Frequency Shift Keying (FSK)

Introduction

In previous lectures we have studied analog modulation. We now turn to digital modulation. The trend in wireless system is overwhelmingly towards digital modulation for a variety of reasons. An obvious reason is that if you want to transmit digital information (e.g., wireless internet) you need to employ digital modulation. However, even for analog signals like voice, digital modulation is very attractive because in the digital domain you can employ *coding* for the purposes of data compression, encryption, and error correction. This results in a system that makes much more efficient use of bandwidth and power and can provide a wider array of services for your customers.

The "real world" is analog (at least according to classical physics). As such we implement digital modulation using analog systems. For the same reasons that FM is superior to AM for analog radio communication, most digital schemes employ frequency- or phase-shift techniques.

Digital Signals

In digital signaling we seek to send one of two logic states. Traditionally we label these states as logic "1" or logic "0". In a digital circuit these states are usually represented by voltages such as 5 V for logic 1 and 0 V for logic 0. In RF it is almost always more convenient to have a symmetric representation, so we will typically use, say 1 V for logic 1 and –1 V for logic 0.

Although we ultimately are interested is communicating a stream of discrete bits, we must implement this with continuous-time signals. Say we wish to send the bit pattern 010… with one bit being sent every T_b seconds. T_b is called the *bit period.* The *bit rate*, or number of bits per second, is $R_b = 1/T_b$. We communicate a bit pattern by forming a continuous-time function $m(t)$ where, say, $m(t) = -1$ for $0 \le t < T_b$, $m(t) = +1$ for $T_b \le t < 2T_b$, and so on. We then modulate $m(t)$ onto an RF channel, transmit it, receive and demodulate it, and finally sample it to reconstruct the bit pattern. In the presence of noise this process may fail and we might end up with logic 1 when logic 0 was sent, or vise versa. In this case we say we have a *bit error.* The fraction of all bits that are in error is called the *bit error rate* or BER. The BER is the same as the probability that a given bit will be in error, which we write as P_e.

The BER will depend on the choice of modulation (amplitude, phase, frequency) that we use to send $m(t)$. The modulation scheme will also affect the required RF bandwidth. One important question is: How much RF bandwidth is required to communicate at a bit rate R_b? A modulation scheme that uses less RF bandwidth is said to be more *bandwidth efficient*. Another important question is: What received power level is required to achieve a given BER? A modulation scheme that requires less power is said to be more *power efficient*. Whether power or bandwidth efficient is more important depends on our application.

Sinc Function

The following type of waveform or *rectangular pulse* arises quite often in digital communication, at least in theory,

$$x(t) = \mathrm{rect}\left(\frac{t}{T_b}\right) A_c \cos(\omega_0 t + \phi) \tag{21.1}$$

where $\mathrm{rect}(t)$ is 1 if $|t| \le 1/2$ and 0 otherwise, and $\omega_0 = 2\pi f_0$. The spectrum of this is

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \\ &= A_c \int_{-T_b/2}^{T_b/2} \cos(2\pi f_0 t + \phi) e^{-j2\pi ft} dt \\ &= \frac{A_c T_b}{2} \left\{ e^{j\phi} \mathrm{sinc}\left[T_b(f - f_0)\right] + e^{-j\phi} \mathrm{sinc}\left[T_b(f + f_0)\right] \right\} \end{aligned} \tag{21.2}$$

where

$$\mathrm{sinc}\, x = \frac{\sin \pi x}{\pi x} \tag{21.3}$$

is the "sinc" function (pronounced "sink"). The sinc function and its square are shown in Fig. 21.1. Note that the sinc is zero whenever its argument is a non-zero integer and $\mathrm{sinc}(0) = 1$. The sinc function has the property that

$$\int_{-\infty}^{\infty} \mathrm{sinc}\, x\, dx = \int_{-\infty}^{\infty} (\mathrm{sinc}\, x)^2\, dx = 1 \tag{21.4}$$

For this reason, and because, as seen in Fig 21.1, the ½ power width of sinc^2 is very close to 1, we can take the characteristic width of the sinc function to be 1. Therefore we will often say that the sinc functions in (21.2) have a bandwidth of $B = 1/T_b$, that is, a bandwidth equal to the bit rate. The spectrum is centered at the carrier frequency f_c or its negative.

