1

INTRODUCTION

PROLOGUE

How does one transmit information when bandwidth is expensive? One can explore new wavelengths. One can use multiple antennas to distinguish more pathways. One can design better modulation and coding for each pathway. This book is about the last.

According to the laws of nature, sending bits of information via an electrical medium requires two main resources: energy and bandwidth. Each bit needs energy, and just as in soccer football, a certain energy is needed to hit the ball far enough. Adding bandwidth to football needs new rules. Suppose every player has a ball and all must pass through a narrow passage on the way to the goal. How difficult is it to reach the goal?

Electrical communication is a game too, played according to nature's laws. In the first half of the twentieth century, several of these became clear. First, much less energy is needed per bit if the signal bandwidth is widened. Second, a certain type of energy pulse, called "orthogonal," is easier to process. These facts were well established in 1949 when Claude Shannon published something different, a formula for the ultimate capacity of a set bandwidth and energy to carry information. The key to approaching that limit is coding, that is, imposing clever patterns on the signals. As much as 90% of signal energy can be saved compared to rudimentary methods, or the information can be sent much further. In principle, whatever configuration of energy and bandwidth was available, it could be coded.

Bandwidth Efficient Coding, First Edition. John B. Anderson.

Nonetheless, those who designed codes in the following decades thought of them, consciously or not, as trading more bandwidth for less energy. Even with a simple trade, a signal could travel much further if its bandwidth were scaled up. In this way—and only this way—signals reached Mars and even Pluto. Neither was crowded with players, so that wide bandwidth signals were practical. There will soon be rovers from several nations on Mars, so that allotting wide bandwidth to each is not quite so practical. Here on Earth the situation is more desperate. Such services as cellular wireless and digital video must share a very crowded spectrum; furthermore, governments have discovered they can force providers to pay astonishing prices for bandwidth. Today, bandwidth costs far more than energy. Everyone needs to minimize it.

In the 1970s, methods of signal coding were discovered that did not increase bandwidth; signals could travel farther and easier without scaling up bandwidth. This was clear in Shannon's 1949 formula from the first day, but the concepts took time to sink in. What his formula really says is that coding leads to large savings no matter what the combination of energy and bandwidth. Today, with bandwidth costly, we want to work at narrow bandwidth and higher energy, not the reverse. The problem is: We know little about how to design the coding. The purpose of this book is to take that next step. How should efficient signal coding work when very narrow bandwidths per bit are available?

1.1 ELECTRICAL COMMUNICATION

This chapter introduces important and potentially troublesome concepts in a mostly philosophical way. Among them are bandwidth and time, pulse shapes, modulation, and coding. The needed formal communication background is Chapter 2, and Shannon's theory appears in Chapter 3. Certain concepts need some evolution to fit modern needs. The outcome of the book is that practical coding schemes exist that work well in a very narrowband world.

Communication transmits messages through time and space. In this book, the medium is electrical signals, and we will restrict ourselves to radio. For our purposes, messages are composed of symbols, and we want to transmit these accurately and efficiently through the physical world. In communication engineering, sending a symbol has three basic costs, energy, bandwidth, and implementation complexity. Each trades off against the others. Once the three costs are set the measure of good performance is most often probability of symbol error.

Even a century ago, it was clear enough that error may be reduced by spending more energy. A basic fact of communication, that first became evident with FM broadcasting, is that for the same performance, energy may be traded for bandwidth. Error-correcting codes, as first used, were thought of as a manifestation of this fact: By transmitting extra check bits, energy could be saved overall. But Claude Shannon's 1949 work [7] implied that energy or bandwidth or both could be reduced while maintaining the same error performance. Each combination of energy and bandwidth had a certain capacity to carry symbols but alas, there was no free lunch, and rapidly diverging "coding" complexity is required to approach this capacity. Today, energy, bandwidth, and complexity are a three-way trade-off.

Since 1949 there have been 60 years of progress and a rich research literature, but coding and its attendant complexity have been associated mainly with relatively low energy and wide bandwidth on a per data bit basis. Economic necessity and some changes in physical transmission paths are forcing changes in that thinking. Two examples are wireless networks and satellite digital broadcasting. Successive mobile network generations have offered more bits per second to their customers and the upcoming fifth Generation Systems will employ a variety of methods to offer even more. Among them are better antennas and shorter paths, both of which increase the symbol energy, making possible narrower bandwidth. According to theory, there is just as much to gain from coding and complexity in this narrowband/high-energy regime as we have enjoyed in wider band applications.

We move now to fundamental concepts and a sketch of the book. Knowledge of history helps make sense of a subject, and so we offer some high points, together with some important published works. The prerequisites for the book are first courses in probability, Shannon theory, signal space theory, and coding methods of the traditional type. Our own treatment of these subjects will be to extend their results to the new narrowband field.

Bandwidth and Coding. Coding, bandwidth, and their interaction lie at the heart of this book.

