6 Signaling Principles

6.1 Introduction

In previous chapters, we learned that the communication system we are designing is based upon the transmission of energy between locations. Engineering practice has given some insights in how to optimize that energy transfer to enable efficient links for numerous users. In the previous chapter, we learned that these users are interested in communicating many different types of information, each type having unique properties that place requirements on the behavior of our communications system.

Putting the concepts of energy transmission and information delivery together yields the concept of modulation. The system engineer has to assure that information is delivered, or the communication network is useless. The choice of how best to transfer that information on the energy beam is the goal of the next two chapters. In this chapter, a review of basic modulation principles is presented with the goal of introducing certain themes for the system designer. The concept of the signal spectrum, signal-to-noise ratio (SNR), and the differences between analog and digital communication systems are important milestones for understanding in this chapter.

In addition, we will address the topic of baseband digital communication as an introduction to digital modulation. Analog communication systems differ from digital systems in the physical goal of the communication link, and this topic will be investigated in detail for the simple baseband digital link. The basic quality performance measure, bit error rate, will be developed and will be shown to depend upon signal design and other link factors not entire under the designer's control.

6.2 Modulation Overview

Modulation is a process whereby the message information is encoded on the energy carrier. If we assume the carrier wave is a cosine wave, we can imagine several means of modulating, or modifying, the carrier. The amplitude, frequency, or phase of the voltage or current waveform could be modified. Thinking of propagating electromagnetic radiation, we could even modulate radiated power or wave polarization. It is important to properly use the terminology. In our treatment, the message *modulates* the carrier.

6.2.1 Carrier Modulation for Information Transmission

The point of wireless communication is to use electromagnetic radiation techniques to communicate information. Guglielmo Marconi's "wireless" telegraph invention, inspired by Heinrich Hertz, was based on radiation emitted by a spark gap transmitter. (Note that Marconi was not the sole inventor.)

For a long time prior to the development of wireless, telecommunication was primarily achieved via wired telegraphy. This system transmitted an electrical code of "dots" and

"dashes" sent down a wire to a receiver. This scheme is classified as a baseband transmission scheme, since the signal spectrum energy is centered around DC.
The wireless telegraphy system used a bandpass transmission system, basically transmitting burst of noise in a relative narrow bandwidth to represent the dots and dashes.

Baseband systems thrive today in communication carried over cables and twisted wire pairs, although they don't use bursts of noise to represent information, but rather pulse waveforms. This "on/off" energy transmission scheme worked well enough for telegraphy, but as demand grew for voice applications the concept of using a modulated carrier began to dominate. Carrier modulated schemes are also classified as bandpass transmission schemes, since the signal spectrum energy is bandpass in nature and centered on a specific center frequency (the carrier frequency).

The term "radio" became associated with carrier modulation systems only after the use of modulated carrier technology for voice transmission became common. This terminology persisted well into the digital modulation age. Interestingly, radio technology is now called wireless again.

Modulated carrier systems offer several advantages over baseband systems, and their use has grown to overwhelmingly dominate the radio world. Figure 6-1 enumerates some comparison criteria between the two techniques, showing the relative advantage of carrier modulation over baseband from the system perspective.

Criterion	Baseband	Carrier Modulated (Bandpass)
Interferers	DC, AC power, and any other baseband signal	Only those signals transmitting in-band
Antenna Dimensions (scales with f)	Huge	Manageable (drives toward higher frequencies)
Filter Design	Difficult to design high-Q filters at low frequencies	Easier to design high-Q filters at high frequencies
Multiplexing	Only TDM possible	TDM and FDM possible
Noise	Higher, shaped background noise PSD at lower frequencies	Relatively low, white background noise PSD at higher frequencies

Figure 6-1 Comparison of baseband and bandpass communication schemes.

6.2.2 Modulated Carrier Communication Systems

The architecture of a modulated carrier system is actually quite simple from one perspective. The system consists of three basic blocks: a transmitter, a channel, and a receiver. The transmitter receives input information (the message) and prepares it for transmission through the channel. This block must prepare the information for reliable transmission, so the design of the transmitter is performed with much consideration of the channel and receiver.

The channel consists of the medium through which the information is transmitted. It might be deep space, ocean, a tape recording, or a wire. The channel may delay or distort a signal, and add interference and noise before the signal arrives at the receiver. The receiver has the hard job of trying to recover the information from the received signal. Theoretically, the receiver's job is to undo the work of the transmitter. The problem is that the channel effects have altered the signal such that the job may seem next to impossible.

6.2.2.1 System Modulation Architecture

This section focuses on the radio communication link proper. There is more to wireless communication than the radio link, of course. There are two major assumptions made concerning the functional operation of the radio communication link. First, the radio link has and requires no knowledge of the content of the message it transmits. Second, the radio link is completely self-contained. In other words, operations within the radio link affect only the radio link.

The communication link architecture has traditionally been described in block diagram form as shown in Figure 6-2 and Figure 6-3, derived from the classic work of Claude Shannon. It is important to note that the message/signal flows along a single path through the functional blocks of the system. Notice that each function is paired, or that each operation performed in the transmitter is un-done in the receiver.

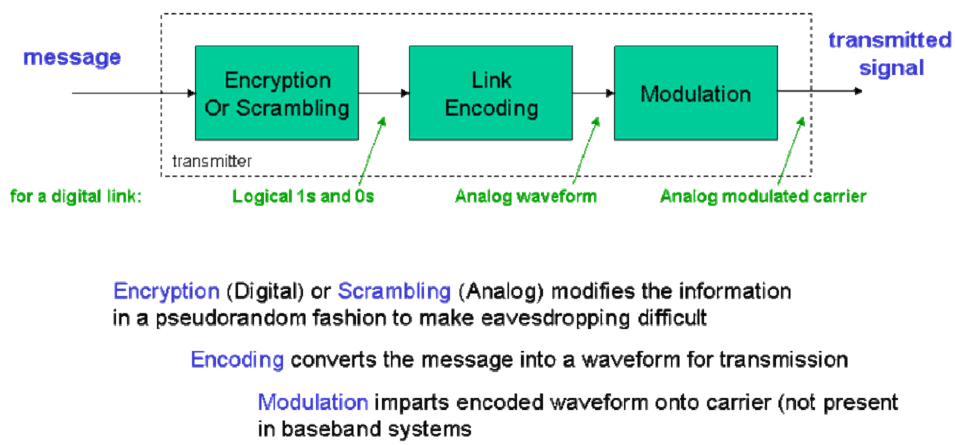

Figure 6-2 Transmitter high-level design.

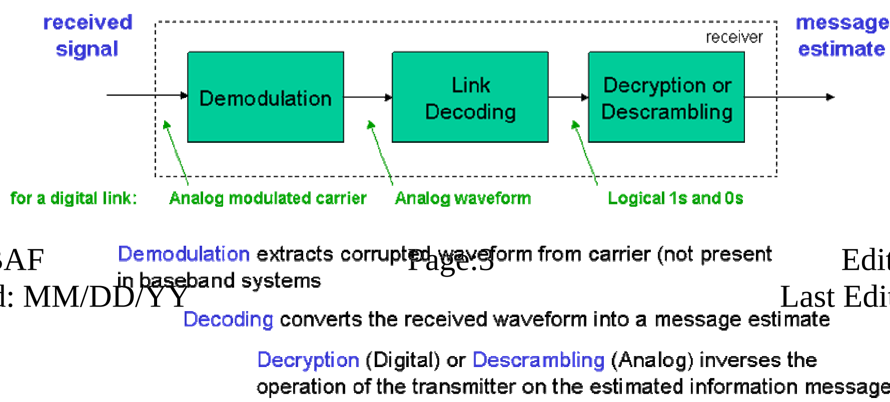

Figure 6-3 Receiver high-level design.

Within the transmitter are four major functions. The first function, encryption and scrambling, provides a means of securing the data during transmission (encryption) and preventing undesirable conditions (such as long strings of “1”s) for modulators and demodulators (scrambling). The second major function is link encoding. This is an operation performed upon the message data whereby overhead is added to allow for error checking and/or correction. Since this is undone in the receiver, it is completely transparent to system above the radio link The third major function is modulation, which is the subject of this chapter. This is the process of imprinting the message onto a carrier waveform for transmission. It turns out that there is much theory and design in the choice of modulation, producing major impact on system performance. The fourth major function, not shown, is preparing the modulated carrier wave for transmission, which is not dealt with here. This might involve amplification and filtering for transmission through the channel.

6.3 Review of AM and FM Principles

The study of modulation often begins with something familiar to students – radio. Broadcast AM and FM have been important media for quite some time. The modulation formats therefore serve as important comparison points for other types of modulation. In this section, the formats are reviewed and investigated for a number of specific performance parameters.

6.3.1 Analog Modulation Principles

Analog modulation refers to the modulation of a carrier waveform by an analog message waveform. An analog waveform is one whose values vary in analogy with some physical measurable quantity. More properly, an analog signal should be classified as a continuous-time, continuous-valued waveform. Music, speech, and ECG waveforms are examples of analog signals.

The general form of an analog modulated transmitted waveform is

$$g(t) = A(t)\cos\left\{2\pi\left[f_c + \Delta f(t)\right]t + \theta(t)\right\}, \qquad (6\text{-}1)$$

where a cosine carrier waveform is assumed, and $f_c=\omega_c/2\pi$ is the carrier frequency. Each of the functions A(t), Δf(t), and θ(t) are continuous functions of time. A(t) is the carrier amplitude, Δf(t) is the frequency deviation, and θ(t) is the carrier phase.

The analog message can modulate the amplitude, frequency, or phase of the carrier, and in some cases, a combination of these parameters. A critical point distinguishing analog modulation from digital modulation systems is that the analog waveform is the quantity being transmitted for the analog system. In a digital system, only the digital data is being transmitted, not the waveform per se. This results in the performance measures for analog systems being quite different from those for digital systems. Analog systems are evaluated with respect to their fidelity, or the degree to which they reproduce the input waveform down to the finest detail, and the extent to which noise degrades the ability to perceive that signal. The key performance parameters of an analog system are thus gain, bandwidth, linearity, and noise immunity.

Digital systems, on the other hand, are concerned only with the rate at which logical bits are transmitted and whether they are received in error, not caring at all about reproducing the original waveform. Additionally, the efficiency of transmitting a high data rate through a limited bandwidth is important. The key performance parameters of a digital system are thus throughput (data rate), bit error rate (BER), and spectral efficiency.

6.3.2 Amplitude Modulation

6.3.2.1 Amplitude Modulation: Waveforms and Spectra

Amplitude modulation is described as modulation of the carrier amplitude by an analog message. The carrier's frequency and phase remain unmodulated. Thus, the transmitted signal model can be modeled as

$$g_{AM}(t) = x(t)\cos(2\pi f_c t) \tag{6-2}$$

The amplitude term x(t) is derived from the message m(t). The specific relationship between x(t) and m(t) specifies the particular form of amplitude modulation, which will be discussed below. The modulation system for AM is very simply modeled as shown in Figure 6-4. This system can be implemented using analog multiplier ICs or mixer circuits to perform the multiplication function.

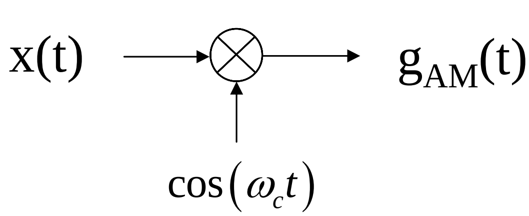

Figure 6-4 Generic amplitude modulator.

The modulation property of the Fourier Transform allows us to predict the spectrum of the amplitude modulated signal. Given the Fourier Transform pair $x(t) \Leftrightarrow X(f)$,

$$g_{AM}(t) = x(t)\cos(2\pi f_c t) \quad \Leftrightarrow \quad G_{AM}(f) = \frac{1}{2}\left[X(f - f_c) + X(f + f_c)\right]. \qquad (6\text{-}3)$$

There are several related modulation schemes which may be loosely termed amplitude modulation. The waveforms and power spectra associated with these formats for an example waveform are shown in Figure 6-5 and Figure 6-6. In the figures, the message used is a triangle wave with fundamental frequency of 20 kHz, and the carrier is a 100 MHz sinusoid.

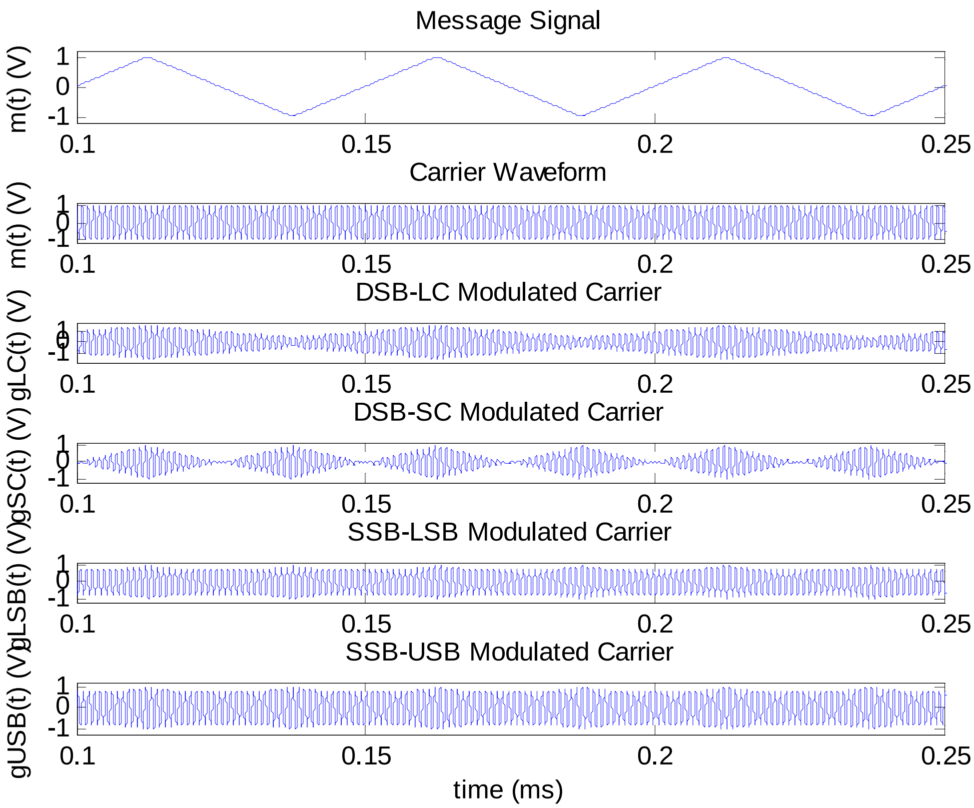

Figure 6-5 Example of waveforms for various amplitude modulation formats.