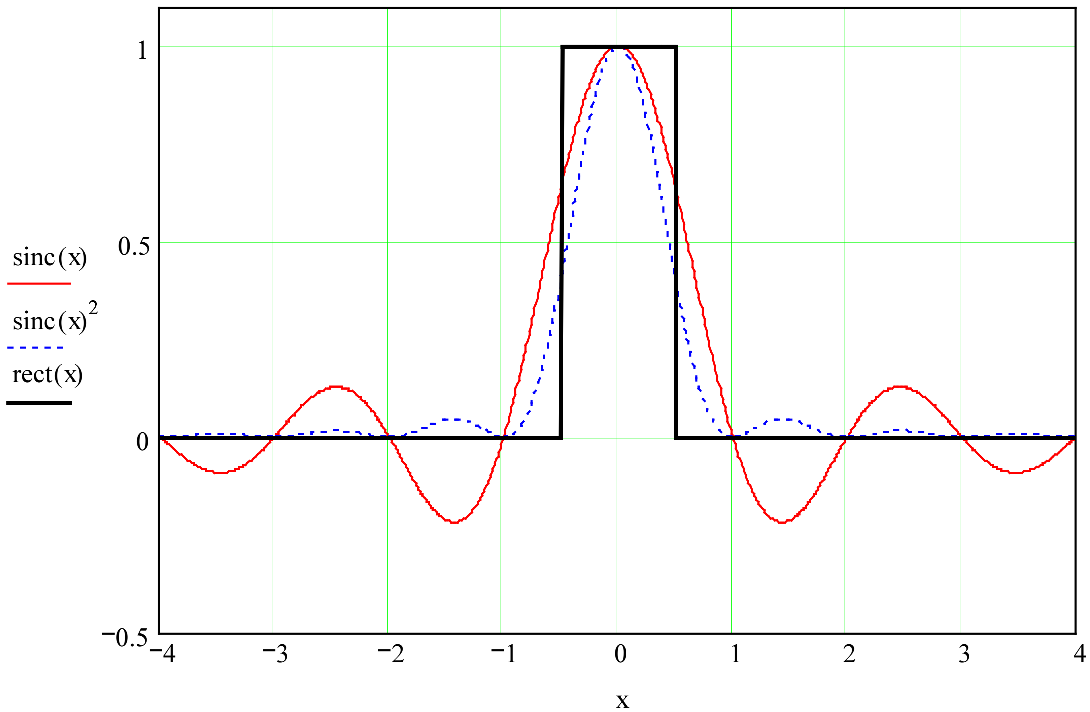

Figure 21.1: The sinc function, sinc function squared, and rect function. All three curves have unit area. About 90% of the area or "energy" of sinc^2 lies in the main lobe (i.e., between –1 and 1). 99% of the energy lies between –10 and 10.

The sinc function is plotted in dB in Fig. 21.2. Note the *sidelobes* in-between the zeros. About 10% of the total energy in sinc^2 lies in the side lobes. The sidelobes are more apparent when viewed on a logarithmic scale as in Fig. 21.2.

The pulse (21.1) carries an average power of $A_c^2/2$ for a time T_b. Power times time is energy, and we define the *bit energy* as

$$E_b = \frac{A_c^2 T_b}{2} \qquad (21.5)$$

Recall also, that if a receiver has noise temperature T_N then the noise spectral density is $N_0 = kT_N$, with k Boltzmann's constant. If the received signal bandwidth is B, then the noise power is $P_N = N_0 B = \langle n(t)^2 \rangle$ (into a 1-Ω load) where $n(t)$ is the noise voltage as a function of time.

