Direct - conversion FM design

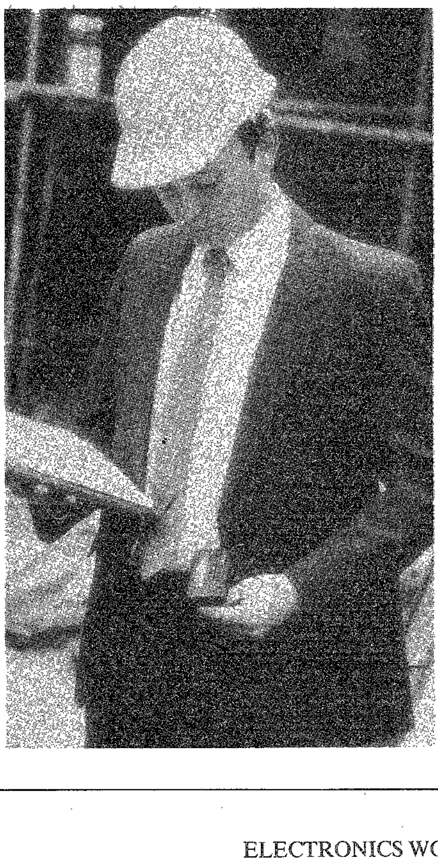
Direct-conversion detection is as old as the hills. Digital electronics has given the technique a new lease of life. By Ian Hickman.

Homodyne or direct-conversion reception has always attracted a good deal of attention, especially in amateur circles[1]. It has the attraction of simplicity, both in principle and in hardware terms. The first figure shows a simple homodyne receiver which could in principle be simplified even further by the omission of the RF amplifier (at the expense of a poorer noise figure) and even of the input tuned circuit or band-pass filter – some filtering might be provided by the aerial, if for example it were a half-wave dipole.

The homodyne has something in common with the superhet, but whereas the latter produces a supersonic intermediate frequency (hence SUPERsonic HETerodyne receiver), in the homodyne the local-oscillator frequency is the same as the signal's carrier frequency, giving an IF of 0Hz.

It is well known that a homodyne receiver can be used for reception of SSB, although in a simple homodyne there is no protection against signals in the unwanted sideband on the other side of the carrier. A small offset between the frequency of the local carrier and that of the SSB signal can, however, be tolerated, at least on speech signals.

Homodyne reception can also be used for the reception of AM, but no frequency offset is permissible and the phase of the local carrier must be identical to that of the incoming carrier, otherwise all the modulation "washes out". This means in practice that the local oscillator must be phase-locked to the carrier of the incoming signal. If the local oscillator circuit is undercoupled so that it barely oscillates, if at all, the incoming signal energy can readily synchronise it, an arrangement universally employed under the name of "reaction" in the days of battery-powered "straight" wireless sets using 2V directly heated valves.



FSK and the homodyne

CW is readily received by a homodyne receiver, but it is not immediately obvious how it could be successfully employed for FM reception. However, it can, as will shortly appear.

The simplicity of the homodyne means that it is potentially a very economical system of reception and, for this reason, there has always been an active interest in the subject on the part of commercial concerns; a number of homodyne receivers has appeared on the market[2]. This paging receiver is actually a data receiver using FSK modulation, which is a type of FM where the information is conveyed by changing the signal frequency rather than its amplitude. One could in principle receive the signal by tuning the local carrier just below (or above) the two tones and picking them out with two appropriate audio frequency filters, but this would be a very poor solution, since there would be no protection from unwanted signals on the other side of the carrier.

The solution adopted was much more elegant, with the local oscillator tuned midway between the two tones, so that both ended up at the same audio frequency, equal to half the separation of the two tones at RF. Now in a simple homodyne receiver, this would simply render the two tones indistinguishable; in a practical system it is necessary to have some way of sorting them out. This is entirely feasible, but it does involve just a little more kit than in a basic Fig. 1 type simple homodyne receiver. Before looking at how it is done, some basic theory is needed, which I have chosen to illus-

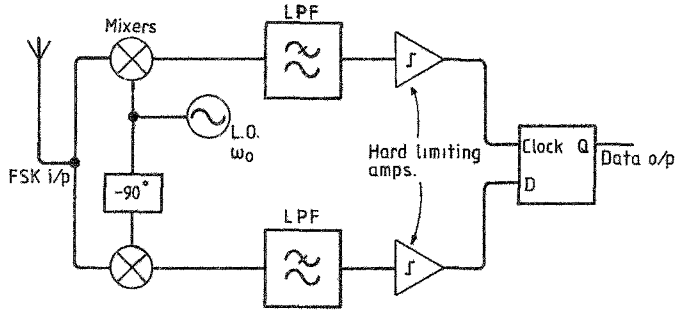


Fig. 1. Principle of the homodyne, in which the received signal is converted directly to audio by setting the local-oscillator frequency equal to that of the signal.

