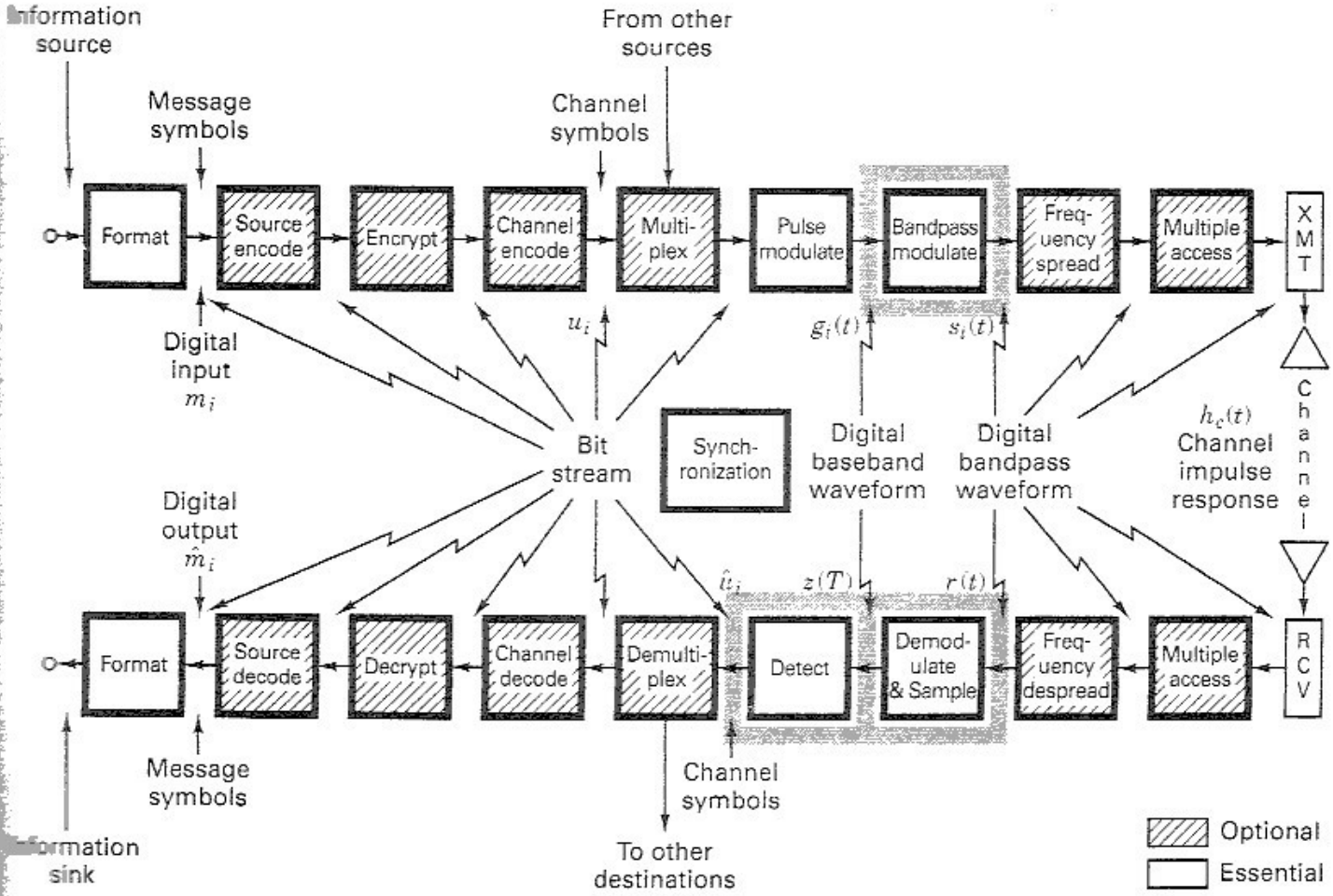
Bandpass Modulation and Demodulation



4.1 WHY MODULATE?

Digital modulation is the process by which digital symbols are transformed into waveforms that are compatible with the characteristics of the channel. In the case of baseband modulation, these waveforms usually take the form of shaped pulses. But in the case of *bandpass modulation* the shaped pulses modulate a sinusoid called a *carrier wave,* or simply a *carrier;* for radio transmission the carrier is converted to an electromagnetic (EM) field for propagation to the desired destination. One might ask why it is necessary to use a carrier for the radio transmission of baseband signals. The answer is as follows. The transmission of EM fields through space is accomplished with the use of antennas. The size of the antenna depends on the wavelength λ and the application. For cellular telephones, antennas are typically $\lambda/4$ in size, where wavelength is equal to c/f, and c, the speed of light, is 3×10^8 m/s. Consider sending a baseband signal (say, $f = 3000$ Hz) by coupling it to an antenna directly without a carrier wave. How large would the antenna have to be? Let us size it by using the telephone industry benchmark of $\lambda/4$ as the antenna dimension. For the 3,000 Hz baseband signal, $\lambda/4 = 2.5 \times 10^4$ $m \approx 15$ miles. To transmit a 3,000 Hz signal through space, *without carrier-wave modulation,* an antenna that spans 15 miles would be required. However, if the baseband information is first modulated on a higher frequency carrier, for example a 900 MHz carrier, the equivalent antenna diameter would be about 8 cm. For this reason, carrier-wave or bandpass modulation is an essential step for all systems involving radio transmission.

Bandpass modulation can provide other important benefits in signal transmission. If more than one signal utilizes a single channel, modulation may be used to separate the different signals. Such a technique, known as *frequency-division multiplexing*, is discussed in Chapter 11. Modulation can be used to minimize the effects of interference. A class of such modulation schemes, known as *spread-spectrum modulation*, requires a system bandwidth much larger than the minimum bandwidth that would be required by the message. The trade-off of bandwidth for interference rejection is considered in Chapter 12. Modulation can also be used to place a signal in a frequency band where design requirements, such as filtering and amplification, can be easily met. This is the case when radio-frequency (RF) signals are converted to an intermediate frequency (IF) in a receiver.

4.2 DIGITAL BANDPASS MODULATION TECHNIQUES

Bandpass modulation (either analog or digital) is the process by which an information signal is converted to a sinusoidal waveform; for digital modulation, such a sinusoid of duration T is referred to as a digital symbol. The sinusoid has just three features that can be used to distinguish it from other sinusoids: amplitude, frequency, and phase. Thus bandpass modulation can be defined as the process whereby the amplitude, frequency, or phase of an RF carrier, or a combination of them, is varied in accordance with the information to be transmitted. The general form of the carrier wave is

$$s(t) = A(t) \cos \theta(t) \tag{4.1}$$

where $A(t)$ is the time-varying amplitude and $\theta(t)$ is the time-varying angle. It is convenient to write

$$\theta(t) = \omega_0 t + \phi(t) \tag{4.2}$$

so that

$$s(t) = A(t) \cos [\omega_0 t + \phi(t)] \tag{4.3}$$

where ω_0 is the *radian frequency* of the carrier and $\phi(t)$ is the *phase*. The terms f and ω will each be used to denote frequency. When f is used, frequency in hertz is intended; when ω is used, frequency in radians per second is intended. The two frequency parameters are related by $\omega = 2\pi f$.

