Digital Signal Processing: A Computer Science Perspective
Jonathan Y. Stein

Print ISBN 0-471-29546-9 Online ISBN 0-471-20059-X

Part IV

Applications

Digital Signal Processing: A Computer Science Perspective
Jonathan Y. Stein

Print ISBN 0-471-29546-9 Online ISBN 0-471-20059-X

18 Communications Signal Processing

In this chapter we will survey various topics in signal processing for communications. Communications, like signal processing itself, is commonly divided into analog and digital varieties. Analog communications consist of techniques for transmitting and receiving speech, music or images as analog signals, as in telephones, broadcast radio and television. Digital communications are methods of transferring digital information, usually in the form of bit streams. Digital communications are often between computers, or between human and computer, although increasingly digital communications are being used between people as well (email). Both analog and digital signal processing may be used for various portions of both analog and digital communications systems.

A device that takes an analog input signal and creates an analog communications signal is called a transmitter, while a receiver inputs an analog communications signal and attempts to recover, as accurately as possible, the original analog message signal. A device that takes a digital input and creates a digital communications signal is usually called a modulator, while a demodulator inputs a digital communications signal and attempts to recover, with as few bit errors as possible, the original digital message. Transmitters and receivers are sometimes packaged together and called *transceivers*; for digital communications it is almost universal to package the modulator and demodulator together, and to call the combined device a *modem*.

Digital communications systems include such diverse objects as fax machines, telephone-grade modems, local area networks, wide area networks, private digital telephone exchanges, communications satellites and their ground stations, the public switched telephone network (yes, it too has become digital), and the Internet. Although the history of data communications is relatively short, the present scope of its theory and application is huge, and we will have to stringently restrict the scope of our treatment.

After a historical introduction we will start our survey with an overview of analog communications, including AM and FM transmitters and receivers. We then briefly study information and communications theory, including error correcting codes. We then design our first modem, the rest of the chapter being devoted to successive improvements to its design. We roughly follow the chronological development of telephone-grade modems that increased bit rates from 300 b/s to 56 Kb/s. Along the way we will learn about FSK, PSK, QAM, MCM, TCM, and PCM modems, and master the basic algorithms needed for modem implementation.

18.1 History of Communications

Let's go back over 2000 years and imagine ourselves at the foot of the great Temple in Jerusalem. It is the thirtieth day of the month, and the Calendar Council is in session, waiting for witnesses to come to testify that they had seen the new moon. A group of people approach running. They are ushered into the chamber and interrogated by experts in astronomy and mathematics. If their testimony is found to be genuine, the new month is declared to have begun; if no reliable witnesses arrive the new month only starts the next day. Now that information must be disseminated quickly to those living as far away as Babylon. Only one bit of information must be transmitted—whether the new month has commenced—but telephones, radio, and even telegraph lines do not yet exist.

Now it is not really difficult to transmit the single bit of information to nearby locations. One need only do something that can be reliably seen from afar. So the Council orders a bonfire to be lit on the top of a nearby mountain. On a neighboring mountain an official is waiting. When he sees the beacon he lights a fire of his own, which is observed at the first mountain and recognized as an acknowledgment that the message has been received. It is also observed at another mountain further away, where the next beacon in the chain is lit. In this way the message that the new month has commenced is quickly and reliably transmitted. This technique was in use until thwarted by the (good?) Samaritans, who maliciously lit beacons at inappropriate times in order to create confusion.

Similar communications techniques were used by other pretelegraph peoples. Native Americans would burn wet grass under a blanket, which when removed would send up a blast of dark smoke that could be seen from afar. Natives of western Africa used tomtom drums that could be heard throughout the jungle (where visibility is limited). Mariners used signaling lamps that could be seen from miles away.

What can we do if we need to transmit more than one bit of information? The native Americans would simultaneously light two or three separate fires, and the number of columns of smoke signified the urgency of the message. The Africans used drums of variable pitch, and could send intricate messages by varying the sounds of their drumming. At sea mariners would open and close shutters on the signaling lamps, thus sending entire messages.

These methods of communications suffer from several drawbacks. First, they work over limited distances, requiring relay operators for larger range. Second, they are not reliable; after the battle of Waterloo a signal lamp message over the English channel was received as 'At Waterloo Nelson defeated ...' with 'Napoleon' covered up by the fog. Nathan Rothschild made a fortune buying up stocks on the plunging London exchange, knowing the truth through more reliable carrier pigeons. Third, these communications media are all *broadcast*, meaning that they can be intercepted by all. Although this is sometimes required it can also be a disadvantage. Settlers of the American West spotted Indian smoke signals and recognized that the enemy was close at hand. Finally, all these methods are *multiple access* with no signature, and can thus be easily forged (as the Samaritans did).

The discovery of electric current by Stephen Gray of London in 1729 produced a new medium for reliable communications over distances, removing many of the disadvantages of previous methods. In 1747, William Watson laid 1200 feet of wire over Westminster bridge, touching one end to the water of the Thames, and the other to a charged Leiden jar; a man touching the jar with his feet in the river received a shock. It took a while longer to realize that the flow of current could be detected by its lighting a light or moving an armature. In 1844, Samuel Morse telegraphed the message 'What hath God wrought?' over an electric cable, ushering in a new era for humankind. Morse's *telegraph* could distinguish between two states, current flowing or not, and so Morse had to devise a code to efficiently send letters of the alphabet using only two-state signals. The Morse code represents letters using combinations of $s = 0$ and $s = 1$ values; $s = 0$ are used as dividers, while $s = 1$ may occur in short durations (called a dot) or three times that duration (called a dash). The letter 'E' is encoded as a dot, that is, by a single $s = 1$, and thus only requires 1 time unit to transmit (although it must be followed by a single $s = 0$ inside a word and by three consecutive $s = 0$ at the end of a word). The letter 'Q' is encoded as dash, dash, dot, dash, occupying 13 basic time intervals. The entire Morse code is presented in Table 18.1. In 1866, the first transatlantic cable was laid, for the first time linking America and Europe by an almost instantaneous communications medium (unfortunately, it failed within a month).

A	·–		K	–·–		U	··–		0	–––––	
B	–···		L	·–··		V	···–		1	·––––	
C	–·–·		M	––		W	·––		2	··–––	
D	–··		N	–·		X	–··–		3	···––	
E	·		O	–––		Y	–·––		4	····–	
F	··–·		P	·––·		Z	––··		5	·····	
G	––·		Q	––·–		.	·–·–·–		6	–····	
H	····		R	·–·		,	––··––		7	––···	
I	··		S	···		?	··––··		8	–––··	
J	·–––		T	–		-	–···–		9	––––·	

Table 18.1: The Morse code. Every letter, number, or punctuation mark is assigned a unique combination of dots and dashes.

Telegraphy using Morse code still had a few disadvantages. It was relatively slow and error prone. It required skilled telegraphers at both ends and could not be directly used by individuals. Unless special codes were employed the messages could be read by others, and it was difficult to authenticate the sender's identity. For some time people strived to mechanize the transfer of text using the Morse code, but this was a difficult task due to the variable-length characters. In 1875, Emile Baudot from France created a new code, one optimized for mechanized text transfer. In the Baudot code each letter took five equal time units, where each unit could be current flow (*mark*) or lack thereof (*space*). Actual commercial exploitation of this code began in early twentieth century, under the trademark name *teletype.*

A further breakthrough was announced within a year of Baudot's code when, on March 10, 1876, Dr. Alexander Graham Bell in Boston and Elisha Gray in Chicago both filed for patents for a new invention, later to be called the *telephone.* Like the telegraph it used voltage signals traveling over a wire, but rather than being simple on-off, these signals carried a voice. Eventually, Dr. Bell won the protracted legal battle that reached the level of the U.S. Supreme Court. The telephone could be placed in every home, and used by anyone without the need for intervention of skilled middlemen. For the first time,point-to-point communication was direct, reliable, relatively private, and the voice of the person at the other end could be recognized.

Another development was born out of a purely mathematical insight. In 1865, James Clerk Maxwell wrote down differential equations describing all that was then known about electricity and magnetism. These equations described how an electric charge created an electric field (Coulomb's law),

how an electric current created a magnetic field (Ampère's law), and how a changing magnetic field created an electric field. Far away from currents and charges the equations

$$\begin{aligned} \nabla \cdot \underline{E} &= 0 & \qquad \nabla \cdot \underline{B} &= 0 \\ \nabla \times \underline{E} &= -\tfrac{1}{c}\tfrac{\partial \underline{B}}{\partial t} & \qquad \nabla \times \underline{B} &= 0 \end{aligned}$$

were obviously not symmetric. To make them completely symmetric Maxwell hypothesized that a changing electric field could induce a magnetic field,

$$\begin{aligned} \nabla \cdot \underline{E} &= 0 & \qquad \nabla \cdot \underline{B} &= 0 \\ \nabla \times \underline{E} &= -\tfrac{1}{c}\tfrac{\partial \underline{B}}{\partial t} & \qquad \nabla \times \underline{B} &= +\tfrac{1}{c}\tfrac{\partial \underline{E}}{\partial t} \end{aligned}$$

a phenomenon that had not previously been observed. These new equations admitted a new type of solution, a changing electric field inducing a changing magnetic field reinforcing the original changing electric field. This *electromagnetic* field could travel at the speed of light (not surprising since light is exactly such a field) and carry a signal far away without the need for wires. In 1887, Hertz performed an experiment to test Maxwell's purely theoretical prediction. He made sparks jump between two polished brass knobs separated by a small gap, and detected the transmitted electromagnetic waves using a simple receiver of looped wire and similar knobs several meters away.