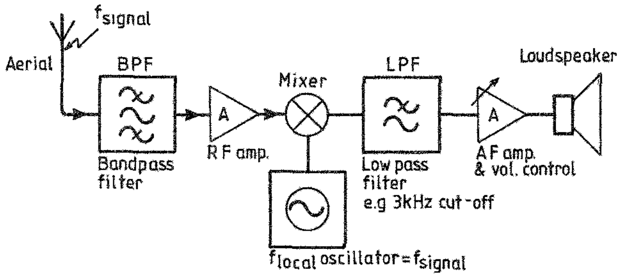


Fig. 3. Homodyne receiver for frequency-shift keyed transmission.

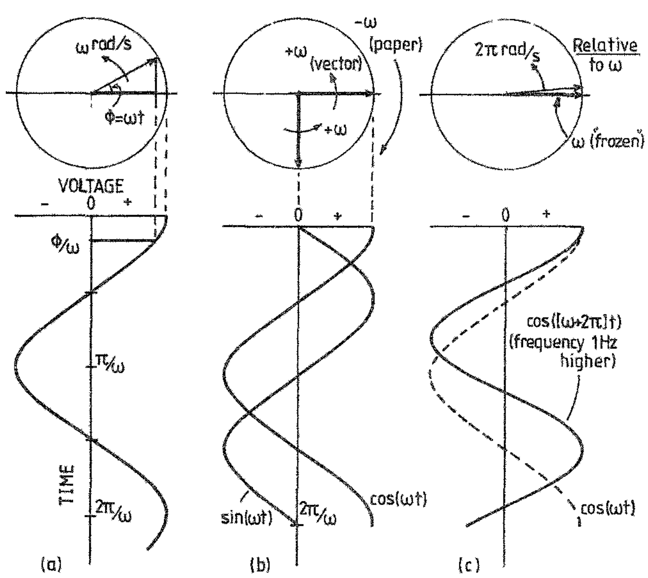


Fig. 2. Vector representation of sinusoidal waveforms. Waveform at (a) is derived from projection of rotating vector onto horizontal axis, while at (b) two such vectors are shown in quadrature, producing sine and cosine waves. Slightly different frequencies produce the effect seen at (c).

trate graphically rather than with algebra and trigonometry, though the results are of course the same.

Figure 2(a) shows a sinusoidal waveform and illustrates how its instantaneous value is equal to the projection onto the horizontal axis of a vector of fixed length, rotating (by convention) anticlockwise. Figure 2(b) carries the idea a little further and shows two such vectors, representing a sinewave and a cosine wave. As in (a), both vectors should be imagined as rotating at an angular speed of ω rad/s, that is ($\omega/2\pi$) Hz. If they really were, there would be a blur at anything above a few tens of Hertz, so further imagine the paper they are drawn on to be rotating in the opposite direction, i.e. clockwise, at ω rad/s, thus freezing the motion and enabling us to see what is going on.

Furthermore, with this convention one can picture what happens when a slightly different-frequency sinewave is also present, say at a frequency of ($\omega + 2\pi$) rad/s or 1Hz higher. This can be represented on the vector diagram as a vector rotating anticlockwise at a velocity of 2π radians, or one complete revolution, per second relative to the frozen ω vector, Figure 2(c). Had the frequency of the second sinewave been ($\omega - 2\pi$) rad/s, that is to say 1Hz lower than ω, then its relative rotation would have been clockwise.

The method used by the paging receiver mentioned earlier to distinguish between the equal-frequency baseband tones produced when the homodyne receiver is tuned midway between the two radio frequencies is shown in **Fig. 3**, a block diagram of the receiver. Incoming signal is applied to two mixers, each supplied with a local oscillator drive at a frequency of f_o, but the drive to one mixer is phase shifted by 90° relative to the other.

Referring back to Fig.2(c), a vector rotating anticlockwise at $f_{sh}/2$ relative to f_o (where f_{sh} is the frequency shift between the two FSK tones) will come into phase with the sine component of the local oscillator, $\sin(\omega_o t)$, a quarter of a cycle before coming into phase with $\cos(\omega_o t)$. On the other hand, when the incoming signal is $f_{sh}/2$ lower in frequency than f_o, then the clockwise rotation of the vector in Fig. 2(c) indicates that it will come into phase with $\cos(\omega_o t)$ a quarter of a cycle before $\sin(\omega_o t)$.

Now relative phases are preserved through a frequency changer or mixer, so the audio signal in the Q channel will be in quadrature with that in the I channel. Furthermore, one channel will lead the other or vice versa, according as the incoming RF tone is above or below f_o. The two audio paths include filters to suppress frequencies much above $f_{sh}/2$ (these filters must be reasonably well phase-matched, obviously) after which the signals are amplified and then turned into square waves by comparators. As the square waves are in quadrature, the edges of the I channel waveform occur midway between those in the Q channel, so the D input of the flip-flop will be either positive or negative when the clock edge occurs, depending upon whether the RF tone is currently higher or lower in frequency than f_o, i.e. whether the signal represents a logical 1 or a 0.

