5 ANGLE MODULATION AND DEMODULATION

As discussed in the previous chapter, a carrier modulation can be achieved by modulating the amplitude, frequency, and phase of a **sinusoidal carrier** of frequency f_c. In that chapter, we focused on various linear amplitude modulation systems and their demodulations. Now we discuss nonlinear frequency modulation (FM) and phase modulation (PM), often collectively known as angle modulation.

5.1 NONLINEAR 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. As we have seen, the other two parameters of the carrier sinusoid, namely its frequency and phase, can also be varied in proportion to the message signal as frequency-modulated and phase-modulated signals, respectively. We now describe the essence of frequency modulation (FM) and phase modulation (PM).

False Start

In the 1920s, broadcasting was in its infancy. However, there was an active search for techniques to reduce noise (static). Since the noise power is proportional to the modulated signal bandwidth (sidebands), efforts were focused on finding a modulation scheme that would reduce the bandwidth. More important still, bandwidth reduction also allows more users, and there were rumors of a new method that 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 $m(t)$, was quite intriguing. The carrier angular 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 understanding was that controlling the constant parameter k can control the modulated signal bandwidth. While this is true, there was also the hope that by using an arbitrarily small k, we could make the information bandwidth arbitrarily small. This possibility was seen as a passport to communication heaven. Unfortunately, experimental results showed that the underlying reasoning was seriously wrong. The FM bandwidth, as it turned out, is always greater than (at best equal to)

the AM bandwidth. In some cases, its bandwidth was several times that of AM. Where was the fallacy in the original reasoning? We shall soon find out.

The Concept of Instantaneous Frequency

While AM signals carry a message with their varying amplitude, FM signals can vary the instantaneous frequency in proportion to the modulating signal $m(t)$. This means that the carrier frequency is changing continuously every instant. Prima facie, this does not make much sense, since 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 a beginning mechanics course. Until the presentation of derivatives via Leibniz and Newton, 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. We never forget, however, the wonder and amazement that was caused by the contemplation of derivative and instantaneous velocity when these concepts were first introduced. A similar experience awaits the reader with respect to **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 a straight line $\omega_c t + \theta_0$, as shown in Fig. 5.1. A hypothetical case general angle of $\theta(t)$ happens to be tangential to the angle $(\omega_c t + \theta_0)$ at some instant t. The crucial point is that, around t, 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 angular frequency of $\varphi(t)$ is ω_c. Because $(\omega_c t + \theta_0)$ is tangential to $\theta(t)$, the angular 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 define that the instantaneous frequency ω_i at any instant t is the slope of $\theta(t)$ at t. Thus, for

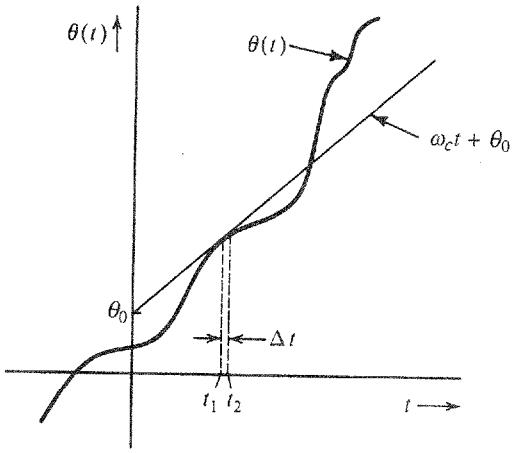

Figure 5.1 Concept of instantaneous frequency.

$\varphi(t)$ in Eq. (5.1), the instantaneous angular frequency and the generalized angle are related via

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

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

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

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

The instantaneous angular 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 angular 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 angular 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

