From 'RADIO RECEIVERS', edited by W.Gosling;
published by Peter Peregrinus for I.E.E.,1986

Chapter 1

# Origins and types of radio receivers

Professor D. G. Tucker

## 1.1 Origins

The origins of the radio receiver obviously lie in the experimental discovery of electromagnetic radiation as such, and in the means used for demonstrating or detecting its existence. Thus the spark-gap detector of Heinrich Hertz [1], used in his classic experiments of 1886–88, could be regarded as the first radio receiver. An even earlier origin could be claimed if David E. Hughes's demonstration in 1880 of the detection of a remote transient current by a loose-contact microphone unconnected to the generating circuit could be accepted as radio [2]. Certainly Hughes's detector was nearer to the type of detector used in the early radio-telegraphy experiments of Marconi and others in the 1890s, and quite widely during the first decade of the 20th century, i.e. the coherer. It must be emphasised that in all early radio systems the purpose of the receiver was to *detect* the presence or absence at the aerial of the electromagnetic wave representing the signal. The coherer, which was some sort of loose-contact device, generally comprising metal particles between two electrodes in a tube, detected the wave by the change of resistance between the electrodes; with most metals the resistance fell as the RF voltage made the particles 'cohere', and this could enable a visual or aural indicator to be operated in a local-battery circuit [3].

More important in early commercial radio telegraphy from, say, 1902 to the First World War, was the magnetic detector in which the received wave changed the state of a piece of magnetic material and thus enabled a local circuit to be activated as in the coherer receivers. Many different physical phenomena were exploited in various kinds of detector [4]. Many had the disadvantage, shared with the coherer and magnetic detectors, of not being inherently self-restoring after each wave packet had been received; many coherers, for example, had a trembler-bell type of tapper to reset them.

The kind of detector which really laid the foundation of the modern radio receiver was the rectifier detector. The origins of this are not very clear. Early electrolytic detectors, such as that of Pupin in 1898–99 [5], were effectively biased rectifiers; and some thermojunction-type detectors may well have operated at least partially as rectifiers. Probably the first detector which was designed consciously to use a rectifier effect was the thermionic diode of J. A. Fleming [7], first used in 1904, followed within a couple of years by the crystal detector [8]. It was realised that these could be used without a local-battery circuit by 'detecting' the envelope components of the RF wave (e.g. by means of a telephone earpiece), or even the DC component (e.g. by means of a galvanometer).

The detector was one essential component of the early radio receiver; the other was the tuning arrangement. In early experiments using spark transmitters, the signal transmitted had a very wide spectrum and the receiver likewise was barely resonant. Thus as soon as radio telegraphy began to be used commercially, neighbouring links would interfere with one another. The solution lay with the system of 'syntony' (i.e. tuning) developed by Lodge during the 1890s [9]. Marconi applied the idea by using aerial circuits loosely coupled to the rest of the system (i.e. to the spark-gap in the transmitter and to the detector in the receiver) so that they could be resonant and only lightly damped [10]. When transmitter and receiver were tuned to the same frequency, interference with other links was greatly reduced. Adjustable inductors provided the means of tuning, as shown in the circuit diagram of Fig. 1.1.

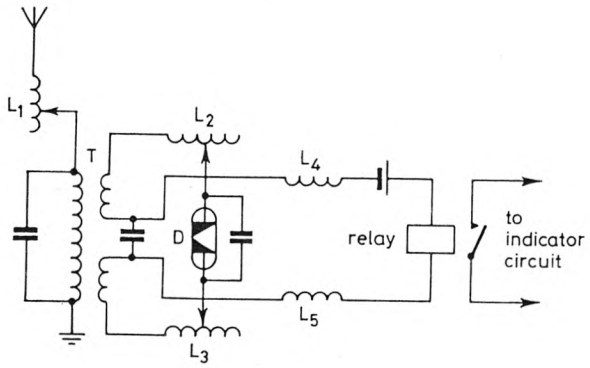


**Fig. 1.1** *Marconi receiver circuit 1900 (based on British Patent No. 7777, April 1900)*
T Transformer
$L_1$ Aerial tuning inductor
$L_2$, $L_3$ Coherer circuit tuning inductors
$L_4$, $L_5$ Chokes to prevent relay circuit clamping tuning
D Coherer



## 1.2 Pre-electronic development

Improvement in transmitters during the first decade of the 20th century permitted certain improvements in receivers to be effective. The replacement of spark-generators by oscillating arcs or by high-frequency alternators led to more coherent transmissions, and not only made sharp tuning worthwhile and a multiplicity of point-to-point links in one area really feasible, but also permitted the effective utilisation of the heterodyne principle invented by Fessenden in 1902 [11], although not adopted into general use until some years later [12].

Fessenden's heterodyne system combined the received signal with the output of a local low-power RF alternator, the frequency of which differed from that of the signal by, say, 1000 Hz. The combining was done in what was effectively a Bell telephone receiver without a permanent magnet, i.e. with a soft-iron core, with a coil for the signal and another for the local oscillation. As the force on the diaphragm is proportional to the square of the flux, then the audible (or 'beat') signal is proportional to

$$i_s\, i_o \cos{(\omega_s - \omega_o)t}$$

where $i_s \cos \omega_s t$ is the signal current and $i_o \cos \omega_o t$ is the local current (the non-alternating and the double-frequency components are, of course, inaudible). Thus the receiving operator hears bursts of tone at around 1000 Hz when mark signals are being transmitted. This system was found to be increased in sensitivity by using the force between the two coils, in a 'dynamometer' receiver, and detection ranges obtained with heterodyne receivers were much improved over those of conventional systems; moreover, the selectivity obtained was much better owing to the discrimination of the telephone receiver and of the ear.

A heterodyne receiver in which the nonlinear interaction needed to obtain the beat signal was obtained not in the telephone, but by using a rectifier as shown in Fig. 1.2; this was the type of circuit used in the Arlington–Salem tests of 1913, described by Hogan [12], in which an improvement of five times in the 'audibility factor' over the previously used electrolytic detector was claimed.