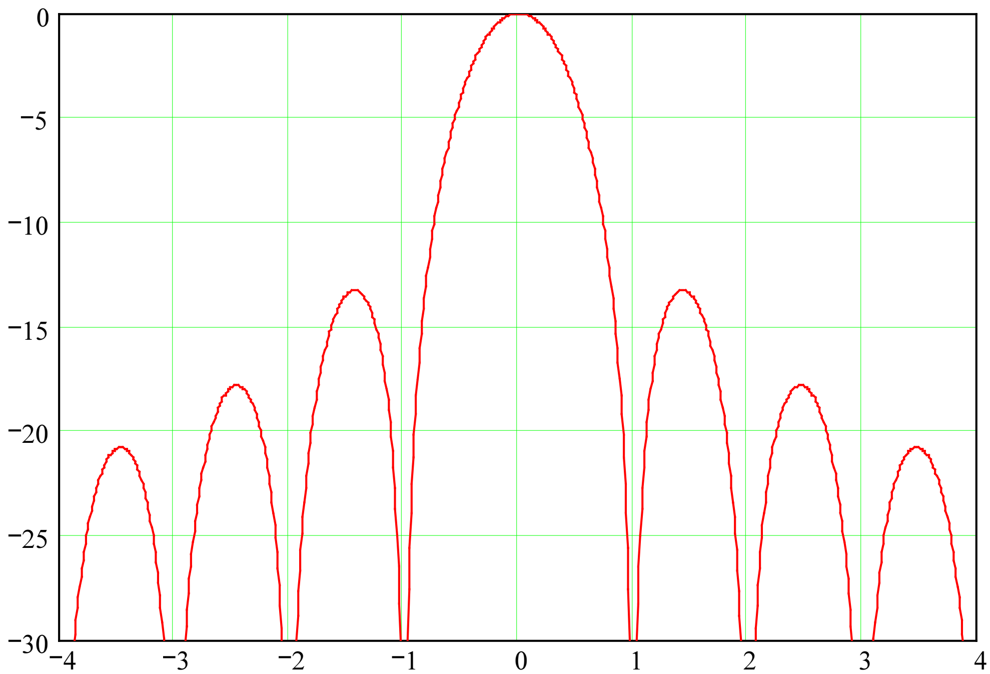

Figure 21.2: $\mathrm{sinc}^2(x)$ *in dB vs.* x.

Frequency Shift Keying (FSK)

Conceptually FSK is very simple. If we wish to send binary data, we associate each of the two logic states "0" and "1" with distinct frequencies, say, f_0 and f_1. To send logic "0" we tune our transmitter to f_0; to send logic "1" we tune our transmitter to f_1. An audio analogy would be to press one of two particular keys of a piano to send "0" or "1". The receiver need only be able to distinguish between the two frequencies. This can be accomplished by, for example, using two band-pass filters, one centered at f_0 and the other at f_1 and seeing which produces the larger output. Errors occur when noise randomly causes the wrong filter to have a larger response.

FSK can, in theory and in practice, be implemented using analog FM transceivers. This is an attractive feature and was especially so in the past when analog FM transceiver technology was much more mature than the corresponding digital technology. To transmit FSK using analog FM we simply apply one of two discrete modulation voltages ± 1 to the transmitter. At the receiver we do an analog FM demodulation and ideally the receiver will output the original bit pattern. Recall that FM produces an RF signal of the form

$$s(t) = A_c \cos\left(2\pi f_c t + 2\pi k_f \int_0^t m(x)dx\right) \quad (21.6)$$

We take $m(t)$ to be a binary signal, i.e., $m = \pm 1$. Since the amplitude of the modulating signal is unity, the frequency deviation is equal to the frequency deviation constant, i.e., $\Delta f = k_f$.

The instantaneous frequency is

$$f(t) = f_c + \Delta f \, m(t) \tag{21.7}$$

This takes on one of the two discrete values $f = f_c \pm \Delta f$. Thus, as we claimed, FSK involves transmitting one of two frequencies to represent one of the two possible logic states. Logic 1 would be represented by a frequency $f_1 = f_c + \Delta f$ and logic 0 by $f_0 = f_c - \Delta f$. An example of an FSK signal is shown in Fig. 21.3.

