5 ANGLE (EXPONENTIAL) MODULATION

In AM signals, the amplitude of a carrier is modulated by a signal $m(t)$, and, hence, the information content of $m(t)$ is in the amplitude variations of the carrier. Because a sinusoidal signal is described by amplitude and angle (which includes frequency and phase), there exists a possibility of carrying the same information by varying the angle of the carrier. This chapter explores such a possibility.

A Historical Note

In the twenties, broadcasting was in its infancy. However, there was a constant search for techniques that will reduce noise (static). Now, since the noise power is proportional to the modulated signal bandwidth (sidebands), the attempt was focused on finding a modulation scheme that will reduce the bandwidth. It was rumored that a new method had been discovered for eliminating sidebands (no sidebands, no bandwidth!). The idea of **frequency modulation (FM)**, where the carrier frequency would be varied in proportion to the message $m(t)$, appeared quite intriguing. The carrier frequency $\omega(t)$ would be varied with time so that $\omega(t) = \omega_c + km(t)$, where k is an arbitrary constant. If the peak amplitude of $m(t)$ is m_p, then the maximum and minimum values of the carrier frequency would be $\omega_c + km_p$ and $\omega_c - km_p$, respectively. Hence, the spectral components would remain within this band with a bandwidth $2km_p$ centered at ω_c. The bandwidth is controlled by the arbitrary constant k, whose value can be selected as we please. By using an arbitrarily small k, we could make the information bandwidth arbitrarily small. This was a passport to communication heaven. Unfortunately, the experimental results showed that something was seriously wrong somewhere. The FM bandwidth was found to be always greater than (at best equal to) the AM bandwidth. In some cases, its bandwidth was several times that of AM. Where is the fallacy in this reasoning? We shall soon find out.

5.1 CONCEPT OF INSTANTANEOUS FREQUENCY

By definition, a sinusoidal signal has a constant frequency, and, hence, the variation of frequency with time appears to be contradictory to the conventional definition of a sinusoidal

signal frequency. We must extend the concept of a sinusoid to a generalized function whose frequency may vary with time.

In FM we wish to vary the carrier frequency in proportion to the modulating signal $m(t)$. This means the carrier frequency is changing continuously every instant. Prima facie, this does not make much sense because to define a frequency, we must have a sinusoidal signal at least over one cycle (or a half-cycle or a quarter-cycle) with the same frequency. This problem reminds us of our first encounter with the concept of **instantaneous velocity** in our beginning mechanics course. Until that time, we were used to thinking of velocity as being constant over an interval, and we were incapable of even imagining that velocity could vary at each instant. But with some mental struggle, the idea gradually sinks in. We never forget, however, the wonder and amazement that was caused by the idea when it was first introduced. A similar experience awaits the reader with the concept of **instantaneous frequency**.

Let us consider a generalized sinusoidal signal $\varphi(t)$ given by

$$\varphi(t) = A \cos \theta(t) \tag{5.1}$$

where $\theta(t)$ is the **generalized angle** and is a function of t. Figure 5.1 shows a hypothetical case of $\theta(t)$. The generalized angle for a conventional sinusoid $A \cos(\omega_c t + \theta_0)$ is $\omega_c t + \theta_0$. This is a straight line with a slope ω_c and intercept θ_0, as shown in Fig. 5.1. The plot of $\theta(t)$ for the hypothetical case happens to be tangential to the angle $(\omega_c t + \theta_0)$ at some instant t. The crucial point is that over a small interval $\Delta t \to 0$, the signal $\varphi(t) = A \cos \theta(t)$ and the sinusoid $A \cos (\omega_c t + \theta_0)$ are identical; that is,

$$\varphi(t) = A \cos (\omega_c t + \theta_0) \qquad t_1 < t < t_2$$

We are certainly justified in saying that over this small interval Δt, the frequency of $\varphi(t)$ is ω_c. Because $(\omega_c t + \theta_0)$ is tangential to $\theta(t)$, the frequency of $\varphi(t)$ is the slope of its angle $\theta(t)$ over this small interval. We can generalize this concept at every instant and say that the instantaneous frequency ω_i at any instant t is the slope of $\theta(t)$ at t. Thus, for $\varphi(t)$ in Eq. (5.1),

$$\omega_i(t) = \frac{d\theta}{dt} \tag{5.2a}$$

$$\theta(t) = \int_{-\infty}^{t} \omega_i(\alpha)\, d\alpha \tag{5.2b}$$

Now we can see the possibility of transmitting the information of $m(t)$ by varying the angle θ of a carrier. Such techniques of modulation, where the angle of the carrier is varied in some manner with a modulating signal $m(t)$, are known as **angle modulation** or **exponential modulation**. Two simple possibilities are: **phase modulation (PM)** and **frequency modulation (FM)**. In PM, the angle $\theta(t)$ is varied linearly with $m(t)$:

$$\theta(t) = \omega_c t + \theta_0 + k_p m(t)$$

where k_p is a constant and ω_c is the carrier frequency. Assuming $\theta_0 = 0$, without loss of generality,

$$\theta(t) = \omega_c t + k_p m(t) \tag{5.3a}$$

The resulting PM wave is

$$\varphi_{PM}(t) = A \cos [\omega_c t + k_p m(t)] \tag{5.3b}$$

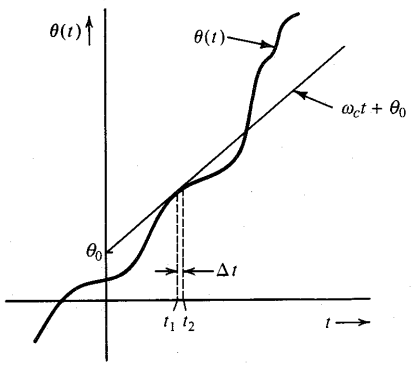


Figure 5.1 Concept of instantaneous frequency.

The instantaneous frequency $\omega_i(t)$ in this case is given by

$$\omega_i(t) = \frac{d\theta}{dt} = \omega_c + k_p \dot{m}(t) \tag{5.3c}$$

Hence, in PM, the instantaneous frequency ω_i varies linearly with the derivative of the modulating signal. If the instantaneous frequency ω_i is varied linearly with the modulating signal, we have FM. Thus, in FM the instantaneous frequency ω_i is

$$\omega_i(t) = \omega_c + k_f m(t) \tag{5.4a}$$

where k_f is a constant. The angle $\theta(t)$ is now

$$\theta(t) = \int_{-\infty}^{t} [\omega_c + k_f m(\alpha)]\, d\alpha$$

$$= \omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha \tag{5.4b}$$

Here we have assumed the constant term in $\theta(t)$ to be zero without loss of generality. The FM wave is

$$\varphi_{FM}(t) = A \cos \left[\omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha\right] \tag{5.4c}$$

Generalized Concept of Angle Modulation

From Eqs. (5.3b) and (5.4c), it is apparent that PM and FM are not only very similar but are inseparable. Replacing $m(t)$ in Eq. (5.3b) with $\int m(t)$ changes PM into FM. Thus, a signal that is an FM wave corresponding to $m(t)$ is also the PM wave corresponding to $\int m(\alpha)\, d\alpha$ (Fig. 5.2a). Similarly, a PM wave corresponding to $m(t)$ is the FM wave corresponding to $\dot{m}(t)$ (Fig. 5.2b). Therefore, by looking at an angle-modulated carrier, there is no way of telling whether it is FM or PM. In fact, it is meaningless to ask an angle-modulated wave whether it is FM or PM. An analogous practical situation would be to ask a person (who is married, with children) whether he is a father or a son. The person would be puzzled because he is both, a father (of his child) and a son (of his father).

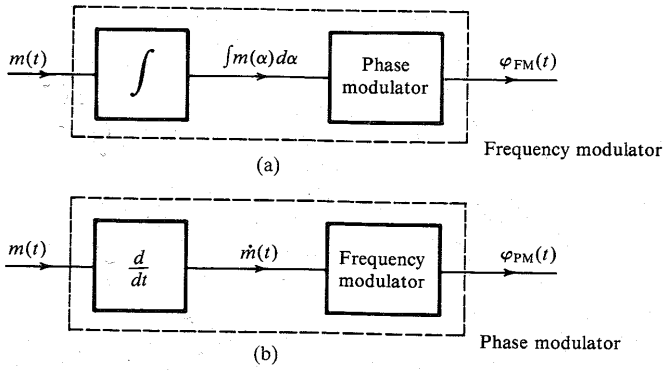


Figure 5.2 Phase and frequency modulation are inseparable.

Equations (5.3b) and (5.4c) show that in both PM and FM the angle of a carrier is varied in proportion to some measure of $m(t)$. In PM, it is directly proportional to $m(t)$, whereas in FM, it is proportional to the integral of $m(t)$. But why should we limit ourselves only to these cases? We have an infinite number of possible ways of generating a measure of $m(t)$. If we restrict the choice to a linear operator, then a measure of $m(t)$ can be obtained as the output of a suitable linear (time-invariant) system with $m(t)$ as its input, as shown in Fig. 5.3. The system transfer function is $H(s)$ and its impulse response is $h(t)$. The output of this system, $\psi(t)$, is a measure of $m(t)$. This is a reversible operation; that is, $m(t)$ can be recovered from $\psi(t)$ by passing it through a system of the transfer function $1/H(s)$.

The generalized angle-modulated carrier $\varphi_{EM}(t)$ can be expressed as

$$\varphi_{EM}(t) = A \cos [\omega_c t + \psi(t)] \tag{5.5a}$$

$$= A \cos \left[\omega_c t + \int_{-\infty}^{t} m(\alpha)\, h(t-\alpha)\, d\alpha\right] \tag{5.5b}$$

If $h(t) = k_p \delta(t)$, this equation reduces to Eq. (5.3b), and we have the conventional PM. Similarly, if $h(t) = k_f u(t)$, the equation reduces to Eq. (5.4c), resulting in conventional FM. Now, FM and PM are just two possibilities (out of an infinite number.) We shall see later that the optimum performance system is neither FM nor PM, but something else, depending on the modulating signal spectrum and the channel characteristics.

