CHAPTER

23 Modulation, noise, and information

In this chapter we examine how noise degrades the accuracy of digital data transmission and the fidelity of analog transmission. We begin with an explanation of matched filtering, a subject which came up in Chapter 22. We show that, for binary data links using matched filtering and coherent detection, the probability of error depends only on the noise level at the input of the receiver and on the energy, but not the shape, of the transmitted pulses. We look at two example systems: BPSK (binary phase-shift keying) with coherent detection and OOK (on–off keying) with envelope detection. The error rates and channel capacities (maximum error-free data rates when using forward error correction coding) are calculated and compared with Shannon's expression for the capacity of a band-limited channel. Finally, traditional AM and FM are examined with respect to their noise characteristics.

23.1 Matched filtering

We stated in Chapter 22 that, in the presence of noise, the post-detection signal-to-noise ratio is maximized when the predetection bandpass shape of the receiver is that of a *matched filter*.

A matched filter is one whose impulse response is proportional to the time-reversed waveform of the incoming signal pulse, as will be shown below. For example, if the input signal is a pulse whose shape is the same as the symmetric impulse response of a root raised-cosine filter, then the receiver should use a root raised-cosine filter or an equivalent cascade of filters. Sometimes we deal with complicated pulses, such as the biphase coded pulses used in pulse compression radar. In these cases, it is common to use a front-end bandpass filter which is a matched filter for the individual subpulses, followed by coherent detection and then a decoder which undoes the plus/minus phase coding. Here the cascade of the front-end filter and the decoder is equivalent to a matched filter for the coded pulses. Note that coherent detection is down-conversion to baseband, a linear operation. We can think of

the entire matched filter as the actual detector, whose output samples are the received symbols.

Let D denote the output of the receiver's filter (which we have just defined as the "detector" output) and let $h(t)$ be the impulse response of this filter, a real function. For a signal pulse unaccompanied by noise, the output from the filter is given by

$$D_{\mathrm{S}} = \int_{-\infty}^{\infty} h(t) V_S(-t) \mathrm{d}t, \tag{23.1}$$

where we have assumed the pulse position to be such that the filter output should be sampled at $t = 0$. The output signal power, in units of volts^2, is given by $P_{\mathrm{S}} = D_{\mathrm{S}}{}^2$. The noise output of the filter can be written in terms of N_0, the standard one-sided noise density, in units of volts^2/Hz, as

$$P_{\mathrm{N}} = \frac{N_0}{2} \int_{-\infty}^{\infty} |H(\omega)|^2 d\omega, \tag{23.2}$$

where $H(\omega)$ is the filter transfer function, i.e., the Fourier transform of $h(t)$. Note that we have assumed that N_0 is a constant (white noise). Using Parseval's theorem, we can write P_{N} in terms of $h(t)$:

$$P_{\mathrm{N}} = \frac{N_0}{2} \int_{-\infty}^{\infty} h(t)^2 dt. \tag{23.3}$$

With these expressions for P_{S} and P_{N}, we can write the output signal-to-noise ratio as

$$\mathrm{SNR} = \frac{P_{\mathrm{S}}}{P_{\mathrm{N}}} = \frac{\left(\int_{-\infty}^{\infty} h(t) V_S(-t) \mathrm{d}t \right)^2}{\frac{N_0}{2} \int_{-\infty}^{\infty} h(t)^2 \mathrm{d}t}. \tag{23.4}$$

At this point, we invoke Schwarz's inequality[1] which gives us

$$\mathrm{SNR} \leq \frac{\int_{-\infty}^{\infty} h(t)^2 \mathrm{d}t \int_{-\infty}^{\infty} V_{\mathrm{S}}(-t)^2 \mathrm{d}t}{\frac{N_0}{2} \int_{-\infty}^{\infty} h(t)^2 \mathrm{d}t} = \frac{1}{N_0} \int_{-\infty}^{\infty} V_{\mathrm{S}}(-t)^2 \mathrm{d}t = \frac{1}{\frac{N_0}{2}} \int_{-\infty}^{\infty} V_{\mathrm{S}}(t)^2 \mathrm{d}t. \tag{23.5}$$

[1] Schwarz's inequality is written as $|\int f(x) g(x) \mathrm{d}x|^2 \leq \int |f(x)|^2 \mathrm{d}x \int |g(x)|^2 \mathrm{d}x$. This can be seen by considering the integral of fg to be the dot product of a multidimensional vector, while the integral $\int |f(x)|^2 \mathrm{d}x$ is the length of the vector f and $\int |g(x)|^2 \mathrm{d}x$ is the length of the vector g. For two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$, we know that $|\boldsymbol{A} \bullet \boldsymbol{B}|^2 \leq |\boldsymbol{A} \bullet \boldsymbol{A}||\boldsymbol{B} \bullet \boldsymbol{B}|$.