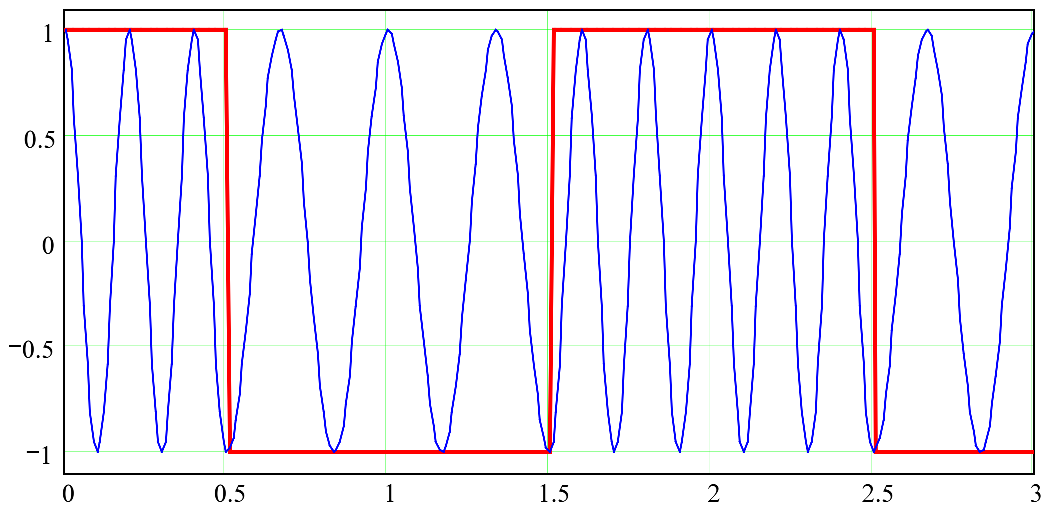

Figure 21.3: Simulated FSK signal.(See Mathcad code in Appendix.) Square wave (thick curve) is the modulating signal. The bit period is 1. The thin waveform is the resulting "RF" signal.

In this example the carrier frequency is 4 and the frequency deviation is 1. So, as seen in the figure, logic 1 is represented by a frequency of $4+1=5$ and logic 0 by $4-1=3$. In a real system there would be many more oscillations during each bit period.

Recall that for FM we defined the modulation index as $\beta = \Delta f / f_m$. For the square-wave bit pattern shown in Fig. 21.3, the period is $2T_b$ so the frequency is $f_m = 1/(2T_b)$ and $\beta = 2\Delta f \, T_b$. Usually the symbol "*h*" is used in the FSK case and we write the *modulation index* as $h = 2\Delta f \, T_b$.

Although an ideal FM receiver would output $m(t)$, in the presence of noise things are more complicated.

BER for FSK

In general, to demodulate an FSK signal and recover the modulation $m(t)$, we need to determine at which of the frequencies f_0, f_1 there is more power present during a given bit period. Ideally all the power would be at one or the other frequency and then we would know which logic state was being sent. However, we are doing this in the presence of noise. With noise it is possible that more power will appear at the wrong frequency than at the right one, and we will have a bit error. We want to figure out how often this happens.

Mathematically, we can take the approach we use to calculate Fourier series coefficients, namely, multiply by a sinusoid and integrate. This can be implemented in hardware using mixers and integrators. Assume a logic "1" is being sent. The received signal will be $A_c \cos \omega_1 t$ for a period T_b. The receiver adds noise to give us a total signal of $A_c \cos \omega_1 t + n(t)$. Now calculate the Fourier coefficient at ω_1. We get

$$\begin{aligned} a_1 &= \frac{2}{T_b} \int_0^{T_b} \left[A_c \cos \omega_1 t + n(t) \right] \cos \omega_1 t \, dt \\ &= A_c + \frac{2}{T_b} \int_0^{T_b} n(t) \cos \omega_1 t \, dt \\ &= A_c + X_1 \end{aligned} \tag{21.8}$$

The second term X_1 is the Fourier coefficient of $\cos \omega_1 t$ for the noise. We assume this is a zero-mean Gaussian random variable. Since the time interval is T_b, spectral components within a bandwidth of $B = 1/T_b$ will contribute to this. This corresponds to a noise power of $N_0 B$ where $N_0 = kT_N$ is the noise spectral density. Therefore, the variance is $\left\langle X_1^2 \right\rangle = N_0 B = N_0 / T_b$. Now let's calculate the Fourier coefficient at ω_0. We get

$$\begin{aligned} a_0 &= \frac{2}{T_b} \int_0^{T_b} \left[A_c \cos \omega_1 t + n(t) \right] \cos \omega_0 t \, dt \\ &= \frac{2}{T_b} \int_0^{T_b} n(t) \cos \omega_0 t \, dt \\ &= X_0 \end{aligned} \tag{21.9}$$

Here we've assumed that $\cos \omega_1 t$ and $\cos \omega_0 t$ are *orthogonal* over this interval. This sets a constraint on the difference of the two frequencies, as we'll see later. X_0 is another Gaussian random variable with variance N_0 / T_b.