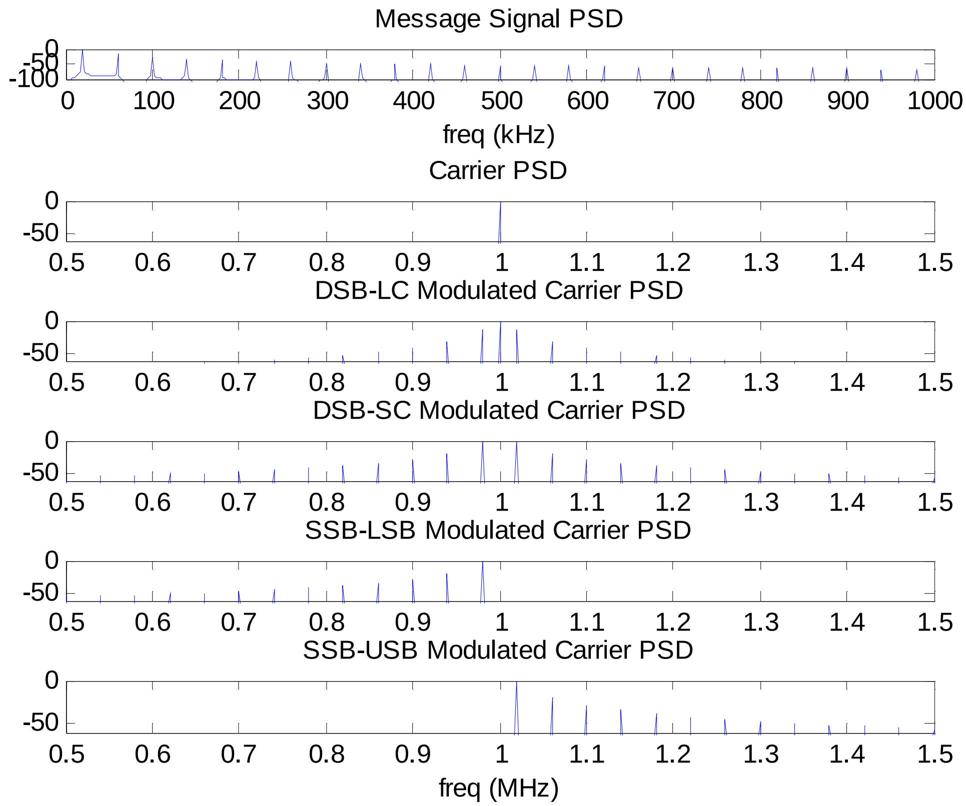

Figure 6-6 Power spectra for the various amplitude modulation formats.

The first two schemes shown are modeled as described above as simple amplitude modulation of a cosine carrier waveform. The first, called Double-Sideband – Large-Carrier, or DSB-LC, is the format used in commercial broadcast AM radio. The time domain waveform has an envelope which normally never dips below zero. This is understood by viewing the model for this format:

$$g_{DSB-LC}(t) = D\left(1+\mu\frac{m(t)}{\max\left[m(t)\right]}\right)\cos(\omega_c t). \qquad (6\text{-}4)$$

This format is most easily distinguished in its spectrum by the presence of the carrier "spike". This term results from the DC offset (the additive "1" in the expression above) multiplying the cosine, thus creating an unmodulated carrier term. The DC offset also creates a waveform envelope which is does not dip below zero, provided the modulation index μ is kept less than unity. This allows the waveform to be demodulated by either a synchronous demodulator or an envelope detector.

The second format is called Double-Sideband – Suppressed-Carrier, or DSB-SC. This format does not have the DC offset present in x(t) as in DSB-LC, and, as a result, the waveform's envelope swings positive and negative. Synchronous demodulation must be used to recover the message from a DSB-SC waveform due to this envelope behavior.

$$g_{DSB-SC}(t) = Dm(t)\cos(2\pi f_c t) . \quad (6\text{-}5)$$

The DSB-SC format features a spectrum without the carrier spike ("suppressed"). It turns out the carrier spike has much to do with the type of transmitters and receivers employed, and invokes a good system-level design discussion (see design note below).

The third and fourth amplitude modulation schemes are called Single-Sideband (SSB) formats, the former being Upper Sideband (USB) and the latter Lower Sideband (LSB). We are not quite ready to discuss the modulation format for these signals, but it is apparent that their transmitted signal spectrum occupies only ½ the bandwidth of the DSB formats. The architecture complexity of the modulator and demodulator has been traded for better bandwidth efficiency in these schemes.

For now, we can design a system to create SSB signals from DSB-SC signals using the architecture shown in Figure 6-7. The drawback is that the requirements placed on the bandpass filter are quite extreme. This filter must show a very steep rolloff at the carrier frequency to separate the upper from the lower sidebands. Later in this section we will introduce another method for producing SSB signals which gets around this painful design requirement.

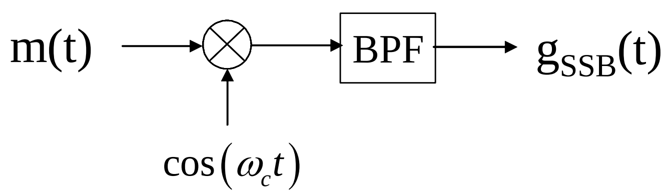

Figure 6-7 Filtering system for creating SSB waveforms.

Notice that the various amplitude modulation formats occupy a specific transmission bandwidth. Given the message bandwidth B, some formats occupy a bandwidth 2B and are thus term double-sideband transmission formats. Others occupy a bandwidth B and are termed single-sideband transmission formats. Yet, each format is able to transmit the message accurately, indicating some information redundancy in the double-sideband formats.

Example – plot the time waveform for specific AM specifications and extract modulation index
Example – Given specifications, plot the spectrum of AM signal

6.3.2.2 AM demodulation

With any communication system architecture, it is found that the receiver design is more complicated than the transmitter. This will be true of demodulators as compared to modulators as subcomponents of these systems. In this system, we describe two chief demodulation techniques for the amplitude modulation schemes described above.

The optimum demodulator for all amplitude modulation formats is the synchronous demodulator shown in Figure 6-8. This system recovers the input message signal provided noise and distortions are minimized and the local oscillator is well synchronized to the received carrier.

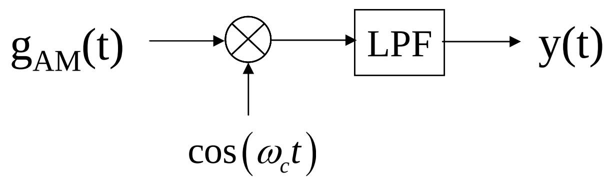

Figure 6-8 Synchronous demodulator for amplitude demodulation.

There are several important features characterizing this demodulator. Since it is the basis for many of the demodulator structures we will encounter, it is useful to discuss its structure. The first feature is its structure. It is composed of a mixing operation (multiplication by a local oscillator) followed by a lowpass filter. The operation of this system is best viewed in the frequency domain using Fourier transform techniques. Another feature is the fact that the local oscillator is assumed to be synchronous with the carrier. This is the source of the demodulator's name. Not shown is the method of synchronization, but it is assumed that the local oscillator is locked (dynamically) in frequency and phase to the received carrier. Finally, there is a lowpass filter, which is used to extract the baseband message from higher frequency components present after mixing. The design constraints on this filter are typically less stringent and based mostly meeting SNR requirements.

Another type of demodulator which is only useful for DSB-LC demodulation is the envelope detector. This detector, an example of which is shown in Figure 6-9, simply tracks the positive envelope of the modulated carrier and outputs that envelope as the baseband output signal.

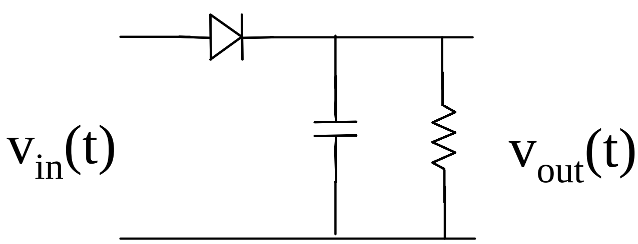

Figure 6-9 Simple envelope detector for DSB-LC waveforms.

This demodulator operates by tracking the envelope of the carrier, and lowpass filtering this envelope signal to extract the message. If the envelope of the carrier is not positive-definite, then the message signal will not be able to be recovered (as would be the case, e.g., if DLB-SC modulation were used, or if DSB-LC with $\mu>1$ were used).

Example – derive output of synchronous demodulator given DSB-SC and DSB-LC input

Design Note: DSB-LC vs DSB-SC example

As an example of how engineers trade modulation formats, that is, decide which format is best for a given application, consider a radio control helicopter application. The first task is to simply control the helicopter from some ground station. In the second task, the helicopter will send an analog signal back to the ground station. Which of the two formats, DSB-LC or DSB-SC, might be best suited to each application, given that the signal bandwidth and fidelity requirements are equal for each case?

At first, it might seem an arbitrary choice. But the system engineer must think about SWaP (Size, Weight, and Power) and not just theory. Since the radio control helicopter is a small flying machine, its total weight and its total DC power consumption (i.e. battery size) must be minimized. For the first task, this is best achieved by using DSB-LC where an envelope detector can be implemented as the receiver on the helicopter. The ground station, which has more DC power available, can afford the relatively inefficient transmission of the large carrier spike of DSB-LC in order to allow the use of the simple and low power envelope detector on the helicopter. For the second application, where the helicopter is transmitting a signal back, DSB-LC is not a good choice. Transmitting the large carrier spike of DSB-LC requires much DC power, which hurts the SWaP of the helicopter. DSB-SC, which transmits much less power, is a much better choice for a helicopter transmitter. The ground station can afford the complexity and power to implement a synchronous demodulator to receive the DSB-SC signal. These examples show that secondary considerations sometimes drive the modulation or signal design trade.

6.3.2.3 AM in Presence of Noise

The above-mentioned modulator and demodulator pairs work wonderfully together. However, every circuit, and every channel, contributes distortion and noise to the signals they transmit. Here we deal with the impact of noise on amplitude modulation systems.

6.3.2.3.1 Noise model overview

When noise is present, the demodulation of signals becomes more difficult. For the case of additive white Gaussian noise (AWGN), however, simple analysis can predict what will happen in the demodulator and lead to a measure of performance universally used to compare recovered signal quality, the Signal-to-Noise Ratio (SNR).

Noise is a random process, arising from some naturally occurring physical process. Noise is differentiated here from interference, which describes the impact of other artificial signals upon the reception of the desired signal. The reason for the differentiation is simple – interfering signals often have structure. This structure can lead to methods of rejecting interference not effective with noise.

Our model for noise must account for two measurable quantities. The first is the distribution of the noise amplitude measurements, which is described by an amplitude probability density function (PDF). The second describes the frequency content of the noise process, and is described by the noise power spectral density (PSD).

For a noise process n(t), we define the PDF to be $p_n(\alpha)$, and the most common distribution function is the Gaussian or Normal distribution. The noise we consider here will be Gaussian noise, meaning that the samples of its amplitude follow a Gaussian or normal distribution. Given a Gaussian noise process with mean μ_n and variance σ_n^2, the PDF $p_n(\alpha)$ for the noise is described by

$$p_n(\alpha) = \frac{1}{\sqrt{2\pi\sigma_n^{\,2}}} \exp\left\{ \frac{-(\alpha - \mu_n)}{2\sigma_n^{\,2}} \right\}. \qquad (6\text{-}6)$$

It is noted that the Gaussian random process is completely specified by its mean and variance.

The power spectral density (PSD) is a function derived from the correlation statistics of the noise, which is beyond the scope of the present discussion. More conveniently, the PSD $S_n(f)$ can be thought of as the distribution of signal power with frequency.
The total power of a function can be calculated from its power spectral density using

$$P_n = \int_{-\infty}^{\infty} S_n(f)\,df \quad \text{(W)}. \qquad (6\text{-}7)$$

Here we use the term P_n to denote the total noise power. Since the noise is assumed zero mean, meaning there is no DC component, the ac power is the same as the average power.

For the noise we are considering, the power is distributed equally in all frequencies. Thinking of white light as a combination of many colors, this type of noise is called white noise. White noise technically has a flat PSD (i.e. $S_n(f)$=constant) over all frequencies. In common practice, it is sufficient to assure that the PSD is flat over the system bandwidth.

It is important to keep track of the type of PSD being used to describe the noise. Spectrum analyzers display a one-sided PSD of the signal. Often, in analyses and textbooks, a two-sided spectrum is given. Of course, the one-sided spectrum is obtained by simply doubling the positive frequency content of the two-sided spectrum, and then displaying only the positive frequency content. The two-sided PSD of white noise is defined using a specific nomenclature:

$$S_{n,white}(f) = \frac{N_0}{2} \quad \text{(W/Hz)}. \tag{6-8}$$

Example – PSD of thermal noise
Example – Power in bandwidth of white thermal noise

6.3.2.3.2 Noise and Linear Systems

One of our goals in this noise discussion is to be able to determine the average noise power at some point in a system. Keeping in mind that noise is a random process, we will need to develop techniques that allow for power computation from the noise PSD. Given a linear system h(t), with input x(t) and output y(t), we can define the basic input/output relationships from system theory:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\gamma) h(t-\gamma) d\gamma , \tag{6-9}$$

$$Y(f) = X(f) H(f). \tag{6-10}$$

.

Here, the Fourier transform pair is denoted as $h(t) \Leftrightarrow H(f)$.

