CHAPTER 1.3
MODULATION

Geoffrey C. Orsak, H. Vincent Poor, John B. Thomas

MODULATION THEORY

As discussed in Chap. 1.1 and shown in Fig. 1.1.1, the central problem in most communication systems is the transfer of information originating in some source to a destination by means of a channel. It will be convenient in this section to call the sent message or intelligence $a(t)$ and to denote the received message by $a^*(t)$, a distorted or corrupted version of $a(t)$.

The message signals used in communication and control systems are usually limited in frequency range to some maximum frequency $f_m = \omega_m/2\pi$ Hz. This frequency is typically in the range of a few hertz for control systems and moves upward to a few megahertz for television video signals. In addition the bandwidth of the signal is often of the order of this maximum frequency so that the signal spectrum is approximately low-pass in character. Such signals are often called *video signals* or *baseband signals*. It frequently happens that the transmission of such a spectrum through a given communication channel is inefficient or impossible. In this light, the problem may be looked upon as the one shown in Fig. 1.1.2, where an encoder E has been added between the source and the channel; however, in this case, the encoder acts to *modulate* the signal $a(t)$, producing at its output the *modulated wave* or signal $m(t)$.

Modulation can be defined as the modification of one signal, called the *carrier*, by another, called the *modulating signal*. The result of the modulation process is a modulated wave or signal. In most cases a frequency shift is one of the results. There are a number of reasons for producing modulated waves. The following list gives some of the major ones.

(***a***) *Frequency Translation for Efficient Antenna Design.* It may be necessary to transmit the modulating signal through space as electromagnetic radiation. If the antenna used is to radiate an appreciable amount of power, it must be large compared with the signal wavelength. Thus translation to higher frequencies (and hence to smaller wavelengths) will permit antenna structures of reasonable size and cost at both transmitter and receiver.

(***b***) *Frequency Translation for Ease of Signal Processing.* It may be easier to amplify and/or shape a signal in one frequency range than in another. For example, a dc signal may be converted to ac, amplified, and converted back again.

(***c***) *Frequency Translation to Assigned Location.* A signal may be translated to an assigned frequency band for transmission or radiation, e.g., in commercial radio broadcasting.

(***d***) *Changing Bandwidth.* The bandwidth of the original message signal may be increased or decreased by the modulation process. In general, decreased bandwidth will result in channel economies at the cost of fidelity. On the other hand, increased bandwidth will be accompanied by increased immunity to channel disturbances, as in wide-band frequency modulation or in spread-spectrum systems, for examples.

(***e***) *Multiplexing.* It may be necessary or desirable to transmit several signals occupying the same frequency range or the same time range over a single channel. Various modulation techniques allow the signals to share the same channel and yet be recovered separately. Such techniques are given the generic name of

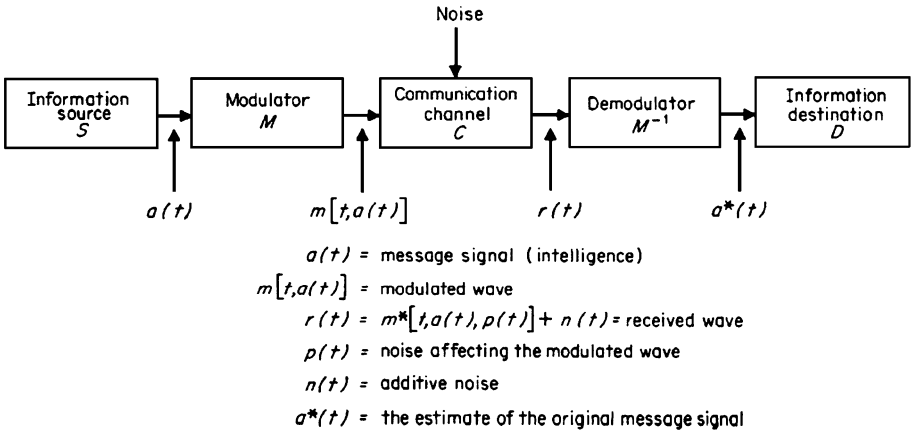


FIGURE 1.3.1 Communication system involving modulation and demodulation.

multiplexing. As will be discussed later, multiplexing is possible in either the frequency domain (frequency-domain multiplexing FDM) or in the time domain (time-domain multiplexing, TDM). As a simple example, the signals may be translated in frequency so that they occupy separate and distinct frequency ranges as mentioned in item (b).