The frequencies of the two RF tones are $f_o+f_{sh}/2$ and $f_o-f_{sh}/2$ and the resultant frequencies out of the mixers are the difference frequencies between these radio frequencies and the local oscillator frequency, or $(f_o+f_{sh}/2)-f_o$ and $(f_o-f_{sh}/2)-f_o$. The first of these audio tones is at a frequency of $f_{sh}/2$, while the second is at $-f_{sh}/2$ and of course by the very nature of an FSK signal, only one is present at any instant. Played through a loudspeaker, they would sound indistinguishable –as indeed they are in themselves. It is only by deriving two versions of, say, $+f_{sh}/2$ using quadrature related local oscillators and comparing them that it can be distinguished from $-f_{sh}/2$.

The ability of the receiver to distinguish between two audio tones of identical frequency, one positive and one negative, indicates that negative frequencies are "for real", in the sense that a negative frequency has a demonstrable significance different from that of its positive counterpart. This can only be observed, however, if both the P and Q (in-phase and quadrature versions) are available; the signal is then said to be a "complex" signal. A complex signal cannot be conveyed on a single wire, unlike an ordinary or "real" signal.

FM reception

In the case of more general FM signals, including analogue voice, more extensive processing of the baseband (i.e. zero-frequency IF) signals is required. Whilst this could, in principle, be carried out in analogue circuitry, it is often nowadays performed with digital signal-processing (DSP). The great attraction here is that one set of digital hardware can provide any required bandwidth and any type of demodulation (rather than having separate hardware filters and detectors for AM, FM, PM etc.) in, say, a professional or military communications or surveillance receiver (at present the arrangement would be unnecessarily expensive in a broadcast FM set). The signals must first be digitised, which in the present state of the art cannot be done economically at RF with enough bits to provide sufficient resolution. A superhet front end translates the signal to a low IF. There it can can be conveniently digitised directly, or alternatively it is translated to zero Hz and then digitised.

There are several examples of receivers using this approach. The STC model STR 8212 is a general-coverage HF receiver with a DSP back end which includes FM in its operating modes. In such a receiver, a non-standard IF bandwidth is easily implemented, requiring only a different filter algorithm in prom, rather than a special design of crystal filter, with the associated time and cost penalties; a rather similar receiver is available from one of the large American communications manufacturers.

Another implementation of a high-performance HF-band receiver with a zero-frequency final IF is described in ref.3. (This did not list FM as one of its modes, but discussion with the authors afterwards confirmed that this mode is indeed included.) At the same venue, a paper[4] from Siemens Plessey Defence Systems described their PV3800 range of broadband ESM receivers covering 0.5 —1000MHz. These include an FM demodulation mode and use a DSP back end; from the brief details given it would seem likely that again a zero-frequency IF is used.

To understand the reception of conventional analogue FM signals by a homodyne receiver, it is time to introduce the general expression for a narrow-band signal centred about a frequency ω_o; this is

$$V(t) = P(t)\cos\omega_o t - Q(t)\sin\omega_o t \quad (1)$$

where P(t) and Q(t) are called the in-phase and quadrature components. It is important to realise that equation (1) is only useful to describe narrow-band systems, such as could pass through a band-pass filter with a bandwidth of not more than a few percent of the centre frequency; for very wide band system it would become mathematically untractable. So bear in mind that the functions of time P(t) and Q(t) are relatively slowly varying functions, that is to say a very large number of cycles of the carrier frequency ω_o will have elapsed by the time there has been a significant change in the values of P(t) and Q(t).

With this proviso, equation (1) can, with suitable values of P and Q, represent any sort of steady state signal, including FM. I am using this expression, following the development in ref. 5, rather than the possibly more usual approach (see box) followed by other writers, e.g. in ref. 6, because it seems to fit in better with the explanation which follows.

Now FSK is a very specific and rather unrepresentative type of frequency modulation, resulting when a discrete waveform representing a digital data stream is used to modulate the frequency of a transmitter, but I introduced it first for the sole purpose of clearing up the question of the existence of negative frequencies. In the more general case, an FM signal results when a continuous waveform representing a voltage varying with time, for example speech or music, is used to modulate the frequency of a transmitter. The resultant RF spectrum is in general very complex, even for modulation with a single sinusoidal tone, unless the m, the "modulation index", is small.

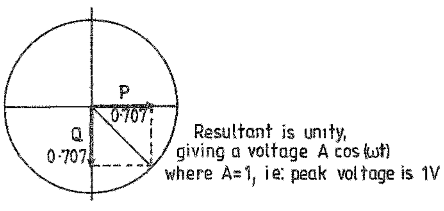

P
0·707
Q
0·707
Resultant is unity,
giving a voltage A cos (ωt)
where A=1, ie: peak voltage is 1V

Fig. 4. In-phase and quadrature components. If P^2+Q^2 is constant, the wave is of constant amplitude.