Since noise is a random process, functions such as N(f), the Fourier transform of n(t), while defined, remain random and thus immune to many of our analytic efforts. However, we can work with PSDs of signals and avoid this inconvenience. Given an input X(f) and output Y(f), the input-output relationship for the PSDs of the signals becomes

$$S_y(f) = \left|H(f)\right|^2 S_x(f). \tag{6-11}$$

Here, we see a convenient relationship between the input and output power spectral density. In our current problem of noise passing through a linear system, we would write

$$S_{n_o}(f) = \left|H(f)\right|^2 S_{n_i}(f). \tag{6-12}$$

Example – noise power after Butterworth LPF

6.3.2.3.3 Signal to Noise Ratio

The usual method of signal quality given the presence of noise is the signal-to-noise ratio, or SNR. The SNR seems like a pretty simple topic, but there needs to be an understanding of how it is measured before it is determined. There are also different types of SNR, as will be described in other parts of this text.

First, we define the SNR as the ratio of signal power to noise power within a specified bandwidth:

$$SNR = \frac{P_S}{P_N}. \qquad (6\text{-}13)$$

It is important to remember that the SNR is defined within a bandwidth, such as the receiver bandwidth. This defines the bandwidth over which the noise is integrated in calculations, and is typically a receiver filter bandwidth.

Often, there are two SNRs of interest in link design: the SNR produced at the input to the receiver, and the SNR produced at the output of the receiver front-end filter. The former, which is associated with the modulated carrier, is called the carrier-to-noise ratio (CNR) and is a design parameter for the link budget. The latter is the SNR and is a design parameter for the receiver and link budget.

6.3.2.3.4 Narrowband Noise Modeling

Since the received signal is often bandpass in nature (such as a modulated carrier signal), and a narrowband bandpass signal more specifically, we will adopt a matching noise model which allows for easier analysis. A narrowband signal is defined as a signal whose bandwidth is small relative to its center frequency. Such a signal could correspond to a modulated carrier scheme such as AM, or a narrowband process such as non-white noise whose PSD is confined to a narrow range of frequencies about some center frequency. The latter is a description of the original wireless telegraphy systems demonstrated in the late 19^{th} century.

Any narrowband process or signal may be modeled using the functional form

$$\begin{aligned} x_{NB}(t) &= x_I(t)\cos(\omega_c t) - x_Q(t)\sin(\omega_c t) \\ &= \mathrm{Re}\left\{A(t)e^{j\theta(t)}e^{j\omega_c t}\right\} \end{aligned}. \qquad (6\text{-}14)$$

Here, the subscripts I and Q refer to in-phase and quadrature, respectively. In-phase denotes association with the cosine component, and quadrature (meaning $\pi/2$ out of phase) denotes association with the sine component. The term A(t) is the slowly-varying envelope of the narrowband signal or process, and $\theta(t)$ is the phase.

Here, we wish to model narrowband noise, so we write the model above as consisting of in-phase and quadrature noise random processes

$$n_{NB}(t) = n_I(t)\cos(\omega_c t) - n_Q(t)\sin(\omega_c t). \qquad (6\text{-}15)$$

The two processes $n_I(t)$ and $n_Q(t)$ are just random noise processes, and are Gaussian if $n_{NB}(t)$ is Gaussian. The narrowband noise $n_{NB}(t)$ is Gaussian, and it is additive, but it is certainly not white. Given that it is Gaussian, its PDF is described by a mean, assumed to be zero, and a variance σ_n^2. The mean of $n_I(t)$ and $n_Q(t)$ will be zero as well. The total noise variance is related to the variance of the I and Q processes in a simple fashion: $\sigma_n^2 = \sigma_{n_I}^2 = \sigma_{n_Q}^2$. At first this seems incorrect, but the example below shows how this result occurs.

The PSD of the two noise components are related to the PSD of the narrowband noise as well.

$$S_{n_I}(f) = S_{n_Q}(f) = LPF\{S_n(f-f_c) + S_n(f+f_c)\}. \qquad (6\text{-}16)$$

The notation LPF{} denotes lowpass filtering. This model allows for modeling the effects of noise in arbitrary modulation schemes and demodulators.

Example – mean-square of nb noise.

6.3.2.3.5 AM demodulation with noise

We are now in a position to determine the quality of the performance of reception for an AM receiver. For this analysis, we assume the use of the synchronous demodulator as shown in Figure 6-10. The input x(t) consists of a received AM signal with AWGN.

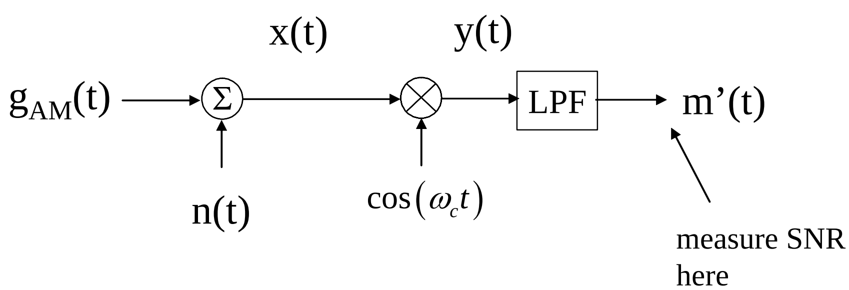

Figure 6-10 Synchronous demodulation of received AM signal corrupted by noise.

The input $g_{AM}(t)$ is assumed to be a strictly bandlimited signal with bandwidth BW. The format is arbitrary, but DSB-SC is assumed for this illustration. The AWGN n(t) is described by its PSD level $N_0/2$. The received signal input to the demodulator is thus modeled as the sum of two independent components:

$$\begin{aligned} x(t) &= g_{AM}(t) + n(t) \\ g_{AM}(t) &= Dm(t)\cos(\omega_c t) \\ n(t) &= n_I(t)\cos(\omega_c t) - n_Q(t)\sin(\omega_c t) \end{aligned}. \qquad (6\text{-}17)$$

The total input noise power in the transmission bandwidth is found by integrating the narrowband noise PSD:

$$\sigma_n^{\ 2} = 2\int_{f_{low}}^{f_{high}} S_n(f)df \ . \tag{6-18}$$

The received signal x(t) is next mixed with the local oscillator. The result of this exercise is an operation on both the components $g_{AM}(t)$ and n(t) which creates signal components at baseband and at twice the carrier frequency.

$$\begin{aligned} y(t) = Dm(t)\left[\frac{1}{2}+\frac{1}{2}\cos(2\omega_c t)\right]+ \\ \frac{1}{2}n_I(t)+\frac{1}{2}n_I(t)\cos(2\omega_c t)+\frac{1}{2}n_Q(t)\sin(2\omega_c t) \end{aligned} \quad .(6\text{-}19)$$

The action of the LPF is to reject the high frequency components in this signal, and to limit the noise power exiting the system:

$$m'(t) = \frac{1}{2}D\cdot LPF\left\{m(t)+n_I(t)\right\}. \tag{6-20}$$

The total signal power exiting the filter, noting that the message is strictly bandlimited to less than the assumed ideal LPF bandwidth, is

$$\left\langle m'^2(t)\right\rangle = \frac{1}{4}D^2\left\langle m^2(t)\right\rangle. \tag{6-21}$$

To find the total noise power, first note that the system through which the signals have passed is linear, and so we can separate the components to extract the output noise term

$$n_o(t) = \frac{1}{2}D\cdot LPF\left\{n_I(t)\right\}. \tag{6-22}$$

For the assumed ideal LPF, the output noise power follows directly

$$\left\langle n_o^{\ 2}(t)\right\rangle = \frac{1}{4}D^2\left\langle n_I^{\ 2}(t)\right\rangle = \frac{1}{4}D^2\sigma_n^{\ 2}. \tag{6-23}$$

The SNR at the receiver output is thus $SNR = \left\langle m^2(t)\right\rangle \big/ \sigma_n^{\ 2}$. SNR values less than 10 dB give poor quality listening experiences. Generally, SNRs greater than 20 dB are desired, with some systems seeking SNRS well above 60 dB.

Example – SNR with Butterworth LPF

6.3.3 Frequency Modulation overview

Analog amplitude modulation systems dominated wireless/radio systems in the first half of the 20th century. Without a doubt, the scheme has earned a prominent place in communications history. However, there are some criticisms of the format that led engineers to develop something better.

AM receivers, while working well with large received signal strengths, suffered in situations where the received noise and interferers impacted the SNR. This was mostly a function of two features of the modulation scheme. First, the fact that the message modulated the amplitude of the carrier meant that any noise corrupting the amplitude of the signal anywhere in the path would corrupt the output of the signal. This lead to problems with excessive static with AM systems. Second, the relatively narrowband nature of the modulation scheme led to problems with interference. Since other interferers were also narrowband, their power spectrum would overlap the power spectrum of the AM signal. It is impossible to remove the effects of the interferer, and AM is especially susceptible to interference.

Frequency modulation was designed to combat these problems. First, by modulating the frequency of the carrier, some of the amplitude noise could be rejected. Also, the nature of the modulation acted to spread the signal energy over a wider bandwidth. This, the design thinking went, would prevent a narrowband interferer from impacting the entire signal. The result was greatly improved reception. The Steely Dan song said it best – "no static at all…FM" [1].

6.3.3.1 Frequency Modulation: Waveform and Spectrum

Analog frequency modulation has dominated the music broadcast industry due to its ability to deliver higher fidelity reception with better noise immunity. Part of this is due to the FM broadcast standard, but much of the quality is due to the modulation format itself.

Frequency modulation is an example of a class of nonlinear modulation called angle modulation, where the phase of the carrier is modulated with a message waveform. The amplitude of the carrier is fixed, and the received signal amplitude is clipped in the receiver to remove amplitude fluctuations in the receiver. The material in this section follows in part the discussion of Stremler [2].

We first need to have a clear understanding of what we mean when we say the phase of the carrier. We repeat the general description of a sinusoidal carrier, here with fixed amplitude and time varying phase:

$$g(t) = A\cos\{\theta(t)\}. \qquad (6\text{-}24)$$

Now, the phase of the carrier $\theta(t)$ is related to its (instantaneous) frequency $f_i(t)$ by the following:

$$\theta(t) = \int_0^t 2\pi\ f_i(\tau)\,d\tau + \theta_0 \qquad\qquad f_i(t) = \frac{1}{2\pi}\frac{d\theta}{dt}. \qquad (6\text{-}25)$$

The term θ_0 is an arbitrary phase constant.

Angle modulation may be viewed as either a modulation of the carrier's instantaneous frequency via $f_i(t)$, or the carrier phase via $\theta(t)$. If the message is linearly proportional to $f_i(t)$ we call the modulation format frequency modulation (FM). If the message is linearly proportional to $\theta(t)$ we call the modulation format phase modulation (PM). We will not address PM directly here.Note that since the transmitted signal g(t) is not linearly derived from the message signal, this modulation format is nonlinear.

For FM, the message directly modulates the carrier instantaneous frequency, hence

$$f_i(t) = f_c + k_f m(t) \qquad\qquad \theta(t) = \omega_c t + \int_0^t k_f m(\tau)d\tau + \theta_0. \qquad (6\text{-}26)$$

The proportionality constant k_f combines with the peak magnitude of m(t) to create a term called the peak frequency deviation Δf. This describes the maximum extent that the carrier frequency deviates from the unmodulated carrier frequency f_c while modulation is occurring.

An example best shows the process. An input triangle waveform message is modulating a sinusoidal carrier in Figure 6-11. The input waveform and output signal are shown, along with the corresponding plot of instantaneous frequency. In the figure, a normalized triangle wave of frequency 20 kHz is used as a message signal. A frequency waveform is derived from the message as $f_i = f_c + k_f m(t)$. The carrier is a 1 MHz sinusoid. The peak frequency deviation is 500 kHz, as is noted in the second plot. As shown in the third plot, the frequency of the carrier is greatly modulated by the message waveform, due to the large frequency deviation. A close-up of the carrier waveform covering one message period is shown in the fourth plot, showing the variations in carrier frequency.

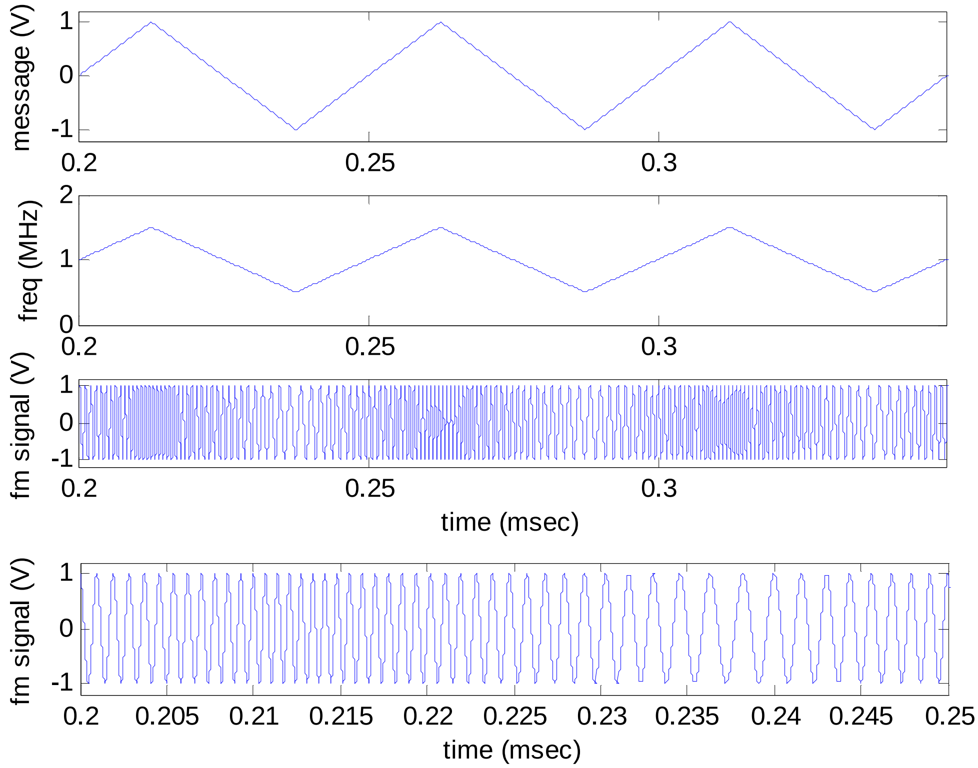

Figure 6-11 Example of FM, showing relationship between triangle wave message, carrier instantaneous frequency, carrier waveform, and a close-up of the carrier waveform.