The basic *bandpass modulation/demodulation* types are listed in Figure 4.1. When the receiver exploits knowledge of the carrier's phase to detect the signals, the process is called *coherent detection;* when the receiver does not utilize such phase reference information, the process is called *noncoherent detection*. In digital communications, the terms *demodulation* and *detection* are often used interchangeably, although demodulation emphasizes waveform recovery, and detection entails the process of symbol decision. In ideal coherent detection, there is available at the receiver a prototype of each possible arriving signal. These prototype waveforms attempt to duplicate the transmitted signal set in every respect, even RF phase. The

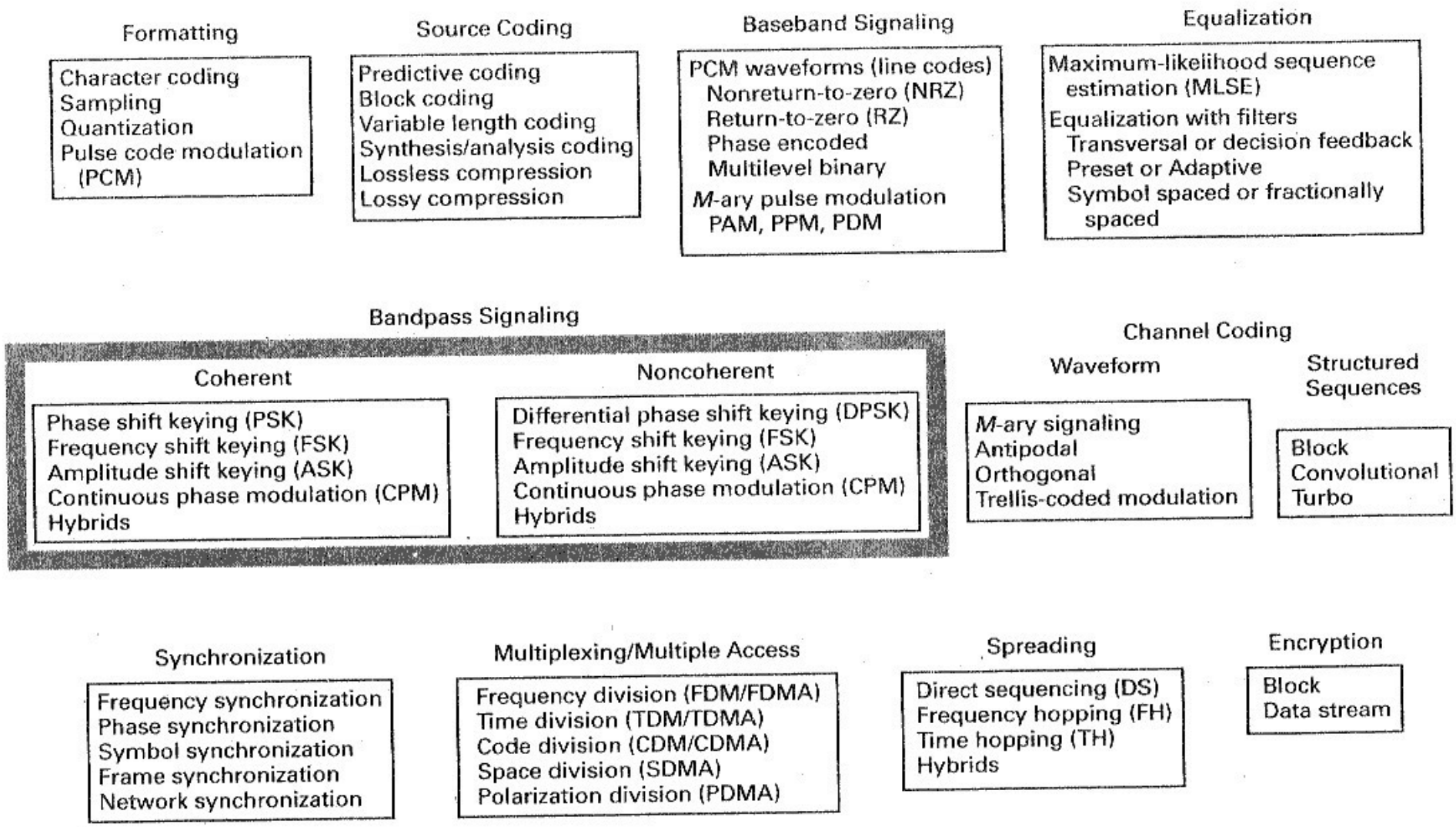


Figure 4.1 Basic digital communication transformations.

receiver is then said to be *phase locked* to the incoming signal. During demodulation, the receiver multiplies and integrates (correlates) the incoming signal with each of its prototype replicas. Under the heading of coherent modulation/demodulation in Figure 4.1 are listed phase shift keying (PSK), frequency shift keying (FSK), amplitude shift keying (ASK), continuous phase modulation (CPM), and hybrid combinations. The basic bandpass modulation formats are discussed in this chapter. Some specialized formats, such as offset quadrature PSK (OQPSK), minimum shift keying (MSK) belonging to the CPM class, and quadrature amplitude modulation (QAM), are treated in Chapter 9.

Noncoherent demodulation refers to systems employing demodulators that are designed to operate without knowledge of the absolute value of the incoming signal's phase; therefore, phase estimation is not required. Thus the advantage of noncoherent over coherent systems is reduced complexity, and the price paid is increased probability of error (P_E). In Figure 4.1 the modulation/demodulation types that are listed in the noncoherent column, DPSK, FSK, ASK, CPM, and hybrids, are similar to those listed in the coherent column. We had implied that phase information is not used for noncoherent reception; how do you account for the fact that there is a form of phase shift keying under the noncoherent heading? It turns out that an important form of PSK can be classified as noncoherent (or differentially coherent) since it does not require a reference in phase with the received carrier. This "pseudo-PSK," termed *differential PSK* (DPSK), utilizes phase information of the prior symbol as a phase reference for detecting the current symbol. This is described in Sections 4.5.1 and 4.5.2.

4.2.1 Phasor Representation of a Sinusoid

Using a well-known trigonometric identity called Euler's theorem, we introduce the complex notation of a sinusoidal carrier wave as follows:

$$e^{j\omega_0 t} = \cos \omega_0 t + j \sin \omega_0 t \tag{4.4}$$

One might be more comfortable with the simpler, more straightforward notation $\cos \omega_0 t$ or $\sin \omega_0 t$. What possible benefit can there be with the complex notation? We will see (in Section 4.6) that this notation facilitates our description of how real-world modulators and demodulators are implemented. For now, let us point to the general benefits of viewing a carrier wave in the complex form of Equation (4.4).

First, within this compact from, $e^{j\omega_0 t}$, is contained the two important quadrature components of any sinusoidal carrier wave, namely the inphase (real) and the quadrature (imaginary) components that are orthogonal to each other. Second, the unmodulated carrier wave is conveniently represented in a polar coordinate system as a unit vector or phasor rotating counterclockwise at the constant rate of ω_0 radians/s, as depicted in Figure 4.2. As time is increasing (i.e., from t_0 to t_1) we can visualize the time-varying projections of the rotating phasor on the inphase (I) axis and the quadrature (Q) axis. These cartesian axes are usually referred to as the I channel and Q channel respectively, and the projections on them represent the

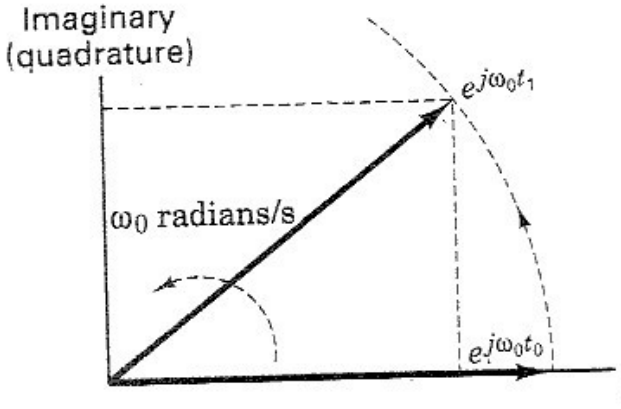


Figure 4.2 Phasor representation of a sinusoid.

signal components (orthogonal to each other) associated with those channels. Third, when it comes time to modulate the carrier wave with information, we can view this modulation as a methodical perturbation of the rotating phasor (and its projections).