The generalized angle modulation concept is useful because it shows the convertibility of one type of angle modulation (such as PM) to another (such as FM). This is quite clear from Fig. 5.2. For instance, we show later that the bandwidth of FM is approximately $2k_f m_p$, where m_p is the peak amplitude of $m(t)$. We can derive the equivalent result for PM by referring to Fig. 5.2b, which shows that PM is actually the FM when the modulating signal is $\dot{m}(t)$. Clearly, the bandwidth of PM is approximately $2k_p m_p'$, where m_p' is the peak amplitude of $\dot{m}(t)$. This

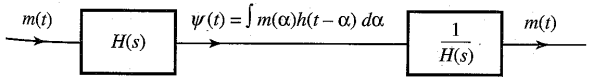


Figure 5.3 Generalized exponential modulation.

shows that if we analyze one type of angle modulation (such as FM), we can readily extend those results to any other kind. Historically, the angle modulation concept began with FM, and in this chapter we shall primarily analyze FM, with occasional discussion of PM. But this does not mean that FM is superior to other kinds of angle modulation. On the contrary, for most practical signals, PM is superior to FM. Actually, the optimum performance is realized neither by PM nor by FM, but by something in between.

This discussion also shows that we need not discuss methods of generation and demodulation of each type of modulation. From Fig. 5.2, it is clear that PM can be generated by an FM generator, and FM can be generated by a PM generator. One of the methods of generating FM in practice (the Armstrong indirect-FM system) actually integrates $m(t)$ and uses it to phase-modulate a carrier (see Fig. 5.6).

EXAMPLE 5.1 Sketch FM and PM waves for the modulating signal $m(t)$ shown in Fig. 5.4a. The constants k_f and k_p are $2\pi \times 10^5$ and 10π, respectively, and the carrier frequency f_c is 100 MHz.

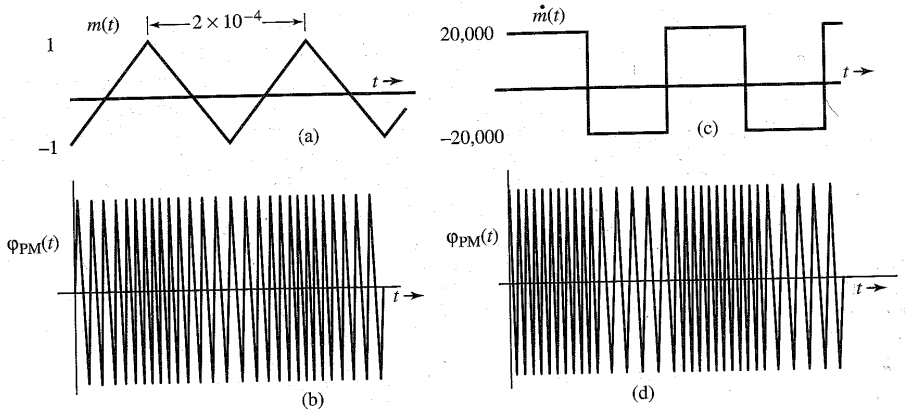


Figure 5.4 FM and PM waveforms.

For FM:

$$\omega_i = \omega_c + k_f m(t)$$

Dividing throughout by 2π, we have the equation in terms of the variable f (frequency in hertz). The instantaneous frequency f_i is

$$f_i = f_c + \frac{k_f}{2\pi} m(t)$$

$$= 10^8 + 10^5 m(t)$$

$$(f_i)_{\min} = 10^8 + 10^5\,[m(t)]_{\min} = 99.9 \text{ MHz}$$

$$(f_i)_{\max} = 10^8 + 10^5\,[m(t)]_{\max} = 100.1 \text{ MHz}$$

Because $m(t)$ increases and decreases linearly with time, the instantaneous frequency increases linearly from 99.9 to 100.1 MHz over a half-cycle and decreases linearly from 100.1 to 99.9 MHz over the remaining half-cycle of the modulating signal (Fig. 5.4b).

For PM: PM for $m(t)$ is FM for $\dot{m}(t)$. This also follows from Eq. (5.3c).

$$f_i = f_c + \frac{k_p}{2\pi} \dot{m}(t)$$

$$= 10^8 + 5\,\dot{m}(t)$$

$$(f_i)_{\min} = 10^8 + 5\,[\dot{m}(t)]_{\min} = 10^8 - 10^5 = 99.9 \text{ MHz}$$

$$(f_i)_{\max} = 10^8 + 5\,[\dot{m}(t)]_{\max} = 100.1 \text{ MHz}$$

Because $\dot{m}(t)$ switches back and forth from a value of $-20{,}000$ to 20,000, the carrier frequency switches back and forth from 99.9 to 100.1 MHz every half-cycle of $\dot{m}(t)$, as shown in Fig. 5.4d.

This indirect method of sketching PM [using $\dot{m}(t)$ to frequency-modulate a carrier] works as long as $m(t)$ is a continuous signal. If $m(t)$ is discontinuous, $\dot{m}(t)$ contains impulses, and this method fails. In such a case, a direct approach should be used. This is demonstrated in the next example.

EXAMPLE 5.2 Sketch FM and PM waves for the digital modulating signal $m(t)$ shown in Fig. 5.5a. The constants k_f and k_p are $2\pi \times 10^5$ and $\pi/2$, respectively, and $f_c = 100$ MHz.

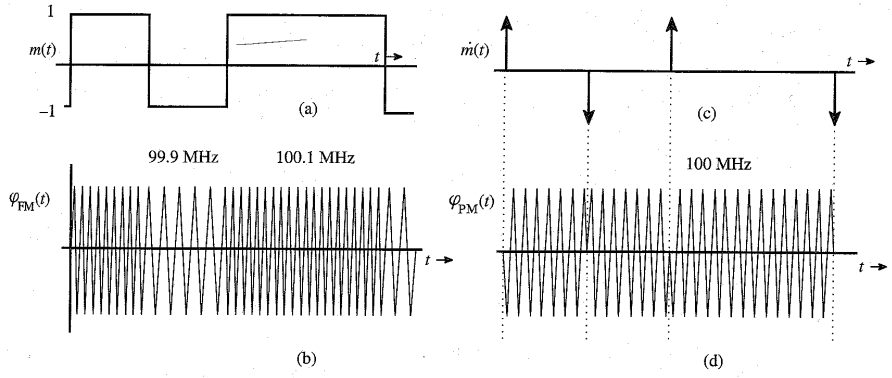


Figure 5.5 FM and PM waveforms.

For FM:

$$f_i = f_c + \frac{k_f}{2\pi} m(t) = 10^8 + 10^5 m(t)$$

Because $m(t)$ switches from 1 to -1 and vice versa, the FM wave frequency switches back and forth between 99.9 MHz and 100.1 MHz, as shown in Fig. 5.5b. This scheme of carrier frequency modulation by a digital signal (Fig. 5.5b) is called **frequency-shift keying (FSK)** because information digits are transmitted by shifting the carrier frequency (see Sec. 7.8).

For PM:

$$f_i = f_c + \frac{k_p}{2\pi}\dot{m}(t) = 10^8 + \frac{1}{4}\dot{m}(t)$$

The derivative $\dot{m}(t)$ (Fig. 5.5c) contains impulses of strength ± 2, and it is not immediately apparent how an instantaneous frequency can be changed by an infinite amount and then changed back to the original frequency in zero time. Let us consider the direct approach:

$$\begin{aligned}\varphi_{\text{PM}}(t) &= A\cos[\omega_c t + k_p m(t)] \\ &= A\cos\left[\omega_c t + \frac{\pi}{2}m(t)\right] \\ &= \begin{cases} A\sin\omega_c t & \text{when } m(t) = -1 \\ -A\sin\omega_c t & \text{when } m(t) = 1 \end{cases}\end{aligned}$$

This PM wave is shown in Fig. 5.5d. This scheme of carrier PM by a digital signal is called **phase-shift keying (PSK)** because information digits are transmitted by shifting the carrier phase. Note that PSK may also be viewed as a DSB-SC modulation by $m(t)$.

The PM wave $\varphi_{\text{PM}}(t)$ in this case has phase discontinuities at instants where impulses of $\dot{m}(t)$ are located. At these instants, the carrier phase shifts by π instantaneously. A finite phase shift in zero time implies infinite instantaneous frequency at these instants. This agrees with our observation about $\dot{m}(t)$.

The amount of phase discontinuity in $\varphi_{\text{PM}}(t)$ at the instant where $m(t)$ is discontinuous is $k_p m_d$, where m_d is the amount of discontinuity in $m(t)$ at that instant. In the present example, the amplitude of $m(t)$ changes by 2 (from -1 to 1) at the discontinuity. Hence, the phase discontinuity in $\varphi_{\text{PM}}(t)$ is $k_p m_d = (\pi/2) \times 2 = \pi$ rad, which confirms our earlier result.

When $m(t)$ is a digital signal (as in Fig. 5.5a), $\varphi_{\text{PM}}(t)$ shows a phase discontinuity where $m(t)$ has a jump discontinuity. We shall now show that in such a case the phase deviation $k_p m(t)$ must be restricted to a range $(-\pi, \pi)$ in order to avoid ambiguity in demodulation. For example, if k_p were $3\pi/2$ in the present example, then

$$\varphi_{\text{PM}}(t) = A\cos\left[\omega_c t + \frac{3\pi}{2}m(t)\right]$$

In this case $\varphi_{\text{PM}}(t) = A\sin\omega_c t$ when $m(t) = 1$ or $-1/3$. This will certainly cause ambiguity at the receiver when $A\sin\omega_c t$ is received. Such ambiguity never arises if $k_p m(t)$ is restricted to the range $(-\pi, \pi)$.

What causes this ambiguity? When $m(t)$ has jump discontinuities, the phase of $\varphi_{\text{PM}}(t)$ changes instantaneously. Because a phase $\varphi_o + 2n\pi$ is indistinguishable from the phase φ_o, ambiguities will be inherent in the demodulator unless the phase variations are limited to the range $(-\pi, \pi)$. This means k_p should be small enough to restrict the phase change $k_p m(t)$ to the range $(-\pi, \pi)$.

No such restriction on k_p is required if $m(t)$ is continuous. In this case the phase change is not instantaneous, but gradual over a time, and a phase $\varphi_o + 2n\pi$ will exhibit n additional carrier cycles over the case of phase of only φ_o. We can detect the PM wave by using an FM demodulator followed by an integrator (see Prob. 5.4-1). The additional n cycles will be detected by the FM demodulator, and the subsequent integration will yield a phase $2n\pi$. Hence, the phases φ_o and $\varphi_o + 2n\pi$ can be detected without ambiguity. This conclusion can also be verified from Example 5.1, where the maximum phase change $\Delta\varphi = 10\pi$.

Because a band-limited signal cannot have jump discontinuities, we can say that when $m(t)$ is band-limited, k_p has no restrictions.

Power of an Angle-Modulated Wave

Although the instantaneous frequency and phase of an angle-modulated wave can vary with time, the amplitude A always remains constant. Hence, the power of an angle-modulated wave (PM or FM) is always $A^2/2$, regardless of the value of k_p or k_f.

