CHAPTER 12.3
PULSE MODULATION/ DEMODULATION

George F. Pfeifer

PULSE MODULATION

Pulse modulation is, in general, the encoding of information by means of varying one or more pulse parameters. It finds application in both the communication and the control fields. The control applications are usually confined to the use of pulse-time modulation (PTM) and pulse-frequency modulation (PFM), where on-off control power can be used to minimize device dissipation. All pulse modulation schemes require sampling analog signals, and some, such as pulse-code modulation (PCM) and delta modulation, require the additional quantization of the analog signals.

In communications, the chief application of pulse modulation is found where it is desired to time-multiplex by interleaving a number of single-channel, low-duty-cycle pulse trains. The pulse trains may, in turn, be used for compound modulation by amplitude or angle modulation of a continuous carrier. In usual applications, subcarriers are pulsed, time-division-multiplexed, and then used to frequency-modulate a carrier.

Since noise is present in all systems, a prime consideration in modulation selection is the choice of a waveform based on its signal-to-noise efficiency. For instance, PTM is more efficient than pulse amplitude modulation (PAM), which offers no improvement over continuous AM; however, PTM is less efficient than PCM or delta modulation. A chief advantage of pulsed systems such as PTM, PCM, and delta is improved signal-to-noise ratio in exchange for increased bandwidth, in the same manner as continuous FM improves over AM.

SAMPLING AND SMOOTHING

An ideal impulse sampler can be considered as the multiplication of an impulse train, period T seconds, with the continuous signal $f(t)$. This is shown in Fig. 12.3.1*a* for the impulse train defined as $p_T(t) = \text{rep}_T[\delta(t)]$, where $\delta(t)$ is an impulse at $t = 0$ and

$$\text{rep}_T[u(t)] = \sum_{n=-\infty}^{\infty} u(t - nT) \tag{1}$$

The output spectrum function is the convolution of $F(f)$, the Fourier transform of the input signal, and the transform of $p_T(t)$, which is $(1/T)\ \text{comb}_{1/T}(1)$, defined as

$$\text{comb}_{1/T}[U(f)] = \sum_{n=-\infty}^{\infty} U\left(\frac{n}{T}\right) \delta\left(f - \frac{n}{T}\right) \tag{2}$$

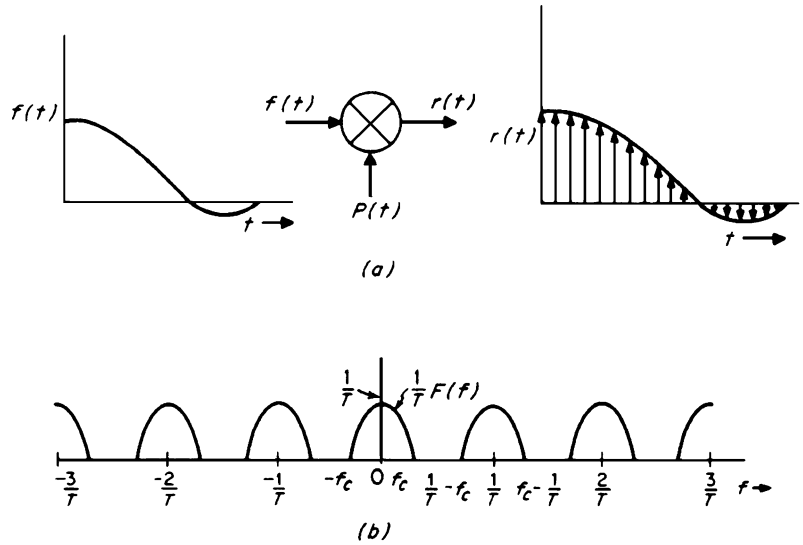

FIGURE 12.3.1 Pulse modulation: (*a*) output spectrum; (*b*) sampling configuration.

Thus the transform $R(f) = F(f) * (1/T)\ \text{comb}_{1/T}(1)$ with spectrum is as shown in Fig. 12.3.1*b*. The result of ideal impulse sampling has been to repeat the original signal spectrum, assumed to be band-limited, each $1/T$ Hz and multiply each by a $1/T$ scale factor. Since all the signal information is present in each lobe of Fig. 12.3.1*b*, it is only necessary to recover a single lobe through filtering in order to recover the signal function reduced by a scale factor.

Consider an ideal low-pass rectangular filter of bandwidth of f_f Hz defined as $T(jf) = A(f)\ \exp\ [-j^{\theta}(jf)]$, where

$$A(f) = \begin{cases} 1 & |f| < f_f \\ 0 & |f| > f_f \end{cases} \text{ and } \quad \theta(jf) = 2\pi\alpha f \quad \text{for all } f$$

The cutoff frequency f_f is adjusted to select the output spectral lobe about zero $f_c < f_f < 1/T - f_c$ and will fall in the guard band between lobes. That portion of filter output $R(f)$ selected is

$$R_0(f) = (1/T)F(f)\ \exp\ (-j2\pi\alpha f) \tag{3}$$

that will inverse transform as

$$r_0(t) = (1/T)f(t - \alpha) \tag{4}$$

which is identical with the signal function, with the amplitude reduced by a scale factor and function shifted by α seconds. If $\alpha = 0$, signifying no delay, the filter is termed a "cardinal data hold"; otherwise, it is an "ideal low-pass filter." Unfortunately, these filters cannot be realized in practice, since they are required to respond before they are excited.

Examination of Fig. 12.3.1*b* gives rise to the sampling theorem accredited to Shannon and/or Nyquist, which states that when a continuous time function with band-limited spectrum $-f_c < f < f_c$ is sampled at twice the highest frequency, $f_s = 2f_c$, the original time function can be recovered. This corresponds to the point where the sampling frequency $f_s = 1/T$ is decreased so that the spectral lobes of Fig. 12.3.1*b* are just touching. To decrease f_s beyond the value of $2f_c$ would cause spectral overlap and make recovery with an ideal filter impossible. A more general form of the sampling theorem states that any $2f$ independent samples per second will

completely describe a band-limited signal, thus removing the restriction of uniform sampling, as long as independent samples are used. In general, for a time-limited signal of T seconds, band-limited to f_c Hz, only $2f_cT$ samples are needed to specify the signal completely.

In practice, the signal is not completely band-limited, so that it is common to allow for a greater separation of spectral lobes, called the *guard band*. This guard band is generated simply by sampling at greater than $2f_c$, as in the case for Fig. 12.3.1*b*. Although the actual tolerable overlap depends on the signal spectral slope, setting the sampling rate at about $3f_c = f_s$ is usually adequate to recover the signal.

In practice, narrow but finite-width pulse trains are used in place of the idealized impulse sampling train.

PULSE-AMPLITUDE MODULATION

Pulse-amplitude modulation is essentially a sampled-data type of encoding where the information is encoded into the amplitude of a train of finite-width pulses. The pulse train can be looked upon as the carrier in much the same way as the sine wave is for continuous-amplitude modulation. There is no improvement in signal-to-noise when using PAM, and furthermore, PAM is not considered wideband in the sense of FM or PTM. Thus PAM would correspond to continuous AM, while PTM corresponds to FM. Generally, PAM is used chiefly for time-multiplex systems employing a number of channels sampled, consistent with the sampling theorem.

There are a number of ways of encoding information as the amplitude of a pulse train. They include both bipolar and unipolar pulse trains for both instantaneous or square-topped sampling and for exact or top sampling. In top sampling, the magnitude of the individual pulses follows the modulating signal during the pulse duration, while for square-topped sampling, the individual pulses assume a constant value, depending on the particular exact sampling point that occurs somewhere during the pulse time. These various waveforms are shown in Fig. 12.3.2.

The top-modulation bipolar sampling case is shown in Fig. 12.3.2*c*; it is simply sampling with a finite-pulse-width train. Carrying out the convolution yields

$$R_{\text{STB}}(f) = \frac{\tau}{T} \sum_{n=-\infty}^{\infty} \left(\text{sinc} \frac{\tau n}{T} \right) F\left(f - \frac{n}{T} \right) \tag{5}$$

The spectrum for top-modulation bipolar sampling, using a square-topped rectangular spectrum for the original signal spectrum, is shown in Fig. 12.3.3*a*. The signal spectrum repeats with a $(\sin x)/x$ scale factor determined by the sampling pulse width, with each repetition a replica of $F(f)$.