The improvement of waveform provide by the oscillating arc and the RF alternator also encouraged the experimental development of radio telephony, although it was demonstrated even with spark-generated waves. Fessenden was the pioneer here too, as far as radio was concerned, although the basic principle of transmitting a speech signal was shown by Leblanc in 1886 [13] for line transmission. Fessenden had several ideas for modulation [14], but the one used in practice was to use the resistance of a carbon microphone to modulate the aerial current directly. For such envelope-modulated signals, the receiver would be as in Fig. 1.2 without the local generator circuit, the rectifier plus $C_2$ (when properly proportioned) producing the envelope (i.e. speech) waveform across the telephone earpieces. However, radio telephony made little progress

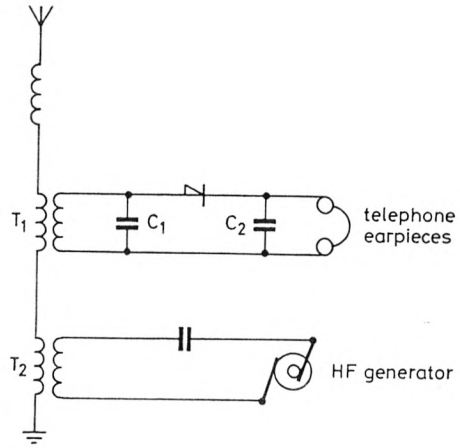


**Fig. 1.2** *Heterodyne receiver, 1913 (based on Fig. 11 of HOGAN, J. L., Proc. Inst. Radio Engs., 1913, 1, p. 87)*
$T_1$, $T_2$ Transformers
$C_1$ Tuning capacitance
$C_2$ 'Telephone condenser' to bypass RF currents

before the adoption of the thermionic valve. The reasons for this are not entirely clear: it was entirely practicable; Ruhmer's textbook [15] of 1907 demonstrates this; and when radio broadcasting became general in the 1920s, crystal-detector receivers without thermionic valves were quite common.

### 1.3 The thermionic valve

The undirectional conductivity between a plate and the filament inside an electric filament lamp had been well explored by 1900, but the first practical application for the effect appears to be the thermionic diode radio detector of Fleming, patented in 1904 [16]. This, when biased to make the anode positive with respect to the filament, was a rectifier of reasonable efficiency, although generally inferior to the crystal detector, which came two years later. The thermionic triode was invented by Lee de Forest [17]; the story is well-known, as is that of the patent litigation between Fleming and de Forest, which, by creating uncertainty as to who had the patent rights, delayed the development of applications of the thermionic valve throughout the patent period. De Forest called his device the 'audion', thus making it apparent that he regarded it as another form of detector; more generally the term 'valve' was used in Britain because of the obvious analogy between the diode and a hydraulic valve, although the term 'tube' or 'vacuum tube' became adopted in the USA.

For some years the audion was used only as a more sensitive detector for radio telegraphy. Its third electrode, or 'grid', was connected by a capacitor to



the aerial circuit without any explicit DC conducting path; the grid was thus floating so that the valve had to be soft; and detection or rectification was the only function the valve could perform. It was not until Lowenstein [18] discovered the favourable effect of a negative continuous bias voltage on the grid (with the concomitant omission of the capacitance between the grid and the external circuit) that the triode could be used as a signal amplifier, and this was therefore the crucial invention that transformed the design of radio receivers. With the understanding of the principles of design of triode circuits which came from Armstrong's formulation [19] of the use of measured characteristic curves, it became possible to use the triode as a reasonably linear amplifier by keeping the signal voltage swing within the range where at the positive extremity the grid current became excessive, and at the negative extremity the valve cut off. Alternatively, by using a larger negative bias, the valve could be operated as a simultaneous detector and amplifier. This gave a more sensitive radio receiver than had previously been possible with equipment of any robustness.

More sensitivity and more frequency selectivity were obtained by using several triode stages in cascade, with tuned intervalve couplings. With more sensitivity it was of course necessary to have the increased selectivity, for signals were being detected that were at a low enough level to be competing with noise, or 'static' as it was called. It is not quite clear when multistage receivers were first used, but Langmuir's patent [20], filed in 1913 but not issued until 1918 and then with some additional coverage, must represent a very early design. It is shown in Fig. 1.3. It was intended for radio telegraphy where each 'mark' signal comprised a group of short bursts of RF wave, the repetition frequency of these bursts being called the group frequency. The first two stages of the receiver amplified the RF wave; the third stage was an amplifier detector, so that the coupling transformer to the fourth stage itself were designed to be effective at the group frequency, which was usually an audio frequency.

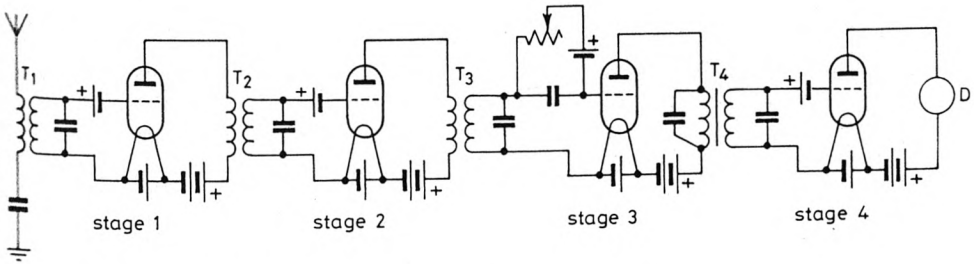


**Fig. 1.3** *Langmuir's four-stage triode-valve receiver, 1913–18*
$T_1$, $T_2$, $T_3$ Transformers tuned to RF
$T_4$ Iron-cored transformer tuned to group (or audio) frequency
D Radio telegraph recorder or telephone receiver
Stages 1 and 2 are RF amplifiers, stage 3 is on amplifier-detector, stage 4 is a group (or audio) frequency amplifier

In later and more effective multistage receivers, the RF transformers had a step-up ratio and the tuning was done on the primary. The tuning (variable) capacitors would be ganged for convenience of tuning, and the system was called TRF (i.e. tuned radio frequency) reception. When many stages of RF amplification were used, difficulties, especially that of self-oscillation, were encountered because of the effect of anode-to-grid self-capacitance in the valves. There were many ways of dealing with this problem [21], and one of the most successful was the 'neutrodyne' circuit devised by Hazeltine around 1923 [22]. This circuit is, in essence, of the type shown in Fig. 1.4*a*, which may be redrawn as a bridge circuit as in Fig. 1.4*b*. In each stage, the anode winding of the output transformer is divided as shown, and one winding feeds back to the grid through a capacitor $C$. Then the bridge circuit of *b* is exactly equivalent, so that if $L_1 = L_2$ and $C$ = self-capacitance from anode to grid, then none of the output voltage across AB appears at the input PO.