This is defined as the peak frequency deviation of the frequency-modulated wave above or below the centre frequency (the unmodulated carrier frequency), divided by the modulating frequency. Thus, if the amplitude of a 1kHz modulating frequency at the input of the transmitter were adjusted for a peak frequency deviation of ±2kHz, then m=2. It is fairly easy to show that, in the case of modulation by a single sinusoidal tone, the peak phase deviation from the phase of the unmodulated carrier is simply equal to m radians. For any modulating waveform, there will be a peak frequency deviation and a corresponding peak phase deviation, but the term modulation index is only really meaningful when talking about a single sinusoidal modulating tone.

Before pursuing the niceties of the FM signal, however, I must explain the significance of P(t) and Q(t). If P is a constant (say unity) and Q is zero or vice versa, the result is a unit-amplitude cosine or sine waveform of angular frequency ω_o (the centre frequency), the only difference being that one is at its positive peak voltage, the other at zero but increasing, at the instant t=0, respectively.

Looking at the effect of other values for the constants, if P=Q=0.707 (I have written just P rather than P(t) here, since P(t) indicates a function of time, i.e. a variable, whereas just at the moment I am considering constants) then, as **Fig. 4** shows, the phase of the sinusoidal waveform is 45° at t=0 and its amplitude (courtesy of Pythagoras) is unity. Note that the phase at t=0 (or at any other time, relative to an undisturbed carrier wave cos ω_t) is given by $\tan^{-1}(Q/P)$ and the amplitude by $(P^2+Q^2)^{1/2}$.

If one insists that even if P and Q are allowed to vary, i.e. are functions of time, they shall always vary in such a way that at every instant (P^2+Q^2) is constant, then there will be a wave of constant amplitude. In this case, since the amplitude modulation index is zero, any information the signal carries is due to its variation of frequency and it can be described by the values of P and Q.

To start with a very simple example, suppose $P(t) = \cos\omega_d$ and $Q(t)=\sin\omega_d$, where $\omega_d=2\pi$ rad/s (say). Since $\cos^2x+\sin^2x=1$ for all possible values of x (including therefore ω_d), the result is a constant amplitude signal. Further, its phase relative to ω_o is $\tan^{-1}(Q(t)/P(t))=\tan^{-1}(\tan\omega_d t)$ or ω_d $_t$imes t. In other words, since the phase of the signal is advancing by $\omega_d=2\pi$ rad/s relative to ω_o, the signal frequency is 1Hz higher than ω_o – a (constant) deviation of +1Hz from the centre frequency. Now if ω_d had been -2π rad/s, then the deviation would have been −1Hz, since cos(-x)=cosx, whereas sin(-x)=-sinx. Thus the deviation is simply the rate of change of phase of the modulated signal with respect to the unmodulated carrier.

If now ω_d itself varies sinusoidally at an audio frequency ω_a, then the result is a frequency modulated wave. But if, like me, you start to get confused as the algebraic symbols go on piling up, take heart; some waveforms are coming in just a moment. However, there is one more expression to look at first, since it forms the basis of the particular form of FM demodulation to be examined.

In FM, the transmitted information is contained in the deviation of the instantaneous frequency from the unmodulated carrier – indeed, the deviation *is* the transmitted information. But the deviation is simply the rate of change of the phase angle of the signal relative to the unmodulated carrier; this phase angle is equal to $\tan^{-1}(Q(t)/P(t))$, or ϕ, say. So the instantaneous frequency of the signal ω_i is

$$\omega_\iota=\omega_o+d\phi/dt.$$

Now ω_o is a constant and so conveys no information: to demodulate the signal evaluate $d\phi/dt$, that is $d\{\tan^{-1}(Q(t)/P(t)]\}/dt$. After a few lines of algebraic manipulation (which aren't given in Ref. 3, but which I have checked out and can vouch for) this turns out to be

$$d\phi/dt=$$

$$\frac{P(t).dQ(t)/dt\ -\ Q(t).dP(t)/dt}{P^2(t)+Q^2(t)} \qquad (3)$$

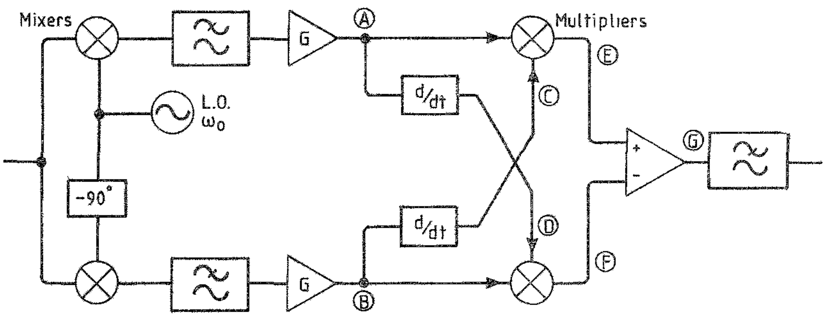


Fig. 5. Sine/cosine demodulator, which produces the numerator of equation (3) at G.