5.2 BANDWIDTH OF ANGLE-MODULATED WAVES

In order to determine the bandwidth of an FM wave, let us define

$$a(t) = \int_{-\infty}^{t} m(\alpha)\, d\alpha \tag{5.6}$$

and

$$\hat{\varphi}_{\text{FM}}(t) = A\, e^{j[\omega_c t + k_f a(t)]} = A e^{jk_f a(t)} e^{j\omega_c t} \tag{5.7a}$$

Now

$$\varphi_{\text{FM}}(t) = \text{Re}\ \hat{\varphi}_{\text{FM}}(t) \tag{5.7b}$$

Expanding the exponential $e^{jk_f a(t)}$ in Eq. (5.7a) in power series yields

$$\hat{\varphi}_{\text{FM}}(t) = A\left[1 + jk_f a(t) - \frac{k_f^2}{2!}a^2(t) + \cdots + j^n\frac{k_f^n}{n!}a^n(t) + \cdots\right]e^{j\omega_c t} \tag{5.8a}$$

and

$$\begin{aligned}\varphi_{\text{FM}}(t) &= \text{Re}\,[\hat{\varphi}_{\text{FM}}(t)] \\ &= A\left[\cos\omega_c t - k_f a(t)\sin\omega_c t - \frac{k_f^2}{2!}a^2(t)\cos\omega_c t + \frac{k_f^3}{3!}a^3(t)\sin\omega_c t + \cdots\right]\end{aligned} \tag{5.8b}$$

The modulated wave consists of an unmodulated carrier plus various amplitude-modulated terms, such as $a(t)\sin\omega_c t$, $a^2(t)\cos\omega_c t$, $a^3(t)\sin\omega_c t$, ... The signal $a(t)$ is an integral of $m(t)$. If $M(\omega)$ is band-limited to B, $A(\omega)$ is also band-limited* to B. The spectrum of $a^2(t)$

* This is because integration is a linear operation equivalent to passing a signal through a transfer function $1/j\omega$. Hence, if $M(\omega)$ is band-limited to B, $A(\omega)$ must also be band-limited to B.

is simply $A(\omega) * A(\omega)/2\pi$ and is band-limited to $2B$. Similarly, the spectrum of $a^n(t)$ is band-limited to nB. Hence, the spectrum consists of an unmodulated carrier plus spectra of $a(t)$, $a^2(t)$, ..., $a^n(t)$, ..., centered at ω_c. Clearly, the modulated wave is not band-limited. It has an infinite bandwidth and is not related to the modulating-signal spectrum in any simple way, as was the case in AM.

Although the theoretical bandwidth of an FM wave is infinite, we shall see that most of the modulated-signal power resides in a finite bandwidth. There are two distinct possibilities in terms of bandwidths—narrow-band FM and wide-band FM.

Narrow-Band Angle Modulation

Unlike AM, angle modulation is nonlinear. The principle of superposition does not apply. This may be verified from the fact that

$$A\cos\{\omega_c t + k_f[a_1(t) + a_2(t)]\} \neq A\cos[\omega_c t + k_f a_1(t)] + A\cos[\omega_c t + k_f a_2(t)]$$

The principle of superposition does not hold. If, however, k_f is very small (that is, if $|k_f a(t)| \ll 1$), then all but the first two terms in Eq. (5.8) are negligible, and we have

$$\varphi_{\text{FM}}(t) \simeq A[\cos\omega_c t - k_f a(t)\sin\omega_c t] \tag{5.9}$$

This is a linear modulation. This expression is similar to that of the AM wave. Because the bandwidth of $a(t)$ is B, the bandwidth of $\varphi_{\text{FM}}(t)$ in Eq. (5.9) is only $2B$. For this reason, the case ($|k_f a(t)| \ll 1$) is called **narrow-band FM (NBFM)**. The **narrow-band PM (NBPM)** case is similarly given by

$$\varphi_{\text{PM}}(t) \simeq A[\cos\omega_c t - k_p m(t)\sin\omega_c t] \tag{5.10}$$

A comparison of NBFM [Eq. (5.9)] with AM [Eq. (4.8a)] brings out clearly the similarities and differences between the two types of modulation. Both cases have a carrier term and sidebands centered at $\pm\omega_c$. The modulated-signal bandwidths are identical (viz., $2B$). The sideband spectrum for FM has a phase shift of $\pi/2$ with respect to the carrier, whereas that of AM is in phase with the carrier. It must be remembered, however, that despite apparent similarities, the AM and FM signals have very different waveforms. In an AM signal, the frequency is constant and the amplitude varies with time, whereas in an FM signal, the amplitude is constant and the frequency varies with time.

Equations (5.9) and (5.10) suggest a possible method of generating narrow-band FM and PM signals by using DSB-SC modulators. The block-diagram representation of such systems is shown in Fig. 5.6.

Wide-Band FM (WBFM): The Fallacy Exposed

If the deviation in the carrier frequency is large enough [i.e., if the constant k_f is chosen large enough so that the condition $|k_f a(t)| \ll 1$ is not satisfied], we cannot ignore the higher order terms in Eq. (5.8b), and the preceding analysis becomes too complicated to lead to a fruitful solution. We shall take here the route of the pioneers, who by their intuitively simple reasoning came to grief in estimating the FM bandwidth. If we could discover the fallacy in their reasoning, we would have a chance of obtaining a better estimate of the wide-band FM bandwidth.

Consider an $m(t)$ that is band-limited to B Hz. This signal is approximated by a staircase signal $\hat{m}(t)$, as shown in Fig. 5.7a. The signal $m(t)$ is now approximated by pulses of constant

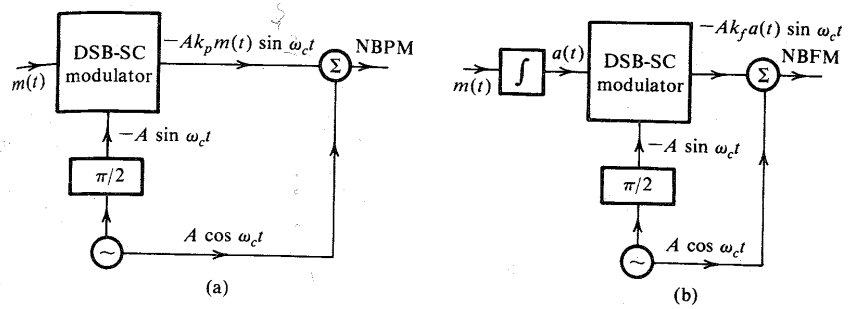


Figure 5.6 Narrow-band PM and FM wave generation.

amplitudes. For convenience, each of these pulses will be called a "cell." It is relatively easy to analyze FM corresponding to $\hat{m}(t)$ because it has constant amplitudes. To ensure that $\hat{m}(t)$ has all the information of $m(t)$, the cell width in $\hat{m}(t)$ must be no greater than the Nyquist interval of $1/2B$ seconds. Thus, $m(t)$ is approximated by constant-amplitude pulses (cells) of width $T = 1/2B$ seconds. Consider a typical cell starting at $t = t_k$. This cell has a constant amplitude $m(t_k)$. Hence, the FM signal corresponding to this cell is a sinusoid of frequency $\omega_c + k_f m(t_k)$ and duration $T = 1/2B$, as shown in Fig. 5.7b. The FM signal for $\hat{m}(t)$ consists of a sequence of such sinusoidal pulses corresponding to various cells of $\hat{m}(t)$.

The FM spectrum for $\hat{m}(t)$ consists of the sum of the Fourier transforms of these sinusoidal pulses corresponding to all the cells. The Fourier transform of a sinusoidal pulse in Fig. 5.7b (corresponding to the kth cell) is a sinc function shown shaded in Fig. 5.7c (see Example 3.12, Fig. 3.22d with $T = 1/2B$). Note that the spectrum of this pulse is spread out on either side of its frequency $\omega_c + k_f m(t_k)$ by $2\pi/T = 4\pi B$. Figure 5.7c shows the spectra of sinusoidal pulses corresponding to various cells. The minimum and the maximum amplitudes of the cells are $-m_p$ and m_p, respectively. Hence, the minimum and maximum frequencies of the sinusoidal pulses corresponding to the FM signal for all the cells are $\omega_c - k_f m_p$ and $\omega_c + k_f m_p$, respectively. Moreover, the spectrum for each sinusoid spreads out on either side of its frequency by $4\pi B$ rad/s, as shown in Fig. 5.7c. Hence, the maximum and the minimum significant frequencies in this spectrum are $\omega_c + k_f m_p + 4\pi B$ and $\omega_c - k_f m_p - 4\pi B$, respectively. The spectrum bandwidth is the difference $2k_f m_p + 8\pi B$.

We can now understand the fallacy in the reasoning of the pioneers. The maximum and minimum carrier frequencies are $\omega_c + k_f m_p$ and $\omega_c - k_f m_p$, respectively. Hence, it was reasoned that the spectral components must also lie in this range, resulting the FM bandwidth of $2k_f m_p$. The implicit assumption was that a sinusoid of frequency ω has its entire spectrum concentrated at ω. Unfortunately, this is true only of the everlasting sinusoid because the Fourier transform of such a sinusoid is an impulse at ω. For a sinusoid of finite duration T seconds, the spectrum is spread out on either side of ω by $2\pi/T$, as shown in Example 3.12. The pioneers had missed this spreading effect.

The deviation of the carrier frequency is $\pm k_f m_p$. We shall denote the carrier frequency deviation by $\Delta\omega$. Thus,

$$\Delta\omega = k_f m_p \tag{5.11}$$

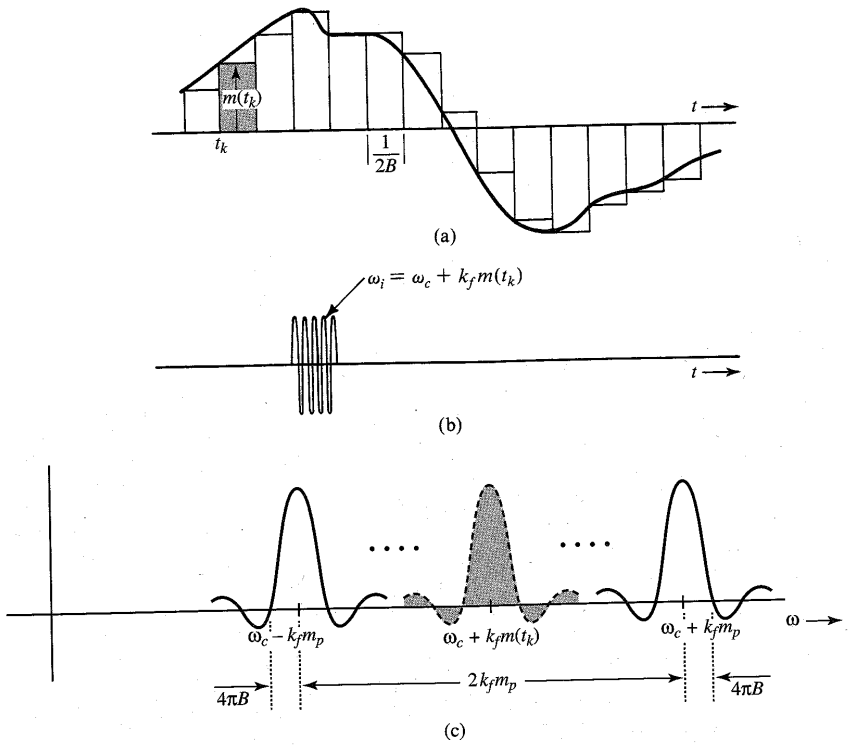


Figure 5.7 Estimation of FM wave bandwidth.