Looking at Equation (23.4), we see that if $h(t) = \beta V_S(-t)$, where β is any constant, then the SNR will be equal to its maximum possible value, the right-hand expression in Equation (23.5). This is the matched filter. Note that, with a matched filter, the signal-to-noise ratio for a pulse is independent of the pulse shape and is simply the time integral of $V_S(t)^2$, i.e., the energy of the pulse, $\mathscr{E}_S$, divided by the noise spectral density N_0. (N_0 has units of volts^2/Hz, which is the same as volts^2 sec.) Thus, when a matched filter is used,

$$\mathrm{SNR} = \frac{1}{N_0/2}\int_{-\infty}^{\infty} V_S(t)^2 \mathrm{d}t = \frac{2\mathscr{E}_s}{N_0}. \tag{23.6}$$

23.2 Analysis of a BPSK link

Coherent detection makes the BPSK link especially easy to analyze. The signal at the output of the matched filter is just the sum of the signal voltage and the noise voltage, which we will assume to have a Gaussian distribution, characteristic of thermal noise. The detected signal voltage will be either A or $-A$, so the Gaussian noise distribution will be centered at either A or $-A$, as shown in Figure 23.1. The one/zero decision threshold will, of course, be set at $V = 0$ for this symmetric situation.

The probability distribution for a zero (the solid curve) is the normal Gaussian distribution function, $p_0 = (2\pi\sigma^2)^{-1/2} \exp[-(V+A)^2/(2\sigma^2)]$. By inspection of the figure we can write

$$p_e = \int_0^{\infty} \frac{e^{-(V+A)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dU = \int_A^{\infty} \frac{e^{-V^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \mathrm{d}U = \frac{1}{\sqrt{2\pi}} \int_{\frac{A}{\sigma}}^{\infty} e^{-u^2/2} \mathrm{d}U. \tag{23.7}$$

Now let us relate this probability of error to the signal-to-noise ratio. We saw in the previous section that, at the output of a matched, the ratio of the square of the signal portion of the sampled output to the average square of the noise portion is $2\mathscr{E}_S/N_0$, where $\mathscr{E}_S$ is the energy of the pulse, i.e., the time integral of the V_S^2, and

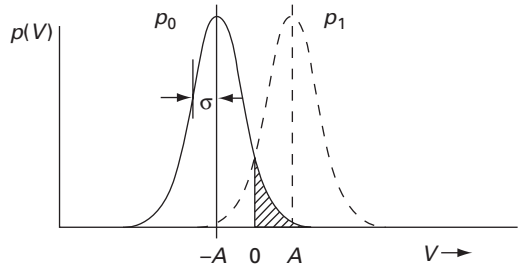

Figure 23.1. Gaussian voltage probability distributions for received BPSK ones and zeros for $A = 1.5\sigma$ (SNR = 1.5^2). The shaded area in the figure is the probability p_e of a transmission error, i.e., the probability that a received zero will be interpreted as a one or that a received one will be interpreted as a zero.

N_0 is the noise power density. Therefore $A^2/\sigma^2 = 2\mathscr{E}_S/N_0$, and we can rewrite Equation (23.7) as

$$p_e = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2\,\mathscr{E}_S/N_0}}^{\infty} e^{\frac{-u^2}{2}} du. \tag{23.8}$$

We have been implicitly working at baseband, i.e., assuming that the signal (and the noise) have been converted down from an RF carrier frequency, ω_0, through multiplying the modulated RF signal by a sine wave in phase with the (normally suppressed) carrier. This, as we have seen, constitutes synchronous detection. In this common situation, the matched filter is a baseband filter.

It is interesting to note that we could alternately have used a matched filter at the RF (or some IF) frequency. In this case, the output of the filter is an RF sine wave, multiplied by the baseband pulse. We can sample this directly, but the sampling must be very precise, so that the samples are taken at points that, with a dc pulse, would be at the peaks of the RF sine wave. The SNR will be the same as for the synchronous conversion to baseband because even though the RF bandwidth is twice the baseband bandwidth and contains twice as much total noise power, the synchronous sampling will respond to only half this noise, e.g., the "cosine" component.

If symbols are arriving at the rate $1/T$ in the minimum baseband bandwidth of $1/(2T)$ that eliminates intersymbol interference (see Chapter 22), then $2\mathscr{E}_S/N_0 = (\mathscr{E}_S/T)/(N_0/[2T]) = P_S/P_N = SNR$ the signal-to-noise ratio. This probability of error from Equation 23.8 is plotted vs. SNR in Figure 23.2.

This figure displays the well-known "dropout" effect in digital communications systems. If, for example, the input SNR drops from ten to five, a factor of only 2, the error rate increases by a factor of 100. The probability of error for the BPSK distributions shown in Figure 23.1, where $A/\sigma = 1.5$, is 0.067.

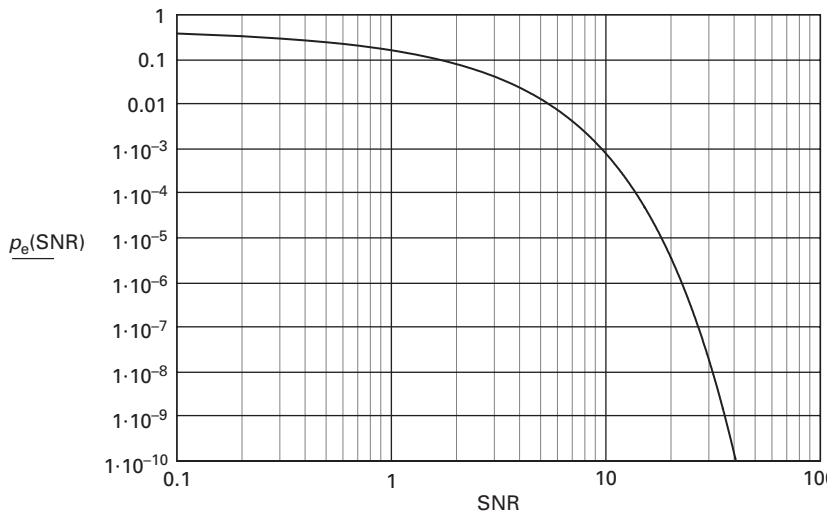

Figure 23.2 Envelope probability distributions $P(E, A, \sigma)$, *for* $\sigma = \sqrt{2}$ and $A = 0$ *or* $1.5\sqrt{2}$.