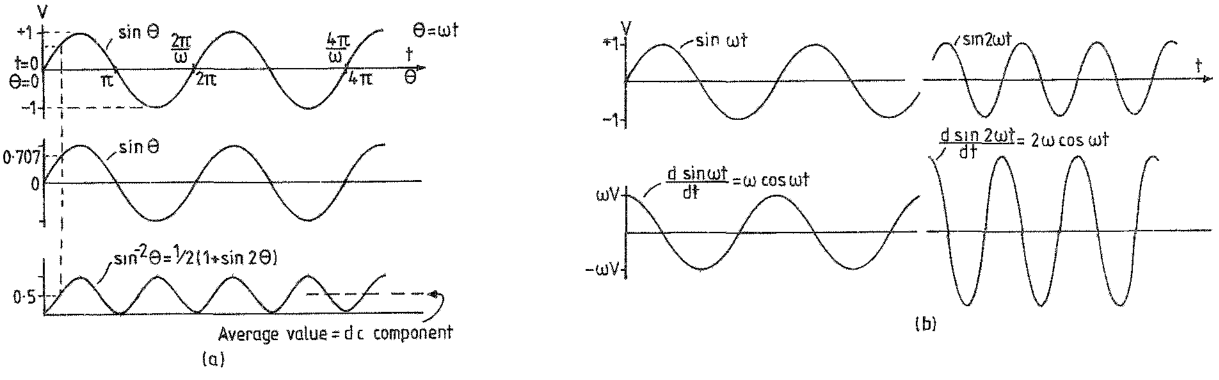
Now as seen earlier, if $P^2(t)+Q^2(t)$ is constant, the result is a constant-envelope wave. For an FM signal, this condition is fulfilled (ignoring fading, for the moment) so, to recover the modulation, a circuit which implements the numerator of the right hand side of equation (3) is needed.

Such a circuit is shown, in block diagram form, in **Fig. 5**. Taking it in easy stages, start with **Fig.6(a)**, which recaps on the basic trig. identity $\sin^2\phi=1/2(1+\sin2\phi)$, as can be seen by multiplying $\sin\phi$ by itself, point by point. Figure 6(b) recalls how $d(\sin a\omega t)/dt=a\cos a\omega t$, i.e. when you differentiate a sinewave, it suffers a 90° phase advance and the amplitude of the resultant is proportional to the frequency of the original.

In Fig. 5, assume that P(t) is fixed at $+2000\pi$ rad/s, and Q likewise. There is thus a fixed frequency offset of 1kHz (2000π rad/s) above the carrier frequency ω_o. In Fig. 5, the frequency of the incoming signal is first changed from being centred on ω_o to being centred on zero by mixing it with a local oscillator signal which is also at ω_o. The two quadrature-related versions of the LO give the in-phase and quadrature baseband versions, P and Q, of the incoming signal. In the upper branch of Fig. 5, the P or in-phase (cosine) component of the signal (now at the original deviation frequency of +1kHz) is multiplied by a differentiated version of the Q or quadrature component. Since these are in phase with each other, the result is a waveform at twice the frequency, and with an offset equal to half its peak-to-peak value, i.e. always positive, as in Fig. 6(a).

Figure 7 shows this and also the waveforms corresponding to the lower branch of the Fig. 5 circuit. Here, the resultant waveform is again at twice the frequency, but in this case, always negative, since $d(\cos\omega_d t)/dt=-\omega_d\sin\omega_d t$. Finally, subtracting Q(t).dP(t)/dt from P(t).dQ(t)/dt, as in Fig. 7, gives a pure DC level. All traces of waveforms at $2\omega_d$ wash out entirely, since when Q(t).dP(t)/dt is zero P(t).dQ(t)/dt is at its maximum and vice versa – provided that the two LO components are exactly in quadrature, that there are no differences in the phase responses of the upper and lower channel of the Fig. 5 circuit and that their gains also match.

Fig. 6. Effects of squaring and differentiating sine waves. Squaring the wave, as at (a), doubles its frequency and produces a DC component. Differentiating, shown at (b), gives a cosine wave with an amplitude proportional to frequency.