If $a_1 > a_0$ then we conclude, correctly, that the logic level is "1". Otherwise we make a bit error. Equivalently, we look at $r = a_1 - a_0$. If this is positive then we assume logic level "1" while if it is negative we assume logic level "0". Now $r = A_c + X_1 - X_0$. X_0, X_1 are both zero-mean Gaussian RVs with variance N_0 / T_b. If they are independent, then statistical theory tells us that their difference with also be a zero-mean Gaussian RV with variance $\sigma^2 = 2N_0 / T_b$. Therefore r will be a Gaussian RV with mean A_c and variance $\sigma^2 = 2N_0 / T_b$ as illustrated in Fig. 21.1.

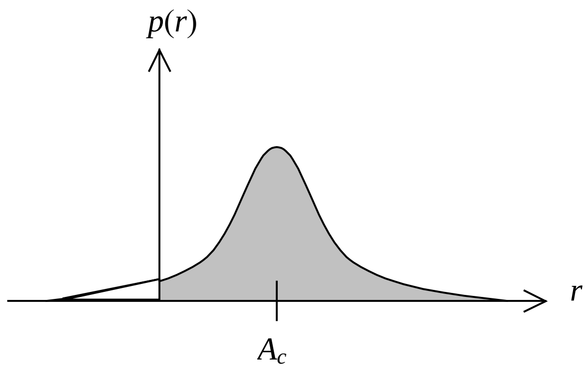

Figure 21.4: PDF of FSK detector output.

We make an error in figuring out the logic state if $r < 0$. The probability of this is

$$P_e = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{0} e^{-\frac{1}{2}\left(\frac{r-A_c}{\sigma}\right)^2} dr$$
$$= Q\left(\frac{A_c}{\sigma}\right) \tag{21.10}$$

If the logic state had been "0", we'd get the mirror image of Fig. 21.1 and find the same error. Therefore (21.10) is the *BER*.

Let's express A_c / σ in a couple of different ways to get some insight into this expression. Using $\sigma^2 = 2N_0 / T_b$ and $E_b = A_c^2 T_b / 2$, we can write $(A_c / \sigma)^2 = A_c^2 T_b / 2N_0 = E_b / N_0$, so

$$P_e = Q\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{21.11}$$

This is plotted in Fig. 21.5. Given a receiver noise temperature, $N_0 = kT_N$ is determined, and this expression tells us the received energy per bit required to achieve a given *BER*. For example, from Fig. 21.5 we can see that to get a *BER* of about 10^{-3} we'd need E_b / N_0 to be about 10 dB.

The fact that E_b / N_0 determines the *BER* has implications for our bit rate $R_b = 1/T_b$. With a fixed N_0 we need to keep E_b fixed to maintain a given *BER*. But $E_b = P_{R,W} T_b = P_{R,W} / R_b$. So, if we increase our bit rate, we must increase received power by the same amount. Conversely, if received power decreases, we can maintain our *BER* by reducing the bit rate. So, the data rate your radio link will operate at is an important consideration in determining the required received power.

Recall that for a bandwidth of $B = 1/T_b$ and a noise spectral density $N_0 = kT_N$, the noise power is $P_{N,W} = N_0 B$. Now

$$\begin{aligned} \frac{E_b}{N_0} &= \frac{P_{R,W} T_b}{N_0} \\ &= \frac{P_{R,W}}{N_0 B} \\ &= \frac{P_{R,W}}{P_{N,W}} \end{aligned} \tag{21.12}$$

That is, E_b / N_0 is also the *S/N* ratio. However, since the noise power depends on the bandwidth, which depends on the data rate, E_b / N_0 is in a sense a more fundamental quantity.