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

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

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

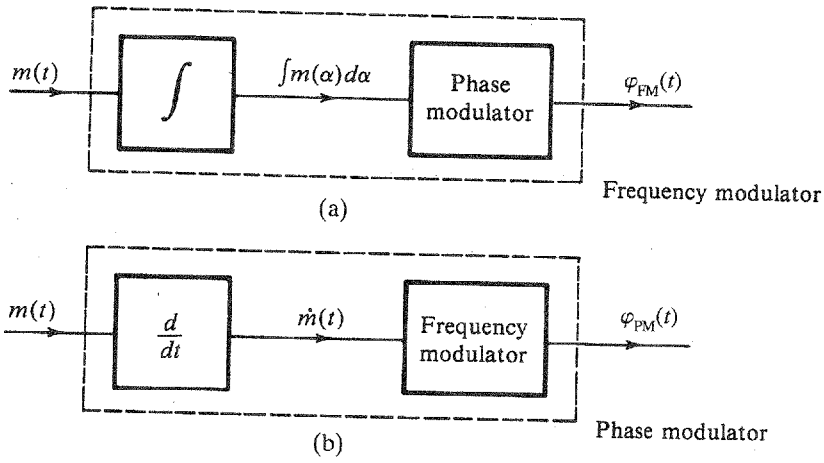

Figure 5.2 Phase and frequency modulation are equivalent and interchangeable.

Relationship between FM and PM

From Eqs. (5.3b) and (5.5), it is apparent that PM and FM not only are very similar but are inseparable. Replacing $m(t)$ in Eq. (5.3b) with $\int m(\alpha)d\alpha$ 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 only at an angle-modulated signal $\varphi(t)$, 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. It is analogous to asking a married man with children whether he is a father or a son. This discussion and Fig. 5.2 also show that we need not separately discuss methods of generation and demodulation of each type of modulation.

Equations (5.3b) and (5.5) 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 $m(t)$. As shown in Fig. 5.2b, a frequency modulator can be directly used to generate an FM signal or the message input $m(t)$ can be processed by a filter (differentiator) with transfer function $H(s) = s$ to generate PM signals. But why should we limit ourselves to these cases? We have an infinite number of possible ways of processing $m(t)$ before FM. If we restrict the choice to a linear operator, then a measure of $m(t)$ can be obtained as the output of an invertible linear (time-invariant) system with transfer function $H(s)$ or impulse response $h(t)$. The generalized angle-modulated carrier $\varphi_{EM}(t)$ can be expressed as

$$\varphi_{EM}(t) = A\cos[\omega_c t + \psi(t)] \tag{5.6a}$$

$$= A\cos\left[\omega_c t + \int_{-\infty}^{t} m(\alpha)\,h(t-\alpha)\,d\alpha\right] \tag{5.6b}$$

As long as $H(s)$ is a reversible operation (or invertible), $m(t)$ can be recovered from $\psi(t)$ by passing it through a system with transfer function $[H(s)]^{-1}$ as shown in Fig. 5.3. Now PM and FM are just two special cases with $h(t) = k_p\delta(t)$ and $h(t) = k_f u(t)$, respectively.

This 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 here in this chapter we shall primarily analyze FM, with occasional discussion of

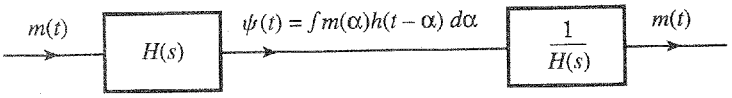

Figure 5.3 Generalized phase modulation by means of the filter $H(s)$ and recovery of the message from the modulated phase through the inverse filter $1/H(s)$.

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 pure PM nor by pure FM, but by something in between.

Power of an Angle-Modulated Wave

Although the instantaneous frequency and phase of an angle-modulated wave can vary with time, the amplitude A 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.

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.