23.3 On–off keying with envelope detection

The BPSK system discussed above, using coherent detection, is suited for data transmission under low signal conditions. (When there is ample signal power, multilevel "M-ary" modulation can obviously transmit data at a faster bit rate.) But let us now consider simple on–off keying with envelope, i.e., noncoherent detection. Before we can calculate the probability of error, we must find the distribution function of the output voltage from the envelope detector.

23.3.1 Envelope probability distributions

As we have seen, the IF signal, containing random noise, is a random variable with Gaussian probability distribution. The presence of a signal offsets the Gaussian curve. The output of the envelope detector is also a random variable. Let us find the probability distribution of the envelope, first without the presence of a signal. Figure 23.3(a) shows the in-phase and quadrature components, $\boldsymbol{N}_I$ and $\boldsymbol{N}_Q$. of the IF noise voltage. The length of their vector sum is $\boldsymbol{E}$, the envelope voltage.

Since the I and Q noise components are uncorrelated, their joint probability distribution is the product of their individual Gaussian distributions.

$$p(V_I, V_Q) = \frac{\mathrm{e}^{-V_I^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \times \frac{\mathrm{e}^{-V_Q^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} = \frac{\mathrm{e}^{-(V_I^2+V_Q^2)/2\sigma^2}}{2\pi\sigma^2} = \frac{\mathrm{e}^{-E^2/2\sigma^2}}{2\pi\sigma^2}. \tag{23.9}$$

To obtain the distribution in terms of E and θ, we note that

$$p(E,\theta)E\mathrm{d}\theta\mathrm{d}E = \frac{\mathrm{e}^{-E^2/2\sigma^2}}{2\pi\sigma^2} E\mathrm{d}\theta\mathrm{d}E, \tag{23.10}$$

from which we identify

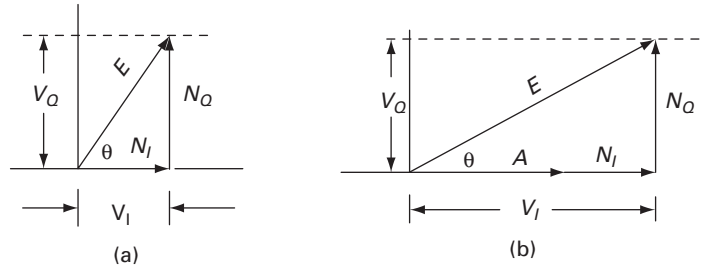

Figure 23.3 (a) *I* and *Q* noise phasors alone; (b) noise phasors together with a sine wave of amplitude *A*.

$$p(E,\theta) = \frac{E\mathrm{e}^{-E^2/2\sigma^2}}{2\pi\sigma^2}. \tag{23.11}$$

Integrating over θ, from 0 to 2π, we find $p(E)$, the envelope probability distribution:

$$p(E) = \int_0^{2\pi} P(E,\theta)\mathrm{d}\theta = \frac{E\mathrm{e}^{-E^2/2\sigma^2}}{\sigma^2}. \tag{23.12}$$

This function is known as the Rayleigh probability distribution. Note: for both the I and the Q noise components, the variance is σ^2. Therefore, the total noise power, which is the expectation of E^2, is $2\sigma^2$.

Now we must find the envelope distribution function when the IF signal is the superposition of noise plus a sinusoidal carrier. This is shown in Figure 23.3(b). Again, both noise voltages, N_I and N_Q, have Gaussian distributions, but now the distribution for N_I is centered at A, the amplitude of the carrier, so we can write

$$\begin{aligned} p(V_I,V_Q) &= \left[\sqrt{2\pi\sigma^2}\right]^{-2}\mathrm{e}^{[-(V_I-A)^2+V_Q^2]/2\sigma^2} \\ &= (2\pi\sigma^2)^{-1}\mathrm{e}^{[-(E\cos(\theta)-A)^2+E\sin(\theta)^2]/2\sigma^2}. \end{aligned} \tag{23.13}$$

Expanding the argument of the exponential on the right-hand side we have

$$p(V_I,V_Q) = (2\pi\sigma^2)^{-1}\mathrm{e}^{-(E^2+A^2)/2\sigma^2}\mathrm{e}^{-2AE\cos(\theta)/2\sigma^2}. \tag{23.14}$$

Following the steps used in the carrier-free case, and integrating over θ, we find

$$\begin{aligned} p(E,A,\sigma) &= \frac{E\mathrm{e}^{-(E^2+A^2)/2\sigma^2}}{\sigma^2}\frac{1}{2\pi}\int_0^{2\pi}\mathrm{e}^{-AE\cos(\theta)/\sigma^2}\mathrm{d}\theta \\ &= \frac{E\mathrm{e}^{-(E^2+A^2)/2\sigma^2}}{\sigma^2}I_0(AE/\sigma^2), \end{aligned} \tag{23.15}$$

where I_0 is the modified Bessel function of order zero. This function, $p(E,A,\sigma)$, known as the *Rician* distribution, is plotted in Figure 23.4 for $A = 0$ (the envelope distribution of noise alone) and for $A = 1.5\sqrt{2}$, for the same noise power and average transmitted power (assuming equal probabilities for transmitted ones and zeros) used for the BPSK example of Figure 23.1.