Coding theory is built upon abstract models, nonphysical concepts such as channel models, information, symbols, and arithmetic operations. Although it is not necessary to its mathematics, the theory can suggest physical conclusions by starting with signals and symbols that are avatars to physical channels and transmissions. In this book, the physical channel is always the white noise linear channel, wherein a white noise stochastic process with a certain power spectrum $N_0/2$ watts/Hz (the *noise*) adds to a analog function of time (the *signal*). Something is needed to convert abstract symbols to the analog domain; this is a *modulator*. This last is explored in Section 1.2. A modulator—demodulator by itself has a certain probability of error. The import of Shannon's work is that another two boxes, the *encoder—decoder*, can be added that can in principle reduce the error rate virtually to zero, provided that the energy–bandwidth combination is sufficient. The distinction between coding and modulation is a tricky one, taken up initially in Section 1.4, and the Shannon theory tools we need are in Chapter 3. That chapter starts from the core result in the 1949 paper, which is the capacity of the additive white Gaussian noise (AWGN) channel. In this channel, an independent Gaussian-distributed noise value with mean zero and variance $N_0/2$ is added to a real signal value. Initially, there were no analog signals; part of Shannon's genius was making the critical jump from the abstract AWGN channel to the channel with analog signals and bandwidth that we need in this book.

While it is true that the promise of narrow band/high-energy coding was already clear in 1949, we have spent most of the time since then developing techniques in the opposite regime. During its first 25 years, coded communication was primarily about parity-check codes and the binary symmetric channel. Codewords contained extra "redundant" bits, there being no other way to distinguish codeword bits from customer bits. With such codes nearly universal, one could easily fall into the belief that coding mostly exchanged extra bits (i.e., bandwidth) for coding gain (reduced energy). With the advent of coded modulation in the mid 1970s, cracks appeared in this belief. New transmission methods such as continuous-phase modulation (CPM) and set-partition coding appeared, which reduced energy without bandwidth expansion, and sometimes did this without redundant bits. By the late 1970s, it was clear that codes could even be used to *reduce* bandwidth, or even bandwidth and energy both.[1] These new methods eventually broke through the coding equals bandwidth-expansion belief, and doubled or tripled the rate of coded communication in a given bandwidth. Today, we would like to double or triple it once more.

More changed in the 1970s and 1980s than the bandwidth/energy regime. If coding was not redundant bits, then what was it? If CPM, which seemed to be modulation, and convolutional codes both required a trellis decoder of similar size, then were they not both coding schemes? Set-partition codes contained a convolutional code within them; was that the code, or was something else? Should removal of intersymbol interference (ISI) be called decoding? It became clear that coding needed more careful definition if there were to be reliable structure and language for our work.

So to bandwidth itself. To the philosophical, bandwidth is a measure of *changeability*, referenced to a unit of time. In the popular mind, it is the number of data bits a system such as a telephone handles, again per unit time. To a prosaic communications engineer, it is simply the width of a Fourier transform. These views are not really different: A signal can only convey information by changing, the more change the more information, and a signal with a narrow Fourier spectrum is one that changes slowly. All three views carry within them a unit of time, in MKS units the second. These views reflect our life experience, but communication theory is based on the *product* of changeability and time, the total change however accumulated. To send a certain quantity of information requires a certain accumulation. The product unit is the Hertz-second, a subtle concept explained more fully in Section 1.3.

This book is about more efficient transmission via coding when the relative bandwidth is narrow. The channel is linear and nonchanging, with simple white noise. We can look forward to second-generation research in synchronization, fading and nonlinear channels in the future, but they are not here yet.

1.2 MODULATION

Modulation is the conversion of symbols to physical signals. In this book, they are analog signals. Effectively, it is digital-to-analog conversion. We may as well assume that data to be transmitted arrive as bits, and the modulator then accepts these $\log_2 M$ bits at a time. We normally think of a modulator as accepting $\log_2 M$-bit groups and

[1] The author recalls adding some provocation to conference presentations in the 1970s by suggesting that the coding schemes presented did not increase bandwidth. Fortunately, the result was only laughter.

applying some process to each one. Most often the modulation process works from a set of M alternatives; for example, it may produce M tones or phases or amplitudes. We associate each M-fold alternative with a piece of transmission time, the *symbol time*, denoted T_s, or when no confusion can result, simply T. The average transmitted energy during T_s is the *symbol energy* E_s.

Referring spectra and energy back to the data bits divides their values by $\log_2 M$. This leads to the soundest analytical picture, but the M that best exploits the channel resources is often nonbinary. We will reserve *transmission symbol* to mean this M-ary set of alternatives.

Within this framework, a formal definition of modulator for this book is "*A conversion of all possible M-ary symbol sequences to analog signals by repeated application of a fundamental operation to each symbol*".

Most modulators in use today are *pulse* modulators, meaning that they associate each symbol with a pulse according to

$$s(t) = \sqrt{E_s} \sum_n u_n h(t - nT), \tag{1.1}$$

where u_n is the symbol and $h(t)$ is the *base pulse*. The symbols are pulse amplitudes and by convention they are independent random variables with zero mean and unit variance; consequently, E_s in Eq. (1.1) is the average symbol energy. The process here is called *linear modulation*, because the pulses simply add. In the classic modulator, the pulses are T-orthogonal, meaning

$$\int h(t)h(t - nT)\,\mathrm{d}t = 0, \qquad n \text{ an integer, } n \neq 0. \tag{1.2}$$

The great majority of modulators in applications heretofore are linear and use orthogonal pulses. Nonlinear modulators have some use, and classic examples are frequency-shift keying and the CPM signaling in Chapter 6. There is little loss in Shannon theory from the linearity requirement, but the same is not true with orthogonality. There is evidence that it leads to loss in many situations and most new methods in this book dispense with it. Its advantage is that it leads to a simple optimum detector; this is explored in Section 2.2.