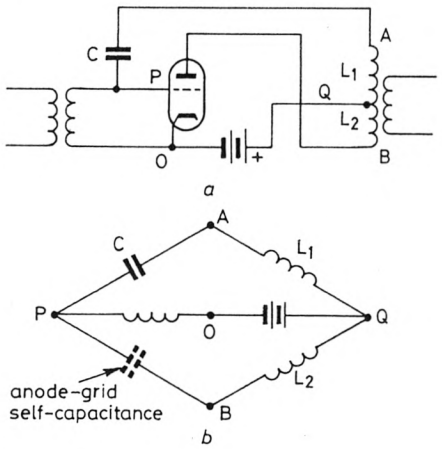


**Fig. 1.4** *The 'neutrodyne' principle*

The TRF system was popular for a time as it was reliable and the ganging of the tuning was straightforward, all the RF tuned circuits having the same capacitance for any given frequency. Nevertheless, the actual mechanics of coupling the different variable capacitors so that control could be by a single knob led to much inventive activity, there being some 62 separate patents in the USA alone [23].

The requirements of the tuning system were greatly eased by a phenomenon which became known as 'detector discrimination'. Provided detection was carried out at a high enough signal level (which was almost certain if there were two or more RF stages), then if the tuning could reduce any interfering amplitude-modulated signal to a level, say, 20 dB below the wanted signal, the nonlinear discrimination of the detector would improve this ratio to 46 dB. The effect was adequately analysed around 1930 [24].



## 1.4 Feedback

### *1.4.1 Reaction or regenerative receiver*

An early discovery in the work on triode-valve circuits, once the invention of negative grid-bias had made triodes into effective amplifiers, was the effect of feedback from output to input. The discovery, and more particularly its understanding and exploitation, may with some confidence be credited to Armstrong early in 1913, although there were many contenders for the priority of invention, and in the USA, various courts alternated the decision between Armstrong and de Forest until the Supreme Court eventually (in 1934) awarded it to de Forest. There were also British and European inventors who did not compete in the US litigation but have good claims. As with the litigation over the invention of the thermionic valve itself, there were millions of dollars at stake, and great bitterness was aroused. The story is told elsewhere [25].

The feedback was positive, i.e. part of the output signal was fed back to reinforce the input. Its importance lay in two methods of exploitation:

1 As a generator of high-frequency oscillations.
2 As an amplifier of very high amplification and selectivity.

In the first application it was vastly more flexible, versatile and convenient than other methods of generating HF waves and gradually displaced them.

In the second application it made a very sensitive radio receiver. A typical circuit arrangement is shown in Fig. 1.5. Armstrong squeezed even more out of it by applying feedback to both RF and audion in an amplifier-detector type

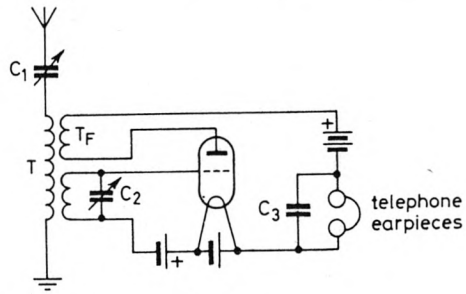


**Fig. 1.5** *Radio receiver with 'reaction', i.e. with feedback*
$T_F$ Feedback winding on aerial tuning transformer T
$C_1$, $C_2$ Tuning capacitors
$C_3$ Bypass capacitor

of receiver, as shown in fig. 1.6, where the legend is taken from his paper of 1915 [26]. In the more usual circuit of Fig. 1.5, control of the feedback (then more usually called 'reaction') would be by adjustment of the mutual coupling between $T_F$ and the grid-circuit coil, and the reaction knob was a prominent

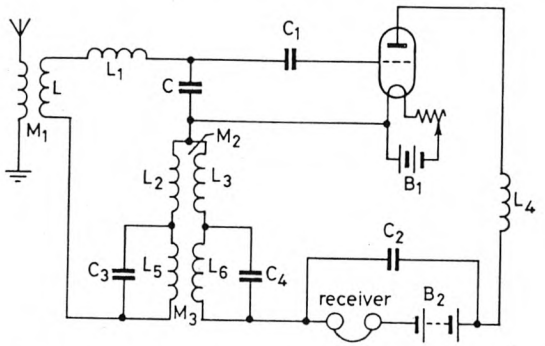


**Fig. 1.6** *Armstrong's circuit of 1915 providing feedback at RF and audio in an Amplifier-detector type of receiver*
'Here $M_2$ represents the coupling for the radio frequencies and the coils are of relatively small inductance. $M_3$ is the coupling for the audio frequencies, and the transformer is made up of coils having an inductance of the order of a henry or more. The condensers $C_3$ and $C_4$ have the double purpose of tuning $M_3$ to the audio frequency, and of by-passing the radio frequencies. The total amplification of weak signals by this combination is about 100 times, with the ordinary bulb. On stronger signals, the amplification becomes smaller as the limit of the audion's response is reached'.

feature of many early broadcast receivers. Obviously it was not possible to preset it, because the adjustment was critical even when the tuning controls were not altered; too much reaction would cause the receiver to oscillate and thus not only prevent reception but also cause radiation of the oscillation and interference with neighbouring receivers; too little reaction would provide insufficient sensitivity and selectivity. Valve characteristics were very sensitive to variations in the battery voltages as well as changeable with age, and the reaction knob had to be frequently reset.

In spite of its operating disadvantages, the feedback (or reaction or regenerative) receiver was immensely important as a stage in the development of radio communication. The very high sensitivity obtained when critically adjusted permitted reception of long-distance (e.g. transatlantic) signals on small aerials, and it was only slowly displaced by the superheterodyne receiver (see following).

### *1.4.2 Self-oscillating receivers: autodyne and homodyne*

In the development of feedback amplifiers and oscillators, some special properties and advantages of oscillating receivers were noticed and developed. Two of these which were closely related were the autodyne and the homodyne.