Unipolar sampling can be implemented by adding a constant bias A to $f(t)$, the signal, to produce $f(t) + A$, where A is large enough to keep the sum positive; that is, $A > |f(t)|$. Sampling the new sum signal by multiplication with the pulse train results in the unipolar top-modulated waveform of Fig. 12.3.2*e*. The spectrum is

$$R_{\text{STU}}(f) = \frac{\tau}{T} \sum_{n=-\infty}^{\infty} \left(\text{sinc} \frac{\tau n}{T} \right) \left[F\left(f - \frac{n}{T} \right) + A\delta\left(f - \frac{n}{T} \right) \right] \tag{6}$$

The delta-function part of the summation reduces to the spectrum function of the pulse train $S(f)$

$$R_{\text{STU}}(f) = AS(f) + \frac{\tau}{T} \sum_{n=-\infty}^{\infty} \left(\text{sinc} \frac{\tau n}{T} \right) F\left(f - \frac{n}{T} \right) \tag{7}$$

The resulting spectrum of top-modulation unipolar sampling is the same as with bipolar sampling plus the impulse spectrum of the sampling pulse train, as shown in Fig. 12.3.3*b*. For square-topped-modulation bipolar sampling, the time-domain result is

$$r_{\text{SSB}}(t) = \text{rect}\ (t/\tau) * \text{comb}_T f(t) \tag{8}$$

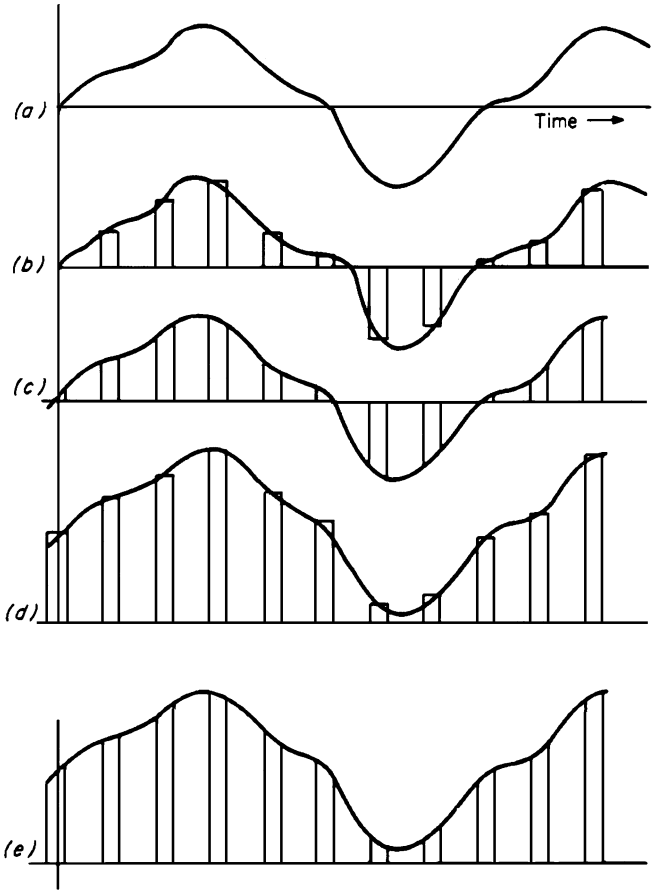

FIGURE 12.3.2 PAM waveforms: (*a*) modulation; (*b*) square-top sampling, bipole pulse train; (*c*) top sampling, bipole pulse train; (*d*) square-top sampling, unipolar pulse train; (*e*) top sampling, unipolar pulse train.

with spectrum function

$$R_{\text{SSB}}(f) = \frac{\tau}{T}(\text{sinc}\, f\tau) \sum_{n=-\infty}^{\infty} F\left(f - \frac{n}{T}\right) \tag{9}$$

In this case, the signal spectrum is distorted by the sinc $f\tau$ envelope, as shown in Fig. 12.3.2*c*. This frequency distortion is referred to as *aperture effect* and may be corrected by use of an equalizer sinc $f\tau$ form, following the low-pass reconstruction filter.

As in the previous case of unipolar sampling, the resulting spectrum for square-topped modulation will contain the pulse-train spectrum, as shown in Fig. 12.3.2*d*. The expression is

$$R_{\text{SSU}}(f) = AS(f) + \frac{\tau}{T}(\text{sinc}\, f\tau) \sum_{n=-\infty}^{\infty} F\left(f - \frac{n}{T}\right) \tag{10}$$

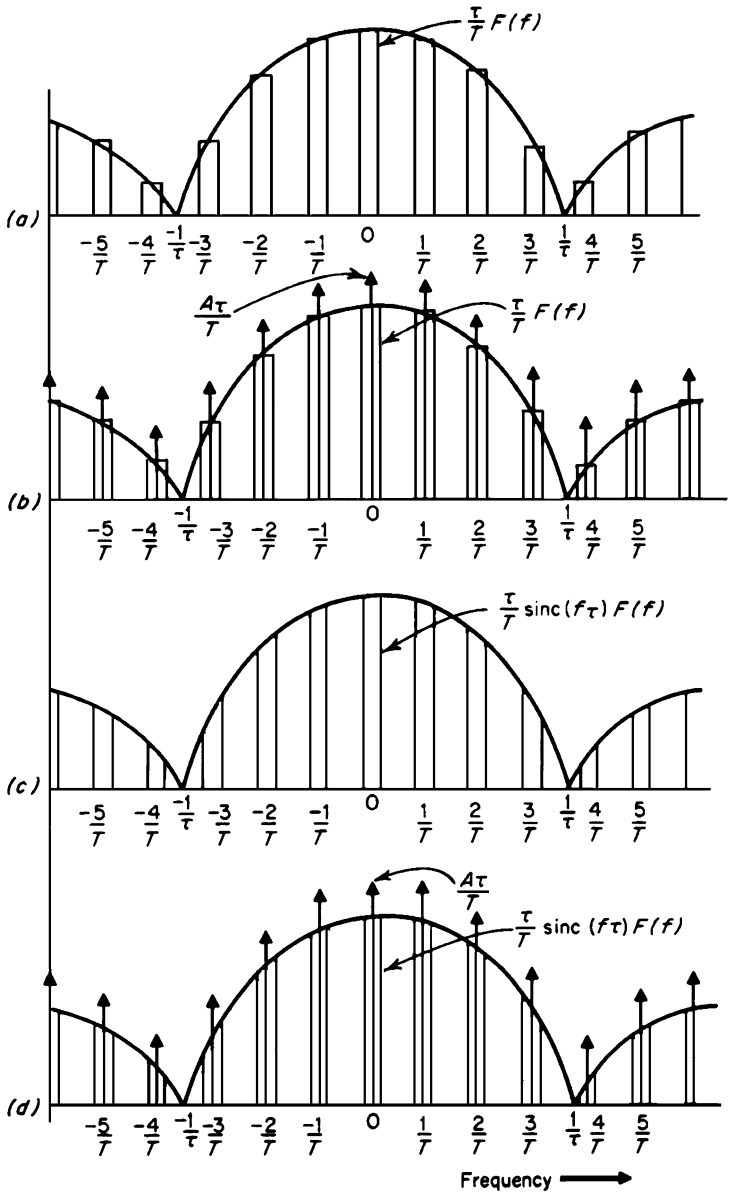

FIGURE 12.3.3 PAM spectra: (*a*) top modulation, bipolar sampling; (*b*) top modulation, unipolar sampling; (*c*) square-top modulation, bipolar sampling; (*d*) square-top modulation, unipolar sampling.

The signal information is generally recovered, in PAM systems, by use of a low-pass filter that acts on the reduced signal energy around zero frequency, as shown in Fig. 12.3.3.

PULSE-TIME, PULSE-POSITION, AND PULSE-WIDTH MODULATION

In PTM the information is encoded into the time parameter instead of, for instance, the amplitude, as in PAM. There are two basic types of PTM: pulse-position modulation (PPM) and pulse-width modulation (PWM), also known as pulse-duration (PDM) or pulse-length (PLM) modulation. The PTM allows the power-driver circuitry to operate at saturation level, thus conserving power loss. Operating driver circuitry full on, full off, is especially important for heavy-duty high-load control applications, as well as for communication applications.

In PPM the information is encoded into the time position of a narrow pulse, generally with respect to a reference pulse. The basic pulse width and amplitude are kept constant, while only the pulse position is changed, as shown in Fig. 12.3.4. There are three cases of PWM which are the modulation of the leading edge, trailing edge, or both edges, as displayed in Fig. 12.3.5. In this case the information is encoded into the width of the pulse, with the pulse amplitude and period held constant. The derivative relationship existing between PPM and PWM can be illustrated by consideration of trailing-edge PWM modulation. The pulses of PPM can be derived from the edges of trailing-edge PWM (Fig. 12.3.5*b*) by differentiation of the PWM signal and a sign change of the trailing-edge pulse. Pulse-position modulation is essentially the same as PWM, with the information-carrying variable edge replaced by a pulse. Thus, when that part of the signal power of PWM that carries no information is deleted, the result is PPM.

Generally, in PTM systems a guard interval is necessary because of the pulse rise times and system responses. Thus 100 percent of the interpulse period cannot be used without considerable channel cross-talk because of pulse overlap. It is necessary to trade off crosstalk versus channel utilization at the system design level.

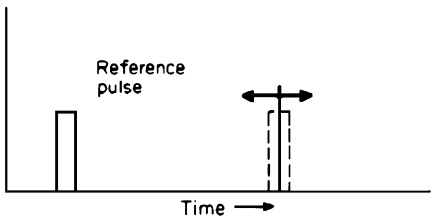

FIGURE 12.3.4 PPM time waveform.