The spectrum for an FM signal is a bit more difficult to predict, since the modulation is nonlinear. This being the case, the use of Fourier Transform properties is not available to quickly predict the spectrum. In fact, only in the case of simple messages such as sinusoids may the spectrum be predicted analytically.

The spectrum of the example message from Figure 6-11 is shown in Figure 6-12. In this figure, the power spectrum of the message is shown in the first plot, and the PSD of the modulated carrier is shown in the second plot. For this example, the carrier frequency was changed to 100 MHz, and the peak frequency deviation was changed to 75 kHz to mimic a commercial FM broadcast. The triangle wave message parameters are the same as those for the previous example.

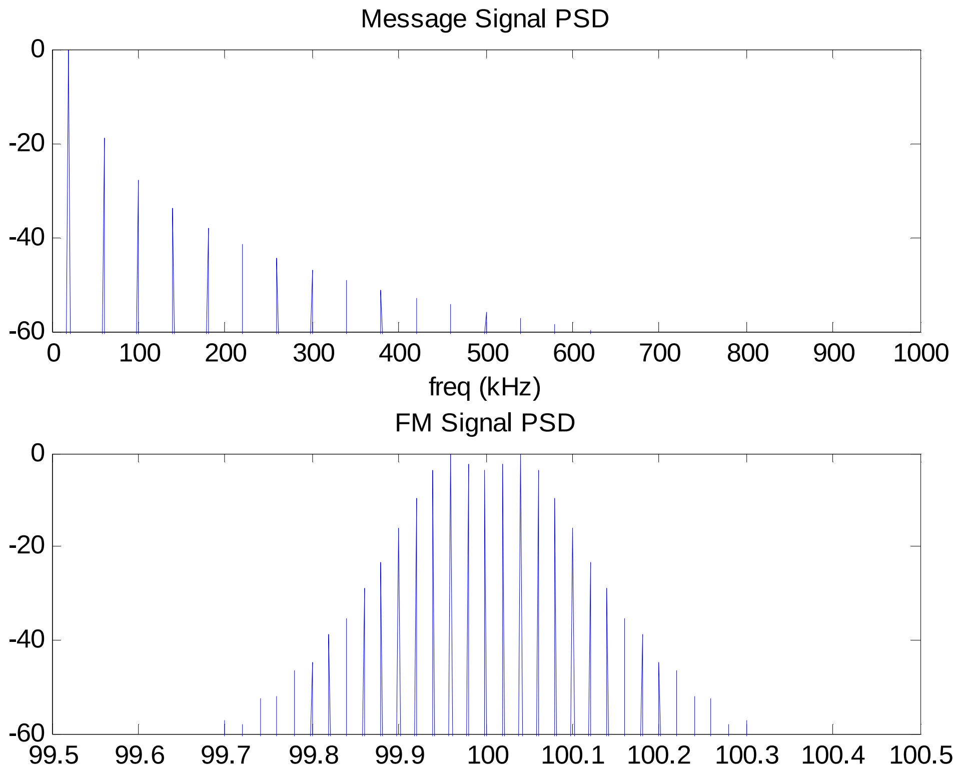

Figure 6-12 Spectrum of triangle wave message and resulting FM modulated carrier.

The peak frequency deviation for this signal is 75 kHz, and the message fundamental frequency is 20 kHz. Yet the bandwidth of the transmitted signal does not seem to correlate with either of these quantities. Estimated the spectrum of FM signals is not nearly as simple as for AM since the modulation format is nonlinear. Fortunately, there is a way to estimate the transmitted signal bandwidth.

6.3.3.2 Signal spectrum spreading – Carson's Rule

In AM, with bandlimited messages, the modulated carrier spectrum is found to be also bandlimited. This is not the case for FM signals. Estimating the bandwidth of an FM signal is difficult, as the spectrum is not typically finite even when the message is bandlimited.

For any signal, the definition of bandwidth must depend upon some agreed-upon standard, e.g. that range of frequencies which contains 90% of the signal power. The FM modulation process is often characterized by a unit-less figure of merit called the FM beta (β) or FM modulation index. This figure was originally defined for the case of sinusoidal modulation only, but can be extrapolated to cases of arbitrary messages. This FOM may be defined in the general case as the peak frequency deviation divided by the message bandwidth:

$$\beta = \frac{\Delta f}{BW_m}. \quad (6\text{-}27)$$

There are two effects happening simultaneously in FM. First, the message amplitude causes the instantaneous frequency of the carrier to vary between extremes defined by the peak frequency deviation and the carrier frequency. Thus, we could imagine the instantaneous frequency following the message as in the second plot of Figure 6-11. At the same time, the message is oscillating at its fundamental frequency, and this message spectrum consists of many harmonics (for the case of a periodic message). This message oscillation combines with the peak frequency deviation effect to produce a much broader spectrum than expected. Part of this is due to the nonlinearity of the format, part is due to the dynamics of the modulation.

The end result is that the FM spectrum is spread out much broader in general than AM signals (for $\beta > 1$). In fact, it is found that the spreading is proportion to the β of the modulation. This is embodied in Carson's rule, which gives an estimate of the 99% power bandwidth of an FM signal:

$$BW_{FM} \approx 2(\beta + 1) BW_m. \quad (6\text{-}28)$$

For the example shown in Figure 6-12, we indicated a peak frequency deviation of 75 kHz. The message bandwidth was roughly 200 kHz. This yields a β of 0.375, which corresponds to a relatively narrowband FM implementation. Carson's rule predicts a transmitted bandwidth of 550 kHz, which agrees favorably with the bandwidth shown in Figure 6-12.

Example – Transmitted BW of broadcast FM and FM standards

6.3.3.3 FM SNR improvement

The major result of FM spectral spreading is that interferers, whose energy is narrowband, do not effect the relatively wideband FM signal. In AM modulation, narrowband interferers' energy greatly overlap the signal's narrowband energy, resulting in severe susceptibility to interference. This results FM's superior performance in the presence of interferers. In addition, FM uses amplitude clipping to remove much of the amplitude noise that plagues AM receivers, thus removing the static – "No static at all".

The analysis of the FM SNR performance for an idealized receiver is beyond the scope of this discussion. The important points will be addressed here. The receiver architecture is shown in Figure 6-13.

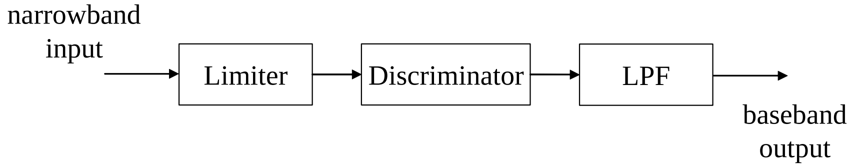

Figure 6-13 Ideal FM receiver.

The limiter is simply an amplitude clipping circuit. The discriminator may be viewed as a filter with frequency response having a linear region in the range of frequencies occupied by the bandwidth of the FM signal of interest. As the signal frequency changes, the magnitude of the output signal changes, and so a signal which varies linearly with frequency is obtained. In comparing the performance of FM to AM, the result indicates the promise of spectral spreading:

$$SNR_{out,FM} = 3\beta^2 SNR_{out,AM} . \qquad (6\text{-}29)$$

Since beta is a measure of the peak frequency deviation, which is important in the amount of spreading of the spectrum, higher beta leads to a wider transmitted spectrum, and thus to a greater improvement in SNR! However, the designer does not always have a wider spectrum in which to transmit the signal.

6.3.4 Quadrature Multiplexing

The quadrature multiplexing (or IQ modulation) architecture is an extension of the double-sideband suppressed carrier architecture, where the two carriers at the same frequency, $\cos(\omega_c t)$ and $\sin(\omega_c t)$, are each modulated independently and then summed to form the transmitted signal. For a properly configured and synchronized system, the transmission scheme can be thought of as providing two independent channels of equal bandwidth. This system will be introduced as an analog modulation system, but finds its real power in digital modulation systems.

6.3.4.1 Quadrature Multiplexing System architecture

The quadrature multiplexing architecture is shown in Figure 6-14. Two inputs are applied to the system, $m_1(t)$ and $m_2(t)$. In practice, these are called the I and Q inputs, for in-phase and quadrature, with the same associations as discussed above in the context of narrowband noise modeling.

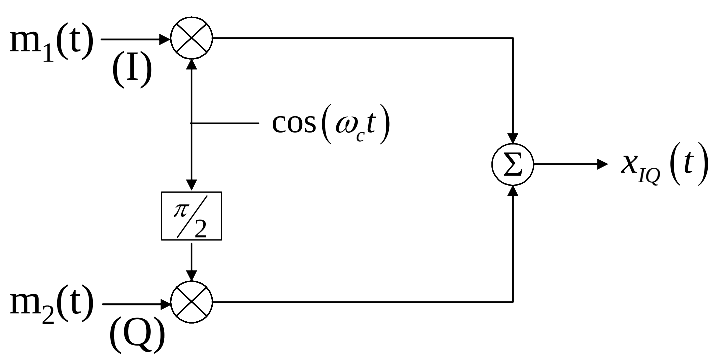

Figure 6-14 IQ modulator architecture.

Analysis of the modulation architecture reveals that the transmitted signal is

$$x_{IQ}(t) = m_1(t)\cos(\omega_c t) - m_2(t)\sin(\omega_c t). \qquad (6\text{-}30)$$

Note that the transmitted signal $x_{IQ}(t)$ is in the form of a narrowband signal. The transmitted signal is in a double-sideband format with bandwidth equal to twice the bandwidth of the baseband inputs.

The IQ demodulation architecture is shown in Figure 6-15. The input to the system is a narrowband modulated signal $x_{IQ}(t)$. Note that it is not important that the signal actually be quadrature modulated, but here it is assumed. For example, a DSB-SC signal could also be demodulated with this system, as is demonstrated in the example below

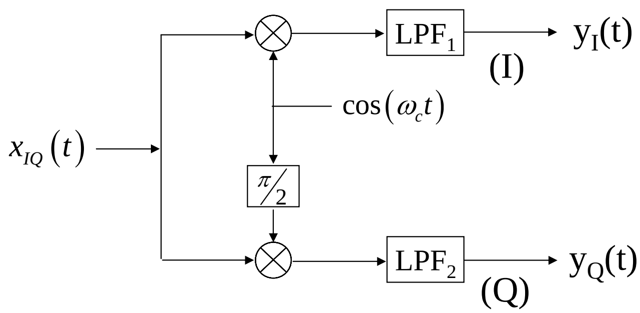

Figure 6-15 IQ demodulator architecture.

If the signal $x_{IQ}(t)$ is applied at the input to the demodulator, the messages $m_1(t)$ and $m_2(t)$ will appear at the I and Q outputs respectively, albeit potentially distorted and corrupted by noise. This assumes, of course, the proper synchronization of the local oscillator and proper lowpass filter design.

Example – demodulate DSB-SC with IQ demod

6.3.4.2 Two signals in one double-sided bandwidth

The system described above operates on the principle of the orthogonality of the sine and cosine functions. Each of these functions is used as a carrier waveform with the same carrier frequency. The two carriers are transmitted simultaneously at the same frequency and occupying the same bandwidth. This leads to a concept of increased efficiency in the use of the spectrum allocated for this carrier. In most of the communications world, transmission bandwidth is an assigned quantity, and the allocated bandwidth often is much smaller than the designer would prefer. The system engineer's job is to make the best use of what is assigned.

To illustrate this increased spectral efficiency, consider the example of the next three figures. Here, we show two input signals m_1 and m_2, which are baseband signals having bandlimited PSDs. These two signals are IQ modulated onto a 1 MHz carrier, with m_1 on the I input and m_2 on the Q input. The time domain waveforms are illustrated in Figure 6-16.

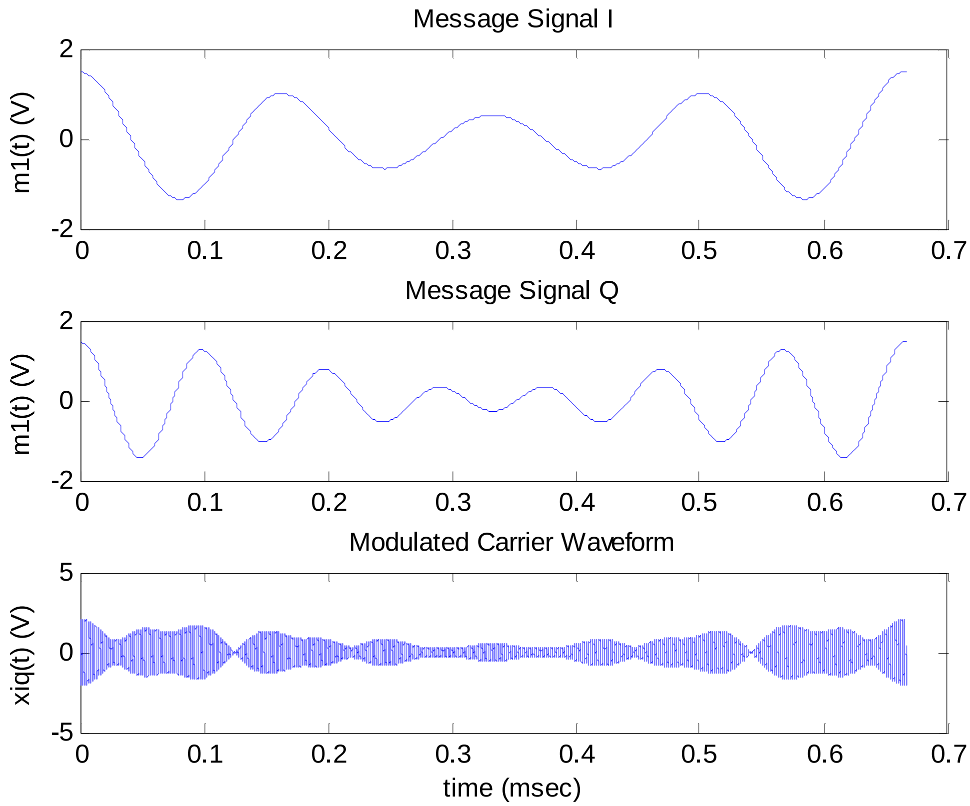

Figure 6-16 Time domain waveform for quadrature multiplexed signal.