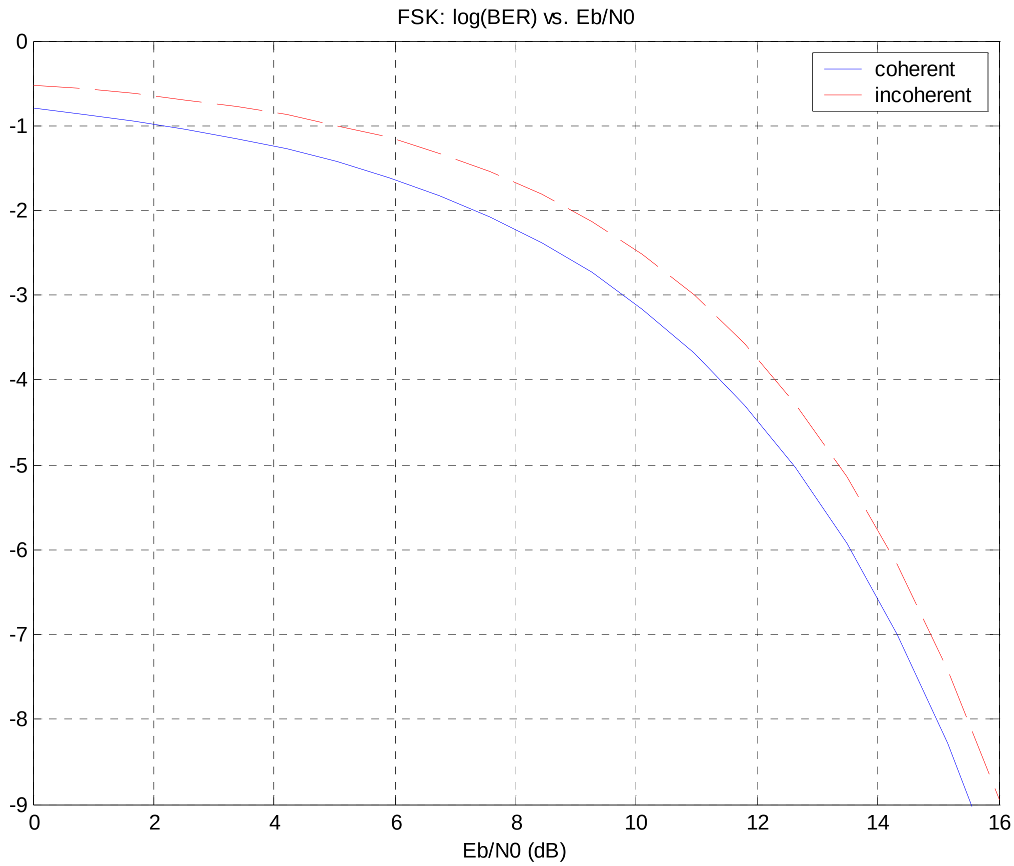

Figure 21.5: BER vs. E_b / N_0. *Solid blue curve corresponds to coherent detection; dashed red curve corresponds to incoherent detection.*

In writing (21.8) and (21.9) we've implicitly assumed that the cosine in the signal and the cosine in our detector had the same phase. This kind of detection scheme we've outlined is a *coherent* scheme. If you can't get phase coherence, you have to use a *noncoherent* detector. For example, you could compare the amplitudes of the output of two bandpass filters, one tuned to f_0 and the other to f_1. The analysis is more difficult than for the coherent case. The result is

$$P_e = \frac{1}{2} e^{-\frac{E_b}{2N_0}} \tag{21.13}$$

This is also plotted in Fig. 21.5. You can see that the *BER* is higher than for the coherent detector. Typically, however, less than 1 dB of increase in signal power will make up the difference.