The carrier frequency deviation in hertz will be denoted by Δf. Thus,

$$\Delta f = \frac{k_f m_p}{2\pi}$$

The estimated FM bandwidth (in hertz) can be expressed as

$$\begin{aligned} B_{\text{FM}} &= \frac{1}{2\pi}(2k_f m_p + 8\pi B) \\ &= 2(\Delta f + 2B) \end{aligned} \tag{5.12}$$

The bandwidth estimate thus obtained is somewhat higher than the actual value because this is the bandwidth corresponding to the staircase approximation of $m(t)$, not the actual $m(t)$, which is considerably smoother. Hence, the actual bandwidth is somewhat smaller than this value. Therefore, we must readjust our bandwidth estimation. In order to make this midcourse

correction, we observe that for the narrow-band case, k_f is very small. Hence, Δf is very small (compared to B). In this case we can ignore the Δf term in Eq. (5.12) with the result

$$B_{\text{FM}} \approx 4B$$

But we have shown earlier that for narrow-band, the FM bandwidth is $2B$ Hz. This indicates that a better bandwidth estimate is

$$B_{\text{FM}} = 2(\Delta f + B) \tag{5.13a}$$

$$= 2\left(\frac{k_f m_p}{2\pi} + B\right) \tag{5.13b}$$

This is precisely the result obtained by Carson,[1] who investigated this problem rigorously for tone modulation [sinusoidal $m(t)$]. This formula goes under the name **Carson's rule** in the literature. Observe that for a truly wide-band case, where $\Delta f \gg B$, Eqs. (5.13) can be approximated as

$$B_{\text{FM}} \approx 2\Delta f \qquad \Delta f \gg B \tag{5.14}$$

Because $\Delta\omega = k_f m_p$, this formula is precisely what the pioneers had used for FM bandwidth. The only mistake was in thinking that this formula will hold for all cases, especially for the narrow-band case, where $\Delta f \ll B$.

We define a deviation ratio β as

$$\beta = \frac{\Delta f}{B} \tag{5.15}$$

Carson's rule can be expressed in terms of the deviation ratio as

$$B_{\text{FM}} = 2B(\beta + 1) \tag{5.16}$$

The deviation ratio controls the amount of modulation and, consequently, plays a role similar to the modulation index in AM. Indeed, for the special case of tone-modulated FM, the deviation ratio β is called the **modulation index.**

Phase Modulation

All the results derived for FM can be directly applied to PM. Thus, for PM, the instantaneous frequency is given by

$$\omega_i = \omega_c + k_p \dot{m}(t)$$

Therefore, the frequency deviation $\Delta\omega$ is given by

$$\Delta\omega = k_p m_p' \tag{5.17a}$$

where*

$$m_p' = [\dot{m}(t)]_{\max} \tag{5.17b}$$

* We are assuming that $|\dot{m}(t)_{\min}| = m_p'$.

Therefore,*

$$B_{\text{PM}} = 2(\Delta f + B) \tag{5.18a}$$

$$= 2\left(\frac{k_p m_p'}{2\pi} + B\right) \tag{5.18b}$$

One interesting aspect of FM is that $\Delta\omega = k_f m_p$ depends only on the peak value of $m(t)$. It is independent of the spectrum of $m(t)$. On the other hand, in PM, $\Delta\omega = k_p m_p'$ depends on the peak value of $\dot{m}(t)$. But $\dot{m}(t)$ depends strongly on the frequency spectrum of $m(t)$. The presence of higher frequency components in $m(t)$ implies rapid time variations, resulting in a higher value of m_p'. Similarly, predominance of lower frequency components will result in a lower value of m_p'. Hence, whereas the WBFM carrier bandwidth [Eq. (5.13)] is practically independent† of the spectrum of $m(t)$, the WBPM carrier bandwidth [Eq. (5.18)] strongly depends on the spectrum of $m(t)$. For $m(t)$ with a spectrum concentrated at lower frequencies, B_{PM} will be smaller than when the spectrum of $m(t)$ is concentrated at higher frequencies.

Verification of FM Bandwidth Relationship

We can verify the bandwidth relations for a specific case of tone modulation; that is, when $m(t)$ is a sinusoid. Let

$$m(t) = \alpha \cos \omega_m t$$

From Eq. (5.6),‡

$$a(t) = \frac{\alpha}{\omega_m} \sin \omega_m t$$

Thus, from Eq. (5.7a), we have

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{\left(j\omega_c t + \frac{k_f \alpha}{\omega_m} \sin \omega_m t\right)}$$

Moreover

$$\Delta\omega = k_f m_p = \alpha k_f$$

and the bandwidth of $m(t)$ is $B = f_m$ Hz. The deviation ratio (or the modulation index, in this case) is

$$\beta = \frac{\Delta f}{f_m} = \frac{\Delta\omega}{\omega_m} = \frac{\alpha k_f}{\omega_m}$$

Hence,

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{j(\omega_c t + \beta \sin \omega_m t)}$$

$$= Ae^{j\omega_c t}(e^{j\beta \sin \omega_m t}) \tag{5.19}$$

* Equation (5.17a) can be applied only if $m(t)$ is a continuous function of time. If $m(t)$ has jump discontinuities, its derivative does not exist. In such a case, we should use the direct approach (discussed in Example 5.2) to find $\varphi_{\text{PM}}(t)$ and then determine $\Delta\omega$ from $\varphi_{\text{PM}}(t)$.

† Except for its weak dependence on B [Eqs. (5.13)].

‡ Here we are assuming that the constant $a(-\infty) = 0$.

The exponential term in parentheses is a periodic signal with period $2\pi/\omega_m$ and can be expanded by the exponential Fourier series, as usual,

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} C_n e^{jn\omega_m t}$$

where

$$C_n = \frac{\omega_m}{2\pi} \int_{-\pi/\omega_m}^{\pi/\omega_m} e^{j\beta \sin \omega_m t} e^{-jn\omega_m t}\, dt$$

Letting $\omega_m t = x$, we get

$$C_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(\beta \sin x - nx)}\, dx$$

The integral on the right-hand side cannot be evaluated in a closed form but must be integrated by expanding the integrand in infinite series. This integral has been extensively tabulated and is denoted by $J_n(\beta)$, the Bessel function of the first kind and nth order. These functions are plotted in Fig. 5.8a as a function of n for various values of β. Thus,

$$e^{j\beta \sin \omega_m t} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{jn\omega_m t} \tag{5.20}$$

Substituting Eq. (5.20) into Eq. (5.19), we get

$$\hat{\varphi}_{\text{FM}}(t) = A \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j(\omega_c t + n\omega_m t)}$$

and

$$\hat{\varphi}_{\text{FM}}(t) = A \sum_{n=-\infty}^{\infty} J_n(\beta) \cos\,(\omega_c + n\omega_m)t$$

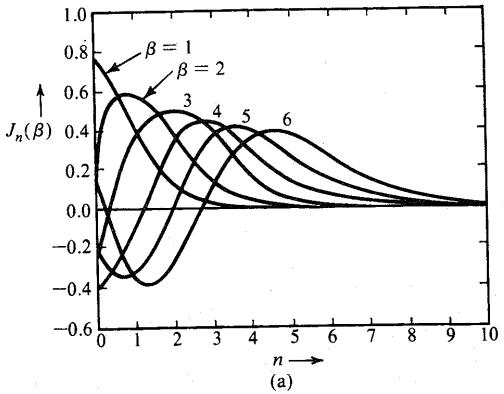
The modulated signal has a carrier component and an infinite number of sidebands of frequencies $\omega_c \pm \omega_m$, $\omega_c \pm 2\omega_m$, ..., $\omega_c \pm n\omega_m$, ..., as shown in Fig. 5.8b. The strength of the nth sideband at $\omega = \omega_c + n\omega_m$ is* $J_n(\beta)$. From the plots of $J_n(\beta)$ in Fig. 5.8a it can be seen that for a given β, $J_n(\beta)$ decreases with n. For a sufficiently large n, $J_n(\beta)$ is negligible, and there are only a finite number of significant sidebands. It can be seen from Fig. 5.8a that $J_n(\beta)$ is negligible for $n > \beta + 1$. Hence, the number of significant sidebands is $\beta + 1$. The bandwidth of the FM carrier is given by

$$B_{\text{FM}} = 2nf_m = 2(\beta + 1)f_m$$

$$= 2(\Delta f + B)$$

which verifies our previous result [Eqs. (5.13)]. When $\beta \ll 1$ (NBFM), there is only one significant sideband and the bandwidth $B_{\text{FM}} = 2f_m = 2B$. It is important to note that this example is a verification, not a proof, of Carson's formula.

* Also $J_{-n}(\beta) = (-1)^n J_n(\beta)$. Hence, the magnitude of the LSB at $\omega = \omega_c - n\omega_m$ is the same as that of the USB at $\omega = \omega_c + n\omega_m$.



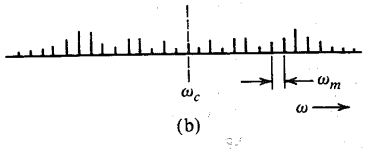


Figure 5.8 (a) Variations of $J_n(\beta)$ as a function of n for various values of β. (b) Tone-modulated FM wave spectrum.

Amplitude modulation is a linear kind of modulation. Hence, most of the results derived for tone modulation are generally valid for other signals. In the literature, tone modulation in FM is often discussed in great details. Unfortunately, angle modulation being a nonlinear kind of modulation, the results derived for tone modulation may have little connection to practical situations. Indeed, these results are meaningless at best and misleading at worst when applied to practical signals. For instance, based on tone modulation analysis, it is often stated that FM is superior to PM by a factor of 3 in terms of the output SNR. We show in Sec. 12.3 that for most of the practical signals, it is PM that is superior to FM. This author feels that too much stress on tone modulation can be misleading. For this reason we have omitted further discussion of tone modulation here.