The *autodyne* arose from the use of the heterodyne principle, described in Section 1.2. The receiver (basically as in Fig. 1.5) was made to generate its local oscillation by causing the amplifier-detector valve also to generate self-oscillations at a frequency differing from that of the incoming wave by a convenient audio frequency. This was, of course, basically a telegraph



receiver. The first inventor appears to be Round in 1913 [27], although Armstrong was close behind and gave a good explanation of the systems [28]. Armstrong noted that the self-oscillation could easily pull-in or synchronise to the incoming wave and that this condition was useless for telegraphic reception. It appears to have been Kendall [29] who realised in 1915 that this synchronised self-oscillating receiver, later called the *homodyne*, could efficiently demodulate a carrier wave envelope-modulated by speech.

The homodyne, unlike the autodyne, was the subject of much interest and development later. In 1923, Hartley [30] gave a mathematical discussion of the subject showing the effect of phase errors in the local oscillation etc. In an article in 1924 Colebrook [31] described the homodyne substantially as Kendall specified it. From about this time onwards, however, the system developed towards one of greater refinement, with the nonlinear oscillation circuit separated from the desirably-linear signal circuit; the name 'synchrodyne' was later coined by the present author. The history of this development from the early 1920s onwards has been separately published [32] and will therefore not be pursued here. The special attraction of the system in the 1950s was higher selectivity coupled with higher audio quality, owing to the filtration being by lowpass filter in the audio section and not by bandpass filter at a relatively high IF.

### *1.4.3 Super-regenerative receiver*

There were other circuits and systems developed which utilised feedback of the regenerative or positive type, and many of these are summarised in Blake's book [33]. The most important was the super-regenerative receiver.

The name was given to the system by Armstrong, who is generally held to be the inventor. It is, however, difficult to see the difference in principle between Armstrong's system [34, 35] and the considerably earlier British one due to Bolitho [36, 37]. A somewhat similar principle was involved in Turner's 'valve relay' [38]. The general idea is to exploit the extremely high gain which is obtained in a feedback valve circuit which is just on the point of oscillating. At this point, the application of the signal, even of almost infinitesimal magnitude, will produce a substantial amplitude of the oscillation, and this amplitude will depend on the signal amplitude. The problem is, of course, that a valve circuit cannot normally be held just at the point of oscillation. It is normal for the amplitude to build up until limited by nonlinear action, at which point the extreme sensitivity to the applied signal has been lost. So the object of all these super-regenerative inventions is to keep the circuit just at the point of oscillation.

This object is achieved by using an auxiliary circuit such that as soon as the oscillation starts to build up it is quenched by a brief alteration of a parameter of the circuit; it then starts to build up again, and so on. If the quenching is caused to occur at intervals which are too close to cause any audible modulation of the output audio signal, but which are sufficiently spaced in

relation to the cycles of the radio signal, then the envelope of the oscillation waveform reproduces the modulation (audio) signal with extremely high amplification and usually acceptable distortion. Turner used an electro-mechanical relay to short-circuit the feedback as the oscillation built up, but this was obviously inadequate for speech. Bolitho and Armstrong used separate valve oscillators to give the periodic quenching, the former by causing the feedback to be largely cancelled as the separate oscillator to the grid of the feedback valve, so that the amplification was cut off on negative half-cycles of the separate oscillator. One of Armstrong's circuits is shown in Fig. 1.7. Here the circuit A is the radio-frequency amplifier and detector, and circuit B is the separate oscillator of lower frequency.

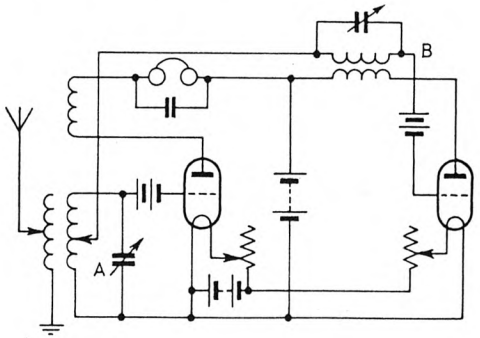


**Fig. 1.7** *Armstrong's super-regenerative receiver*

## 1.5 Superheterodyne receiver

The tuned-radio-frequency (TRF) receiver discussed in Section 1.3 had the disadvantage that its bandwidth in hertz increased in proportion to the frequency of the radio wave; moreover, the high capacitance of early triodes severely limited the use of such amplifiers at the higher radio frequencies (say from a few megahertz upwards). In order to extend the use of receivers into the short-wave region and to improve selectivity even at low frequencies, the idea of adapting the heterodyne principle to give a supersonic difference frequency, which could then be amplified by a TRF receiver of fixed frequency (later called the intermediate frequency or IF) occurred to several people around the end of the First World War. Levy (France) [39], Schottky (Germany [40], and Armstrong (USA) [41–43] are the best known of these, and their circuits were quite similar, basically as shown in Fig. 1.8. Some RF tuning was used to prevent difficulty with image frequencies, and the selection of the frequency to be received was made by suitable tuning of the oscillator A.

At first various IFs were used, and both telephony and telegraphy made use of the system. The system did not at first come into general use because it



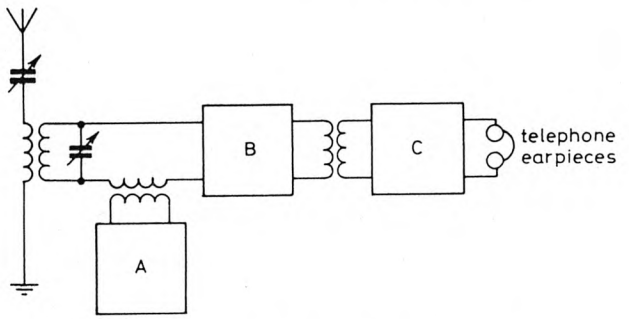


**Fig. 1.8** *Typical arrangement of early superheterodyne receiver*
A Valve oscillator
B Simple valve amplifier-detector
C TRF (fixed frequency) receiver

required a large number of valves. For example, Armstrong used six stages of IF amplification to his trials of 1918–19; after the first tuned stage, *RC* coupling was used. There were also problems of 'tracking' the tuning of the aerial circuit and the oscillator. It did not come into general use for broadcast receivers until the 1930s. Then, for this application, a frequency of 465 kHz was standardised for the IF, and it was usual to have some RF amplification, which had by then become possible even on short waves through the development of tetrode and pentode valve. Design of the IF circuit as a bandpass filter came in then too.