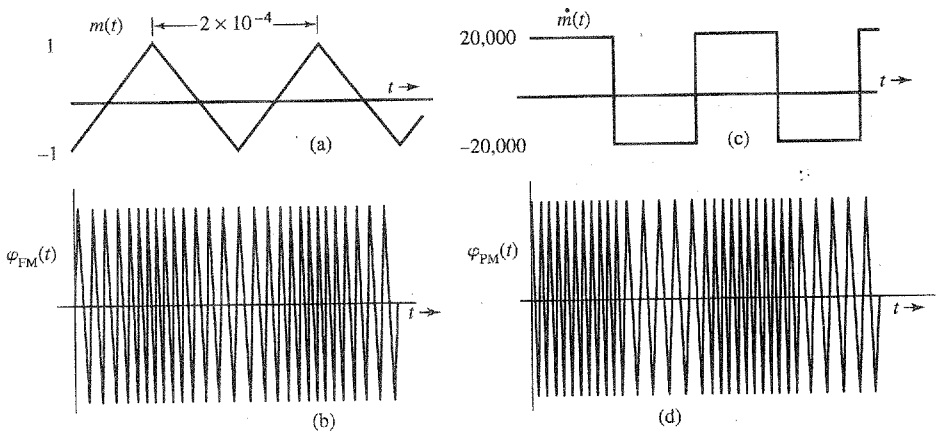

Figure 5.4 FM and PM waveforms.

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

$$\begin{aligned} f_i &= f_c + \frac{k_f}{2\pi} m(t) \\ &= 10^8 + 10^5 m(t) \\ (f_i)_{\min} &= 10^8 + 10^5 [m(t)]_{\min} = 99.9 \text{ MHz} \\ (f_i)_{\max} &= 10^8 + 10^5 [m(t)]_{\max} = 100.1 \text{ MHz} \end{aligned}$$

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).

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

For PM:

$$\begin{aligned} f_i &= f_c + \frac{k_p}{2\pi} \dot{m}(t) \\ &= 10^8 + 5\,\dot{m}(t) \\ (f_i)_{\min} &= 10^8 + 5\,[\dot{m}(t)]_{\min} = 10^8 - 10^5 = 99.9 \text{ MHz} \\ (f_i)_{\max} &= 10^8 + 5\,[\dot{m}(t)]_{\max} = 100.1 \text{ MHz} \end{aligned}$$

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, it means that the PM signal has sudden phase changes and, hence, $\dot{m}(t)$ contains impulses. This indirect method fails at *points of the discontinuity*. In such a case, a direct approach should be used at the point of discontinuity to specify the sudden phase changes. 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.

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 and 100.1 MHz, as shown in Fig. 5.5b. This scheme of carrier

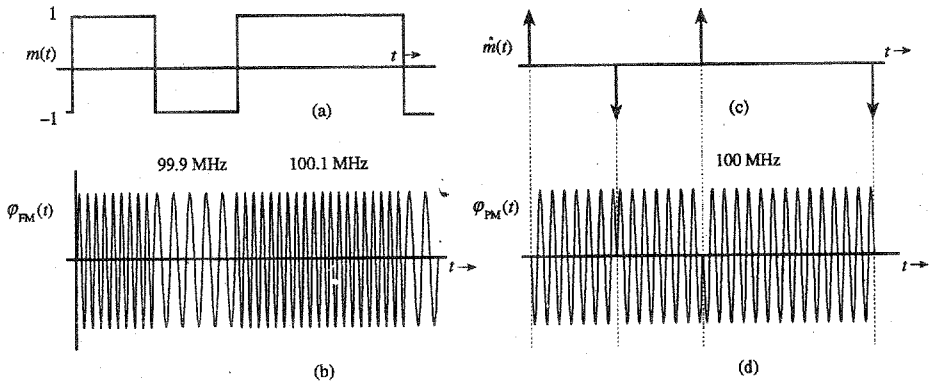

Figure 5.5 FM and PM waveforms.

frequency modulation by a digital signal (Fig. 5.5b) is called **frequency shift keying (FSK)** because information digits are transmitted by keying different frequencies (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) is zero except at points of discontinuity of $m(t)$ where impulses of strength ± 2 are present. This means that the frequency of the PM signal stays the same except at these isolated points of time! 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:

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

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

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

In this case $\varphi_{PM}(t) = A \sin \omega_c t$ when $m(t) = 1$ or $-1/3$. This will certainly cause ambiguity at the receiver when $A \sin \omega_c t$ is received. Specifically, the receiver cannot decide the exact value of $m(t)$. 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 $\varphi_{PM}(t)$ 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 time, and a phase $\varphi_o + 2n\pi$ will exhibit n additional carrier cycles in 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 also say that when $m(t)$ is band-limited, k_p has no restrictions.