Let us calculate the probability p_e that a bit is received incorrectly when we have (arbitrarily) set the decision threshold at the intersection of the ON and OFF envelope distribution functions, which we will denote as E_t. We will also assume that we transmit, on average, equal numbers of ones and zeros. The probability of error is, therefore, $\frac{1}{2}p_{\mathrm{e}0} + \frac{1}{2}p_{\mathrm{e}1}$, where $p_{\mathrm{e}0}$ is the probability that a transmitted zero is received incorrectly and $p_{\mathrm{e}1}$ is the probability that a transmitted one is received incorrectly or

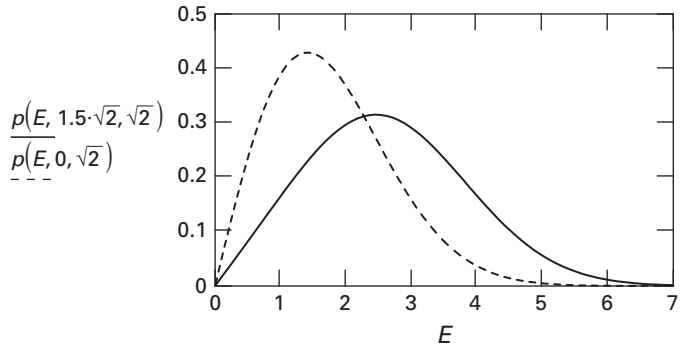

Figure 23.4 Probability distribution $p(E,A,\sigma)$, for the envelope of a signal that is the sum of a sine wave of amplitude $A = 1.5\sqrt{2}$ plus noise with $\sigma = 2$.

$$p_{\mathrm{eOOK}}(A,\sigma) = 1/2 \int_{E_t}^{\infty} p_0(E,A,\sigma)\mathrm{d}E + 1/2 \int_{0}^{E_t} p_1(E,A,\sigma)\mathrm{d}E. \qquad (23.16)$$

Evaluating P_{eOOK} for A $= 1.5\sqrt{2}$ and $\sigma = \sqrt{2}$ for comparison with the BPSK example of Figure 23.1 (same average transmitted power, but twice the noise power (quadrature as well as in-phase components), yields $P_{\mathrm{eOOK}} = 0.34$ for a threshold $E_{\mathrm{t}} = 2.3$, compared to only .067 for P_{eBPSK}.

Channel capacity

We have seen how to calculate the expected bit error rate as a function of the signal and noise powers for two example situations, coherently-detected PSK and envelope-detected on–off keying. We also know that data error rates can be reduced by expanding (encoding) the data so that it contains redundancies.[2] Of course, the net data rate slows when redundant bits are transmitted. These redundancies may be crude, such as transmitting a packet of data several times, or elegant, such as the nested combination of block codes and convolutional codes used in data links and digital broadcasting systems. Given a particular communications link with an optimal encoding scheme, it is obvious to ask how many bits must be transmitted, on average, for each data bit, if we are willing to tolerate a certain average data error rate. One might expect that, as the error tolerance is reduced to zero, the optimal transmission efficiency would also go to zero, i.e., that an infinite number of bits would have to be transmitted for each recovered data bit, as indicated by the crossed out curve in Figure 23.5. Shannon's remarkable "noisy channel theorem" showed that this is not so. For any given communications link, there must exist coding schemes that will permit data transmission with an arbitrarily low error tolerance while still

[2] Making the data redundant enough so that errors can be detected and corrected at the receiver is known as forward error correction (FEC). In reverse error correction, the receiver can only detect errors. To make corrections, it must request repeat transmissions.

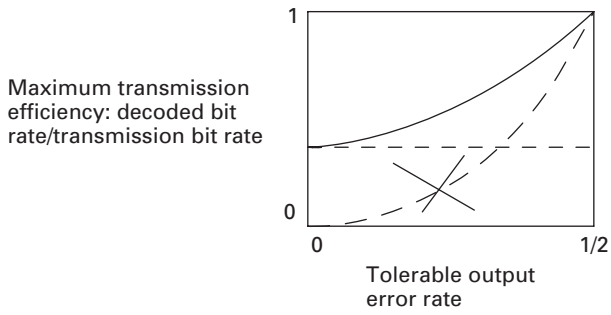

Figure 23.5 Transmission efficiency vs. tolerated output error rate when using optimum channel encoding.

achieving a non-zero efficiency. This situation is shown by the solid curve Figure 23.5, which shows the efficiency for a link using an optimum error coding scheme.