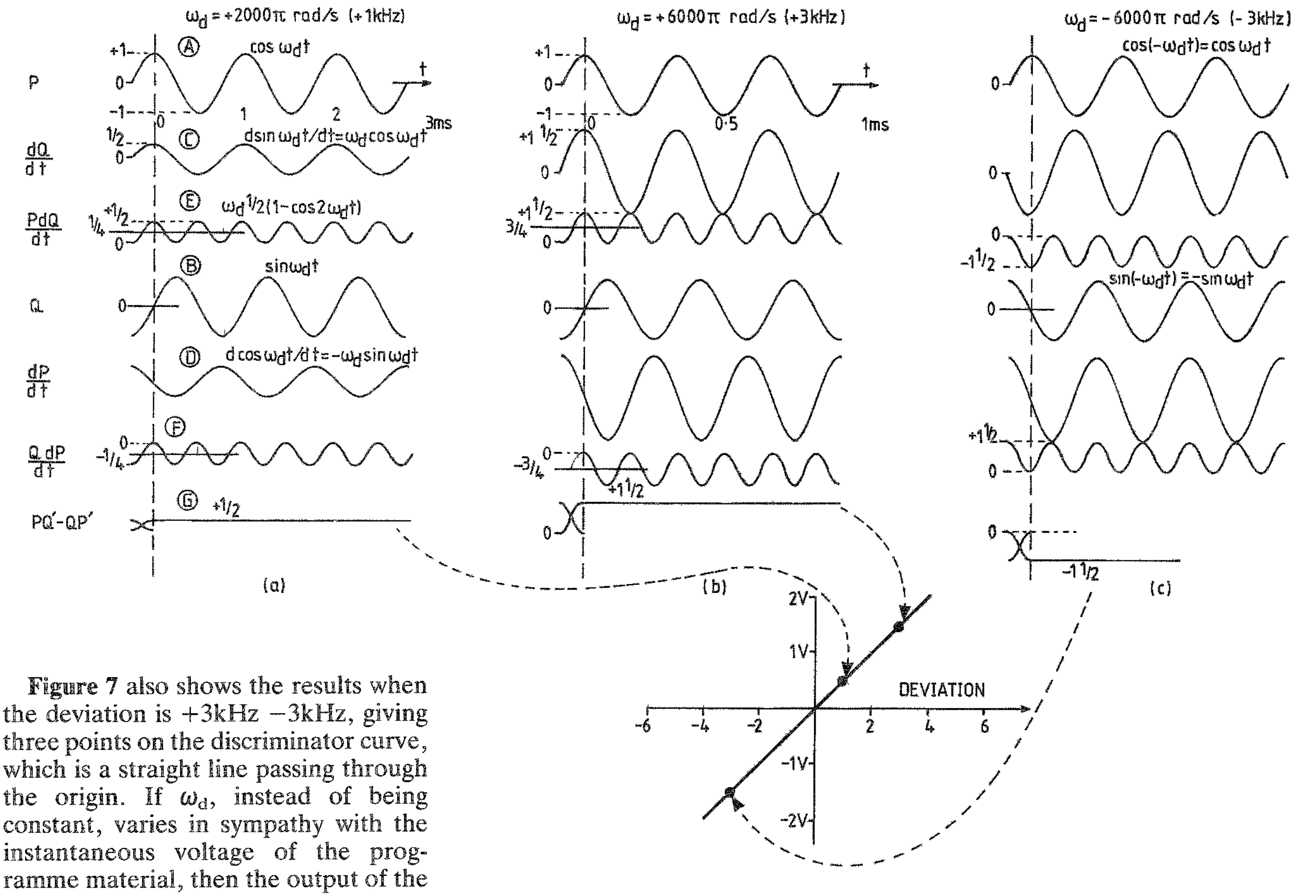


Figure 7 also shows the results when the deviation is +3kHz −3kHz, giving three points on the discriminator curve, which is a straight line passing through the origin. If ω_d, instead of being constant, varies in sympathy with the instantaneous voltage of the programme material, then the output of the circuit will simply be a recovered version of the original modulating signal as broadcast. This is illustrated for modulation by a single sinusoidal tone in **Fig.8.**

Note that, if the LO frequency is not exactly equal to the carrier frequency of the received signal, then the output of the circuit will contain an offset voltage, proportional to the mistuning, but this will not in any way affect the operation of the circuit as described. Indeed, in principle the offset could be equal to the peak output voltage at full modulation, so that the recovered audio would always be of one polarity, providing that the low-pass filters in Fig. 5 had a high enough cut-off frequency to pass twice the maximum deviation frequency.

The offset could be even greater; one could in theory apply expression (3) directly to a received broadcast FM signal at 100MHz, using the signal direct for the P(t) input and a version delayed by a quarter wavelength of coaxial cable for the Q(t) input. However, with the broadcast standard peak deviation limited to ±75kHz, the peak recovered audio would amount to only 0.075% of the standing DC offset, giving a rather poor signal-to-noise ratio.

Fig. 7. Sine-wave demodulator operation with a constant frequency offset. As seen in Fig. 5., subtracting QdP/dt from PdQ/dt gives a DC level proportional to frequency.

Homodyne in practice

The circuit of Fig. 5 could be implemented entirely in analogue circuitry, using double balanced mixers, low-pass filters and op-amps. Differentiation is very simply performed with an op-amp circuit, with none of the drift problems that beset integrators, while the multipliers could be implemented very cheaply using operational transconductance amplifiers (OTAs). An application note in the Motorola Linear handbook explains how to connect the LM13600 as a four-quadrant multiplier. However, as the denominator of (3) was ignored, the output of the circuit will vary in amplitude in sympathy with the square of the strength of the incoming signal; there is no AM suppression. The amplifiers G in Fig. 5 cannot be made into limiting amplifiers since, for the circuit to work, the baseband P and Q signals need to remain sinusoidal. In principle, the amplifiers could be provided with AGC loops, but these would need to track exactly in gain: not very practical.