5.2 BANDWIDTH OF ANGLE-MODULATED WAVES

Unlike AM, angle modulation is nonlinear and no properties of Fourier transform can be directly applied for its bandwidth analysis. To determine the bandwidth of an FM wave, let us define

$$a(t) = \int_{-\infty}^{t} m(\alpha)\, d\alpha \tag{5.7}$$

and define

$$\hat{\varphi}_{FM}(t) = A\, e^{j[\omega_c t + k_f a(t)]} = A e^{jk_f a(t)} e^{j\omega_c t} \tag{5.8a}$$

such that its relationship to the FM signal is

$$\varphi_{FM}(t) = \text{Re}\left[\hat{\varphi}_{FM}(t)\right] \tag{5.8b}$$

Expanding the exponential $e^{jk_f a(t)}$ of Eq. (5.8a) in power series yields

$$\hat{\varphi}_{FM}(t) = A\left[1 + jk_f a(t) - \frac{k_f^2}{2!} a^2(t) + \cdots + j^n \frac{k_f^n}{n!} a^n(t) + \cdots\right] e^{j\omega_c t} \tag{5.9a}$$

and

$$\varphi_{\text{FM}}(t) = \text{Re}\,[\hat{\varphi}_{\text{FM}}(t)]$$
$$= A\left[\cos\,\omega_c t - k_f a(t)\sin\,\omega_c t - \frac{k_f^2}{2!}a^2(t)\cos\,\omega_c t + \frac{k_f^3}{3!}a^3(t)\sin\,\omega_c t + \cdots\right] \quad (5.9b)$$

The modulated wave consists of an unmodulated carrier plus various amplitude-modulated terms, such as $a(t)\sin\,\omega_c t$, $a^2(t)\cos\,\omega_c t$, $a^3(t)\sin\,\omega_c t, \ldots$. The signal $a(t)$ is an integral of $m(t)$. If $M(f)$ is band-limited to B, $A(f)$ is also band-limited* to B. The spectrum of $a^2(t)$ is simply $A(f) * A(f)$ 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 $a(t)$, $a^2(t), \ldots, a^n(t), \ldots$, 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 bandwidth of an FM wave is theoretically infinite, for practical signals with bounded $|a(t)|$, $|k_f a(t)|$ will remain finite. Because $n!$ increases much faster than $|k_f a(t)|^n$, we have

$$\frac{k_f^n a^n(t)}{n!} \simeq 0 \qquad \text{for large } n$$

Hence, we shall see that most of the modulated-signal power resides in a finite bandwidth. This is the principal foundation of the bandwidth analysis for angle-modulations. There are two distinct possibilities in terms of bandwidths—narrowband FM and wideband FM.

Narrowband Angle Modulation Approximation

Unlike AM, angle modulations are nonlinear. The nonlinear relationship between $a(t)$ and $\varphi(t)$ is evident from the terms involving $a^n(t)$ in Eq. (5.9b). When k_f is very small such that

$$|k_f a(t)| \ll 1$$

then all higher order terms in Eq. (5.9b) are negligible except for the first two. We then have a good approximation

$$\varphi_{\text{FM}}(t) \approx A\left[\cos\,\omega_c t - k_f a(t)\sin\,\omega_c t\right] \quad (5.10)$$

This approximation is a linear modulation that has an expression similar to that of the AM signal with message signal $a(t)$. Because the bandwidth of $a(t)$ is B Hz, the bandwidth of $\varphi_{\text{FM}}(t)$ in Eq. (5.10) is $2B$ Hz according to the frequency-shifting property due to the term $a(t)\sin\,\omega_c t$. For this reason, the FM signal for the case of $|k_f a(t)| \ll 1$ is called **narrowband FM (NBFM)**. Similarly, the **narrowband PM (NBPM)** signal is approximated by

$$\varphi_{\text{PM}}(t) \approx A\left[\cos\,\omega_c t - k_p m(t)\sin\,\omega_c t\right] \quad (5.11)$$

NBPM also has the approximate bandwidth of $2B$.