For example, consider a carrier wave that is *amplitude modulated* (AM) with a sinusoid having an amplitude of unity and a frequency ω_m, where $\omega_m \ll \omega_0$. The analytical form of the transmitted waveform is

$$s(t) = \mathrm{Re}\left\{ e^{j\omega_0 t} \left(1 + \frac{e^{j\omega_m t}}{2} + \frac{e^{-j\omega_m t}}{2} \right) \right\} \tag{4.5}$$

where Re{x} is the real part of the complex quantity {x}. Figure 4.3 illustrates that the rotating phasor $e^{j\omega_0 t}$ of Figure 4.2 is now perturbed by two sideband terms—$e^{j\omega_m t}/2$ rotating counterclockwise and $e^{-j\omega_m t}/2$ rotating clockwise. The sideband phasors are rotating at a much slower speed than the carrier-wave phasor. The net result of the composite signal is that the rotating carrier-wave phasor now appears to be growing longer and shorter pursuant to the dictates of the sidebands, but its frequency stays constant—hence, the term "amplitude modulation."

Another example to reinforce the usefulness of the phasor view is that of *frequency modulating* (FM) the carrier wave with a similar sinusoid having a frequency of ω_m radians/s. The analytical representation of *narrowband* FM (NFM) has an appearance similar to AM and is represented by

$$s(t) = \mathrm{Re}\left\{ e^{j\omega_0 t} \left(1 - \frac{\beta}{2} e^{-j\omega_m t} + \frac{\beta}{2} e^{j\omega_m t} \right) \right\} \tag{4.6}$$

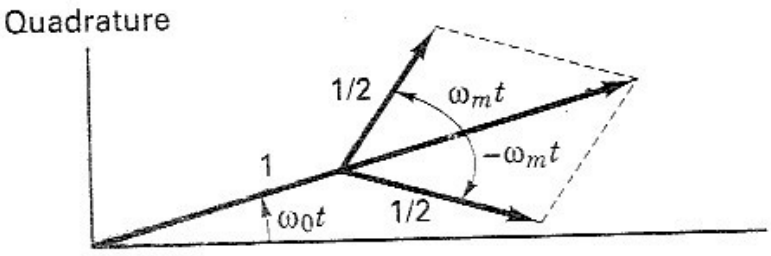


Figure 4.3 Amplitude modulation.

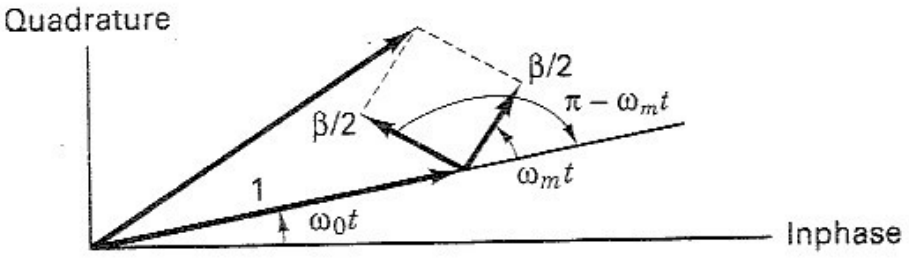


Figure 4.4 Narrowband frequency modulation.

where β is the modulation index [1]. Figure 4.4 illustrates that the rotating carrier-wave phasor is again perturbed by two sideband terms, but because one of the sideband terms carries a minus sign in Equation (4.6), the clockwise and counterclockwise rotating sideband phasors have a different symmetry than in the case of AM. In the case of AM the sideband symmetry results in the carrier-wave phasor growing longer and shorter with time. In NFM, the sideband symmetry (90° different than AM) results in the carrier-wave phasor speeding up and slowing down according to the dictates of the sidebands, but the amplitude says essentially constant—hence, the term "frequency modulation."

Figure 4.5 illustrates examples of the most common digital modulation formats: PSK, FSK, ASK, and a hybrid combination of ASK and PSK (ASK/PSK or APK). The first column lists the analytic expression, the second is a typical pictorial of the waveform versus time, and the third is a vector (or phasor) schematic, with the orthogonal axes labeled $\{\psi_j(t)\}$. In the general M-ary signaling case, the processor accepts k source bits (or channel bits if there is coding) at a time and instructs the modulator to produce one of an available set of $M = 2^k$ waveform types. Binary modulation, where $k = 1$, is just a special case of M-ary modulation.

In Figure 4.2, we represented a carrier wave as a phasor rotating in a plane at the speed of the carrier-wave frequency ω_0 radians/s. In Figure 4.5, the phasor schematic for each digital-modulation example represents a constellation of information signals (vectors or points in the signaling space), where time is not represented. In other words, the constantly rotating aspect of the unmodulated carrier wave has been removed, and only the information-bearing phasor positions, relative to one another, are presented. Each example in Figure 4.5 uses a particular value of M, the set size.

4.2.2 Phase Shift Keying

Phase shift keying (PSK) was developed during the early days of the deep-space program; PSK is now widely used in both military and commercial communications systems. The general analytic expression for PSK is

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos\left[\omega_0 t + \phi_i(t)\right] \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.7}$$

where the phase term, $\phi_i(t)$, will have M discrete values, typically given by

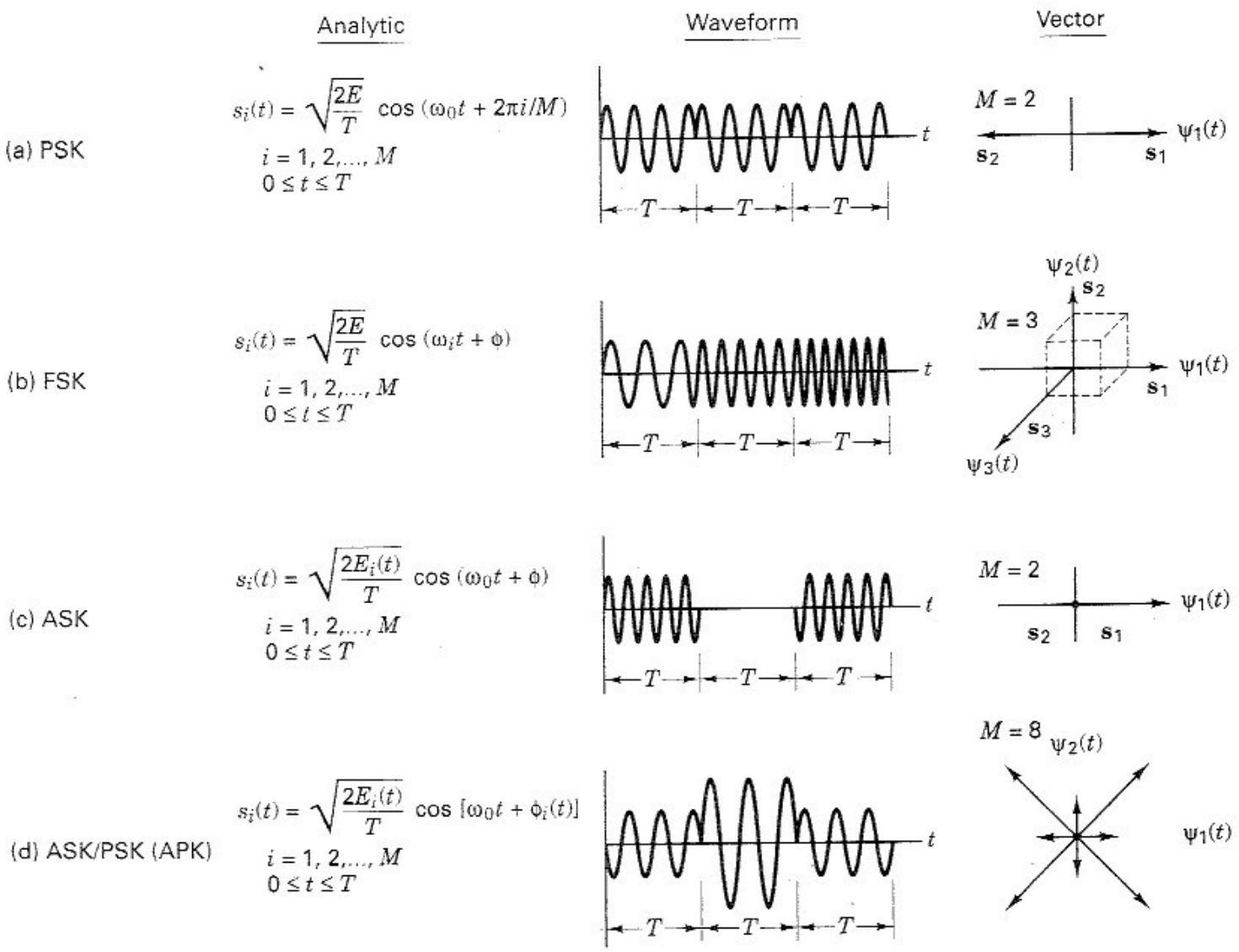


Figure 4.5 Digital modulations. (a) PSK. (b) FSK. (c) ASK. (d) ASK/PSK (APK).