FSK Spectrum

To signal a sequence of logic states, we are sending a series of pulses of the form (21.1) each with frequency $f_0 = f_c - \Delta f$ or $f_1 = f_c + \Delta f$. The spectrum of a single pulse is given by (21.2). We might be inclined to think that the power spectrum of an FSK signal would therefore consist of two squared sincs, one centered at f_0 and one at f_1. This is more-or-less true if the frequency deviation is greater than or about equal to the width of one sinc, namely, $1/T_b$. For the waveform shown in Fig. 21.3, the power spectrum is as shown in Fig. 21.6

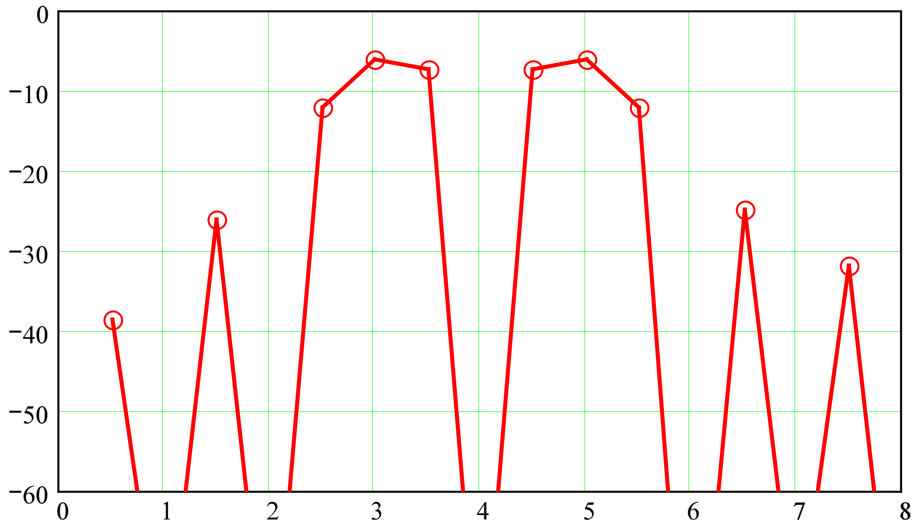

Figure 21.6: Power spectrum of the FSK signal shown in Fig. 21.3. Power in dB vs. frequency.

For this situation we can estimate the bandwidth as the difference of the frequencies f_0 and f_1, which is $2\Delta f$, plus half the width of the lower sinc, plus half the width of the upper sinc, i.e.,

$$B = 2\Delta f + 1/T_b \tag{21.14}$$

For the situation illustrated in Fig. 21.6, this would be 3. Or we can use Carson's rule $B = 2(\Delta f + f_m)$. In this case we have to realize that if our bits are flipping back and forth between –1 and 1, then $m(t)$ has a period of $2T_b$, corresponding to a frequency of $f_m = 1/2T_b$. Putting this into Carson's rule gives (21.14).

So, we see that our BER is given by (21.11) and our bandwidth by (21.14). Since the BER doesn't (apparently) depend on B, why not let $\Delta f \to 0$ so that we only need to use a bandwidth $1/T_b$? Clearly something must be in our way, because $\Delta f = 0$ would mean that there would be no modulation and hence no information being sent.

MSK & GMSK

What is the minimum bandwidth (21.14) that we can use for FSK? Implicit in (21.9) is the orthogonality of $\cos\omega_0 t$ and $\cos\omega_1 t$:

$$\begin{aligned}\frac{2}{T_b}\int_0^{T_b}\cos\omega_1 t\cos\omega_0 t dt &= \frac{1}{T_b}\int_0^{T_b}\cos(\omega_1+\omega_0)t\,dt + \frac{1}{T_b}\int_0^{T_b}\cos(\omega_1-\omega_0)t\,dt \\ &= \frac{\sin(\omega_1+\omega_0)T_b}{(\omega_1+\omega_0)T_b} + \frac{\sin(\omega_1-\omega_0)T_b}{(\omega_1-\omega_0)T_b} \\ &= \operatorname{sinc}\left[(f_1+f_0)2T_b\right] + \operatorname{sinc}\left[(f_1-f_0)2T_b\right] \\ &= 0\end{aligned} \quad (21.15)$$

The first term is zero (or extremely close) because $f_1 + f_0$ is very large. For the second term to be zero, however, we require $(f_1 - f_0)2T_b$ to be at least 1. Since $f_1 - f_0 = 2\Delta f$, this means

$$\begin{aligned}\Delta f &= \frac{1}{4T_b} \\ &= \frac{R_b}{4}\end{aligned} \quad (12.16)$$

that is, the frequency deviation must be at least one-fourth the bit rate. This defines the minimum frequency deviation that results in orthogonal signals. FSK with this frequency deviation is referred to as *minimum shift keying* or MSK. For MSK we have $f_1 - f_0 = 1/2T_b$ and the modulation index is $h = 1/2$. Therefore, during a period T_b, $\cos\omega_1 t$ will go through an extra ½ of a sine wave compared to $\cos\omega_0 t$. An example of an MSK signal is shown in Fig. 21.7.

