restriction is about 15 V, giving a total output swing of 110 V and current swing 53 mA, the output power therefore being $110 \times 53/8 = 730$ mW only. The fact that the signal amplitude has to be so much restricted in order to keep within the 5% limit shows that the load resistance was incorrectly chosen. Going

* Fig. 20.2 is therefore what is called a *dynamic* characteristic curve: that is to say, a curve in which account is taken of the load resistance as well as $r_a$.

through the same process of drawing load-line, investigating permissible grid-swing before the distortion limit is reached, and calculating from the current and voltage swings the power delivered to the speaker, enables us to find the power that can be delivered into each of a series of loads of different resistance. The results are given as a curve in Fig. 20.13.

The *optimum load*, being that into which the greatest power can be delivered, is evidently about 5·2 kΩ. The corresponding load-line is drawn as DOF on Fig. 20.12. To achieve this power the grid requires a signal that swings it from 0 to − 60 V, giving a swing in anode current from $12\frac{1}{2}$ to 67 mA. The two amplitudes

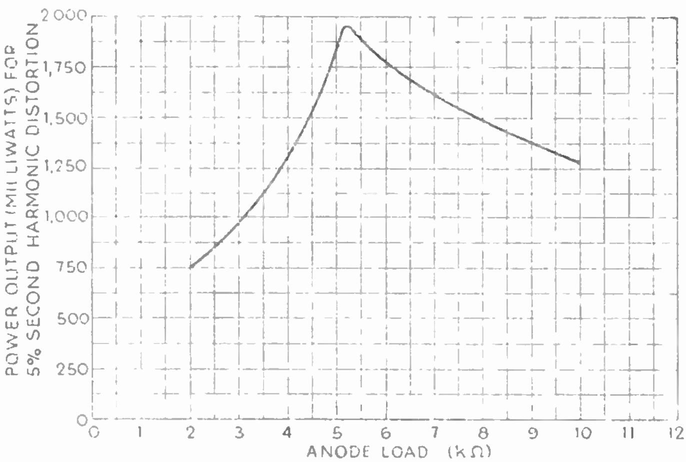



**Fig. 20.13—The results of calculating the output power at 5% second-harmonic distortion in Fig. 20.12 for various values of load resistance**

from O are in the ratio 9 to 11, showing that 5% second-harmonic distortion has just been reached. The power available for the loudspeaker is now

$$\frac{(67 - 12\frac{1}{2}) \times (378 - 94)}{8} = \frac{54\frac{1}{2} \times 284}{8} = 1{,}935 \text{ mW}.$$

This result, it must be remembered, applies only to point O, representing maximum voltage and current ratings for the valve. At other points, the optimum load is likely to be different.

In general, the user of a valve is not obliged to go through this elaborate examination of valve curves, for the makers' recommendations as to anode voltage and current, grid bias and optimum load, are set forth in the technical data they supply on request. This is just as well, because actually, although there is a middle range of frequencies over which a loudspeaker is an approximately resistive load, at the extreme frequencies the reactance predominates, and this makes the matter too complex to deal with here. The resistive load line is, all the same, a useful guide.

In providing the optimum load the user can do no more than ask the maker of his chosen loudspeaker to supply it with a transformer suited to the valve he proposes to use. The ratio of the transformer (p. 98) should be $\sqrt{(R_p/R_s)}$ where $R_p$ and $R_s$ are respectively the optimum load and the mean impedance of the moving coil.

## Output Tetrodes and Pentodes

The triode we have just been considering gave, at best, 1·935 W output. To achieve this it had to be supplied with 37 mA at 250 V, which is 9·25 W. Actually the drain is rather larger when the valve is delivering full power, because the current rises slightly as a result of the distortion. Even if that is left out of account, the power efficiency is $1{\cdot}935/9{\cdot}25 = 21\%$. So only about one fifth of the power put in is made useful. For this reason triodes are seldom used as output valves.

Looking at Fig. 20.10 we can see why the efficiency of a triode is low: the swing of the anode voltage from the working point $O_a$ is limited by the " $V_g = 0$ " line when it is still far above zero. When considering pentodes we saw (Fig. 17.8, dotted line) that the $I_a/V_a$ curve rises steeply from zero and then flattens out. Although at that time we were not interested in power efficiency, we see now that such a shape is almost ideal in permitting a large signal voltage swing for a given input power. But clearly some modification is required to increase $I_a$ from the very few milliamps shown in Fig. 17.8 to the many needed to give an output of several watts. On the other hand, there is no need to screen the anode from the control grid so thoroughly as in a r.f. pentode.

An a.f. output pentode is therefore larger and has wider grid-wire spacings; Fig. 20.14 shows a set of typical $I_a/V_a$ curves. Although the anode resistance ($r_a$) is lower than in a r.f. type of pentode, the nearly horizontal curves show that it is very much higher than in a comparable triode. The spacing of the lines indicates a normal $I_a/V_g$ slope ($g_m$), so the amplification factor ($\mu$) is high. Compared with the triode, then, the pentode offers two advantages: not only is it more efficient, in that a greater proportion of the power drawn from the h.t. supply is convertible into a.c. power, but it is more *sensitive*, meaning that a volt of signal applied to its grid releases a larger output.

If all tetrodes had the secondary-emission kink in the curves as shown in Fig. 17.7 they would obviously be even more unsuitable as output valves than for r.f. amplification. But by various design expedients, such as careful adjustment of the distance from $g_2$ to anode, and constricting the electrons passing through the control grid into beams, like rays of light through a grating, with the $g_2$ wires aligned to come in the " shadows " so that they catch very few electrons, it has been found possible to establish a space charge between $g_2$ and anode which serves the purpose of the suppressor

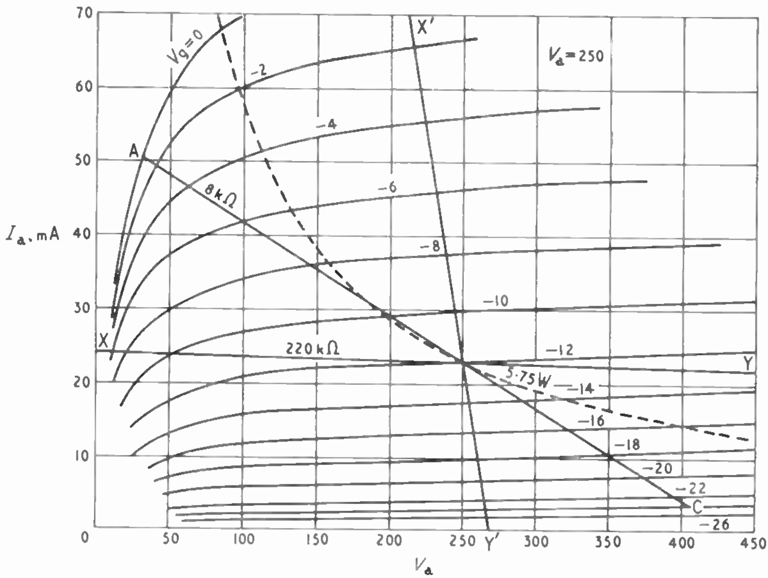



Fig. 20.14—Typical $I_a$–$V_a$ curves for a small output pentode. The line AOC represents a usual load

grid ($g_3$) in a pentode. The characteristics of these *beam tetrodes* are almost indistinguishable from those of output pentodes, and in what follows " pentodes " should be understood to include these " kinkless " tetrodes.

The optimum load resistance for a triode is nearly always greater than its $r_a$; often about double. Let us see how this works out with our specimen pentode, Fig. 20.14. At a likely working point O ($V_a = V_{g2} = 250$ V, $V_{g1} = -12$ V, $I_a = 23$ mA) the $r_a$ indicated by the slope of the valve curve is some 110 kΩ. A load-line drawn through O to represent 220 kΩ is XOY. Towards X it cuts the curves for $V_{g1} = -10$ to $V_{g1} = 0$ in rapid succession, while towards Y it is far from reaching the curves for $V_{g1} = -14$ to $V_{g1} = -24$. With a load such as this, a signal swinging the control grid from 0 to − 24 V would clearly result in the most appalling non-linearity distortion, together with phenomenally high voltage peaks at the anode.

At the opposite extreme the load line X'OY', representing a very low resistance, cuts the curves for high bias in rapid succession, while the intercepts with the low bias curves are widely spaced. Since these two types of distortion occur at opposite ends of the signal swing, it is reasonable to expect some intermediate load to be best. We reach the same conclusion if we consider the power developed, for with load XOY there is a large voltage swing and very little current, while with X'OY' the result is the opposite. For both voltage and current to be reasonably large, an intermediate load resistance is needed.

Let us try 8 kΩ, represented by the line AOC. The power delivered to this load when a signal swings $V_{g_1}$ from 0 to − 24 V can be obtained, as with the triode, by multiplying together the peak-to-peak current and voltage swings and dividing by 8:

$$\frac{(405 - 30) \times (50{\cdot}5 - 2{\cdot}5)}{8} = \frac{375 \times 48}{8} = 2{\cdot}250 \text{ mW}.$$

If the power supplied to the valve is reckoned as $250 \times 23 = 5{\cdot}75$ W, the efficiency works out at $2{\cdot}25/5{\cdot}75 = 39\%$. In this case not only will the anode current be rather more than 23 mA when full power is being delivered, but one ought to add $I_{g2}$, which would probably be about 4 mA. Even so, the efficiency is notably higher than the triode's.

## Harmonic Distortion in the Pentode

How about distortion? With the triode, as we have seen, second-harmonic distortion predominates. With the pentode we have to take into account distortion resulting in both second and third harmonics of the original signal.

A load resistance can generally be found which makes the positive and negative half-cycles equal, so that second harmonic is almost or entirely absent; third harmonic then predominates, and can be calculated by a method somewhat similar to that described on p. 306 for second harmonic. This is not necessarily the best load resistance, however; and the calculation or measurement of distortion when there are two or more harmonics at the same time is beyond the scope of this book. Fig. 20.15 gives some results,

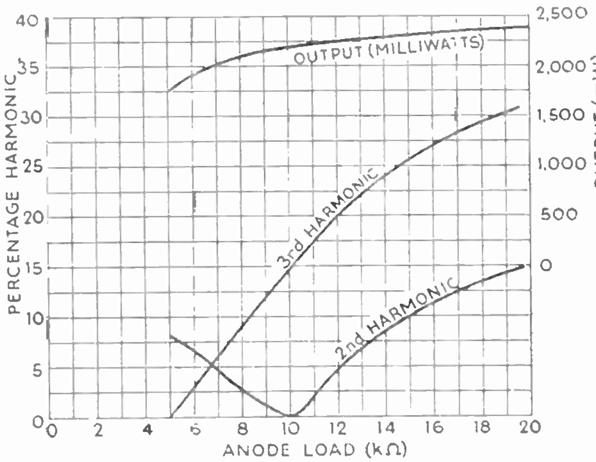



**Fig. 20.15—Output and harmonic distortion related to load resistance for a pentode**

showing how the output power and the percentage of second and third harmonic vary with load resistance, given a fixed input (grid) voltage.

That there is a medium load resistance making the second-harmonic distortion zero is very clearly brought out by this diagram. The third harmonic increases steadily with increasing resistance, as does also (though less steeply) the power delivered. Bearing in mind that third-harmonic distortion is more unpleasant than second, one would be inclined to recommend a load of about 6 kΩ in this case. It must be remembered, however, that this graph applies to only one particular working point, and that the optimum load resistance varies according to the working point selected.

It is after comparing the curves for a number of alternative working points that the final operating data for a valve are determined by its designer.

In selecting a working point one must of course take care not to exceed the power that the valve can get rid of safely as heat. The dotted line in Fig. 20.14 marks all points corresponding to an anode input power of 5·75 W. Such lines are usually included on valve curve sheets to show the limit beyond which the working point must not be placed.

An important general conclusion for the user to note is that with pentodes there is less latitude in the choice of load resistance than with triodes; so the fact that the impedance of a loudspeaker load varies considerably with frequency is more likely to result in distortion when the output valve is a pentode.

## Negative Feedback

The pentode's greater proneness to distortion offsets to some extent its advantages—its relatively high gain and high efficiency. If a reduction in the gain of the valve can be tolerated, however, it is possible to decrease the distortion very considerably without decreasing the power available. This is done by feeding back into the grid circuit some of the amplified voltage present at the anode, in the opposite polarity to that which would tend to cause oscillation. It is therefore called *negative feedback*. Fig. 20.16 shows the idea in principle.

If $A$ denotes, as before, the gain without negative feedback, and $B$ is the fraction fed back in series with the input, then for every signal volt applied to the input terminals, $-A$ volts appear at the output terminals, and $-AB$ volts are fed back. The input source is now required to supply, not 1 volt, but $1 + AB$ volts to produce the same output. The gain with negative feedback (which we might call $A'$) is thus $A/(1 + AB)$.

Taking as an example the load line AOC in Fig. 20.14 we have $A$ = anode-swing/grid-swing = 375/24 = 15·6. If one quarter of the output is fed back, $B = 1/4$, $AB = 3{\cdot}9$, and $AB + 1 = 4{\cdot}9$; so the preceding stage must deliver 4·9 times as much signal voltage as if there were no negative feedback, and $A' = 15{\cdot}6/4{\cdot}9 = 3{\cdot}2$. The amount of feedback applied is usually reckoned as the reduction in gain caused thereby, expressed in decibels—in this case 13·8 dB. Since the signal applied between grid and cathode is as before (the increase being solely on account of having to balance out the fed-

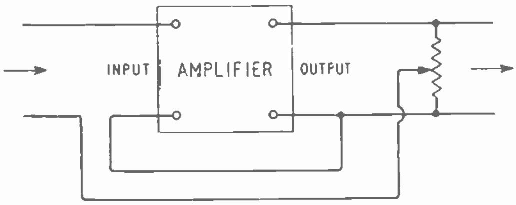



**Fig. 20.16—Showing in principle how negative voltage feedback is applied to an amplifier**

back voltage), the valve does not actually handle any increased signal and so requires only its normal bias.

Provided that the drop in gain from $A$ to $A'$ in the output stage is made good by a sufficiently increased input signal, there need be no drop in output power. But consider now what happens to harmonics or any other superfluous features generated within the output stage itself. The effect of feeding them back negatively is to reduce them in exactly the same proportion as the signal proper, but since they are not present in the input signal the increase in the input does nothing to restore them. So negative feedback reduces harmonic distortion in the same ratio as it reduces the gain of the stage.

It also reduces frequency distortion. Suppose, in the example just considered, that at the highest signal frequency the output stage gain ($A$) is reduced by 25% (from 15·6 to 11·7) by stray capacitance or some other cause. Then $A'$ at this frequency is $11{\cdot}7/(1 + 11{\cdot}7/4) = 2{\cdot}99$, which is only $6\frac{1}{2}$% less than the normal $A'$ (3·2).

In the example considered, the preceding stage has to deliver 4·9 times as much signal voltage as without feedback, namely $4{\cdot}9 \times 12$ = nearly 60 V peak, or nearly 120 V swing. If much negative feedback is applied across the output stage, care must be taken that removal of distortion therein does not cause distortion in the preceding stage by overworking it. This risk can be avoided, and the benefits extended, by feeding back over more than one stage. The voltage to be fed back is then usually small enough to be derivable from the secondary side of the output transformer, in which case transformer distortion also is lessened. Fig. 20.17 shows one way in which this can be arranged.

The more the feedback is extended, however, the greater is the risk that at some frequency (usually right outside the normal a.f. limits) the total phase shift round the feedback loop will amount to

180° more or less than the 180° intentionally provided, converting negative feedback into positive, sufficient perhaps to set up continuous oscillation. This problem of ensuring that the amplifier is stable receives much attention in the design of high-quality audio amplifiers.

There is another important effect of negative feedback. Depending on whether the voltage fed back is arranged so as to be proportional to the output voltage or the output current, the apparent resistance of the output valve is either reduced or increased. When, as is usual, this valve is a pentode, it is very desirable for its

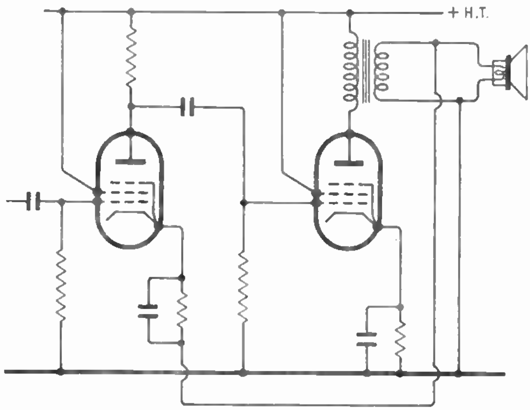



**Fig. 20.17—In this negative feedback circuit the whole of the voltage across the loudspeaker is fed back to the last stage but one**

resistance to be reduced. The reason is that loudspeakers inevitably resonate or " ring " at certain frequencies, accentuating sounds of those frequencies unnaturally, as well as prolonging them beyond their duration in the original. By shunting the loudspeaker with a low resistance, such resonance can be damped down just as in any tuned circuit. The usual resistance of a triode output valve—a fraction of the load resistance—is low enough to serve this purpose, but without feedback a pentode's $r_a$ is not.

From the point of view of the load, the resistance of an output valve with negative voltage feedback appears to be

$$r_a' = \frac{r_a}{1 + \mu B}$$

instead of $r_a$ as it would be without feedback. Note that this reduction is greater than the reduction of gain and distortion, because $\mu$ is greater than $A$; very much greater if the valve is a pentode. In our example (Fig. 20.14), $r_a$ is about 110 kΩ and $\mu$ 360 at the working point O, so the apparent resistance $r_a'$ when $B = 1/4$

is $1 \cdot 2$ kΩ; which is lower than the probable resistance of a triode that would be used for a load of, say, 6–8 kΩ.

So by applying negative feedback to a pentode the low distortion and effective damping of a triode can be combined with the high power efficiency and small bias voltage of the pentode.

## The Cathode Follower

In all valve circuits we have considered up to this point, the load has been between anode and + h.t. If, instead, it is connected between cathode and − h.t. we get an extreme case of negative feedback, because the whole of the output voltage is fed back. In its simplest form the circuit is as Fig. 20.18.

Suppose a signal of 1 volt to be applied between grid and cathode. This causes $A$ signal volts to appear across R. When the grid is sent positive, it causes more anode current to flow, increasing the voltage drop across R, and therefore making the cathode more positive too. The $A$ volts across R are therefore in series with the 1 volt between grid and cathode, making a total of $1 + A$ volts to be supplied by the driving source: 1 volt actually to drive the valve and $A$ volts to oppose the $A$ volts fed back across R.