The transmitted signal spectrum is shown in Figure 6-17. The spectrum looks a little strange compared to the individual message spectra, and that can be explained by a more in-depth analysis of the IQ modulation format. However, all is well, and the signal is passed to the demodulator.

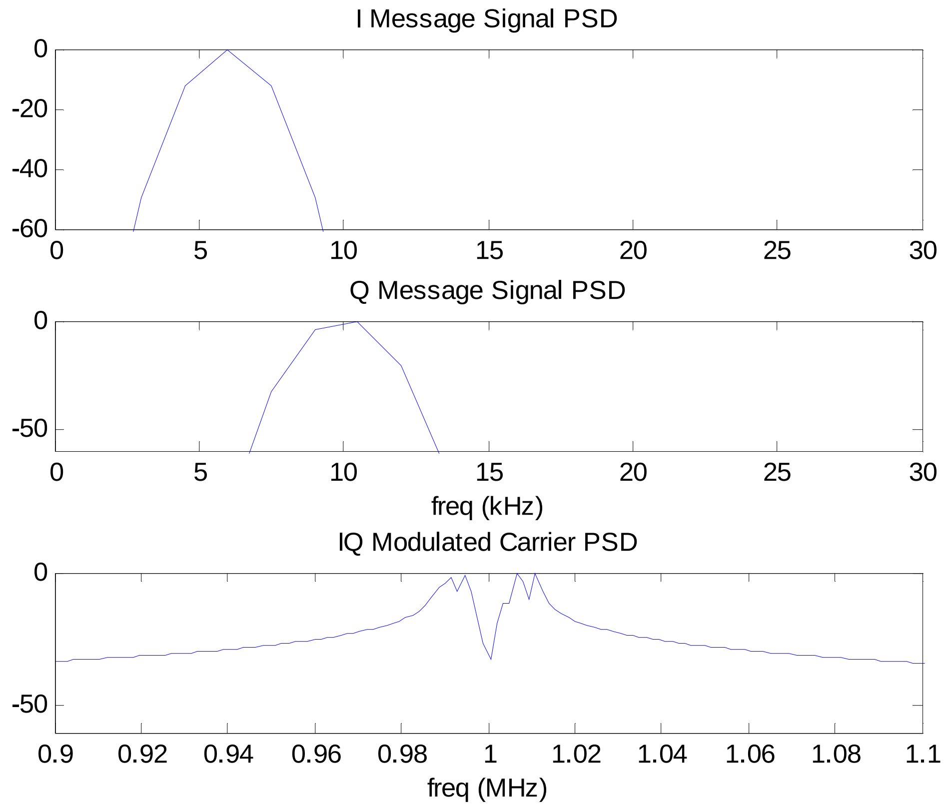

Figure 6-17 Transmitted IQ Signal Power Spectral Density.

The input and output spectra, using the architectures of Figure 6-14 and **Figure 6-15**, are compared in Figure 6-18. The demodulator filter is a Butterworth filter, and some slight distortion of the message signal spectrum is observed. Otherwise, the signals are recovered well.

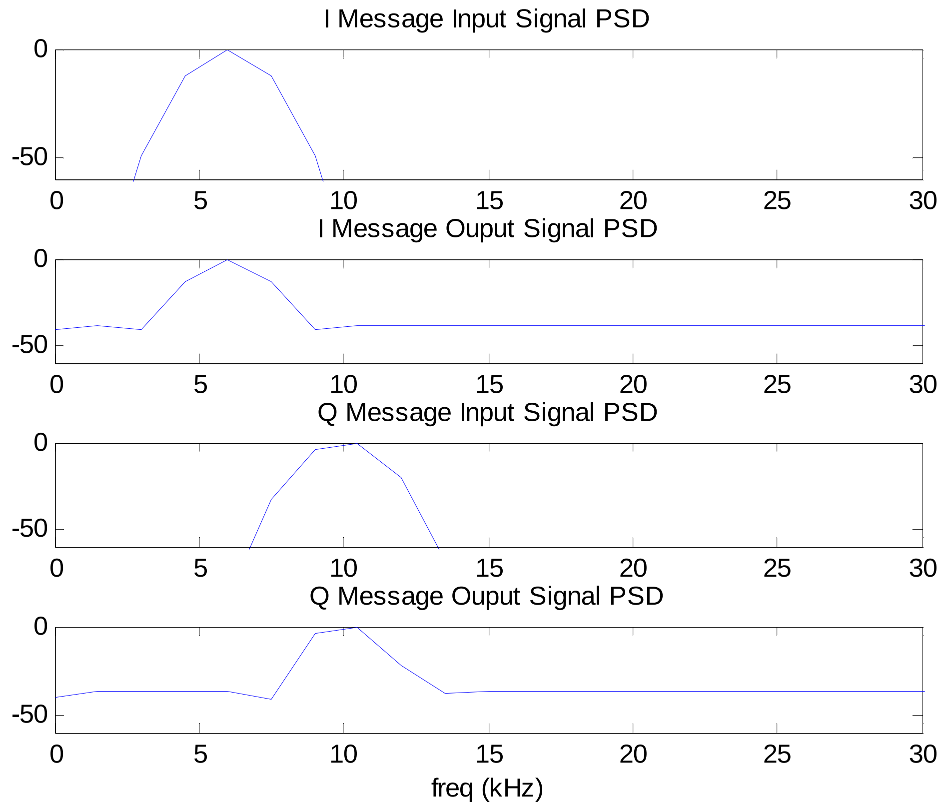

Figure 6-18 Comparison of message and recovered signal spectra for quadrature multiplexing example.

What is interesting about this modulation scheme is that interference between the two carriers is avoided only by careful synchronization in the receiver. Note that the cosine and sine (I and Q) carriers have a fixed phase relationship, being in quadrature. However, the local oscillator must somehow become synchronized with the carrier frequency and phase for the results to appear nicely as above. Suppose there is a phase error between the receiver local oscillator and the carrier waveforms. What happens to the I and Q output? There must be a phase reference, so assume that the local oscillator is the zero-phase reference, and that the carrier phase leads by some angle ψ. Then the received signal, which is the input to the IQ demodulator, is

$$x_{IQ}(t) = m_1(t)\cos(\omega_c t + \psi) - m_2(t)\sin(\omega_c t + \psi). \qquad (6\text{-}31)$$

The outputs of the IQ demodulator are found to be

$$\begin{aligned} y_I(t) &= \frac{1}{2}\left\{m_1(t)\cos\psi + m_2(t)\sin\psi\right\} \\ y_2(t) &= \frac{1}{2}\left\{m_2(t)\cos\psi - m_1(t)\sin\psi\right\} \end{aligned}. \qquad (6\text{-}32)$$

The two channels are no longer separate, emphasizing the requirement for precise phase synchronization between the receiver's local oscillator and the carrier. This requirement is one aspect of the difficulty in designing and building precision receivers.

The effect of a phase error of 30 degrees is shown in Figure 6-19 for the example messages demonstrated above. Certainly the impact is significant, and so the problem of carrier synchronization demands attention.

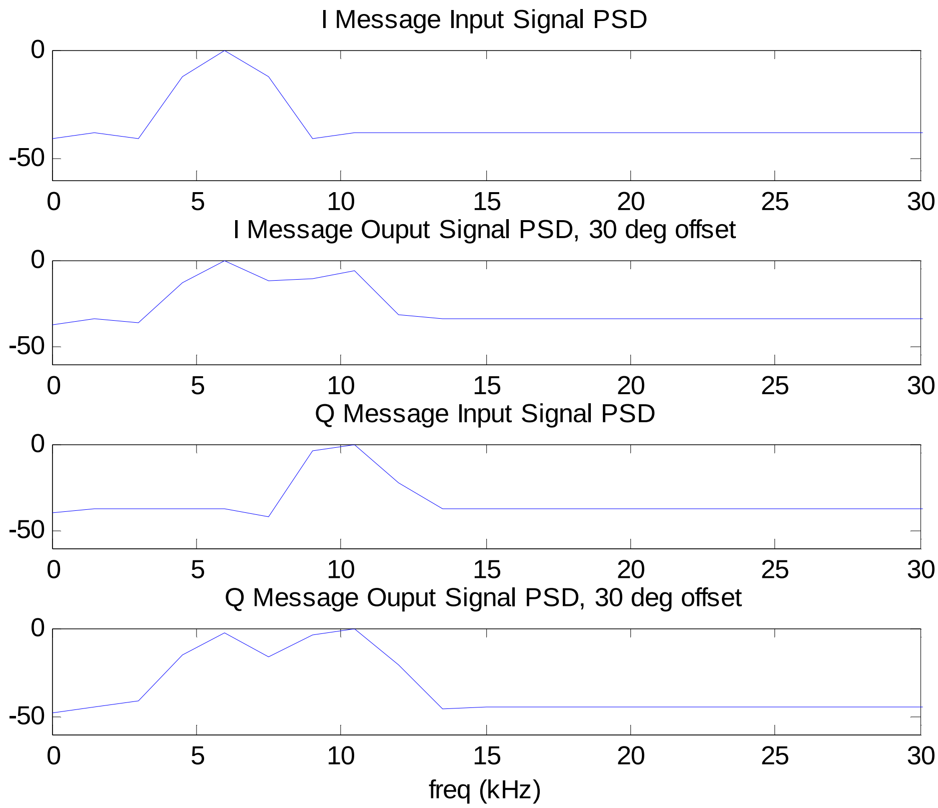

Figure 6-19 Comparison of message and recovered signal spectra with local oscillator out of phase with carrier by 30 degrees.

6.4 Baseband Pulse Detection

In order to understand the performance of digital systems, it is best to look at baseband digital communication links before modulated links. This section describes the basic architecture and techniques, performance analysis, and typical improvement schemes for baseband pulse communication.

6.4.1 Baseband Digital Communication Architecture

The architecture for a baseband digital system is actually quite simple. There is no modulation or demodulation, as this is baseband signaling, just as would be seen between digital logic chips. The chief goal of the transmitter is to translate a stream of logical data into an analog waveform which can be sent along a wire. This waveform is called a line code, and this is the most basic means of communicating digital data. (Note that, in reality, there are no digital links – as soon as the line code exists, we are working with an analog link!)

The line code is a code for representing digital data. The code consists of symbols transmitted one per symbol period. Each symbol can represent 1 bit of information, in which case the code would be called a binary line code, since the symbol can have only two voltage values during the symbol period. Some line codes can represent several bits, and thus would have several possible voltage values during the symbol period.

There are many line codes for representing digital data. Some common forms are shown in Figure 6-20. A form known to many students of digital circuits is shown in the first plot, which displays a unipolar non-return-to-zero code. The code gets its name from the fact that the voltage level does not return to zero during the symbol period. A polar code simply represents a “1” with a positive voltage and “0” with a negative voltage during the symbol period. These codes are important in the digital modulation schemes discussed later in this chapter.

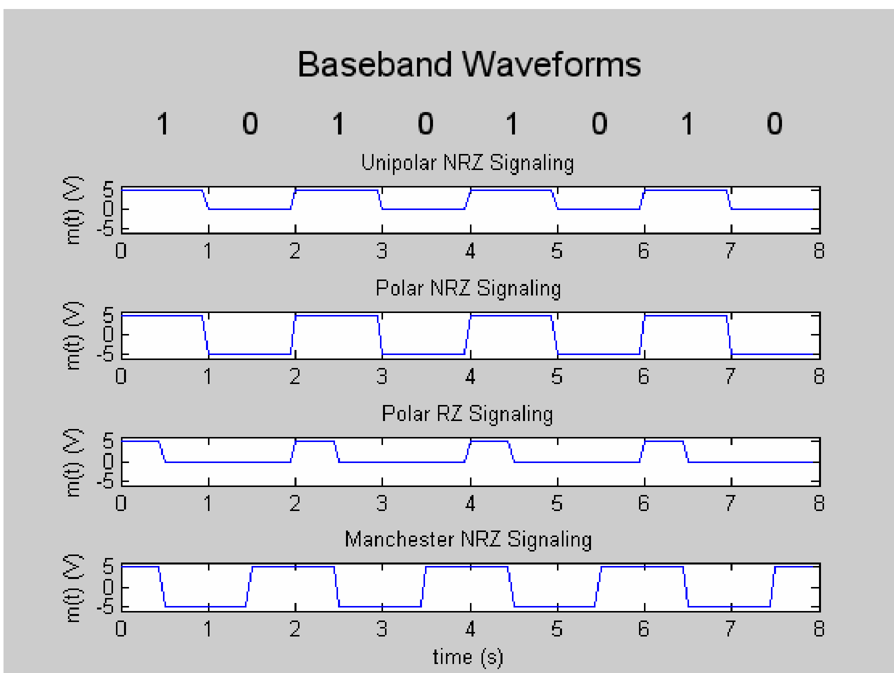

Figure 6-20 Examples of line codes used to represent a stream of digital data.

The third plot shows a return-to-zero code, where the waveform is forced to return to zero during each symbol period. Here, a duty cycle must be specified, with a 50% duty cycle RZ waveform displayed in the figure. The Manchester code shown in the fourth plot uses a low-high symbol to represent a “0” and a high-low symbol for a “1”.

Each of these linecodes has advantages and disadvantages. The unipolar NRZ code is simple to implement. The polar line code is more complex, requiring two power supplies in the transmitter, but produces better noise immunity. The Manchester code has more edges for synchronization, but is more complicated to encode and decode. The symbols shown here are formed from rectangle pulses, but the pulses may have other shapes. We will be especially interested in pulse shaping as we investigate bandwidth-efficient digital modulation later in the chapter.

Since these waveforms are unmodulated, their energy is centered around DC. Thus, we consider these baseband signals, and the problem of detecting and decoding is one of baseband pulse detection.

6.4.2 Baseband Pulse Detection Architecture and Analysis

In this section, we address the problem of detecting pulses in the presence of noise. The problem is simple. A baseband digital communication system can be thought of as sending pulses to represent “1”s and “0”s. The receiver must somehow decide whether a “1” or “0” was sent. The analysis which follows is classic, and closely follows that of Ziemer and Peterson [3].