Thus, the process of modulation can be considered as a form of encoding used to match the message signal arising from the information source to the communication channel. At the same time it is generally true that the channel itself has certain undesirable characteristics resulting in distortion of the signal during transmission. A part of such distortion can frequently be accounted for by postulating noise disturbances in the channel. These noises may be additive and may also affect the modulated wave in a more complicated fashion, although it is usually sufficient (and much simpler) to assume additive noise only. Also, the received signal must be decoded (demodulated) to recover the original signal.

In view of this discussion, it is convenient to change the block diagram of Fig. 1.1.2 to that shown in Fig. 1.3.1. The waveform received at the demodulator (receiver) will be denoted by $r(t)$, where

$$r(t) = m^*[t, a(t), p(t)] + n(t) \tag{1}$$

where $a(t)$ is the original message signal, $m[t, a(t)]$ is the modulated wave, $m^*[t, a(t), p(t)]$ is a corrupted version of $m[t, a(t)]$, and $p(t)$ and $n(t)$ are noises whose characteristics depend on the channel. Unless it is absolutely necessary for an accurate characterization of the channel, we will assume that $p(t) \equiv 0$ to avoid the otherwise complicated analysis that results.

The aim is to find modulators M and demodulators M^{-1} that make $a^*(t)$ a "good" estimate of the message signal $a(t)$. It should be emphasized that M^{-1} is not uniquely specified by M; for example, it is not intended to imply that $MM^{-1} = 1$. The form of the demodulator, for a given modulator, will depend on the characteristics of the message $a(t)$ and the channel as well as on the criterion of "goodness of estimation" used.

We now take up a study of the various forms of modulation and demodulation, their principal characteristics, their behavior in conjunction with noisy channels, and their advantages and disadvantages. We begin with some preliminary material or signals and their properties.

ELEMENTS OF SIGNAL THEORY

A real time function $f(t)$ and its Fourier transform form a Fourier transform pair given by

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt \tag{2}$$

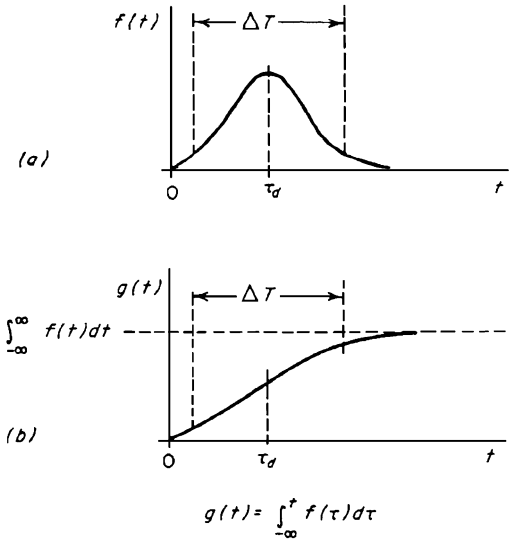


FIGURE 1.3.2 Duration and delay: (*a*) typical pulse; (*b*) integral of pulse.

and

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}\,d\omega \tag{3}$$

It follows directly from Eq. (2) that the transform $F(\omega)$ of a real-time function has an even-symmetric real part and an odd-symmetric imaginary part.

Consider the function $f(t)$ shown in Fig. 1.3.2*a*. This might be a pulsed signal or the impulse response of a linear system, for example. The time ΔT over which $f(t)$ is appreciably different from zero is called the *duration* of $f(t)$, and some measure, such as τ_d, of the center of the pulse is called the *delay* of $f(t)$. In system terms, the quantity ΔT is the system *response time* or *rise time*, and τ_d is the system delay. The integral of $f(t)$, shown in Fig. 1.3.2b, corresponds to the step-function response of a system with impulse response $f(t)$.

If the function $f(t)$ of Fig. 1.3.2 is nonnegative, the new function

$$\frac{f(t)}{\int_{-\infty}^{\infty} f(t)\,dt}$$

is nonnegative with unit area. We now seek measures of duration and delay that are both meaningful in terms of communication problems and mathematically tractable. It will be clear that some of the results we obtain will not be universally applicable and, in particular, must be used with care when the function $f(t)$ can be negative for some values of t; however, the results will be useful for wide classes of problems.