$$\phi_i(t) = \frac{2\pi i}{M} \qquad i = 1, \ldots, M$$

For the binary PSK (BPSK) example in Figure 4.5a, M is 2. The parameter E is symbol energy, T is symbol time duration, and $0 \le t \le T$. In BPSK modulation, the modulating data signal shifts the phase of the waveform $s_i(t)$ to one of two states, either zero or π (180°). The waveform sketch in Figure 4.5a shows a typical BPSK waveform with its abrupt phase changes at the symbol transitions; if the modulating date stream were to consist of alternating ones and zeros, there would be such an abrupt change at each transition. The signal waveforms can be represented as vectors or phasors on a polar plot; the vector length corresponds to the signal amplitude, and the vector direction for the general M-ary case corresponds to the signal phase relative to the other $M - 1$ signals in the set. For the BPSK example, the vector picture illustrates the two 180° opposing vectors. Signal sets that can be depicted with such opposing vectors are called *antipodal signal sets.*

4.2.3 Frequency Shift Keying

The general analytic expression for FSK modulation is

$$s_i(t) = \sqrt{\frac{2E}{T}} \cos(\omega_i t + \phi) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.8}$$

where the frequency term ω_i has M discrete values, and the phase term ϕ is an arbitrary constant. The FSK waveform sketch in Figure 4.5b illustrates the typical frequency changes at the symbol transitions. At the symbol transitions, the figure depicts a gentle shift from one frequency (tone) to another. This behavior is only true for a special class of FSK called continuous-phase FSK (CPFSK) which is described in Section 9.8. In the general MFSK case, the change to a different tone can be quite abrupt, because there is no requirement for the phase to be continuous. In this example, M has been chosen equal to 3, corresponding to the same number of waveform types (3-ary); note that this $M = 3$ choice for FSK has been selected to emphasize the mutually perpendicular axes. In practice, M is usually a nonzero power of 2 (2, 4, 8, 16, . . .). The signal set is characterized by Cartesian coordinates, such that each of the mutually perpendicular axes represents a sinusoid with a different frequency. As described earlier, signal sets that can be characterized with such mutually perpendicular vectors are called *orthogonal* signals. Not all FSK signaling is orthogonal. For any signal set to be orthogonal, it must meet the criterion set forth in Equation (3.69). For an FSK signal set, in the process of meeting this criterion, a condition arises on the spacing between the tones in the set. The necessary frequency spacing between tones to fulfill the orthogonality requirement is discussed in Section 4.5.4.

4.2.4 Amplitude Shift Keying

For the ASK example in Figure 4.5c, the general analytic expression is

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cos(\omega_0 t + \phi) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.9}$$

where the amplitude term $\sqrt{2E_i(t)/T}$ will have M discrete values, and the phase term ϕ is an arbitrary constant. In Figure 4.5c, M has been chosen equal to 2, corresponding to two waveform types. The ASK waveform sketch in the figure can describe a radar transmission example, where the two signal amplitude states would be $\sqrt{2E/T}$ and zero. The vector picture utilizes the same phase–amplitude polar coordinates as the PSK example. Here we see a vector corresponding to the maximum-amplitude state, and a point at the origin corresponding to the zero-amplitude state. Binary ASK signaling (also called on–off keying) was one of the earliest forms of digital modulation used in radio telegraphy at the beginning of this century. Simple ASK is no longer widely used in digital communications systems, and thus it will not be treated in detail here.

4.2.5 Amplutude Phase Keying

For the combination of ASK and PSK (APK) example in Figure 4.5d, the general analytic expression

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cos\left[\omega_0 t + \phi_i(t)\right] \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.10}$$

illustrates the indexing of both the signal amplitude term and the phase term. The APK waveform picture in Figure 4.5d illustrates some typical simultaneous phase and amplitude changes at the symbol transition times. For this example, M has been chosen equal to 8, corresponding to eight waveforms (8-ary). The figure illustrates a hypothetical eight-vector signal set on the phase-amplitude plane. Four of the vectors are at one amplitude, and the other four vectors are at a different amplitude. Each of the vectors is separated by 45°. When the set of M symbols in the two-dimensional signal space are arranged in a rectangular constellation, the signaling is referred to as quadrature amplitude modulation (QAM); examples of QAM are considered in Chapter 9.

The vector picture for each of the modulation types described in Figure 4.5 (except the FSK case) is characterized on a plane whose *polar* coordinates represent signal *amplitude* and *phase*. The FSK case assumes orthogonal FSK (see Section 4.5.4) and is characterized in a *Cartesian* coordinate space, with each axis representing a *frequency tone* ($\cos \omega_i t$) from the M-ary set of orthogonal tones.

4.2.6 Waveform Amplitude Coefficient

The waveform amplitude coefficient appearing in Equations (4.7) to (4.10) has the same general form $\sqrt{2E/T}$ for all modulation formats. The derivation of this expression begins with

$$s(t) = A \cos \omega t \tag{4.11}$$

where A is the peak value of the waveform. Since the peak value of a sinusoidal waveform equals $\sqrt{2}$ times the root-mean-square (rms) value, we can write

$$\begin{aligned} s(t) &= \sqrt{2} A_{\text{rms}} \cos \omega t \\ &= \sqrt{2A_{\text{rms}}^2} \cos \omega t \end{aligned}$$

Assuming the signal to be a voltage or a current waveform, A_{rms}^2 represents average power P (normalized to 1 Ω). Therefore, we can write

$$s(t) = \sqrt{2P} \cos \omega t \tag{4.12}$$

Replacing P watts by E joules/T seconds, we get

$$s(t) = \sqrt{\frac{2E}{T}} \cos \omega t \tag{4.13}$$

We shall use either the amplitude notation A in Equation (4.11) or the designation $\sqrt{2E/T}$ in Equation (4.13). Since the *energy* of a received signal is the key parameter in determining the error performance of the detection process, it is often more convenient to use the amplitude notation in Equation (4.13) because it facilitates solving directly for the probability of error P_E as a function of signal energy.

4.3 DETECTION OF SIGNALS IN GAUSSIAN NOISE

The bandpass model of the detection process is virtually identical to the baseband model considered in Chapter 3. That is because a received bandpass waveform is first transformed to a baseband waveform before the final detection step takes place. For linear systems, the mathematics of detection is unaffected by a shift in frequency. In fact, we can define an *equivalence theorem* as follows: Performing bandpass linear signal processing, followed by heterodyning the signal to baseband yields the same results as heterodyning the bandpass signal to baseband, followed by baseband linear signal processing. The term "heterodyning" refers to a frequency *conversion* or *mixing* process that yields a spectral shift in the signal. As a result of this equivalence theorem, all linear signal-processing simulations can take place at baseband (which is preferred for simplicity), with the same results as at bandpass. This means that the performance of most digital communication systems will often be described and analyzed as if the transmission channel is a baseband channel.