The problem of ganging the RF tuning with the tuning of the local oscillator was not straightforward, as they were at different frequencies; the same frequency shift was required on each, so that tuning required different proportional changes. Basically it was the oscillator frequency which determined which RF channel was selected, but it was necessary that the RF tuning should agree within reasonable limits with the channel thus chosen. The problem was called 'tracking', and the methods of dealing with it were electrical circuit methods rather than mechanical; they are described in the textbooks [44] having originated in the early 1930s [45]. The basis was the correct proportioning of the inductance of the oscillator circuit and of a capacitance in series with the tuned circuit of the oscillator.

As the control of radio receivers became simplified by the abandonment of reaction in favour of multivalve 'superhets' (as superheterodyne receivers quickly became called), so the need for automatic volume control arose. It came into use in the early 1930s. It was essentially a system whereby a direct voltage proportional to the amplitude of the carrier signal at the output of the IF stage was used to control the gain of the amplifying valves, thus greatly reducing the variations in the output of the receiver when the input signal fluctuated owing to such factors as fading. Its operation depended on the 'variable-mu' valve, which was introduced just before 1930; this is a thermionic valve in which the grid has unequal spacing of the wires so that the curve of

anode current against grid voltage is greatly prolonged along the negative grid-voltage axis. By this means the amplification is gradually reduced as the grid bias is made more negative. When the direct control voltage, described above, is fed negatively to the grids of the valves of the receiver, their overall amplification is reduced by a factor much greater than that by which the output has increased. For example, a fluctuation of one-hundredfold in the input amplitude leads to a fluctuation of the audio output of only perhaps two or three times.

The invention of this kind of automatic volume control was almost certainly due to Wheeler in 1932 [46], although the US Supreme Court declared his patent invalid in 1941 [47], awarding priority to a semi-mechanical system of Espenschied and Bown disclosed in 1921 [48].

The superhet, with IF selectivity and automatic volume control, has survived to the present day in substantially the same form, with, of course, the replacement of valves by transistors.

## 1.6 Sidebands

For nearly two decades from the first successful radio experiments of Marconi and others, radio workers seem to have been entirely unaware of the concept of sidebands and of the bandwidth requirements of a modulated signal. So also in the wire transmission field, knowledge of such matters seems to have arisen only with the advent of carrier telephony. As far as the present author has been able to discover, Carson was the first in the electrical communication field to discover and elucidate the concept of sidebands. In a patent [49] of 1915 he gives a full mathematical theory of the production of sidebands of various types in a nonlinear device (i.e. a modulator) and also shows the benefit, not only of carrier suppression, but also of single-sideband operation. This work was later published as a paper [50]. Not far behind Carson was England [51], who also defined sidebands, mentioning carrier suppression and re-insertion at the receiver.

This work led directly to the frequency-division-multiplex era in point-to-point communications, where single-sideband suppressed-carrier operation became standard, both in line and radio communication [52]. An experimental one-way transatlantic radio telephone link was set up on this basis in 1923 on a carrier frequency of 60 kHz [53]. The problems of receiver design, involving the re-introduction of the carrier, were more complex than others we have discussed, and we cannot enter into detail here. At low carrier frequencies it was possible to make the local carrier oscillators at transmitter and receiver sufficiently stable to maintain the re-introduced carrier within a few hertz of the carrier used at the transmitter, and this was sufficient to demodulate the



single sideband without serious distortion. In general it was necessary to transmit a pilot carrier at low level in order, first, to synchronise the carrier oscillator at the receiver, and secondly, to provide a signal on which automatic volume control (in this context usually called automatic gain control) could be based. The design problems were well described by Reeves [54].

In view of much later developments in radio point-to-point communications, it is interesting to note that the concept of group modulation of a series of individually-modulated carriers was developed before 1920, for in that year Ryan, Tolmie and Bach [55] described systems (with some experimental results of trials) which included one with 'mono-radio-frequency modulated at several super-audio frequencies, each of which is itself modulated at audio frequency'.

We have said that sidebands in the electrical communication field were discovered in 1915. The curious thing is that sidebands were well-known to scientists in the field of acoustics as far back as 1875.

It is probable that the first reference to sidebands (although not by that name) is in the paper by Mayer [56] published in 1875. He describes experiments in acoustics in which he modulates or interrupts the sound from a tuning fork by means of a rotating screen with holes in it. As the speed of rotation is increased from zero he notices two additional sounds appear:

> 'On revolving the perforated disk, two additional or secondary sounds appear – one slightly above, the other slightly below the pitch of the fork'.

This is a very clear picture of sidetones.

Lord Rayleigh did some further work on this subject, and set out the theory clearly in the second edition of Volume 1 of his famous book in 1894 [57]. He gives in effect the equation:

$$2(1 + \cos mt)\cos 2\, nt = 2\cos 2\, nt + \cos 2\,(n + m)t + \cos 2\,(n - m)t$$

and explains that this is only a particular case. As an example of a more complex modulation he expands:

$$4\cos^4 mt \cos 2\, nt$$

showing that this has four secondary sounds.

Rayleigh also gives a very satisfying physical explanation of the secondary sounds. He describes first of all an experiment in which a tuning fork of frequency 128 is driven by a current which is interrupted at frequency 128 by a fork-driven interrupter. This current can also be interrupted by another independent interrupter of frequency 4. When the second interrupter was inoperative, the fork had a strong response in its normal tuning of 128, but scarcely any when tuned to 124 or 132. When the second interrupter was working, however, the fork would respond powerfully when tuned to 124 or 132 as well as when tuned to 128, but not when tuned to intermediate pitches, such as 126 or 130.

The physical explanation which Rayleigh gives is:

'When a fork of frequency 124 starts from rest under the influence of a force of frequency 128, the impulses cooperate at first, but after ⅛ of a second the new impulses begin to oppose the earlier ones. After ¼ of a second another series of impulses begins whose effect agrees with that of the first, and so on. Thus if all these are allowed to act, the resultant effect is trifling; but if every alternative series is stopped off, a large vibration accumulates'.

This is a very helpful way of looking at sidetones.

## 1.7 Frequency modulation

The original invention of frequency modulation appears to be that of Ehret in 1902 [58], fully discussed in an earlier paper by the present author [59]. His claim is a very concise definition of FM:

'The method of transmitting intelligence, which consists in generating electroradiant energy, modifying the frequency of said energy in accordance with the signal to be sent and receiving the energy in a device responsive to changes in the frequency of the transmitted energy'.