Radio waves can carry Morse or Baudot code by transmitting or not transmitting (on-off keying). They can also carry voice by continuously changing some characteristic of the field, such as its amplitude (AM) or frequency (FM). In the next section we will learn how this can be done.

EXERCISES

18.1.1 Compute the time durations of the 26 letters in Morse code. What is the average duration assuming all characters are equally probable? What is the average duration assuming that the letter probabilities are roughly E:12%, TAOINS:8%, HRDLU:4%, MCFGPB:2%, and all the rest 1%. Is Morse code better or worse than Baudot code for actual text?

18.1.2 Write a program that inputs a text file and outputs Morse code. You will need a computer with minimal sound capabilities. Whenever $s = 1$ play a tone (1000 Hz is good). Make the speed an adjustable parameter, specified in words per minute (figure an average word as 5 characters). Add an option to your program to output two different tones, a high-frequency tone for $s = 1$ and a low-frequency one for $s = 0$.

18.1.3 Modify the above program to output a file with sampled signal values (use a sampling rate of 8000 Hz and a tone of 750Hz). Now write a program that inputs this file and decodes Morse code (converts signal values back to text). Improve your program to take into account small amounts of noise and small variabilities in speed (and add these features to the generating program). Do you think you could write a program to read Morse code sent by hand on a noisy channel?

18.2 Analog Modulation Types

In our historical discussion we carefully avoided using the word 'modulation'; we now just as carefully define it.

Definition: modulation
Modulation is the exploitation of any observable characteristic of a signal to carry information. The signal whose characteristics are varied is called a *carrier*. We *modulate* the carrier by the information signal in order to create the modulated signal, and *demodulate* the modulated signal in order to recover the information signal. The systems that perform these functions are called the *modulator* and *demodulator*, respectively. ∎

Modulation is used whenever it is not possible or not convenient to convey the information signal directly. For example, a simple two station intercom will probably directly transmit the voice signals (after amplification) from one station to another over a pair of wires. This scenario is often called *baseband transmission*. A more sophisticated intercom system may modulate a radio signal, or the AC power signal, in order to eliminate the need for wires. The public switched telephone network uses wires, but maximizes their utilization by modulating a single base signal with a large number of subscriber signals.

Perhaps the simplest signal that is used as a carrier is the sinusoid.

$$s(t) = A\cos(2\pi f t + \phi) \tag{18.1}$$

For example, the very existence of the carrier can be used to send Morse or Baudot code. This is called **On-Off K**eying (OOK) and mathematically is represented by

$$s_{\text{OOK}}(t) = A(t)\cos(2\pi f_c t) \tag{18.2}$$

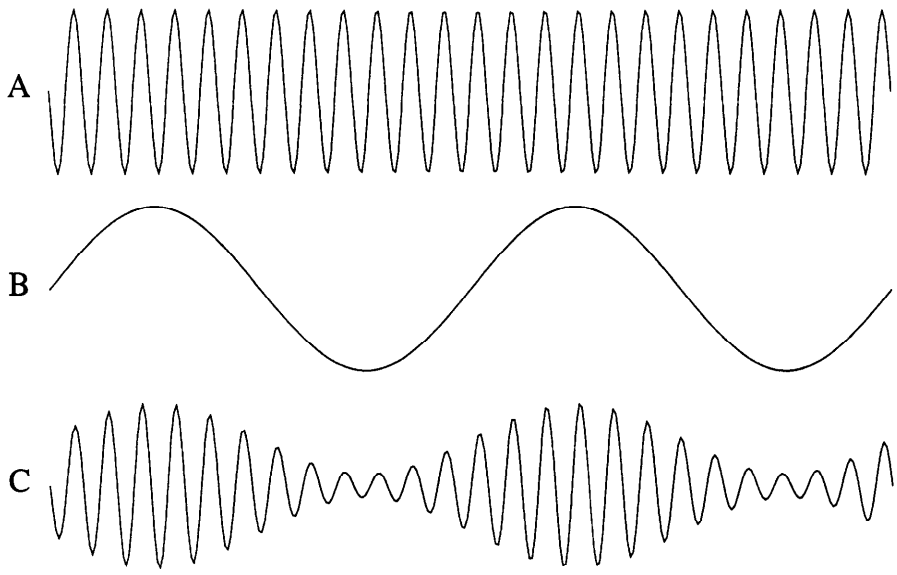

Figure 18.1: Amplitude modulation changes the amplitude of a carrier in accordance to a modulating signal. In (A) we see the carrier, in (B) a sinusoidal modulating signal, and in (C) the resulting AM signal (the modulation index was 75%).

where $A(t)$ takes the values zero or one, f_c is the carrier frequency, and (without limiting generality) we choose the phase to be zero. In order to carry voice or other acoustic modulating signals $v(t)$, we need more freedom. Now equation (18.2) is strongly reminiscent of the instantaneous representation of a signal of equation (4.66); but there the amplitude $A(t)$ was a continuously varying function. This leads us to the idea of conveying a continuously varying analog signal $v(t)$ by varying the carrier's amplitude

$$s_{\text{AM}}(t) = A_0 \left(1 + m_{\text{AM}}\, v(t)\right) \cos(2\pi f_c t) \tag{18.3}$$

where we assume $|v(t)| \leq 1$. This modulation technique, known as **A**mplitude **M**odulation (AM), is depicted in Figure 18.1. The coefficient $0 < m_{\text{AM}} \leq 1$ is known as the modulation index, and is often specified as a percentage.

Amplitude is not the only signal characteristic that one can modulate. The sinusoidal carrier of equation (18.1) has two more characteristics that may be varied, the frequency f and the phase ϕ. Morse- or Baudot-encoded text may be sent by **F**requency **S**hift **K**eying (FSK), that is, by jumping between two frequencies.

$$s_{\text{FSK}}(t) = A \cos\left(2\pi f(t)\, t + \phi\right) \tag{18.4}$$

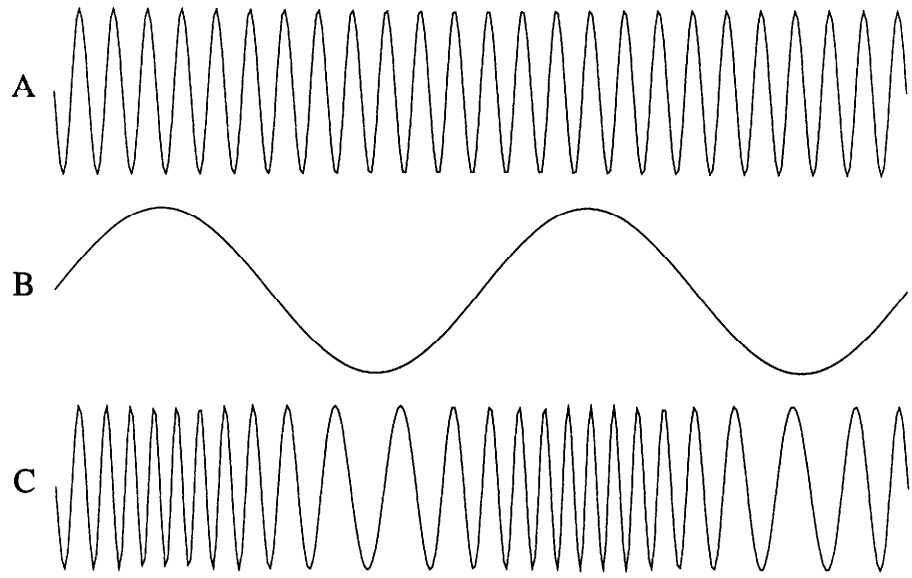

Figure 18.2: Frequency modulation changes the frequency of a carrier in accordance to a modulating signal. In (A) we see the carrier, in (B) a sinusoidal modulating signal, and in (C) the resulting FM signal.

Here it is $f(t)$ that can take on two different values. The third alternative is called **P**hase **S**hift **K**eying (PSK),

$$s_{\text{PSK}}(t) = A\cos\Big(2\pi f_c t + \phi(t)\Big) \tag{18.5}$$

where $\phi(t)$ can take on two values (e.g., 0° and 180°). Similarly, voice can be transmitted by **F**requency **M**odulation (FM) and by **P**hase **M**odulation (PM), as will be explained in the next section. For example, in Figure 18.2 we see the frequency of a sinusoid continuously varying in sinusoidal fashion.

We have still not exhausted the possibilities for modulation. The sinusoid, although the most prevalent carrier, is not the only signal that can be modulated. An alternative is to start with a train of pulses and modify their amplitudes (PAM), their relative timing (PPM) or their pulse widths (PWM). Another common occurrence is *secondary modulation* where modulated signals are used to modulate a second signal. For example, several AM-modulated voice signals may be used to frequency modulate a wideband radiotelephone link carrier. Sometimes it seems that the number of different modulation techniques that have been used in communications systems equals the number of communications systems designers.

EXERCISES

18.2.1 Why is equation (18.3) not simply $A_0 v(t) \cos(2\pi f_c t)$? Plot sinusoidally modulated AM signals for various values of modulation index. What index do you think should be used?

18.2.2 Write a program that generates an AM-modulated wave. (For concreteness you may assume a sampling frequency of 2.048 MHz, a carrier of 455 KHz, and take the modulating signal to be a sinusoid of frequency 5 KHz.) Plot 1 millisecond of signal. What does the spectrum look like?