To maintain the unit variance property on u_n, a binary symbol alphabet needs to be $\{+1, -1\}$. A uniformly spaced 4-ary modulator has alphabet $(1/\sqrt{5})\{\pm 3, \pm 1\}$ and an 8-ary $(1/\sqrt{21})\{\pm 7, \pm 5, \pm 3, \pm 1\}$. These three standard modulators will be referred to by their traditional names 2PAM, 4PAM, and 8PAM, where PAM means pulse-amplitude modulation. There can be a small advantage from nonuniform spacing, especially in bandpass schemes, but we will not pursue this in the book.

Equation (1.1), without a sinusoidal carrier, is said to be in *baseband* form. The base pulse $h(t)$ and the signal $s(t)$ ordinarily have a lowpass spectrum. Most applications employ carrier modulation, which is the same except that the spectrum of $s(t)$ is translated up by the f_c Hz, the carrier frequency. The translation is performed through multiplication by either $\sin 2\pi f_c t$ or $\cos 2\pi f_c t$, and the final signal has the form

$$s(t) = \sqrt{2E_s}\,[I(t)\cos 2\pi f_c t - Q(t)\sin 2\pi f_c t]. \tag{1.3}$$

Here $I(t)$ and $Q(t)$ are both baseband, that is, lowpass, signals called respectively, the in-phase and quadrature signals, and the outcome is a signal with a narrow spectrum centered at f_c Hz. $I(t)$ and $Q(t)$ satisfy

$$\mathcal{E}\Big[\int_T [I^2(t) + Q^2(t)]\,dt\Big] = 1,$$

where $\int_T$ is over a signal interval; consequently E_s is the symbol energy, this time for a dual symbol and two signals. Equation (1.3) is said to be a *passband* signal, written in the in-phase and quadrature, or I/Q, form.[2] Observe that one passband signal corresponds to two baseband signals. Each can take an independent Eq. (1.1) and they can be detected independently if $\cos 2\pi f_c t$ is known. Any passband signal can be constructed in this way.

Passband signals are essential because they allow many signals to be sent through the same medium and the properties of different wavelengths to be exploited. But for several reasons we will treat primarily baseband signals in this book. The chief one is that with perfect phase synchronization, $I(t)$ and $Q(t)$ are obtainable and independent and there is no reason to add the complexity of the passband form. A passband notation is a statement that there is imperfect synchronization or that there exist channel distortions that affect I and Q differently. These are interesting topics, but the schemes in this book have not yet reached that state of the art.

Pulse Properties. At first it may seem that a practical pulse is one that takes place wholly in its own interval, but it became clear by the 1950s that not much reduction in implementation comes from this and the bandwidth properties are far worse. The serious study of pulse shapes began with Harry Nyquist in 1924 [5], who studied pulses that have zeros at all integer multiples of T. This property is now called the *Nyquist pulse criterion* (NPC). He proved that a sufficient condition for a symmetric NPC pulse is

Property 1.1 (**The Spectral Antisymmetry Condition**) *A sufficient condition that a symmetrical $h(t)$ be NPC is that $H(f)$, the Fourier transform, is antisymmetric about the points $[H(f), f] = [H(0)/2, 1/2T]$ and $[H(0)/2, -1/2T]$.*

Note that a symmetric pulse has a real, symmetric transform. Later researchers found the necessary and sufficient condition and removed the requirement that $h(t)$ be time symmetric. These issues are discussed in introductory texts ([1], Section 2.2; [2], [3]). However, there is little reason in theory or in applications to abandon

[2] A common alternate framework takes $s(t)$ as the real part of $h(t)\exp j2\pi f t$, where $h(t)$ is now complex. See, for example, the text [3]. The method is used because complex numbers mimic the needed operations, not because there are unreal signals.

symmetric pulses, and we will use them and the antisymmetry condition almost exclusively.

In a later paper [6], Nyquist observed that there seemed to be a lower limit of about $1/2T$ Hz to the bandwidth of an NPC pulse. Research by others eventually developed the following theorem and the closely allied sampling theorem:

Property 1.2 *The bandwidth of any NPC pulse with zeros at nT cannot be narrower than* $1/2T$ *Hz, and the narrowest pulse is* $h(t) = A\,\mathrm{sinc}(t/T)$*, A is a real constant.*

Here sinc(x) is defined as $\sin(\pi x)/\pi x$. The pulse is clearly NPC; that no narrower pulse has zeros at nT is shown in the text references. The sinc pulse plays a major role in communication theory and will come back in Section 1.3.

Although the zero crossing property was important in Nyquist's time, today it is pulse orthogonality that matters, because such pulses have a simple optimum receiver. But the two concepts are closely related. Equation (1.2) is simply a statement that the autocorrelation of $h(t)$ is itself an NPC pulse. The Fourier transform of this autocorrelation is always $|H(f)|^2$, whether or not h is symmetric. This leads to two properties:

Property 1.3 (**Nyquist Orthogonality Criterion**) *$h(t)$ is an orthogonal pulse if and only if its autocorrelation function is NPC.*

Property 1.4 *A sufficient condition that $h(t)$ is orthogonal is that $|H(f)|^2$ is antisymmetric about the points* $[|H(f)|^2, f] = [|H(0)|^2/2, 1/2T]$ *and* $[|H(0)|^2/2, -1/2T]$.