This brief consideration reveals several features of the arrangement. The first is that since the input is $1 + A$ volts and the output is $A$ volts, the voltage amplification of the stage as a whole

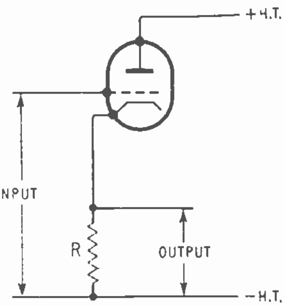



**Fig. 20.18—Simplest form of cathode-follower circuit**

must be less than 1, no matter what the characteristics of the valve and the value of $R$. The value of $B$ in the formula given on p. 312 is 1, so the " gain " is $A/(1 + A)$ as just found. If $A$ is made large by using a high-$\mu$ valve and high $R$, the output is very nearly as much as the input, but never quite as much. A stage that gives out less voltage than is put in may not appear worthy of further attention; but it is too soon to jump to such a conclusion.

The next point is that as the " live " side of the output—the cathode—goes more positive when the grid is made more positive (and vice versa), the output is not an inversion of the input as in the anode-coupled amplifier, but is in the same phase.

A third result follows from this; namely, that as the cathode signal voltage is normally very nearly the same as that on the grid, the signal voltage between them is small in relation to the input. It is this that earns for the device the name *cathode follower*; the cathode follows the changes of grid voltage. This is just the opposite of the demoniacal intervention explained with the aid of Fig. 17.3, in which the input capacitance of the valve is inconveniently multiplied. The demon is a friendly one this time, connecting the $A$-volt battery the opposite way round so as to reduce the charging current and *divide* the effective grid-to-cathode capacitance ($C_{gk}$) by $A + 1$. On the anode side, the Miller effect is eliminated because the anode is at a fixed potential, so the grid-to-anode capacitance ($C_{ga}$) is just normal. The contrast can be illustrated by an example. Suppose the real $C_{ga}$ and $C_{gk}$ are each 5 pF and the amplification $A$ is 40. Then with the load resistance in the anode side $C_{gk}$ becomes multiplied by 41, making 205 pF, while $C_{gk}$ is still 5 pF (because the cathode potential is fixed). The total input capacitance is thus 210 pF. With the resistance transferred to the cathode side, $C_{ga}$ remains 5 pF, and $C_{gk}$ is divided by 41, giving 0·12 pF. Total, 5·12 pF. This feature is of great value when amplifying signals having a wide range of frequencies (such as in television) from a high-impedance source, as any substantial input capacitance in the next stage would provide a low-impedance shunt at the high frequencies, cutting them down in relation to the low.

A fourth feature is the extreme reduction in apparent internal resistance of the valve. The formula on the last page but one stated that this resistance is equal to $r_a/(1 + \mu B)$ in a valve with negative feedback. In the cathode follower $B$ is 1, so we are left with $r_a/(1 + \mu)$, which is approximately equal to $1/g_m$. For instance, if $g_m$ is 5 mA/V, which is 0·005 A/V, the valve resistance is $1/0{\cdot}005 = 200\ \Omega$.

Another way of looking at this is to consider what happens when the load impedance is much reduced, say by feeding into some low impedance. If $r_a$ is very large, as in a simple pentode amplifier, the output voltage falls almost in proportion with the fall in load impedance. But in the cathode follower such fall in voltage reduces the voltage fed back, so that more of the input voltage is available for driving the valve, thus increasing the output. There is thus a strong compensating action tending to keep the output steady regardless of changes in load impedance; which is exactly what was meant by saying that the valve has a low effective output resistance.

This, too, is a valuable feature when amplifying a wide range of frequencies and feeding them to some distant point via a line, which inevitably has a considerable amount of shunt capacitance. If the source had much internal resistance, it would cause the higher frequencies to be discriminated against; but the resistance of a cathode follower is so low that the circuit it feeds into has little effect unless its impedance is lower still.

We have, then, a very useful device for feeding a low-impedance load from a high-impedance source. The fact that there is a slight loss in voltage is outweighed by the great step-up in current. Compared with the cathode follower as an impedance-changing device, a transformer sacrifices far more voltage in yielding a current step-up, is difficult to design for a very wide range of frequencies, and is much more liable to distort and to introduce stray coupling.

One can see that R in Fig. 20.18 incidentally provides grid bias (p. 368). If less bias is needed, it can be tapped off as shown in

Fig. 20.19—A more practical form of cathode-follower

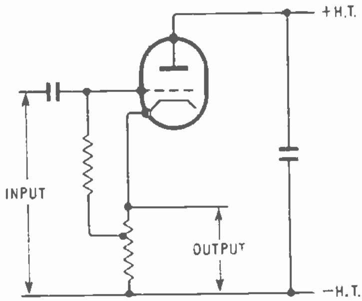



Fig. 20.19 through a resistance too high in proportion to the input source to divert an appreciable proportion of the signal or to reduce the feedback. The grid capacitor is to prevent a conductive input circuit from shorting out the effect of this bias connection. A very large capacitance is commonly connected across the anode supply as shown, to help keep the anode potential constant.

A point worth noting is that if a pentode is used in a cathode-follower circuit and no special device (p. 367) is used to tie the signal-frequency potential of the second grid to the cathode instead of the anode, it is thereby converted into a triode.

## Valves in Parallel and in Push-Pull

If more power is wanted than can be provided by a single output valve, two (or more) may be used. By simply adding a second valve of the same type in parallel with the first, connecting grid to grid and anode to anode, the swings of voltage at the anode are left unchanged, but the current swings are doubled. So, therefore, is the power; while the load resistance needed for two valves is half that needed for one. The $V_a/I_a$ curve for one valve can be made to apply to a number in parallel merely by multiplying the figures on the $I_a$ scale by the number of valves. In practice it is necessary to be careful to choose valves with closely similar characteristics.

Alternatively, a pair of output valves may be connected in *push-pull*, as shown in Fig. 20.20. Here the inter-stage coupling is a transformer $T_1$, in which the mid-point, instead of one end, of the

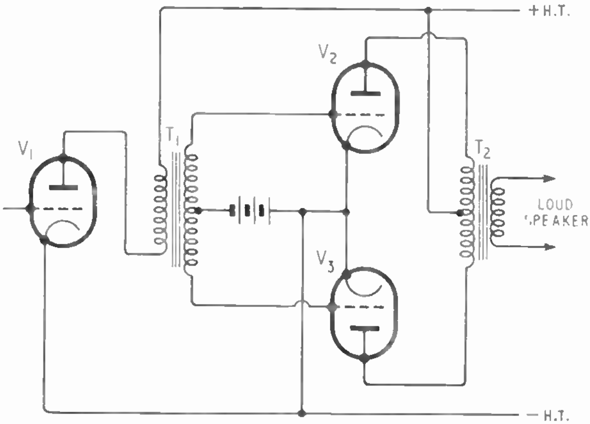



Fig. 20.20—Two output valves, $V_2$ and $V_3$, in push-pull

secondary is earthed. At an instant when, with the normal connection, the " live " end of the secondary would be at (say) + 20 V, the other (earthed) end being zero potential, the centre-point of the winding would be at + 10 V. With the push-pull arrangement this centre-point is brought to earth potential, the two ends therefore being respectively + 10 V and − 10 V. Thus each valve receives half the available voltage, the two halves always being in opposite phase.

The resulting opposite-phase anode signal currents, which would cancel one another if passed in the same direction through a transformer, are made to add by causing them to flow through separate halves of the centre-tapped primary of $T_2$. The voltage induced in the secondary, and hence the current flowing in the loudspeaker, is due to the combined currents of the two valves.

This mode of connection has several advantages over the more obvious parallel arrangement. These are:

(1) The steady anode currents, since they pass in opposite directions through their respective half-primaries of the output transformer, cancel one another as regards magnetization of the transformer core. The permeability of the core, on which the inductance of the transformer depends, falls off very seriously when the magnetization exceeds a certain amount, this effect being described as magnetic saturation. To avoid it a larger core must be used. So push-pull connection enables a smaller transformer to be used than would be needed for the same two valves in parallel.

(2) Signals fed through the common h.t. connection cancel; valves in push-pull therefore do not feed magnified signals into the h.t. line of a set, and so are less likely to cause undesired feedback. Conversely, disturbances on the h.t. line (hum, etc.) cancel in the two valves.

(3) *Second*-harmonic distortion produced by either valve is cancelled by equal and opposite distortion from the other. Two *triodes* in push-pull, therefore, give a greater undistorted output than if connected in parallel.

*Third*-harmonic distortion does not cancel in this way. Pentodes, whose output is limited by third harmonics, therefore do not share in this particular advantage. They can be made to do so, however, by connecting their second grids to suitable tappings on the primary of the output transformer, in what is rather absurdly called the " ultra-linear " system.

## Phase Splitters

In Fig. 20.20 the two stages are shown coupled by a transformer $T_1$, because the push-pull principle is more easily seen in this arrangement. But for the reasons we noted in considering intervalve transformers, resistance couplings are usually preferred. The problem is to provide two signal voltages in opposite phase. Several methods have been devised, of which one example will now be given.

In the ordinary resistance-coupled amplifier the resistor is on the anode side and the inverted output is taken from the anode. In the cathode follower the resistor is on the cathode side, and the output taken from the cathode is " right way up ". If we had both of these

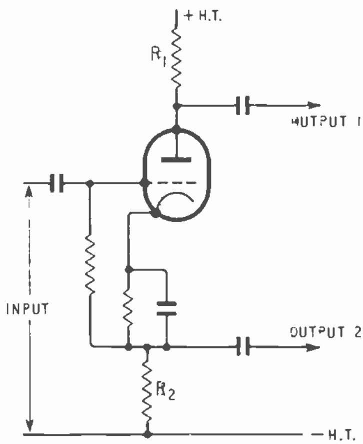



**Fig. 20.21—The " concertina " phase-splitter circuit as an alternative to the transformer $T_1$ in Fig. 20.20**

outputs simultaneously they would be in opposite phase and therefore suitable for driving a push-pull pair. There is no reason why we should not have them, simply by splitting the coupling resistance into two equal parts and putting one on the anode side and the other on the cathode, as in Fig. 20.21.

Because the same signal current passes through $R_1$ and $R_2$, making their resistances equal ensures that the two outputs are equal, or balanced—a matter of great importance in a push-pull system. To facilitate matching these resistances, the grid biasing arrangement (which will be dealt with in Chapter 23) is independent, and is made so by shorting out the bias resistor to signal currents by means of a high-capacitance shunt capacitor.

It will be seen that half the total output is fed back to the grid circuit, and therefore each of the two outputs is slightly less in voltage than the input (p. 315).

Looking at the circuit diagram, and imagining the two output points moving in and out in potential, one can appreciate the aptness of the name " concertina circuit ".

## The Loudspeaker

Since the commonest type of load in the output stage of a receiver is the loudspeaker, we shall end this chapter with a glance at how it works. Its object, of course, is to convert the electrical power delivered to it from the output stage into acoustical power having as nearly as possible the same waveforms.

A great many types have been devised, but the only one in general use to-day is the moving-coil. As the name suggests, it works on essentially the same principle as the moving-coil meter (p. 61). The mechanical force needed to stir up air waves is obtained by making use of the interaction between two magnetic fields (p. 60).

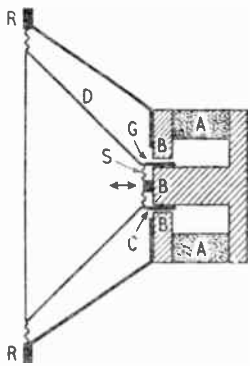



**Fig. 20.22—Cross-section of a moving-coil loudspeaker**

One of the fields is unvarying and very intense, and the other is developed by passing the signal currents through a small movable coil of wire.

Fig. 20.22 shows a cross-section of a typical speaker, in which the steady magnetization is provided by a ring-shaped permanent magnet A. Its magnetic circuit is completed by iron pole-pieces BB, except for a narrow circular gap G. The high permeability of the iron enables the magnet to set up a very intense magnetic flux across the gap. In this gap is suspended the moving coil C, through which the alternating signal currents from the secondary of the output transformer are passed. The mechanical force is proportional to both the flux density in the gap (which is large and constant) and the current in the moving coil. As the latter is alternating, the force is alternating, acting along the directions of the double-headed arrow in Fig. 20.22 and so tending to move the coil to and fro in the gap. The coil is attached to a conical diaphragm D, which is flexibly supported at its centre by a " spider " S and at its circumference by a ring R, so the vibrations of the coil are communicated to the diaphragm and hence to the air.

It is more difficult to avoid distortion due to the loudspeaker than any other in the whole system. As already mentioned (p. 314) the moving system as a whole has its own resonant frequency, often in the region of 80 c/s; its bad effects can be alleviated by a low output-stage resistance. To reproduce the lowest frequencies fully, it is necessary to move a large volume of air. If this is done by permitting a large amplitude of movement, either a very large magnet system is needed, or there is non-linearity due to the coil moving beyond the uniform part of the magnetic field in the gap. If it is obtained by using a very large cone, this makes it impossible for the cone to vibrate as a whole at the high audio frequencies and they are badly reproduced. The design is therefore a compromise. For the highest quality of reproduction it is usual to have two or more loudspeakers, each reproducing a part of the whole a.f. band.

It will be evident that at an instant when the diaphragm in Fig. 20.22 is moving to the left, there will be compressed air in front of it and rarefied air behind it. If the period of one cycle of movement of the diaphragm is long enough to give the resulting air wave time to travel round its edge from front to back, these pressures will equalize and not much sound will be sent out. To prevent this loss, evidently worst at the lowest frequencies, the loudspeaker is mounted so that it " speaks " through a hole in a *baffle*. This consists of a piece of wood, flat or in the form of a cabinet, designed to lengthen the air-path from front to back of the diaphragm and so to ensure that the bass is adequately radiated.

CHAPTER 21

# Cathode-Ray Tubes: Television and Radar

## Description of Cathode-Ray Tube

THE PRECEDING CHAPTERS have covered the general principles underlying all wireless systems (and incidentally quite a number of other things such as deaf aids and industrial r.f. heating), together with sufficient examples of their application to provide at least the outlines of a complete picture of the broadcasting and reception of sound programmes. There remains one device which is so important for such purposes as television and radar that this book would be incomplete without some account of it: the cathode-ray tube.

In many respects it resembles the valve; the differences arise from the different use made of the stream of electrons attracted by the anode. Both devices consist of a vacuum tube containing a cathode for emitting electrons, an anode for attracting them, and (leaving diodes out of account) a grid for controlling their flow. But whereas the valve is designed for making use of its anode current outside, the cathode-ray tube uses it inside by directing the stream of electrons against one end, where its effects can be seen. This stream was originally known as a cathode ray; hence the name of the tube.

Fig. 21.1 shows a section of a typical c.r. tube. The collection of electrodes at the narrow end is termed the *gun*, because its purpose is to produce the stream of high-speed electrons. There is the cathode with its heater, and close to it the grid. Although the names of these electrodes follow valve practice, their shapes are

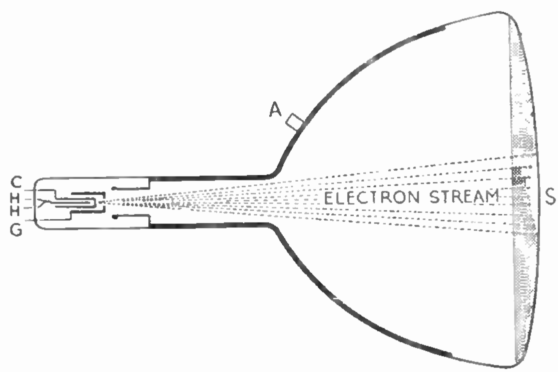



**Fig. 21.1—Section of typical cathode-ray tube. H = heater; C = cathode; G = grid; A = anode; S = fluorescent screen**

modified; the emitting part of the cathode is confined to a spot at its tip, and the " grid " is actually more like a small cup with a hole in the bottom. Provided that the grid is not so negative with respect to the cathode as to turn back all the emitted electrons, a narrow stream emerges from this hole, attracted by the positive anode voltage, which for television is usually between 7 and 17 kV. Many different shapes of anode have been favoured, almost the only feature common to all being a hole in the middle for allowing the electrons, greatly accelerated by such a high voltage, to pass on to the enlarged end of the tube.

Although this end is generally called the screen, the name has nothing to do with screening as we considered it in connection with valves (p. 233). The glass is coated with a thin layer of a substance called a *phosphor*, which glows when it is bombarded by electrons. This effect is known as *fluorescence*, and is essential in a c.r. tube; but there is also an effect, called *phosphorescence*, which is a continuance of the glow after bombardment. Phosphorescence can be controlled by choice of phosphor, and in some tubes dies away in a few thousandths of a second, and in others (for " long after-glow " tubes) lasts as much as a minute. The colour of glow also can be controlled by choice of phosphor.

When the electrons have done their bombarding they have to be withdrawn from the screen, or they would soon charge it so negatively that the following electrons would be repelled from it. In some types the anode is extended in the form of a coating of carbon inside the tube, as shown in Fig. 21.1. The bombardment of the screen causes secondary emission (p. 236), and enough secondary electrons are drawn away to the anode to balance those arriving from the gun. In modern television tubes the problem is solved in another way, by covering the phosphor with an extremely thin film of aluminium, which provides a conducting path to the anode. The film is too thin to stop the bombardment, but it is sufficiently mirror-like to reflect the glow and thereby increase its brightness as seen from outside the tube.

The only thing that can be done with the tube as described so far is to vary the brightness of the large patch of light on the screen by varying the grid bias. To make it useful, two more things are needed: a means of *focusing* the beam of electrons so that the patch of light on the screen can be concentrated into a small spot; and a means of *deflecting* the beam so that the spot can be made to trace out any desired path or pattern. Both of these results are obtained through the influence of electric or magnetic fields. The fact that electrons are so infinitesimally light that they respond practically instantaneously, so that the trace of the spot on the screen faithfully portrays field variations corresponding to frequencies up to millions per second, is the reason for the great value of the cathode-ray tube, not only in television but in almost every branch of scientific and technical research.

In television tubes both focusing and deflection are often done by magnetic fields, but as the electric methods are rather easier to understand we shall consider them first.

## ELECTRIC FOCUSING

When considering electric fields we visualized them by imagining lines of force mapping out the paths along which small charges such as electrons would begin to move under the influence of the field forces (p. 51). For example, if two parallel circular plates, A and B in Fig. 21.2*a*, were maintained at a difference of potential, the lines would be somewhat as marked—parallel and uniformly distributed except near the edges. As an electron moves from, say, A to B, its potential at first is that of A and at the end is that of B; on the journey it passes through every intermediate potential, and the voltages could be marked, like milestones, along the way. If this were done for every line of force we could join up all the points

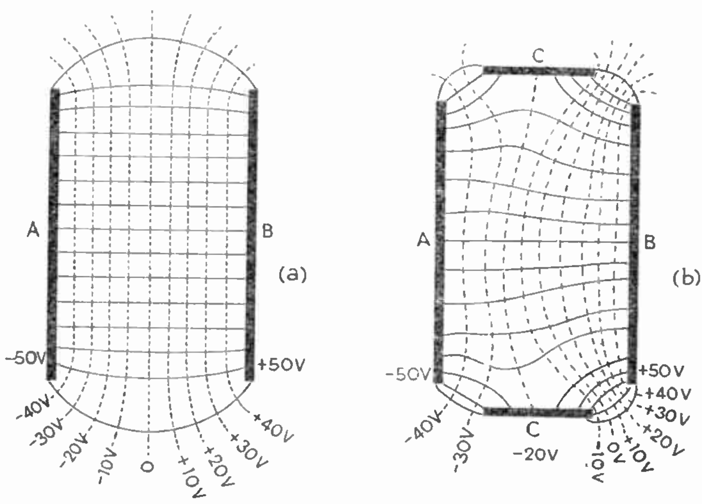



**Fig. 21.2—Diagrams to illustrate electrostatic focusing**

marked with the same voltage, and the result would be what is called an equipotential line. If potential is analogous to height above sea-level (p. 44) equipotential lines are analogous to contour lines.