18.2.3 Why do we prefer sinusoidal carriers to other waveforms (e.g., square waves)?

18.2.4 Can we simultaneously modulate with AM and FM? AM and PM? FM and PM?

18.3 AM

Now that we know what modulation is, we can commence a more systematic study of modulated signals and the signal processing systems used to modulate and demodulate. For now we are only interested in modulating with continuous analog signals such as speech; digital modulation will be treated later.

How can we create an amplitude modulated signal using analog electronics? The simplest way would be to first create the carrier using an oscillator set to the desired frequency. Next the output of this oscillator is input to an amplifier whose gain is varied according to the modulating signal (see Figure 18.3). Since both oscillators and variable gain amplifiers are standard electronic devices, building an AM transmitter in analog electronics

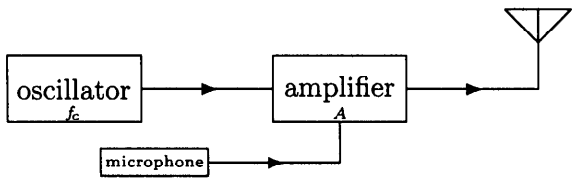

Figure 18.3: The basic analog AM transmitter built from an oscillator and a variable gain amplifier. The oscillator has a single parameter f_c that is not varied during transmission. The amplifier's gain parameter A is varied according to the signal. The inverted triangle at the top right is the conventional graphic representation of an antenna.

is straightforward. Of course there are lots of technical details to be dealt with, such as guaranteeing oscillator frequency stability, ensuring that the microphone's output is sufficiently strong, keeping the amplifier in its linear range, band-pass filtering the signal to avoid interference to nearby receivers, matching the input impedance of the amplifier with the output impedance of the oscillator, etc. Failing to properly cope with any of these details will result in inefficiency, low or distorted audio, or interference.

Wouldn't it be simpler to implement the AM transmitter using DSP? The analog oscillator and amplifier could be replaced with digital ones, and using correct digital techniques there will be no problems of efficiency, frequency stability, amplifier stability, impedance matching, etc. Although in principle this approach is correct, there are two practical problems. First, a digital amplifier by itself will only be sufficient for very low-power applications; in order to supply the high power usually needed (from about ten watts for mobile radios to many thousands of watts for broadcast stations) an additional analog *power amplifier* will usually be needed. Second, the bandwidth BW of the audio frequencies (AF) is usually much lower than the radio frequency (RF) of f_c. Directly implementing Figure 18.3 digitally would require us to operate at a sampling rate over twice $f_c + BW$, which would be extremely wasteful of computational power. Instead we can perform all the computation at an intermediate frequency (IF) and then upmix the signal to the desired radio frequency. Figure 18.4 shows a hybrid AM transmitter that utilizes digital techniques for the actual modulation and analog electronics for the upmixing and power amplification.

Now that we know how to transmit AM we need a receiver to demodulate our AM transmission. The simplest analog receiver is the *envelope detector*,

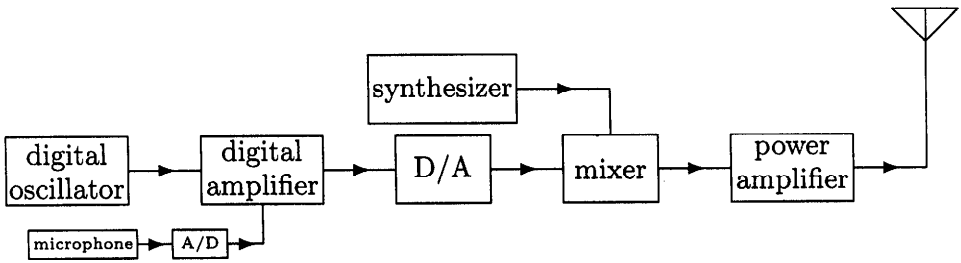

Figure 18.4: The basic hybrid digital-analog AM transmitter. The digital components operate at an intermediate frequency and at low power. After conversion to the analog domain the signal is upmixed to the desired carrier frequency and amplified to the required output power. The synthesizer is a (digital) local oscillator.

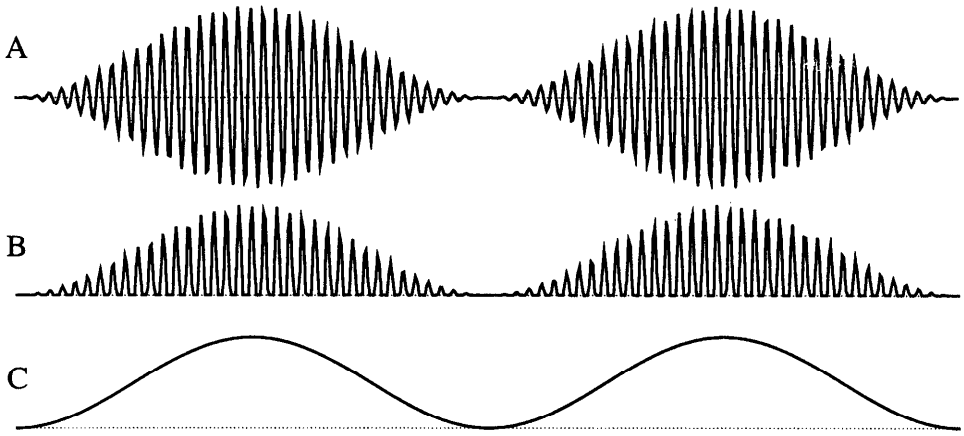

Figure 18.5: The basic analog envelope detector for the demodulation of AM signals. In (A) we see the AM signal to be demodulated. After half wave rectification the signal depicted in (B) results. Subsequent low-pass filtering removes the RF and leaves (C) the desired AF to within DC.

the operation of which can be best understood by studying Figure 18.5. Since the desired signal is the 'envelope' of the received signal, it can be retrieved from either the top or bottom of Figure 18.5.A by connecting the peaks. Choosing to use the top half, half wave rectification results in the signal of Figure 18.5.B. We next low-pass filter this signal in order to remove the high-frequency RF, leaving only the envelope as in Figure 18.5.C (with a strong DC component). This filtering is performed by placing the rectified signal onto a capacitor that charges up to the voltage peaks and slowly interpolates between them. Finally a DC blocking filter is used to remove the 1 from $1 + v(t)$.

Unfortunately, the envelope detector is ill suited to digital implementation. It assumes f_c to be very high compared to f_m, otherwise the envelope will not be well sampled, and thus downmixing to a low IF will decrease its efficacy. More importantly, in order to actually see the analog signal's peaks in its digital representation, a sampling frequency much higher than Nyquist is required. Even sampling at several times Nyquist we can not expect most of the sampling instants to fall close enough to the peaks.

A better way of digitally performing AM demodulation is to use the instantaneous representation of Section 4.12. There are two closely related ways of doing this. The first is to apply the Hilbert transform to the IF signal

and to obtain the instantaneous amplitude by the square root of the sum of the squares. The second involves a complex downmix to zero including a complex low-pass filter to remove everything except the frequency components from zero to BW. We can then proceed to obtain the instantaneous amplitude as before. These methods of digital AM demodulation do not require high f_c and function with sampling frequencies close to Nyquist.

Up to now we have been thinking of AM only in the time domain. What does the spectrum of an AM signal look like? We'll first consider modulating with a single sinusoid, so that equation (18.3) becomes

$$s_{\mathrm{AM}}(t) = A_0 \Big(1 + m_{\mathrm{AM}} \cos(\omega_m t)\Big) \cos(\omega_c t) \tag{18.6}$$

where ω_m and ω_c are the modulating and carrier angular frequencies. A little algebra proves

$$\begin{aligned} s_{\mathrm{AM}}(t) &= A_0 \cos(\omega_c t) + A_0\, m_{\mathrm{AM}} \cos(\omega_m t) \cos(\omega_c t) \\ &= A_0 \cos(\omega_c t) + m_{\mathrm{AM}} \frac{A_0}{2} \Big(\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t \Big) \end{aligned} \tag{18.7}$$

so that the spectrum contains three discrete lines, one corresponding to the original carrier frequency, and two lines at the carrier plus and minus the modulation frequency (Figure 18.6.A).

What if we modulate the carrier not with a single sinusoid but with a general signal $v(t)$? The modulating signal can be Fourier analyzed into a collection of sinusoids each of which causes two lines spaced f_m away from the carrier. We thus obtain a carrier and two *sidebands* as depicted in Figure 18.6.B. The two sidebands are inverted in frequency with respect to each other but contain precisely the same information.

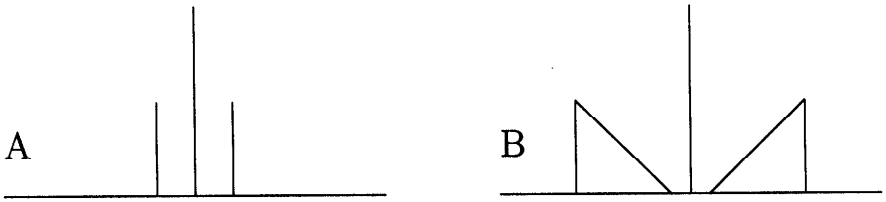

Figure 18.6: The generation of sidebands of an AM signal. In (A) we modulate a sinusoid of frequency f_c by a single sinusoid of frequency f_m to obtain an AM signal with three frequency lines, f_c, $f_c \pm f_m$. In (B) we modulate a sinusoid by a signal with an entire spectrum of frequencies, conventionally depicted as a triangle. We obtain the carrier and two sidebands.