4.3.1 Decision Regions

Consider that the two-dimensional signal space in Figure 4.6 is the locus of the noise-perturbed prototype binary vectors $(\mathbf{s}_1 + \mathbf{n})$ and $(\mathbf{s}_2 + \mathbf{n})$. The noise vector, $\mathbf{n}$, is a zero-mean random vector; hence the received signal vector, $\mathbf{r}$, is a random vector with mean $\mathbf{s}_1$ or $\mathbf{s}_2$. The detector's task after receiving $\mathbf{r}$ is to decide which of the signals ($\mathbf{s}_1$ or $\mathbf{s}_2$,) was actually transmitted. The method is usually to decide on the signal classification that yields the minimum expected P_E, although other strategies are possible [2]. For the case where M equals 2, with $\mathbf{s}_1$ and $\mathbf{s}_2$ being equally likely and with the noise being an additive white Gaussian noise (AWGN) process, we will see that the minimum-error decision rule is equivalent to choosing the signal class such that the distance $d(\mathbf{r}, \mathbf{s}_i) = \| \mathbf{r} - \mathbf{s}_i \|$ is minimized, where $\| \mathbf{x} \|$ is called the *norm* or *magnitude* of vector $\mathbf{x}$. This rule is often stated in terms of decision regions. In Figure 4.6, let us construct decision regions in the following way. Draw a line connecting the tips of the prototype vectors $\mathbf{s}_1$ and $\mathbf{s}_2$. Next, construct the perpendicular bisector of the connecting line. Notice that this bisector passes through the origin of the space if $\mathbf{s}_1$ and $\mathbf{s}_2$ are equal in amplitude. For this $M = 2$ example, in Figure 4.6, the constructed perpendicular bisector represents the locus of points equidistant between $\mathbf{s}_1$ and $\mathbf{s}_2$; hence, the bisector describes the boundary between decision region 1 and decision region 2. The *decision rule* for the detector, stated in terms of *decision regions*, is as follows: Whenever the received signal $\mathbf{r}$ is located

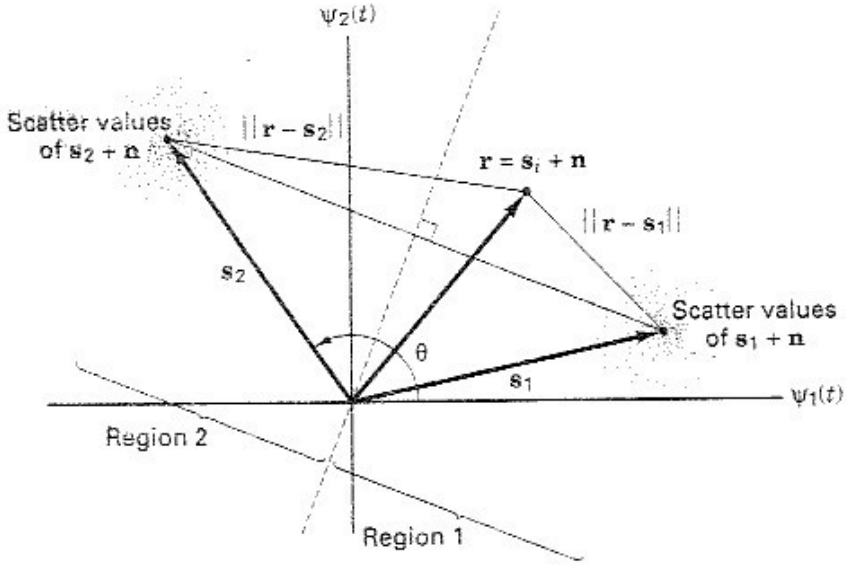


Figure 4.6 Two-dimensional signal space, with arbitrary equal-amplitude vectors $\mathbf{s}_1$ and $\mathbf{s}_2$.

in region 1, choose signal $\mathbf{s}_1$; when it is located in region 2, choose signal $\mathbf{s}_2$. In Figure 4.6, if the angle θ equals 180°, then the signal set $\mathbf{s}_1$ and $\mathbf{s}_2$ represents BPSK. However, in this figure, θ was purposely chosen to be less than 180° in order to emphasize the idea of decision regions in general.

4.3.2 Correlation Receiver

In Section 4.2 we treated the detection of *baseband* binary signals in Gaussian noise. Since the detection of *bandpass* signals employs the same concepts, we shall summarize the key findings of that section. We focus particularly on that realization of a matched filter known as a *correlator.* In addition to binary detection, we also consider the more general case of M-ary detection. We assume that the only performance degradation is due to AWGN. The received signal is the sum of the transmitted prototype signal plus the random noise:

$$r(t) = s_i(t) + n(t) \qquad \begin{matrix} 0 \le t \le T \\ i = 1, \ldots, M \end{matrix} \tag{4.14}$$

Given such a received signal, the detection process consists of *two basic steps* as shown in Figure 3.1. In the first step, the received waveform, $r(t)$, is reduced to a *single random variable* $z(T)$, or to a *set of random variables* $z_i(T)$ $(i = 1, \ldots, M)$, formed at the output of the demodulator and sampler at time $t = T$, where T is the symbol duration. In the second step, a symbol decision is made on the basis of comparing $z(T)$ to a threshold or on the basis of choosing the maximum $z_i(T)$. Step 1 can be thought of as transforming the waveform into a point in the decision space. This point can be referred to as the *predetection point,* the most critical reference point in the receiver. When we talk about received signal power, or received interfering noise, or E_b/N_0, their values are always considered with reference to this predetection point. Sometimes such received parameters are loosely described with reference to the *receiver input* or the *receiving antenna.* But when described in such ways, there is an underlying assumption of no loss in SNR or E_b/N_0 between these inputs and the predetection point. At each symbol time, the signal that is available at the predetection point is a sample of a baseband pulse. We do not yet have bits.

Does the fact that the energy-per-bit versus N_0 is *defined* at a place where there are no bits strike the reader as a paradox? Actually, it really should not, because this is a convenient reference point, where the baseband pulse—even before bit-decisions are made can be considered to *effectively* represent bits. Step 2 can be thought of as determining *in which decision region* the point is located. For the detector to be optimized (in the sense of minimizing the error probability), it is necessary to optimize the waveform-to-random-variable transformation, by using matched filters or correlators in step 1, and by also optimizing the decision criterion in step 2.

In Sections 3.2.2 and 3.2.3 we found that the matched filter provides the maximum signal-to-noise ratio at the filter output at time $t = T$. We described a correlator as one realization of a matched filter. We can define a *correlation receiver* comprised of M correlators, as shown in Figure 4.7a, that transforms a received waveform, $r(t)$, to a sequence of M numbers or correlator outputs, $z_i(T)$ $(i = 1, \ldots, M)$. Each correlator output is characterized by the following product integration or correlation with the received signal:

$$z_i(T) = \int_0^T r(t)s_i(t)\,dt \qquad i = 1, \ldots, M \tag{4.15}$$

The verb "to correlate" means "to match." The correlators attempt to match the incoming received signal, $r(t)$, with each of the candidate prototype waveforms, $s_i(t)$, known a priori to the receiver. A reasonable decision rule is to choose the waveform, $s_i(t)$, that *matches best* or has the *largest correlation* with $r(t)$. In other words, the decision rule is

$$\begin{array}{l}\text{Choose the } s_i(t) \text{ whose index} \\ \text{corresponds to the max } z_i(T)\end{array} \tag{4.16}$$

Following Equation (3.10), any signal set, $\{s_i(t)\}$ $(i = 1, \ldots, M)$, can be expressed in terms of some set of basis functions, $\{\psi_j(t)\}$ $(j = 1, \ldots, N)$, where $N \leq M$. Then the bank of M correlators in Figure 4.7a may be replaced with a bank of N correlators, shown in Figure 4.7b, where the set of basis functions $\{\psi_j(t)\}$ form *reference signals.* The decision stage of this receiver consists of logic circuitry for choosing the signal, $s_i(t)$. The choice of $s_i(t)$ is made according to the best match of the coefficients, a_{ij}, seen in Equation (3.10), with the set of outputs $\{z_j(T)\}$. When the prototype waveform set, $\{s_i(t)\}$, is an orthogonal set, the receiver implementation in Figure 4.7a is identical to that in Figure 4.7b (differing perhaps by a scale factor). However, when $\{s_i(t)\}$ is *not* an orthogonal set, the receiver in Figure 4.7b, using N correlators instead of M, with reference signals $\{\psi_j(t)\}$, can represent a cost-effective implementation. We examine such an application for the detection of multiple phase shift keying (MPSK) in Section 4.4.3.