In Fig. 21.2*a* the equipotential lines (dotted) are labelled with their voltages, on the assumption that the total p.d. is 100 V. An important relationship between the lines of force and the equipotential lines is that they must cross at right angles to one another.

Now introduce a third electrode, C in Fig. 21.2*b*, consisting of a cylindrical ring maintained at say $-20$ V. As C is a conductor, the whole of it must be at this voltage; so the equipotential lines between C and B must be very crowded. Along the axis between

A and B, where C's influence is more remote, they open out almost as they were in *a*. When the lines of force are drawn it is seen that in order to be at right angles to the equipotential lines they must bend inwards towards the centre of B.

This is the principle employed in electric focusing. At least one extra anode (corresponding to C) is adjusted to such a voltage as to make the convergence bring all the electrons to the same spot on the screen. The process is analogous to the focusing of rays of light by a lens, and in fact the subject in general is called electron optics. Just as a lens has to be made up of more than one piece of glass if it is to give a really fine focus, so a good electron lens generally contains at least three anodes at different voltages. The focus is usually adjusted by varying one of the voltages.

## Deflection

Electric deflection is a simpler business. If two small parallel plates are placed, one above and the other below the beam as it emerges from the gun ($Y_1Y_2$ in Fig. 21.3), and there is a difference of potential between the plates, the electric field between them will

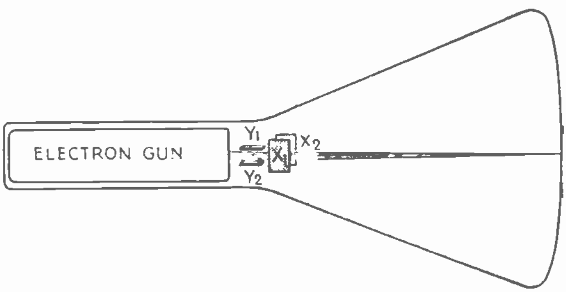



**Fig. 21.3—Arrangement of deflector plates in a cathode-ray tube**

tend to make the electrons move from the negative to the positive plate. This tendency, combined with the original axial motion, will deflect the beam upwards or downwards, depending on which plate is positive. A second pair of plates, $X_1X_2$, is placed on each side of the beam for deflecting it sideways. By applying suitable voltages between $Y_1$ and $Y_2$ and between $X_1$ and $X_2$ the spot of light can be deflected to any point on the screen.

Magnetic deflection depends on the force which acts on an electric current flowing through a magnetic field (p. 61). In this case the current consists of the electron beam, and the field is produced by coils close to the neck of the tube. The force acts at right angles to the directions of both the current and the field, so if the coils are placed at the sides of the beam as in Fig. 21.4, so as to set up a magnetic field in the same direction as the electric field between the X plates in Fig. 21.3, the beam is deflected, not sideways, but up or down. A second pair of coils, above and below, provides sideways deflection.

With electric deflection, the angle through which the beam is deflected is inversely proportional to the final anode voltage. So if, say, 100 V between the plates is sufficient to deflect the spot to the

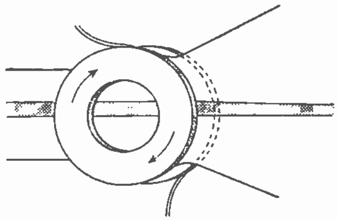


**Fig. 21.4—Arrangement of deflector coils in a cathode-ray tube**

edge of the screen when the voltage on the final anode is 1,000, raising the anode voltage to 2,000 makes it necessary to raise the deflecting voltage to 200. With magnetic deflection, the angle is inversely proportional to the square root of the anode voltage; so if, say, 20 mA in the deflecting coils was sufficient in the first case, it would only have to be raised to 28·3 mA in the second.

Raising the anode voltage thus necessitates more deflecting current or voltage, but it brightens the glow on the screen and generally improves the focus.

## Magnetic Focusing

For focusing magnetically, it is necessary for the direction of the field to be along the tube's axis. Electrons travelling exactly along the axis are therefore parallel to the field and experience no force. Those that stray from this narrow path find themselves cutting across the field and being deflected by it. That much is fairly obvious, but it is not at all easy to predict from first principles the final result of such deflection. Actually, by suitably adjusting the field strength, the electrons can be made to converge to a focus, but instead of doing so in the same plane as the divergence (as they would in electric focusing) the path is a spiral one. As with electric focusing, the beam is subjected to the focusing field before being deflected.

The magnetic field can be produced by a coil wound around the neck of the tube, in which case the focus is adjusted by varying the current through the coil. It is now more usual to use a ring-shaped permanent magnet, with a movable yoke for focus control.

## Operation of Cathode-Ray Tube

Although the final anode voltage is high by receiver valve standards, the current in the electron beam is small—of the order of microamps. As with valves, the grid is normally kept negative and takes negligible current. A variable bias voltage is used for adjusting brightness. The focus is adjusted as already described, according to the type of tube. Sometimes there are shift controls, for adjusting the initial position of the spot on the screen by applying bias voltages between the plates.

It is obvious that an electrically-deflected c.r. tube with these auxiliaries can be used as a voltmeter, by applying the voltage to be measured between a pair of deflection plates. It has the advantage of being a true voltmeter, since current is not drawn; but, of course, the anode voltage must be accurately maintained or the reading will be affected.

If the deflector voltage alternates at any frequency above a few cycles per second, the movement of the spot is too rapid to be followed by eye; what one sees is a straight line of light, the length of which is proportional to the peak-to-peak voltage. But the possibilities of the c.r. tube are more fully realized when voltages are simultaneously applied to both pairs of plates. For example, if a source of test input voltage to an amplifier is applied to one pair and the output is applied to the other, the appearance of the line or trace on the screen is very informative. If the amplifier is linear and free from phase shift, it is a diagonal straight line. Non-linearity shows up as curvature of the line, and phase shift as an opening out into an ellipse.

A particularly useful range of tests can be performed if to the horizontally-deflecting (or X) pair of plates is applied a voltage that increases at a steady rate. The c.r. tube then draws a time graph of any voltage applied to the other pair, and in this way the waveform can be seen. The usual procedure is to arrange that when the " time " voltage has moved the spot right across the screen it returns it very rapidly to the starting point and begins all over again. If the time of traverse is made equal to that of a small whole number of cycles of the waveform to be examined, the separate graphs coincide and appear to the eye as a steady " picture ".

A c.r. tube unit designed for test purposes, and especially for examining waveforms, is called an *oscilloscope*. The apparatus for producing the deflection proportional to time is a *time base generator*, and is so important both for oscilloscopes and television that it deserves further consideration.

## Time Bases

For depicting a waveform graph with a linear time scale, it is necessary for the deflecting voltage (or current, if magnetic deflection is used) to increase at a uniform rate; then, to avoid losing part of the waveform to be observed, its return should be as nearly as possible instantaneous. So the ideal waveform of the X deflecting voltage is as in Fig. 21.5; which explains why a time base generator is often called a sawtooth generator.

The subject of time bases is a very large one, and hundreds of circuits have been devised. The two main problems are to obtain a linear working stroke and a rapid flyback. The basis of many of the methods is to charge a capacitor at a controllable rate through a resistor, and then discharge it quickly by short-circuiting it. With an ordinary resistor the charging is not linear; it follows the exponential curve shown in Fig. 3.6*b*, the reason being that as the voltage across the capacitor rises the charging voltage falls off and so (with an ordinary resistor) does the current.

There are two main ways of overcoming this defect: one is to distort the non-linear sawtooth into a linear one, and the other is to use some sort of resistor that passes a constant current regardless of

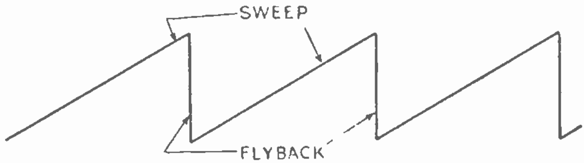



**Fig. 21.5—Sawtooth waveform used for c.r. tube time-base deflection**

the decline in voltage across it. If we look at the characteristic curves of a pentode (Fig. 20.14) we see that over a wide range of anode voltage the current changes very little. So if the capacitor is charged through a pentode, the voltage across it rises at a nearly constant rate, which can be adjusted by controlling the voltage of the control grid or the screen grid.

## Gas-filled Valves

So much for the charging; how about the rapid discharge? To switch the low resistance across the capacitor with the required frequency and regularity, an electronic device is needed. One of the simplest is a triode in which, instead of as high a vacuum as possible, a small amount of gas (such as argon, or mercury vapour) is put. Although the gas is actually quite rarified, this kind of valve is called gas-filled, or alternatively " soft "; and it is distinguished in diagrams by a black spot. When its anode voltage is gradually increased from zero the speed of the electrons drawn across to it increases correspondingly. At a certain speed the collisions between them and the gas molecules knock electrons out of the molecules. Molecules with electrons missing are positively-charged ions (p. 30), which neutralize the space charge of an equal number of electrons. Although the ions are naturally attracted to the cathode, they are far heavier than electrons and move comparatively slowly. So each one remains in the space long enough for hundreds of electrons to cross, and thus neutralizes the space charge not only of the electron torn from it but also of many of those emitted from the cathode. So the negative space charge—which in an ordinary valve is chiefly reponsible for its apparent resistance, $r_a$—is suddenly eliminated. In fact, the space between cathode and anode (except for a thin layer close to the cathode, across which practically the whole of the applied voltage is available to bring the emitted electrons up to enough speed to *ionize* the gas molecules) is converted into a good conductor, like a metal.

The $I_a/V_a$ characteristic of this type of valve is therefore quite different from that of a high-vacuum or " hard " valve; compare Fig. 21.6 with Fig. 20.14 or even Fig. 9.3. Up to the ionizing or " striking " point the anode current is quite small, for the anode voltage is insufficient to move the electrons fast enough to ionize the

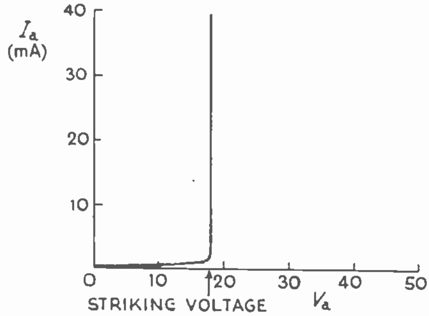



**Fig. 21.6—Characteristic curve of a "soft" valve. When it is passing current; the voltage across it is nearly constant, and $r_a$ is therefore nil**

gas. Beyond that point the slope resistance is nil, and if the current were not limited by the external circuit resistance it would very soon destroy the valve. This striking voltage is therefore the voltage drop across the valve, almost regardless of the current flowing.

In a soft diode, the striking voltage depends only on the kind of gas present (and to some extent its temperature); but in a soft triode, or *thyratron*, the use of negative bias raises the striking voltage. If, for example, each negative grid volt raises it by 30 V, 30 is called the *control ratio*. In some ways the control ratio corresponds to $\mu$ in a hard valve, but not altogether. For when the valve does strike, the negative grid becomes blanketed by the suddenly created swarm of positive ions and loses control of the anode current. The only way to stop it is to reduce the anode voltage below the point at which it can maintain ionization.

Fig. 21.7 shows the essentials of a simple time base circuit, using a pentode ($V_1$) as a nearly-linear charging resistor and a thyratron ($V_2$) as a discharger. When the voltage across C has reached the striking voltage of $V_2$, it strikes and short-circuits C, discharging it almost instantaneously. Its anode voltage having thus been cut off, $V_2$ de-strikes, and the whole cycle starts afresh. The striking voltage—and hence the amplitude of the sawtooth—can be controlled by the grid bias, as shown.

The capacitor across the amplitude control maintains the bias voltage fairly constant in spite of the varying current.

## Hard-Valve Time Bases

Thyratrons have been introduced because they represent an important branch of electronics which has many applications, and

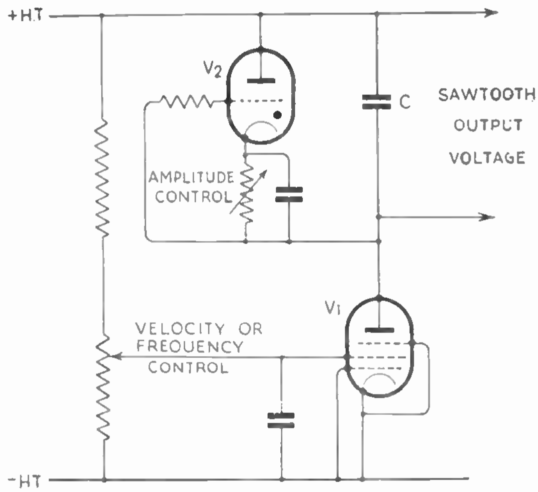



**Fig. 21.7—Outline circuit of sawtooth generator**

because they illustrate time-base action in the simplest way; but in practice they are now seldom used in time bases, mainly for two reasons. The ionization and de-ionization processes take time; only a little, it is true, but enough to limit the frequency to something like 50 kc/s. And because their life is less than that of hard valves they are no longer favoured for television receivers with their many hours of use.

Hard valves lack the natural sudden change-over to a highly-conducting state, so it is necessary to make the circuit produce an extremely rapid change from cut-off to zero—or even positive—grid bias. This is usually arranged by means of an oscillator with what would for other purposes be regarded as a grossly excessive amount of positive feedback.

Fig. 21.8 shows the simplest circuit of this kind; very commonly used, and (for a reason which will appear) known as the *blocking oscillator*. As we shall see, at one stage in its cycle of operation C becomes charged in the polarity that makes the grid negative—so

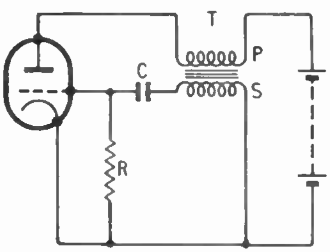



**Fig. 21.8—Blocking oscillator circuit**

highly charged that the valve is biased far beyond anode-current cut-off. It is discharging exponentially through R. Directly its voltage falls to the point at which anode current *just* begins to flow, that current flowing through the primary coil of the close-coupled transformer T induces a secondary voltage in the direction that opposes the remaining voltage of C, reducing the negative bias and increasing the anode current. The process is self-accelerating, and in less than a millionth of a second has driven the grid positive, so the grid side of C is being charged negatively. The rate of charge, through the low resistance from cathode to grid, is far faster than the discharge through R. But as C gains voltage the grid voltage falls, until it starts to cut off anode current. This reverses the induced voltage in T, ensuring a rapid cut-off of both grid current and anode current. There is now no voltage from T, but the highly charged C makes the grid highly negative, " blocking " the valve. And this is where we came in.

The periods of slow discharge of C through R, alternating with rapid recharge through the valve and sudden cut-off, give the voltage across R a sawtooth waveform. Because it is exponential it has to be linearized; one method being to pass it through an amplifier. The fastest rate of discharge is when the voltage is most negative, which is where the slope of the amplifier valve curve is least. As the voltage becomes less negative, the amplification increases, tending to compensate for the slower rate of change.

In another much-used circuit, known as the *multivibrator*, two resistance-coupled valves are each connected " head to tail ", so either can be regarded as supplying the other with amplified positive feedback, with results similar to those just described.

## Application to Television

In cinematography the appearance of motion is produced by showing successive still pictures at a sufficiently high rate to deceive the eye—say 25 per second. Each picture can be regarded as made up of a great number of small areas of light and shade; the greater the number of these *picture elements*, the better the definition. For entertainment, hardly less than 200,000 is acceptable. Now it is obviously impracticable to communicate as many as 200,000 radio signals, indicating degrees of light and shade, simultaneously; the only way is to send them in succession. If the whole lot has to be transmitted 25 times per second, the total picture element frequency is $25 \times 200{,}000 = 5{,}000{,}000$ per second. If light and dark picture elements alternate, each pair necessitates one cycle of signal; so the signal frequency is 2·5 Mc/s. In practice this calculation has to be slightly modified to take account of other things, but it does give some idea of the problem.

These picture-element or *video-frequency* (v.f.) signals cannot be directly radiated, because they are liable to have any frequency from 2·5 Mc/s down to almost zero; so they must be used to modulate a carrier wave just like a.f. signals. But for reasons discussed in Chapter 18 it is necessary for the carrier-wave frequency to be a good many times higher than the highest modulation frequency. That is why television cannot be satisfactorily broadcast on frequencies less than about 40 Mc/s, and the bandwidth is not a mere 20 kc/s or so as in sound broadcasting but extends to about 5,000 kc/s (or about half of this if most of one sideband is suppressed, as it usually is in practice).

To deal with picture elements successively, it is necessary to scan the scene in such a way as to include them all. The usual method is like that of the eye in reading a page of printed matter; it moves in lines from left to right, with rapid flyback between lines; and this horizontal movement is combined with a much slower vertical

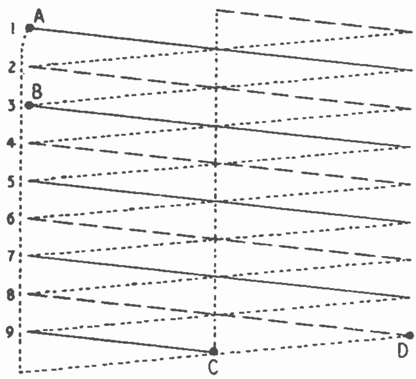



**Fig. 21.9—Nine-line interlaced "raster". The dotted lines are fly-backs, which normally are invisible**

movement down the page. It should be clear that scanning of this kind can be achieved in a cathode-ray tube by means of a high-frequency time base connected to the horizontally-deflecting plates or coils, and a low-frequency vertical time base. The spot then traces out a succession of lines, and if its brightness is meanwhile being controlled by the v.f. signals in proportion to the brightness of the corresponding point in the scene being broadcast, a picture will be produced on the screen of the c.r. tube.

To cover all the picture elements, the scene has to be scanned in some hundreds of lines—in the British system 405—and as a complete picture is covered 25 times per second the line frequency is $25 \times 405 = 10{,}125$ c/s. The picture frequency, 25 c/s, is low enough to cause noticeable flicker, so to avoid this the picture is scanned 50 times per second, odd and even lines being covered in alternate sweeps or "frames". The frequency of the vertical (frame) time base is therefore 50 c/s, and its flyback occurs after every $202\frac{1}{2}$ lines. This subdivision of the scanning is called *interlacing*.