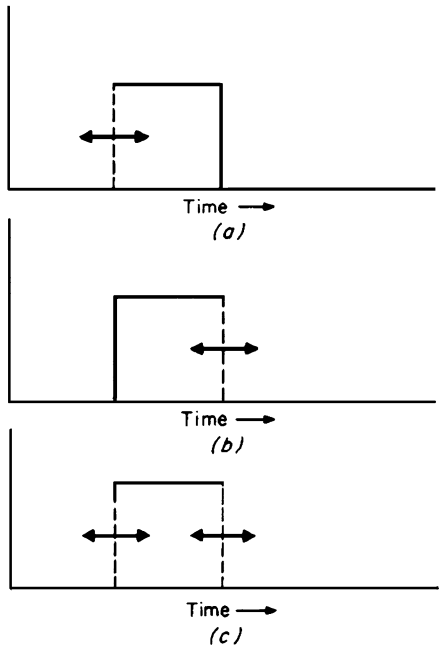

FIGURE 12.3.5 PWM time waveforms: (*a*) leading-edge modulation; (*b*) trailing-edge modulation; (*c*) both-edge modulation.

Another consideration is that the information sampling rate cannot exceed the pulse repetition frequency and would be less for a single channel of a multiplexed system where channels are interwoven in time.

Generation of PTM

There are two basic methods of pulse-time modulation: (1) based on uniform sampling in which the pulse-time parameter is directly proportional to the modulating signal at uniformly sampled points and (2) in which there is some distortion of the pulse-time parameter because of the modulation process. Both methods of modulation are illustrated in Fig. 12.3.6 for PWM. Basically, PPM can be derived from trailing-edge PWM, as shown in Fig. 12.3.6*c* by use of an edge detector or differentiator and a standard narrow-pulse generator.

In the uniform sampling case for PWM of Fig. 12.3.6*a*, the modulating signal is sampled uniformly in time and the special PAM derived by a sample-and-hold circuit as shown in Fig. 12.3.7*a*. This PAM signal provides a pedestal for each of the three types of sawtooth waveforms producing leading, trailing, or double-edge PWM, as shown in Fig. 12.3.7*c*, *e*, and *g*, respectively. The uniform sampled PPM is shown in Fig. 12.3.7*h*, as derived from the trailing-edge modulation of *g*.

Nonuniformly sampled modulation, termed *natural sampling* by some authors, is shown in Fig. 12.3.8, and results from the method of Fig. 12.3.8*b*, where the sawtooth is added directly to the modulating signal. In this case the modulating waveform influences the time when the samples are actually taken. This distortion is small when the modulating-amplitude change is small during the interpulse period *T*. The distortion is caused by the modulating signal distorting the sawtooth wave-form when they are added, as indicated in Fig. 12.3.6*b*. The information in the PPM waveform is similarly distorted because it is derived from the PWM waveform, as shown in Fig. 12.3.7*h*.

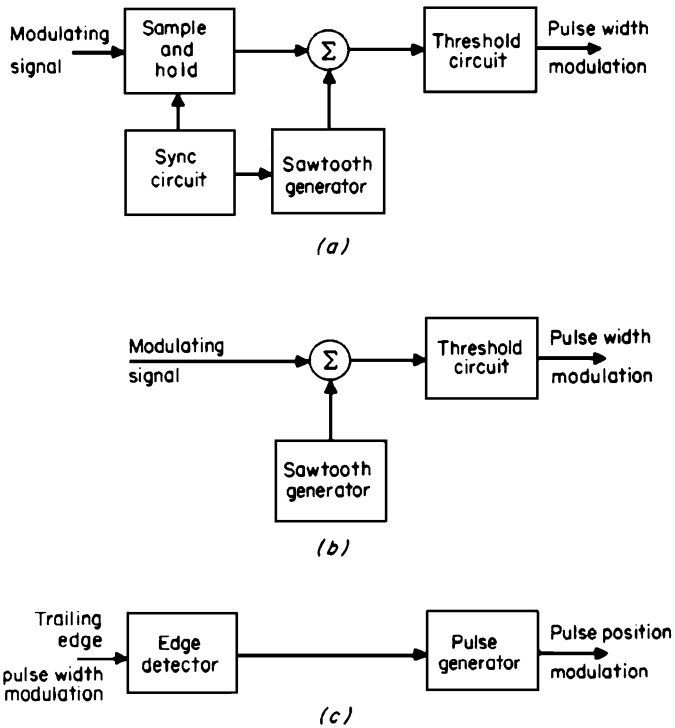

FIGURE 12.3.6 PTM generation: (*a*) pulse-width-modulation generation, uniform sampling; (*b*) pulse-width-modulation generation, nonuniform sampling; (*c*) pulse-position-modulation generation.

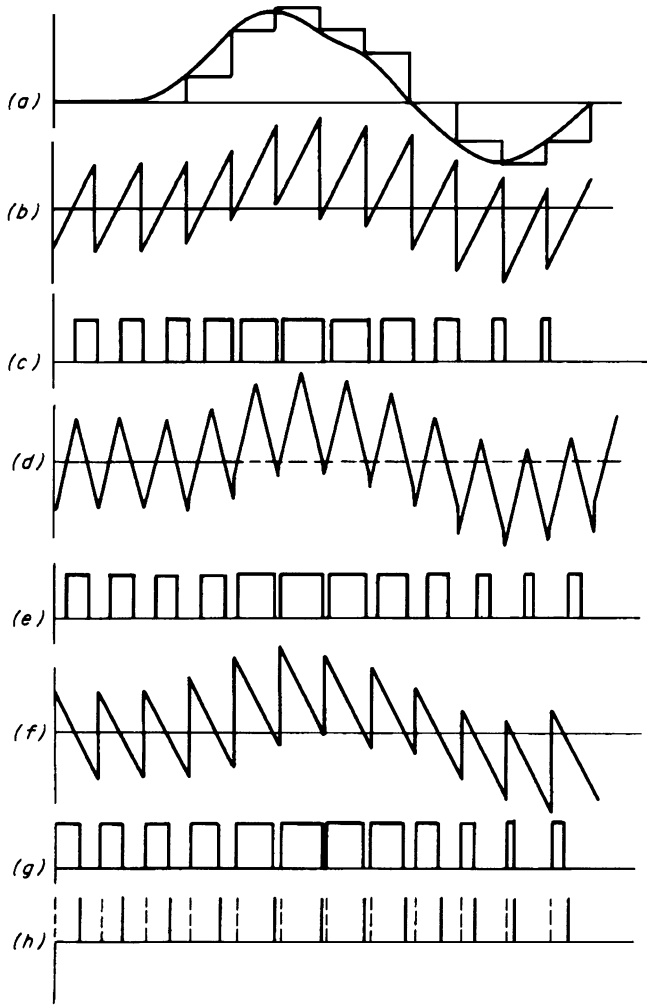

FIGURE 12.3.7 Pulse-time modulation, uniform sampling: (*a*) modulating signal and sample-and-hold waveform; (*b*) sawtooth added to sample-and-hold waveform; (*c*) leading-edge modulation; (*d*) sawtooth added to sample-and-hold waveform; (*e*) double-edge modulation; (*f*) sawtooth added to sample-and-hold waveform; (*g*) trailing-edge modulation; (*h*) pulse-position modulation (reference pulse dotted).

PULSE-TIME MODULATION SPECTRA

The spectra are smeared in general, for most modulating signals, and are difficult to derive; however, it is possible to get some idea of what happens to the spectra with modulation by considering a sinusoidal modulation of form

$$A \cos 2\pi f_s t \tag{11}$$

The amplitude $A < T/2$, where T is the interpulse period, assuming no guard band.

For PPM with uniform sampling and unity pulse amplitude, the spectrum is given by

$$\begin{aligned} x(t) = {} & \frac{\tau}{T} + \frac{2\tau}{T}\sum_{m=1}^{\infty} (\text{sinc } mf_0) J_0(2\pi A m f_0) \cos\ 2\pi\ mf_0 t \\ & + \frac{2\tau}{T}\sum_{n=1}^{\infty} \text{sinc } (nf_s) J_n(2\pi A n f_s) \cos\left(2\pi\ nf_s t - \frac{n\pi}{2}\right) \\ & + \frac{2\tau}{T}\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\left\{\text{sinc } (mf_0 + nf_s) J_n[2\pi A(mf_0 + nf_s)] \cos\left[2\pi(mf_0 + nf_s)t - \frac{n\pi}{2}\right]\right. \\ & \left. + \text{sinc } (nf_s - mf_0) J_n[2\pi A(nf_s - mf_0)] \cos\left[2\pi(nf_s - mf_0)t - \frac{n\pi}{2}\right]\right\} \end{aligned} \tag{12}$$

where τ = pulse width
T = pulse period
f_s = modulation frequency
J_n = Bessel function of first kind, nth order
$f_0 = 1/T$

As is apparent, all the harmonics of the pulse-repetition frequency and the modulation frequency are present, as well as all possible sums and differences. The dc level is τ/T, with the harmonics carrying the modulation. The pulse shape effects the line amplitudes as a sinc function, reducing the spectra for higher frequencies.