He envisaged modulation by speech, for his description of his invention included the following:

'It comprises, further, a method of modifying and varying the frequency of the electroradiant energy in a manner corresponding and in accordance with the signal to be transmitted.

It resides also in an additional step of modifying the energy to be transmitted and received by and in accordance with sound-waves, such as speech.

It comprises, further, a method of receiving the modified transmitted energy and causing the reproduction of speech and other signals by the effects of variations or changes in the frequency of the received energy'.

The process of modulation was to modify the frequency of the second stage of a two-stage spark transmitter, e.g. by connecting the microphone across a tuning inductance. The receiver used the variation of voltage across an inductance connected in series with the aerial or across a shunt-tuned circuit coupled to it. It is perhaps unlikely that the system could have worked well with Ehret's proposed circuits, but the idea of the system is perfectly clear and correct. The patent also very clearly covers frequency-shift telegraphy.

The idea of frequency modulation as a means of *reducing* the bandwidth required for the transmission of a speech signal by radio was prevalent after the discovery of sidebands in radio (see Section 1.6), but in his paper of 1922, Carson [60] effectively disposed of this idea and showed that FM produced a



series of sidebands and that the minimum possible bandwidth was the same as for AM. It was Armstrong's pioneering experiments from 1933 onwards, with his paper [61] of 1936, which finally showed the real properties and value of FM.

Armstrong showed that the basis of reception of FM signals was to tune and amplify as in a superheterodyne, then to subject the signal to an amplitude limiter, then to apply it to a frequency discriminator which would convert the FM to AM and so by rectification to audio. The amplitude limiter was to remove any amplitude variations which were, of course, irrelevant to the message but which would otherwise appear in the output. It could comprise a grid-limited triode (or other) valve, or the rectifier-type of limiter that was favoured later. A discriminator used by Armstrong is shown in Fig. 1.9, and its method of operation can be explained by reference to Fig. 1.10. The carrier frequency is $f_0$, and the modulation varies the instantaneous frequency within

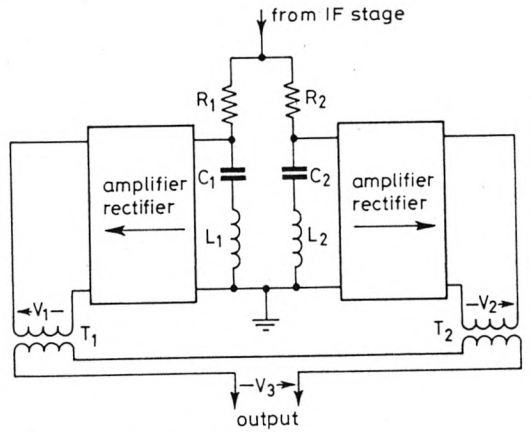


**Fig. 1.9** *Early Armstrong discriminator*

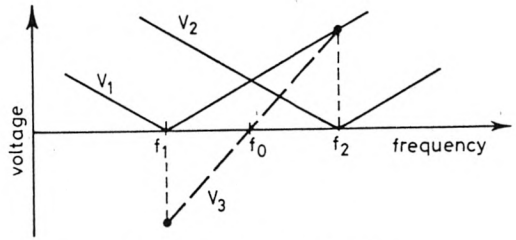


**Fig. 1.10** *To show operation of Armstrong discriminator*

the limits $f_1$ and $f_2$. The resistances $R_1$ and $R_2$ are of high value so that the tuned circuits are effectively operated on a constant-current basis. $L_1C_1$ resonates at $f_1$ and $L_2C_2$ at $f_2$. After amplification and rectification, therefore, the voltages $V_1$ and $V_2$ vary with frequency as shown in Fig. 1.10. The windings of the audio transformers $T_1$ and $T_2$ are so connected that $V_1$ and $V_2$ are opposed in the

output circuit, giving a variation of $V_3$ with frequency as shown also in Fig. 1.10. Thus the audio modulation of the transmitted FM signal is correctly reproduced as audio. Of course, the graphs in Fig. 1.10 have been simplified, following Armstrong's own presentation; the graph of $V_3$ between frequencies $f_1$ and $f_2$ would not in practice be quite a straight line.

A later and simpler form of discriminator (one of many which were developed) is shown in Fig. 1.11 and its operation in Fig. 1.12. $L_0C_0$ is tuned to $f_0$, $L_1C_1$ to $f_1$, and $L_2C_2$ to $f_2$. Although the graphs of $V_1$ and $V_2$ against frequency are curved, they do add up to an approximately linear variation.

A point which Armstrong appears to have at first failed to appreciate is that the minimum bandwidth required in the system is not just twice the maximum frequency deviation (i.e. $f_2 - f_1$), but needs to be this *plus* twice the maximum modulating frequency.

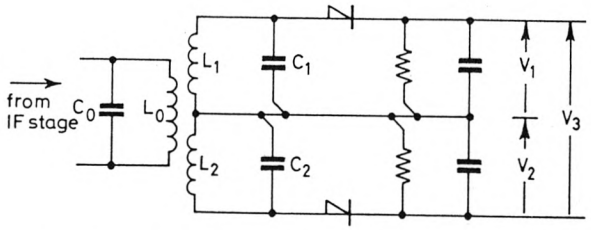


**Fig. 1.11** *Later form of double-circuit discriminator*

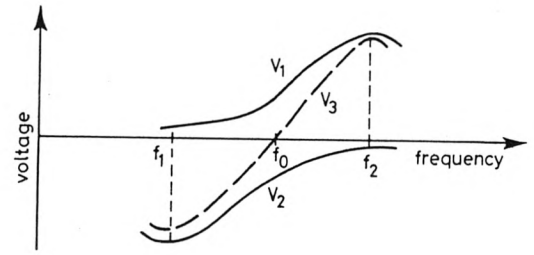


**Fig. 1.12** *To show operation of discriminator of Fig. 1.11*

The most important feature of frequency modulation discovered by Armstrong was the favourable effect of using a wide bandwidth for the signal. In contrast to the normally accepted view that bandwidths should be the minimum required to transmit a signal in order to minimise the noise admitted to the receiver, in an FM system there is a signal-to-noise advantage in making the system utilise the widest band possible. This is effected by having a very high modulation index. To show in a simple way how this works, following to some extent Armstrong's own reasoning, let us consider two systems, one in which the maximum frequency deviation is 10 kHz and the other in which it is 100 kHz. Assume that the highest audio frequency in the modulating signal, or



passed by the output circuit (or audible to the ear), is 10 kHz. The required bandwidth of the first system is about 40 kHz; of the second system about 220 kHz. Fig. 1.13 shows the discriminator responses in the two cases.