A scanning diagram showing all 405 lines would be difficult to follow, so Fig. 21.9 shows a *raster* (as it is called) of only 9 lines. The principle is the same, however. Beginning at A, the line time base scans line No. 1, at the end of which the flyback (shown dotted) brings the spot of light back to the left-hand edge. But not to A, for the frame time base has meanwhile moved it downward two spaces to B. In the same way all the odd lines are scanned, until No. 9, half-way along which, at C, the frame time base flies back to the top of the picture. Although for clearness it is shown as doing so vertically, this is not practicable as it would mean infinitely high speed; actually the time occupied by 28 of the 405 lines is allowed. During this time—as during all the line fly-backs—the cathode ray is cut off, so nothing appears. The line time base can now fill in all the even lines, shown dashed. At the end of the last (D) the second frame fly-back returns the spot to the starting point A, though again not by the idealized dotted line.

At the sending end an image of the scene is focused as in a photographic camera; but in place of the film the television camera contains a plate inside a special form of cathode-ray tube. This plate is covered with a vast number of spots which become electrically charged in proportion to the amount of light reaching them from the scene. The electron beam is caused to scan the plate in the manner already described, and picks off the charges; their voltages are amplified, and used to modulate the carrier wave of the sender.

An essential addition to the system is some means for ensuring that all the receiving scanners work in exact synchronism with the scanner in the camera; otherwise the picture would be all muddled up. Synchronizing signals are transmitted during the flyback periods of the camera scanner, and the receivers separate these from the picture signals and apply them to the time base generators to control their exact moments of flyback.

## The Television Receiver

We have considered sound receivers in some detail as an illustration of the principles which are the subject of this book. Television receivers are based on the same principles, and their details are too numerous to go into fully here. But since a television receiver is now reckoned as almost basic equipment of a home, an outline of its contents is shown as Fig. 21.10.

The aerial is usually a dipole, with or without elaborations, as described in Chapter 15. The five " blocks " from aerial to loudspeaker comprise a superheterodyne sound receiver, and by now should be old friends. But the first two, at least, are used also for the vision signal and need special consideration. We have just seen that in order to give a reasonably clear picture the modulation frequencies have to extend to several Mc/s. If both sidebands were transmitted, the total bandwidth would be twice that number of Mc/s, but fortunately the visible effects of omitting most of one of them are negligible, so it is usually suppressed at the sending end. Just clear of this one very wide sideband, the sound part of the programme is transmitted in the usual manner. The tuning of the

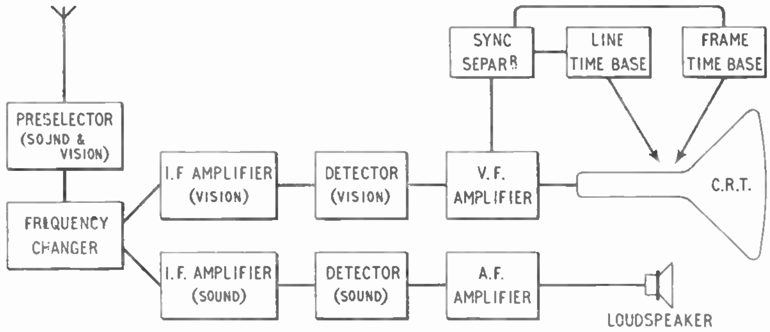



**Fig. 21.10—Block diagram of a television receiver**

preselector stage must be broad enough to accept both, which means that its selectivity must be low; in fact, so far from aiming for very low damping in the tuning circuits one usually has to increase the damping by shunting them with resistances of only a few thousand ohms. Fortunately the radio frequencies over 40 Mc/s, which are needed to carry picture modulation, are limited in range, so it is usually possible for all the senders receivable at any one place to be given widely different frequencies.

Only a single oscillator is used in the frequency changer, so the intermediate frequencies of the sound and picture (or vision) signals are spaced by the same amount as the incoming signals. For example, London television works on 41·5 Mc/s for sound and 45·0 Mc/s for vision, and a typical oscillator frequency for receiving it is 79·65 Mc/s. The frequency of the sound carrier wave at the output of the frequency changer is therefore 79·65 – 41·5 = 38·15 Mc/s, and of vision is 79·65 – 45·0 = 34·65 Mc/s. These are not always separated at once as shown in Fig. 21.10; sometimes part of the i.f. amplifier is common to both. The purely sound i.f. amplifier can be much more selective, so there is no difficulty in cutting out the vision signals; but the vision part has to cover nearly as wide a band as the earlier stages, so it is necessary to provide highly selective circuits to cut out the sound.

In principle, the vision detector and v.f. amplifier are similar to the sound detector and a.f. amplifier; but the component values for vision must be chosen to prevent the higher frequencies from being weakened by stray capacitances as discussed in Chapters 16 and 17. Besides using a comparatively low coupling resistance, a favourite device is a small inductance to neutralize the stray capacitance near the highest video frequency, where its effect would otherwise be most felt. The v.f. signal is finally imparted to either cathode or grid of the c.r. tube, according to whether the "brightest" signal is the most negative or the most positive.

Meanwhile two time-base generators are working at the correct frequencies to make the beam scan the c.r. screen as described with the aid of Fig. 21.9. During the periods occupied by the fly-backs of the camera at the sending end, the video signal (which is then not required for its main purpose) is formed into special pulses that can be separated out in the receiver by means of the " sync. separator " and applied to the time-base generators, nudging their elbows as it were, to let them know exactly when it is time for them to fly back. The line and frame synchronizing pulses are different, so that they can be distinguished by the sync. separator and delivered to the right time base.

## Colour Television

Many different systems have been devised for achieving television in colour. The most favoured of them build up the picture from three primary colours. For instance, if in place of each frame (or line, or picture element) transmitted in plain television, there were three, representing these colours, and they were superimposed at the receiver, they would reproduce the picture as in colour printing. This has been done quite successfully, but has the serious disadvantage of multiplying the already vast bandwidth occupied by each programme by three. For broadcasting, therefore, much ingenuity has resulted in substantial economy in bandwidth, at the cost of complication in both sender and receivers.

It is easy enough to produce c.r. tube phosphors having three suitable colours; the problem is to combine them in one tube in such a manner that they are finely enough distributed to give clear definition, yet at the same time respond only to a beam controlled by signals of the right " colour ". Although this problem too has been solved sufficiently to give very acceptable results, as with the rest of the receiver the cost is relatively high.

## Application to Radar

If one makes a sound and hears an echo $t$ seconds later, one knows that the object reflecting the wave back is $550t$ feet distant, for sound travels through air at about 1,100 feet per second and in this case it does a double journey. Radar is based on the fact that radio waves also are reflected by such objects as aeroplanes, ships, buildings and coastlines. If the time taken for the echo to return to the point of origin is measured, the distance of the reflecting object is known to be $93{,}141t$ miles away, the speed of radio waves through space being 186,282 miles per second. Since distances which it is useful to measure may be less than a mile, it is obvious that one has to have means for measuring very small fractions of a second. In radar, time is measured in microseconds ($\mu$sec).

This is where the cathode-ray tube again comes in useful. With a time base of even such a moderate frequency as 10 kc/s one has a scale that can be read to less than 1 $\mu$sec, and a much faster time base can easily be used if necessary. The time between sending out a wave and receiving its echo can be measured by making both events produce a deflection at right angles to the time base line. This can best be understood by considering a typical sequence of operations.

We begin with the spot at A in Fig. 21.11, just starting off on a stroke of the time base. Simultaneously a powerful sender is caused to radiate a burst or pulse of waves. This is picked up by a receiver on the same site, connected to the Y plates so that the spot is deflected, say to B. The pulse must be very short, so that echoes from the nearest objects to be detected do not arrive while it is still going on. In practice it may have to be a fraction of a microsecond, and as it has to consist of a reasonable number of r.f. cycles, their frequency obviously has to be many Mc/s.

The time base voltage continues to move the spot across the screen, and if the receiver picks up one or more echoes they are made visible by deflections such as C. When the spot reaches the

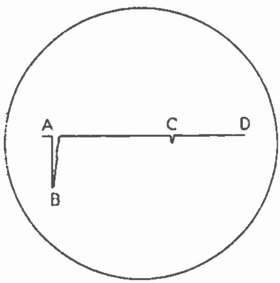



Fig. 21.11—Appearance of trace on cathode-ray tube in one type of radar receiver

end of its track, at D, it flies back to the start; then, either at once or after a short interval, begins another stroke. The sender, being synchronized, registers another deflection at A, and the echo is again received at C. The repetition frequency is normally high enough for the trace of the spot to be seen steadily as a whole.

The time for an echo to return from a distance of one mile being 1/93,141 second, it is equal to the time occupied by one cycle of a 93,141 c/s oscillation. By switching the Y plates over to a 93,141 c/s oscillator, so that the cycles are seen as a wavy line, the time base can be marked off in distances corresponding to miles.

This is only a mere outline of one of many ways in which the c.r. tube is used in radar. In one of the most-used, the time base begins at the centre and moves radially outwards like a spoke of a wheel. Successive sweeps occur at slightly different angles, so as to cover the whole screen during each revolution. Normally the glow is suppressed by negative bias on the c.r. tube grid; echoes reduce this bias, bringing up the glow on the tube at the appropriate radial distance and at an angle which indicates the angle at which the outgoing waves were being radiated.

To concentrate transmission and reception exclusively into the direction indicated on the screen, the aerial—usually the same one is used for both—must be very highly directional. This in itself necessitates a high frequency (p. 198), which is usually in the category of "microwaves", running into thousands of Mc/s. At such frequencies the techniques have to be very different from those we have considered; to mention one thing, capacitors are so inductive and inductors are so capacitive that it is not practicable to make separate provision for them in circuits—or indeed make circuits at all in the sense understood hitherto. Instead of wiring, there is "plumbing". This is another specialized subject for which appropriate books must be consulted.

CHAPTER 22

# More About Transistors

## Valves and Transistors Compared

SINCE BEING INTRODUCED to valves and transistors, in Chapters 9 and 10 respectively, we have chosen valves rather than transistors as examples, because their behaviour is freer from complications that might distract attention when studying the various kinds of circuits for the first time. Now, however, we can take up the transistor story again and see how it compares with what we know about vacuum valves.

One complication, as we have already noted, is that germanium transistors at least are considerably affected by temperature, so that special provision has to be made against this in circuit design. Another is that their performance tends to fall off at a much lower frequency. But perhaps the most fundamental difference is that under working conditions the transistor conducts at the input as well as the output—in fact the input resistance is much lower than the output—and changes in output current affect the input current as well as vice versa. As a result, the " equivalent generator " is less simple than for a valve (p. 142).

Lest all this seem to put the transistor in an unfavourable light, it must be said that the complications concern only the designer; the user benefits by the small size, long life, and low consumption of the transistor.

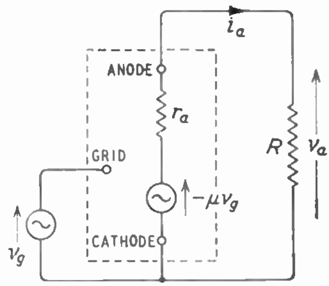
It will help us to see what is involved in the transistor equivalent generator if we first review the relatively simple valve equivalent. We became acquainted with this on p. 143, and Fig. 22.1 is a repetition of the diagram for a triode. Although the valve has three electrodes, as commonly used (with negative grid bias) the only current path is between cathode and anode. So there is only one resistance, $r_a$. The signal current made to flow in this circuit is the same as if $r_a$ were in series with a generator giving $\mu$ times the signal voltage $v_g$ applied between cathode and grid. Because the cathode is common to both input and output circuits, and is usually more or less earthed, voltages are reckoned from it; so to express the fact that a positive $v_g$ causes a negative output voltage it is necessary to use a minus sign: $-\mu v_g$. This synthetic valve enables one to calculate the amplification given in any circuit more easily than from a sheet of characteristic curves, provided that the valve is being used within reasonably linear limits.

## The Equivalent Current Generator

Fig. 22.1 is not the only possible equivalent circuit even for a valve. Instead of our generator giving a specified voltage ($-\mu v_g$)

regardless of the impedance of the rest of the circuit, we can imagine a generator giving a specified current regardless of the impedance of the rest of the circuit. To fulfil the condition, the voltage generator must have no impedance of its own; if it had, any flow of current would cause some of the generator e.m.f. to be used up in driving the current through itself. The current generator, by contrast, must be imagined as having infinite imped-

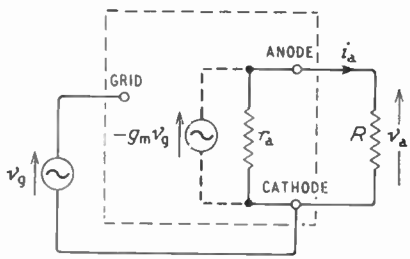
**Fig. 22.1 — Voltage-generator equivalent of a valve (repeated from Fig. 11.3)**




ance; otherwise any impedance of the circuit it was feeding would reduce the current. Because no real generator conforms to this description, it is more difficult to imagine than the voltage generator; but it is nevertheless a very helpful idea in some circumstances. Various symbols for a current generator are in use; the one shown in Fig. 22.2 is the same as for the voltage generator except that the leads are dotted to indicate that the shunt circuit formed by the generator has an infinite impedance.

How do we know that the generator current is $-g_m v_g$? Let us suppose we do not know. Our guiding principle is that this

**Fig. 22.2 — Current-generator equivalent of a valve**




equivalent circuit must give the same answers as the voltage-generator circuit. We have already calculated the output voltage ($v_a$) with the help of the voltage-generator circuit (p. 144); it is

$$v_a = \frac{-\mu v_g R}{r_a + R}$$

If we remember that $\mu = g_m r_a$ (p. 122) we can write the same thing as

$$v_a = \frac{-g_m v_g r_a R}{r_a + R}$$

In Fig. 22.2, $v_a$ is equal to the current from the current generator, multiplied by the resistance of $r_a$ and $R$ in parallel. As this parallel resistance is $r_a R/(r_a + R)$ (p. 38), the above equation shows that the current must be $-g_m v_g$.

The $r_a$ of a pentode is usually so much greater than $R$, the load resistance, that its shunting effect is often neglected, and the effective resistance through which the current $-g_m v_g$ passes regarded as just $R$. In other words, $v_a \simeq -g_m v_g R$, so the voltage gain $(= v_a/v_g)$

$$A \simeq -g_m R$$

which is the same as we found on p. 239. Because it shows up the effect of very high $r_a$ so clearly, the current-generator equivalent is especially suitable for pentodes.

## The Three Configurations

Another thing we have found about valves is that there are three basic ways in which they can be connected as amplifiers. These ways (called configurations) are named after the electrode that is common to both input and output:

(1) Common cathode (p. 14), Fig. 22.3*a*;
(2) Common grid (p. 276), Fig. 22.3*b*;
(3) Common anode (or cathode follower) (p. 315), Fig. 22.3*c*.

Valves that have only three electrodes separately involved in output and input circuits (including pentodes as well as triodes, because their second and third grids are connected to fixed potentials) can have no more than these three configurations, unless we make three more by interchanging output and input, but those three do not amplify so are not usually counted.

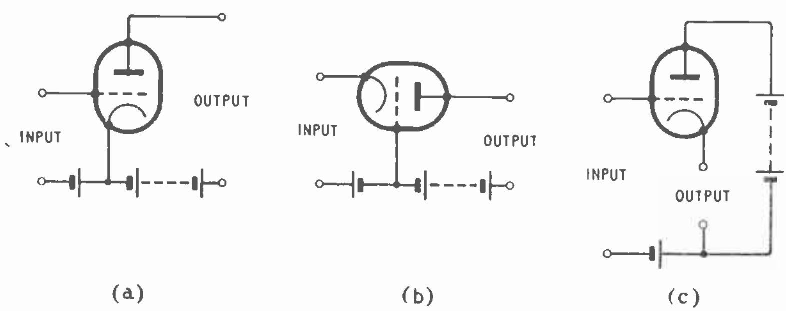



**Fig. 22.3—The three basic amplifier configurations of a valve**

Each of the three configurations has its own voltage-generator equivalent; the one we have used so far, with its $r_a$ and $\mu$, applies to the common-cathode, which is by far the most used. Note that what we have been calling $v_g$ is more precisely $v_{kg}$—the signal voltage between cathode and grid—and $r_a$ means the resistance between cathode and anode. For the common-anode configuration we could use the same form of equivalent as Fig. 22.1, but because the input voltage is $v_{ag}$ instead of $v_{kg}$ its amplification factor and internal resistance are different. They could be given special symbols, but since this configuration is the exception and the common-cathode is the rule they are almost always given in terms of the common-cathode $\mu$ and $r_a$; they are in fact $\mu/(\mu+1)$ and $r_a/(\mu+1)$ respectively, which show that the common-anode amplification factor must be less than 1 and the internal (or output) resistance is relatively low (p. 315–6).

The equivalent circuit for the common-grid is more complicated, because the output current passes through the input circuit and consequently the resistance the valve appears to have depends on the resistance of the input circuit—the source of the signal voltage. In this respect a valve " in common-grid " is more like a transistor than it is in the other two configurations.

There are also, of course, current-generator equivalents for all three configurations.

The transistor has three corresponding configurations: common emitter, common base, and common collector; shown in Fig. 22.4. They correspond to common cathode, grid and anode respectively, not only in the way the electrodes are connected, but also very

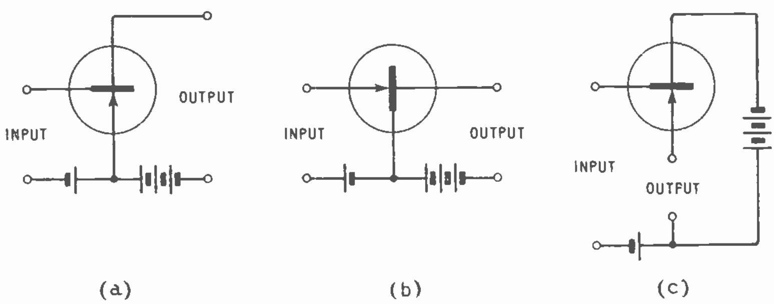



**Fig. 22.4—The three basic amplifier configurations of a transistor, corresponding to Fig. 22.3**

largely in the way they behave, as we shall see on p. 351. If the valve example were followed, all three configurations would be calculated in terms of common-emitter parameters. But historically (p. 138) the common-base came first, and " $\alpha$ " without qualification means the current amplification in common-base. Similarly " $r_c$ " means the collector resistance in that configuration. These and other common-base parameters are the ones most commonly given in transistor data, and they are used in the kind of equivalent circuits (for all configurations) most like the valve's. So we can hardly avoid taking notice of them, much though we might prefer to follow the valve example and base everything on common-emitter.