The spectrum for PWM is similar to that of PPM, and for uniformly sampled trailing-edge sinusoidal modulation is given by

$$\begin{aligned} x(t) = {} & \frac{1}{2} + \frac{1}{\pi T}\sum_{m=1}^{\infty}\frac{1}{mf_0}\cos\left[2\pi\ mf_0 t + \frac{\pi}{2}(2m-1)\right] \\ & + \frac{1}{\pi T}\sum_{m=1}^{\infty}\frac{1}{mf_0} J_0(2\pi A m f_0) \cos\left(2\pi\ mf_0 t - \frac{\pi}{2}\right) \\ & + \frac{1}{\pi T}\sum_{n=1}^{\infty}\frac{1}{nf_s} J_n(2\pi A n f_s) \cos\left[2\pi\ nf_s t - (n+1)\frac{\pi}{2}\right] \\ & + \frac{1}{\pi T}\sum_{\substack{m=1 \\ n=1}}^{\infty}\left\{\frac{1}{mf_0 + nf_s} J_n[2\pi A(mf_0 + nf_s)] \cos\left[2\pi(mf_0 + nf_s)t - (n+1)\frac{\pi}{2}\right]\right. \\ & \left. + \frac{1}{nf_s - mf_0} J_n[2\pi A(nf_s - mf_0)] \cos\left[2\pi(nf_s - mf_0)t - (n+1)\frac{\pi}{2}\right]\right\} \end{aligned} \tag{13}$$

The same comments apply for PWM as for PPM.

A more compact form is given for PPM and PWM, respectively, as

$$x(t) = \frac{1}{T}\sum_{\substack{m=\infty \\ n=\infty}}^{\infty} (-j)^n J_n[2\pi A(mf_0 + nf_s)] P(mf_0 + nf_s) \exp\ [j2\pi(mf_0 + nf_s)t] \tag{14}$$

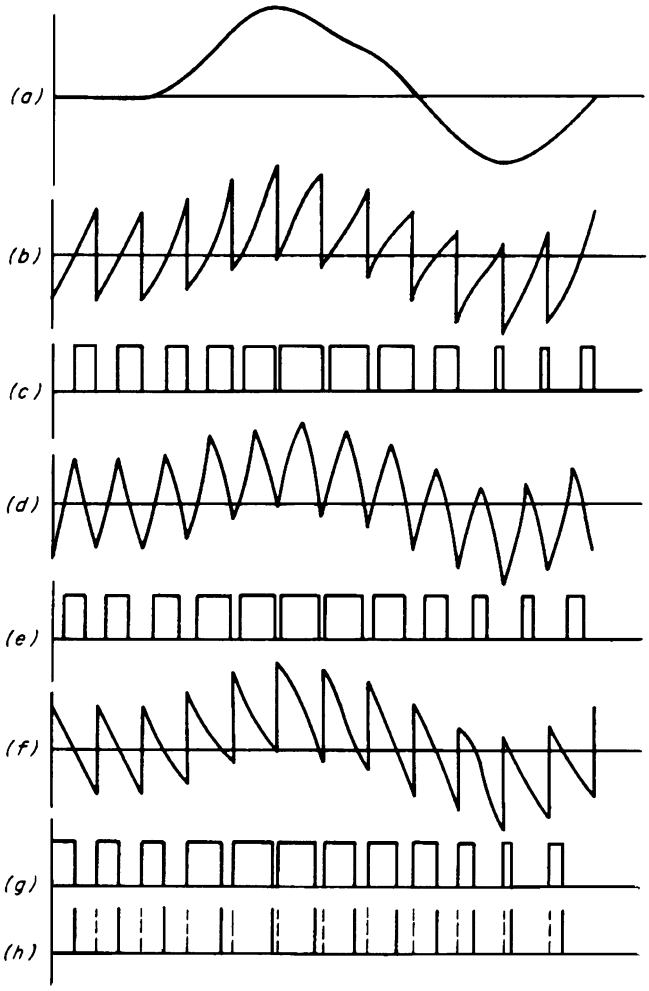

FIGURE 12.3.8 Pulse-time modulation, nonuniform sampling; (*a*) modulating signal; (*b*) sawtooth added to modulation; (*c*) leading-edge modulation; (*d*) sawtooth added to modulation; (*e*) double-edge modulation; (*f*) sawtooth added to modulation; (*g*) trailing-edge modulation; (*h*) pulse-position modulation.

where $P(f)$ is the Fourier transform of the pulse shape $p(t)$, and

$$x(t) = \frac{1}{2} + \frac{1}{T} \sum_{\substack{m=-\infty \\ m \neq 0}}^{\infty} j^{2m-1} \frac{e^{j2\pi\, mf\, 0t}}{2\pi\, mf_0} - \frac{1}{T} \sum_{\substack{m=-\infty \\ n=-\infty \\ |m| + |n| \neq 0}}^{\infty}$$

$$(-j)^{n+1} \frac{J_n[2\pi A(mf_0 + nf_s)]}{2\pi(mf_0 + nf_s)} \exp\,[j2\pi(mf_0 + nf_s)t] \tag{15}$$

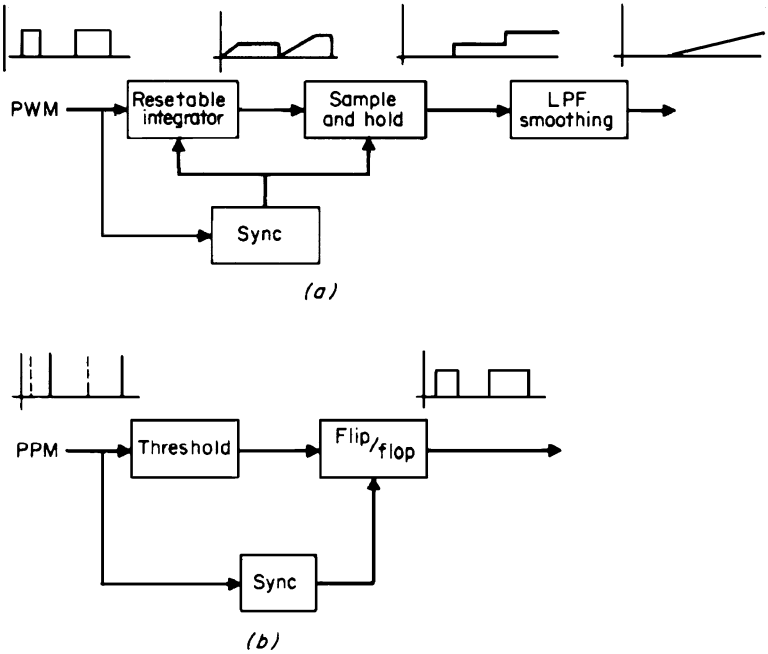

FIGURE 12.3.9 Pulse-time demodulation: (*a*) PWM demodulation; (*b*) PPM to PWM for demodulation.

DEMODULATION OF PTM

Demodulation of PWM or PPM can be accomplished by low-pass filtering if the modulation is small compared with the impulse period. However, in general, it is best to demodulate on a pulse-to-pulse basis that usually requires some form of synchronization with the pulses. The distortion introduced by nonuniform sampling cannot be eliminated and will be present in the demodulated waveform. However, if the modulation is small compared with the interpulse period T, the distortion will be minimized.

To demodulate PWM each pulse can be integrated and the maximum value sampled and held and low-pass-filtered, as shown in Fig. 12.3.9*a*. To sample and reset the integrator, it is necessary to derive sync from the PWM waveform, in this case trailing-edge-modulated.

Generally, PPM is demodulated by conversion to PWM and then demodulated as PWM. Although in some demodulation schemes the actual PWM waveform may not exist as such, the general demodulation scheme is the same. PPM can be converted to PWM by the configuration of Fig. 12.3.9*b*. The PPM signal is applied to an amplitude threshold, usually termed a *slicer*, that rejects noise except near the pulses. The pulses are applied to a flip-flop synchronized to one particular state by the reference pulse, and it generates the PWM as its output.

PULSE FREQUENCY MODULATION

In PFM the information is contained in the frequency of the pulse train, which is composed of narrow pulses. The highest frequency possible ideally occurs when there is no more interpulse spacing left for finite-width pulses. This frequency, given by $1/\tau$, where τ is the pulse width, will not be achieved in practice, owing to the pulse rise time. The lowest frequency is determined by the modulator, usually a voltage-controlled oscillator

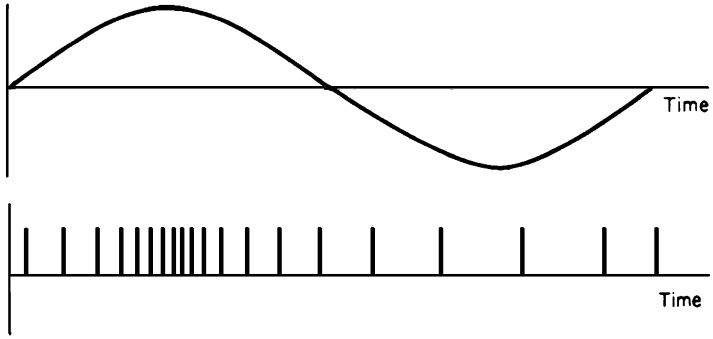

FIGURE 12.3.10 PFM modulation.

(VCO), in which in practice a 100:1 ratio of high to low frequency is easily achievable. Examination of Fig. 12.3.10 indicates why PFM is used mostly for control purposes rather than communications. The wide variation and uncertainty of pulse position do not lend themselves to time multiplexing, which requires the interweaving of channels in time. Since one of the chief motivations of pulse modulation in communication systems is to be able to time-multiplex a number of channels, PFM is not used. On the other hand, PFM is a good choice for on-off control applications, especially where fine control is required. A classic example of PFM control is for the attitude control of near-earth satellites that have on-off gas thrusters where a very close approximation to a linear system response is achievable.