Consider now a frequency function $F(\omega)$, which will be assumed to be real. If $F(\omega)$ is not real, either $|F(\omega)|^2 = F(\omega)F(-\omega)$ or $|F(\omega)|$ can be used. Such a function might be similar to that shown in Fig. 1.3.3*a*. The radian frequency range $\Delta\Omega$ (or the frequency range ΔF) over which $F(\omega)$ is appreciably different from zero is called the *bandwidth* of the function. Of course, if the function is a *bandpass* function, such as that shown in Fig. 1.3.3*b*, the bandwidth will usually be taken to be some measure of the width of the positive-frequency

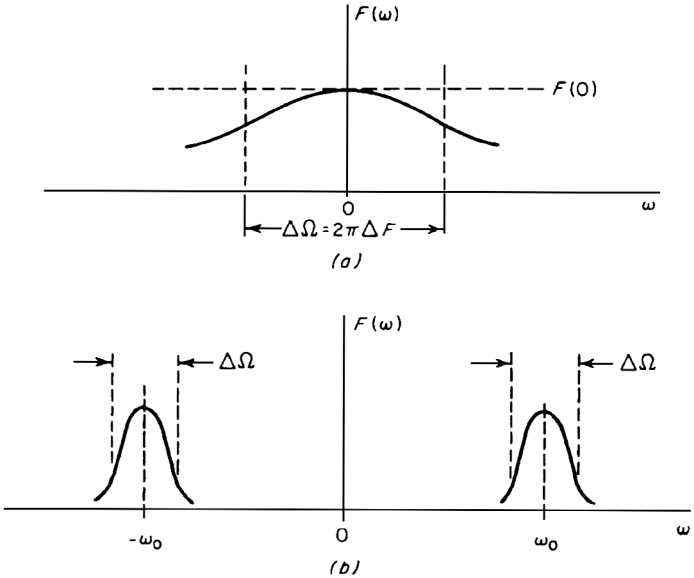


FIGURE 1.3.3 Illustrations of bandwidth: (*a*) typical low-pass frequency function; (*b*) typical bandpass frequency function.

(or negative-frequency) part of the function only. As in the case of the time function previously discussed, we may normalize to unit area and consider

$$\frac{F(\omega)}{\int_{-\infty}^{\infty} F(\omega)\,d\omega}$$

Again this new function is nonnegative with unit area.

Consider now the Fourier pair $f(t)$ and $F(\omega)$ and let us change the time scale by the factor a, replacing $f(t)$ by $af(at)$ so that both the old and the new signal have the same area, i.e.,

$$\int_{-\infty}^{\infty} f(t)\,dt = \int_{-\infty}^{\infty} af(at)\,dt \tag{4}$$

For $a < 1$, the new signal $af(at)$ is stretched in time and reduced in height; its "duration" has been increased. For $a > 1$, $af(at)$ has been compressed in time and increased in height; its "duration" has been decreased. The transform of this new function is

$$\int_{-\infty}^{\infty} af(at)e^{-j\omega t}\,dt = \int_{-\infty}^{\infty} f(x)e^{-j(\omega/a)x}\,dt = F\left(\frac{\omega}{a}\right) \tag{5}$$

The effect on the bandwidth of $F(\omega)$ has been the opposite of the effect on the duration of $f(t)$. When the signal duration is increased (decreased), the bandwidth is decreased (increased) in the same proportion. From the discussion, we might suspect that more fundamental relationships hold between properly defined durations and bandwidths of signals.

DURATION AND BANDWIDTH–UNCERTAINTY RELATIONSHIPS

It is apparent from the discussion above that treatments of duration and bandwidth are mathematically similar although one is defined in the time domain and the other in the frequency domain. Several specific measures of these two quantities will now be found, and it will be shown that they are intimately related to each other through various *uncertainty relationships*. The term "uncertainty" arises from the *Heisenberg uncertainty principle* of quantum mechanics, which states that it is not possible to determine simultaneously and exactly the position and momentum coordinates of a particle. More specifically, if Δx and Δp are the uncertainties in position and momentum, then

$$\Delta x\, \Delta p \geq h \tag{6}$$

where h is a constant. A number of inequalities of the form of Eq. (6) can be developed relating the duration ΔT of a signal to its (radian) bandwidth $\Delta\Omega$. The value of the constant h will depend on the definitions of duration and bandwidth.

Equivalent Rectangular Bandwidth DW1 and Duration DT1. The equivalent rectangular bandwidth DW1 of a frequency function F(w) is defined as

$$\Delta\Omega_1 = \frac{\int_{-\infty}^{\infty} F(\omega)\, d\omega}{F(\omega_0)} \tag{7}$$

where ω_0 is some characteristic center frequency of the function $F(\omega)$. It is clear from this definition that the original function $F(\omega)$ has been replaced by a rectangular function of equal area, width $\Delta\Omega_1$, and height $F(\omega_0)$. For the low-pass case ($\omega_0 \equiv 0$), it follows from Eqs. (2) and (3) that Eq. (7) can be rewritten

$$\Delta\Omega_1 = \frac{2\pi f(0)}{\int_{-\infty}^{\infty} f(t)\, dt} \tag{8}$$

where $f(t)$ is the time function which is the inverse Fourier transform of $F(\omega)$.