EXERCISES

18.3.1 Our basic analog AM receiver assumed that only a single signal is received at the antenna, while in fact many signals are received simultaneously. One method of isolating the signal of interest uses a band-pass filter centered at f_c; the more conventional method uses a mixer and a band-pass filter centered at an *intermediate frequency* (IF). Diagram the two methods and discuss their advantages and disadvantages.

18.3.2 Diagram an entire AM receiver including antenna, local oscillator and mixer, IF filter, a half wave rectifier, a low-pass filter, DC blocking filter, and speaker. Show representative signals at the output of each block.

18.3.3 Implement a digital envelope detector. Create a sinusoidally modulated signal with $f_c = 50$, $f_m = 2$, and sampling frequency $f_s = 500$. Compare the demodulated signal with the correct modulating signal. Now decrease f_s to 200. Finally decrease f_c to 10. What do you conclude?

18.3.4 Show that half of the energy of an AM signal with index of modulation $m_{\text{AM}} = 1$ is in the carrier and one-quarter is in each of the sidebands.

18.3.5 Double sideband (DSB) is a more energy-efficient variant of AM, whereby the carrier is removed and only the two sidebands are transmitted. Diagram a transmitter and receiver for DSB.

18.3.6 Single sideband (SSB) is the most efficient variant of AM, whereby only a single sideband is transmitted. Diagram a transmitter and receiver for SSB.

18.3.7 Can AM demodulation be performed by a filter? If yes, what is its frequency response? If not, what portion of the analog and digital detectors is not a filter?

18.4 FM and PM

You might expect that frequency modulation of a carrier $A\cos(\omega_c t)$ with a signal $v(t)$ would be accomplished by

$$s(t) = A\cos\Big(\big(\omega_c + m\,v(t)\big)\,t\Big) \tag{18.8}$$

where m is the index of modulation. Indeed the amplitude is constant and the frequency varies around the carrier frequency according to the modulating signal; yet this is *not* the way FM is defined. To see why not, assume that

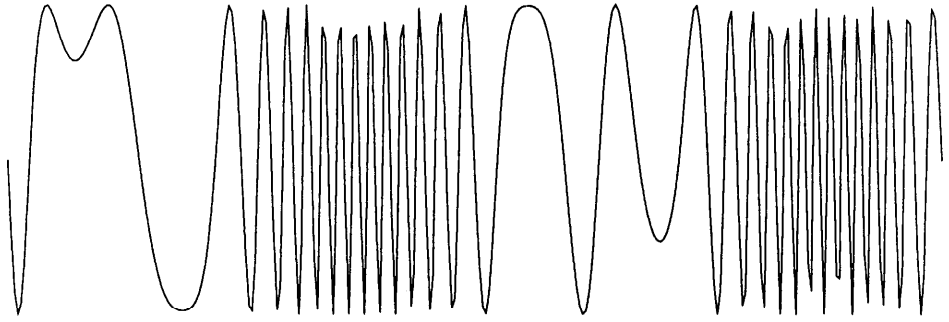

Figure 18.7: Frequency modulation according to the naive equation (18.8) has obvious artifacts. True frequency modulation should look sinusoidal.

the modulating signal $v(t)$ is a sinusoid (let's use sine rather than cosine this time) of frequency ω_m.

$$s(t) = A\cos\Big(\big(\omega_c + m\sin(\omega_m t)\big)\,t\Big) \tag{18.9}$$

Plotting this for time close to $t = 0$ results in a picture similar to 18.2.C, but for longer times we observe artifacts as in Figure 18.7. This is not what we expect from FM; in particular we want all the extrema of the signal to be those of the underlying carrier, whereas here we observe obviously nonsinusoidal extrema as well!

The reason for this errant behavior is not hard to see. The signal can be rewritten

$$s(t) = A\cos\Big(\omega_c t + m\,t\sin(\omega_m t)\Big)$$

and so has phase swings that increase linearly with time. For large t the phase swings completely dominate the argument of the sine except in the immediate vicinity of the modulating sinusoid's zeros, thus completely destroying the overall sinusoidal behavior. The solution to this problem is easy to see as well—we simply move the modulating sinusoid so that it is not multiplied by time

$$s(t) = A\cos\Big(\omega_c t + m\sin(\omega_m t)\Big) \tag{18.10}$$

or for a more general modulating signal

$$s(t) = A\cos\Big(\omega_c t + m_{\text{PM}}\,v(t)\Big) \tag{18.11}$$

which is known as **Phase Modulation** (PM).

There is a more direct way to arrive at PM. We think of a carrier signal $A\cos(\omega_c t + \phi)$ as having a degree of freedom not previously exploited—the phase ϕ. Having the phase vary with the modulating signal will create an information-bearing signal from which the modulating signal may be later retrieved, at least assuming the phase does not vary too much. We can use the PM index of modulation m_{PM} to ensure that the phase deviation does not exceed 2π, the point where ambiguity would set in.

True frequency modulation is similar to phase modulation, but not identical. Recalling equation (4.72) we realize that we can make the instantaneous frequency vary with a modulating signal by phase modulating by that signal's integral. If that is done, the information-bearing signal has no unwanted artifacts, and phase recovery followed by differentiation indeed restores the modulating signal.

$$s(t) = A\cos\left(\omega_c t + m_{\text{FM}}\int_{-\infty}^{t} v(\tau)d\tau\right) \tag{18.12}$$

For a modulating signal that consists of a single sinusoid, the entire difference between PM and FM is a phase shift and a change in the modulation index; for a more general modulating signal, FM and PM are less compatible. The integral of $v(t) = \sin(\omega_m t)$ is $-\frac{1}{\omega_m}\cos(\omega t)$, and so high-frequency Fourier components of $v(t)$ are much weaker in FM than in PM, a phenomenon known as *de-emphasis*. A PM signal heard on an FM receiver has too much treble and sounds 'tinny', while using a receiver designed for PM to intercept an FM signal produces a 'bassy' sound. FM may be generated using a PM transmitter, if *pre-emphasis* is performed on the modulating audio in order to compensate for the later loss of high frequencies.

The PM/FM transmitter is very similar to the AM one, with the exception that the amplified microphone voltage is used to vary the phase rather than the amplitude of the carrier; but how do we make an analog FM receiver? One way is to use frequency-to-voltage conversion to convert the received FM signal into an AM one. An *FM discriminator* is a circuit with gain that varies linearly with frequency, and can thus be used for the frequency-to-voltage conversion.

The digital FM receiver can derive the instantaneous frequency from the instantaneous phase through differentiation. Were we to drastically oversample we could get by with the simple difference, since

$$\phi(t+\delta t) - \phi(t) \approx \frac{d}{dt}\phi(t)\,\delta t$$

as long as the phase behaves approximately linearly in the time interval δt.

For more rapidly varying phases we must use a true differentiation filter (see Section 7.3).

The instantaneous phase signal is bounded in the interval between $-\pi$ and π (or perhaps $[0 \ldots 2\pi]$) and has discontinuities when it crosses these boundaries. These phase jumps have no physical meaning, they are simply artifacts of the nonuniqueness of inverse trigonometric functions. Differentiation of such discontinuities would give rise to tremendous unphysical spikes in the frequency demodulation. Hence we must first *unwrap* the phase before differentiation. This can be done by setting a phase change threshold, and adding $\pm 2\pi$ whenever the phase jumps by more than this threshold. For oversampled signals this threshold can be relatively small, but close to Nyquist it must be carefully chosen in order to avoid unwrapping legitimate changes in phase.

The unwrapped phase signal resulting from the above operation is considerably smoother than the original phase. If, however, the signal has not been correctly mixed down to zero frequency, the residual carrier frequency causes linear phase increase or decrease, which will eventually cause the phase to overflow. In sophisticated implementations one models this phase change by linear regression and corrects the mixer frequency accordingly. A simpler technique to avoid phase overflow is not to correct the phase at all, only the *phase difference*. Differentiation of the phase difference signal gives the frequency difference, and the actual frequency is found by adding the frequency difference to the previous frequency. This frequency is in the vicinity of the residual carrier frequency, and thus never overflows.

An alternative method of phase differentiation is called the dual differentiator method. It exploits the fact that the specific differentiation to be performed is

$$\frac{d}{dt}\Phi(t) = \frac{d}{dt}\tan^{-1}\left(\frac{y(t)}{x(t)}\right) = \frac{\dot{y}x - \dot{x}y}{A^2(t)} \tag{18.13}$$

where $A(t) = \sqrt{x^2(t) + y^2(t)}$ is the amplitude detection. If we are interested in the frequency alone, we can *limit* the input signal (giving a constant amplitude) and then the above is directly proportional to the instantaneous frequency. If the amplitude is to be calculated in any event, it should be done first, and then a division carried out.