## Transistor Equivalent Circuits: T Networks

Fig. 22.5 shows what are called the T-network transistor equivalent circuits for common-base configuration; the one with an internal current generator (*a*) is put first because of a transistor

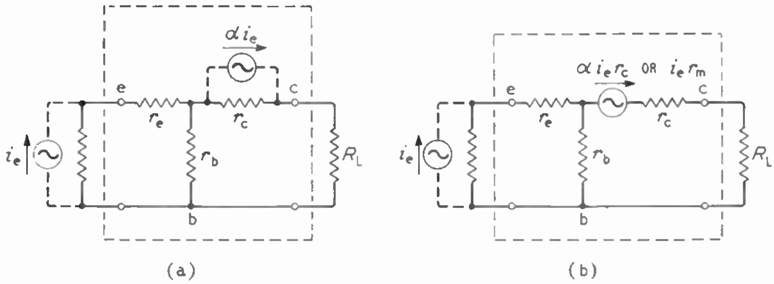



**Fig. 22.5—Equivalent T circuits of a common-base-connected transistor: (*a*) current generator; (*b*) voltage generator**

being a current amplifier. In the same sort of way as we found the valve equivalent current generator from its voltage generator (p. 339) we can find the transistor equivalent voltage generator (*b*). Note the special symbol $r_m$ (corresponding upside-down to a valve's $g_m$) for $\alpha r_c$, for it is often given in transistor data.

So far as the collector " arm " only is concerned, the forms are the same as the valve equivalent circuits. The two additional parameters, $r_e$ and $r_b$, make matters very different, however. Both being comparatively low resistances, they present the ingoing signal with a low resistance. But the most awkward thing for calculations is that $r_b$ is common to both input and output circuits. So part of $i_e$ goes through $r_c$ into the output circuit, represented by the load resistance $R_L$, quite apart from the efforts of the internal generator. Because $r_c + R_L$ is very much larger than $r_b$, this effect is usually small enough to neglect. But part of the output current flows through the input circuit, causing a certain amount of feedback. And both input and output resistances of the transistor depend to some extent on the resistances of load and signal source.

Fig. 22.6 shows the corresponding diagrams for the common-emitter configuration. The collector resistance and the current or voltage of the internal generator are stated in terms of the common-base current amplification factor and collector resistance, $\alpha$ and $r_c$, for the sole reason that these are the quantities specified in transistor data, and not because common-base is more important than common-emitter. It would have been better, in fact, if the symbols $\alpha$ and $r_c$ had been allocated to what are shown in Fig. 22.6 as $\alpha/(1-\alpha)$ and $r_c(1-\alpha)$, instead of to the corresponding quantities in the less-used common-base circuit. One system of symbols denotes $\alpha/(1-\alpha)$ and $r_c\,(1-\alpha)$ as $\alpha'$ and $r_c'$ (with $\alpha''$ and $r_c''$ for the common-collector counterparts). Alternatively the three different

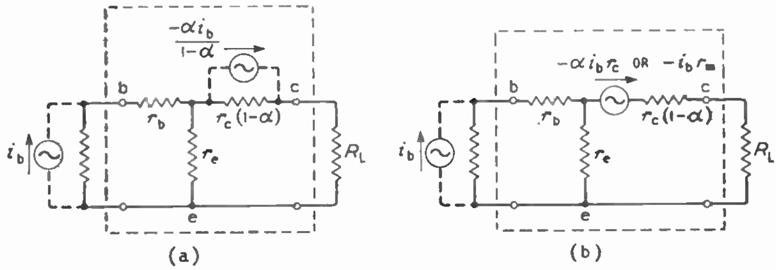
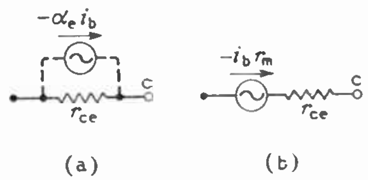



**Fig. 22.6—Common-emitter T circuits corresponding to the common-base circuits of Fig. 22.5, using the same symbols**

$\alpha$'s can be distinguished by a subscript letter indicating the common electrode: $\alpha_e$, $\alpha_b$ and $\alpha_c$. On the same principle we would have $r_{ce}$, $r_{cb}$ and $r_{cc}$. The collector-arm parts of Fig. 22.6 would then read as in Fig. 22.7.

One has to be wary about the signs of the currents and voltages, especially as different writers express the same facts in different ways. Here, they are drawn to correspond most closely with valve practice. This has the advantage that the negative sign of the internal generator in the common-emitter circuit agrees with the reversal of signal polarity during amplification. But in some books the arrow points the other way, so the minus sign is plus. In others, one is left to guess which direction an invisible arrow is pointing. Remember, too, that in *p-n-p* transistors the feed currents going in

**Fig. 22.7 —' If common-emitter symbols are used, these collector arms should be substituted for those in Fig. 22.6**




at base and collector terminals are negative, so a positive $i_b$ is one that makes the base current less negative; changes it, say, from —0·3 mA to —0·2 mA. This reminds us of signal voltages at the grid of a valve. It is easier to become confused at the transistor output: a positive $i_b$ causes an amplified negative current in the direction of the arrow, and because this is in the opposite direction to the negative current from the collector battery it reduces that negative current. The negative voltage drop in $R$ therefore falls, so the potential of terminal c becomes nearer that of the collector battery; i.e., becomes more negative. Having once worked it out in detail this way to be quite sure, we can henceforth ignore the collector battery and feed current and arrive at the same result direct from the equivalent circuit by saying that a positive $i_b$ causes a negative current into $R$ which makes the output terminal (c) negative.

## Other Transistor Parameters

Although these T equivalent circuits look the most natural to people who are used to the valve equivalents, and tie up most closely with the explanation of what goes on inside a transistor, and comprise data commonly specified, they have two serious disadvantages. One is that the parameters $r_b$, $r_e$, etc., cannot be found from the usual characteristic curves or directly measured, and the other is that the formulae for the things one wants to know —input and output resistances, and amplification—are complicated when given in terms of these data. Unfortunately the number of different equivalent circuits and sets of parameters devised to overcome these drawbacks is bewilderingly large. They all have their pros and cons, which cannot be gone into fully in a book like this.

We can, however, take a glance at the general idea, which is to regard the transistor as a sealed box with pairs of input and output terminals (one terminal of each pair being common), and to measure what happens at one pair when some current or voltage is applied to the same or the other pair. It is not really essential to draw an equivalent circuit to represent what is inside the box, but we may note in passing that instead of the minimum four parameters being three resistances and one current or voltage from an imaginary generator, as in our T networks, there can be two resistances and two generators.

The fact that what is connected to one end of the box affects the other must be kept in mind. For example, if one of the parameters in our chosen set of four is the resistance measured between the output terminals (i.e., the output resistance) it is necessary to specify the impedance (to signal currents) connected to the input terminals. The most obvious choices are zero and infinity.

One commonly used set consists of what are called the hybrid parameters:

$h_{11}$ = input resistance with output short-circuited to a.c.

$h_{21}$ = current amplification with output short-circuited to a.c.

$h_{22}$ = output conductance with input open-circuited to a.c.

$h_{12}$ = voltage feedback ratio, with input open-circuited to a.c.

These hybrid parameters are easily measurable, and the formulae embodying them are simpler than those with T parameters, but more complicated than for a vacuum valve. Take the output resistance, for example. which for a valve is simply $r_a$, but for a transistor is

$$r_{out} = h_{22} - \frac{1}{\dfrac{h_{12}h_{21}}{h_{11}+R_S}}$$

Without the details of how they are derived, here for reference are the others:

$$\text{Input resistance} = r_{in} = h_{11} - \frac{h_{12}h_{21}}{h_{22}+1/R_L}$$

$$\text{Current amplification} = A_I = \frac{h_{21}}{h_{22}R_L+1}$$

$$\text{Voltage amplification} = A_V = \frac{h_{21}}{h_{11}(h_{22}+1/R_L)-h_{12}h_{21}} = \frac{A_I R_L}{r_{in}}$$

$R_L$ is the load resistance and $R_s$ the resistance of the signal source, as shown in Fig. 22.8. The power amplification, $A_P$, is of course $A_I A_V$. The above formulae serve for all three configurations, and the definitions of the $h$ parameters are the same, but their values differ. Without distinguishing mark, as above, they would be taken as denoting the common-base values. The common-emitter values are distinguished as $h'_{12}$ or $h_{12e}$, etc.

Typical values for a transistor with resistance coupling (and therefore rather low $I_c$) are: $h_{11e} = 2{,}000$; $h_{21e} = 45$; $h_{22e} = 25 \times 10^{-6}$; $h_{12e} = 8 \times 10^{-4}$. For the sake of example let us suppose $R_L$ and $R_S$ are both 5,000 Ω. Substituting these values in the formulae we get: $r_{in} = 1{,}840\ \Omega$; $r_{out} = 50\ \text{k}\Omega$; $A_I = 40$; $A_V = 109$; $A_P = 4{,}360$, or 36·4 dB. It must be realized that these gains are reckoned from terminal b in Fig. 22.8 to terminal c, and

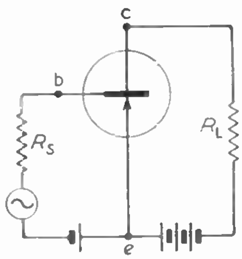



**Fig. 22.8—Theoretical common-emitter transistor circuit showing source and load resistances, $R_S$ and $R_L$**

take no account of any loss in $R_S$. If this stage were followed by another identical one, $R_L$ would be shunted by the $r_{in}$ of the next stage, which would bring it down to 1,350. raising $A_I$ to 43·5 but reducing $A_V$ to 30 and $A_P$ to 1,300 (31 dB).

Because the input current/voltage relationship of a transistor is markedly non-linear, its input resistance can be regarded as a constant quantity only within narrow limits; in other words, these parameter calculations all assume that the amplitude of the signal is small enough for variations in value of the parameters during each signal cycle to be neglected.

Most books on transistors tabulate formulae for converting from $h$ parameters (and perhaps others not mentioned here) to T parameters and vice versa.

## Effects of Frequency

Although we have been thinking chiefly of a.c. signals, we have been easing our first approach to the subject by considering all impedances as resistances. In practical circuits, however, reactances have to be taken into account. The methods of doing so are much the same as for valve circuits, so will not be repeated. The transistor itself, viewed from its input and output terminals, is in general not purely resistive. Not only do the factors $R_S$ and $R_L$—which ought to be $Z_S$ and $Z_L$—appear in the formulae for $r_{in}$ and $r_{out}$, but the transistor itself has self-capacitance, and (owing to the relatively slow speeds of electrons and holes in solid material) transit-time

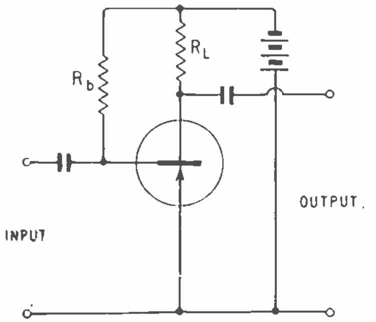



**Fig. 22.9—This method of applying base current bias to a transistor, via $R_b$, is unsatisfactory in practice**

effects, which modify its parameters at the higher frequencies. Some very complicated equivalent circuits have been devised to imitate these high-frequency effects.

The only notice we shall take of them here is to add $f_\alpha$ to our stock of transistor data. It is the frequency at which $\alpha$ is 3 dB (or nearly 30%) below the low-frequency value normally quoted. This is the way we reckoned the " cut-off " frequency of amplifier circuits (p. 302). $f_\alpha$ shows whether a transistor is suitable for r.f. use or only for a.f. One with a $f_\alpha$ of 100 kc/s might be excellent for a.f. amplification but would clearly be useless for i.f., say 470 kc/s.

## Biasing Arrangements

Although Fig. 22.8 illustrates the principle of resistance-coupled amplification by transistor, it is a quite unpractical circuit. No one would want to use a second battery for applying base bias. And it must be remembered that for a transistor it is bias *current* that has to be applied in the required amount. Another feature about transistor biasing in contrast to valve biasing is that its polarity is the same as the "h.t."; namely, negative for a *p-n-p* transistor. So it is no use trying to apply the method shown in Fig. 20.17, by means of a resistor in series with the emitter. It would supply voltage bias, of the wrong polarity.

A very simple method is to supply the bias current from the collector battery through a suitable resistance, $R_b$ in Fig. 22.9. Supposing for example that the required bias is 50 μA and the battery is 6 V, $R_b$ should be $6/50 = 0{\cdot}12$ MΩ. In this calculation the fact that the voltage across $R_h$ differs from 6 V by the amount of the base voltage is ignored, because the discrepancy is very small. It is, in fact, far less than differences between individual transistors of the same type—or the same transistor at different temperatures. Because of these wide differences between transistors nominally of the same type, this simple bias scheme also is unpractical. $R_b$ would have to be adjusted to suit each individual transistor, and (worse still) varied to compensate for temperature changes. Otherwise the virtually fixed bias current supplied by $R_b$ might pass too much collector current at one extreme or cut it off at the other.

A popular solution is to apply too much bias to the base and reduce it to the required amount by emitter bias of the kind we dismissed above because it was of the wrong polarity—$R_e$ in Fig. 22.10. Just like the cathode biasing resistor for a valve, this emitter resistor automatically adjusts the bias to compensate for

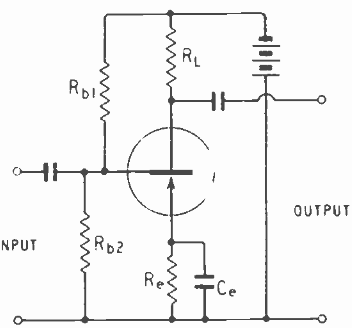



**Fig. 22.10—This method of applying base current bias is self-compensating so is commonly used**

too much or too little emitter current (which is nearly the same as the collector current). Suppose it is too much; then it develops a relatively large voltage across $R_e$, which opposes the base bias voltage to a large extent, so that the net bias voltage (and consequently current) is small, tending to reduce the collector current. It would not succeed in doing this if the base bias were supplied through $R_b$ only, as in Fig. 22.9, because (as we have seen) that supplies a practically constant bias current. It is necessary to keep the base *voltage* as constant as possible, so that changes in emitter voltage are effective in controlling base current. This is done in Fig. 22.10

by the potential divider $R_{b1}$ and $R_{b2}$. If $R_{b1}$ is much lower in resistance than $R_b$, the voltage of the base (relative to the bottom end of $R_{b2}$) will be very little affected by changes in the comparatively small base current. But there is a limit to what can be done in this direction, for we do not want the resistance to be so low as to be a serious drain on the battery, or to draw off appreciable signal current from the base. Typical values are 27 kΩ for $R_{b1}$, 3·3 kΩ for $R_{b2}$, and 220 Ω for $R_e$. $C_e$ is used to short-circuit $R_e$ to signal currents, which would otherwise apply negative feedback (p. 315).

## A Practical Amplifier Circuit

That is not to say that negative feedback is not wanted. In a.f. amplifiers it usually is, to counteract the non-linearity distortion that arises at large signal amplitudes. But one wants to be able to adjust it independently of bias requirements, and to apply it over more than one stage as in valve circuits.

Fig. 22.11 shows a typical complete a.f. amplifier circuit. Except for the biasing arrangements, just described, and the values of some of the components, it could quite well be a valve circuit. We note the resistance-coupled two stages of amplification leading to a transformer-coupled push-pull output stage driving a loudspeaker. The signal voltage across the loudspeaker feeds back current through the 68 kΩ resistor to the input of the second stage. The coupling and by-passing capacitances are large by valve standards, because the circuit resistances—especially the transistor inputs—are small.

## Transformer Coupling

Although the reasons given on p. 303 for usually preferring resistance to transformer coupling in valve a.f. amplifiers apply also to transistors, there is rather more to be said in favour of transformers for transistors than for valves. The much lower input resistance of the transistor necessitates far fewer secondary turns, so the transformer can be cheaper. The resulting voltage step-down is a current step-up, whereas with resistance coupling the signal current available for the input of one transistor cannot be more (and in practice is less) than the current from the output of the previous one. In other words, a considerably greater gain is obtainable by matching the very unequal output and input resistances by means of a transformer.

For perfect matching, the transformer ratio $n$, and therefore the current step-up, would be equal to $\sqrt{r_{out}/r_{in}}$ (p. 98). This is awkward to calculate from the formulae on p. 345, but (bearing in mind that the variable nature of transistor parameters makes precise calculation futile) a reasonable approximation can be obtained by neglecting the negative terms in the formulae. The

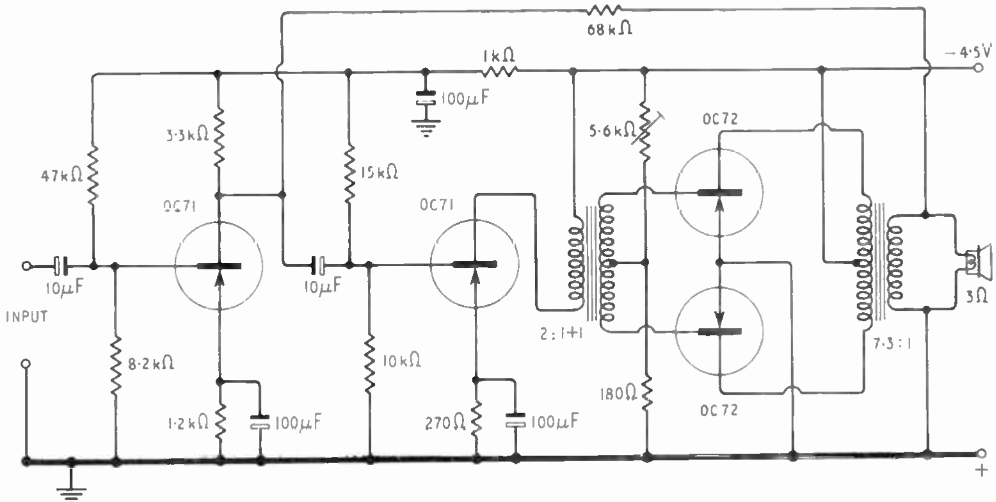



Fig. 22.11—Circuit diagram of a three-stage transistor a.f. amplifier, similar in configuration to a valve amplifier but requiring only a small fraction of the power supply. Junction transistors with common emitter are used for all stages. (*Courtesy of Messrs. Mullard Ltd.*)

ratio is then simply $1/\sqrt{h_{11}h_{22}}$. The effective current gain of the stage from b1 to b2 in Fig. 22.12 (call it $A_{I(S)}$) is equal to $A_I$ for the transistor alone, multiplied by $n$:

$$A_{I(S)} = A_I n \simeq \frac{A_I}{\sqrt{h_{11}h_{22}}}$$

Since the transistor output resistance (approximately $1/h_{22}$) is matched to the load resistance $R_L$, $1/h_{22} = R_L$, so $h_{22}R_L = 1$ and our formula for $A_I$ reduces to $h_{21}/2$, and

$$A_{I(S)} \simeq \frac{h_{21}}{2\sqrt{h_{11}h_{22}}}$$

On the assumption that the input resistances of the two transistors are equal, the power gain is equal to the square of the current gain:

$$A_{P(S)} \simeq \frac{h_{21}^2}{4\, h_{11}h_{22}}$$

Filling in the values we took for the example on p. 345 we get $A_{P(S)} \simeq 45^2/(4 \times 2{,}000 \times 25 \times 10^{-6}) = 10{,}000$ (40 dB), compared with 1,300. The actual values with transformer coupling, owing to the much lower resistance to collector current, would probably be rather different, but the general effect similar. The above approximation underestimates the theoretical gain, and thereby fortunately helps to allow for transformer losses.