The same procedure can be followed in the time domain, and the *equivalent rectangular duration* ΔT_1 of the signal $f(t)$ can be defined by

$$\Delta T_1 = \frac{\int_{-\infty}^{\infty} f(t)\, dt}{f(t_0)} \tag{9}$$

where t_0 is some characteristic time denoting the center of the pulse. For the case where $t_0 \equiv 0$, it is clear, then, from Eqs. (8) and (9) that equivalent rectangular duration and bandwidth are connected by the uncertainty relationship

$$\Delta T_1\, \Delta\Omega_1 = 2\pi \tag{10}$$

Second-Moment Bandwidth DW2 and Duration DT2. An alternative uncertainty relationship is based on the second-moment properties of the Fourier pair $F(\omega)$ and $f(t)$.

A second-moment bandwidth $\Delta\Omega_2$ can be defined by

$$(\Delta\Omega_2)^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} (\omega - \bar{\omega})^2 |F(\omega)|^2\, d\omega/\varepsilon \tag{11}$$

and a *second-moment duration* ΔT_2 by

$$(\Delta T_2)^2 = \int_{-\infty}^{\infty} (t-\bar{t})^2 |f(t)|^2 dt/\varepsilon \tag{12}$$

where the total energy ϵ is given by

$$\epsilon = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(\omega)|^2 d\omega = \int_{-\infty}^{\infty} |f(t)|^2 dt = 1$$

These quantities are related by the inequality

$$\Delta\Omega_2 \, \Delta T_2 \geq 1/2 \tag{13}$$

This expression is a second uncertainty relationship connecting the bandwidth and duration of a signal. Many other such inequalities can be obtained.

CONTINUOUS MODULATION

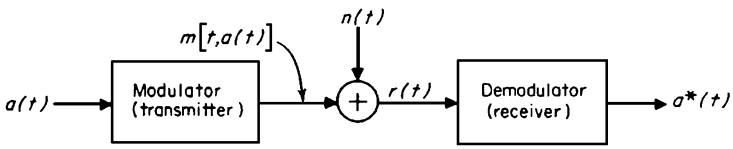
Modulation can be defined as the modification of one signal, called the *carrier*, by another, called the *modulation, modulating signal,* or *message signal*. In this section we will be concerned with situations where the carrier and the modulation are both continuous functions of time. Later we will treat the cases where the carrier and/or the modulation have the form of pulse trains.

For our analysis, Fig. 1.3.1 can be modified to the system shown in Fig. 1.3.4 in which the message is sent through a modulator (or transmitter) to produce the modulated continuous signal $m[t, a(t)]$. This waveform is corrupted by additive noise $n(t)$ in transmission so that the received (continuous) waveform $r(t)$ can be written

$$r(t) = m[t, a(t)] + n(t) \tag{14}$$

The purpose of the demodulator (or receiver) is to produce some best estimate $a^*(t)$ of the original message signal $a(t)$. As pointed out earlier a more general model of the transmission medium would allow corruption of the modulated waveform itself so that the received signal was of the form of Eq. (1). For example, in wireless systems, multiplicative disturbances can result because of multipath transmission or fading so that the received signal is of the form

$$r_1(t) = p(t)m[t, a(t)] + n(t) \tag{15}$$



$a(t)$ = message signal (intelligence)
$m[t,a(t)]$ = modulated signal
$r(t) = m[t,a(t)] + n(t)$ = received signal
$n(t)$ = additive noise
$a^*(t)$ = the estimate of the original message signal

FIGURE 1.3.4 Communication-system model for continuous modulation and demodulation.

where both $p(t)$ and $n(t)$ are noises. However, we shall not treat such systems, confining ourselves to the simpler additive-noise model of Fig. 1.3.4.