The method for finding the spectrum of a tone-modulated FM wave can be used for finding the spectrum of an FM wave when $m(t)$ is a general periodic signal. In this case,

$$\hat{\varphi}_{\text{FM}}(t) = Ae^{j\omega_c t}[e^{jk_f a(t)}]$$

Because $a(t)$ is a periodic signal, $e^{jk_f a(t)}$ is also a periodic signal, which can be expressed as an exponential Fourier series in the preceding expression. After this, it is relatively straightforward to write $\varphi_{\text{FM}}(t)$ in terms of the carrier and the sidebands.

EXAMPLE 5.3 **(a)** Estimate B_{FM} and B_{PM} for the modulating signal $m(t)$ in Fig. 5.4a for $k_f = 2\pi \times 10^5$ and $k_p = 5\pi$.

(b) Repeat the problem if the amplitude of $m(t)$ is doubled [if $m(t)$ is multiplied by 2].

(a) The peak amplitude of $m(t)$ is unity. Hence, $m_p = 1$. We now determine the essential bandwidth B of $m(t)$. It is left as an exercise for the reader to show that the Fourier series for this periodic signal is given by

$$m(t) = \sum_n C_n \cos n\omega_0 t \qquad \omega_0 = \frac{2\pi}{2 \times 10^{-4}} = 10^4\pi$$

where

$$C_n = \begin{cases} \dfrac{8}{\pi^2 n^2} & n \text{ odd} \\ 0 & n \text{ even} \end{cases}$$

It can be seen that the harmonic amplitudes decrease rapidly with n. The third harmonic is only 11% of the fundamental, and the fifth harmonic is only 4% of the fundamental. This means the third and fifth harmonic powers are 1.21 and 0.16%, respectively, of the fundamental component power. Hence, we are justified in assuming the essential bandwidth of $m(t)$ as the frequency of the third harmonic, that is, $3(10^4/2)$ Hz. Thus,

$$B = 15\,\text{kHz}$$

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = \frac{1}{2\pi}(2\pi \times 10^5)(1) = 100\,\text{kHz}$$

and

$$B_{\text{FM}} = 2(\Delta f + B) = 230\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{100}{15}$$

and

$$B_{\text{FM}} = 2B(\beta + 1) = 30\left(\frac{100}{15} + 1\right) = 230\,\text{kHz}$$

For PM: The peak amplitude of $\dot{m}(t)$, is 20,000, and

$$\Delta f = \frac{1}{2\pi} k_p m_p' = 50\ \text{kHz}$$

Hence,

$$B_{\text{PM}} = 2(\Delta f + B) = 130\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{50}{15}$$

and

$$B_{\text{PM}} = 2B(\beta + 1) = 30\left(\frac{50}{15} + 1\right) = 130\,\text{kHz}$$

(b) Doubling $m(t)$ doubles its peak value. Hence, $m_p = 2$. But its bandwidth is unchanged so that $B = 15$ kHz.

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = \frac{1}{2\pi}(2\pi \times 10^5)(2) = 200\,\text{kHz}$$

and

$$B_{\text{FM}} = 2(\Delta f + B) = 430\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{200}{15}$$

and

$$B_{\text{FM}} = 2B(\beta + 1) = 30\left(\frac{200}{15} + 1\right) = 430\,\text{kHz}$$

For PM: Doubling $m(t)$ doubles its derivative so that now $m_p' = 40,000$, and

$$\Delta f = \frac{1}{2\pi} k_p m_p' = 100\text{kHz}$$

and

$$B_{\text{PM}} = 2(\Delta f + B) = 230\,\text{kHz}$$

Alternately, the deviation ratio β is given by

$$\beta = \frac{\Delta f}{B} = \frac{100}{15}$$

and

$$B_{\text{PM}} = 2B(\beta + 1) = 30\left(\frac{100}{15} + 1\right) = 230\,\text{kHz}$$

Observe that doubling the signal amplitude [doubling $m(t)$] roughly doubles the bandwidth of both FM and PM waveforms.

EXAMPLE 5.4 Repeat Example 5.3 if $m(t)$ is time-expanded by a factor of 2; that is, if the period of $m(t)$ is 4×10^{-4}.

Recall that time expansion of a signal by a factor of 2 reduces the signal spectral width (bandwidth) by a factor of 2. We can verify this by observing that the fundamental frequency is now 2.5 kHz, and its third harmonic is 7.5 kHz. Hence, $B = 7.5$ kHz, which is half the previous bandwidth. Moreover, time expansion does not affect the peak amplitude so that $m_p = 1$. However, m_p' is halved, that is, $m_p' = 10,000$.

For FM:

$$\Delta f = \frac{1}{2\pi} k_f m_p = 100\,\text{kHz}$$

$$B_{\text{FM}} = 2(\Delta f + B) = 2(100 + 7.5) = 215\,\text{kHz}$$

For PM:

$$\Delta f = \frac{1}{2\pi} k_p m_p' = 25\,\text{kHz}$$

$$B_{\text{PM}} = 2(\Delta f + B) = 65\,\text{kHz}$$

Note that time expansion of $m(t)$ has very little effect on the FM bandwidth, but it halves the PM bandwidth. This verifies our observation that the PM spectrum is strongly dependent on the spectrum of $m(t)$.

EXAMPLE 5.5 An angle-modulated signal with carrier frequency $\omega_c = 2\pi \times 10^5$ is described by the equation

$$\varphi_{\text{EM}}(t) = 10\cos(\omega_c t + 5\sin 3000t + 10\sin 2000\pi t)$$

(a) Find the power of the modulated signal.
(b) Find the frequency deviation Δf.
(c) Find the deviation ratio β.
(d) Find the phase deviation $\Delta\phi$.
(e) Estimate the bandwidth of $\varphi_{\text{EM}}(t)$.

The signal bandwidth is the highest frequency in $m(t)$ (or its derivative). In this case $B = 2000\pi/2\pi = 1000$ Hz.

(a) The carrier amplitude is 10, and the power is

$$P = 10^2/2 = 50$$

(b) To find the frequency deviation Δf, we find the instantaneous frequency ω_i, given by

$$\omega_i = \frac{d}{dt}\theta(t) = \omega_c + 15,000\cos 3000t + 20,000\pi\cos 2000\pi t$$

The carrier deviation is $15,000\cos 3000t + 20,000\pi\cos 2000\pi t$. The two sinusoids will add in phase at some point, and the maximum value of this expression is $15,000 + 20,000\pi$. This is the maximum carrier deviation $\Delta\omega$. Hence,

$$\Delta f = \frac{\Delta\omega}{2\pi} = 12,387.32\text{ Hz}$$

(c)

$$\beta = \frac{\Delta f}{B} = \frac{12,387.32}{1000} = 12.387$$

(d) The angle $\theta(t) = \omega t + (5\sin 3000t + 10\sin 2000\pi t)$. The phase deviation is the maximum value of the angle inside the parentheses, and is given by $\Delta\phi = 15$ rad.

(e)

$$B_{EM} = 2(\Delta f + B) = 26,774.65 \text{ Hz}$$

Observe the generality of this method of estimating the bandwidth of an angle-modulated waveform. We need not know whether it is FM, PM, or some other kind of angle modulation. It is applicable to any angle-modulated signal.

A Historical Note: Edwin H. Armstrong (1890–1954)

Today, nobody doubts that FM has a place in broadcasting and communication. As recently as the late sixties, the future of FM broadcasting seemed doomed because of uneconomical operations.

The history of FM is full of strange ironies. The impetus behind the development of FM was the necessity to reduce the transmission bandwidth. Superficial reasoning showed that it was feasible to reduce the transmission bandwidth by using FM. But the experimental results showed otherwise. The transmission bandwidth of FM was actually larger than that of AM. Careful mathematical analysis by Carson showed that FM indeed required a larger bandwidth than AM. Unfortunately, Carson did not recognize the compensating advantage of FM in its ability to suppress noise. Without much basis, he concluded that FM introduced inherent distortion and had no compensating advantages whatsoever.[1] In a later paper he says "In fact, as more and more schemes are analyzed and tested, and as the essential nature of the problem is more clearly perceivable, we are unavoidably forced to the conclusion that static (noise), like the poor, will always be with us."[2] The opinion of one of the ablest mathematicians of the day in the communication field, thus, set back the development of FM by more than a decade. The noise-suppressing advantage of FM was later proved by Major Edwin H. Armstrong,[3] a brilliant engineer whose contributions to the field of radio systems are comparable with those of Hertz and Marconi. It was largely the work of Armstrong that was responsible for rekindling the interest in FM.

Although Armstrong did not invent the concept of FM, he must be considered the father of modern FM. To quote from the early British text *Frequency Modulation Engineering* by Christopher E. Tibbs: "The subject of frequency modulation as we understand it today may be considered to date from Armstrong's paper of 1936. It is true that a good deal of the knowledge of the subject existed prior to that date, but Armstrong was the first to point out in a truly remarkable paper those peculiar characteristics to which modern technique owes its value."[4]

Armstrong was one of the leading architects who laid the groundwork for the mass-communication system. His work on FM came toward the close of his career. Before that, he was well known for several breakthrough contributions to the radio field. *Fortune* magazine says[5]: "Wideband frequency modulation is the fourth, and perhaps the greatest, in a line of Armstrong inventions that have made most of modern broadcasting what it is. Major Armstrong is the acknowledged inventor of the regenerative 'feedback' circuit, which brought radio art out of the crystal-detector headphone stage and made the amplification of broadcasting possible; the superheterodyne circuit, which is the basis of practically all modern radio; and the super-regenerative circuit now in wide use in ... shortwave systems."

Armstrong was the last of the breed of the lone attic inventors. For the sake of establishing FM broadcasting, he fought a long and costly battle with the radio broadcast establishment,

which, abetted by the Federal Communications Commission (FCC), fought tooth and nail to resist FM. In 1944, the FCC, on the basis of erroneous testimony of a technical expert, abruptly shifted the allocated bandwidth of FM from the 42–50-MHz range to 88–108 MHz. This dealt a crippling blow to FM by making obsolete all the equipment (transmitters, receivers, antennas, etc.) that had been built and sold for the old FM bands. Armstrong continued to fight the decision, and in 1947 he succeeded in getting the technical expert to admit his error. In spite of all this, the FCC allocations remained unchanged. Armstrong spent a sizable fortune that he made from previous inventions in legal struggles. The broadcast industry, which so strongly resisted FM, turned around and used his inventions without paying him royalties. Armstrong spent nearly half of his life in the law courts in some of the longest, most notable, and acrimonious patent suits of the era.[4] In the end, with his funds depleted, his energy drained, and his family life shattered, a despondent Armstrong committed suicide (in 1954) by walking out of a window 13 stories above the street.