In the case of *binary detection,* the correlation receiver can be configured as a single matched filter or product integrator, as shown in Figure 4.8a, with the reference signal being the difference between the binary prototype signals, $s_1(t) - s_2(t)$. The output of the correlator, $z(T)$, is fed directly to the decision stage.

For binary detection, the correlation receiver can also be drawn as two matched filters or product integrators, one of which is matched to $s_1(t)$, and the

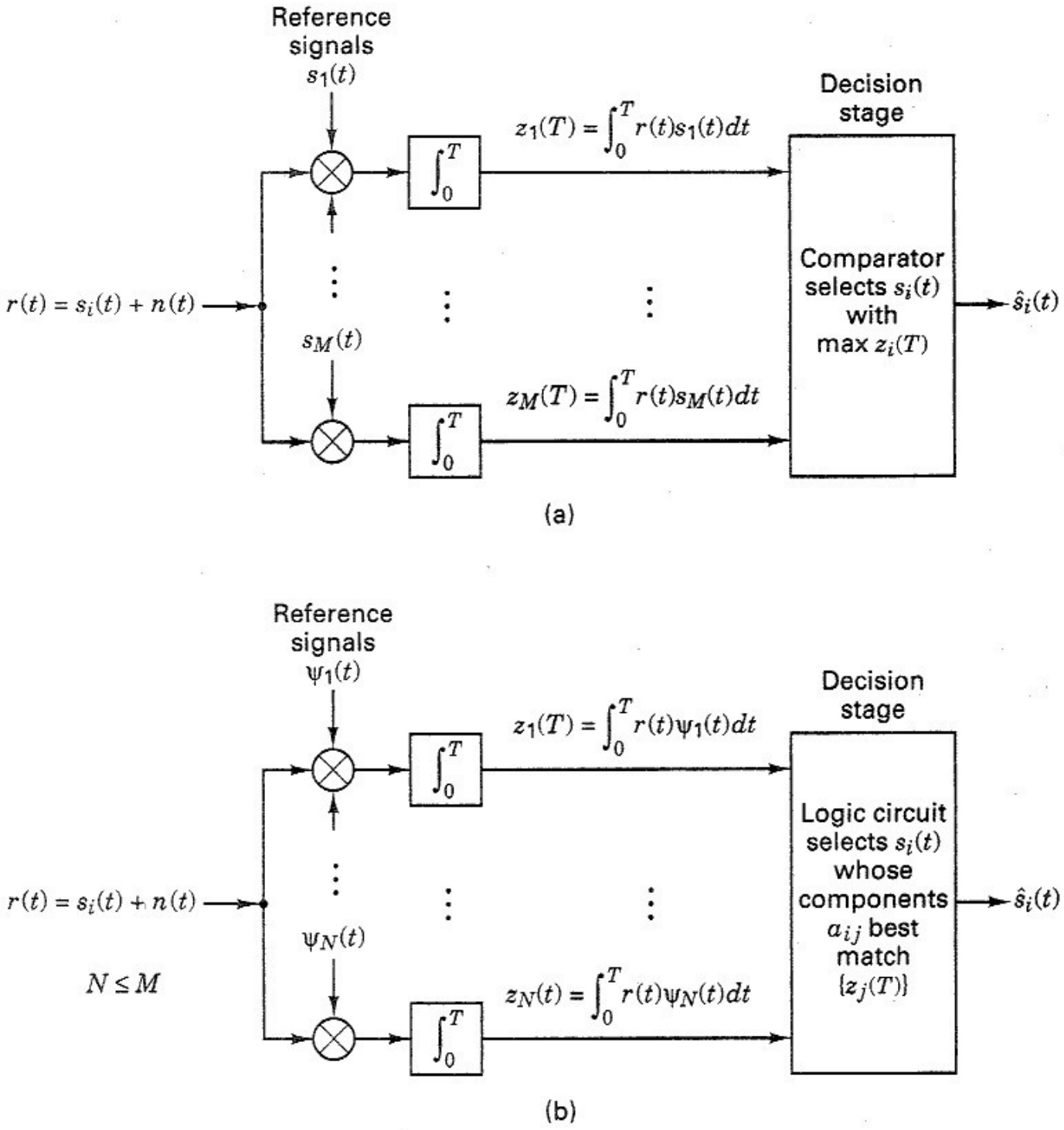


Figure 4.7 (a) Correlator receiver with reference signals $\{s_i(t)\}$. (b) Correlator receiver with reference signals $\{\psi_j(t)\}$.

other is matched to $s_2(t)$. (See Figure 4.8b.) The decision stage can then be configured to follow the rule in Equation 4.16, or the correlator outputs $z_i(T)$ $(i = 1, 2)$ can be differenced to form

$$z(T) = z_1(T) - z_2(T) \tag{4.17}$$

as shown in Figure 4.8b. Then, $z(T)$, called the *test statistic,* is fed to the decision stage, as in the case of the single correlator. In the *absence of noise,* an input waveform $s_i(t)$ yields the output $z(T) = a_i(T)$, a signal-only component. The input noise

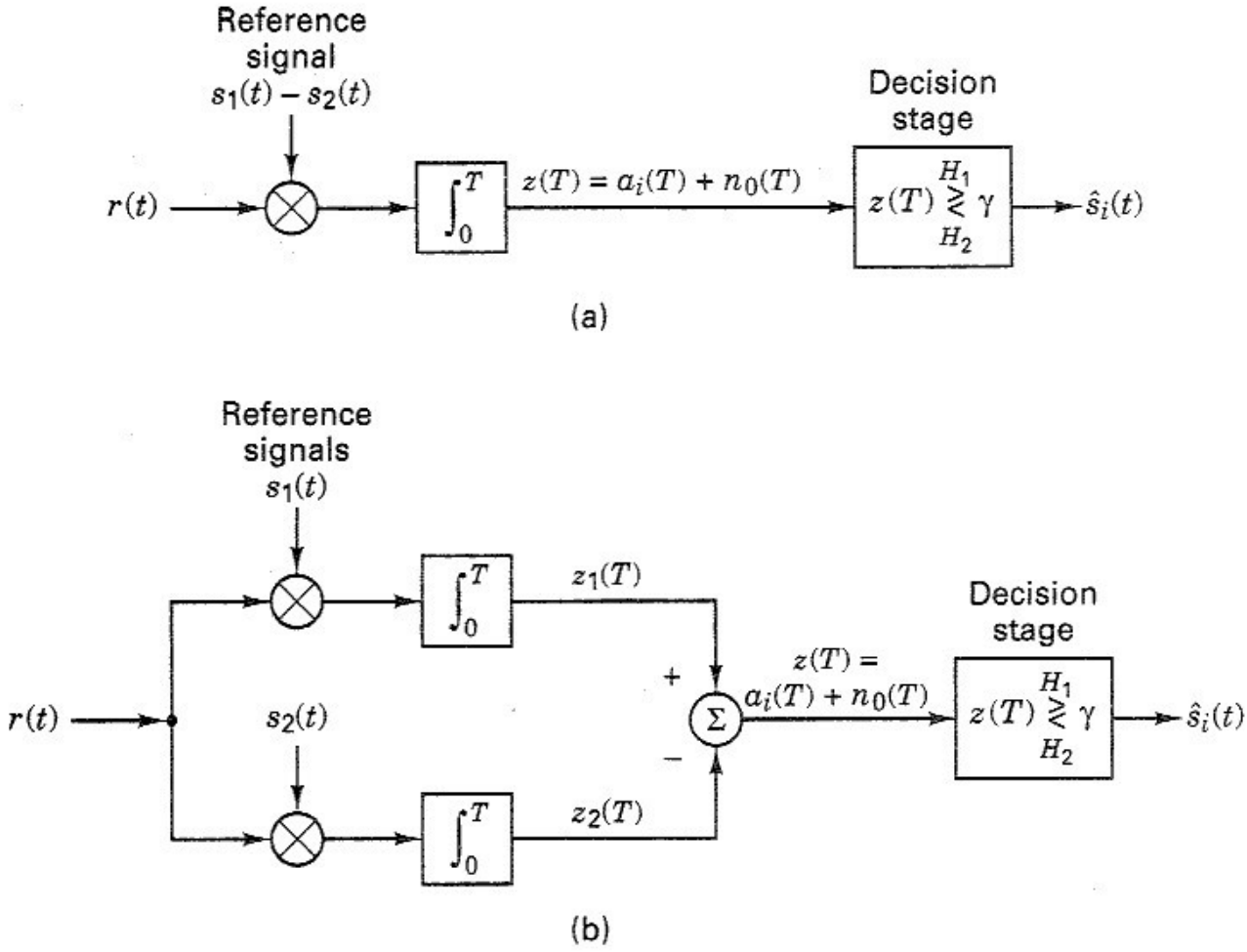


Figure 4.8 Binary correlator receiver. (a) Using a single correlator. (b) Using two correlators.