The maximum efficiency is unity if we are willing to tolerate an output error rate of 1/2, but this is the point at which no net information is transmitted since a one is randomly declared a one or a zero and vice versa. But as the error rate tolerance is lowered to zero, the maximum efficiency approaches an asymptotic *non*zero value. This asymptotic maximum efficiency value is called the *channel capacity* (bits/bit) and, when multiplied by the bit rate, in bits/sec, the result, in units of bits/sec, is a rate and is also referred to as the channel capacity. (Context usually resolves any confusion between the two.) The ususal statement of Shannon's theorem is that it is possible to transmit data at a rate equal to or less than the channel capacity, with an arbitrarily low data error rate. While Shannon's work shows that optimum codes must exist, it does not show how to construct then. However, it does show how to calculate the channel capacity for a given link, the highest standard against which we can judge the efficiency of any coding scheme proposed for the link. If the arbitrarily low data error rate is set too low, the encoded data must be in the form of extremely long blocks. Fortunately, acceptably low error rates can be achieved with acceptably short data blocks.

Binary symmetric channel

The coherent BPSK link discussed above is an example of a *binary symmetric channel*. The transmitted symbols are either one or zero, and every received signal is deemed either a one or a zero. From the symmetry of Figure 23.1, it is obvious that the probability p_e that a transmitted one will be received in error as a zero is equal to the probability that a transmitted zero will be received in error as a one, as indicated in Figure 23.6. This figure, a sort of probability flow graph for a binary channel, is equivalent to a 2×2 matrix whose coefficients are probabilities. The binary symmetric channel corresponds to a symmetric channel matrix.

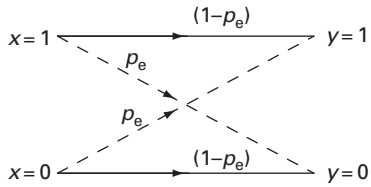

Figure 23.6. Binary symmetric channel: p_e is the probability that a symbol x_i is wrongly received as a symbol y_j.

The channel capacity for this link, which depends only on p_e and on p_i, the probability that the i-th transmitted bit is a one, is given by

$$C(p_e, p_i) = p_e \log_2(p_e) + (1 - p_e) \log_2(1 - p_e) \\ -p_i \log_2(p_i) - (1 - p_i) \log_2(1 - p_i). \tag{23.17}$$

This expression is always greatest for $p_i = 1/2$ (equal probability of transmitted ones and zeros). Figure 23.7 shows C(p_e,1/2) plotted versus p_e.

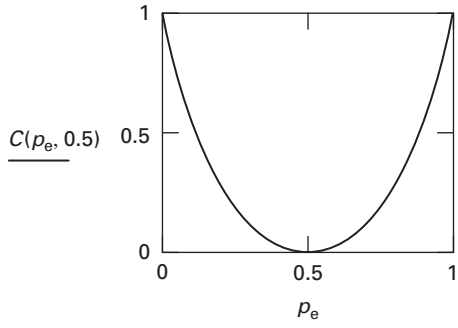

Figure 23.7 Channel capacity of a binary symmetric channel vs. transmission error rate, assuming ones and zeros are transmitted with equal probability.

Note that the channel capacity falls to zero at $p_e = 1/2$, where the receiver is equally likely to interpret a transmitted one as a zero and vice versa. Then C rises again to unity for $p_e = 1$, where the receiver always mistakes a one for a zero and zero for a one. In this case, the message is transmitted faithfully, assuming the receiver realizes that it should complement the bits. (It takes only a single transmitted bit to resolve this ambiguity.)

The reader will find excellent treatments of information theory, channel capacity, and coding in the texts listed at the end of this chapter but the origin of Equation (23.17) warrants at least a brief discussion, in the limited context of a binary channel. The discussion should at least serve to show that, while the concepts are subtle, the mathematics is simple. In the binary channel, the transmitter sends a symbol which is denoted either x_1 (a one) or x_2 (a zero). The *information* in a transmitted symbol is defined as $I(x_i) = \log_2 (p(x_i)^{-1})$, where $p(x_i)$ is the probability of x_i. Note that if $p(x_i)$ is nearly one, the information contained in x_i is nearly zero. If the transmitted symbols are almost always ones (or zeros), very little information is transmitted by sending a one (or zero), as we would have expected the one (or the zero). But if $p(x_i)$ is a very small number, the occurrence of x_i is highly informative, as we would not have expected it. A second definition is that of *mutual information*,

$I(x_i; y_j) = \log_2[p(x_i|y_j)|p(x_i)]$, where $p(x_i|y_j)$ is the conditional probability that x_i was transmitted, given that y_j was received. Note that while the distribution of transmitted symbols is generally flat (uniform), the distribution of $I(x_i; y_j)$ for a given j is not uniform, but peaked around the value of i most likely to have resulted in y_j.