Generation of PFM

Basically, PFM is generated by modulation of a VCO as shown in Fig. 12.3.11*a*. A constant reference voltage is added to the modulation so that the frequency can swing above and below the reference-determined value. For control applications it is usually required that the frequency follow the magnitude of the modulation, its sign determining which actuators are to be turned on, as shown in Fig. 12.3.11*b*.

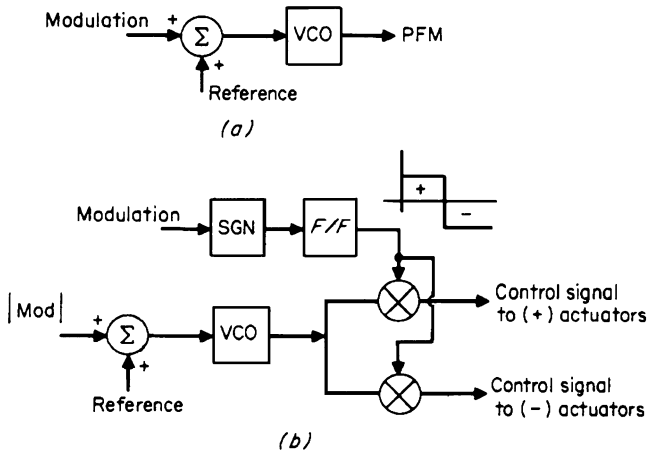

FIGURE 12.3.11 Generation of PFM: (*a*) PFM modulation; (*b*) PFM for control.

PULSE-CODE MODULATION

In PCM the signal is encoded into a steam of digits. This differs from the other forms of pulse modulation by requiring that the sample values of the signal be quantized into a number of levels and subsequently coded as a series of pulses for transmission. By selecting enough levels, the quantized signal can be made to approximate closely the original continuous signal at the expense of transmitting more bits per sample. The PCM scheme lends itself readily to time multiplexing of channels and will allow widely different types of signals; however, synchronization is strictly required. This synchronization of the system can be on a single-sample or code-group basis. The synchronizing signal is most likely inserted with a group of samples from different channels, on a frame or subframe basis to conserve space.

The motivation behind modern PCM is that improved implementation techniques of solid-state circuitry allow extremely fast quantization of samples and translation to complex codes with reasonable equipment constraints. PCM is an attractive way to trade bandwidth for signal-to-noise and has the additional advantage of transmission through regenerative repeaters with a signal-to-noise ratio that is substantially independent of the number of repeaters. The only requirement is that the noise, interference, and other disturbances be less than one-half a quantum step at each repeater. Also, systems can be designed that have error-detecting and error-correcting features.

PCM CODING AND DECODING

Coding is the generation of a PCM waveform from an input signal, and decoding is the reverse process. There are many ways to code and many code groups to use: hence standardization is necessary when more than one user is considered. Each sample value of the signal waveform is quantized and represented to sufficient accuracy by an appropriate code character. Each code character is composed of a specified number of code elements. The code elements can be chosen as two-level, or binary; three-level, or ternary; or n-ary. However, general practice is to use binary, since it is not affected as much by interference introduced by the required increased bandwidth. An example of binary coding is shown in Fig. 12.3.12 for 3-bit or eight levels of quantization. Each code group is composed of three pulses, with the pulse trains shown for on-off pulses in Fig. 12.3.12*b* and bipolar pulses in Fig. 12.3.12*c*.

A generic diagram of a complete system is shown in Fig. 12.3.13. The recovered signal is a delayed copy of the input signal degraded by noise because of sources such as sampling, quantization, and interference. For this type of system to be efficient, both sending and receiving terminals must be synchronized. The synchronism is

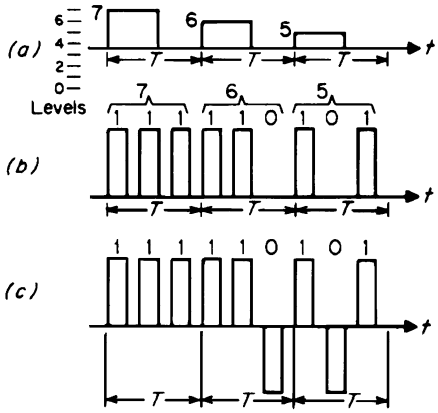

FIGURE 12.3.12 Binary pulse coding: (*a*) quantized samples; (*b*) on-off coded pulses; (*c*) bipolar coded pulses.

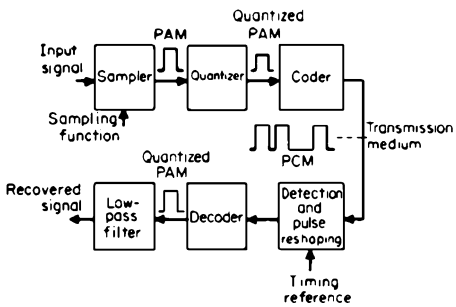

FIGURE 12.3.13 Basic operations of a PCM system.

required to be monitored continuously and be capable of establishing initial synchronism when the system is out of frame. The synchronization is usually accomplished by use of special sync pulses that establish frame, subframe, or word sync.

There are three basic ways to code, namely, feedback and subtraction, pulse counting, and parallel comparison. In *feedback subtraction* the sample value is compared with the most significant code-element value and that value subtracted from the sample value if the element value is less. This process of comparison and subtraction is repeated for each code-element value down to the least significant bit. At each subtraction the appropriate code element or bit is selected to complete the coding. In *pulse counting* a gate is established by using the PWM pulse corresponding to a sample value. Clock pulses are gated using the PWM gate and are connected in a counter. The output of a decoding network attached to the counter is read out as the PCM. *Parallel comparison* is the fastest method since the sampled value is applied to a number of different threshold values. The thresholders are read out as the PCM.

SYSTEM CONSIDERATIONS FOR PCM

Quantization introduces an irremovable error into the system, referred to as *quantization noise*. This kind of noise is characterized by the fact that its magnitude is always less than one-half a quantum step, and it can be treated as uniformly distributed additive noise with zero mean value and rms value equal to $1/\sqrt{12}$ times the total height of a quantum step. When the ratio of signal power to quantization noise power at the quantizer output is used as a measure of fidelity the improvement with quantizer levels is as shown in Fig. 12.3.14 for different kinds of signals.

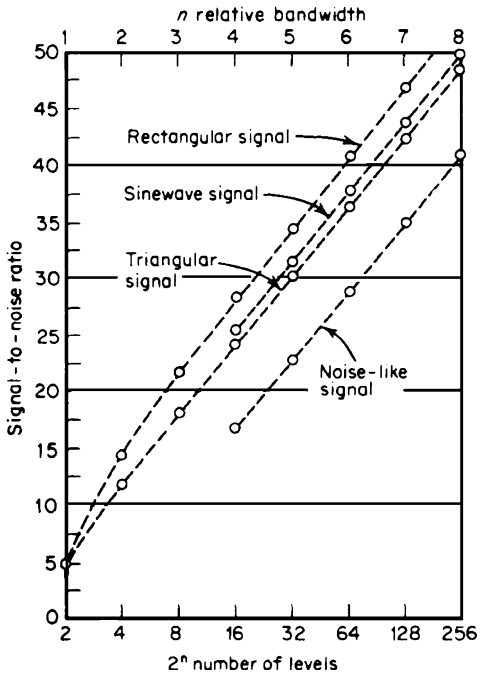

FIGURE 12.3.14 PCM signal-to-noise improvement with number of quantization levels.

In general, using an n-ary code with m pulses allows transmission of n^m values. For the binary code this reduces the 2^m values which approximate the signal to 1 part in $2^m - 1$ levels. Encoding into pulse and minus pulses, assuming either pulse is equally likely, results in an average power of $A^2/4$, which is half the on-off power of $A^2/2$, where the total pulse amplitude, peak to peak, is A. The channel capacity for a system sampled at the Nyquist rate of $2f_m$ and quantized into s levels is

$$C = 2f_m \log_2 s \quad \text{(bits/s)} \tag{16}$$

or for m pulses of n values each

$$C = mf_m \log_2 n^2 \quad \text{(bits/s)} \tag{17}$$

Since the encoding process squeezes one sample into m pulses, the pulse widths are effectively reduced by $1/m$; thus the transmission bandwidth is increased by a factor of m, or $B = mf_m$.

The maximum possible ideal rate of transmission of binary bits is

$$C = B \log_2 (1 + S/N) \quad \text{(bits/s)} \tag{18}$$

according to Shannon. For a system sampled at the Nyquist rate, quantized to $K\sigma$ per level and using the plus and minus pulses, the channel capacity is

$$C = B \log_2(1 + 12S/K^2N) \quad \text{(bits/s)} \; N = \sigma^2 \tag{19}$$

where S is the average power over large time interval and σ is the rms noise voltage at decoder input.

DELTA MODULATION

Delta modulation (DM) is basically a one-digit PCM system where the analog waveform has been encoded in a differential form. In contrast to the use of n digits in PCM, simple DM uses only one digit to indicate the changes in the sample values. This is equivalent to sending an approximation to the signal derivative. At the receiver the pulses are integrated to obtain the original signal. Although DM can be simply implemented in circuitry, it requires a sampling rate much higher than the Nyquist rate of $2f_m$ and a wider bandwidth than a comparable PCM system. Most of the other characteristics of PCM apply to DM.