Orthogonal Pulse Examples. An obvious example is the square pulse

$$v(t) = \sqrt{T}, \quad -T/2 < t \le T/2,$$
$$0, \quad \text{otherwise}, \tag{1.4}$$

shown here as a unit energy pulse. Its spectrum is $H(f) = \sqrt{1/T}\,\mathrm{sinc}(f/T)$. This pulse has very poor spectral properties, since $|H(f)|$ decays only as $\approx 1/fT$, which is far too slowly for applications. The spectrum does not have the spectral antisymmetry (Property 1.4), but $h(t)$ is nonetheless T-orthogonal.

An important practical example is the *root-raised-cosine* (root RC) pulse, so named because its square spectrum $|(H(f)|^2$ obeys Property 1.4 with an antisymmetric piece of a raised-up cosine. The unit-energy time pulse is

$$h(t) = \frac{\sin[\pi(1-\alpha)t/T] + (4\alpha t/T)\cos[\pi(1+\alpha)t/T]}{\sqrt{T}(\pi t/T)[1-(4\alpha t/T)^2]}, \qquad t \ne 0,\ t \ne \pm T/4\alpha;$$
$$(1/\sqrt{T})[1-\alpha+4\alpha/\pi], \qquad t = 0;$$
$$(\alpha/\sqrt{2T})[(1+2/\pi)\sin(\pi/4\alpha) + (1-2/\pi)\cos(\pi/4\alpha)], \qquad t = \pm T/4\alpha. \tag{1.5}$$

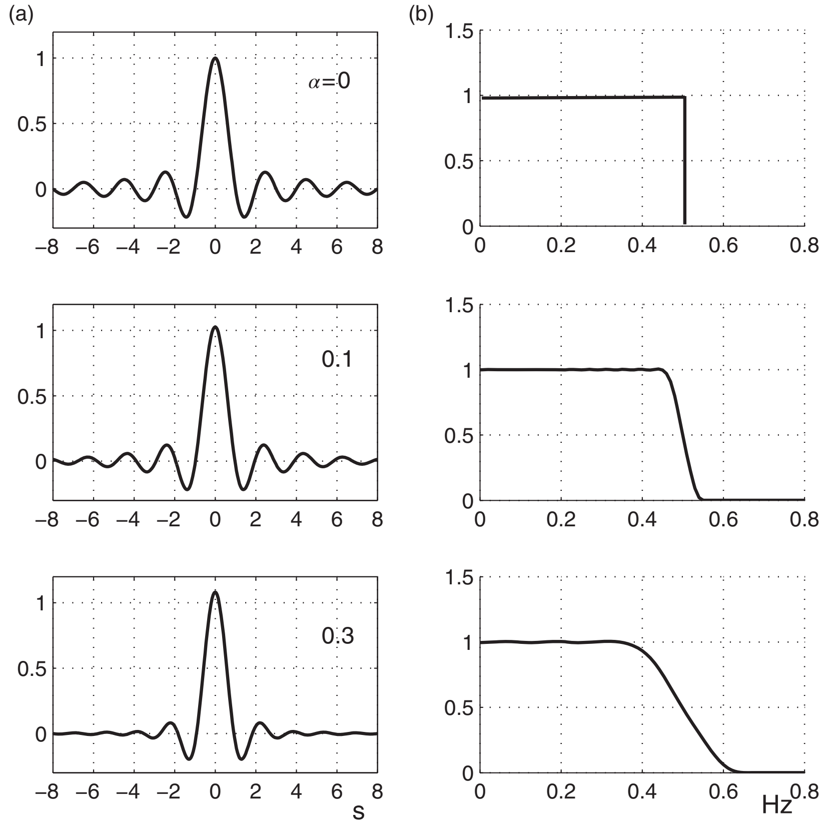

FIGURE 1.1 Root RC orthogonal pulses (a) and their spectra (b) with excess bandwidth $\alpha = 0, 0.1, 0.3$. Unit energy, $T = 1$.

The spectrum is

$$
\begin{aligned}
|H(f)|^2 = \;& 1, \qquad 0 \le f \le (1-\alpha)/2T; \\
& \cos^2[\frac{\pi T}{2\alpha}(f - \frac{1-\alpha}{2T})], \qquad (1-\alpha)/2T < f < (1+\alpha)/2T; \\
& 0, \qquad \text{elsewhere.}
\end{aligned} \tag{1.6}
$$

Here $\alpha \ge 0$, the excess bandwidth factor is the fraction by which the pulse bandwidth exceeds $1/2T$ Hz. Figure 1.1 shows the pulse and spectrum for $\alpha = 0, 0.1, 0.3$, the $\alpha = 0$ case being the sinc pulse. The time main lobe does not differ much but the side lobes rapidly diminish as α grows. The 30% case is arguably the most common pulse in applications, and it will be the standard pulse in most of the book.