We turn now to the spectrum of PM and FM signals, wondering whether there are sidebands here as there were in the AM case. Even if there are sidebands, they must be much different than those we saw for AM. For example, assume the power of the modulating signal increases. For AM the carrier remains unchanged and the sideband energy increases; for PM/FM

the total power must remain unchanged (otherwise there would be unwanted AM!) and thus an increase in sideband power must result in a decrease in carrier power. At some point the carrier will even have to entirely disappear! Using the same type of algebra that led to equation (18.7) we find

$$\begin{aligned} s(t) &= A\cos\Big(\omega_c t + m\sin(\omega_m t)\Big) \\ &= A\Big(\cos(\omega t)\cos\Big(m\sin(\omega_m t)\Big) - \sin(\omega t)\sin\Big(m\sin(\omega_m t)\Big)\Big) \end{aligned}$$

where m means m_{PM} or m_{FM}. Now $\cos(m\sin(\omega_m t))$ and $\sin(m_{\text{FM}}\sin(\omega_m t))$ are easily seen to be periodic signals with frequency ω_m. It turns out that these periodic functions have expansions in terms of the *Bessel functions* $J_0, J_1, \ldots$ (see A.1).

$$\begin{aligned} \sin(m\sin(\omega t)) &= 2\Big(J_1(m)\sin(\omega t) + J_3\sin(3\omega t) + \cdots\Big) \\ \cos(m\sin(\omega t)) &= J_0(m) + 2\Big(J_2(m)\sin(2\omega t) + J_4(m)\sin(4\omega t) + \cdots\Big) \end{aligned}$$

Plugging these in, and using the trigonometric product identities (A.32) multiple times, we obtain the desired spectral representation.

$$\begin{aligned} s(t) = A\Big(& J_0(m)\cos(\omega_c t) \\ & + J_1(m)\Big(\cos\Big((\omega_c+\omega_m)t\Big) - \cos\Big((\omega_c-\omega_m)t\Big)\Big) \\ & + J_2(m)\Big(\cos\Big((\omega_c+2\omega_m)t\Big) + \cos\Big((\omega_c-2\omega_m)t\Big)\Big) \\ & + J_3(m)\Big(\cos\Big((\omega_c+3\omega_m)t\Big) - \cos\Big((\omega_c-3\omega_m)t\Big)\Big)\Big) \\ & + \cdots \end{aligned} \tag{18.14}$$

This is quite different from equation (18.7) with its sidebands at $\omega_c \pm +\omega_m$! Here we have an infinite number of sidebands at $\omega_c \pm k\omega_m$ with amplitudes varying according to the Bessel functions. The carrier amplitude is proportional to J_0 and thus starts at unity for zero modulation index and decreases as m increases. All the sidebands start at zero amplitude for $m = 0$ and at first increase, but later oscillate. Of course, for constant modulation index m, the amplitude of the sidelobes tends to decrease with distance from the carrier. As a rough estimate we can say that $J_n(m)$ is close to zero for $n > m$, so that the number of significant sidebands is $2n$ and the bandwidth is given by $\text{BW} \approx 2n\omega_m$.

EXERCISES

18.4.1 Prove that equation (18.9) has extrema other than those of the carrier by differentiating and setting equal to zero.

18.4.2 Diagram an analog transmitter and receiver for FM.

18.4.3 Find the spectral representation of the PM signal.

$$s(t) = A\cos\Big(\omega_c t + m\cos(\omega_m t)\Big)$$

18.4.4 AM reception suffers from noise more than FM does, for the simple reason that additive wideband noise directly changes the received signal's amplitude, while most noise does not masquerade as frequency or phase changes. This is the reason FM is commonly used for high quality music broadcasting. Explain why FM receivers use a hard-limiter before the demodulator.

18.4.5 Communications-grade FM receivers come equipped with a *squelch* circuit that completely silences the receiver when no FM signal is present. Explain how this works and why such a circuit is not used in AM receivers.

18.4.6 What happens when two AM signals transmit too close together in frequency? What happens with FM?

18.5 Data Communications

Communications systems tend to be extremely complex. For example, a phone call starts with someone picking up the receiver (i.e., the telephone goes *off-hook*). This causes current to flow thus informing the local **C**entral **O**ffice (CO) that service has been requested. The CO responds by sending a signal composed of two sinusoids of 350 and 440 Hz called *dial tone* to the customer and starting up a rotary dialing pulse decoder and a DTMF receiver. The customer hears the dial tone and starts dialing. As soon as the CO notes activity it stops sending dial tone and starts decoding and collecting the digits. At some point the CO realizes that the entire number has been dialed and decides whether the call is local, long distance, overseas, etc. If the called party belongs to the same CO the appropriate line must be found, and whether it is presently in use must be checked. If it *is* in use a busy signal (480+620 Hz one half-second on and one half-second off) is returned to the calling party; if not, an AC *ring voltage* is placed on it, and a *ring-back* signal (440+480 Hz one second on and three seconds off) returned until someone answers by going off-hook. However, if the phone call

must be routed to another CO, complex optimization algorithms must be called up to quickly determine the least expensive available way to connect to the desired party. The calling CO then informs the called CO of the caller and callee phone numbers along a digital link using multifrequency tones or digital messages. The called CO then checks the called number's line and either returns an indication of busy, or places ring voltage and returns an indication of ringing. Of course we haven't mentioned caller identification, call waiting, billing, voicemail, etc.

If making a telephone call is *that* complex behind the scenes, just think of what happens when you surf the Internet with a web browser! In order to facilitate comprehension of such complex systems, they are traditionally divided into layers. The popular **O**pen **S**ystems **I**nterconnection (OSI) reference model delineates seven distinct layers for the most general data communications system, namely physical, datalink, network, transport, session, presentation, and application layers. At each layer the source can be considered to be communicating with the same layer of the destination via a *protocol* defined for that layer. In reality information is not transferred directly between higher layers; rather it is passed down to the physical layer, sent over the communications channel, and then passed up through the layers. Hence, each layer requires all the layers under it in order to function, directly accessing functions of the layer immediately beneath it and providing functionality to the layer immediately above it. The physical layer contains specifications of the cables and connectors to be employed, the maximum allowed voltage levels, etc. It also defines the 'line code' (i.e., the modulation type that determines how the digital information influences the line voltage). The datalink layer specifications are responsible for detecting and correcting errors in the data over a link (between one node in a network and the next), while the network layer routes information from the point of origin through the network to the destination, and ensures that the network does not become congested. The transport layer guarantees reliable source-to-destination transport through the network, while the session layer is where an entire dialog between the two sides is established (e.g., a user logs on to a computer) and maintained. The presentation layer translates data formats and provides encryption-decryption services and, finally, the application (e.g., email, file transfer, remote log-on, etc.) is the most abstract layer, providing users with a comprehensible method of communicating. Most of these layers do not require DSP. Their main function is packaging information into various-size 'chunks', tacking headers onto them, and figuring out where to send them.

EXERCISES

18.5.1 Assume that someone uses a dial-up modem to connect to the World Wide Web. Try to identify as many communications protocols as you can, and at what OSI layer(s) they operate. (Hint: The modem has connection, physical layer transfer and perhaps error correction facilities. The application on the user's computer uses a serial protocol to communicate with the service provider. The Internet is based on TCP/IP. The web sits above the Internet.)

18.5.2 Do we really need to divide communications systems into layers? If not, what are there advantages and disadvantages?

18.6 Information Theory

Digital communications involves reliably sending information-bearing signals from a *source* through a *channel* to a *destination.* Were there no channel this would be a trivial task; the problem is that the channel distorts and adds noise to the signal, adversely affecting the reliability. Basic physics dictates the (usually negative) effects of the channel, and signal processing knowledge helps design signals that get through these channels with minimal damage.

As anyone who has ever been on the Internet knows, we always want to send the information from source to destination as quickly as possible. In order to measure the speed of the information transfer we need to know how much information is in an arbitrary message. This is the job of information theory.

The basic tenet of information theory is that information content can always be quantified. No matter what form the information takes, text, speech, images, or even thoughts, we can express the amount of information in a unique way. We will always measure information content in bits. The rate of information transfer is thus measured in bits per second.

Suppose that I am thinking of a number x between 0 and 255 (for definiteness, $x = 137$); how much information is transferred when I tell you that number? You probably know the answer to that question, exactly eight bits of information. Formally, the reason that a number between 0 and 255 contains eight bits of information is that in general eight individual yes-no questions must be asked in order to find the number. An optimal sequence of questions is as follows:

Q_1: Is the $x_1 = x$ greater than or equal to 128? A_1: Yes ($x_1 = 137 \geq 128$).
Q_2: Is $x_2 = x_1 - 128$ greater than or equal to 64? A_2 No ($x_2 = 9 < 64$).
Q_3: Is $x_3 = x_2$ greater than or equal to 32? A_3: No ($x_3 = 9 < 32$).
Q_4: Is $x_4 = x_3$ greater than or equal to 16? A_4: No ($x_4 = 9 < 16$).
Q_5: Is $x_5 = x_4$ greater than or equal to 8? A_5: Yes ($x_5 = 9 \geq 8$).
Q_6: Is $x_6 = x_5 - 8$ greater than or equal to 4? A_6: No ($x_6 = 1 < 4$).
Q_7: Is $x_7 = x_6$ greater than or equal to 2? A_7: No ($x_7 = 1 < 2$).
Q_8: Is $x_8 = x_7$ equal to 1? A_8: Yes ($x_7 = 1$).

Only the number 137 will give this particular sequence of yes-no answers, and interpreting *yes* answers as 1 and *no* answers as 0 produces the binary representation of x from MSB to LSB. Similarly we can determine the number of bits of information in arbitrary messages by constructing a set of yes-no questions that uniquely determines that message.

Let's assume a source wishes to convey to the destination a message consisting of an integer between 0 and 255. The transmitter needn't wait for the receiver to ask the questions, since the questioning tactic is known. All the transmitter needs to do is to transmit the answers A_1 through A_8.