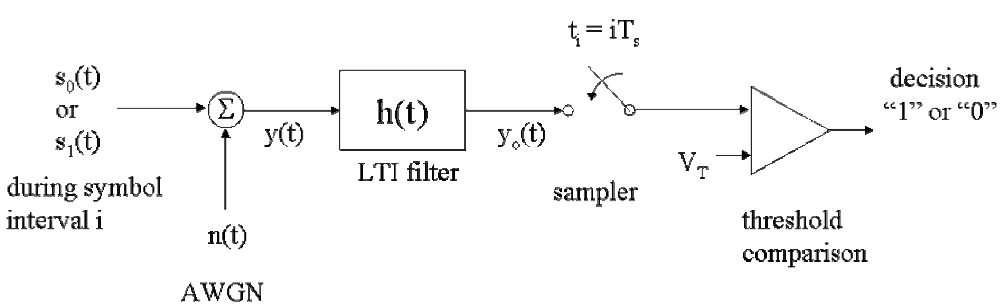

Figure 6-21 Baseband pulse receiver architecture.

A diagram of the system to be analyzed is shown in Figure 6-21. The diagram shows a simplified baseband pulse receiver. The filter h(t) is an LTI system but otherwise unspecified at this point.

The analysis focuses on one symbol interval (of duration T_s), denoted by the index i. The transmitter has sent the waveform representing a logical “0” or “1”, which has resulted in the baseband signal $s_0(t)$ or $s_1(t)$ arriving at the receiver, respectively. The signals $s_o(t)$ and $s_1(t)$ are baseband pulses as indicated by the chosen linecode. For convenience, we focus on the case of the signal $s_0(t)$ being transmitted in the current symbol interval.

We model the effect of noise in the detection process as AWGN adding to the signal at the input to the receiver. Thus, the input to the LTI filter h(t) is the combined signal y(t) = s_0(t) + n(t). Since h(t) is an LTI system, we can predict some information about the output y(t) = s_{0o}(t) + n_o(t). The output of the filter due to the received symbol is found by convolution

$$s_{0o}(t) = s_0(t) * h(t) = \int_{-\infty}^{\infty} s_0(\gamma) h(t-\gamma)\, d\gamma \,. \qquad (6\text{-}33)$$

In the absence of noise, the sampled value of the output signal is a single value of the waveform exiting the filter due only to the input signal $S_{0o} = s_{0o}(iT_s)$. The synchronization of the sampler will be discussed later.

The output of the filter due to noise can be described only stochastically. We have the input noise description as zero mean AWGN. The noise will remain Gaussian distributed (not proven). However, the noise PSD will be affected by the filter:

$$G_{n_o}(f) = G_{n_i}(f)\left|H(f)\right|^2 = \frac{N_0}{2}\left|H(f)\right|^2 . \qquad (6\text{-}34)$$

The average noise power exiting the filter can then be calculated:

$$P_{n_o} = \sigma_{n_o}{}^2 = \int_{-\infty}^{\infty} G_{n_o}(f)\, df = \frac{N_0}{2}\int_{-\infty}^{\infty}\left|H(f)\right|^2 df \,. \qquad (6\text{-}35)$$

In the absence of a signal, the noise at the output of the filter is sampled, yielding a sample value N_o. This describes obtaining a sample from a Gaussian random process having zero mean and variance P_{no} as calculated above.

However, in our receiver analysis, both signal and noise are present. The resulting combined signal $y_{0o}(t) = s_{0o}(t) + n(t)$ is sampled, producing a sample value $Y_{0o} = y_{0o}(iT_s) = S_{0o} + N_o$. In the general case, where it is not known whether s_0 or s_1 was sent, the sample value is Y_o, and can correspond to either Y_{0o} or Y_{1o}.

Examining this sample value Y_{0o}, we find that S_{0o} is simply a constant, while N_o is a sample from a zero mean Gaussian random process with variance $\sigma_{n_o}{}^2$. Thus, we conclude that Y_{0o} must be a Gaussian random process having mean S_{0o} and variance $\sigma_{n_o}{}^2$. If we were to examine a number of these samples, we would find their PDF to be Gaussian, modeled as

$$p_{Y_{0o}}(\alpha) = \frac{1}{\sqrt{2\pi\sigma_{n_o}^{\ 2}}} \exp\left\{ \frac{-(\alpha - S_{0o})}{2\sigma_{n_o}^{\ 2}} \right\}. \tag{6-36}$$

Similarly, if a logical “1” were transmitted, the sample value would be a Gaussian distributed random variable with the same noise variance $\sigma_{n_o}^{\ 2}$, but with mean S_{1o}.

$$p_{Y_{1o}}(\alpha) = \frac{1}{\sqrt{2\pi\sigma_{n_o}^{\ 2}}} \exp\left\{ \frac{-(\alpha - S_{1o})}{2\sigma_{n_o}^{\ 2}} \right\}. \tag{6-37}$$

We can now ask about the quality of the communication of “1”s and “0”s. It is important to understand that, as opposed to analog reception, the digital receiver shown in Figure 6-21 does not have the job of recreating a waveform. It merely has to enable a decision – “Was $s_0(t)$ or $s_1(t)$ sent?”, implying the transmission of “0” or “1” respectively. In our simplified architecture, the decision is made by the comparator. If the sampled value Y_o is less than the threshold V_T, the receiver decides that s_0 was sent. Otherwise, the receiver decides that s_1 was sent.

This decision, however, is muddied by two problems. First, we have yet to select the threshold value V_T. But, more importantly, the noise adds considerable uncertainty to the value of Y_o, and thus is a source of error in the decision process. If we plot the PDFs for Y_{0o} and Y_{1o}, the error probability situations become apparent. Consider Figure 6-22, which shows the PDF for Y_{0o}, the sample given that the waveform $s_0(t)$ was received with noise. In this demonstration, polar signaling is assumed, and the threshold is set at 0 V.

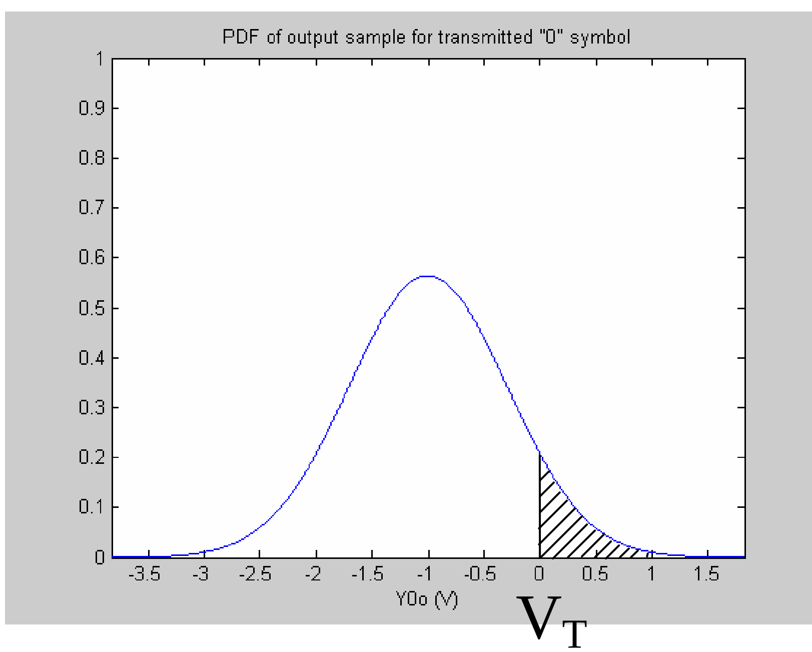

Figure 6-22 PDF for output sample given a transmitted zero symbol (threshold = 0 V).

If there were no noise (modeled as σ_n^2=0), the only possible value of the sample would be S_{0o}. The presence of noise ($\sigma_n^2 \neq 0$) adds uncertainty to the measurement, as indicated by the non-zero width of the PDF. The threshold value V_T is indicated on the α axis, and so any sample of Y_{0o} having value greater than V_T would correspond to a decision error.

Mathematically, the probability of that error occurring is found by integrating the PDF over the interval V_T to infinity, which corresponds to calculating the shaded area in the figure

$$\Pr\{error \mid s0\} = \int_{V_T}^{\infty} P_{Y_{0o}}(\alpha)\, d\alpha \,. \tag{6-38}$$

Similarly, the error probability for the situation where $s_1(t)$ was received is found to be

$$\Pr\{error \mid s_1\} = \int_{-\infty}^{V_T} P_{Y_{1o}}(\alpha)\, d\alpha \tag{6-39}$$

To find the total probability of error for this pulse detection problem, we have to include the a priori probabilities that signals s_1 and s_0 were sent, denoted by p and q, respectively, where p=1-q. The probabilities p and q are called the source probabilities. Then, the total error probability P_e is

$$P_e = p\Pr\{error \mid s_1\} + q\Pr\{error \mid s_0\}\,. \tag{6-40}$$

Of course, the threshold for the decision circuit is not an arbitrary value. To optimally choose the threshold, we differentiate the P_e expression with respect to V_T, set the result equal to zero, and solve for the optimum threshold value V_{Topt} to find

$$V_{Topt} = \frac{\sigma_{n_o}{}^2}{S_0 - S_1}\ln\frac{p}{q} + \frac{S_0 + S_1}{2}\,. \tag{6-41}$$

If the source probabilities are equal, p=q=0.5, the optimum value of the threshold becomes $V_{Topt} = (S_0+S_1)/2$.

We thus have two important terms to consider early in our problem. First, we must know the source probabilities to accurately calculate both the optimum threshold and the error probability. Then, we must guarantee that the threshold is set to the optimum value.

The integral of the Gaussian PDF must be performed numerically, and so several techniques have arisen for efficiently determining results. First, we will modify the integral into a standard form. Given a Gaussian random process with mean μ and variance σ^2, the probability that a sample from that random process will be less than some value b is

$$\Pr\{x \le b\} = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\beta-\mu)}{2\sigma}\right\} d\beta = Q\left(\frac{\mu - b}{\sigma}\right). \quad (6\text{-}42)$$

The Q-function is defined as shorthand notation for the rather cumbersome integral of the Gaussian PDF, otherwise known as the cumulative distribution function (CDF). Values of the Q-function are available in either tabular or chart form. It is noted that the Q-function is a decreasing function of increasing argument.

Mathematical tools often define a numerical function called the complementary error function, or erfc, which can be used to compute values of the Q-function according to the relationship

$$Q(z) = \frac{1}{2} erfc\left(\frac{z}{\sqrt{2}}\right). \quad (6\text{-}43)$$

We can thus rewrite the error probability relationship in a more compact form:

$$P_e = pQ\left(\frac{V_T - S_{0o}}{\sigma_{n_o}}\right) + qQ\left(\frac{-V_T + S_{1o}}{\sigma_{n_o}}\right). \quad (6\text{-}44)$$

This relationship allows for the error probability to be determined in the general case of arbitrary threshold, source probabilities, and signal waveforms.

For the special case of equal source probabilities, assuming the use of the optimum threshold for that condition, we find that the error probability relationship reduces to

$$\begin{aligned} P_e &= \frac{1}{2} Q\left(\frac{\frac{S_{0o}+S_{1o}}{2} - S_{0o}}{\sigma_{n_o}}\right) + \frac{1}{2} Q\left(\frac{-\frac{S_{0o}+S_{1o}}{2} + S_{1o}}{\sigma_{n_o}}\right) \\ &= \frac{1}{2} Q\left(\frac{S_{1o}-S_{0o}}{2\sigma_{n_o}}\right) + \frac{1}{2} Q\left(\frac{S_{1o}-S_{0o}}{2\sigma_{n_o}}\right) = Q\left(\frac{S_{1o}-S_{0o}}{2\sigma_{n_o}}\right) \end{aligned} . \quad (6\text{-}45)$$

This result emphasizes the fact that the error probability is a strong function of signal distance (S_{1o}-S_{0o}). The larger this distance, the smaller the error probability, and thus the better the transmission quality. (This is why polar signaling outperforms unipolar for the same amplitude parameter A.)

Obviously, larger noise power will lead to a larger P_e. What can cause the noise power to be larger? A larger AWGN input noise PSD is one cause. A more likely culprit would be the bandwidth of the filter h(t) – in an attempt to make the filter bandwidth wide enough to pass all of the signal power, a large noise power is passed (thus reducing the SNR!)

One final note. This analysis is predictive, and its result is an error probability prediction for the system and assumptions shown. The usual performance measure associated with links is the Bit Error Rate, or BER, which is an *a posteriori* measure of the number of bits in error divided by the total number of bits transferred. If the theoretical model is accurate, and enough trials with enough samples per trial are performed, we expect the P_e to estimate the BER. In our analyses, we will often use P_e and BER as if interchangeable, assuming that these caveats are addressed.

Example – Pe calculation.

6.4.3 The Matched Filter

The choice of filter design for the LTI filter h(t) in the receiver of Figure 6-21 is not arbitrary. There should be some design goal for that filter. The goal could be very simple, such as to pass the signal without distortion. But this would ignore the effect of noise in the detection process.

A better design goal would be to consider the role of the filter in the purpose of the detector. We could choose a filter which maximizes the SNR at the filter output. Or, we could choose a filter which minimizes the P_e. Both of these goals are laudable, but both also need more specification in order to lead to filter specification. Perhaps the two goals are effectively the same.

We have the advantage here in that the received signal waveform shape is known in advance. This means that we can predict the output of the filter due to the received waveform representations of logical "1"s and "0"s, as discussed in the previous section. In the following, we assume a unipolar NRZ baseband code, where the symbol for "0" is zero volts during the symbol duration. We will generalize the derivation afterwards. We define a *matched filter* as a filter which is designed to maximize the SNR_o at the filter output for a particular signal waveform, where the SNR_o is defined as

$$SNR_o = \frac{s_{1o}^{\,2}\left(t_s\right)}{\sigma_{n_o}^{\,2}}. \qquad \text{(6-46)}$$

The term $s_{1o}^2(t_s)$, which is similar to a power, is the peak value of the output waveform squared, where the peak value defines the sample instant.

The impulse response of the matched filter is specified in terms of the received signal waveform, and becomes (assuming white noise)

$$h_m(t) = s_1^*(t_0 - t). \quad (6\text{-}47)$$

The delay t_0 is the time of the peak signal output from the filter.