The root RC pulse family shows a central fact about pulse design, that bandwidth trades off against time duration. If small interferences with neighboring channels matter, even a small change in bandwidth has a major effect. The uncertainty principle of Fourier analysis states that signal time and bandwidth have a constant product. Beyond this, if ever narrower bandwidth is demanded for a fixed symbol time T, a point must be reached where pulses cannot be orthogonal (it is $\approx 1/2T$ Hz).

The spectra of modulated signals will be taken up in the next section; the error performance and optimum receivers are the subject of Sections 2.1–2.2.

1.3 TIME AND BANDWIDTH

Signal spectra are a crucial issue in this book. Transmission capacity is more sensitive to bandwidth than to energy or signal complexity. Not only width matters but also the *shape* of the spectrum, and particularly, the stop-band side-lobes. The sensitivity heightens as the signal bandwidth efficiency grows. Since bandwidth efficiency is the reason for the book, spectra are a central issue.

The Hertz-Second. Information is of itself timeless. Yet we live in a world where activities are measured by time, and as communication engineers we measure signal bandwidth. In the transmission of information by signals, these resources trade off against each other. For a given parcel of information, if we want to send it faster, we use less time and more bandwidth; if bandwidth is scarce, we take more time. Earlier in the chapter simple modulation signals were defined. By scaling time A-fold faster and power A-fold larger, symbols transmit in the same energy per bit, but are A-fold faster and in A-fold wider bandwidth. If we think in terms of a *joint* time–bandwidth resource, the scaling sets the latency of the transmission but nothing else changes, at least in free space and with sufficient technology. If there are many parcels to send through the same channel, the time–bandwidth available can be divided among the parcels in many ways until the total resource is consumed. The time–frequency consumed by a given transmission will be called its *occupancy* later in the book.

Humans are not timeless, wideband beings with no opinion about latency, but in every application delays up to a limit are acceptable. Thus, some time–bandwidth trade-off is possible and we often accept a time–bandwidth product view. Another compelling reason for such a view is communication theory, which expresses itself most fundamentally and yields its best insight in terms of this product. The view dominates this book. Its unit is the Hertz-second (Hz-s). According to Webster, the Hertz is a “ … unit of frequency of a periodic process equal to one cycle per second.” A Hz-s is thus dimensionless in the sense that it does not refer to an arbitrarily defined unit like a second. An alien being will see the same quantity that we do.

A fundamental unit of efficiency in communication theory is bits per Hertz-second, abbreviated b/Hz-s. We will call this *bit density* to distinguish it from the more common and more loosely used word *rate*, which can mean input bits per output bit or per second depending on the context. We will avoid common measures of bandwidth efficiency such as bits per Hertz, or its reciprocal Hertz per bit, since these are dimensionally incorrect, and they assume that a second of time has elapsed—they will confuse an alien friend.

Bandwidth Criteria. To form a time–bandwidth product, one must measure bandwidth, but this is not straightforward. To begin with, a finite-time signal has an infinitely wide spectrum. Second, one wishes a single-number measure, whereas the spectrum of a signal is an entire function of frequency. As well, one wants a measure that makes sense with both practical and theoretical signals. Several criteria have evolved and find use in the book.

- *Half-Power Frequency*. Also called 3 dB down frequency; the 3 dB can be another useful value, such as 10 dB. However, because of the spectral antisymmetry Property 1.1 that applies to orthogonal pulses, the 3 dB version is particularly useful: All reasonable orthogonal pulses with the same symbol time T have the same 3 dB down frequency.
- *Power in Band (PIB)*. This refers to the fraction of signal power that lies in the band $[-f, f]$, $f > 0$, expressed usually in percent. The measure applies to baseband signals. For example, the 99% PIB frequency is f inside which lies 99% of the average signal power. The PIB measure is useful because f for 99% or 99.99%, depending on the situation, mark the bandwidths outside of which there will be little interference with neighboring channels. The terms power and energy are used interchangeably, but technically, power applies to an ongoing signal and energy to a single pulse. An alternate measure is the power *out* of band fraction (POB), which refers to the percent outside $[-f, f]$.
- The signal space distance between critical signals can be expressed as a function of frequency and a bandwidth measure derived from that. This method will be taken up in Section 2.5.2.

An analogous problem is measuring the diameter of the sun. The textbook value is 1392680 km, but actually the sun declines in density and has no edge. The size of a kilometer is not in doubt, nor is the Hz, but a density or other feature needs to be specified. In this book, it is most often convenient to use the half-power frequency, because for the important class of orthogonal pulses it has a fixed relation to the symbol time, and the other criteria do not differ much. In some applications and with some pulses, the 99% PIB is more important.

On a mathematical level, all these measures deal in some way with the fundamental problem that a signal or its spectrum or both must lack finite support, while physical signals and spectra clearly have finite support. In an important 1975 paper [11], Slepian discusses this puzzle and asserts that the total bandwidth $\mathcal{W}$ and time $\mathcal{T}$ that one allocates to a signal must always be approximate. Some small fractions ϵ_f and ϵ_t of the energy must lie outside the nominal $\mathcal{W}$ and $\mathcal{T}$. Fixing the fractions defines an occupied bandwidth and time. This is the POB idea.