The expected value of the mutual information is the sum of the $I(x_i; y_j)$'s, weighted by their probabilities, i.e.,

$$\langle I(x_i; y_j)\rangle = \sum_i \sum_j p(x_i, y_j) I(x_i; y_j), \tag{23.18}$$

where $p(x_i, y_j)$ is the joint probability of x_i and y_j, i.e., $p(x_i\, y_j) = p(x_i)\, p(y_j|x_i)$. For the binary symmetric channel discussed above, $p(y_j, x_i) = p_e$ for $i \neq j$ and $(1 - p_e)$ for $i = j$. The channel capacity is given by the maximum value of $\langle I(x_i; y_j)\rangle$ with respect to the set of values $p(x_i)$, by

$$C = \max\langle I(x_i; y_j)\rangle \text{ w.r.t } \{p(x_i)\}. \tag{23.19}$$

For the binary symmetric channel, the maximum occurs when equal numbers of ones and zeros are transmitted, i.e., when $p(x_1) = p(x_2) = 1/2$. Evaluation of Equation (23.19) yields Equation (23.17), the expression for the channel capacity. Note that the channel capacity is based on the relative frequency of transmitted ones and zeros and on the values of $p(y_j \mid x_i)$, which is the conditional probability that the receiver produces y_j when the transmitter sent x_i. However, to calculate the values of $I(x_i; y_j)$, we need to know $p(x_i \mid y_j)$, the conditional probability that when the receiver produces y_j, the transmitter had sent x_i. The two conditional probabilities are related through Bayes' theorem: $p(x_i, y_j) = p(x_i)\, p(y_j \mid x_i) = p(y_j)\, p(x_i \mid y_j)$.

Channel capacity of the BPSK and OOK example channels

For the BPSK example shown in Figure 23.1, with $A/\sigma = 1.5$, we found the probability of error to be $p_e = .067$. Using this value in Equation 23.17, the channel capacity is 0.645 data bits/transmitted bit. For the comparable asymmetric OOK channel of Figure 23.4 the two integrals in Equation 23.16 are, respectively, the probability that a transmitted zero is received as a one and vice versa. If we calculate the channel capacity using Equation 23.19, we find that it maximizes at about C = 0.08, if the threshold is set at $E_t = 2.5$ and the proportion of transmitted zeros is 52%. In addition to low channel capacity, another disadvantage of the OOK system is that the threshold must be changed when either the signal or noise level changes.

Channel capacity of a bandpass channel

Shannon also presented a formula for the *maximum* channel capacity for a band-limited channel with added Gaussian white noise – the channel most amenable to analysis, and a channel often encountered in practice, for example, in space-craft telemetry links. This formula states that the maximum channel capacity is given by

$$C = B\ \log_2[1 + S/N] = B\ \log_2[1 + S/(BN_0)] \qquad (23.20)$$

where S is the received signal power, N is the noise power which is equal (for white noise) to BN_0, where B is the channel bandwidth, and N_0 is the noise spectral density at the receiver in Watts/Hz. The figure of merit for any particular modulation scheme is how close its channel capacity approaches this ideal maximum channel capacity. To see that Equation 23.20 is reasonable note that, when S/N is high, it is essentially the practical number of quantization levels, n, and $\log_2(n)$ converts this into bits. The bandwidth factor, B, will be the symbol rate or close to it.

For illustration, suppose the signal power available from a spacecraft is 10^{-16} Watts, contained within a bandwidth of 100Hz, and the noise density at the receiver input is 4×10^{-19} Watts/Hz. With these numbers, Equation 23.20 produces $C = 181$, which is the maximum rate (bits/sec) at which information could be transmitted over the channel at an arbitrarily small error rate, if the modulation scheme and signal coding are optimum.

Let us look at our example BPSK channel in the light of Equation 23.20. In that example, the SNR at the output of the detector was $(A/\sigma)^2 = 1.5^2$. If we are using a matched filter, we saw that $(A/\sigma)^2$ be equal to $2\mathscr{E}/N_0$. Let us assume we transmit bits at a rate of $1/T$ and use receiver bandwidth of $1/(2T)$. The received power is $S = \mathscr{E}/T$ and the noise at the receiver is $N_0/(2T)$, from which we have $S/N = 2\mathscr{E}/N_0$. Using Equation 23.20, we see that the maximum channel capacity will be $C = 1/(2T)^{-1}\ \log_2(1 + 2\mathscr{E}/N_0) = .850/T$ bits/sec. But, from the bit error probability, a function of A/σ, we had calculated the actual channel capacity of this BPSK link to be .645 bits/bit. If we multiply this by the bit rate, 1/T, the channel capacity becomes $.645/T$ bits/second. This BPSK link, therefore, has .645/.850 = 76% of the maximum possible channel capacity.

For much larger values of S/N, we would find that, while the ideal channel capacity increases, our BPSK channel capacity saturates at the value $1/T$. This is simply due to staying with two-level binary modulation, as opposed to N-level PAM modulation.

Noise in analog FM and AM systems

Let us first look at the noise produced at the output of an FM receiver. The instantaneous received signal voltage, V_{SIG}, is accompanied by noise, V_N, as shown in Figure 23.8.

This is the same as the diagram in Figure 23.3(b), except that now we are interested in the phase angle rather than the magnitude of the vector sum of the signal plus noise. This "phase noise," ϕ_N, will cause noise in the detected output of an FM or PM detector. Clearly the angle ϕ_N becomes smaller if the signal strength is increased. But there is another way to defeat the noise. The signal-to-noise ratio at the detector output depends on the ratio of the signal's modulation phase excursions to the phase noise. If the modulation level is increased, even without increasing the signal strength, the output signal-to-noise ratio will be improved.

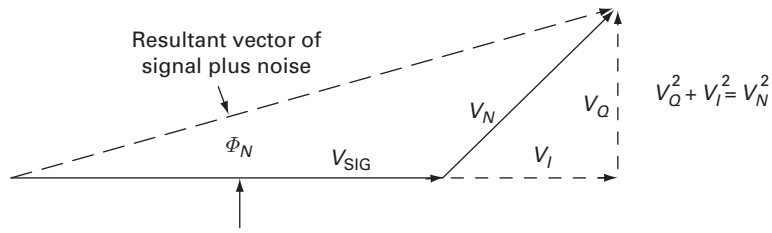

Figure 23.8. Signal and noise voltages.