## "Class B" for A.F. Amplification

One of the main advantages of transistors over vacuum valves is economy in power consumption. This is due partly to the absence of heater and partly to the shapes of the characteristic curves. In Fig. 10.14, for example, it is easy to place a load line so that the signal can swing from a small fraction of one volt $V_c$ to a small fraction of one milliamp $I_c$, thus avoiding the causes of low efficiency in valve amplifiers. So the tendency is to use valves where power can be drawn from the mains—so cheaply that economy is unimportant—and transistors where the enormously more costly battery power has to be used; for example, in deaf aids and portable receivers.

In valve amplifiers of the types described so far, power is being consumed all the time at almost the same rate whether the maximum signal output is being delivered or none at all. We came up against this on p. 161, in connection with r.f. power amplifiers, and found one solution in the so-called Class B system, which uses enough grid bias to reduce the anode current almost to zero. At first it might seem that this solution would be useless for a.f. amplifiers, because one half of every signal cycle would be cut off, causing quite intolerable distortion. And that is true enough for a single valve or transistor. But if this " over "-biasing is used in a push-pull amplifier (p. 317) the half-cycles that are cut off by one valve are amplified by the other, so the full cycles are re-assembled in the output transformer and all is well.

Since the vital difference between Class A amplification (in which each valve amplifies the whole cycle) and Class B is the amount of grid bias, it does not show up in circuit diagrams. In fact, the same push-pull stage could be used for either; for example, in Fig. 22.11 by different settings of the pre-set resistor that controls the amount of base bias for the output stage. But in practice the change from Class A to Class B not only affects the bias arrangements but also the design of the input and output transformers and

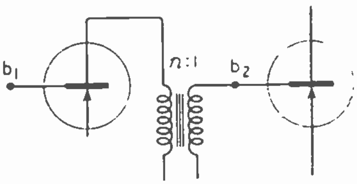



**Fig. 22.12—Transformer coupling for matching the output of one transistor to the input of the next**

other things, so would not usually be left optional. Owing to the difficulties of exactly matching the two halves of the stage and of ensuring correct transfer from one to another at the working points on the bottom bends, there is almost inevitably greater distortion than in Class A. Class B is sometimes used in valve a.f. amplifiers, especially the high-power ones for modulators and public-address installations, but the advantages are more conspicuous in transistor amplifiers, partly because of the importance of power economy in battery-driven equipment and partly because the shape of transistor characteristics enables the efficiency to be obtained by Class B to be so much higher. The signal power output can be something like five times the maximum power loss allowable in the transistors.

## Common Base and Collector Configurations

Although the common emitter configuration is the most used because (for one thing) it gives the most power gain, sometimes one of the other two is preferable. The common-collector circuit, or emitter follower, corresponds to the valve cathode follower and has similar uses; for example, for matching a high impedance to a low. But its input impedance depends very largely on the load impedance at the output, and the output impedance on what is connected to the input. Like the cathode follower, its voltage gain is less than 1, but it gives considerable current gain—roughly equal to $\alpha_e$. And there is no phase reversal between input and output.

This last feature applies also to common-base circuits, but as regards the transistor's input and output impedances it is at the opposite extreme: the input impedance is lower even than in common-emitter, and the output impedance is higher, so it is suitable for matching a low impedance to a high. These input and output impedances depend less on the impedances connected to output and input respectively than in common-emitter, and much less than in common-collector. Because of this smallness of interaction between input and output circuits, and also because the amplification is maintained to a higher frequency, common-base circuits are often used for r.f. amplification.

## Neutralization

Nevertheless, there is some interaction even in common-base circuits. The load or coupling of a transistor r.f. stage is of course normally a tuned circuit. Over the band of frequencies centred on the resonant frequency, the impedance of a tuned circuit changes rapidly from capacitive through pure resistance (at resonance) to inductive. So the phase of the output current, and consequently the current fed back to the input circuit, varies widely at different frequencies close to resonance. Just as with Miller effect in valves, the result at some frequency may be to make the net resistance at the input negative, causing risk of oscillation. Even short of actual oscillation, adjustment of the tuning is difficult if every tuned circuit affects the others.

One remedy, as with valves, is neutralization. There are several possible methods; Fig. 22.13 shows one of them. The neutralizing components, $C_n$ and $R_n$, are adjusted until the tuning of the collector

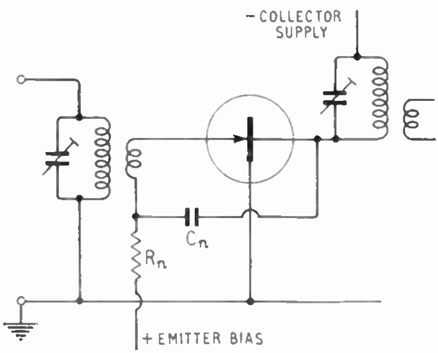



**Fig. 22.13—Method of neutralizing a transistor stage of r.f. amplification**

circuit has little or no effect on the base circuit. Note that tuned r.f. step-down transformers are used to match the high output resistance of a common-base-connected transistor to the very low input impedance of the next.

## Transistor Oscillators

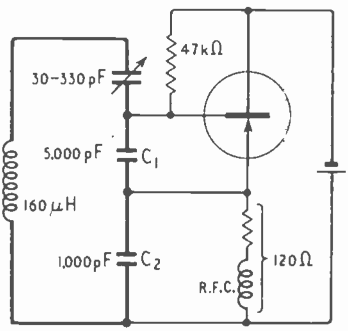
The various valve oscillator circuits of Chapter 12 have their transistor counterparts, the circuit diagrams for which may actually be identical except for the transistor itself and (with *p-n-p* types) the polarity of the supplies. What we have learnt about transistors as compared with valves will however have shown the desirability of some differences in circuit design.

The most important, perhaps, is the low input impedance. (The output impedance, being comparable with that of a valve, calls for no comment.) One would therefore use a considerable impedance step-down in the positive feedback path. For instance, if the circuit of Fig. 12.5 were being adapted for a transistor, the " reaction coil " would be given very few turns compared with the tuned coil. Similarly in Figs. 12.6 and 12.7 the input (base to emitter) would embrace only a small proportion of the whole tuned circuit.

The same policy is often applied to a valve. This is not, of course, because of the low input impedance of the valve—on the contrary, except at very high frequencies that is large—but deliberately to mismatch the feedback coupling in order to prevent it from being tighter than is necessary just to maintain oscillation at the desired amplitude. Owing to the amplification of a valve, a good match would feed back far too much and cause violent distortion and perhaps " squegging " (p. 369). Moreover, since the interelectrode capacitances of a valve vary with voltages, etc., the less they are coupled to the tuned circuit the more constant the frequency is likely to be.

These considerations apply with even greater force to transistors, because their impedance is much more variable than a valve's,

Fig. 22.14—A transistor r.f. oscillator, using the Gouriet circuit to stabilize frequency




especially with changes in temperature. And whereas the input impedance of a valve is so high at all negative grid voltages that driving it beyond cut-off makes little difference, driving a transistor beyond cut-off reduces its input resistance from perhaps less than 1 kΩ to more than 100 kΩ. If this impedance were closely coupled to the tuned circuit, one half of each cycle would be much more heavily damped than the other and the waveform would be distorted.

So between the tuned circuit (which is what ought to determine the frequency exclusively) and the transistor—especially its input—

the coupling should be as loose as will no more than just maintain oscillation. Fig. 22.14 shows a circuit, devised by G. G. Gouriet and J. K. Clapp independently for valve oscillators, which fulfils this requirement admirably for transistors also. In effect, it is a Colpitts circuit in which the valve or transistor is tapped across small portions of the impedance of the tuned circuit. The capacitances $C_1$ and $C_2$ are therefore made relatively large. Component values shown are suitable for the medium-frequency band, centred on about 1 Mc/s, and give good frequency stability. The $f_\alpha$ of the OC45 transistor is about 6 Mc/s.

CHAPTER 23

# Power Supplies

## THE POWER REQUIRED

HAVING COME ACROSS frequent references to power supplies needed for valves and transistors, we finally consider them in detail.

The power that has to be fed to sender and receiver valves can be divided into three classes: cathode-heating, anode (including screen-grid) supplies, and grid bias.

So far as cathode heating is concerned, it is purely a matter of convenience that it is done electrically. In principle there is no reason why it could not be done by a bunsen burner or by focusing the sun on it with a lens. But there are overwhelming practical advantages in electric heating. Cathode heating power has to be supplied all the time the valves are required for work, whether or not they are actually doing any.

Power supplied to the anode is more directly useful, in that a part of it is converted into the r.f. or a.f. output of the valve. This is generally not true of power taken by auxiliary electrodes, so valves are designed to reduce that to a minimum.

Not all valves need any grid bias; and, of those that do, not all have to be provided with it from a special supply. And only in large sending valves is any appreciable power consumed, the reason being that with this exception care is taken to keep grid current as nearly as possible down to nil (p. 124). This restriction, desirable for most purposes, would unduly limit r.f. power output.

Transistors have the great advantages of needing no heater power at all and of making more economical use of what has to be supplied. Usually the sole source of power is a small dry battery, or (where available) a car battery.

## BATTERIES

A battery consists of a number of *cells*, and although a single cell is commonly called a battery it is no more correct to do so than to call a single gun a battery.

A cell consists of two different sorts of conducting plates or electrodes separated by an "exciting" fluid or *electrolyte*. The e.m.f. depends solely on the materials used, not on their size, and cannot greatly exceed 2 V. The only way of obtaining higher voltage is to connect a number of cells in series to form a battery.

Other things being equal, a small cell has a higher internal resistance than a large one, so the amount of current that can be drawn from it without reducing its terminal voltage seriously is small. To obtain a large current a number of cells of equal voltage must be connected in parallel, or (preferably) larger cells used.

Many types of cell have been devised, but only two are commonly used for radio. They represent the two main classes of batteries—primary and secondary.

Primary cells cease to provide current when the chemical constituents are exhausted; whereas secondary cells, generally called *accumulators*, can be "recharged" by passing a current back through them by means of a source of greater e.m.f.

The most commonly used primary batteries are the so-called dry batteries—a badly chosen term, because any really dry battery would not work. The name refers to cells in which the electrolyte is in the form of a paste instead of a free liquid. The essential ingredient is ammonium chloride or "sal-ammoniac", and the plates are zinc (−) and carbon (+). The e.m.f. is about 1·4 V per cell. When current is drawn, each coulomb causes a certain quantity of the zinc and electrolyte to be chemically changed. Towards the end of its life, the e.m.f. falls and internal resistance

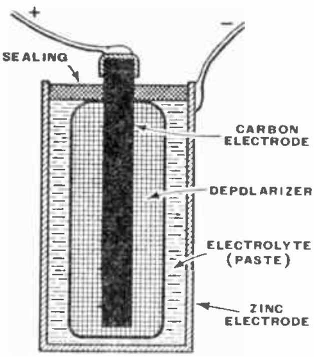



**Fig. 23.1—Cross-section of a dry cell, showing the essential parts**

rises. Ideally, the cell would last indefinitely when not in use, but in practice a certain amount of "local action" goes on inside, causing it to deteriorate; in time, therefore, a cell becomes exhausted even if never used.

If only the ingredients named above were included in a dry cell, its voltage would drop rapidly when supplying current, owing to the formation and dispersal of a layer of hydrogen bubbles on the surface of the carbon. To reduce this layer a *depolarizer* is used, consisting largely of manganese dioxide around the carbon electrode. Fig. 23.1 gives an idea of the usual construction of dry cells.

The most commonly used secondary cell is the lead-acid type, in which the plates consist of lead frames or grids filled with different compounds of lead, and the electrolyte is dilute sulphuric acid. Accumulators have an e.m.f. of 2 V per cell and possess the great advantages of very low internal resistance and of being rechargeable.

IRVINE — POWERED

IRVINE — POWERED

IRVINE — POWERED

IRVINE — POWERED

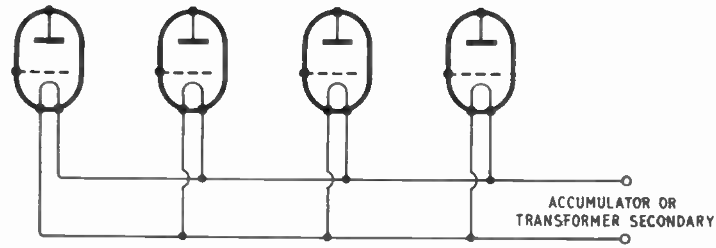
Against this there are certain drawbacks. The acid is very corrosive, and liable to cause much damage if it leaks out or " creeps " up to the terminals. If the cell is allowed to stand for many weeks, even if unused, it becomes discharged, and if left in that condition it " sulphates ", that is to say, the plates become coated with lead sulphate which cannot readily be removed and which permanently reduces the number of ampere-hours the cell can yield on one charge. If the terminals are short-circuited, the resistance of the accumulator is so low that a very heavy current—possibly hundreds of amps—flows and is liable to cause permanent damage. Some accumulators, such as those used in motor cars, are designed to supply heavy currents for short periods. Others, called " block " or " mass " batteries, are intended to give small currents for long periods.

## Cathode Heating

As stated in Chapter 9, the cathode of a valve may be either directly or indirectly heated. The power for either type can be obtained from a battery, but in practice directly-heated valves are then used because they require only about one-tenth of the power needed by the indirectly-heated types, and therefore a given battery will run them for much longer, or alternatively a much smaller battery can be used. These points are important, because battery power is far more expensive than that drawn from the mains, and high-power batteries are very heavy. Seeing then that the chief use of batteries is to enable the set to be carried about, it is desirable for the power required to be as small as possible. At one time, small accumulators were much used for filament heating, but they have been superseded by dry cells.

The vast majority of public electricity supplies are a.c. of 200–250 V at a frequency of 50 c/s, which can be stepped down to any convenient voltage for valve heating by means of a transformer. If an attempt were made to run directly-heated valves in this way it would be unsuccessful, because the slight variation in temperature between peak and zero of the a.c. cycle, and also the variations of potential between grid and filament from the same cause, would produce variations in anode current, which would be amplified by the following stages and cause a loud low-pitched hum. An occasional exception is a carefully arranged output stage, which of course is not followed by any amplification.

The temperature of the relatively bulky indirectly-heated cathode changes too slowly to be affected by 50 c/s alternations, and as it carries none of the a.c. its potential is the same all over. Incidentally this improves the valve characteristics, and so does its larger area compared with a filament. Further, its greater rigidity allows the grid to be mounted closer to it, resulting in still greater mutual conductance. One may therefore expect a mains-driven set to be more sensitive than a battery set with the same number of valves.



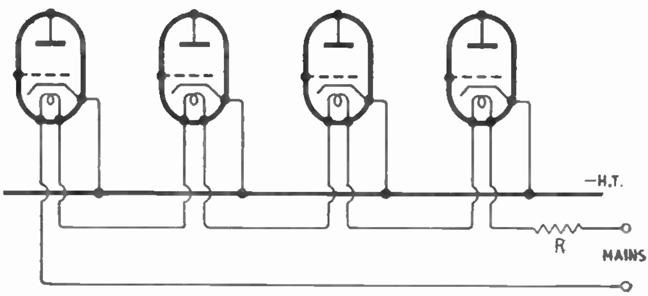




**Fig. 23.2—Diagram *a* shows the usual method of connecting the filaments of battery valves or heaters of a.c. valves, namely, in parallel. To run heaters direct from either d.c. or a.c. mains, they are connected in series, as at *b*. Note that in spite of the different potentials of the heaters all cathodes can be joined to — b.t.**

The most usual rating for a.c. receiver valve heaters is 6·3 V. This odd voltage is to enable these valves to be used alternatively in cars, whose batteries average 6·3 V or 12·6 V under working conditions. They are usually connected in parallel like battery valves, as in Fig. 23.2*a*, but if they are all designed to take the same current, so that they can be run in series (Fig. 23.2*b*), the rated voltage is anything from 6·3 to 40 according to type. A common practice is to connect all the valve heaters of a set in series, together with sufficient resistance to enable them to be run direct from 200–250 V mains, either a.c. or d.c.

## Anode Current from A.C. Mains

The cost of power from the mains is, at most, not more than about one-fortieth that of power from h.t. batteries. So one can afford in a mains-driven set to use plenty of anode current; which means, in turn, a more ample output and less need to run the output stage on the verge of distortion.

The power is there; the problem lies in making use of it. Obviously the a.c. cannot be used as it is, because during one half of each cycle the anodes would be negative and the valves would be inoperative. Before we can use it, we have to convert the current from alternating to direct.

We have already seen, in Chapter 16, how to convert a.c. into d.c. The first requirement is a rectifier. If the a.c. is of sine waveform it can be represented by Fig. 23.3*a*, and the result of connecting a rectifier in series is to abolish half of every cycle, as shown at *b*.

The average of this is only about one-third of the peak alternating current (p. 209), and in any case would be useless for running radio sets because it is not continuous. By using a reservoir capacitor (p. 211) this defect can be removed and at the same time the average is brought almost up to the full peak voltage. The conditions are

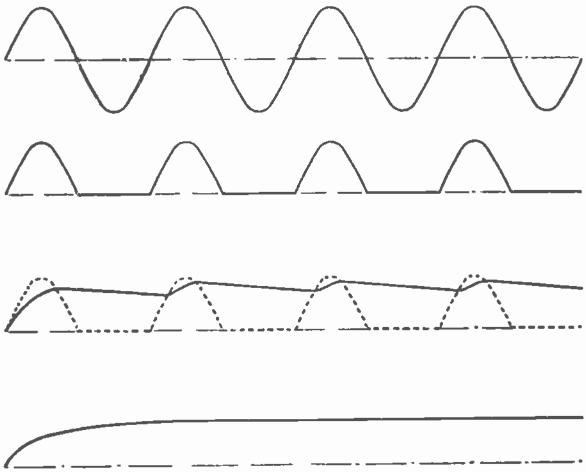



**Fig. 23.3—(*a*) Original a.c.; (*b*) after rectification; (*c*) the effect of a reservoir capacitor; (*d*) after passing through a smoothing filter**

not quite the same as in a detector, because there the load resistance is usually quite high, whereas a typical sound receiver taking about 62 mA at 250 V is a load resistance of 4 kΩ. Unless the reservoir capacitor is enormous, therefore, it has time for the discharge to cause an appreciable loss of voltage between one half-cycle and the next, causing the current supply to vary, as in Fig. 23.3*c*.

This is described as d.c. with a ripple, or unsmoothed d.c.; and would cause a hum if used to feed a receiver. The third requirement is a smoothing circuit, which is a filter (p. 217) designed to impede the a.c. ripple while permitting the d.c. to pass freely. If the filter is effective, the result is as in Fig. 23.3*d*, practically indistinguishable from current supplied by a battery.