Alternatively, the whole of the processing following the mixer low-pass filters in Fig. 5 can be performed by digital signal-processing circuitry; the P and Q baseband signal would be popped into A-to-D converters and digitized at a suitable sample rate. This would have to be at least twice the frequency of the highest audio modulation frequency, even for narrow-band FM. For wide-band FM, the sampling frequency would have to be at least twice the highest frequency deviation to cope with the P and Q signals at points A and B in Fig. 5. In practice, it would need to be higher still to allow for some mistuning of the LO, resulting in the positive peak deviation being greater than the negative or vice versa, and also to allow for practical rather than "brick-wall" low-pass filters following the mixers.

All the mathematical operations indicated in (3) can be performed by a digital signal processor, resulting in a digital output data stream which only needs popping into a D-to-A converter to recover the final audio. In addition to evaluating the numerator of (3) on a sample by sample basis, the DSP can also calculate $P^2(t)+Q^2(t)$ likewise. By dividing each sample by this value, the amplitude of the value of the final data samples is normalised; that is, the amplitude is now independent of variations of the incoming RF signal amplitude – AM suppression has been achieved. Naturally, this only works satisfactorily if the signals going into the A-to-D converters are large enough to provide a reasonable number of bits in the samples, or excessive quantisation noise will result.

I do not know of any homodyne FM receivers working on the principles outlined in this article, in either an analogue or digital implementation, other than the special case of the FSK paging receiver described earlier. Here I am limiting the term "homodyne" to receivers which translate the received signal directly from the incoming RF to baseband, that is to an IF of 0Hz. In this sense, a homodyne is a heterodyne receiver, though not a "superhet".

However, the homodyne principle as described can be and is used as the final IF stage in a double or triple superhet, the penultimate IF being translated down to the final IF of 0Hz, and there digitised. The following DSP section provides all the usual demodulation modes, including narrow band FM implemented as indicated using expression (3) in full. ■

References

1. P. Hawker. Keep it simple — direct-conversion HF receivers. Proc. Conf. on radio receivers, IERE, July 1978, p.137.
2. I.A.W. Vance and B.A.Bidwell. A new radio pager with monolithic receiver. Proc. Conf. on communications equipment, IEE, April 1982, p.138.
3. Coy, Smith and Smith. Use of DSP within a high-performance HFband receiver. Proc. Fifth International Conference on Radio Receivers and Associated Systems, Cambridge, July 1990. Conference Publication No. 325.
4. Dawson and Wayland. A broadband radio receiver designed for ESM applications. *Ibid.*
5. J.H.Roberts. Angle Modulation. Peter Peregrinus Ltd, 1977.
6. Tibbs and Johnstone. Frequency Modulation Engineering, 2nd edn. Chapman and Hall,1956.