If, for example, the rms phase noise is 1/10 radian and the modulation index is 1 radian, the phase SNR is 100. If the deviation is increased to produce a modulation index of 5 radians, the phase SNR increases to 2500. The improvement has been obtained not by increasing the signal power but by increasing the signal bandwidth. In the case of amplitude modulation, the SNR depends on the noise modulating the length of the vector $V_{SIG} + V_N$. Since V_N is fixed in any given situation, the only way to improve the SNR in AM is to increase V_{SIG}, i.e., increase the transmitted power.

Analysis of the SNR improvement in FM

A quantitative analysis of the SNR improvement in FM is simpler if we take the noise to be the background noise produced by the detector when the signal is unmodulated, i.e., the total power in the hiss coming from the loudspeaker or other output device. For the signal we will take an audio sine wave with 100% modulation (maximum deviation). We can represent the noise, V_N, by in-phase (I) and quadrature (Q) noise components, V_I and V_Q where $V_I^2 + V_Q^2 = V_N^2$ as shown in Figure 23.8. Both V_Q, and V_I are phasors rotating at ω_0, the frequency of the unmodulated carrier. Their amplitudes are random and independent. The I component of the noise is the most effective in causing amplitude fluctuations and therefore contributes noise in AM demodulation. But it is mostly the Q component, since it is perpendicular to V_{SIG}, that causes angle fluctuations and therefore contributes noise in the FM demodulation. For $V_N \ll V_{SIG}$, the instantaneous angle noise, $\phi_N(t)$, is just $V_Q(t)$ /V_{SIG} radians. Since V_{SIG}, the carrier amplitude, is constant, the power spectrum of ϕ_n (call it S_ϕ) is proportional to the power spectrum of V_Q. The spectral distribution of V_Q can be assumed uniform (white) so S_ϕ is also uniform. The integral of S_ϕ over the IF band gives the mean square phase fluctuation, $\langle(\phi_n(t))^2\rangle$, so we can write $S_\phi = \langle(\phi_n(t))^2\rangle/B_{IF} = \langle V^2_Q(t)\rangle/(V^2_{SIG}B_{IF}) = \langle V^2_N(t)/2\rangle/(V^2_{SIG}\, B_{IF})$ where B_{IF}, the IF bandwidth, is twice the maximum deviation. An FM demodulator produces an output spectral density proportional to the time derivative of the phase, $\langle(\mathrm{d}\phi_n(\mathrm{t})/\mathrm{d}t)^2\rangle$. The spectral density of the noise in the (one-sided) audio band at the detector output is therefore given by $2\omega_a{}^2 S_\phi(t)$, and the total noise power is the integral of this spectral density over the output bandwidth of the detector, i.e., the audio band (0 to B_a radians):

$$\text{Output noise power} = \int_0^{B_a} 2\omega^2 S_\phi \mathrm{d}\omega = S_\phi \int_0^{B_a} 2\omega^2 \mathrm{d}\omega = \frac{2}{3} S_\phi {B_a}^3 = \frac{{B_a}^3}{6 k_{osc} A_{MAX}} \frac{\langle {V_N}^2 \rangle}{V_{SIG}^2}. \quad (23.21)$$

The maximum amplitude of the sine-wave signal at the output of the detector is $k_{osc}A_{MAX}$, the maximum deviation, so the maximum signal power is just $P_{SIG} = {k_{osc}}^2 {A_{MAX}}^2/2$. Taking the noise power from Equation (23.21), the signal-to-noise ratio at the detector output is:

$$\text{Maximum output SNR} = 3\left(\frac{k_{osc}A_{MAX}}{B_a}\right)^3 \frac{V_{SIG}^2}{\langle {V_N}^2 \rangle}. \quad (23.22)$$

Note that the output SNR improves as the cube of the ratio of the maximum deviation to the full audio bandwidth.

Output signal-to-noise ratio for an AM signal with the same carrier power

Let us consider an AM system with the same carrier power and the same audio bandwidth, in order to compare its output SNR to that of the FM system. Again the modulation will be a single sine-wave tone and the amplitude of the carrier will be V_{SIG}. At 100% modulation, the amplitude of the sine wave modulation envelope will also be V_{SIG} so the audio signal power at the AM detector output will be ${V_{SIG}}^2/2$. For the AM system, the IF bandwidth needs to be only $2B$, twice the audio bandwidth (wide enough to accommodate the highest audio frequency but, to minimize noise, no wider). The noise voltage at the detector output will be V'_I, the in-phase component of V'_N, where the primes distinguish the noise voltages in the AM IF band from the noise voltages in the wider FM IF band. The noise power will be given by $\langle {V'_I}^2 \rangle = \langle {V'_N}^2 \rangle/2$. The SNR at the AM detector output is therefore

$$\text{AM output SNR} = \frac{{V_{SIG}}^2}{\langle {V'_N}^2 \rangle}. \quad (23.23)$$