Delta modulation differs from differential PCM in which the difference in successive signal samples is transmitted. In DM only 1 bit is used to express and transmit the difference. Thus DM transmits the sign of successive slopes.

Coding and Decoding DM

There are a number of coding and decoding variations in DM, such as single-integration, double-integration, mixed-integration, delta-sigma, and high-information DM (HIDM). In addition, companding the signal which is compressing the signal at transmission and expanding it at reception is also used to extend the limited dynamic range. The simple single-integration DM of the coding-decoding scheme is shown in Fig. 12.3.15. In the encoder the modulator produces positive pulses when the sign of the difference signal $\epsilon(t)$ is positive and negative pulses otherwise; and the output pulse train is integrated and compared with the input signal to provide an error signal $\epsilon(t)$, thus closing the encode feedback loop. At the receiver the pulse train is integrated and filtered to produce a delayed approximation to the signal, as shown in Fig. 12.3.16. The actual circuit implementation with operational amplifiers and logic circuits is very simple.

By changing to a double-integration network in the decoder, a smoother replica of the signal is provided. This decoder has the disadvantage, however, of not recognizing changes in the slope of the signal. This gave rise to a scheme to encode differences in slope instead of amplitude, leading to coders with double integration; however, systems of this type are marginally stable and can oscillate under certain conditions. Waveforms of a double-integrating delta coder are shown in Fig. 12.3.17. Single and double integration can be combined to give improved performance while avoiding the stability problem. These mixed systems are often referred to in the literature as *delta modulators* with double integration. A comparison of waveforms is shown in Fig. 12.3.18.

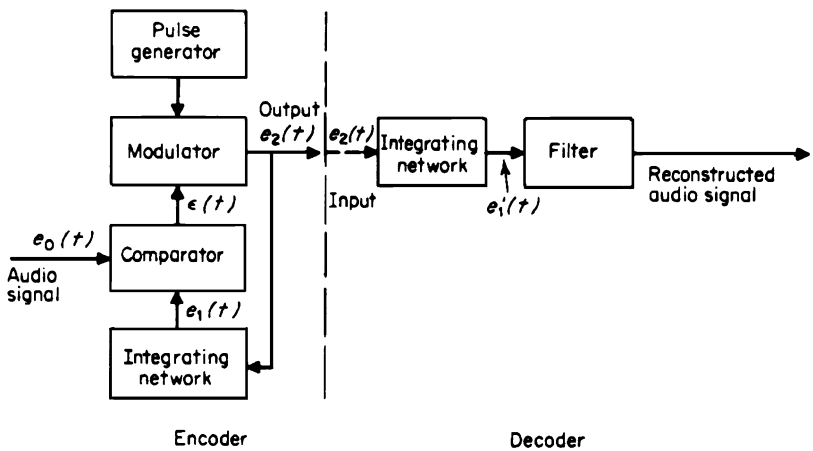

FIGURE 12.3.15 Basic coding-decoding diagram for DM.

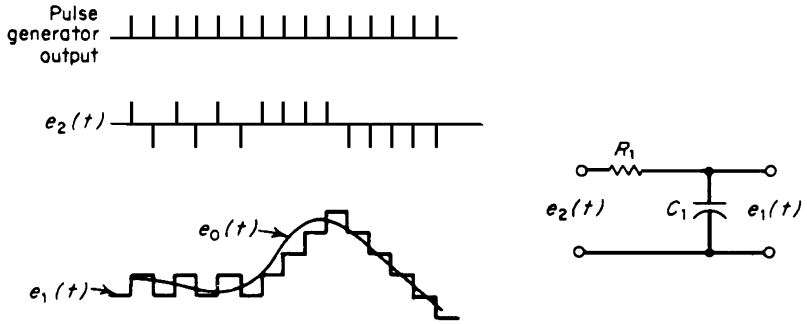

FIGURE 12.3.16 Delta-modulation waveforms using single integration.

System Considerations for DM

The synthesized waveform can change only one level each clock pulse; thus DM overloads when the slope of the signal is large. The maximum signal power will depend on the type of signal, since the greatest slope that can be reproduced is the integration of one level in one pulse period. For a sine wave of frequency f, the maximum-amplitude signal is

$$A_{\max} = f_s \sigma / 2\pi \tag{20}$$

where f_s is the sampling frequency and σ is one quantum step.

It has been observed that a DM system will transmit a speech signal without overloading if the amplitude of the signal does not exceed the maximum permissible amplitude of an 800-Hz sine wave. The DM

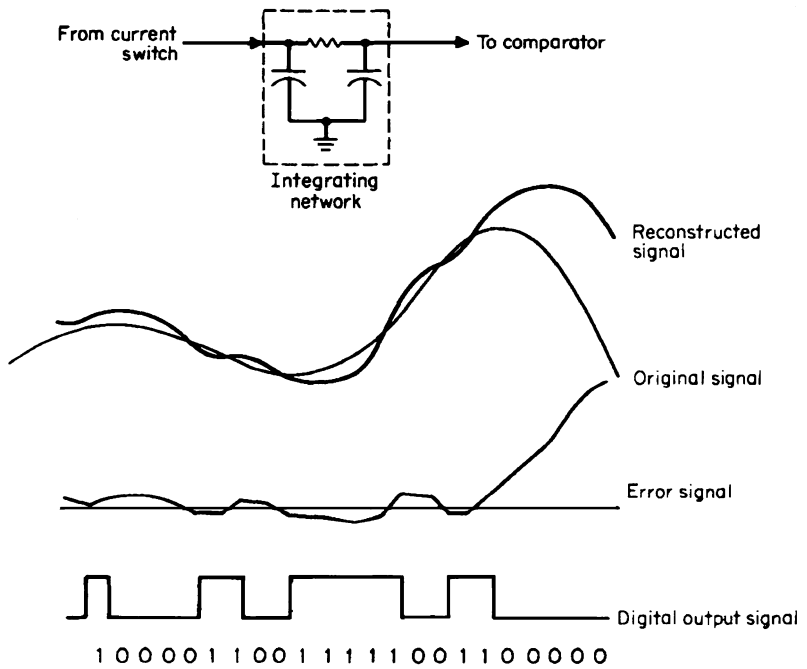

FIGURE 12.3.17 Waveforms for delta coder with double integration.

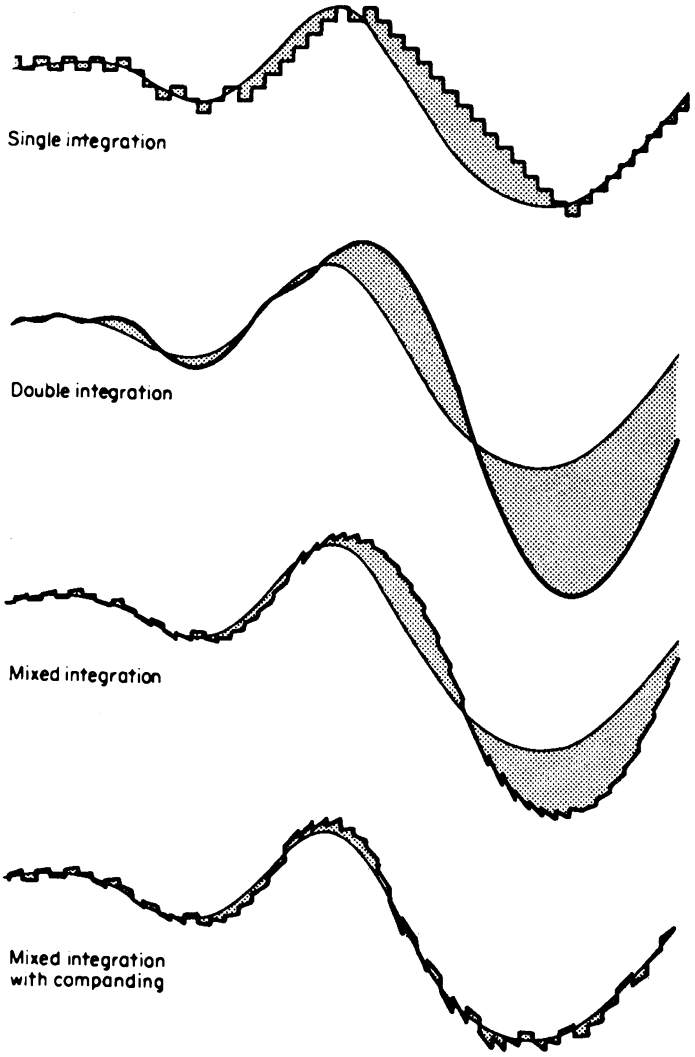

FIGURE 12.3.18 Waveforms for various integrating systems.

coder overload characteristic is shown in Fig. 12.3.19 along with the spectrum of a human voice. Notice that they decrease in frequency together, indicating that DM can be used effectively with speech transmission. Generally speaking, transmission of speech is the chief application of DM, although various modifications and improvements are being studied to extend DM to higher frequencies and transmission of the lost dc component.

Among these techniques is delta-sigma modulation, where the signal is integrated and compared with an integrated approximation to form the error signal similar to $\epsilon(t)$ of Fig. 12.3.15. The decoding is accomplished with a low-pass filter and requires no integration.