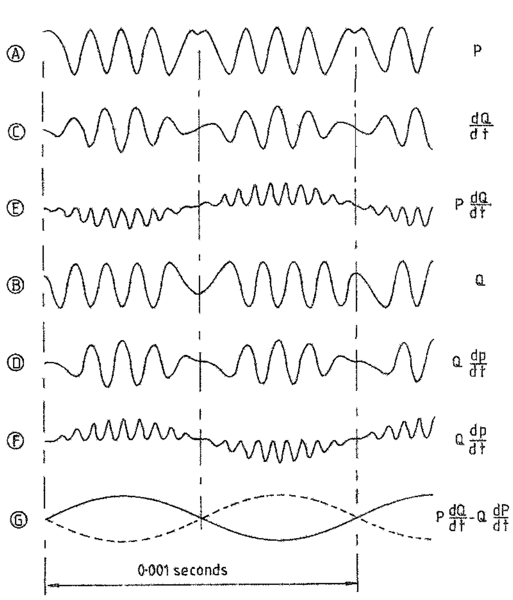
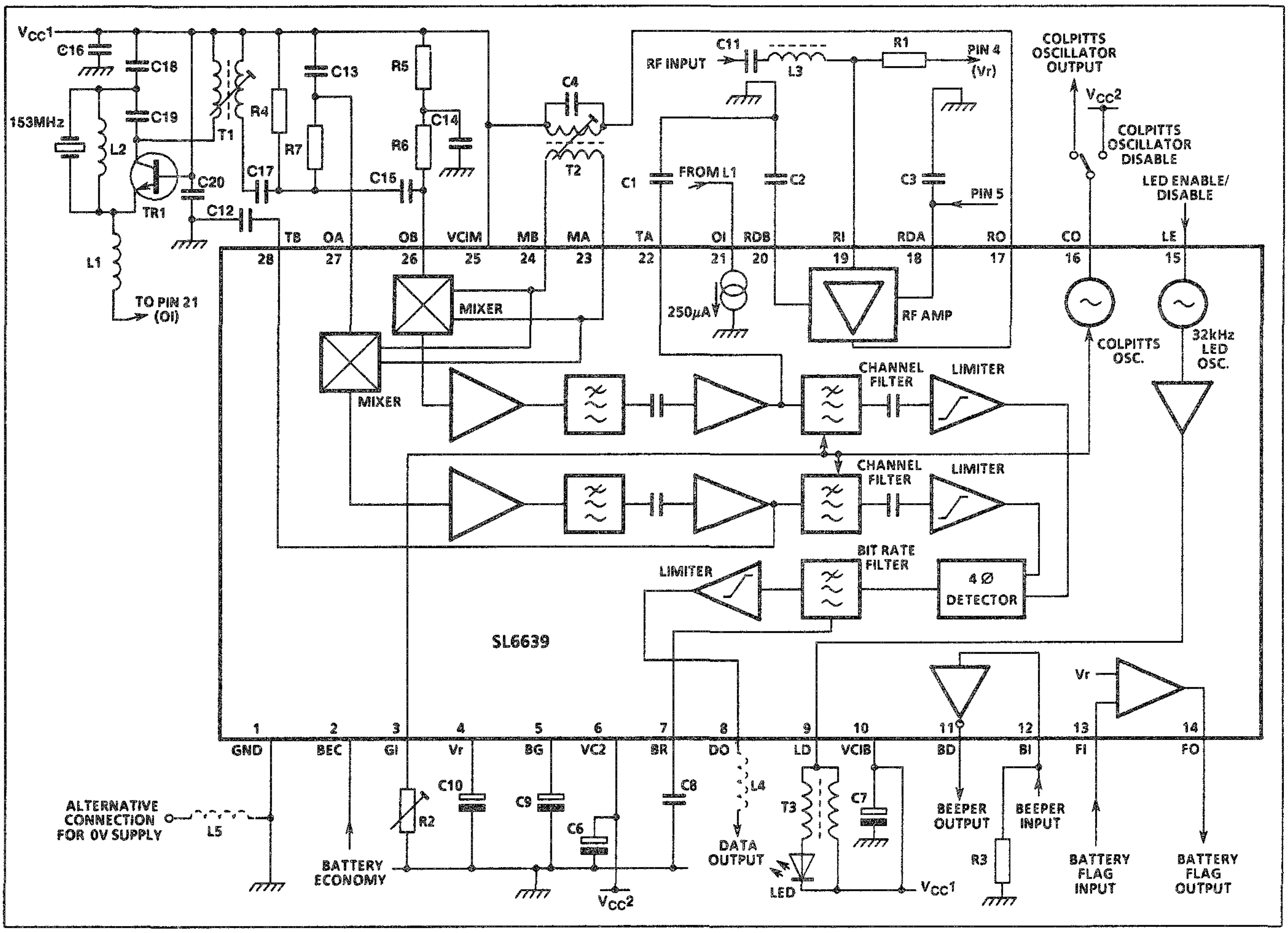


Fig. 8. Waveforms seen in the demodulator of Fig. 5. with a 1kHz FM signal of peak deviation 7kHz.

Fig. 9. Practical application of the SL6639 direct-conversion FSK data receiver chip from Plessey — a 153MHz receiver for a data rate of 512b/s.



Vcc1
C16
C18
C13
R5
C4
RF INPUT
C11
L3
R1
PIN 4 (Vr)
COLPITTS OSCILLATOR OUTPUT
Vcc2
COLPITTS OSCILLATOR DISABLE
LED ENABLE/ DISABLE
153MHz
C19
R4
T1
L2
R7
C14
R6
T2
C1
FROM L1
C2
C3
PIN 5
C20
C17
C15
TR1
C12
L1
TB
OA
OB
VCIM
MB
MA
TA
OI
RDB
RI
RDA
RO
CO
LE
28
27
26
25
24
23
22
21
20
19
18
17
16
15
TO PIN 21 (OI)
MIXER
250µA
RF AMP
COLPITTS OSC.
32kHz LED OSC.
MIXER
CHANNEL FILTER
LIMITER
CHANNEL FILTER
LIMITER
BIT RATE FILTER
LIMITER
4 Ø DETECTOR
SL6639
Vr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
GND
BEC
GI
Vr
BG
VC2
BR
DO
LD
VCIB
BD
BI
FI
FO
ALTERNATIVE CONNECTION FOR 0V SUPPLY
L5
C10
C9
C6
C8
L4
T3
C7
R2
BATTERY ECONOMY
Vcc2
DATA OUTPUT
LED
Vcc1
BEEPER OUTPUT
BEEPER INPUT
R3
BATTERY FLAG INPUT
BATTERY FLAG OUTPUT