Comparison of noise, FM vs. AM under strong signal conditions

All that remains in order to compare the FM and AM systems is to note that the ratio of the IF noise powers is just the ratio of the IF bandwidths, i.e.,

$$\frac{\langle {V_N}^2 \rangle}{\langle {V'_N}^2 \rangle} = \frac{2kA_{max}}{2B_a} = \frac{kA_{max}}{B_a}. \quad (23.24)$$

Using Equations (23.22)–(23.24), we can write

$$\frac{\text{FM SNR}}{\text{AM SNR}} = \frac{3{k_{osc}}^2 {A_{MAX}}^2}{{B_a}^2} = 3(\text{Deviation ratio})^2. \quad (23.25)$$

In FM broadcasting the deviation ratio is about five so the output SNR for a maximum-amplitude audio tone will be higher with FM than with AM by a factor of $3 \cdot 5^2 = 75$ or about 19 dB. The FM sound in television uses a deviation ratio only 1/3 as large as that of FM broadcasting so the signal-to-noise improvement is correspondingly less. Remember that the above analysis is only for the situations of high signal-to-noise. When the signal is comparable to or lower than the noise, the phase is almost totally determined by the noise and the FM system is useless.

FM, AM and channel capacity

The improvement in a signal-to-noise ratio that is possible with wideband FM is an example of increasing the channel capacity of a communications channel by increasing the bandwidth.

Note that S/N_O has units of bandwidth and can be considered a "natural" bandwidth for a given N_0 and S. N_0 is determined by the noise added along the channel, such as atmospheric noise and noise added by the receiver. The signal power, S, is determined by the transmitter power, transmitter and receiver antenna gains, and propagation loss. Channel capacity, from Equation (23.20), vs. bandwidth is plotted in Figure 28.9. Both are normalized to S/N_0.

Note that, for $b > 1$, the channel capacity has essentially reached an asymptotic value of 1.44 S/N_0. If we are below the knee of the curve, we can increase the channel capacity significantly at no cost in transmitter power by (somehow) using more bandwidth. We have seen that FM broadcasting does just this. On the other hand, Equation (23.20) shows that it is expensive (and ultimately impractical) to increase channel capacity by increasing power since the log term increases slowly.

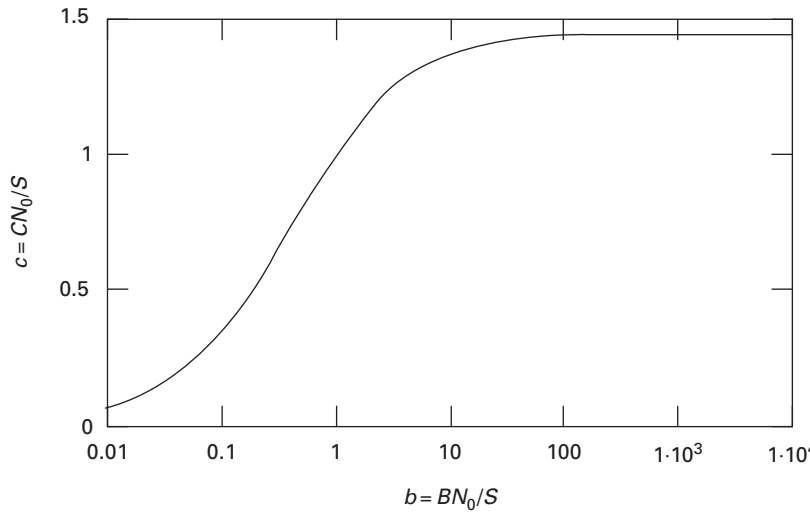

Figure 23.9. Channel capacity vs. bandwidth (both are normalized to S/N_0, the "natural bandwidth").

In AM, the bandwidth is fixed by the highest modulation frequency; the total bandwidth in standard full-carrier DSB AM is twice the highest modulation frequency. When a weak AM station is received, the signal-to-noise ratio is low enough to put us beyond the knee of the channel capacity vs. bandwidth curve and there would be no gain in going to a modulation scheme that increases the bandwidth. But if the station is strong, we are probably far below the knee of the curve. In this case, changing to FM modulation (without changing transmitter power) can bring us up the curve where the higher channel capacity allows a higher signal-to-noise ratio.

Problems

Problem 23.1. Show that the differential equation $x^2y'' + xy' - x^2y = 0$ (modified Bessel's equation of order zero) is satisfied by the function

$$I_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{x\cos(\theta)}\mathrm{d}\theta.$$

Problem 23.2. Consider an asymmetric binary channel in which the probability that a one is correctly received is 0.9 but the probability that a zero is correctly received is 1/2. (A transmitted zero is equally likely to be declared a zero or a one.) Suggest a simple coding method to transmit data reliably through this channel.

Problem 23.3. Consider the situation where a rectangular pulse of length T is transmitted, but the receiver instead of using a $\sin(1/2\ \omega T)/(1/2\ \omega T)$ matched filter, uses a rectangular filter of bandwidth β/T. The noise at the input is Gaussian. Find the value of β that maximizes the SNR at the output of the filter. Compare this to the value that would have been obtained if the matched filter had been used.

References

[1] Lahti, B. P., *Modern Digital and Analog Communication Systems*, 3rd edn, Oxford: Oxford University Press, 1998.
[2] Proakis, J. B., *Digital Communications* – 4th edn, New York: McGraw-Hill, 2000.