LINEAR, OR AMPLITUDE, MODULATION

In a general way, *linear* (*or amplitude*) *modulation* (AM) can be defined as a system where a *carrier wave* $c(t)$ has its amplitude varied linearly by some *message signal* $a(t)$. More precisely, a waveform is linearly modulated (or amplitude-modulated) by a given message $a(t)$ if the partial derivative of that waveform with respect to $a(t)$ is independent of $a(t)$. In other words, the modulated $m[t, a(t)]$ can be written in the form

$$m[t, a(t)] = a(t)c(t) + d(t) \tag{16}$$

where $c(t)$ and $d(t)$ are independent of $a(t)$. Now we have

$$\frac{\partial m[t, a(t)]}{\partial a(t)} = c(t) \tag{17}$$

and $c(t)$ will be called the *carrier*. In most of the cases we will treat, the waveform $d(t)$ will either be zero or will be linearly related to $c(t)$. It will be more convenient, therefore, to write Eq. (16) as

$$m[t, a(t)] = b(t)c(t) \tag{18}$$

where $b(t)$ will be either

$$b_1(t) \equiv 1 + a(t) \tag{19}$$

or

$$b_2(t) \equiv a(t) \tag{20}$$

Also, at present it will be sufficient to allow the carrier $c(t)$ to be of the form

$$c(t) = C \cos(\omega_0 t + \theta) \tag{21}$$

where C and ω_0 are constants and θ is either a constant or a random variable uniformly distributed on $(0, 2\pi)$.

Whenever Eq. (19) applies, it will be convenient to assume that $b_1(t)$ is nearly always nonnegative. This implies that if $a(t)$ is a deterministic signal, then

$$a(t) \geq -1 \tag{22}$$

It also implies that if $a(t)$ is a random process, the probability that $a(t)$ is less than -1 in any finite interval $(-T, T)$ is arbitrarily small; i.e.,

$$p[a(t) < -1] \leq \epsilon \ll 1 \qquad -T \leq t \leq T \tag{23}$$

The purpose of these restrictions on $b_1(t)$ is to ensure that the carrier is not overmodulated and that the message signal $a(t)$ is easily recovered by simple receivers.

We can take the Fourier transform of both sides of Eq. (18) to obtain an expression for the frequency spectrum $M(\omega)$ of the general form of the linear-modulated waveform $m[t, a(t)]$:

$$M(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} B(\omega - \nu) C(\nu)\, d\nu \tag{24}$$

where $B(\omega)$ and $C(\omega)$ are Fourier transforms of $b(t)$ and $c(t)$, respectively. If the carrier $c(t)$ is the sinusoid of Eq. (21), then

$$C(\omega) = \pi C e^{j\theta}(\omega - \omega_0) + \pi C e^{-j\theta}\delta(\omega + \omega_0) \tag{25}$$

and Eq. (24) becomes

$$M(\omega) = (C/2)e^{j\theta}B(\omega - \omega_0) + (C/2)e^{-j\theta}B(\omega + \omega_0) \tag{26}$$

Thus, for a sinusoidal carrier, linear modulation is essentially a symmetrical frequency translation of the message signal through an amount ω_0 rad/s, and no new signal components are generated. On the other hand, if $c(t)$ is not a sinusoid, so that the spectrum $C(\omega)$ has nonzero width, then $M(\omega)$ will represent a spreading and shaping as well as a translation of $B(\omega)$.

Suppose that the message signal $a(t)$ [and hence $b(t)$] is low-pass and strictly band-limited $(0, \omega_s)$ rad/s where $\omega_s < \omega_0$. Then the frequency spectrum of $b(t)$ obeys $B(\omega) = 0, |\omega| > \omega_s$, and $B(\omega - \omega_0)$ and $B(\omega + \omega_0)$ do not overlap. The spectrum $B(\omega + \omega_0)$ occupies only a range of positive frequencies while $B(\omega - \omega_0)$ occupies only negative frequencies.

The *envelope* of the modulated waveform $m[t, a(t)]$ is the magnitude $|b(t)|$. If, as mentioned earlier, the function $b(t)$ is restricted to be nonnegative almost always, then this envelope becomes just $b(t)$. In such cases, an *envelope detector* will uniquely recover $b(t)$ and hence $a(t)$. We now consider the common forms of simple amplitude, or linear, modulation.

DOUBLE-SIDEBAND AMPLITUDE MODULATION (DSBAM)

In this case the function $b(t)$ is given by Eq. (13) so that

$$m[t, a(t)] = C[1 + a(t)] \cos(\omega_0 t + \theta) \tag{27}$$

The transform of $b_1(t)$ is just

$$B_1(\omega) = \mathfrak{F}[1 + a(t)] = 2\pi\delta(\omega) + A(\omega) \tag{28}$$

where $A(\omega)$ is the Fourier transform of $a(t)$ and $\delta(\omega)$ is the Dirac delta function. It follows, therefore, from Eq. (26) that the frequency spectrum of $m[t, a(t)]$ is given by

$$M(\omega) = \frac{C}{2}e^{j\theta}[2\pi\delta(\omega - \omega_0) + A(\omega - \omega_0)] + \frac{C}{2}e^{-j\theta}[2\pi\delta(\omega + \omega_0) + A(\omega + \omega_0)] \tag{29}$$

Depending on the form of $a(t)$ a number of special cases can be distinguished, but in any case the spectrum is given by Eq. (26).