Signals that carry information appear to be random to some extent. This is because information is only conveyed by surprising its receiver. Constant signals, constant amplitude and frequency sinusoids or square waves, convey no information, since one can predict exactly what the signal's value will be at any time. Yet consider a signal that can take only two values, say $s = 0$ or $s = 1$, that can change in value every T seconds, but remains constant between kT and $(k+1)T$. Such a signal is often called a **N**on **R**eturn to **Z**ero (NRZ) signal, for reasons that will become clear shortly. If the signal jumps in an apparently randomly fashion between its two values, one can interpret its behavior as a sequence of bits, from which text, sound, or images may be derived. If one bit is inferred every T seconds, the information transfer rate is $\frac{1}{T}$ bits per second.

According to this point of view, the more random a signal is, the higher its information transfer rate. Longer T implies a lower information transfer rate since the signal is predictable for longer times. More complex predictable behavior also reduces the information transfer rate. For example, a **R**eturn to **Z**ero (RZ) signal (see Figure 18.8.B) is similar to the NRZ signal described above, but always returns to $s = 0$ for odd k (we count from $k = 0$). Since an unpredictable signal value only appears every $2T$ seconds, the information is transferred at half the rate of the NRZ signal. Predictability may be even more subtle. For example, the *Manchester* signal used in Ethernet LANs

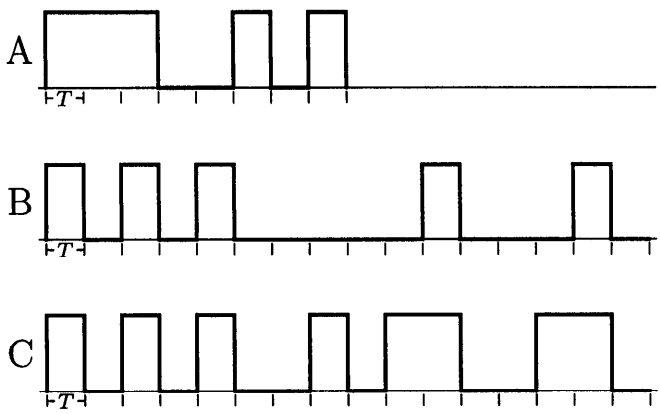

Figure 18.8: Comparison of (A) NRZ, (B) RZ, and (C) Manchester signals. The message is 11100101 and our channel bandwidth requires transitions to be spaced T seconds apart. Using NRZ this message requires $8T$ seconds. RZ and Manchester both require $16T$ seconds to transmit the same message.

(see Figure 18.8.C) encodes a binary one by having $s = 1$ for even k and $s = 0$ for the subsequent $k + 1$ interval; a zero is encoded by $s = 0$ followed by $s = 1$. Once again the information transfer rate is only half that of the NRZ signal, although the lack of randomness is less obvious. Whereas the NRZ signal has no correlation between signal values spaced T seconds apart, the Manchester signal never allows odd k intervals to have the same value as the previous even k interval.

The moral is that any correlation between signal values at different times reduces the amount of information carried. An infinite amount of information is carried by a signal with no correlation between different times (i.e., by white noise). Of course a true white noise signal, which has frequency components up to infinite frequency, cannot pass unaltered through a channel with finite bandwidth. Thus for a finite bandwidth channel, the signal with maximal information content is one whose sole predictability is that caused by the bandwidth constraint. Such a signal has a spectrum that is flat in the allowed pass-band.

We can similarly define the information transfer rate when the signal may take on many values (called symbols), not just zero and one. A signal that can jump randomly every T seconds, but that is a constant $s = 0, 1, 2,$ or 3 in between these jumps, obviously carries 2 bits every T seconds, or $\frac{2}{T}$ bits per second.

What if the different symbols are not equally probable? For example, a signal that takes on 26 values corresponding to a message containing text

in English would use the symbol corresponding to 'E' much more frequently than that corresponding to 'Q'. Information theory tells us that a consistent measure of information of a single symbol s is the *entropy*

$$H(s) = -\left\langle \log_2 p(s) \right\rangle = -\sum_s p(s) \log_2 p(s) \tag{18.15}$$

where s represents the possible signal values, and the triangular brackets stand for the expected value (see Appendix A.13).

To understand this result let's return to the simple case of a sending a message that consists of a number x between 0 and 255. Before transmission commences, the receiver has no information as to the value of x other than the fact that it is between 0 and 255. Thus the receiver assigns an equal probability of $\frac{1}{256}$ to each of the integers $0 \ldots 255$. A priori the transmitter may send a first symbol of 0 or 1 with probability $\frac{1}{2}$. In the previous example it would send a 1; immediately the receiver updates its probability estimates, now $0 \ldots 127$ have zero probability and $128 \ldots 255$ have probability 1 in 128. The receiver's uncertainty has been reduced by a factor of two, corresponding to a single bit of information. Now the second answer (in our example a zero) is sent. Since the second answer is independent of the first, the probability of both answers is the product of the individual probabilities $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Similarly, the probability of any particular sequence of three answers is $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. In general it is clear that after each subsequent answer is received the probability of the message is halved, as is the uncertainty of the receiver. After eight answers have been received the probability of the message has been reduced to $\frac{1}{256}$ and all uncertainty removed.

Now we prefer to think of information as being *added* after each answer has been received, although the probabilities were *multiplied.* The only way of making an arbitrary multiplication into an addition is to employ a logarithmic relation, such as (18.15). If we wish each reduction of probability by a factor of one half to correspond to the addition of a single bit, the base of the logarithm must be 2 and a minus sign must be appended (since $-\log_2 \frac{1}{2} = +\log_2 2 = 1$). Thus, for our simple example, each answer A_i contributes

$$I(A_i) = -\log_2 p(A_i) = \log_2 \tfrac{1}{2} = 1$$

bits of information. The information of the sequence of answers is

$$I(x) = \sum_{i=1}^{8} I(A_i) = -\sum_{i=1}^{8} \log_2 p(A_i) = 8$$

bits, as we claimed.

The guessing game with yes-no questions is not restricted to determining numbers; it can be played for other types of messages. For example, in trying to ascertain which person in a group is intended we could progressively ask 'male or female?', 'tall or short?', 'light or dark hair?', etc. until only one person remains. Indeed after a little thought you will become convinced that every well-defined message can be encoded as a series of answers to yes-no questions. The minimal number of such questions needed to unambiguously recover the message intended is defined to be the information content of that message in bits. In communications we are mostly interested in the rate at which information can be transferred from source to destination, specified in bits per second.

EXERCISES

18.6.1 Consider a signal that can take one of two values, $s = 0$ with probability p and $s = 1$ with probability $1 - p$. Plot the entropy of a single value as a function of p. Explain the position and value of the extrema of this graph.

18.6.2 Compute the entropy in bits per character of English text. Use the probabilities from exercise 18.1.1 or collect histograms using some suitably large on-line text to which you have access. Is a byte required to encode each letter?

18.6.3 Use a file compression program to reduce the size of some English text. What is the connection between final file size and entropy?

18.6.4 Repeat the previous two exercises for other languages that use the same alphabet (French, Spanish, Italian, German, etc.). Can these probabilities be used to discriminate between different languages?

18.6.5 What are the most prevalent *pairs* of letters in English? How can letter pairs be used to aid text compression? To aid in language identification?

18.6.6 Using Table 18.1 compute the time durations of Morse code letters and sort them in increasing order. Did Morse maximize the information transfer rate?

18.7 Communications Theory

We have seen that all information can be converted into bits (i.e., into digital information). Thus all communications, including those of an analog nature, *can* be performed digitally. That does not imply that all communications *should* be performed digitally, since perhaps the conversion of analog

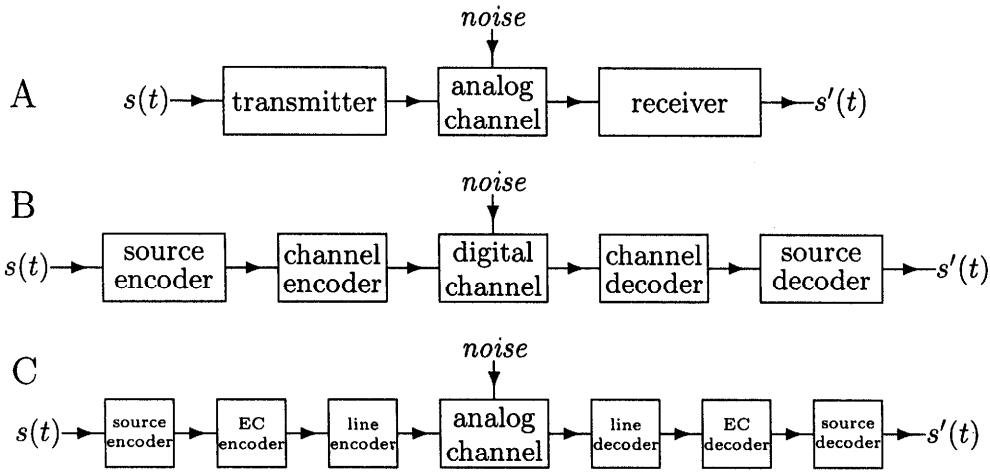

Figure 18.9: The conversion of an analog communications system into digital one. In (A) we see the original analog system. In (B) we have performed the separation into source and channel coding guaranteed by Shannon's theorem. In (C) we add line coding in order to utilize an analog channel (EC stands for error correction).

information into digital data, its transmission, reception, and reconversion into analog data would lead to a loss in quality or efficiency. In previous sections we learned how to transmit analog signals using AM, FM, and other forms of analog modulations. With such robust analog techniques at our disposal it does not seem likely that the conversion to digital communications would be useful.