The signal-to-quantization noise ratio for single-integration DM is given by

$$S/N = 0.2 f_s^{3/2} / f f_0^{1/2} \tag{21}$$

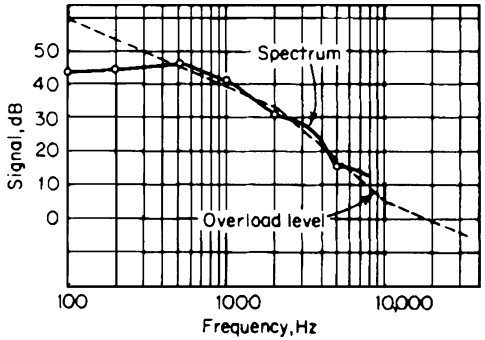

FIGURE 12.3.19 Spectrum of the human voice compared with delta-coder overload level.

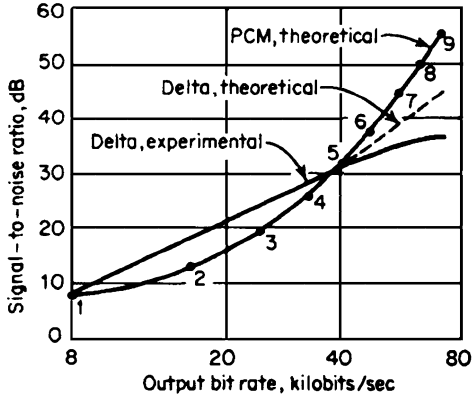

FIGURE 12.3.20 Signal-to-noise ratio for DM and PCM.

where f_s = sampling frequency
f = signal frequency
f_0 = signal bandwidth

For double or mixed DM

$$S/N = 0.026 f_s^{5/2} / f f_0^{3/2} \tag{22}$$

A comparison of signal-to-noise ratio (SNR) for DM and PCM is shown in Fig. 12.3.20, along with an experimental DM system for voice application. Note that DM at 40 kbits/s sampling rate is equal in performance with a 5-bit PCM system.

Extended-Range DM

A system termed *high-information* DM (HIDM, developed by M. R. Winkler in 1963) falls in the category of companded systems and encodes more information in the binary sequence than normal DM. Basically, the method doubles the size of the quantization step when two identical, consecutive binary values appear and takes one-half of the step after each transition of the binary train. The HIDM system is capable of reproducing the signal with smaller quantization and overload errors. This technique also increases the dynamic range. The response of HIDM compared with that of DM is shown in Fig. 12.3.21.

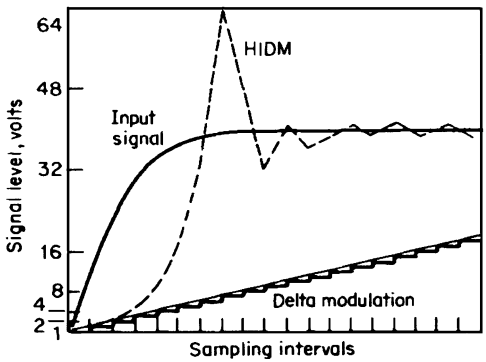

FIGURE 12.3.21 Step response for a high-information delta modulation.

Implementation of HIDM is similar to that of DM, as shown in Figs. 12.3.22 and 12.3.23, with the difference only in the demodulator. The flip-flop of Fig. 12.3.23 changes state on the polarity of the input pulses. While the impulse generator initializes the experimental generators each pulse time, the flip-flop selects either the positive or negative one. The integrator adds and smooths the exponential waveforms to form the output signal. The scheme has a dynamic range with slope limiting of 11.1 levels per pulse period, which is much greater than DM and is equivalent to a 7-bit linear-quantized PDM system.

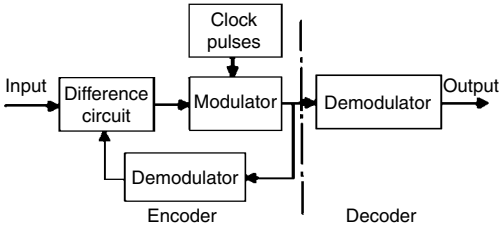

FIGURE 12.3.22 Block diagram of HIDM system.

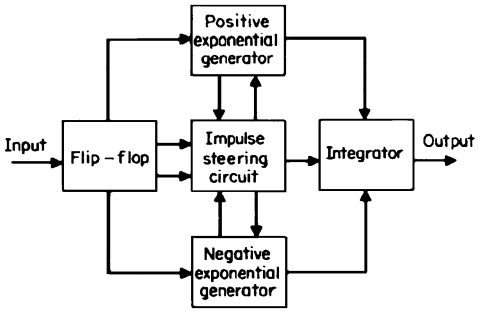

FIGURE 12.3.23 Block diagram of HIDM demodulator.

DIGITAL MODULATION

Digital modulation is concerned with the transmission of a binary pulse train over some medium. The output of, say, a PCM coder would be used to modulate a carrier for transmission. In PCM systems, for instance, the high-quality reproduction of the analog signal is a function only of the probability of correct reception of the pulse sequences. Thus the measure of digital modulation is the probability of error resulting from the digital modulation. The three basic types of digital modulation, amplitude-shift keying (ASK), frequency-shift keying (FSK), and phase-shift (PSK), are treated below.

AMPLITUDE-SHIFT KEYING

In ASK the carrier amplitude is turned on or off, generating the waveform of Fig. 12.3.24 for rectangular pulses. Pulse shaping, such as raised cosine, is sometimes used to conserve bandwidth. The elements of a binary ASK receiver are shown in Fig. 12.3.25. The detection can be either coherent or noncoherent; however, if the added complexity of coherent methods is to be applied, a higher performance can be achieved by using one of the other methods of digital modulation.

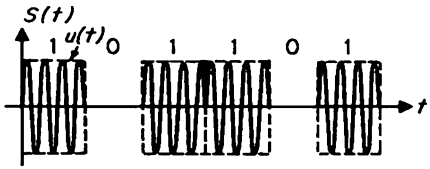

FIGURE 12.3.24 ASK modulation.

The error rate of ASK with noncoherent detection is given in Fig. 12.3.26. Note that the curves approach constant values of error for high signal-to-noise ratios.

The probability of error for the coherent detection scheme of Fig. 12.3.25*c* is shown in Fig. 12.3.27. The coherent-detection operation is equivalent to bandpass filtering of the received signal plus noise, followed by synchronous detection, as shown. At the optimum threshold shown in Fig. 12.3.27, the probability of error of marks and spaces is the same. The curves also tend toward a constant false-alarm rate, as in the noncoherent case.

FREQUENCY-SHIFT KEYING

In FSK the frequency is shifted rapidly between one of two frequencies. Generally, two filters are used in favor of a conventional FM detector to discriminate between the marks and spaces, as illustrated in Fig. 12.3.28. As with ASK, either noncoherent or coherent detection can be used, although in practice coherent detection is not often used. This is because it is just as easy to use PSK with coherent detection and achieve superior performance.

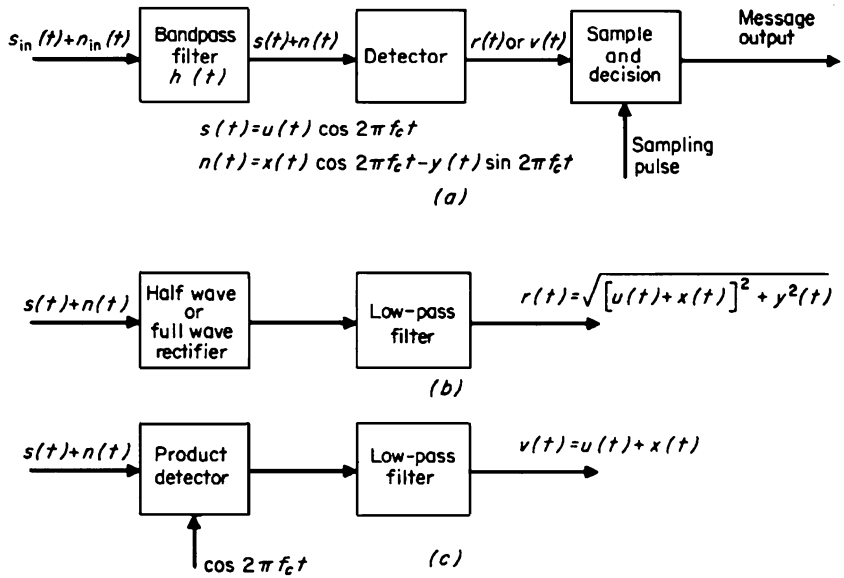

FIGURE 12.3.25 Elements of a binary digital receiver: (*a*) elements of a simple receiver; (*b*) noncoherent (envelope) detector; (*c*) coherent (synchronous) detector.

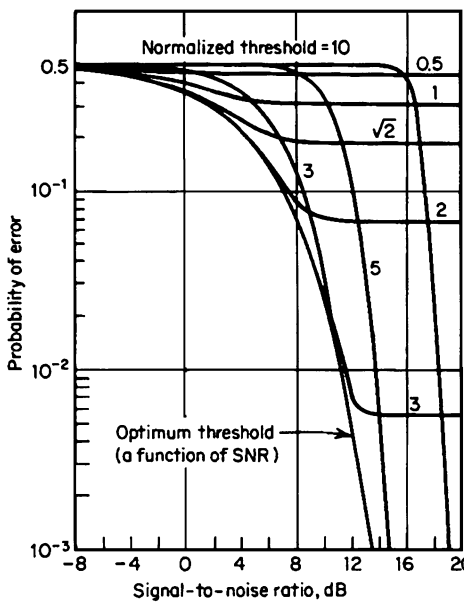

FIGURE 12.3.26 Error rate for on-off keying, noncoherent detection.