The simplest case is where $a(t)$ is the *periodic function* given by

$$a(t) = \eta \cos \omega_s t \qquad \omega_s < \omega_0 \text{ and } 0 \le \eta \le 1 \tag{30}$$

Then the frequency spectrum $A(\omega)$ is

$$A(\omega) = \pi\eta\delta(\omega - \omega_s) + \pi\eta\delta(\omega + \omega_s) \tag{31}$$

For convenience, let us take the phase angle θ to be zero so that

$$M(\omega) = \pi C\left[\delta(\omega - \omega_0) + \delta(\omega + \omega_0) + \frac{\eta}{2}\delta(\omega - \omega_0 - \omega_s) + \frac{\eta}{2}\delta(\omega - \omega_0 + \omega_s)\right.$$

$$\left. + \frac{\eta}{2}\delta(\omega + \omega_0 - \omega_s) + \frac{\eta}{2}\delta(\omega + \omega_0 + \omega_s)\right] \tag{32}$$

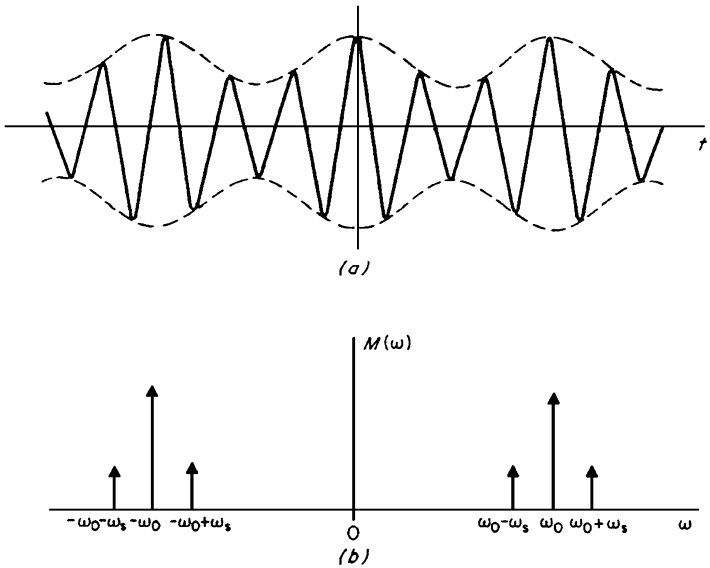


FIGURE 1.3.5 (*a*) Double-sideband AM signal and (*b*) frequency spectrum.

This spectrum is illustrated in Fig. 1.3.5 together with the corresponding modulated waveform

$$m[t, a(t)] = C(1 + \eta \cos \omega_s t) \cos \omega_0 t \tag{33}$$

Note that this last equation can also be written as

$$m[t, a(t)] = \underbrace{C \cos \omega_0 t}_{\text{carrier}} + \underbrace{(C/2)\eta \cos (\omega_0 - \omega_s)t}_{\text{lower sideband}} + \underbrace{(C/2)\eta \cos (\omega_0 + \omega_s)t}_{\text{upper sideband}} \tag{34}$$

In this form it is easy to distinguish the *carrier*, the *lower sideband,* and the *upper sideband*.

Suppose that $a(t)$ is not a single sinusoid but is periodic with period P. Then it may be expanded in a Fourier series and each term of the series treated as in Eq. (32).

DOUBLE-SIDEBAND AMPLITUDE MODULATION, SUPPRESSED CARRIER

If $b(t)$ is given by Eq. (20) so that

$$m[t, a(t)] = Ca(t) \cos (\omega_0 t + \theta) \tag{35}$$

then the carrier is suppressed and Eq. (29) reduces to

$$M(\omega) = (C/2)e^{j\theta} A(\omega - \omega_0) + (C/2)e^{-j\theta} A(\omega + \omega_0) \tag{36}$$

The principal advantage of this technique (DSBAM-SC) over DSBAM is that the carrier is not transmitted in the former case, with a consequent saving in transmittal power. The principal disadvantages relate to problems in generating and demodulating the suppressed-carrier waveform.