In the late 1940s, Claude Shannon laid the mathematical framework for digital communications. Logically the first result, already deep and perhaps surprising, is that digital communications can always be used without sacrificing quality or efficiency. More precisely, Shannon showed that one could separate any communications problem, including an analog one, into two independent parts, *without sacrificing quality.* He called these parts *source coding* and *channel coding.* Source encoding refers to the process of efficiently converting the source message into digital data (i.e., representing the message as a bit stream with minimal number of bits). Channel encoding means the method of selecting signals to be sent over the communications channel. The inverse operations are channel decoding and source decoding, which convert the received signals back into digital data and convert the digital data back into the original message, respectively. By using this model, rather than directly transmitting an analog signal over an analog channel (Figure 18.9.A), we can efficiently convert an analog signal into a digital one,

send this essentially without error over a digital channel, and then recover the original signal (Figure 18.9.B).

It should be stressed that Shannon's separation of communications into two parts is fundamentally different from the OSI separation of communications into seven layers. There is no theory stating that the division of the OSI model does not impair the communications system; the layers are only separated in order to facilitate human comprehension. In a similar fashion the channel coding of Shannon's theorem is often further divided into two separate parts, *error correction coding* and *line coding*. An error correction code converts digital data into protected digital data, which can be transmitted over a digital channel with less fear of corruption due to noise. Of course all real transmission channels are analog, and so digital channels are actually an abstraction. The conversion of the (protected) digital signal into an analog one suitable for the physical transmission line is called *line coding*. The entire process is thus that of Figure 18.9.C. The division of the channel code into error correction code and line code is performed solely as an aid to the designers (it's hard to find one person expert in both fields!) but is not guaranteed to be conserve optimality. Indeed one can increase performance by combining the two (see Section 18.19).

Shannon's theorem, although in many ways satisfying, has not yet convinced us to convert over to digital communications systems. All we have seen is that we have nothing to lose by converting; we have yet to see that we have something to gain. Can digital systems actually *increase* bandwidth efficiency, *improve* the quality, *reduce* the cost, or provide any other measurable advantage as compared with analog communications? Shannon affirmatively answered these questions in a series of theorems about source and channel coding. Source coding theorems are beyond the scope of our present treatment, yet we can readily understand how proper source and channel coding can help us attain some of these goals.

For maximal efficiency source coding should produce a bit stream with no more bits than absolutely needed. We know that the minimal number of bits required to encode a message is the information (entropy), and thus the ideal source coder produces no more bits than entropy requires. For example, speech can be source encoded into 8 Kb/s or less (see Chapter 19) and there are modems (line codes) of over 32 Kb/s; hence using digital techniques one can transfer four conversations over a single telephone line. Thus proper source encoding can increase bandwidth efficiency.

Digital compact disks have replaced analog long playing records mainly due to their superior audio quality. This quality is obtained because of the use of digital error correcting channel codes that guarantee accurate re-

production of the original sound. Analog music signals that have become contaminated with noise cannot generally be corrected, and the noise manifests itself as various hisses and pops. Thus proper channel encoding can indeed increase signal quality.

While we will not delve into all of Shannon's theorems, there is one that will be essential for us. Before Shannon, engineers knew that noise and interference on digital channels cause errors in the reconstructed bit stream; and they thought that there was only one way of overcoming this problem, by increasing the power of the communications signal. The principle in which all designers believed was that no matter what the noise or interference is like, if we transmit a strong enough signal it will wipe them out. Then there was the separate issue of bandwidth; the higher the bandwidth the more data one could reliably transfer in a given time. Thus common wisdom stated that the probability of error for digital communications was a function of the SNR, while the speed was determined by the bandwidth. Shannon's capacity theorem completely changed this picture; by explaining that the SNR and bandwidth establish a maximum transmission rate, under which information could be transferred with arbitrarily low error rate. This result will be the subject of the next section.

EXERCISES

18.7.1 Shannon introduced entropy (defined in the previous section) in connection with source coding. The ultimate purpose of source coding is to produce no more bits than required by the entropy content of the source. When is simple A/D conversion the optimal source coding for an analog signal? What should one do when this is not the case?

18.7.2 In order to achieve the maximum efficiency predicted by Shannon, source coding is often required even for digital data. Explain and give several examples. (Hint: Data compression, fax.)

18.7.3 The Baudot code and ASCII are source codes that convert letters into bits. What are the essential differences between them? Which is more efficient for the transfer of plain text? How efficient is it?

18.7.4 In today's world of industrial espionage and computer hackers sensitive data is not safe unless encrypted. Augment the diagram of Figure 18.9.C to take encryption into account.

18.7.5 We often want to simultaneously send multiple analog signals (for example, all the extensions of an office telephone system) over a single line. This process is called *multiplexing* and its inverse *demultiplexing*. Show how this fits into Figure 18.9.C.

18.8 Channel Capacity

The main challenge in designing the physical layer of a digital communications system is approaching the *channel capacity*. By channel capacity we mean the maximum number of information bits that can be reliably transferred through that channel in a second. For example, the capacity of a modern telephone channel is about 35,000 bits per second (35 Kb/s); it is possible to transfer information at rates of up to 35 kilobits per second without error, but any attempt at perfectly transferring more data than that will surely fail.

Why is there a maximal channel capacity? Why can't we push data as fast as we wish through a digital link? One might perhaps believe that the faster data is transmitted, the more errors will be made by the receiver; instead we will show that data can be received essentially without error up to a certain rate, but thereafter errors invariably ensue. The maximal rate derives from two factors, noise and finite bandwidth. Were there to be no noise, or were the channel to have unlimited bandwidth, there would be unlimited capacity as well. Only when there are both noise *and* bandwidth constraints is the capacity finite. Let us see why this is the case.

Assume there is absolutely no noise and that the channel can support some range of signal amplitudes. Were we to transmit a constant signal of some allowable amplitude into a nonattenuating noiseless channel, it would emerge at the receiver with precisely the same amplitude. An ideal receiver would be able measure this amplitude with arbitrary accuracy. Even if the channel does introduce attenuation, we can precisely compensate for it by a constant gain. There is also no fundamental physical reason that this measurement cannot be performed essentially instantaneously. Accordingly we can achieve errorless recovery of an infinite amount of information per second. For example, let's assume that the allowable signal amplitudes are those between 0 and 1 and that we wish to transmit the four bits 0101. We simply define sixteen values in the permissable range of amplitudes, and map the sixteen possible combinations of four bits onto them. The simplest mapping method considers this string of bits as a value between 0 and 1, namely the binary fraction 0.0101_2. Since this amplitude may be precisely measured by the receiver in one second, we can transfer at least four bits per second through the channel. Now let's try to transmit eight bits (e.g., 01101001). We now consider this as the binary fraction 0.01101001_2 and transmit a constant signal of this amplitude. Once again this can be exactly retrieved in a second and thus the channel capacity is above eight bits per second. In similar fashion we could take the complete works of Shakespeare,

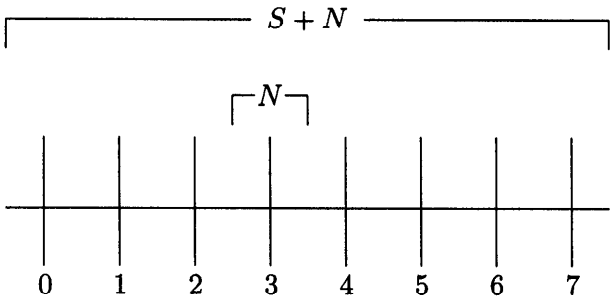

Figure 18.10: The effect of noise on amplitude resolution. The minimum possible spacing between quantization levels is the noise amplitude N, and the total spread of possible signal values is the peak-to-peak signal amplitude S plus the noise N. The number of levels is thus the ratio between the signal-plus-noise and the noise, and the number of bits is the base-two logarithm of this ratio.

encode the characters as bytes, and represent the entire text as a single (rather lengthy) number. Normalizing this number to the interval between 0 and 1 we could, in principle, send the entire text as a single voltage in one second through a noiseless channel. This demonstrates that the information-carrying capacity of a noiseless channel is infinite.

What happens when there *is* noise? The precision to which the amplitude can be reliably measured at the receiver is now limited by the noise. We can't place quantization levels closer than the noise amplitude, since the observed signals would not be reliably distinguishable. As is clarified by Figure 18.10 the noise limits the number of bits to the base-two logarithm of the signal-plus-noise-to-noise ratio, SNNR = SNR + 1.

Of course, even if the noise limits us to sending b bits at a time, we can always transmit more bits by using a time varying signal. We first send b bits, and afterwards another b bits, then yet another b, and so on. Were the channel to be of unlimited bandwidth we could abruptly change the signal amplitude as rapidly as we wish. The transmitted waveform would be piecewise constant with sharp jumps at the transitions. The spectral content of such jump discontinuities extends to infinite frequency, but since our channel has infinite bandwidth the waveform is received unaltered at the receiver, and once again there is no fundamental limitation that hinders our receiver from recovering all the information. So even in the presence of noise, with no bandwidth limitation the channel capacity is effectively infinite.

Signals that fluctuate rapidly cannot traverse a channel with finite bandwidth without suffering the consequences. The amount of time a signal must

remain relatively constant is inversely proportional to the channel bandwidth, and so when the bandwidth is BW our piecewise constant signal cannot vary faster than BW times per second. Were we to transfer an NRZ signal through a noisy finite-bandwidth channel we would transfer BW bits per second. By using the maximum number of levels the noise allows, we find that we can send $\text{BW}\log_2\text{SNNR}$ bits per second. Slightly tightening up our arguments (see the exercises at the end of this section) leads us to Shannon's celebrated channel capacity theorem.