## Types of Rectifier

The vacuum diode (p. 117) is an obvious rectifier, and is perhaps the most commonly used of any. Power rectifiers differ from diode detectors in having larger cathodes to supply the much greater emission needed, and larger anodes to dissipate the greater heat.

A typical rectifier for a sound receiver has a 6·3-V 0·6-A heater and is capable of supplying 90 mA at 350 V.

High-power rectifiers are made for supplying up to several amps at tens of thousands of volts. To stand the high reverse voltage when the diode is not passing current, the anode must be spaced some distance from the cathode. Then during the other half-cycle, when a very heavy current is flowing, the space-charge (p. 118) is bound to be large, necessitating a large voltage between anode and cathode during the conductive half-cycle to neutralize it. This voltage is deducted from the output and thereby lost, merely causing an undesirable amount of heat at the anode.

You may however remember the interesting way in which the space charge can be largely neutralized by positive ions, if a small amount of a suitable gas or vapour is admitted inside the bulb of the diode (p. 238). The voltage drop across these gas-filled or "soft" diodes is only 12–16 V, regardless of the amount of current flowing, which may be very large. So the power lost in the rectifier is comparatively small, and one not much bigger than a receiving valve can provide an output of a kilowatt. On the other hand they need rather more care in use than vacuum types: the cathode must be allowed to heat up before the h.t. is applied, and for mercury-vapour types the temperature at the start must be within certain limits.

There are also various types of semi-conductor rectifiers, which need no cathode heating and can be designed for almost any voltage. The germanium and silicon signal diodes considered in Chapter 10 have their power-handling counterparts. Because of its lower intrinsic conduction, silicon can operate at higher temperatures—and consequently give more output for a given size—than germanium, and passes less reverse current. The older established metal rectifiers, consisting of disks coated with copper oxide or selenium and clamped together in series to handle the required voltage, are still commonly used both as alternatives to vacuum or gas-filled diodes or for lower voltages.

## Rectifier Circuits

There are various ways in which rectifiers can be connected in power unit circuits, and one of the simplest is indicated in Fig. 23.4*a*. A transformer T is shown for stepping the mains voltage up or down as required, and a single rectifier—the symbol denotes any type—is connected in series with it and a reservoir $C_1$. A simple filter circuit consisting of a choke L and capacitor $C_2$ completes the apparatus.

This is called a *half-wave* rectifier circuit, because only one-half of each a.c. cycle is utilized, the other being suppressed (Fig. 23.3*b*). If, as is desirable, the capacitance of $C_1$ is large enough for the voltage not to drop much between half-cycles, the proportion of each cycle during which the alternating voltage exceeds it is very small; and, as the recharging of $C_1$ has to take place during these brief moments, the rectifier is obliged to pass a current many times greater than that drawn off steadily at the output (Fig. 23.4*b*). If a large output current is needed, then the reservoir capacitance must be very large, the rectifier must be able to pass very heavy peak currents, and there are difficulties in designing the transformer to

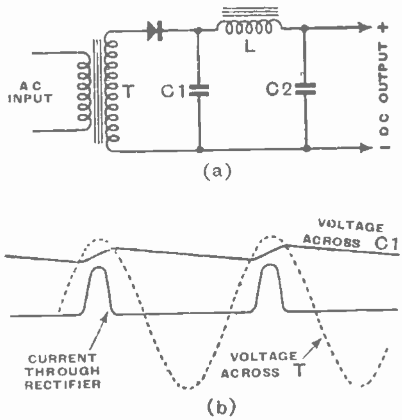



**Fig. 23.4—The voltages and rectifier current in a simple half-wave rectifier power-unit circuit (*a*) are shown in diagram (*b*)**

work under these conditions. So this circuit is used mainly for low-current high-voltage applications and in a.c./d.c. sets (p. 358), which are fed direct from the mains without a transformer. When such a unit is used on d.c., the rectifier serves no useful purpose. The set must of course be connected to d.c. mains in the way that enables the rectifier to pass current.

In Fig. 23.5 this half-wave circuit (*a*) is shown along with more elaborate arrangements; and for comparison the transformer is assumed to supply the same peak voltage in all of them.

(*a*) *Half-wave.* The no-load output voltage (i.e., the voltage when no current is being drawn) is very nearly equal to the peak input voltage; but, for the reasons given above, the voltage tends to drop considerably on load.

(*b*) *Full-wave.* By centre-tapping the transformer secondary, two type *a* circuits can be arranged, in series with the a.c. source, to feed the load in parallel. For the same voltage from the whole secondary, the output voltage is halved; but for a given rectifier rating the current is doubled, and at the same time is steadier, because each rectifier takes it in turn to replenish the reservoir. The resulting ripple, being at twice the a.c. frequency, is easier for the filter to smooth out; but it is more audible. The rectifier cathodes are " common ", and, as one is always out of use while the other is working, a single cathode can serve both anodes without any increase in emission being needed.

This is sometimes described as a 2-phase circuit (the transformer gives two phases 180° apart); and though less suitable than *a* for

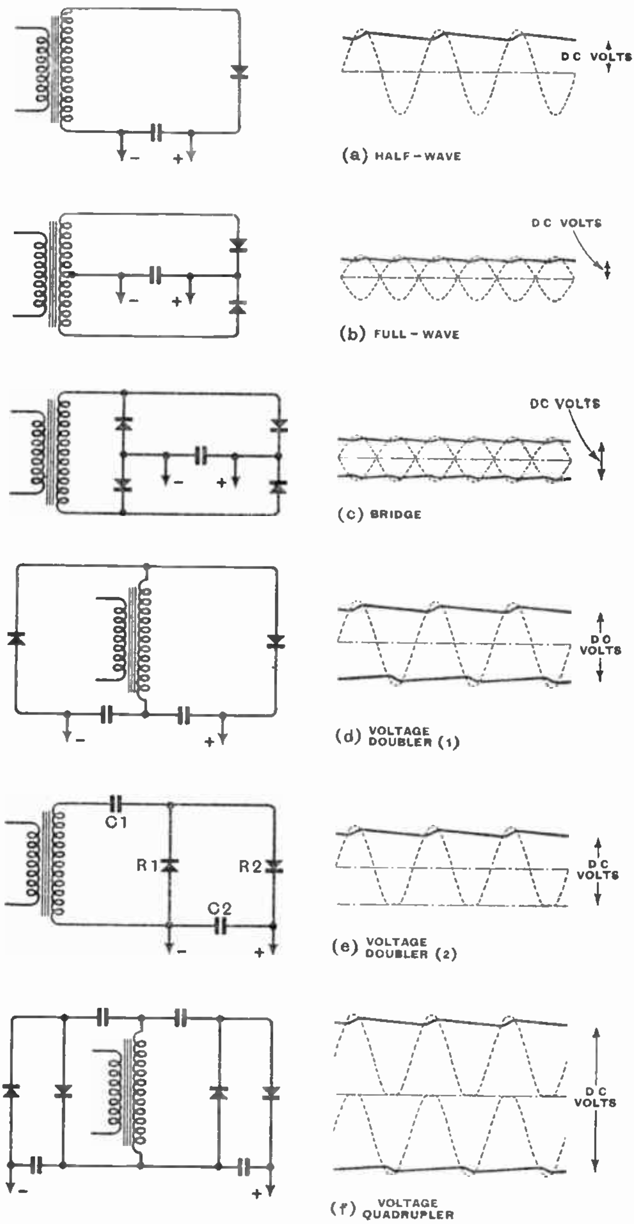



Fig. 23.5—Comparison of various rectifier circuits, the voltage supplied from the transformer being the same in all. Note the comparative output voltages

high voltage is better for large current. It is commonly used with valve rectifiers in a.c. mains-fed receivers and a.f. amplifiers.

(*c*) *Bridge*. Circuit *b* yields a terminal which is either positive or negative with respect to the centre tap by a voltage slightly less than half the total transformer peak voltage. By connecting a second pair of rectifiers in parallel with the same transformer winding, but in the opposite polarity, a terminal of opposite polarity is obtained. The total voltage between these two terminals is therefore twice that between one of them and the centre tap; i.e., it is nearly equal to the peak voltage across the whole transformer. The outputs of the two pairs of rectifiers are effectively in series. Unless a half-voltage point is wanted, the centre-tap can be omitted.

This arrangement gives approximately the same no-load output voltage as *a*, but with the advantages of full-wave rectification. Unfortunately three separate cathodes are needed, so the circuit is not commonly used with valve rectifiers; it is, however, quite usual with metal rectifiers, especially for low voltages.

(*d*) *Voltage Doubler*: *first method*. In contrast to *b*, the rectifiers are fed in parallel off the source to give outputs in series. As compared with *a*, the voltage is doubled and the ripple is twice the frequency. The voltage drop on load tends to be large, because each reservoir is replenished only once per cycle.

Like *c* (with centre tap), this circuit gives two supplies, one positive and one negative, from one transformer winding.

This circuit is a usual one for obtaining receiver h.t. with metal rectifiers.

(*e*) *Voltage Doubler*: *second method*. $R_1$ and $C_1$ act, as explained in connection with Fig. 16.6, to bring the negative peaks (say) to earth potential, so that the positive peaks are twice as great a voltage with respect to earth. $R_2$ and $C_2$ employ this as the input of a type *a* system, giving an output which, on no-load at least, rises to almost twice the transformer secondary peak voltage. Compared with *d*, one output terminal is common to the source, which may be convenient for some purposes. As $R_2C_2$ is a half-wave rectifier this system suffers from the disadvantages of that type, as well as the losses in $R_1$, so is confined to high-voltage low-current power units.

(*f*) *Voltage Quadrupler*. By connecting a second type *e* circuit in parallel with the transformer, to give an equal voltage output of opposite polarity, the total output voltage tends towards four times the peak of the supply. The conversion of *e* to *f* is analogous to the conversion of *a* to *d*.

## A.C. FROM D.C.

The advantage of an a.c. supply is that if its voltage is unsuitable for feeding the load direct it can easily be stepped up or down as required (p. 69). Where the supply is d.c. it may first have to be converted to a.c. in order to change the voltage. A common example is the 6- or 12-V d.c. system which is all that is available for car radio. In Fig. 6.2 we saw how a battery supply could be used to generate a square-wave a.c. supply by rapid to-and-fro switching. For convenience, the battery itself is made to do the switching. In one type of d.c.-to-a.c. converter, called a vibrator, the switching action is the same as in an electric bell. The current from the battery, passing through a magnet coil, attracts a piece of iron on a springy metal reed, which is really the switch arm. Its movement cuts off current from the coil, allowing the switch to close again; the process then repeats continuously. Alternatively the same thing can be done without mechanical movement, by means of some kind of electronic oscillator, as explained in Chapter 12. Transistors are particularly suitable, because the voltage drop across them when they are passing full current can be very small.

The basic principle is very simple, as Fig. 23.6 shows. S represents the switch—mechanical or electronic—and L the primary of the transformer. Let us assume that the resistance of this coil is neg-

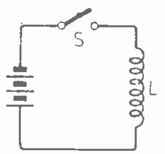



**Fig. 23.6—The basic principle of inductive d.c.-to-a.c. converters can be illustrated by this simple circuit**

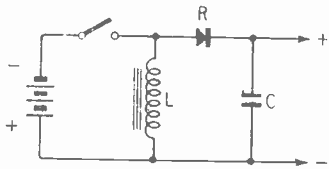



**Fig. 23.7—Addition of rectifier to Fig. 23.6 to convert back to d.c.**

ligible. The moment S is closed, current starts to rise. Its rise generates in L an e.m.f. opposing the battery e.m.f. (p. 66). Since there is no appreciable resistance in the circuit, the current has to rise at a steady rate. The greater the inductance of L, the slower the rise sufficient to balance the battery voltage. Before the current has had time to grow too much, S is opened and the current starts to fall. The e.m.f. generated by L is therefore reversed, and arrangements are made to use it. When the current has fallen to zero, S is again closed, and the cycle repeats. The frequency is usually about 100 c/s in a vibrator and 1,000 c/s in a transistor converter.

For simplicity, let us assume a 1 : 1 transformer ratio, so that an auto-transformer can be used, as shown in Fig. 23.7. During the "on" phase of the cycle the two extra components make no difference, because the rectifier is connected so that it passes no current. During the "off" phase, when the voltage across L is reversed, it conducts and the falling current through L charges C. The variations in voltage across C are smoothed by a filter before being passed to the load. The circuit is, in fact, essentially the same as Fig. 23.4*a* except that a switched battery takes the place of the a.c. supply. Introducing a two-winding transformer for a step-up does not affect the basic principle. And elaborations such as those shown in Fig. 23.5 can be made.

Fig. 23.8 shows one way in which a transistor can be used as an automatic switch. So long as current is rising through L, the e.m.f. generated in $L_b$ (which is our old friend the " reaction coil ") makes the base sufficiently negative to ensure a highly conducting

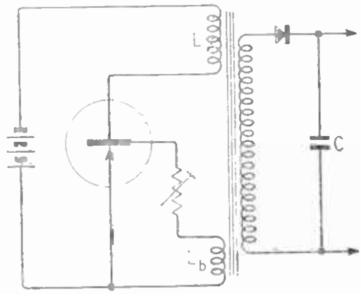



**Fig. 23.8—Transistor d.c. voltage raiser embodying the principle illustrated in Fig. 23.7**

path through the *p-n-p* transistor. When it can rise no more, the e.m.f. ceases, the transistor is cut off, and the falling current charges C. When it ceases, the small $I_{c0}$ through L starts the cycle all over again.

## E.H.T. FOR TELEVISION RECEIVERS

Although in the devices described in the last section the voltage waveform is square, the steadily rising and falling currents constitute a triangular waveform. If what might be called the utilization phase of the cycle is much shorter than the rest, the rate of current change during it must be much higher, and the waveform approaches a sawtooth (p. 328). The e.m.f. generated is correspondingly high.

This suggests the method commonly used to generate the extra-high voltage needed for the c.r. tube in a television receiver. The flyback between one line scan and the next necessitates a reversal of the full deflection current in less than one hundred-thousandth of a second. The rate of current change is thus very high, and when passed through a suitable step-up transformer generates the required voltage, which is rectified by a vacuum diode. Because of the high voltage between its cathode and the others in the set, it is heated by current from a small winding on the e.h.t. transformer.

Fig. 23.9 shows the general arrangement, in which $V_1$ is the valve supplying current to the line deflection coils DD, T is the e.h.t. transformer, and $V_2$ the rectifier.

In various ingenious ways the system is also made to boost its own h.t. voltage for $V_1$ above the rather inadequate level obtainable from the mains in an a.c./d.c. arrangement (which obviously has no step-up transformer). One of these is included in Fig. 23.9;

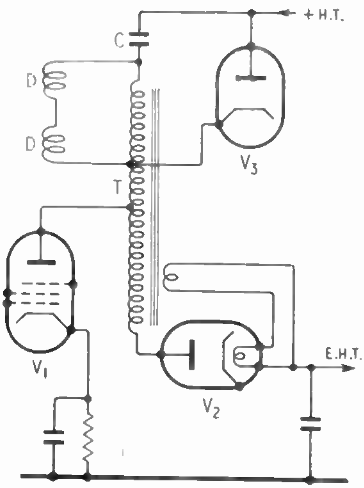



**Fig. 23.9—Typical e.h.t. generator in television receiver, making use of line fly-back**

the voltage developed across the deflection coils is rectified by $V_3$, charging the reservoir C in the polarity that aids the h.t. supply.

## FILTERS

We have already used a simple filter circuit ($R_1$ and $C_1$ in Fig. 16.10) to smooth out the unwanted r.f. left over from the rectification process in a detector. The same circuit is often used in power units, but the values of $R_1$ and $C_1$ are greatly different. The frequency of the ripple to be smoothed following a half-wave rectifier is usually 50 c/s, and 100 c/s with full-wave. For effective smoothing, the reactance of $C_1$ at this frequency must be low compared with the resistance $R_1$. And because of the relatively heavy current in a power unit, $R_1$ must itself be quite low if there is not to be an excessive wastage of volts.

For this reason an iron-cored choke coil is often substituted. It has to be carefully designed if the core is not to be saturated by the d.c. (p. 318) which greatly reduces the inductance and therefore the effectiveness of the coil for smoothing. A typical filter, $LC_2$, is seen in the half-wave rectifier circuit in Fig. 23.4.

Although a reservoir capacitor brings the rectified voltage almost up to the *peak* input voltage when no current is being taken, the voltage falls off fairly steeply with increase of load current. In technical language, the regulation is not very good. For purposes where it is important that the voltage should remain steady in spite of a fluctuating load current—for example, high-power Class B amplifiers—what is called a choke-input filter is used. In the circuit diagram this looks the same as, say, Fig. 23.4*a*, except that the reservoir capacitor is missing. This is not the only difference, however, for the choke has to be of a special kind, known as a swinging choke. Instead of being designed so that its inductance varies as little as possible, its inductance increases substantially as current falls, with the aim of maintaining a nearly constant voltage across it. The output voltage falls very rapidly up to, say, one tenth full load current (below which the current is not allowed to fall) and thereafter is relatively steady.

## Decouplers

Because of the cost and weight of a choke having a high reactance at low frequencies, a resistor is used for smoothing whenever possible, and for this reason electrolytic capacitors of very high capacitance—32–100 μF or even more—are now customary.

Even so, their impedance to the signal-frequency currents flowing from + to − h.t. may not be negligible, and a potential difference set up across this impedance by (say) the signal current in the output stage is passed on to all the other valves fed from the same supply and may seriously upset the working, possibly even causing

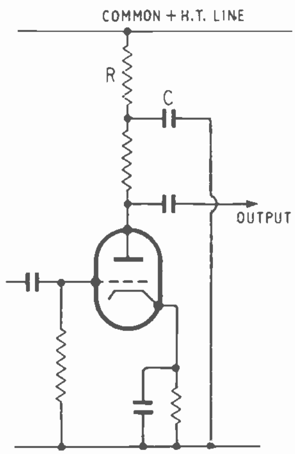



**Fig. 23.10—Decoupling a valve from the h.t. line is performed by inserting R to block signal currents, and providing C to give them a path back to earth**

continuous oscillation. We have, in fact, a form of feedback.

To obviate such undesirable conditions, *decoupling* is used. It is simply an individual filter for those valves liable to be seriously affected. As those valves are generally the preliminary stages, taking only a small proportion of the total anode current, a resistor having sufficient impedance at the frequencies concerned can be used. The loss in voltage may actually be desirable, as these stages are often required to run at lower voltages than the power output stage; in which case a single cheap component is made to serve the double purpose of a voltage-dropping resistance and (in conjunction with a capacitor) a decoupler. Of such is the essence of good commercial design.

These decoupling components are shown as R and C in Fig. 23.10. The signal currents tend to take the easy path, through C, back to the cathode, rather than through R and the h.t. source. The larger *C* and *R*, the more complete the decoupling.