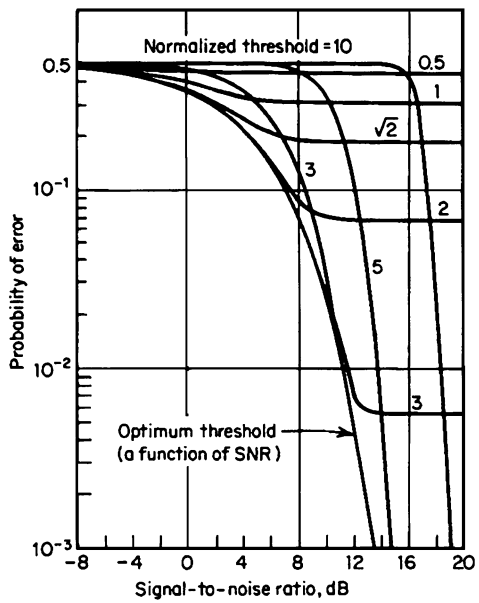

FIGURE 12.3.27 Error-rate for on-off keying, coherent detection.

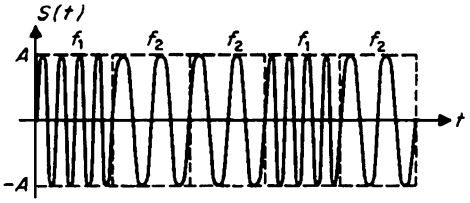

FIGURE 12.3.28 FSK waveform, rectangular pulses.

In the noncoherent FSK system shown in Fig. 12.3.29*a*, the largest of the output of the two envelope detectors determines the mark-space decision. Using this system results in the curve for noncoherent FSK in Fig. 12.3.30. Comparison of the noncoherent FSK error with that of the noncoherent ASK results in the conclusion that both achieve an equivalent error rate at the same average SNR at low error rates. FSK requires twice the bandwidth of ASK because of the use of two tones. In ASK, in order to achieve this performance, it is required to optimize the detection threshold at each SNR. The FSK system threshold is independent of SNR, and thus is preferred in practical systems where fading is encountered.

By synchronous detection of FSK (Fig. 12.3.29*b*) is meant the availability of an exact replica of each possible transmission at the receiver. The coherent-detection process has the effect of rejecting a portion of the bandpass noise. Coherent FSK involves the same difficulties as phase-shift keying but achieves poorer performance. Also, coherent FSK is significantly advantageous over noncoherent FSK only at high error rates. The probability of error is shown in Fig. 12.3.30.

PHASE-SHIFT KEYING

Phase-shift keying is optimum in the minimum-error-rate sense from a decision-theory point of view. The PSK of a constant-amplitude carrier is shown in Fig. 12.3.31, where the two states are represented by a phase difference of π rad. Thus PSK has the form of a sequence of plus and minus rectangular pulses of a continuous sinusoidal carrier. It can be generated by double-sideband suppressed-carrier modulation by a bipolar rectangular waveform or by direct phase modulation. It is also possible to phase-modulate more complex signals than a sinusoid.

There is no performance difference in binary PSK between the coherent detector and the normal phase detector, both of which are shown in Fig. 12.3.32. Reference to Fig. 12.3.32 shows that there is a 3-dB design advantage for ideal coherent PSK over ideal coherent FSK, with about the same equipment requirements. Practically, PSK can suffer if very much phase error $\Delta\phi$ is present in the system, since the signal is reduced by cos $\Delta\phi$. This phase error can be introduced by relative drifts in the master oscillators at transmitter or receiver

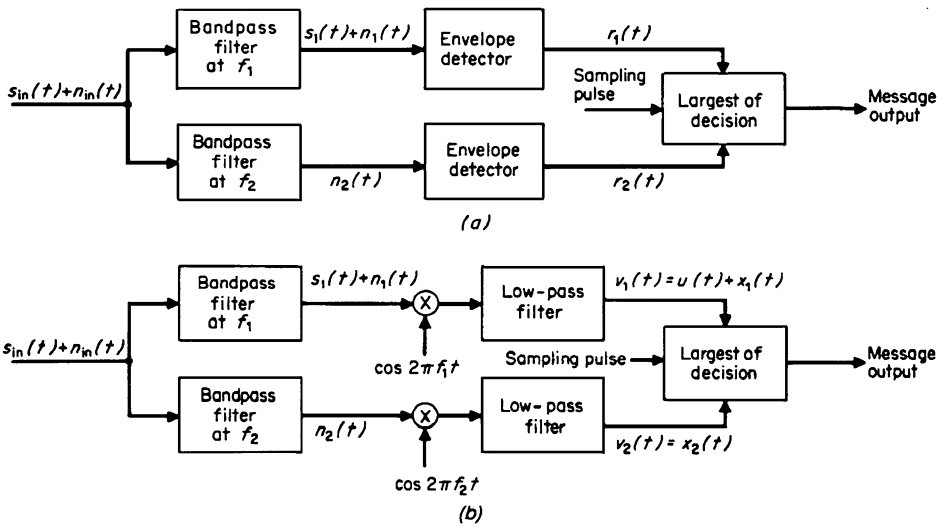

FIGURE 12.3.29 Dual-filter detection of binary FSK signals: (*a*) noncoherent detection tone f_1 signaled; (*b*) coherent detection tone f_1 signaled.

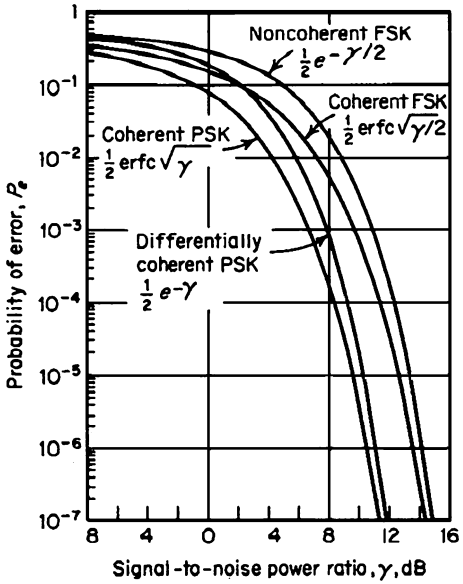

FIGURE 12.3.30 Error rates for several binary systems.

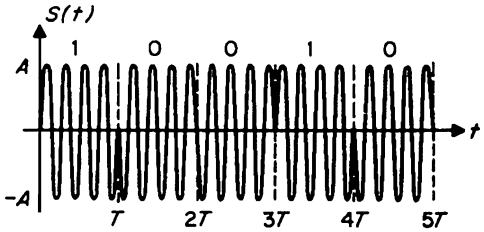

FIGURE 12.3.31 PSK signal, rectangular pulses.

or be a result of phase drift or fluctuation in the propagation path. In most cases this phase error can be compensated at the expense of requiring long-term smoothing.

An alternative to PSK is differential phase-shift keying (DPSK), where it is required that there be enough stability in the oscillators and transmission path to allow negligible phase change from one information pulse to the next. Information is encoded differentially in terms of phase change between two successive pulses. For instance, if the phase remains the same from one pulse to the next (0° phase shift), a mark would be indicated; however, a phase shift of π from the previous pulse to the next would indicate a space. A coherent detector is still required where one input is the current pulse with the other input the previous pulse.

The probability of error is shown in Fig. 12.3.30. At all error rates DPSK requires 3 dB less SNR than noncoherent FSK for the same error rate. Also, at high SNR, DPSK performs almost as well as ideal coherent PSK at the same keying rate and power level.

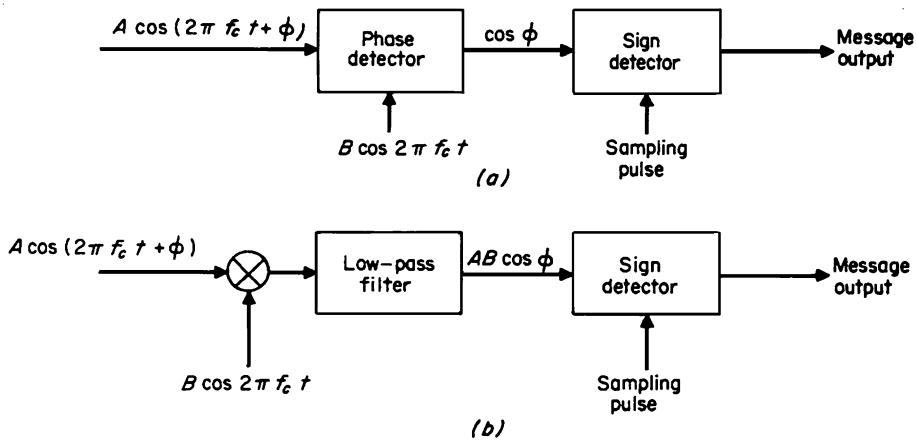

FIGURE 12.3.32 Two detection schemes for ideal coherent PSK: (*a*) phase detection; (*b*) coherent detection.