Features of Angle Modulation

FM (and angle modulation in general) has a number of unique features that recommend it for various radio systems. The transmission bandwidth of AM systems cannot be changed. Because of this AM systems do not have the feature of exchanging signal power for transmission bandwidth. PCM systems have such a feature, and so do angle-modulated systems. In angle modulation, the transmission bandwidth can be adjusted by adjusting Δf. It is shown in Chapter 12 that for angle-modulated systems, the SNR is roughly proportional to the square of the transmission bandwidth B_T. Recall that in PCM, the SNR varies exponentially with B_T and is, therefore, superior to angle modulation.

Immunity of Angle Modulation to Nonlinearities: Another interesting feature of angle modulation is its constant amplitude, which makes it less susceptible to nonlinearities. Consider, for instance, a second-order nonlinear device whose input $x(t)$ and output $y(t)$ are related by

$$y(t) = a_1 x(t) + a_2 x^2(t)$$

If

$$x(t) = \cos [\omega_c t + \psi(t)]$$

then

$$\begin{aligned} y(t) &= a_1 \cos [\omega_c t + \psi(t)] + a_2 \cos^2 [\omega_c t + \psi(t)] \\ &= \frac{a_2}{2} + a_1 \cos [\omega_c t + \psi(t)] + \frac{a_2}{2} \cos [2\omega_c t + 2\psi(t)] \end{aligned}$$

For the FM wave $\psi(t) = k_f \int m(\alpha)\, d\alpha$, and

$$y(t) = \frac{a_2}{2} + a_1 \cos \left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] + \frac{a_2}{2} \cos \left[2\omega_c t + 2k_f \int m(\alpha)\, d\alpha\right]$$

The dc term is filtered out to give the output that contains the original signal plus an additional FM signal, whose carrier frequency as well as frequency deviation are multiplied by 2. Note, however, that the information $m(t)$ is intact in both terms. Thus, the nonlinearity has not

distorted the information in any way. Because of the property of multiplying the carrier frequency, such nonlinear devices are also called **frequency multipliers**.

In the preceding case, because the device was of second order, it multiplied the frequency by 2. We can generalize this result for an nth-order multiplier (nonlinear device). Any nonlinear device, such as a diode or a transistor, can be used for this purpose. The characteristic of these devices can be expressed as

$$y(t) = a_0 + a_1 x(t) + a_2 x^2(t) + \cdots + a_n x^n(t) \tag{5.21}$$

If $x(t) = A \cos \left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right]$, then using trigonometric identities, we can readily show that $y(t)$ is of the form

$$y(t) = c_o + c_1 \cos \left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] + c_2 \cos \left[2\omega_c t + 2k_f \int m(\alpha)\, d\alpha\right]$$
$$+ \cdots + c_n \cos \left[n\omega_c t + nk_f \int m(\alpha)\, d\alpha\right] \tag{5.22}$$

Hence, the output will have spectra at ω_c, $2\omega_c, \ldots,$ $n\omega_c$, with frequency deviations Δf, $2\Delta f$, $\ldots$, $n\Delta f$, respectively. Hence, the nonlinearity generates components at unwanted frequencies. But the desired term $\cos [\omega_c t + \psi(t)]$ is undistorted, and by using a bandpass filter centered at ω_c, we can suppress all unwanted terms in $y(t)$ and obtain the desired signal component without distortion. Note that even the unwanted terms have the desired information intact, and any one of the unwanted terms can be used to extract information. The term $\cos [2\omega_c t + 2k_f \int m(\alpha)\, d\alpha]$, for instance, has twice the original carrier frequency and twice the original frequency deviation. Hence, such nonlinear devices can be used to increase the carrier frequency as well as the frequency deviation.

A similar nonlinearity in AM not only causes unwanted modulation with carrier frequencies $n\omega_c$ but also causes distortion of the desired signal. For instance, if a DSB-SC signal $m(t) \cos \omega_c t$ passes through a nonlinearity $y(t) = a\, x(t) + b\, x^3(t)$, the output is

$$y(t) = am(t) \cos \omega_c t + bm^3(t) \cos^3 \omega_c t$$
$$= \left[am(t) + \frac{3b}{4} m^3(t)\right] \cos \omega_c t + \frac{b}{4} m^3(t) \cos 3\omega_c t$$

Passing this signal through a bandpass filter yields $[am(t) + (3b/4)m^3(t)] \cos \omega_c t$. Observe the distortion component $(3b/4)m^3(t)$ present along with the desired signal $am(t)$.

Immunity from nonlinearity is the primary reason why angle modulation is used in microwave radio relay systems, where power levels are high. This requires highly efficient nonlinear class C amplifiers. In addition, the constant amplitude of FM gives it a kind of immunity against rapid fading. The effect of amplitude variations caused by rapid fading can be eliminated by using automatic gain control and bandpass limiting (discussed in Sec. 5.4). These features make FM attractive for microwave radio relay systems. Angle modulation is also less vulnerable than AM to small signal interference from adjacent channels. Finally, as stated earlier, FM is capable of exchanging SNR for the transmission bandwidth.

In telephone systems, several channels are multiplexed using SSB signals. The multiplexed signal is frequency modulated and transmitted over a microwave radio relay system with many links in tandem. In this application, however, FM is used not to realize the noise reduction but to realize other advantages of constant amplitude, and, hence, NBFM rather than WBFM is used.

WBFM is used widely in space and satellite communication systems. The large bandwidth expansion reduces the required SNR and thus reduces the transmitter power requirement —which is very important because of weight considerations in space. WBFM is also used for high-fidelity radio transmission over rather limited areas.

5.3 GENERATION OF FM WAVES

Basically, there are two ways of generating FM waves: **indirect generation** and **direct generation**.

Indirect Method of Armstrong

In this method, NBFM is generated by integrating $m(t)$ and using it to phase modulate a carrier, as shown in Fig. 5.6b [or Eq. (5.9)]. The NBFM is then converted to WBFM by using frequency multipliers (discussed earlier), as shown in Fig. 5.9. Thus, if we want a 12-fold increase in the frequency deviation, we can use a 12th-order nonlinear device or two second-order and one third-order devices in cascade. The output has a bandpass filter centered at $12\omega_c$, so that it selects only the appropriate term, whose carrier frequency as well as the frequency deviation Δf are 12 times the original values. Generally, we require to increase Δf by a very large factor n. This increases the carrier frequency also by n. Such a large increase in the carrier frequency may not be needed. In this case we can use frequency mixing (see Example 4.2, Fig. 4.7) to shift down the carrier frequency to the desired value (recall that a frequency mixer shifts the carrier frequency).

The NBFM generated by Armstrong's method (Fig. 5.6b) has some distortion because of the approximation of Eqs. (5.8) by Eq. (5.9) (see Example 5.6). The output of the Armstrong NBFM modulator, as a result, also has some amplitude modulation. Amplitude limiting in the frequency multipliers removes most of this distortion.

A simplified diagram of a commercial FM transmitter using Armstrong's method is shown in Fig. 5.10. The final output is required to have a carrier frequency of 91.2 MHz and $\Delta f = 75$ kHz. We begin with NBFM with a carrier frequency $f_{c_1} = 200$ kHz generated by a crystal oscillator. This frequency is chosen because it is easy to construct stable crystal oscillators as well as balanced modulators at this frequency. The deviation Δf is chosen to be 25 Hz in order to maintain $\beta \ll 1$, as required in NBPM. For tone modulation $\beta = \Delta f / f_m$. The baseband spectrum (required for high-fidelity purposes) ranges from 50 Hz to 15 kHz. The choice of $\Delta f = 25$ Hz is reasonable because it gives $\beta = 0.5$ for the worst possible case ($f_m = 50$).

In order to achieve $\Delta f = 75$ kHz, we need a multiplication of $75{,}000/25 = 3000$. This can be done by two multiplier stages, of 64 and 48, as shown in Fig. 5.10, giving a

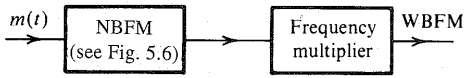


Figure 5.9 Simplified block diagram of Armstrong indirect FM wave generator.

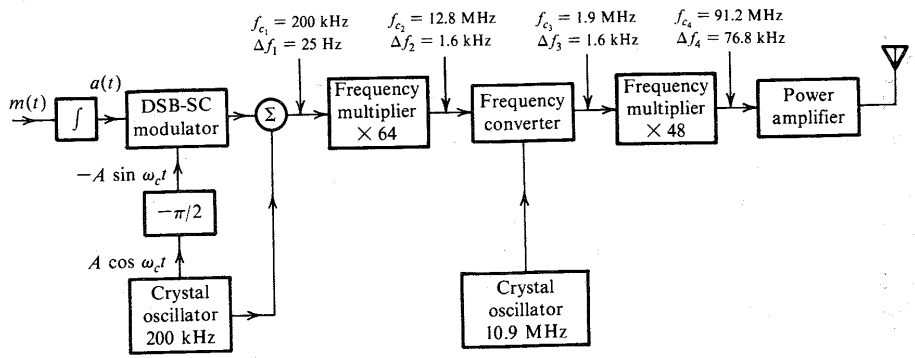


Figure 5.10 Armstrong indirect FM transmitter.

total multiplication of $64 \times 48 = 3072$, and $\Delta f = 76.8$ kHz.* The multiplication is effected by using frequency doublers and triplers in cascade, as needed. Thus, a multiplication of 64 can be obtained by six doublers in cascade, and a multiplication of 48 can be obtained by four doublers and a tripler in cascade. Multiplication of $f_c = 200$ kHz by 3072, however, would yield a final carrier of about 600 MHz. This difficulty is avoided by using a frequency translation, or conversion, after the first multiplier (Fig. 5.10). The first multiplication by 64 results in the carrier frequency $f_{c_2} = 200\,\text{kHz} \times 64 = 12.8$ MHz, and the carrier deviation $\Delta f_2 = 25 \times 64 = 1.6$ kHz. We now shift the entire spectrum using a frequency converter (or mixer) with carrier frequency 10.9 MHz. This results in a new carrier frequency $f_{c_3} = 12.8 - 10.9 = 1.9$ MHz. The frequency converter shifts the entire spectrum without altering Δf. Hence, $\Delta f_3 = 1.6$ kHz. Further multiplication, by 48, yields $f_{c_4} = 1.9 \times 48 = 91.2$ MHz and $\Delta f_4 = 1.6 \times 48 = 76.8$ kHz.