Figure 1.2 illustrates the concept for a signal that is a sequence of the ten 10% root RC pulses with symbols $[1, -1, -1, -1, 1, -1, 1, -1, -1, 1]$ and $T = 1$. The actual time signal and spectrum are shown. The inner 10 s $\times$ 0.5 Hz box shows the nominal time–bandwidth occupancy of the signal, which is 5 Hz-s. The outer box shows the occupancy that occurs when only fractions $\epsilon_f = \epsilon_t = 0.01$ are allowed outside the

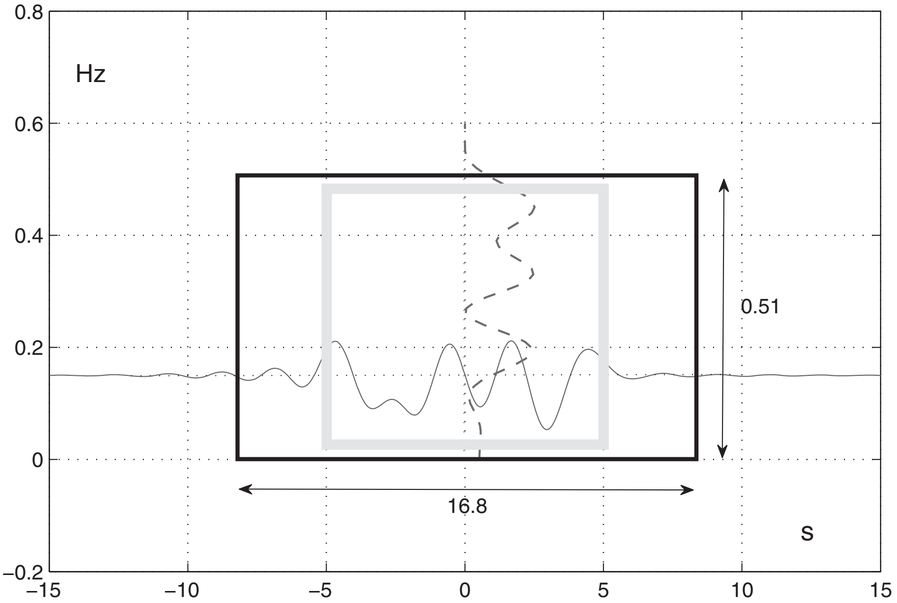

FIGURE 1.2 Time–bandwidth occupancy of a baseband transmission: Nominal (inner box) and actual when only 1% of the pulse energy is allowed out of band (outer box). Ten binary symbols and 10% root RC linear modulation; actual signal (solid line) and spectrum (dashed) are sketched in the boxes.

box.[3] It is clear that the time occupancy now extends far beyond the nominal 10 s, and the occupancy increases to 8.5 Hz-s. For such a short message, a pulse with a better trade-off of frequency and time is needed. There is a time–bandwidth optimization problem here: Given the POB fractions ϵ_t and ϵ_f and a 10-symbol message, what pulse shape minimizes the time–bandwidth occupancy?

We take up occupancy in Chapter 7. The chief conclusions are that the sinc pulse is not optimum and the best pulse is close to a prolate spheroidal wave function.

Signal Spectra. For a linear modulation signal of Eq. (1.1), every set of N bits produces a signal with its own spectrum. Some are wider band than others. They will interfere more with a neighboring channel, or if the channel in use has strict band limitation, they will be damaged more by the limitation. There is a worst case under a given band limitation; this view is taken up in Sections 2.5 and 7.2.2.

The standard view of modulation spectra is to let N grow large and compute the spectrum averaged over the symbol probability distribution. This is called the average power spectral density, denoted as PSD. This need not be the spectrum of

[3] The calculation is performed for one 10% pulse, not for the whole waveform; see Chapter 7.

any concrete signal, but it is easily computed and universally used. Much of its convenience stems from the following:

Theorem 1.1 (**Linear Modulation Spectrum**) *Suppose the same pulse $h(t)$ is used for all N symbols and the symbols u_n are IID with mean zero. Then the PSD is*

$$(1/T)\,\mathcal{E}[|u_1|^2]\,|H(f)|^2. \tag{1.7}$$

That is, the PSD of a linear modulation is the same as the spectrum of its pulse. The proof is simple and available in most digital communication texts ([1], p. 30). Note that the spectrum of the fixed signal in Figure 1.2 is not $|H(f)|^2$ because it is not an average.

The $h(t)$ is primarily a baseband pulse in this book, but the theorem holds if h is a carrier pulse. The IID/mean zero requirement is ordinarily true in practice: If the mean is nonzero, energy will be wasted sending a carrier component; if the symbols are not at least pseudorandom, synchronization will be damaged. When the modulation is coded it is a good assumption, universally applied in this book, that the code symbols are at least uncorrelated, so that the pulse spectrum carries over as well to the coded case.

When the symbols are correlated or the modulation is nonlinear, the PSD is more difficult to compute. Methods exist that are based on cyclostationary random process theory or a random time offset to the signal; the aim is to define a stationary random process with an autocorrelation, which has therefore a spectral density (see [4], p. 61ff). We need the calculation only for CPM coded modulation in Chapter 6.

With the PSD 3 dB bandwidth as a measure, an orthogonal M-ary modulator has time–bandwidth $T(1/2T) = 1/2$ Hz-s, regardless of T and M. The bit density is $2\log_2 M$ b/Hz-s.