An interpretation of the matched filter is that its impulse response is the signal pulse waveform played backward with delay. This will make much more sense in the context of correlative receivers, but that is for a later discussion. The frequency response of this filter is the Fourier transform of its impulse response:

$$H(f) = KS_1(f)e^{-j\omega t_s}. \quad (6\text{-}48)$$

Passing the signal s (t) through its matched filter gives an output resulting from a simple convolution integral:

$$s_o(t) = \int_{-\infty}^{\infty} s_1(\lambda)h(t-\lambda)\,d\lambda = K\int_{-\infty}^{\infty} s_1(\lambda)s_1^*(t_0 - t + \lambda)\,d\lambda. \quad (6\text{-}49)$$

At the sample instant $t=t_0$, the output $s_o(t)$ is found to be the energy in the input signal

$$s_o(t_0) = K\int_{-\infty}^{\infty} s_1(\lambda)s_1^*(\lambda)\,d\lambda = KE_s. \quad (6\text{-}50)$$

Thus, the peak output voltage of the matched filter, which occurs at the time t_s, is found to be the energy of the signal waveform E_s (J). The output signal power will be equal to the voltage squared (into 1 W), and so the signal power is found to be

$$P_s = K^2E_s^{\ 2}. \quad (6\text{-}51)$$

Next, the noise is passed through that same filter (recall that it is a linear system). The AWGN noise is described by its PSD $S_{ni}(f)=N_0/2$. The noise PSD exiting the filter is found using the system relationship (6-12):

$$S_{n_o}(f) = \left|H(f)\right|^2 S_{n_i}(f) = K^2S_1^{\ 2}(f)\frac{N_0}{2}. \quad (6\text{-}52)$$

The total noise power is found by integrating this spectral density

$$\begin{aligned} P_{n_o} &= \int_{-\infty}^{\infty} S_{n_o}(f)df = \int_{-\infty}^{\infty} K^2S_1^{\ 2}(f)\frac{N_0}{2}df \\ &= K^2\frac{N_0}{2}\int_{-\infty}^{\infty} S_1^{\ 2}(f)df = K^2\frac{N_0}{2}E_s \end{aligned}. \quad (6\text{-}53)$$

In this expression, Parseval's relation has been used to extract the signal energy. The SNR produced at the output of the matched filter, assuming white input noise, is thus

$$SNR_{m,o} = \frac{K^2 E_s{}^2}{K^2 \frac{N_0}{2} E_s} = \frac{2E_s}{N_0} . \quad (6\text{-}54)$$

Note that the SNR is independent of the particular pulse shape as long as the filter is matched. The SNR is simply dependent upon the signal waveform energy and the white noise PSD level. The SNR can be increased by increasing the signal amplitude, for example, which is an issue for the link budget.

Example – rectangular pulse with AWGN

To deal with a general linecode, we must extend our definition of the matched filter to account for the different symbols for logical "1" and "0", neither of which is zero. Examples of this situation include the polar or Manchester line codes. The matched filter impulse response changes to take into account the fact that two symbols are possible:

$$h_m(t) = C\left[s_1(t_0 - t) - s_0(t_0 - t)\right] . \quad (6\text{-}55)$$

We define the difference signal energy at the receiver input to be

$$E_d = \int_0^T \left[s_1(t) - s_0(t)\right]^2 dt . \quad (6\text{-}56)$$

This term makes the algebra more compact and analogous to the derivation above. The optimized SNR at the output of this matched filter then becomes

$$SNR_{m,o} = \frac{K^2 E_d{}^2}{K^2 \frac{N_0}{2} E_d} = \frac{2E_d}{N_0} . \quad (6\text{-}57)$$

6.4.4 Receiver Performance BER vs Eb/N0

The performance of the receiver is primarily judged by its ability to reproduce the transmitted bit stream at some level of accuracy, as indicated by the Bit Error Rate (BER). Note that while it is never possible to have a P_e of identically zero, it is possible to push the BER arbitrarily close to zero through the application of signal power and transmitter/receiver complexity (and cost). For example, providing a link budget which delivers more received power will improve the BER. However, money does not grow on trees…

The usual means of expressing the tradeoff between BER performance and received power is a plot called a "waterfall curve". This plot shows the relationship between BER and received power or SNR for the choice of signaling and hardware implementation chosen for the communication system.

Before proceeding with this discussion, we must define the signal energy in a form which allows for easy comparison between signaling formats (line codes). We define the average bit energy as the time averaged energy corresponding to the bitstream being transmitted. Given the source probabilities p for the probability that a 1 was transmitted and q=1-p that a "0" was transmitted; the average bit energy is defined as:

$$E_b = p\left[E_{s1}\right] + q\left[E_{s0}\right]. \tag{6-58}$$

For example, consider the unipolar NRZ and polar NRZ line codes. For each case, the symbol exists in one symbol period of duration T_s. For the unipolar NRZ codes, the amplitude of the symbol pulses are A and 0, corresponding to a "1" and a "0", whereas for the polar code the amplitudes are A and –A.

The energy for a "1" symbol is A^2T_s for the unipolar code, and the same for the polar code. However, the energy for a "0" symbol is 0 for the unipolar code, but A^2T_s for the polar code. Thus, the average bit energies differ for the two codes (assuming p=q):

$$\begin{aligned} E_{b,unipolar} &= \frac{1}{2}A^2T_s \\ E_{b,polar} &= A^2T_s \end{aligned}. \tag{6-59}$$

In the analysis of the error probability of the baseband pulse detection system of Figure 6-21, it was found that the error probability was described, for equal source probabilities and optimum threshold, by

$$P_e = Q\left(\frac{S_{1o} - S_{0o}}{2\sigma_{n_o}}\right) = Q\left(\sqrt{\frac{\left(S_{1o} - S_{0o}\right)^2}{4\sigma_{n_o}^{\;2}}}\right). \tag{6-60}$$

In this expression, we have assumed $S_{1o} > V_T > S_{0o}$. We can rewrite the P_e expression for the case of arbitrary linecode signaling with an appropriate matched-filter receiver to obtain

$$P_e = Q\left(\sqrt{\frac{E_d}{2N_0}}\right). \tag{6-61}$$

Then, by relating the average bit energy and signal difference energies, we can produce the traditional relationship between P_e and the SNR.

The term E_b/N_0 is a useful means of comparing linecode performance. It is the ratio of bit energy during the symbol duration to noise energy in that duration, and is related to SNR. Each linecode has a distinctive E_b/N_0 produced at the output of its matched filter receiver, and thus a distinctive relationship between P_e and SNR. The average bit energies, difference energies, and error probabilities for the line codes of Figure 6-20 are given in Figure 6-23. Equal source probabilities are assumed in calculation of these values.

Line Code	E_b (J)	E_d (J)	P_e
Unipolar NRZ	$\frac{1}{2}A^2T_s$	A^2T_s	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$
Polar	A^2T_s	$4A^2T_s$	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$
Unipolar RZ (50% duty cycle)	$\frac{1}{4}A^2T_s$	$\frac{1}{2}A^2T_s$	$Q\left(\sqrt{\frac{E_b}{N_0}}\right)$
Manchester	A^2T_s	$4A^2T_s$	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$

Figure 6-23 Comparison of average bit energy, signal difference energy, and error probability for the example baseband linecodes.

From this comparison, the performance advantage of certain linecodes is obvious. Unipolar codes, while easy to produce, suffer from small signal distance and thus smaller SNR and P_e. Polar signaling, and its enhanced Manchester cousin, offer larger signal distance and thus superior P_e performance. The catch is that generation of the linecode is more difficult, for example requiring two power supplies as opposed to the one required for unipolar signaling. This gives the designer the ability to think ahead and trade complexity and cost for performance in the link.

The most common means of representing the performance capability of the line code is to plot the error probability versus E_b/N_0, a presentation sometimes called a waterfall curve. Such a comparison is provided in **Figure 6-24**.

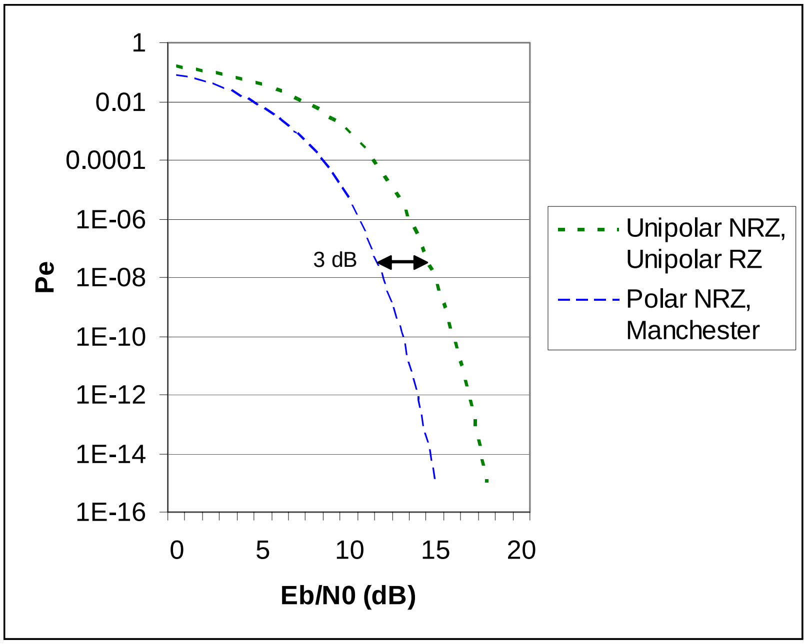

Figure 6-24 Waterfall curve comparison of baseband signaling theoretical performance with matched-filter reception.

It is critical that the reader be able to understand and read this plot. The form of the curve gives the waterfall name to relationship. The plot shows the theoretical P_e performance for the example baseband linecode signaling formats indicated. First, the reader must understand the conditions of the comparison. The three major conditions are that equal source probabilities are assumed, the optimum threshold for this condition is assumed, and that matched-filter is applied at the receiver input.

The plot indicates that as E_b/N_0 is increased (or, as the SNR is increased), the error probability drops appreciably. Notice that when well into the "waterfall", a difference of three dB (a factor of 2) makes an enormous difference in the achieved error probability. Also notice that the performance of the unipolar curves are consistently 3 dB worse that those for polar NRZ and Manchester coding. To understand this, consider the performance trade: to achieve theoretical performance of 10^{-8} error probability one could choose unipolar signaling with double the transmit power, or polar signaling with its increased transmitter complexity. This is an example of one of the considerations system engineers use to select the signaling format for baseband communication systems. Actual performance, of course, will not meet theoretical expectations. Often, a receiver's waterfall curve will plot out with the same form as the theoretical curve, but shifted right by 1-2 dB. This shift is called the implementation loss of the receiver, and has to do with excess loss, distortion, and other design impediments in the receiver implementation.

6.4.5 Intersymbol interference

As symbol pulses propagate through the transmitter, channel, and early stages of the receiver, the pulse energy and shape will be altered. Losses and gains in the system will decrease or increase the energy, and the frequency response of the various system components will alter the pulse shape. In the analysis above, the only importance placed on pulse shape was that the pulses were assumed ideal, that is perfectly rectangular, or that the received pulse shape was matched in the matched-filter receiver analysis. Also, the analysis focused on individual symbols and pulses, when in reality, the symbols are transmitted in continuous streams.

As the pulses travel through the system (channel and receiver through filter), the shape of the pulse will change from the transmitted pulse shape to received pulse shape. Dispersion, which describes a non-uniform delay variation with frequency, governs this transformation. Each pulse can be considered as composed of many different frequency components linearly combined. It is reasonable to expect that if we wish the same pulse shape as transmitted to arrive at the receiver, then each spectral component should arrive at the same time. If some spectral components arrive slower than others at the receiver, the result is a pulse spreading. There are two major undesirable effects of dispersion. First, ignoring loss, the pulse energy is conserved. Since the pulse is spread out, the peak is reduced, which may make detection more difficult. Second, the energy of the pulse may spill over into adjacent symbol periods in a continuous stream of received pulses. Thus, energy from a specific symbol can interfere with many other symbols and vice-versa. This is called *intersymbol interference* or ISI.

One way of displaying the effect of ISI is the eye diagram. This is a display of many received symbol pulse waveforms, with each overlapping in the display. The display normally covers a time corresponding to 2-3 symbol periods to allow all of the necessary information to be obtained. Typical eye diagrams are shown in Figure 6-25 for a bit rate of 1000 bps using unipolar NRZ. In the left diagram, the receive filter bandwidth is 1000 Hz, while 500 Hz bandwidth is represented in the right. In the left diagram, the “eye” is wide open, corresponding to a low-ISI situation, whereas the eye is closing in the right diagram, displaying ISI impact.

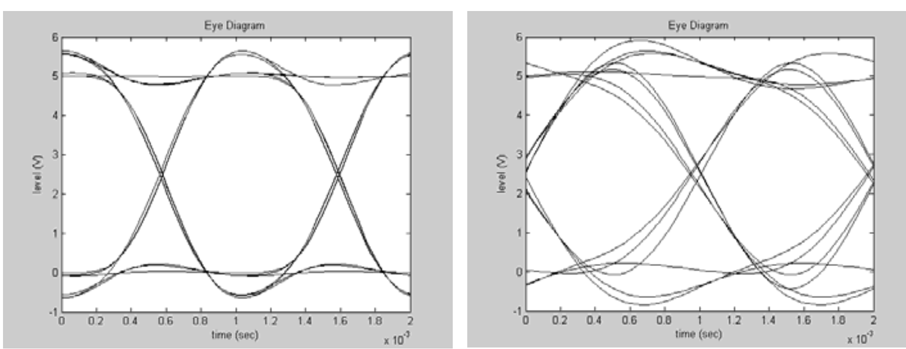

Figure 6-25 Example eye diagrams for a bit rate of 1 kbps and receive filter bandwidth of a) 1 kHz, and b) 500 Hz.

A stylized eye diagram is shown in Figure 6-26 [4]. Some of the information able to be obtained from the diagram is indicated in the figure. This quantitative information can be used to determine the impact ISI might have on BER.