This scheme has an advantage of frequency stability, but it suffers from inherent noise caused by excessive multiplication and distortion at lower modulating frequencies, where $\Delta f/f_m$ is not small enough.

EXAMPLE 5.6 Discuss the nature of distortion inherent in the Armstrong indirect FM generator.

Two kinds of distortions arise in this scheme: amplitude distortion and frequency distortion. The NBFM wave is given by [Eq. (5.9)]

$$\varphi_{\text{FM}}(t) = A[\cos \omega_c t - k_f a(t) \sin \omega_c t]$$
$$= AE(t) \cos [\omega_c t + \theta(t)]$$

where

$$E(t) = \sqrt{1 + k_f^2 a^2(t)} \qquad \text{and} \qquad \theta(t) = \tan^{-1}[k_f a(t)]$$

* If we wish Δf to be exactly 75 kHz instead of 76.8 kHz, we must reduce the narrow-band Δf from 25 Hz to 25(75/76.8) = 24.41 Hz.

Amplitude distortion occurs because the amplitude $AE(t)$ of the modulated waveform is not constant. This is not a serious problem, because amplitude variations can be eliminated by a bandpass limiter discussed in the next section (see Fig. 5.12). Ideally, $\theta(t)$ should be $k_f a(t)$. Instead, the phase $\theta(t)$ in the preceding equation is

$$\theta(t) = \tan^{-1}[k_f a(t)]$$

and the instantaneous frequency $\omega_i(t)$ is

$$\omega_i(t) = \dot{\theta}(t) = \frac{k_f \dot{a}(t)}{1 + k_f^2 a^2(t)}$$
$$= \frac{k_f m(t)}{1 + k_f^2 a^2(t)}$$
$$= k_f m(t)[1 - k_f^2 a^2(t) + k_f^4 a^4(t) - \cdots]$$

Ideally, the instantaneous frequency should be $k_f m(t)$. The remaining terms in this equation are the distortion.

Let us investigate the effect of this distortion in tone modulation, where $m(t) = \alpha \cos \omega_m t$, $a(t) = \alpha \sin \omega_m t/\omega_m$, and the modulation index $\beta = \alpha k_f/\omega_m$. Hence,

$$\omega_i(t) = \beta \omega_m \cos \omega_m t (1 - \beta^2 \sin^2 \omega_m t + \beta^4 \sin^4 \omega_m t - \cdots)$$

It is evident from this equation that this scheme has odd-harmonic distortion, the most important term being the third harmonic. Ignoring the remaining terms, this equation becomes

$$\omega_i(t) \simeq \beta \omega_m \cos \omega_m t (1 - \beta^2 \sin^2 \omega_m t)$$
$$= \beta \omega_m \left(1 - \frac{\beta^2}{4}\right) \cos \omega_m t + \frac{\beta^3 \omega_m}{4} \cos 3\omega_m t$$
$$\simeq \underbrace{\beta \omega_m \cos \omega_m t}_{\text{desired}} + \underbrace{\frac{\beta^3 \omega_m}{4} \cos 3\omega_m t}_{\text{distortion}} \qquad \text{for } \beta \ll 1$$

The ratio of the third harmonic distortion to the desired signal is $\beta^2/4$. For the generator in Fig. 5.10, the worst possible case occurs at the lower modulation frequency of 50 Hz, where $\beta = 0.5$. In this case the third harmonic distortion is 1/16, or 6.25%.

Direct Generation

In a voltage-controlled oscillator (VCO), the frequency is controlled by an external voltage. The oscillation frequency varies linearly with the control voltage. We can generate an FM wave by using the modulating signal $m(t)$ as a control signal. This gives

$$\omega_i(t) = \omega_c + k_f m(t)$$

One can construct a VCO using an operational amplifier and an hysteric comparator[6] (such as a Schmitt trigger circuit). Another way of accomplishing the same goal is to vary one of the reactive parameters (C or L) of the resonant circuit of an oscillator. A reverse-biased

semiconductor diode acts as a capacitor whose capacitance varies with the bias voltage. The capacitance of these diodes, known under several trade names (such as varicaps, varactors, or voltacaps), can be approximated as a linear function of the bias voltage $m(t)$ over a limited range. In Hartley or Colpitt oscillators, for instance, the frequency of oscillation is given by

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

If the capacitance C is varied by the modulating signal $m(t)$, that is, if

$$\begin{aligned} C &= C_0 - km(t) \\ \omega_0 &= \frac{1}{\sqrt{LC_0\left[1 - \dfrac{km(t)}{C_0}\right]}} \\ &= \frac{1}{\sqrt{LC_0}\left[1 - \dfrac{km(t)}{C_0}\right]^{1/2}} \\ &\approx \frac{1}{\sqrt{LC_0}}\left[1 + \frac{km(t)}{2C_0}\right] \qquad \frac{km(t)}{C_0} \ll 1 \end{aligned}$$

Here we have used the binomial approximation $(1+x)^n \approx 1 + nx$ for $|x| \ll 1$. Thus,

$$\begin{aligned} \omega_0 &= \omega_c\left[1 + \frac{km(t)}{2C_0}\right] & \omega_c &= \frac{1}{\sqrt{LC_0}} \\ &= \omega_c + k_f m(t) & k_f &= \frac{k\omega_c}{2C_0} \end{aligned}$$

Because $C = C_0 - km(t)$, the maximum capacitance deviation is

$$\Delta C = km_p = \frac{2k_f C_0 m_p}{\omega_c}$$

Hence,

$$\frac{\Delta C}{C_0} = \frac{2k_f m_p}{\omega_c} = \frac{2\Delta f}{f_c}$$

In practice, $\Delta f/f_c$ is usually small, and, hence, ΔC is a small fraction of C_0, which helps limit the harmonic distortion that arises because of the approximation used in this derivation.

We may also generate direct FM by using a saturable core reactor, where the inductance of a coil is varied by a current through a second coil (also wound around the same core). This results in a variable inductor whose inductance is proportional to the current in the second coil.

Direct FM generation generally produces sufficient frequency deviation and requires little frequency multiplication. But this method has poor frequency stability. In practice, feedback is used to stabilize the frequency. The output frequency is compared with a constant frequency generated by a stable crystal oscillator. An error signal (error in frequency) is detected and fed back to the oscillator to correct the error.

5.4 DEMODULATION OF FM

The information in an FM signal resides in the instantaneous frequency $\omega_i = \omega_c + k_f m(t)$. Hence, a frequency-selective network with a transfer function of the form $|H(\omega)| = a\omega + b$ over the FM band would yield an output proportional to the instantaneous frequency (Fig. 5.11a).* There are several possible networks with such characteristics. The simplest among them is an ideal differentiator with the transfer function $j\omega$.

If we apply $\varphi_{FM}(t)$ to an ideal differentiator, the output is

$$\begin{aligned} \dot{\varphi}_{FM}(t) &= \frac{d}{dt}\left\{A\cos\left[\omega_c t + k_f\int_{-\infty}^{t} m(\alpha)\,d\alpha\right]\right\} \\ &= A\left[\omega_c + k_f m(t)\right]\sin\left[\omega_c t + k_f\int_{-\infty}^{t} m(\alpha)\,d(\alpha)\right] \end{aligned} \tag{5.23}$$

The signal $\dot{\varphi}_{FM}(t)$ is both amplitude and frequency modulated (Fig. 5.11b), the envelope being $A[\omega_c + k_f m(t)]$. Because $\Delta\omega = k_f m_p < \omega_c$, $\omega_c + k_f m(t) > 0$ for all t, and $m(t)$ can be obtained by envelope detection of $\dot{\varphi}_{FM}(t)$ (Fig. 5.11c).

The amplitude A of the incoming FM carrier is assumed to be constant. If the amplitude A were not constant, but a function of time, there would be an additional term containing

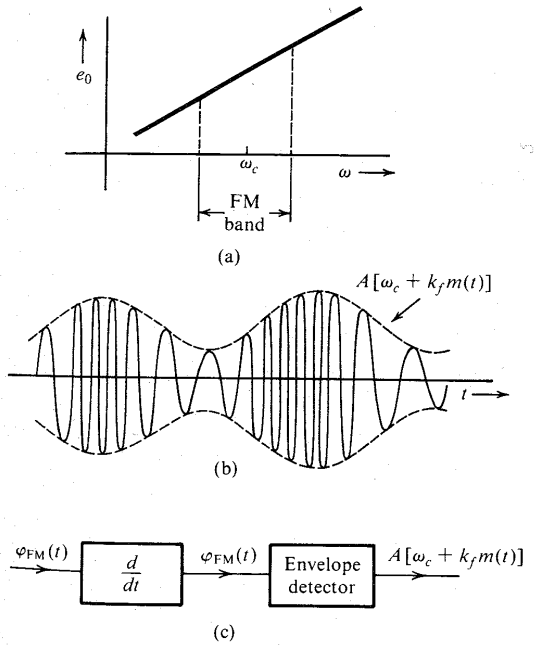


Figure 5.11 (a) FM demodulator frequency response. (b) Output of a differentiator to the input FM wave. (c) FM demodulation by direct differentiation.

* Provided the variations of ω_i are slow in comparison to the time constant of the network.

dA/dt on the right-hand side of Eq. (5.23). Even if this term were neglected, the envelope of $\dot{\varphi}_{\text{FM}}(t)$ would be $A(t)[\omega_c + k_f m(t)]$, and the envelope-detector output would be proportional to $m(t)A(t)$. Hence, it is essential to maintain A constant. Several factors, such as channel noise, fading, and so on, cause A to vary. This variation in A should be removed before applying the signal to the FM detector.