VESTIGIAL-SIDEBAND AMPLITUDE MODULATION (VSBAM)

It is apparent from Eq. (29) that the total information regarding the message signal $a(t)$ is contained in either the upper or lower sideband in conventional DSBAM. In principle the other sideband could be eliminated, say by filtering, with a consequent reduction in bandwidth and transmitted power. Such a procedure is actually followed in single-sideband amplitude modulation (SSBAM), discussed next. However, completely filtering out one sideband requires an ideal bandpass filter with infinitely sharp cutoff or an equivalent technique. A technique intermediate between the production of DSBAM and SSBAM is vestigial sideband amplitude modulation (VSBAM), where one sideband is attenuated much more than the other with some saving in bandwidth and transmitted power. The carrier may be present or suppressed.

The principal use of VSBAM has been in commercial television. The video (picture) signal is transmitted by VSBAM with a consequent reduction in the total transmitted signal bandwidth and in the frequency difference that must be allowed between adjacent channels.

SINGLE-SIDEBAND AMPLITUDE MODULATION (SSBAM)

This important type of AM can be considered as a limiting form of VSBAM when the filter for the modulated waveform is an ideal filter with infinitely sharp cutoff so that one sideband, e.g., the lower, is completely eliminated.

BANDWIDTH AND POWER RELATIONSHIPS FOR AM

It is clear that the bandwidth of a given AM signal is related in a simple fashion to the bandwidth of the modulating signal since AM is essentially a frequency translation. If the modulating signal $a(t)$ is assumed to be low-pass and to have a bandwidth of W Hz, then the bandwidth ΔF of the modulated signal $m[t, a(t)]$ must satisfy

$$W \le \Delta F \le 2W \tag{37}$$

The upper limit of $2W$ Hz holds for double-sideband modulation and the lower limit of W Hz for single-sideband. For the concept of bandwidth to be meaningful for a single-frequency sinusoidal modulating signal, it is assumed, of course, that a low-frequency sinusoid of frequency W Hz has a bandwidth W. Actually, what is really being assumed is that this sinusoid is the highest-frequency component in the low-pass modulating signal.

The only case where intermediate values of the inequality of Eq. (37) are encountered is in VSBAM. In principle any bandwidth between W and $2W$ Hz is possible. In this respect SSBAM can be considered as a limiting case of VSBAM. From a practical point of view, however, it is difficult (or expensive) to design a filter whose gain magnitude drops off too rapidly.

Power relationships are also simple and straightforward for AM signals. Consider Eq. (27) which is a general expression for the modulated AM waveform $m(t)$. Let the average power in this signal be denoted by P_{av}. The phase angle θ will be fixed if $a(t)$ is deterministic and will be taken to be a random variable uniformly distributed on $(0, 2\pi)$ if $a(t)$ is a random process. Also, it will be assumed that $a(t)_{\text{av}}$ is zero. It is clear that P_{av} is given by

$$P_{\text{av}} = (C^2/2)[1 + a^2(t)_{\text{av}}] \tag{38}$$

where $a^2(t)_{\text{av}}$ is the average power in $a(t)$. The first term, $C^2/2$, is the carrier power, and the second term, $C^2/2)a^2(t)_{\text{av}}$, is the signal power in the upper and lower sidebands. If $|a(t)| \le 1$ to prevent overmodulation, then at least half the transmitted power is carrier power and the remainder is divided equally between the upper and lower sideband.

In double-sideband, suppressed-carrier operation, the carrier power is zero, and half the total power exists in each sideband. The fraction of information-bearing power is $^1/_2$. For single-sideband, suppressed-carrier systems,

all the transmitted power is information-bearing, and, in this sense, SSB-SC has maximum transmission efficiency.

The disadvantages of both suppressed-carrier and single-sideband operation lie in the difficulties of generating the signals for transmission and in the more complicated receivers required. Demodulation of suppressed-carrier AM signals involve the reinsertion of the carrier or an equivalent operation. The local generation of sinusoid at the exact frequency and phase of the missing carrier is either difficult or impossible unless a pilot tone of reduced magnitude is transmitted with the modulated signal for synchronization purposes or unless some nonlinear operation is performed on the suppressed-carrier signal to regenerate the carrier term at the receiver. In SSBAM, not only is the receiver more complicated, but transmission is considerably more difficult. It is usually necessary to generate the SSB signal at a low power level and then to amplify with a linear power amplifier to the proper level for transmission. On the other hand, DSB signals are easily generated at high power levels so that inefficient linear power amplifiers need not be used.

ANGLE (FREQUENCY AND PHASE) MODULATION

In angle modulation, the carrier $c(t)$ has either its phase angle or its frequency varied in accordance with the intelligence $a(t)$.

The result is not a simple frequency translation, as with AM, but involves both translation and the production of entirely new frequency components. In general, the new spectrum is much wider than that of the intelligence $a(t)$. The greater bandwidth may be used to improve the signal-to-noise performance of the receiver. This ability to *exchange* bandwidth for signal-to-noise enhancement is one of the outstanding characteristics and advantages of angle modulation.