$n(T)$ is a Gaussian random process. Since the correlator is a *linear* device, the output noise is also a Gaussian random process [2]. Thus, the output of the correlator, sampled at $t = T$, yields

$$z(T) = a_i(T) + n_0(T) \qquad i = 1, 2$$

where $n_0(T)$ is the noise component. To shorten the notation we sometimes express $z(t)$ as $a_i + n_0$. The noise component n_0 is a zero-mean *Gaussian random variable,* and thus $z(T)$ is a *Gaussian random variable* with a mean of either a_1 or a_2, depending on whether a binary one or binary zero was sent.

4.3.2.1 Binary Decision Threshold

For the random variable $z(T)$, Figure 4.9 illustrates the two conditional probability density functions (pdfs), $p(z|s_1)$ and $p(z|s_2)$, with mean value of a_1 and a_2, respectively. These pdfs, also called the *likelihood* of s_1 and the *likelihood* of s_2, respectively, were presented in Section 3.1.2, and are rewritten as

$$p(z|s_1) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_1}{\sigma_0}\right)^2\right] \tag{4.18a}$$

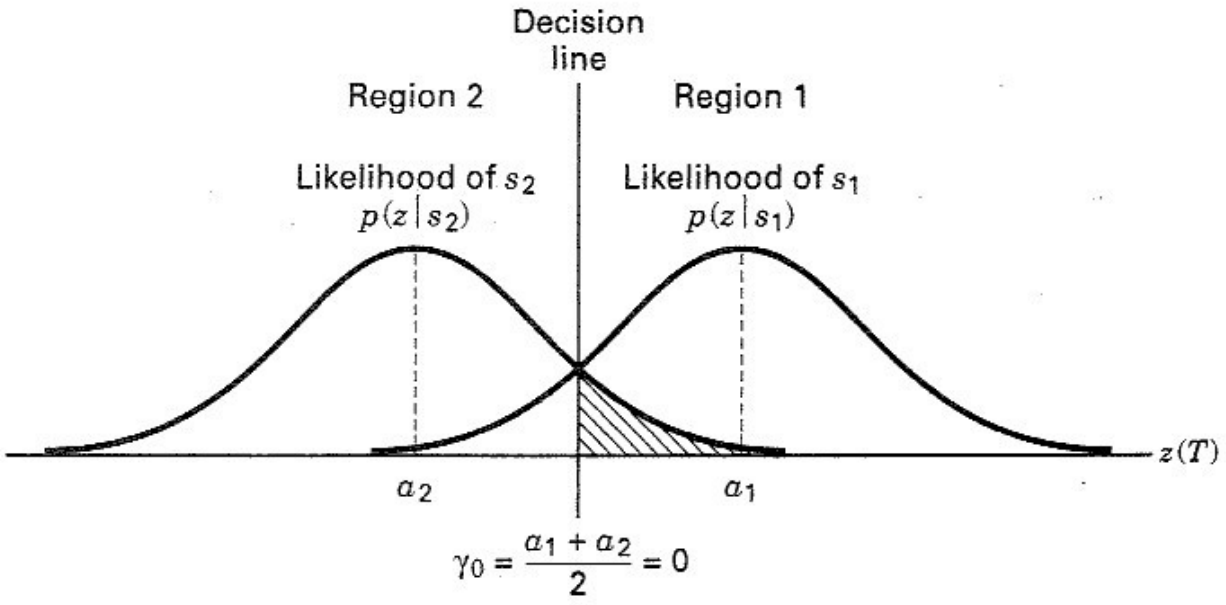


Figure 4.9 Conditional probability density functions: $p(z/s_1)$, $p(z/s_2)$.

and

$$p(z|s_2) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_2}{\sigma_0}\right)^2\right] \tag{4.18b}$$

where σ_0^2 is the noise variance. In Figure 4.9 the rightmost likelihood $p(z|s_1)$ illustrates the probability density of the detector output $z(T)$, given that $s_1(t)$ was transmitted. Similarly, the leftmost likelihood $p(z|s_2)$ illustrates the probability density of $z(T)$, given that $s_2(t)$ was transmitted. The abscissa $z(T)$ represents the full range of possible sample output values from the correlation receiver shown in Figure 4.8.

With regard to optimizing the binary decision threshold for deciding in which region a received signal is located, we found in Section 3.2.1 that the *minimum error* criterion for equally likely binary signals corrupted by Gaussian noise can be stated as

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1 + a_2}{2} = \gamma_0 \tag{4.19}$$

where a_1 is the signal component of $z(T)$ when $s_1(t)$ is transmitted, and a_2 is the signal component of $z(T)$ when $s_2(t)$ is transmitted. The threshold level γ_0 represented by $(a_1 + a_2)/2$ is the *optimum threshold* for minimizing the probability of making an incorrect decision given equally likely signals and symmetrical likelihoods. The decision rule in Equation (4.19) states that hypothesis H_1 should be selected [equivalent to deciding that signal $s_1(t)$ was sent] if $z(T) > \gamma_0$, and hypothesis H_2 should be selected [equivalent to deciding that $s_2(t)$ was sent] if $z(T) < \gamma_0$. If $z(T) = \gamma_0$, the decision can be an arbitrary one. For equal-energy, equally likely antipodal signals, where $s_1(t) = -s_2(t)$ and $a_1 = -a_2$, the optimum decision rule becomes

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} 0, \tag{4.20a}$$

or

$$\begin{aligned} &\text{decide } s_1(t) \quad \text{if } z_1(T) > z_2(T) \\ &\text{decide } s_2(t) \quad \text{otherwise} \end{aligned} \tag{4.20b}$$

4.4 COHERENT DETECTION

4.4.1 Coherent Detection of PSK

The detector shown in Figure 4.7 can be used for the coherent detection of any digital waveforms. Such a correlating detector is often referred to as a *maximum likelihood detector.* Consider the following binary PSK (BPSK) example: Let

$$s_1(t) = \sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi) \qquad 0 \le t \le T \tag{4.21a}$$

$$\begin{aligned} s_2(t) &= \sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi + \pi) \\ &= -\sqrt{\frac{2E}{T}} \cos(\omega_0 t + \phi) \qquad 0 \le t \le T \end{aligned} \tag{4.21b}$$

and

$$n(t) = \text{zero-mean white Gaussian random process}$$

where the phase term ϕ is an arbitrary constant, so that the analysis is unaffected by setting $\phi = 0$. The parameter E is the signal energy per symbol, and T is the symbol duration. For this antipodal case, only a single basis function is needed. If an orthonormal signal space is assumed in Equations (3.10) and (3.11) (i.e., $K_j = 1$), we can express a basis function $\psi_1(t)$ as

$$\psi_1(t) = \sqrt{\frac{2}{T}} \cos \omega_0 t \qquad \text{for } 0 \le t \le T \tag{4.22}$$