The Sinc Pulse. This time pulse has appeared as the Nyquist limit to orthogonal pulses in Section 1.2, and it will appear again in Shannon's capacity calculation in Chapter 3. Controversy surrounds its use, which is worth comment. The sinc and its dual the square pulse are not physically realizable, but to the extent that they can be approximated, they find use. A true time sinc has an attractive spectrum but many disadvantages. Symbol timing cannot be obtained by ordinary means ([1], Section 4.7.2); if symbol timing is not perfect, the sum of the detector ISI is a divergent series. Discrete-time models for faster than Nyquist transmission are difficult to define, and the sinc is far from the solution to Slepian's problem. The heart of the problem is the side lobes in the time domain, which decay only as $1/t$.

The square time pulse has a sinc-shaped spectrum with the dual outcome, that the 1% POB frequency is very large, $\approx 9.5/T$ Hz.

Generally speaking, these outcomes are not acceptable. Sinc pulses are truncated in time, which raises spectral side lobes outside the nominal $[-1/2T, 1/2T]$ Hz bandwidth, and square pulses are truncated in frequency, which creates ISI. The smoother a pulse is in time (the larger root RC α it has) the easier it is to control these effects. Still, even practical smooth pulses can suffer. Truncating a 30% root

RC pulse to time $[-2.5T, 2.5T]$ throws 0.1% of its energy well outside the nominal $[-0.65/T, 0.65/T]$ bandwidth of the pulse ([4], p. 62). The percent may seem small but it is unacceptable interference in many systems.

Wide- and Narrowband Transmission Methods. Whether coded or not, transmission methods are classed as *wideband* if their bit density is less than 2 b/Hz-s. This is because they achieve their rate in b/Hz-s primarily by consuming bandwidth. Everyday examples are space communication and high-quality FM broadcasting. The value 2 b/Hz-s is the density of simple orthogonal pulse binary modulation, as will be shown in Section 2.1. Methods with higher bit density are *high energy*, because they depend chiefly on a high E_b/N_0. These generally need about 3 dB more energy per one-bit increase in density. Short-range wireless links are examples of high energy systems. Note that the ratio E_b/N_0 is what is high, not the bit energy E_b. The choice of wideband or high energy depends on the relative cost of each; there is nothing inherently wrong with either regime.

We will see this distinction in Chapter 2 for modulations, but it is also apparent in Shannon's capacity in Chapter 3.

1.4 CODING VERSUS MODULATION

Not all modulations produce simple pulses and the need to reduce bandwidth can lead to rather complicated signals. It can be a subtle exercise to distinguish coding from modulation, especially when bandwidth plays a role. Some controversy surrounds how to do this. To avoid paradoxes and false hopes, here is a discussion.

Through the development of coding, several concepts have arisen. Coding can be

(i) the imposition of signal patterns, such as a trellis structure or those imposed by memory;

(ii) the addition of redundancy, especially through parity check bits;

(iii) the expansion of a signaling alphabet, followed by a selection of words that represent the data, which has a smaller alphabet; and

(iv) selection of a set of some but not all of the possible modulator sequences.

Concept (i) is not suitable for us; we will see that modulators more band limited than orthogonal pulse schemes create trellis-structured signals and require trellis demodulation, even though they are not encoders. Concept (ii) is troublesome because a number of schemes we would like to call coding do not add redundant symbols; one can say that parity check symbols are an artifact of the binary symmetric channel, in which coding can happen no other way. Signal alphabet expansion (iii) is problematical for several reasons: The modulator sets the alphabet, not the coding; the alphabet of some channels, like the AWGN, is the whole real line. Is there a definition of coding wide enough to include all of the schemes we would like to consider?

In his Gaussian channel papers Shannon evolves toward concept (iv). His 1949 paper [7] that introduced Gaussian channel coding concentrates on philosophy, capacity, and his "$2WT$" theorem, which shows that signals over time $\mathcal{T}$ and bandwidth $\mathcal{W}$ span $2\mathcal{W}T$ orthogonal dimensions. This crucial result is discussed in Section 3.1. It makes possible the capacity theorem and the definition of a code. By 10 years later Shannon would define a code as follows:

> … a real number may be chosen at the transmitting point. This number is transmitted to the receiving point but is perturbed by an additive Gaussian noise, so that the ith real number, s_i, is received as $s_i + x_i$ … A *code word* of length n for such a channel is a sequence of numbers $(s_1, s_2, \ldots, s_n)$. This may be thought of geometrically as a point in n-dimensional Euclidean space. ([8], p. 611)

This very nearly captures concept (iv) as it will be implemented in this book. Modulator signals have an expression in terms of orthogonal basis functions, weighted by real numbers. Each number represents a "channel use." The receiver seeks the least-distant whole sequence in Euclidean space. Gallager [9] and Wozencraft and Jacobs [10] in their classic texts essentially concur.

One hesitates to second-guess these authorities, and in any case definition (iv) fits our needs. Earlier, in Section 1.2, the definition of modulator included the idea that *all* M-ary transmission symbols produce outputs. Building on this, we define a channel code as *a set of some but not all of the possible modulator sequences*. If there are N channel uses, the code has rate per use

$$R = (1/N)\log_2(\text{subset cardinality}) \qquad \text{b/channel use,} \tag{1.8}$$

which is R/T in bits/second with a modulator alone. Note that the modulator need not come first in the transmitter, so long as there is some way to connect codewords and modulator outputs.