In the form of angle modulation, which will be called *phase modulation* (PM), the phase of the carrier is varied linearly with the intelligence $a(t)$. Thus, the modulated signal is given by

$$m(t) = C \cos [\omega_0 t + \theta + k_p a(t)] \tag{39}$$

where k_p is a constant and the *modulation index* $\varnothing_m$ is defined by

$$\varnothing_m = \max |k_p a(t)| \qquad \text{(rad)} \tag{40}$$

In *frequency modulation*, the instantaneous frequency is made proportional to the intelligence $a(t)$. The modulated signal is given by

$$m(t) = C \cos\left[\omega_0 t + k_f \int_{-\infty}^{t} a(\tau)\, d\tau \right] \tag{41}$$

where k_f is a constant. The *maximum deviation* $\Delta\omega$ is given by

$$\Delta\omega = \max |k_f a(t)| \qquad \text{(rad/s)} \tag{42}$$

and, as before, a *modulation index* $\varnothing_m$ by

$$\varnothing_m = \max \left| k_f \int_{-\infty}^{t} a(\tau)\, d\tau \right| \tag{43}$$

In general, the analysis of angle-modulated signals is difficult even for simple modulating intelligence. We will consider only the case where the modulating intelligence $a(t)$ is a sinusoid. Let $a(t)$ be given by

$$a(t) = \eta \cos \omega_s t \qquad \omega_s < \omega_0 \tag{44}$$

Then the corresponding PM signal is

$$m(t) = C \cos (\omega_0 t + k_p \eta \cos \omega_s t) \tag{45}$$

where the phase angle θ has been set equal to zero. In the same way, the FM signal is

$$m(t) = C \cos [\omega_0 t + (k_f \eta / \omega_s) \sin \omega_s t] \tag{46}$$

Let us now consider the FM signal of this last equation. Essentially the same results will be obtained for PM. The equation can be expanded to yield

$$m(t) = C \cos (\varnothing_m \sin \omega_s t) \cos \omega_0 t - C \sin (\varnothing_m \sin \omega_s t) \sin \omega_0 t \tag{47}$$

where the modulation index $\varnothing_m$ is given by $\varnothing_m = k_f \eta / \omega_s$. Sinusoids with sinusoidal arguments give rise to Bessel functions, and Eq. (47) can be expanded and rearranged to obtain

$$m(t) = \underbrace{CJ_0(\varnothing_m)\cos \omega_0 t}_{\text{carrier}} + C\sum_{n=1}^{\infty} J_{2n}(\varnothing_m) \underbrace{[\cos(\omega_0 + 2n\omega_s)t}_{\text{USB}} + \underbrace{\cos(\omega_0 - 2n\omega_s)t]}_{\text{LSB}}$$

$$+ C\sum_{n=1}^{\infty} J_{2n-1}(\varnothing_m) \underbrace{\{\cos[\omega_0 + (2n-1)\omega_s]t}_{\text{USB}} - \underbrace{\cos[\omega_0 - (2n-1)\omega_s]t\}}_{\text{LSB}} \tag{48}$$

where $J_m(x)$ is the Bessel function of the first kind and order m. This expression for $m(t)$ is relatively complicated even though the modulating intelligence is a simple sinusoid. In addition to the carrier, there is an infinite number of upper and lower sidebands separated from the carrier (and from each other) by integral multiples of the modulating frequency ω_s. Each sideband, and the carrier, has an amplitude determined by the appropriate Bessel function. When there is more than one modulating sinusoid, the complexity increases rapidly.

In the general case, an infinite number of sidebands exist dispersed throughout the whole frequency domain. In this sense, the bandwidth of an FM (or PM) signal is infinite. However, outside of some interval centered on ω_0, the magnitude of the sidebands will be negligible; this interval may be taken as a practical measure of the bandwidth. An approximation can be obtained by noting that $J_n(\varnothing_m)$, considered as a function of n, decreases rapidly when $n < \varnothing_m$. Therefore only the first $\varnothing_m$ sidebands are significant. If the highest-frequency sideband of significance is $\varnothing_m \omega_s$, then the bandwidth is given approximately by

$$BW \approx 2\omega_s \varnothing_m \qquad \text{(rad/s)} \tag{49}$$

A more accurate rule of thumb is the slightly revised expression

$$BW \approx 2\omega_s(\varnothing_m + 1) \qquad \text{(rad/s)} \tag{50}$$

which may be considered a good approximation when $\varnothing_m \geq 5$.