Thus, we may express the transmitted signals $s_i(t)$ in terms of $\psi_1(t)$ and the coefficients $a_{i1}(t)$ as follows:

$$s_i(t) = a_{i1}\psi_1(t) \tag{4.23a}$$

$$s_1(t) = a_{11}\psi_1(t) = \sqrt{E}\,\psi_1(t) \tag{4.23b}$$

$$s_2(t) = a_{21}\psi_1(t) = -\sqrt{E}\,\psi_1(t) \tag{4.23c}$$

Assume that $s_1(t)$ was transmitted. Then the expected values of the product integrators in Figure 4.7b, with reference signal $\psi_1(t)$, are found as

$$\mathbf{E}\{z_1|s_1\} = \mathbf{E}\left\{ \int_0^T \sqrt{E}\,\psi_1^2(t) + n(t)\psi_1(t)\,dt \right\} \tag{4.24a}$$

$$\mathbf{E}\{z_2|s_1\} = \mathbf{E}\left\{ \int_0^T -\sqrt{E}\,\psi_1^2(t) + n(t)\psi_1(t)\,dt \right\} \tag{4.24b}$$

$$\mathbf{E}\{z_1|s_1\} = \mathbf{E}\left\{ \int_0^T \frac{2}{T}\sqrt{E}\cos^2\omega_0 t + n(t)\sqrt{\frac{2}{T}}\cos\omega_0 t\,dt \right\} = \sqrt{E} \tag{4.25a}$$

and

$$\mathbf{E}\{z_2|s_1\} = \mathbf{E}\left\{ \int_0^T -\frac{2}{T}\sqrt{E}\cos^2\omega_0 t + n(t)\sqrt{\frac{2}{T}}\cos\omega_0 t\,dt \right\} = -\sqrt{E} \tag{4.25b}$$

where $\mathrm{E}\{\cdot\}$ denotes the ensemble average, referred to as the *expected value.* Equation (4.25) follows because $\mathbf{E}\{n(t)\} = 0$. The decision stage must decide which signal was transmitted by determining its location within the signal space. For this example, the choice of $\psi_1(t) = \sqrt{2/T}\cos\omega_0 t$ normalizes $\mathbf{E}\{z_i(T)\}$ to be $\pm\sqrt{E}$. The prototype signals $\{s_i(t)\}$ are the same as the reference signals $\{\psi_j(t)\}$ except for the normalizing scale factor. The decision stage chooses the signal with the largest value of $z_i(T)$. Thus, the received signal in this example is judged to be $s_1(t)$. The error performance for such coherently detected BPSK systems is treated in Section 4.7.1.

4.4.2 Sampled Matched Filter

In Section 3.2.2, we discussed the basic characteristic of the matched filter—namely, that its impulse response is a delayed version of the mirror image (rotated on the $t = 0$ axis) of the input signal waveform. Therefore, if the signal waveform is $s(t)$, its mirror image is $s(-t)$, and the mirror image delayed by T seconds is $s(T-t)$. The impulse response $h(t)$ of a filter matched to $s(t)$ is then described by

$$h(t) = \begin{cases} s(T-t) & 0 \le t \le T \\ 0 & \text{elsewhere} \end{cases} \tag{4.26}$$

Figures 4.7 and 4.8 illustrate the basic function of a correlator to product-integrate the received noisy signal with each of the candidate reference signals and determine the best match. The schematics in these figures imply the use of analog hardware (multipliers and integrators) and continuous signals. They do not reflect the way that the correlator or matched filter (MF) can be implemented using digital techniques and sampled waveforms. Figure 4.10 shows how an MF can be implemented using digital hardware. The input signal $r(t)$ comprises a prototype signal $s_i(t)$, plus noise $n(t)$, and the bandwidth of the signal is $W = 1/2T$, where T is the symbol time. Thus, the minimum Nyquist sampling rate is $f_s = 2W = 1/T$, and the sampling time T_s needs to be equal to or less than the symbol time. In other words, there must be at least one sample per symbol. In real systems, such sampling is usually performed at a rate that exceeds the Nyquist minimum by a factor of 4 or more. The only cost is processor speed, not transmission bandwidth. At the clock

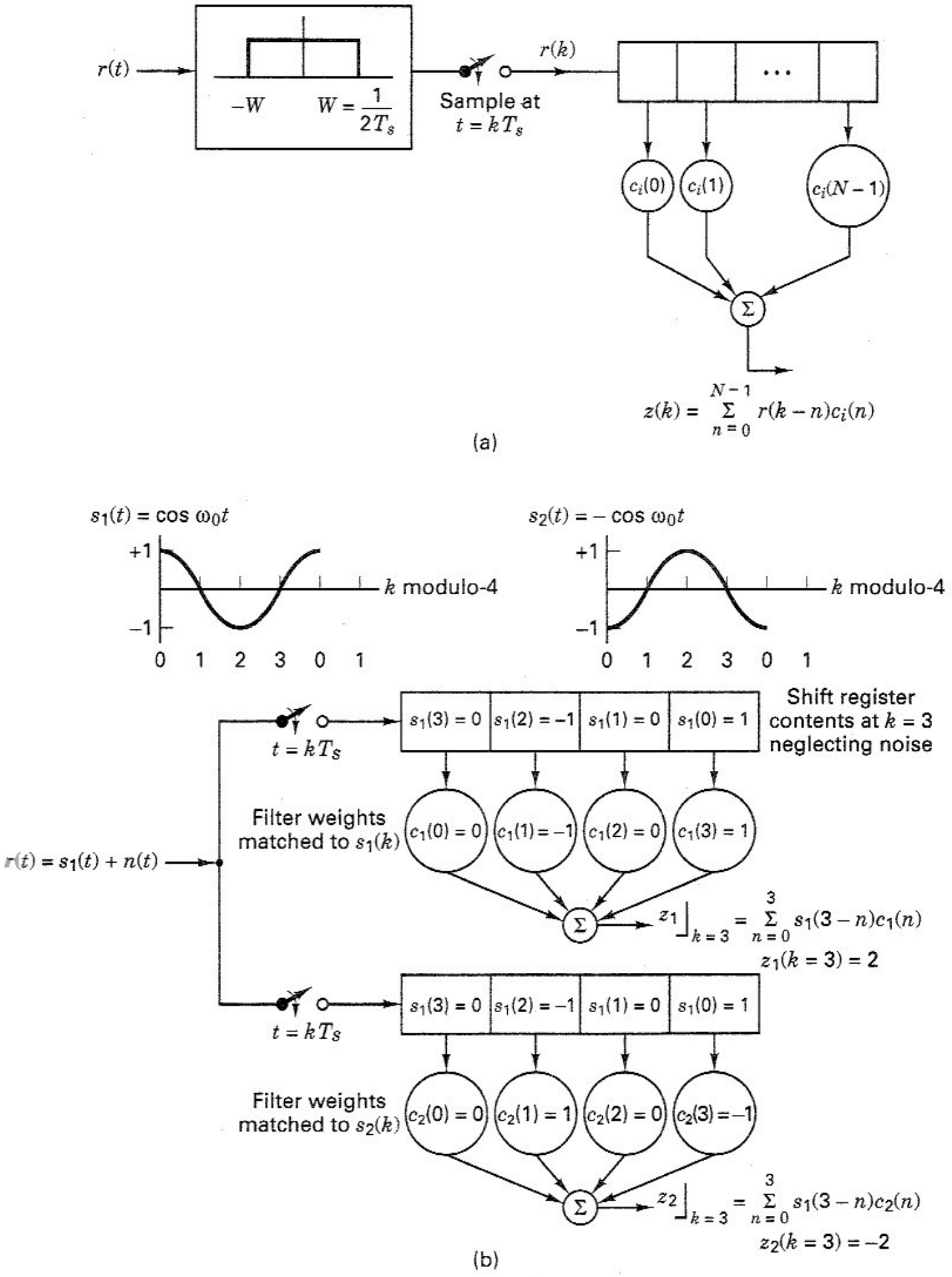


Figure 4.10 (a) Sampled matched filter. (b) Sampled matched filter detection example, neglecting noise.