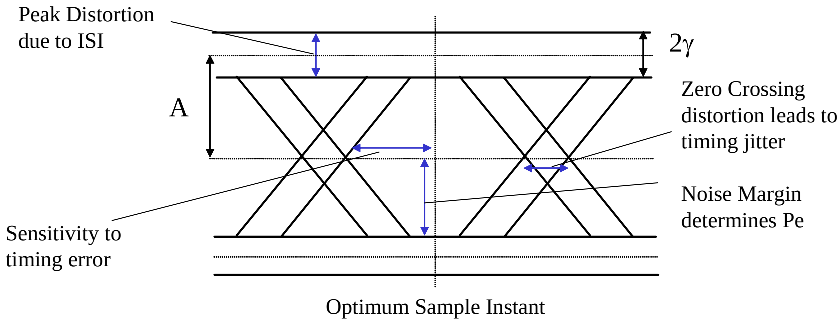

Figure 6-26 Stylized eye diagram showing information obtained from display.

The actual impact of ISI, that is the modification of the sample value due to contributions from adjacent pulses, is evident in the width of the "rails" of the display (at the sample instant), called the *peak distortion*. The width of the center of the "X" indicates uncertainty in the location of the zero crossing, which leads to *timing jitter*. Since most timing recovery systems use an instantaneous measure related to the zero crossing to synchronize to bit intervals, timing jitter is made worse as the width of this feature increases. The width of the eye, along with the slope of the arms of the "X", indicates the sensitivity of the system to jitter. Finally, the *noise margin* is the actual indicator of system performance. The larger this eye opening, the higher the SNR and the better the BER.

6.4.5.1 Effect of ISI on BER

The effect of ISI on BER can be quantified using information obtained from either the filter output pulse waveform or the eye diagram. This analysis follows that of Black [5]. Suppose the pulse waveform exiting the receiver filter in Figure 6-21, that is the pulse waveform to be sampled, is p(t) as shown in Figure 6-27. The pulse waveform has been oriented in time such that the t=0 reference has been set to the sampling instant for the desired symbol, which is the peak of the waveform.

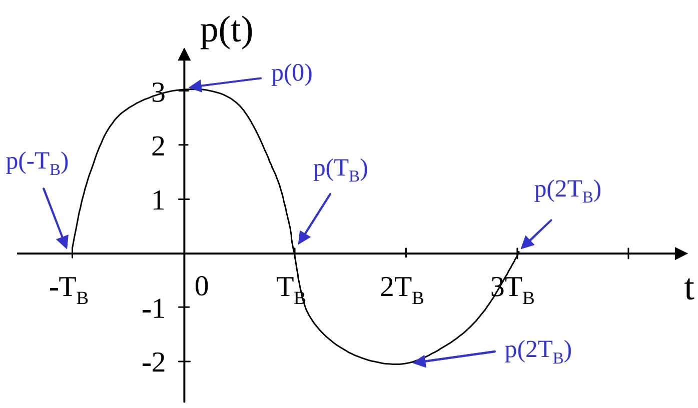

Figure 6-27 Sample waveform p(t) exiting the receiver filter, showing the values which contribute to bit samples at the sampling instants.

Now, the only values of the waveform of any interest are those which occur at the sampling instants. These values are denoted in the figure as p(kT_s), where k is an integer. The sampling interval is T_s, the symbol period, and it is assumed that the sampler has been properly synchronized. The values of the waveform at the sampling instant for the current symbol is denoted by k=0, or p(0). Note that the waveform for the current symbol has values which occur at the sample times for adjacent symbols. These values are the source of ISI for adjacent symbol samples.

For example, for symbol i+1, the previous bit will contribute a value of p(Ts), and the bit previous to that will contribute a value of p(2Ts). In general then, at each sampling instant a value is obtained for the ith symbol which has the value

$$y(iT_s) = \sum_k p\left[(i-k)T_s\right]. \qquad (6\text{-}62)$$

To finalize this model, we note that the waveform p(t) corresponds to binary digital information. Thus, there will be some constant multiplying the waveform p(t) to distinguish logical ones from zeros. For example, in polar signaling, the value of the multiplier could be +A for a logical one and –A for a logical zero. This can be described by adding a multiplier to the expression as follows:

$$y(iT_s) = \sum_k a_k p\left[(i-k)T_s\right]. \qquad (6\text{-}63)$$

By expanding this expression, we can see the ISI terms directly:

$$y(iT_s) = a_0 p(0) + \sum_{k \neq i} a_k p\left[(i-k)T_s\right]. \qquad (6\text{-}64)$$

The first term is the sample value obtained without ISI, and the remaining sum is the cumulative effect of ISI from all other symbols upon the desired symbol's sample. Obviously, the effect of the ISI contribution can either be helpful or harmful. Should the sum be the same sign as the first term, the total sample value magnitude will increase and thus produce a larger SNR. However, if the sum term has opposite sign, the SNR is degraded. Of course, both conditions are equally probable given random source data, so ISI always acts to degrade SNR.

One quantitative measure of the severity of the ISI is the Peak Distortion *D*:

$$D = \frac{\sum_{k \neq 0} \left| p(kT_s) \right|}{p(0)}. \qquad (6\text{-}65)$$

The larger the peak distortion, the more severe the SNR degradation. This peak distortion may be obtained from the eye diagram as well. Using the terminology defined in Figure 6-26, we can rewrite the definition of D:

$$D = \frac{\gamma}{A}. \quad (6\text{-}66)$$

The *eye opening*, a measure of the quality of the bit recovery, is defined as 1-D.

Example – ISI calculation and eye diagram display

6.4.5.2 Use of Pulse Shaping to Control ISI

The problem of ISI has been bothering systems from the time before organized system engineering existed. One of the best examples of system analysis can be found in Nyquist's analysis of the problem of ISI in telegraphy systems [3]. This groundbreaking work led to the concept of pulse shaping to control the problem of ISI long before digital systems became popular.

To make the discussion brief, refer back to (6-64), where the sample value for a given symbol is shown to consist of the desired symbol sample plus interference from adjacent symbols (ISI). Any pulse shape which produces no ISI is obviously the goal of the signal designer. But what is that pulse that is being discussed? Recall that this pulse is the pulse exiting the receiver filter. That is, this is not the transmitted pulse, but the transmitted pulse, distorted by the channel and filter. Of course, if the system and data rate are designed properly, it is possible that the pulse exiting the receiver filter is essentially the same as the transmitted pulse. However, pulses will most often experience some distortion in traveling through the system.

What, then, is the condition for no ISI to be present? Obviously, no ISI will occur if the adjacent symbol contributions are zero, so we require

$$\sum_{k \neq i} a_k p\left[\left(i-k\right)T_s\right] = 0. \quad (6\text{-}67)$$

One pulse type for which this is true is a sinc pulse whose nulls are chosen to coincide with the sample instants. Specifically,

$$p_{NO\ ISI}\left(t\right) = \text{sinc}\left(\frac{t}{T_S}\right). \quad (6\text{-}68)$$

However, it is found that it can be quite difficult to create a perfect since pulse. Once class of pulses which meets the no-ISI condition and is easier to approximate is the raised-cosine pulse (RCP) set. This pulse is defined in the time domain by

$$p_{RCP}(t) = \text{sinc}\left(\frac{t}{T_S}\right)\frac{\cos\left(\pi\alpha\frac{t}{T_S}\right)}{\left(1-4\alpha^2\left(\frac{t}{T_S}\right)^2\right)}. \qquad (6\text{-}69)$$

The term α is called the rolloff factor, and is a factor controlling the pulse baseband bandwidth. The symbol period is T_s. A plot of example pulses are shown in Figure 6-28 for various values of α. As α approaches zero, the pulse converges to a since pulse.

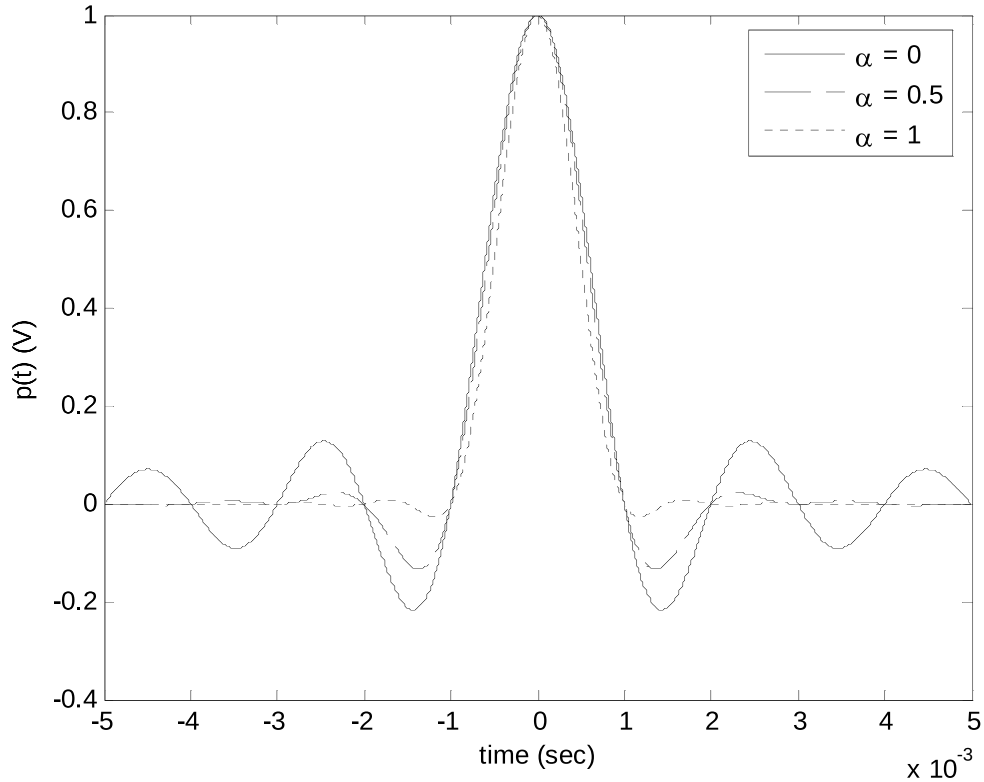

Figure 6-28 Raise-cosine pulse waveform for various values of the rolloff factor.

The baseband spectrum of the raised cosine pulse is given by

$$P_{RCP}(f) = \begin{cases} T_s & 0 \le f < \frac{1-\alpha}{2T_S} \\ \frac{T_S}{2}\left\{1-\sin\left[\frac{\pi T_S}{\alpha}\left(|f| - \frac{1}{2T_S}\right)\right]\right\} & \frac{1-\alpha}{2T_S} \le f < \frac{1+\alpha}{2T_S} \\ 0 & f \ge \frac{1+\alpha}{2T_S} \end{cases} \quad . \quad (6\text{-}70)$$

Note that the RCP spectrum is strictly bandlimited to the value $(1+\alpha)/(2T_s)$. The RCP PSD of the example pulses are plotted in Figure 6-29. Raised-cosine pulses have found widespread use in applications such as computer modems.

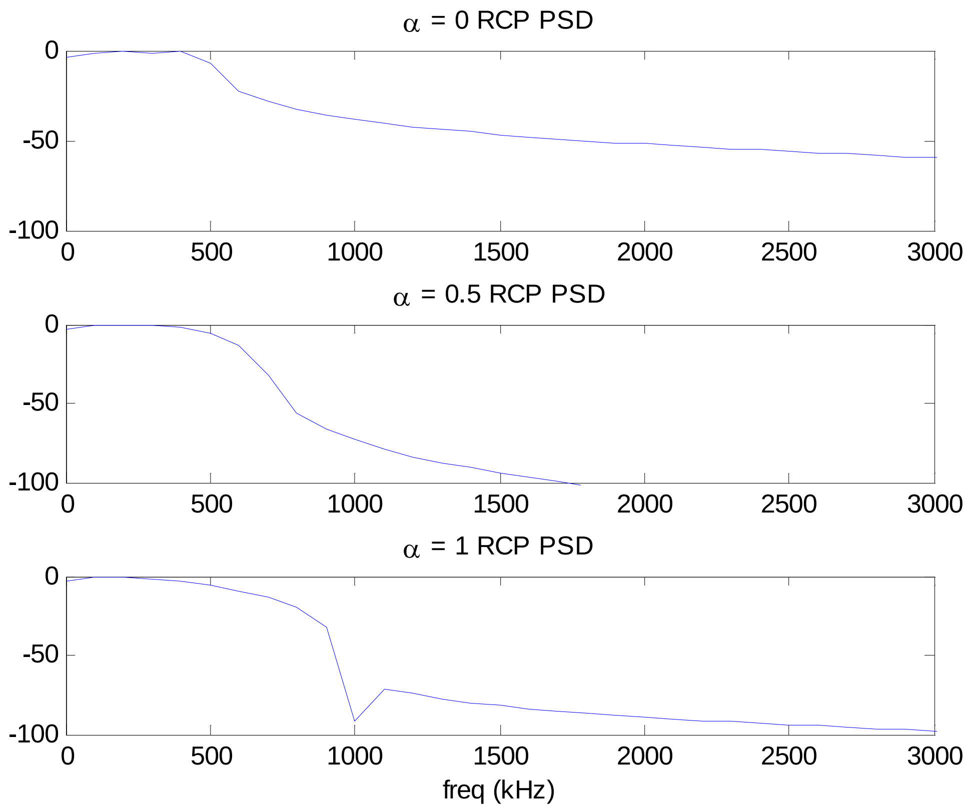

Figure 6-29 Power spectral density of example raised cosine pulses, showing increasing bandwidth.

6.5 Summary

Throughout this chapter the focus has been on the fundamental principles of signal design for both analog and digital systems. Analog systems have the goal of preservation in detail of the transmitted waveform, while digital systems have the goal of information preservation without regard for the linecode representing the information. This is one of the major reasons why digital systems are preferred over analog systems. However, the cost and limitations of digitization sometimes limit the applicability of digital links to perform in some more exacting applications such as high fidelity or wideband signal transmission.

The system engineer has been introduced to some basic figures of merit for communication links, such as SNR, BER, bandwidth, and factors contributing to these. As the system engineer discusses the design with the link analyst (signaling and link budget) and subsystem designers (antenna, transmitter, receiver), the trades for cost and complexity versus performance are performed in the real environment of regulations and schedule.

6.6 References

1. FM, D. Fagan and W. Becker,
2. Stremler, F.G., Introduction to Communication Systems, 3rd ed., 1990, Addison-Wesley.
3. Nyquist, H., Certain Topics in Telegraph Transmission Theory, Proceedings of the IEEE, v. 90, n. 2, pp. 280-305, 2002.