Theorem: The Channel Capacity Theorem
Given a transmission channel bandlimited to BW by an ideal band-pass filter, and with signal-to-noise ratio SNR due to additive white noise:

- there is a way of transmitting digital information through this channel at a rate up to

$$C = \text{BW}\log_2(\text{SNR}+1) \tag{18.16}$$

bits per second, which allows the receiver to recover the information with negligible error;
- at any transmission rate above C bits per second rate no transmission method can be devised that will eliminate all errors;
- the signal that attains the maximum information transfer rate is indistinguishable from white noise filtered by the channel band-pass filter. ■

As an example of the use of the capacity theorem, consider a telephone line. The SNR is about 30 dB and the bandwidth approximately 3.5 KHz. Since $\text{SNR} \gg 1$ we can approximate

$$C = \text{BW}\log_2(\text{SNR}+1) \approx \text{BW}\log_2\text{SNR} = \text{BW}\frac{\text{SNR}_{dB}}{10\log_{10}2} \approx \text{BW}\frac{\text{SNR}_{dB}}{3}$$

and so C is about 35 Kb/s.

What the channel theorem tells us is that under about 35 Kb/s there *is* some combination of modulation and error correcting techniques that can transfer information essentially error-free over telephone lines. We will see later that V.34 modems presently attain 33.6 Kb/s, quite close to the theoretical limit. There will occasionally be errors even with the best modem, but these are caused by deviations of the channel from the conditions of the theorem, for example, by short non-white noise spikes. The reader who presently uses 56 Kb/s modems or perhaps DSL modems that transmit over telephone lines at rates of over 1 Mb/s can rest assured these modems exploit more bandwidth than 3.5 KHz.

The last part of the capacity theorem tells us that a signal that optimally fills the channel has no structure other than that imposed by the channel. This condition derives from the inverse relation between predictability and information. Recall from Section 5.2 that white noise is completely unpredictable. Any deviation of the signal from whiteness would imply some predictability, and thus a reduction in information capacity. Were the signal to be of slightly narrower bandwidth, this would mean that it obeys the difference equation of a band-pass filter that filters it to this shape, an algebraic connection between sample values that needlessly constrains its freedom to carry information.

The channel capacity theorem as expressed above is limited by two conditions, namely that the bandwidth is filtered by an *ideal* band-pass filter, and that the noise is completely *white*. However, the extension to arbitrary channels with arbitrary stationary noise is (at least in principle) quite simple. Zoom in on some very small region of the channel's spectrum; for a small enough region the attenuation as a function of frequency will be approximately constant and likewise the noise spectrum will be approximately flat. Hence for this small spectral interval the channel capacity theorem holds and we can compute the number of bits per second that could be transferred using only this part of the total spectrum. Identical considerations lead us to conclude that we can find the capacities of all other small spectral intervals. In principle we could operate independent modems at each of these spectral regions, dividing the original stream of bits to be transmitted between the different modems. Hence we can add the information rates predicted by the capacity theorem for all the regions to reach an approximate prediction for the entire spectrum. Let's call the bandwidth of each spectral interval δf, and the signal-to-noise ratio in the vicinity of frequency f we shall denote $\mathrm{SNR}(f)$. Then

$$C = \sum_f \log_2(\mathrm{SNR}(f) + 1)\ \delta f$$

and for this approximation to become exact we need only make the regions infinitesimally small and integrate instead of adding.

$$C = \int \log_2(\mathrm{SNR}(f) + 1)\ df \tag{18.17}$$

We see that for the general case the channel capacity depends solely on the frequency-dependent signal-to-noise ratio.

From the arguments that lead up to the capacity theorem it is obvious that the SNR mentioned in the theorem is to be measured at the receiver, where the decisions must be made. It is not enough to specify the transmitted

power at the frequency of interest $P(f)$ (measured in watts per Hz), since for each small spectral region it is this transmitted power times the line attenuation $A(t)$ that must be compared to the noise power $N(t)$ (also in watts per Hz) at that frequency. In other words, the SNR is $P(f)A(t)/N(f)$, and the total information rate to be given by the following integral.

$$C = \int \log_2 \left(\frac{P(f)A(t)}{N(f)} + 1 \right) df \tag{18.18}$$

Unfortunately, equation (18.18) is not directly useful for finding the maximal information capacity for the common case where we are given the line attenuation $A(t)$, the noise power distribution $N(f)$ and the total transmitted power P.

$$P = \int P(f)\, df \tag{18.19}$$

In order to find the maximal capacity we have to know the optimal transmitter power distribution $P(f)$. Should we simply take the entire power at the transmitter's disposal and spread it equally across the entire spectrum? Or can we maximize the information rate of an arbitrary channel by transmitting more power where the attenuation and noise are greater? A little thought leads us to the conclusion that the relevant quantity is the noise-to-attenuation ratio $N(f)/A(f)$. In regions where this ratio is too high we shouldn't bother wasting transmitted power since the receiver SNR will end up being low anyway and the contribution to the capacity minimal. We should start spending power where the N/A ratio is lower, and expend the greatest amount of power where the ratio is lowest and thus the received SNR highest.

In other words, we should distribute the power according to

$$P(f) = \begin{cases} \Theta - \frac{N(f)}{A(f)} & \frac{N(f)}{A(f)} < \Theta \\ 0 & \frac{N(f)}{A(f)} > \Theta \end{cases} \tag{18.20}$$

where the value of Θ is determined by the requirement (18.19) that the total Power should equal P. Gallager called this the 'water pouring criterion'. To understand this name, picture the attenuation to noise distribution ratio as an irregularly shaped bowl, and the total amount of power to be transmitted as the amount of water in a pitcher (Figure 18.11). Maximizing signal capacity is analogous to pouring the water from the pitcher into the bowl. Where the bowl's bottom is too high no water remains, where the bowl is low the height of water is maximal.

Figure 18.11: The water pouring criterion states that the information rate is maximized when the amount of power available to be transmitted is distributed in a channel in the same way as water fills an irregularly shaped bowl.

With the water pouring criterion the generalized capacity theorem is complete. Given the total power and the attenuation-to-noise ratio, we 'pour water' using equation (18.20) to find the power distribution of the signal with the highest information transfer rate. We can then find the capacity using the capacity integral (18.18). Modern modems exploit this generalized capacity theorem in the following way. During an initialization phase they probe the channel, measuring the attenuation-to-noise ratio as a function of frequency. One way of doing this is to transmit a set of equal amplitude, equally spaced carriers and measuring the received SNR for each. This information can then be used to tailor the signal parameters so that the power distribution approximates water pouring.

EXERCISES

18.8.1 SNR always refers to the power ratio, not the signal value ratio. Show that assuming the noise is uncorrelated with the signal, the capacity should be proportional to $\frac{1}{2}\log_2 SNR$.

18.8.2 Using the sampling theorem, show that if the bandwidth is W we can transmit $2W$ pulses of information per second. Jump discontinuities will not be passed by a finite bandwidth channel. Why does this not affect the result?

18.8.3 Put the results of the previous examples together and prove Shannon's theorem.

18.8.4 When the channel noise is white its power can be expressed as a *noise power density* N_0 in watts per Hz. Write the information capacity in terms of BW and N_0.

18.8.5 Early calculations based on Shannon's theorem set the maximum rate of information transfer lower than that which is now achieved. The resolution of this paradox is the improvement of SNR and methods to exploit more of the bandwidth. Calculate the channel capacity of a telephone line that passes from 200 Hz to 3400 Hz and has a signal-to-noise ratio of about 20–25 dB. Calculate the capacity for a digital telephone line that passes from 200 Hz to 3800 Hz and encodes using logarithmic PCM (12-13 bits).

18.8.6 The 'maximum reach' of a DSL modem is defined to be the distance over which it can function when the only source of interference is thermal white noise. The attenuation of a twisted pair of telephone wires for frequencies over 250 KHz can be approximated by

$$A(f) = e^{-s\left(\kappa_1\sqrt{f}+\kappa_2 f\right)L}$$

where L is the cable length in Km. For 24-gauge wire $\kappa_1 = 2.36 \cdot 10^{-3}, \kappa_2 = -0.34 \cdot 10^{-8}$ and for thinner 26-gauge wire $\kappa_1 = 2.98 \cdot 10^{-3}, \kappa_2 = -1.06 \cdot 10^{-8}$. Assume that the transmitter can transmit 13 dBm between 250 KHz and 5 MHz and that the thermal noise power is -140 dBm per Hz. Write a program to determine the optimal transmitter power distribution and the capacity for lengths of 1, 2, 3, 4, and 5 Km.

18.9 Error Correcting Codes

In order to approach the error-free information rate guaranteed by Shannon, modem signals and demodulators have become extremely sophisticated; but we have to face up to the fact that no matter how optimally designed the demodulator, it will still sometimes err. A short burst of noise caused by a passing car, a tone leaking through from another channel, changes in channel frequency characteristics due to rain or wind on a cable, interference from radio transmitters, all of these can cause the demodulator to produce a bit stream that is not identical to that intended. Errors in the reconstructed bit stream can be catastrophic, generating annoying clicks in music, causing transferred programs to malfunction, producing unrecoverable compressed files, and firing missile banks when not intended. In order to reduce the probability of such events, an *error correcting code* (ECC) may be used.

Using the terminology of Section 18.7, an ECC is a method of channel encoding designed to increase reliability. Error correcting codes are independent of the signal processing aspects of the bit transfer (line coding); they are purely mathematical mechanisms that detect whether bits have become