The rate R must be less than $\log_2 M$. Shannon shows that there is another smaller rate called channel capacity, such that a set and a decoder exist that achieve arbitrarily small error probability as $N \to \infty$.

The codewords in the set can be selected in many ways. Shannon imagined that the letters were chosen at random, and this turned out to be a powerful idea. In Chapter 4, we borrow a convolutional encoder to make the choice; this can be viewed as a pseudorandom selection procedure, and just as Shannon predicted, it works very well. Alternately, words can also be specified on a trellis or graph structure, or as the solutions of equations, as they are in parity-check codes. However the set is selected, some sort of block length N is essential, and it must grow large if R is near capacity.

1.5 A TOUR OF THE BOOK

Chapter 2 introduces the communication theory needed for the book, with emphasis on the issues that play a special role. These include error events, calculation of minimum distance (which predicts the error rate), suitable receiver structures, the

BCJR algorithm (essential in iterative decoding), signal phase (which affects decoding complexity), and the performance of modulators, both simple ones and those producing complicated narrowband signals.

Chapter 3 introduces relevant AWGN-channel Shannon theory. The chapter derives the capacity of this channel and finds from it the Shannon limit to communication as a function of the density (rate) per Hz-s, the signal PSD shape, and the signal energy per data bit. Of these three, density plays the dominant role.

Modulators have associated with them the same three quantities, and an error probability is computed in terms of them in Section 2.1. The simple ones provide a benchmark for low-complexity transmission. A good rule of thumb is that their bit error rates as a function of E_b/N_0 lie about 10 dB from the respective capacity. The performance of coding schemes lies in between these two limits.

Chapter 4 introduces faster than Nyquist signaling (FTN), the most successful method of narrowband coding at present. The term has a 40-year history, and it originally meant an orthogonal pulse modulator with symbol time accelerated; the pulses were no longer orthogonal but the error performance was undiminished. Today FTN means that modulation pulses are nonorthogonal, for whatever reason. FTN methods can be coded, and Chapter 3 shows that they have a better Shannon limit. Chapter 4 explores many aspects that arise in this new technology, including simplified receivers, design of good codes, and error performance analysis.

Classical FTN signals were accelerated in time, but the idea extends to compression of subcarriers in frequency. The outcome occupies less bandwidth and is similar to orthogonal frequency division multiplex (OFDM), but with nonorthogonal subcarriers. Chapter 5 presents this idea and a number of variations. Since OFDM is a favored method in fourth-generation wireless telephony, this "non-O FDM" is attracting interest for fifth-generation systems.

Chapter 6 compares these new methods to older coded modulation methods. These older ideas have bandwidth consumption in between the new ones and binary error-correcting codes. Some FTN implementations in the literature are also reviewed, including chip designs.

Chapter 7 explores alternate ideas about the design of the modulation base pulse itself. One analysis, by Slepian, seeks the pulse with the least time and frequency occupancy. The outcome is related to the IOTA pulse, a popular pulse in OFDM. Another analysis seeks the pulse with the best modulator error performance for a given bandwidth.

1.6 CONCLUSIONS

Where does all this lead? The evidence in this book strongly supports certain conclusions:

Very narrow band energy-efficient transmission cannot occur without *both* (*i*) *complicated modulation–demodulation and* (*ii*) *significant decoding complexity*. These work together.

Nonorthogonal modulation pulses are necessary. Narrowband transmission is built upon narrowband pulses. Their response is much longer than the data symbol time and it leads to significant ISI.

To perform reasonably near the Shannon limit requires iterative decoding. No other method is available today.

REFERENCES

1. *J.B. Anderson, *Digital Transmission Engineering*, 2nd ed., Wiley–IEEE Press, Piscataway, NJ, 2005.
2. *M. Schwartz, *Information Transmission, Modulation, and Noise*, 4th ed., McGraw-Hill, New York, 1990.
3. *J.G. Proakis, *Digital Communication*, 4th and later eds., McGraw-Hill, New York, 1995.
4. J.B. Anderson and A. Svensson, *Coded Modulation Systems*, Kluwer-Plenum, New York, 2003.
5. H. Nyquist, Certain factors affecting telegraph speed, *Bell Syst. Tech. J.*, pp. 324–346, 1924.
6. H. Nyquist, Certain topics on telegraph transmission theory, *Trans. AIEE*, **47**, pp. 617–644, 1928.
7. C.E. Shannon, Communication in the presence of noise, *Proc. IRE*, **37**, pp. 10–21, 1949; reprinted in *Claude Elwood Shannon: Collected Papers*, Sloane and Wyner, eds, IEEE Press, New York, 1993.
8. Probability of error for optimal codes in a Gaussian channel, *Bell Syst. Tech. J.*, **38**, pp. 611–656, 1959; reprinted in *Claude Elwood Shannon*, *ibid.*, 1993.
9. R.G. Gallager, *Information Theory and Reliable Communication*, McGraw-Hill, New York, 1968.
10. *J.M. Wozencraft and I.M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.
11. D. Slepian, On bandwidth, *Proc. IEEE*, **64**, pp. 292–300, 1976.

*References marked with an asterisk are recommended as supplementary reading.