Bandpass Limiter

The amplitude variations of an angle-modulated carrier can be eliminated by what is known as a **bandpass limiter**, which consists of a hard limiter followed by a bandpass filter (Fig. 5.12a). The input-output characteristic of a hard limiter is shown in Fig. 5.12b. Observe that the bandpass limiter output to a sinusoid will be a square wave of unit amplitude regardless of the incoming sinusoidal amplitude. Moreover, the zero crossings of the incoming sinusoid are preserved in the output because when the input is zero, the output is also zero (Fig. 5.12b). Thus an angle-modulated sinusoidal input $v_i(t) = A(t)\cos\theta(t)$ results in a constant-amplitude, angle-modulated square wave $v_o(t)$, as shown in Fig. 5.12c. As we have seen earlier, such a nonlinear operation preserves the angle modulation information. When $v_o(t)$ is passed through a bandpass filter centered at ω_c, the output is a constant-amplitude, angle-modulated wave. To show this, consider the incoming angle-modulated wave

$$v_i(t) = A(t)\cos\theta(t)$$

where

$$\theta(t) = \omega_c t + k_f \int_{-\infty}^{t} m(\alpha)\, d\alpha$$

The output $v_o(t)$ of the hard limiter is $+1$ or -1, depending on whether $v_i(t) = A(t)\cos\theta(t)$ is positive or negative (Fig. 5.12c). Because $A(t) \geq 0$, $v_o(t)$ can be expressed as a function of θ:

$$v_o(\theta) = \begin{cases} 1 & \cos\theta > 0 \\ -1 & \cos\theta < 0 \end{cases}$$

Hence, v_o as a function of θ is a periodic square-wave function with period 2π (Fig. 5.12d), which can be expanded by a Fourier series [see Eq. (2.76)],

$$v_o(\theta) = \frac{4}{\pi}\left(\cos\theta - \frac{1}{3}\cos 3\theta + \frac{1}{5}\cos 5\theta + \cdots\right)$$

This is valid for any real variable θ. At any instant t, $\theta = \omega_c t + k_f \int m(\alpha)\, d\alpha$, and the output is $v_o[\omega_c t + k_f \int m(\alpha)\, d\alpha]$. Hence, the output v_o as a function of time is given by

$$\begin{aligned} v_o[\theta(t)] &= v_o\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \\ &= \frac{4}{\pi}\left\{\cos\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] - \frac{1}{3}\cos 3\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right]\right. \\ &\quad \left. + \frac{1}{5}\cos 5\left[\omega_c t + k_f \int m(\alpha)\, d\alpha\right] \cdots\right\} \end{aligned}$$

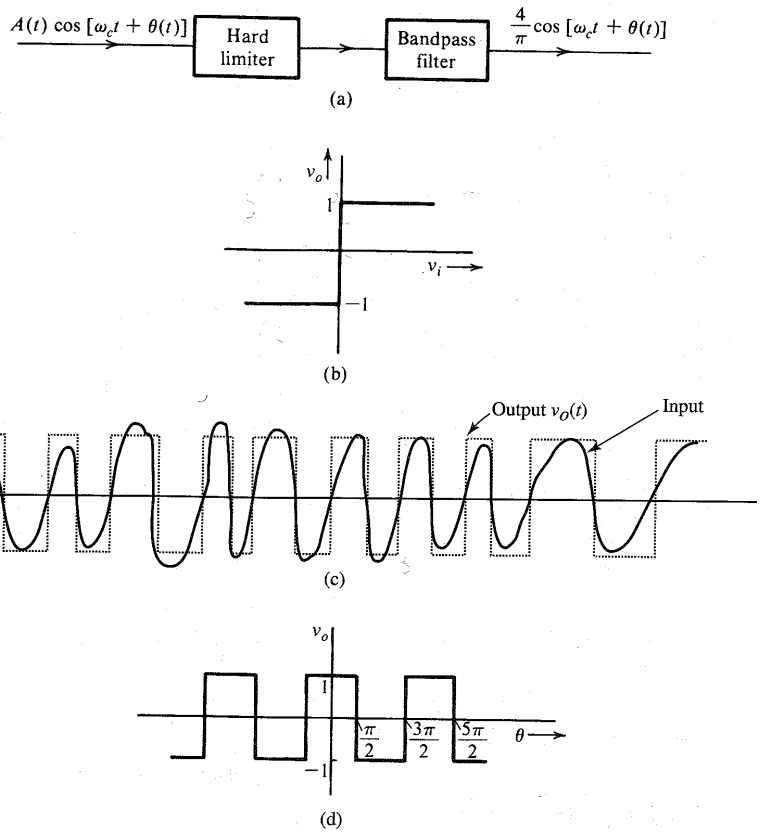


Figure 5.12 (a) Hard limiter and bandpass filter used to remove amplitude variations in FM wave. (b) Hard limiter input-output characteristic. (c) Hard limiter input and the corresponding output. (d) Hard limiter output as a function of θ.

The output, therefore, has the original FM wave plus a frequency-multiplied FM wave with multiplication factors of 3, 5, 7, ... We can pass the output of the hard limiter through a bandpass filter with a center frequency ω_c and a bandwidth B_{FM}, as shown in Fig. 5.12a. The filter output $e_o(t)$ is the desired angle-modulated carrier with a constant amplitude,

$$e_o(t) = \frac{4}{\pi}\cos\left[\omega_c(t) + k_f \int m(\alpha) d\alpha\right]$$

Although we derived these results for FM, this applies to PM (angle modulation in general) as well. The bandpass filter not only maintains the constant amplitude of the angle-modulated carrier but also partially suppresses the channel noise when the noise is small.[7]

Practical Frequency Demodulators

One can use an operational amplifier differentiator as an FM demodulator. A simple tuned circuit followed by an envelope detector can also serve as a frequency detector because its frequency response $|H(\omega)|$ below (or above) the resonance frequency is approximately linear of the form $a\omega + b$. Since the operation is on the slope of $|H(\omega)|$, this method is also called **slope detection**. It suffers from the fact that the slope of $|H(\omega)|$ is linear over only a small band and, hence, causes considerable distortion in the output. This fault can be partially corrected by a **balanced discriminator**.

Another balanced demodulator, the **ratio detector**, also widely used in the past, offers better protection against carrier amplitude variations than does the discriminator. For many years ratio detectors were standard in almost all FM receivers.[8]

Zero-crossing detectors are also used because of advances in digital integrated circuits. These are the **frequency counters** designed to measure the instantaneous frequency by the number of zero crossings. The rate of zero crossings is equal to the instantaneous frequency of the input signal.

Phase-Locked Loop (PLL): Because of their low cost and superior performance, especially when the SNR is low, FM demodulation using PLL is the most widely used method today. In Chapter 4, we saw how a PLL tracks the incoming signal angle and instantaneous frequency. Consider the PLL in Fig. 5.13a. The output $e_o(t)$ of the loop filter $H(s)$ acts as an input to the VCO (Fig. 5.13a). The free-running frequency of VCO is set at the carrier frequency ω_c. The instantaneous frequency of the VCO is given by [see Eq. (4.25)]

$$\omega_{\text{VCO}} = \omega_c + ce_o(t)$$

If the VCO output is $B\cos[\omega_c t + \theta_o(t)]$, then its instantaneous frequency is $\omega_c + \dot{\theta}_o(t)$. Therefore,

$$\dot{\theta}_o(t) = ce_o(t) \tag{5.24}$$

where c and B are constants of the PLL.

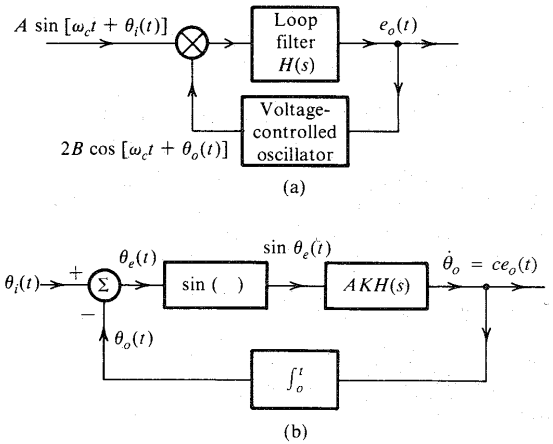


Figure 5.13 Phase-locked loop and its equivalent circuit.

Let the incoming signal (input to the PLL) be $A\sin[\omega_c t + \theta_i(t)]$. If the incoming signal happens to be $A\sin[\omega_o t + \psi(t)]$, it can still be expressed as $A\sin[\omega_c t + \theta_i(t)]$, where $\theta_i(t) = (\omega_o - \omega_c)t + \psi(t)$. Hence, the analysis that follows is general and not restricted to equal frequencies of the incoming signal and the free-running VCO signal.

The multiplier output is

$$AB\sin(\omega_c t + \theta_i)\cos(\omega_c t + \theta_o) = \frac{AB}{2}[\sin(\theta_i - \theta_o) + \sin(2\omega_c t + \theta_i + \theta_o)]$$

The sum frequency term is suppressed by the loop filter. Hence, the effective input to the loop filter is $\frac{1}{2}AB\sin[\theta_i(t) - \theta_o(t)]$. If $h(t)$ is the unit impulse response of the loop filter,

$$\begin{aligned} e_o(t) &= h(t) * \frac{1}{2}AB\sin[\theta_i(t) - \theta_o(t)] \\ &= \frac{1}{2}AB\int_0^t h(t-x)\sin[\theta_i(x) - \theta_o(x)]\,dx \end{aligned} \tag{5.25}$$

Substituting Eq. (5.24) in Eq. (5.25),

$$\dot{\theta}_o(t) = AK\int_0^t h(t-x)\sin\theta_e(x)\,dx \tag{5.26}$$

where $K = \frac{1}{2}cB$ and $\theta_e(t)$ is the phase error, defined as

$$\theta_e(t) = \theta_i(t) - \theta_o(t)$$

These equations [along with Eq. (5.24)] immediately suggest a model for the PLL, as shown in Fig. 5.13b.

When the incoming FM carrier* is $A\sin[\omega_c t + \theta_i(t)]$,

$$\theta_i(t) = k_f\int_{-\infty}^t m(\alpha)\,d\alpha \tag{5.27}$$

Hence,

$$\theta_o(t) = k_f\int_{-\infty}^t m(\alpha)\,d\alpha - \theta_e$$

and, assuming a small error θ_e,

$$e_o(t) = \frac{1}{c}\dot{\theta}_o(t) \simeq \frac{k_f}{c}m(t) \tag{5.28}$$

Thus, the PLL acts as an FM demodulator. If the incoming signal is a PM wave, $\theta_o(t) = \theta_i(t) = k_p m(t)$ and $e_o(t) = k_p\dot{m}(t)/c$. In this case we need to integrate $e_o(t)$ to obtain the desired signal. A detailed analysis of PLL is given next for two special cases.

Small-Error Analysis

In this case, $\sin\theta_e \simeq \theta_e$, and the block diagram in Fig. 5.13b reduces to the linear (time-invariant) system shown in Fig. 5.14a. Straightforward calculation gives

* Here we are using $\sin[\omega_c t + \theta_i(t)]$ rather than the usual $\cos[\omega_c t + \theta_i(t)]$. This is really immaterial, because a cosine can be expressed as a sine with a $\pi/2$ phase addition. Because the final step [Eq. (5.28)] involves differentiation of the angle, the constant phase vanishes.