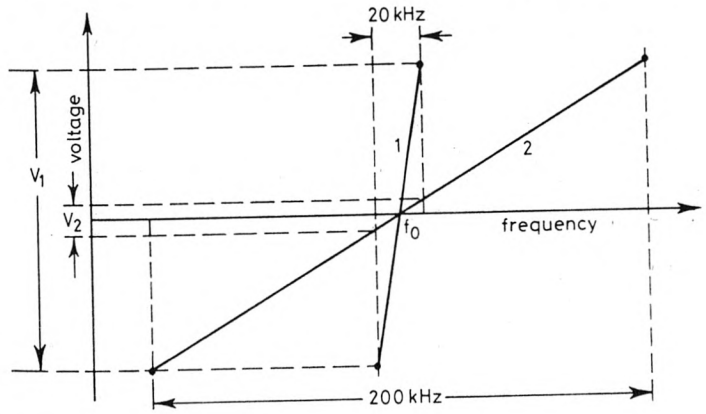


**Fig. 1.13** *To show effect of bandwidth on FM receiver*

Assume that any interfering signal has an amplitude $x$ relative to the wanted signal, where $x \ll 1$. Assume also, for simplicity, that the wanted signal is, at the time, unmodulated. Then the limiter effectively converts the interfering signal into a phase modulation of the wanted wave, which becomes proportional to $\cos(\omega_o t - x \sin \omega_i t)$, where $\omega_o$ is the angular frequency of the wanted signal and $\omega_i$ is the difference between $\omega_o$ and that of the interference. The instantaneous frequency is therefore $\omega_o - x\omega_i \cos \omega_i t$, which has a maximum frequency deviation of $x\omega_i$ at a modulating frequency of $\omega_i$.

It is important to note that the effect of the interference is proportional not only to its amplitude $x$, but also the frequency difference $\omega_i$. Consider for clarity the exaggerated case of $x\omega_i/2\pi = 10$ kHz, still with the assumption that the signal modulation is, at the time, zero. Then curve 1 in Fig. 1.13 shows that in the 10 kHz system the voltage swing in the output is $V_1$. In the 100 kHz system curve 2 applies and the voltage swing is $V_2$, which is only one-tenth of $V_1$. There is thus a prima-facie conclusion that the effect of interference is reduced to one-tenth in amplitude in the system which uses ten times the frequency deviation and more than five times the overall bandwidth. The use of the wider bandwidth does not increase the interferences by introducing more noise power or more interfering signals, because difference frequencies $(\omega_i/2\pi)$ above 10 kHz have been assumed to be inaudible anyway.

Since, as we have said, the interfering output is proportional to the difference frequency $(\omega_i/2\pi)$, it is clear that for the lower audio frequencies the disturbance is much lower, and this feature in suitable circumstances gives FM an advantage over AM even apart from the other effects described.

A fuller explanation of the signal-to-noise advantages of FM, occupying some 16 pages, was given by Sturley [62]. The argument given above is, of course, highly simplified and must not be taken as more than an indication of the type of mechanism involved. A mathematical analysis, still with the simplifying assumptions used here, is given in textbooks [63].

## 1.8 Conclusion

It is hoped that the origins, and practical development into the 1930s, of all the principal processes involved in radio receivers have been covered, or at least indicated, in the foregoing sections. There was, of course, a good deal of theoretical work being done even during these early years and it has not been possible to cover most of this. Then in the nearly 50 years since the period covered there have been innumerable important developments, including that of the transistor and of microelectronics and the extension of the frequency spectrum up to microwaves, which have drastically changed the physical form of radio receivers. Nevertheless, the principles established by the mid-1930s are still remarkably persistent.

It is evident from the foregoing historical account that, while there were numerous important inventors and innovators in the field, none made a more outstanding contribution than the American, Armstrong (1890–1954), who was a significant figure from his undergraduate days in 1913. Our account has necessarily been concerned with the technical aspects only, but it is worth remembering that the inventors had to deal with the business and legal aspects also, and these generated much bitterness. Fleming, de Forest and Armstrong all suffered greatly from this, perhaps none more than Armstrong, who finally took his own life in despair. This aspect of Armstrong's life has been well described in a recent article [64].

## References


1 LODGE, O.: 'The work of Hertz and some of his successors', Lecture delivered at the Royal Institution on 1 June 1894, 'The Electrician' Printing and Publishing Co Ltd, London 1894
2 MARSH, J. O., and ROBERTS, R. G.: 'David Edward Hughes: inventor, engineer and scientist', *Proc. Inst. Elect. Engrs.*, 1979, **126,** pp. 929–935
3 PHILLIPS, V. J.: 'Early radio wave detectors' (Peter Peregrinus, London, 1980)
4 *Ibid.*
5 PHILLIPS, V. J.: 'Early radio wave detectors' (Peter Peregrinus, 1980), P. 70
6 AITKEN, H. G. J.: 'Syntony and spark', p. 106 and Note 44 on p. 127
7 FLEMING, J. A.: British Patent No. 24,850, 1904
8 DUNWOODY, H. H. C.: British Patent No. 5332, 23 March 1906; PICKARD, G. W.: US Patent No 836,531 filed August 1906 and subsequent patents
9 AITKEN, H. G. J.: 'Syntony and spark', Chapter 4
10 BLANCHARD, J.: 'The history of electrical resonance', *Bell Syst., Tech. J.*, 1941, **20,** pp. 415–433