---

* This is because integration is a linear operation equivalent to passing a signal through a transfer function $1/j2\pi f$. Hence, if $M(f)$ is band-limited to B, $A(f)$ must also be band-limited to B.

A comparison of NBFM [Eq. (5.10)] with AM [Eq. (5.9a)] brings out clearly the similarities and differences between the two types of modulation. Both have the same modulated bandwidth $2B$. 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 the apparent similarities, the AM and FM signals have very different waveforms. In an AM signal, the oscillation frequency is constant and the amplitude varies with time, whereas in an FM signal, the amplitude stays constant and the frequency varies with time.

Wideband FM (WBFM) Bandwidth Analysis: The Fallacy Exposed

Note that an FM signal is meaningful only if its frequency deviation is large enough. In other words, practical FM chooses the constant k_f large enough that the condition $|k_f a(t)| \ll 1$ is not satisfied. We call FM signals in such cases **wideband FM (WBFM)**. Thus, in analyzing the bandwidth of WBFM, we cannot ignore all the higher order terms in Eq. (5.9b). To begin, 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 (wideband) FM bandwidth.

Consider a low-pass $m(t)$ with bandwidth B Hz. This signal is well approximated by a staircase signal $\hat{m}(t)$, as shown in Fig. 5.6a. The signal $m(t)$ is now approximated by pulses of constant amplitude. For convenience, each of these pulses will be called a "cell." To ensure

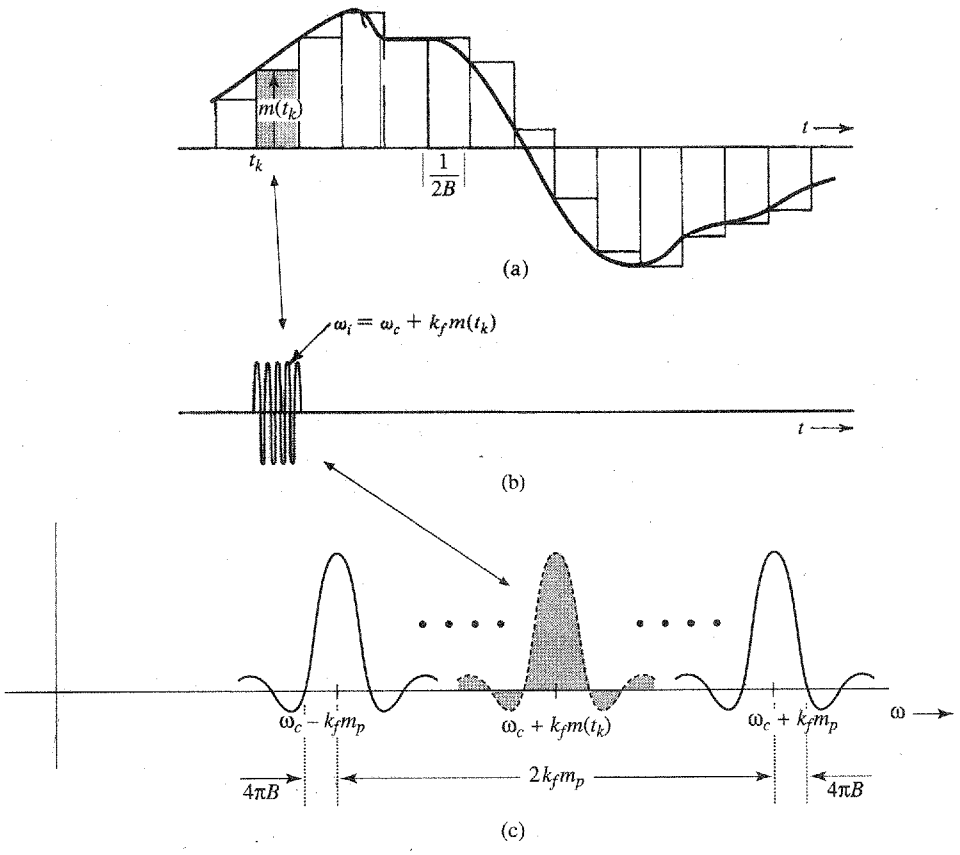

Figure 5.6 Estimation of FM wave bandwidth.