Similar arrangements are used for keeping screens and such auxiliary electrodes at the necessary constant potentials.

It is the extensive use of decoupling and filtering that renders multi-stage circuit diagrams so alarming for the novice to contemplate. But once their purpose has been grasped it is easy to sort them out from the main signal circuits.

## Grid and Base Bias

From time to time we have seen various methods of providing bias voltage, and it only remains to review them.

The most obvious way is by means of a battery in series with the circuit between grid (or base) and cathode (or emitter), as in Figs. 12.5 and 22.8.

A slight modification is the parallel-feed system, shown in Fig. 12.6, where *R* is a resistance so high that it causes negligible loss of the signal applied through $C_g$. A suitable choke coil is sometimes used instead of R; for example, if there is a likelihood of grid current and variation of the effective bias by a substantial voltage drop across R would be undesirable.

Quite obviously it would be possible to replace a bias battery by a mains power unit designed to give the correct voltage. Except for high power apparatus such as senders this is hardly ever done, because the power involved is so small and there are more convenient alternatives.

We came across one of these alternatives in Fig. 16.12, where a resistor $R_3$ inserted between the cathode of an amplifying valve and the common negative line was described as a biasing resistor. The space current (i.e., the whole current passing through the valve, to anode, and screen grid if any) causes a voltage drop across it, positive at the cathode end so that relative to the cathode the grid is biased negatively. The required amount of bias is obtained by choice of resistance.

One advantage of this method is that it is to some extent self-adjusting. If for any reason the valve passes more than the normal current, it thereby increases its bias, which tends to correct the condition. Another advantage is that it simplifies the providing of each valve in a set with its most suitable bias.

On p. 347 we saw that this method provides the wrong polarity for transistor base bias, but that its automatic regulating properties can be used to control the bias from a potential divider.

If only a simple resistor is used, it carries not only the steady (or z.f.) component of current but also the signal current, which produces a voltage in opposition to the input. We have, in fact, negative feedback (p. 312). If that is wanted, well and good; but if not, then something has to be done to provide a path of relatively negligible impedance for the signal currents, without disturbing the z.f. resistance. A by-pass capacitor, $C_3$ in Fig. 16.12, serves the purpose; for r.f. a value of 0·1 μF or even less is enough, but for the lowest audio frequencies nothing less than 25 or 50 μF will do.

This individual cathode bias system is inapplicable to battery sets in which all the filaments are connected in parallel. In such sets a common method is to connect a resistor between the filament battery and "— h.t.", and to tap off from it the required bias to each grid circuit.

The principle of still another biasing method emerged when we considered Fig. 16.7*b*. All that is needed is a grid capacitor and leak having a time constant that is long relative to one signal cycle (p. 212). The peculiarity of the method is that the bias voltage is approximately equal to the peak value of the signal reaching the grid. With a.f. signals it is sometimes used instead of cathode bias for the stage immediately following a detector (Fig. 16.13).

A very common application of grid-leak bias is for oscillators, where it has the advantages of giving a high valve conductance for starting and an automatically adjusted bias for running (p. 163). If the positive feedback in the oscillator is very large and the time constant equal to the duration of many cycles, the first burst of oscillation may generate such a large bias that anode current is cut off, so quenching the oscillation, which cannot resume until some of the negative charge has had time to leak off the grid. Oscillation then restarts, is again quenched, and so on. This phenomenon of intermittent oscillation is known as *squegging*, and is sometimes usefully employed. Elsewhere, it is a nuisance, which can be cured by reducing the positive feedback or the time constant of the grid circuit or both. The blocking oscillator (p. 330) is sometimes considered as an extreme form of squegging.

Both the grid leak and the cathode resistor methods of developing grid bias are referred to as automatic grid bias or *auto-bias*.

# Appendix 1

## Alternative Technical Terms

The reader of books and articles on radio is liable to be confused by the use of different terms to mean the same thing. In the following list the first alternatives to be mentioned are those preferred in this book. The associated terms are not necessarily exact equivalents. Terms distinctively American are printed in italics. The page numbers refer to where the terms are defined or explained.

| | Page |
|---|---|
| Accumulator—Secondary battery | 356 |
| Aerial—*Antenna* | 193 |
| Amplification—Gain | 142 |
| Anode—*Plate* | 117 |
| Anode a.c. resistance—Anode incremental resistance—Valve impedance —*Plate impedance* | 119 |
| Anode battery—H.T.—"*B*" *battery* | 117 |
| Atmospherics—Strays—X's—*Static* | 200 |
| Audio frequency (a.f.)—Low frequency (l.f.)—Speech frequency—Voice frequency | 18 |
| Auto-bias—Self bias | 369 |
| Automatic gain control (a.g.c.)—Automatic volume control (a.v.c.) | 290 |
| Beam tetrode—Kinkless tetrode | 310 |
| Capacitance—Capacity | 50 |
| Capacitor—Condenser | 53 |
| Characteristic resistance—Characteristic impedance—Surge impedance | 177 |
| Coaxial—Concentric | 174 |
| Common cathode (or grid, anode, emitter, base, collector)—Earthed cathode (etc.)—*Grounded cathode* (etc.) | 340 |
| Detection — Rectification — *Demodulation* (British usage originally reserved this term for a different phenomenon) | 206 |
| Dielectric—Insulating material | 53 |
| Dynamic resistance—*Antiresonant impedance* | 109 |
| Earth—*Ground* | 44 |
| Filament battery—L.T.—"*A*" *battery* | 117 |
| Frame aerial—*Loop antenna* | 204 |
| Frequency—Periodicity | 18 |
| Frequency-changer—Mixer—First detector | 262 |
| Grid battery—Bias battery—G.B.—"*C*" *battery* | 124 |
| Hard—High-vacuum | 329 |
| Harmonic—Overtone—Partial | 297 |
| Heptode—Pentagrid | 268 |
| Image interference—Second-channel interference | 273 |
| Inductor—Coil | 64 |
| Interference—Jamming | 252–3 |
| Intermediate frequency (i.f.)—Supersonic frequency | 261 |
| Lead—Connecting wire | 35 |
| Loss—Attenuation | 148 |
| Moving coil (of loudspeaker)—Speech coil—*Voice coil* | 320 |
| Moving-coil loudspeaker—*Dynamic loudspeaker* | 320 |
| Mutual conductance—Slope—*Transconductance* | 121 |
| Negative feedback—*Degeneration* | 312 |
| Noise—Machine interference—*Man-made static* | 270 |
| Parallel—Shunt | 36 |
| Peak value—Crest value—Maximum value | 73 |
| Permittivity—Dielectric constant—Specific inductive capacity (s.i.c.) | 53 |
| Picofarad—Micromicrofarad | 53 |
| Q—Q factor—Magnification—Storage factor | 102 |
| Quality (of reproduced sound)—Fidelity | 294 |

# Appendix 2

## SYMBOLS AND ABBREVIATIONS

### General Abbreviations

| | |
|---|---|
| a.c. | alternating current |
| a.f. | audio frequency |
| a.g.c. | automatic gain control |
| c.r. | cathode ray |
| d.c. | direct current |
| e.h.t. | extra-high tension |
| e.m.f. | electromotive force |
| h.f. | high frequency |
| h.t. | high tension |
| i.f. | intermediate frequency |
| l.f. | low frequency |
| m.m.f. | magnetomotive force |
| p.d. | potential difference |
| r.f. | radio frequency |
| r.m.s. | root-mean-square |
| v.f. | video frequency |
| v.h.f. | very high frequency |
| z.f. | zero frequency |

### Greek Letters

| *Letter* | *Name* | *Usual Meaning* |
|---|---|---|
| $\alpha$ | alpha | current amplification factor (of transistor) |
| $\epsilon$ | epsilon | permittivity |
| $\theta$ | theta | an angle |
| $\lambda$ | lambda | wavelength |
| $\mu$ | mu | (1) permeability<br>(2) voltage amplification factor (of valve)<br>(3) one millionth of (as a prefix to a unit symbol) |
| $\pi$ | pi | $\frac{\text{circumference}}{\text{diameter}}$ of circle ($=3{\cdot}14159\ldots\ldots$) |
| $\rho$ | rho | resistivity |
| $\phi$ | phi (small) | angle of phase difference |
| $\Phi$ | (capital) | magnetic flux |
| $\omega$ | omega (small) | $2\pi f$ |
| $\Omega$ | (capital) | ohm |

### Quantities and Units

| *Quantity* | *Symbol* | *Unit* | *Abbreviation for unit* |
|---|---|---|---|
| Time | $t$ | second | s or sec |
| Time period of one cycle | $T$ | second | s or sec |
| Frequency | $f$ | cycle per second | c/s |
| Wavelength | $\lambda$ | metre | m |
| Electromotive force | $E$ | volt | V |
| Potential difference | $V$ | volt | V |
| Current | $I$ | ampere | A |
| Power | $P$ | watt | W |
| Capacitance | $C$ | farad | F |
| Self inductance | $L$ | henry | H |
| Mutual inductance | $M$ | henry | H |
| Resistance | $R$ | ohm | Ω |
| Reactance: Capacitive | $X$: $X_C$ | ohm | Ω |
| Reactance: Inductive | $X$: $X_L$ | ohm | Ω |
| Impedance | $Z$ | ohm | Ω |
| Conductance | $G$ | mho | ℧ |
| Quantity of electricity, or charge | $Q$ | coulomb | C |

| *Quantity* | *Symbol* | *Unit* | *Abbreviation for unit* |
|---|---|---|---|
| Magnetic flux | $\Phi$ | weber | Wb |
| $Q$ factor, $X/R$ | $Q$ | | |
| Signal gain or loss | | decibel | dB |

*Note.*—Small letters ($e$, $v$, $i$, etc.) are used to indicate instantaneous values.

## Unit Multiple and Submultiple Prefixes

| *Symbol* | *Read as* | *Means* |
|---|---|---|
| M | mega- | one million ( $\times 10^6$) |
| k | kilo- | one thousand ( $\times 10^3$) |
| m | milli- | one thousandth ( $\times 10^{-3}$) |
| $\mu$ | micro- | one millionth ( $\times 10^{-6}$) |
| n | nano- | one thousand-millionth ( $\times 10^{-9}$) |
| p (or $\mu\mu$) | pico- (or micromicro-) | one billionth ( $\times 10^{-12}$) |

Examples: 1 Mc/s = 1,000,000 c/s
1 kΩ = 1,000 Ω
1 mA = 0·001 A
1 μH = 0·000001 H
1 pF = 0·000000000001 F

## Valve Abbreviations

| | |
|---|---|
| k | cathode |
| g | grid |
| $g_1$ | first grid (nearest cathode) |
| $g_2$ | second grid; and so on |
| a | anode |
| $\mu$ | voltage amplification factor |
| $r_a$ | anode a.c. resistance |
| $g_m$ | mutual conductance |
| $g_c$ | conversion conductance (of frequency-changer) |

*Note.*—Symbols are frequently combined, thus—

$I_a$ = anode current
$V_{g2}$ = voltage at second grid
$R_a$ = resistance connected externally to anode
$c_{gk}$ = internal capacitance from grid to cathode

Capital letters are used for associated items outside the valve; small letters for items inside the valve itself.

## Transistor Abbreviations

CONFIGURATION

| *Common emitter* ($e$) | *Common base* ($b$) | *Common collector* ($c$) | |
|---|---|---|---|
| $\alpha'$, $\alpha_e$ | $\alpha$, $\alpha_b$ | $\alpha''$, $\alpha_c$ | current amplification factor |
| $r'_c$, $r_{ce}$ | $r_c$, $r_{cb}$ | $r''_c$, $r_{cc}$ | collector resistance in T network |

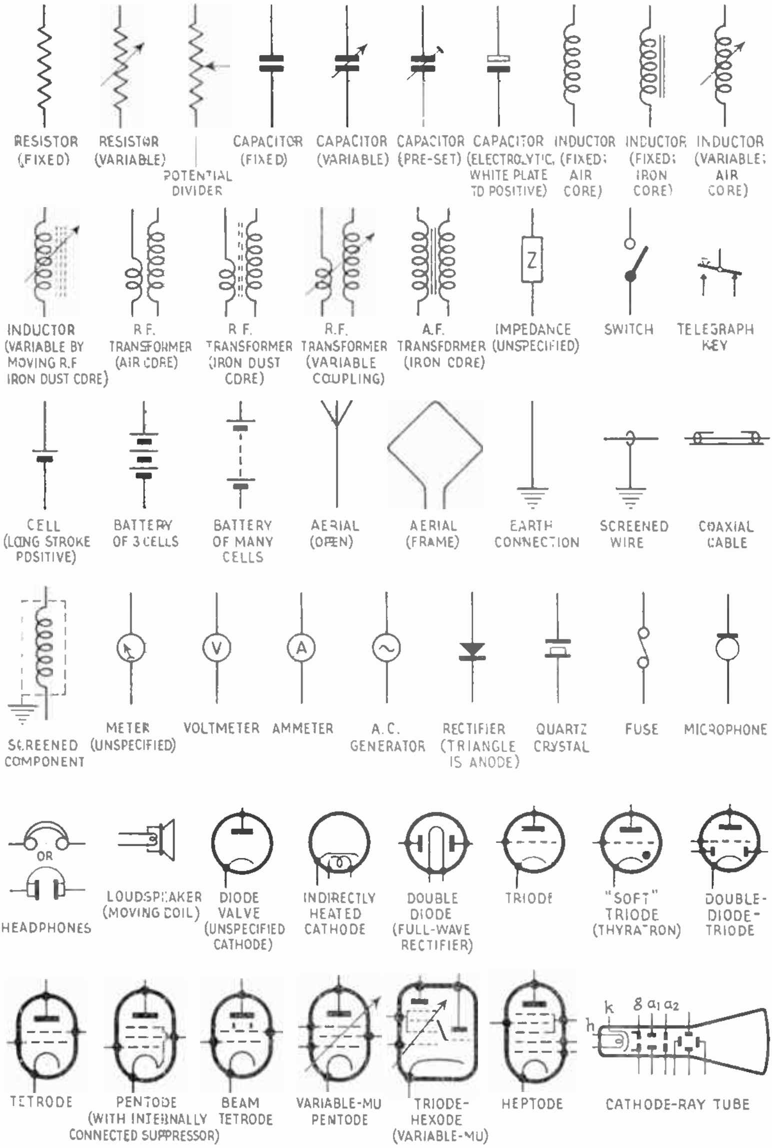
And similarly for other parameters; see Chapter 22.

## Special Abbreviations used in this Book

| | |
|---|---|
| $f_r$ | frequency of resonance |
| $f_o$ | frequency of oscillation |
| $f'$ | frequency off-tune |
| $A$ | voltage amplification |
| $B$ | feedback factor |

# Appendix 3

## Circuit Symbols




Transistors: see Fig. 10.16

# Appendix 4

## Decibel Table

The decibel figures are in the centre column: figures to the left represent decibel loss, and those to the right decibel gain. The voltage and current figures are given on the assumption that there is no change in impedance.

| *Voltage or current ratio* | *Power ratio* | ← − *dB* + → | *Voltage or current ratio* | *Power ratio* |
|---|---|---|---|---|
| 1·000 | 1·000 | 0 | 1·000 | 1·000 |
| 0·989 | 0·977 | 0·1 | 1·012 | 1·023 |
| 0·977 | 0·955 | 0·2 | 1·023 | 1·047 |
| 0·966 | 0·933 | 0·3 | 1·035 | 1·072 |
| 0·955 | 0·912 | 0·4 | 1·047 | 1·096 |
| 0·944 | 0·891 | 0·5 | 1·059 | 1·122 |
| 0·933 | 0·871 | 0·6 | 1·072 | 1·148 |
| 0·912 | 0·832 | 0·8 | 1·096 | 1·202 |
| 0·891 | 0·794 | 1·0 | 1·122 | 1·259 |
| 0·841 | 0·708 | 1·5 | 1·189 | 1·413 |
| 0·794 | 0·631 | 2·0 | 1·259 | 1·585 |
| 0·750 | 0·562 | 2·5 | 1·334 | 1·778 |
| 0·708 | 0·501 | 3·0 | 1·413 | 1·995 |
| 0·668 | 0·447 | 3·5 | 1·496 | 2·239 |
| 0·631 | 0·398 | 4·0 | 1·585 | 2·512 |
| 0·596 | 0·355 | 4·5 | 1·679 | 2·818 |
| 0·562 | 0·316 | 5·0 | 1·778 | 3·162 |
| 0·501 | 0·251 | 6·0 | 1·995 | 3·981 |
| 0·447 | 0·200 | 7·0 | 2·239 | 5·012 |
| 0·398 | 0·159 | 8·0 | 2·512 | 6·310 |
| 0·355 | 0·126 | 9·0 | 2·818 | 7·943 |
| 0·316 | 0·100 | 10 | 3·162 | 10·00 |
| 0·282 | 0·0794 | 11 | 3·55 | 12·6 |
| 0·251 | 0·0631 | 12 | 3·98 | 15·9 |
| 0·224 | 0·0501 | 13 | 4·47 | 20·0 |
| 0·200 | 0·0398 | 14 | 5·01 | 25·1 |
| 0·178 | 0·0316 | 15 | 5·62 | 31·6 |
| 0·159 | 0·0251 | 16 | 6·31 | 39·8 |
| 0·126 | 0·0159 | 18 | 7·94 | 63·1 |
| 0·100 | 0·0100 | 20 | 10·00 | 100·0 |
| $3·16 \times 10^{-2}$ | $10^{-3}$ | 30 | $3·16 \times 10$ | $10^{3}$ |
| $10^{-2}$ | $10^{-4}$ | 40 | $10^{2}$ | $10^{4}$ |
| $3·16 \times 10^{-3}$ | $10^{-5}$ | 50 | $3·16 \times 10^{2}$ | $10^{5}$ |
| $10^{-3}$ | $10^{-6}$ | 60 | $10^{3}$ | $10^{6}$ |
| $3·16 \times 10^{-4}$ | $10^{-7}$ | 70 | $3·16 \times 10^{3}$ | $10^{7}$ |
| $10^{-4}$ | $10^{-8}$ | 80 | $10^{4}$ | $10^{8}$ |
| $3·16 \times 10^{-5}$ | $10^{-9}$ | 90 | $3·16 \times 10^{4}$ | $10^{9}$ |
| $10^{-5}$ | $10^{-10}$ | 100 | $10^{5}$ | $10^{10}$ |
| $3·16 \times 10^{-6}$ | $10^{-11}$ | 110 | $3·16 \times 10^{5}$ | $10^{11}$ |
| $10^{-6}$ | $10^{-12}$ | 120 | $10^{6}$ | $10^{12}$ |

# INDEX

# Index

*This index, besides having an exceptionally large number of references to help the reader find what he wants quickly, is unusual because it includes not only the technical terms actually used in the book but also the equivalents listed in Appendix 1. So it is also, in effect, a glossary of terms, American as well as British, and can be consulted to find the meanings of terms encountered elsewhere. If the index reference is to page 371 or 372, the subject should be looked up again under the first alternative given there.*