11 FESSENDEN, R. A.: US Patent No. 706,740, 1902
12 HOGAN, J. L.: 'The heterodyne receiving system, and notes on the recent Arlington-Salem tests', *Proc. I.R.E.*., 1913, **1**, pp. 75–102
13 LEBLANC, M.: 'Etude sur le telephone multiplex', *La Lumiere Electrique*, 1886, **20,** pp. 97–102
14 FESSENDEN, R. A.: US Patent No. 753,863, September 1901
15 RUHMER, E.: 'Wireless telephony in theory and practice' (published in Germany 1907; English translation by ERSKINE-MURRAY, J., Crosby, Lockwood & Co., London, 1908)
16 FLEMING, J. A.: British Patent No. 24,850, 1904
17 de FOREST, L.: 'The audion', *Trans. Amer. Inst. Elect. Engrs.*, 1906, **25,** p. 735
18 LOWENSTEIN, F.: US Patent No 1,231,764, filed April 1912, issued July 1917
19 ARMSTRONG, E. H.: 'Operating features of the audion', *Electrical World*, 1914, **64,** pp. 1149–52
20 LANGMUIR, I.: US Patent No 1,282,439, October 1918
21 PALMER, L. S.: 'Wireless principles and practice', (Longmans Green, London, 1928), pp. 362–374
22 HAZELTINE, L. A.: 'Tuned radio-frequency amplification with neutralization of capacity coupling', *Proc. Radio Club Amer.*, 1923, **2,** pp. 3–8
23 HARRISON, A. P.: 'Single-control tuning: an analysis of an innovation, *Technology & Culture*, 1979, **20,** pp. 296–321
24 APPLETON, E. V. and BOOHARIWALLA, D.: 'The mutual interference of wireless signals in simultaneous detection', *Wireless Engr. & Exptl. Wireless*, 1932, **9,** p. 136
25 TUCKER, D. G.: The history of positive feedback', *Radio and Electronic Engr.*, 1972, **42,** pp. 69–80
26 ARMSTRONG, E. H.: 'Some recent developments in the audion receiver', *Proc. Inst. Radio. Engrs.*, 1915, **3,** pp. 215–247
27 ROUND, H. J.: British Patent No. 28,413, December 1913
28 As Reference 26, pp. 224–6
29 KENDALL, B. W.: US Patent No. 1,330,471, November 1915
30 HARTLEY, R. V. L.: 'Relations of carrier and side-bands in radio transmission', *Proc. Inst. Radio. Engrs.*, 1923, **11,** p. 34
31 COLEBROOK, F. M.: 'Homodyne', *Wireless World & Radio Rev.*, 1924, **13.** p. 645
32 TUCKER, D. G.: 'The history of the homodyne and synchrodyne', *J. Brit. Inst. Radio Engrs.*, 1954, **14,** p. 143
33 BLAKE, G. G.: 'History of radio telegraphy and telephony' (Radio Press, London, 1926)
34 ARMSTRONG, E. H.: US Patent No. 1,424,065, June 1921
35 ARMSTRONG, E. H.: 'Some recent developments of regenerative circuits', *Proc. Inst. Radio Engrs.*, 1922, **10,** pp. 244–266
36 BOLITHO, J. B.: British Patent No. 156,330, October 1919
37 'The Bolitho circuit', *Wireless World & Radio Rev.*, 1923, **12,** p. 266
38 TURNER, L. B.: British Patent No. 130,408, February 1918
39 LEVY, L.: British Patents No. 143,583, August 1917; 133,306, October 1918; and 150,352, August 1919
40 Original reference not found, but fact stated by DALTON, W. M. M. 'The story of radio' (Adam Hilger, Bristol, 1975), Vol. 2, p. 109
41 ARMSTRONG, E. H.: US Patent No. 1,342,885, 1920
42 ARMSTRONG, E. H.: 'A new system of short wave amplification', *Proc. Inst. Radio Engrs*, 1921, **9,** pp. 3–27 (presented as lecture in New York 3rd December 1919)
43 ARMSTRONG, E. H.: 'The superheterodyne – its origin, development, and some recent improvements', *Proc. Inst. Radio Engrs.*, 1924, **12,** pp. 539–552
44 LOVERING, W. F.: 'Radio communication' (Longmans Green, London, 1958), pp. 372–7
45 LANDON, V. D., and SVEEN, E. A.: 'A solution of the superheterodyne tracking problem', *Electronics*, Aug. 1932, pp. 250–1

46 WHEELER, H. A.: US Patent No 1,879,863, 1932
47 MACLAURIN, W. R.: 'Invention and innovation in the radio industry' (Macmillan, New York, 1949), p. 182
48 ESPENSCHIED, L., and BOWN, R.: US Patent No 1,447,773, September 1921
49 CARSON, J. R.: US Patent No. 1,449,382, December 1915 (also British Patent No. 102,503 issued November 1917)
50 CARSON, J. R.: 'A Theoretical study of the three-element vacuum tube', *Proc. Inst. Radio Engrs.*, 1919, **7,** pp. 187–200
51 ENGLUND, C. R.: US Patent No. 1,245,446, March 1916
52 ESPENSCHIED, L.: 'Application to radio of wire transmission engineering', *Proc. Inst. Radio Engrs.*, 1922, **10,** pp. 344–368
53 NICHOLLS, H. W.: 'Transoceanic wireless telephony', *J. Inst. Elect. Engrs.*, 1922–3, **61,** pp. 812–22
54 REEVES, A. H.: 'The single side-band system applied to short-wave telephone links', *J. Inst. Elect. Engrs.*, 1933, **73,** pp. 245–279
55 RYAN, F. M., TOLMIE, J. R. and BACH, R. O.: 'Multiplex radio telegraphy and telephony', *Proc. Inst. Radio Engrs.*, 1920, **8,** pp. 451–467
56 MAYER, A. M.: 'Researches in acoustics, Part 6', *Phil. Mag.*, 1875, **49,** pp. 352–65
57 Lord RAYLEIGH: 'The theory of sound' – Vol. 1' (Macmillan, London, 1894), 2nd edn
58 EHRET, C. D.: US Patents Nos. 785,805 and 785,804, February 1902
59 TUCKER, D. G.: 'The invention of frequency modulation in 1902', *Radio and Elect. Engr.*, 1970, **40,** pp. 33–37
60 CARSON, J. R.: 'Notes on the theory of modulation', *Proc. Inst. Radio Engrs.*, 1922, **10,** pp. 57–64
61 ARMSTRONG, E. H.: 'A method of reducing disturbances in radio signalling by a system of frequency modulation', *Proc. Inst. Radio Engrs.*, 1936, **24,** pp. 689–740
62 STURLEY, K. R.: 'Frequency modulation' (BBC, London, 1955)
63 TAUB, H., and SCHILLING, D. L.: 'Principles of communication systems' (McGraw-Hill, 1971), pp. 295–304
64 HOPE, A.: 'The battles of Armstrong – radio's forgotten man', *New Scientist*, 1st Feb. 1979. pp. 306–9