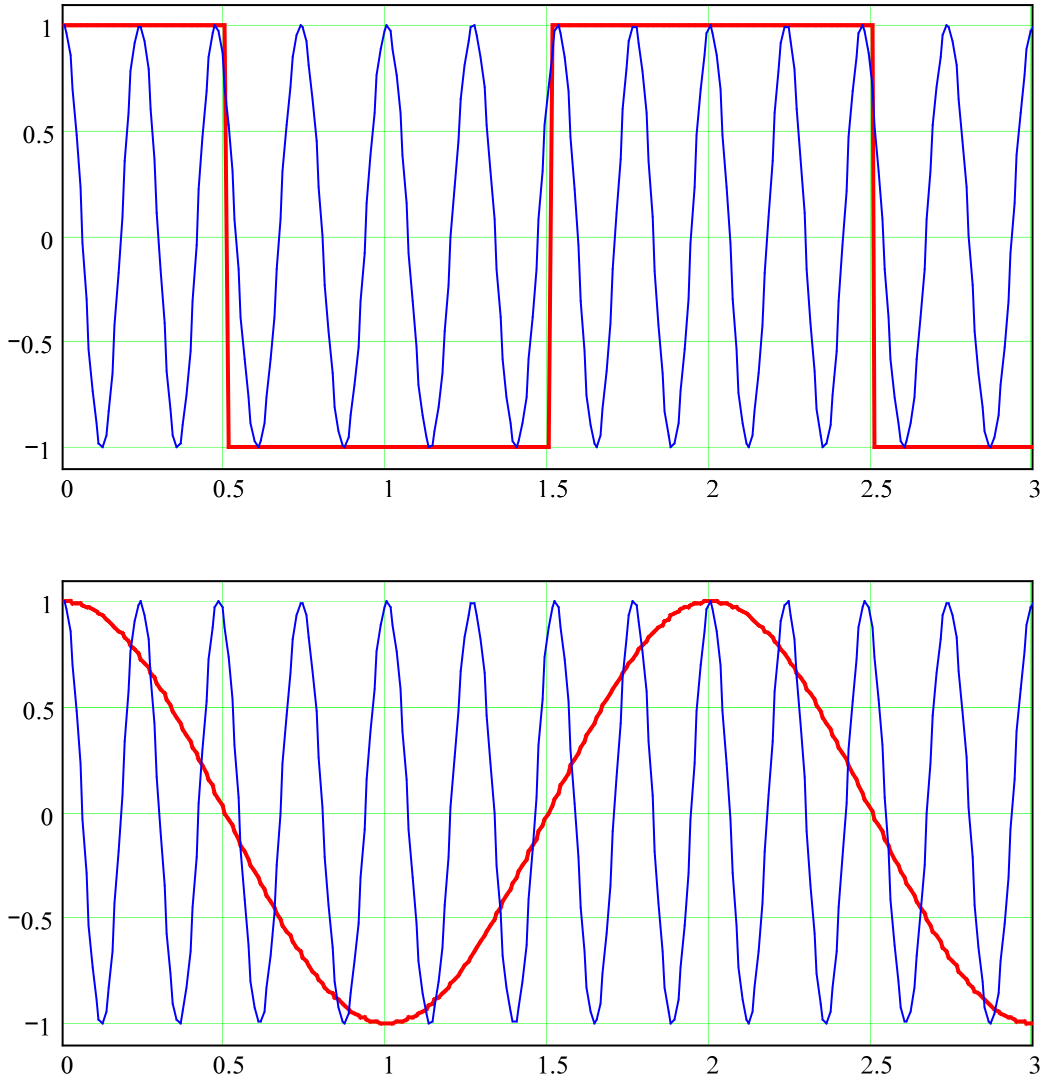

Figure 21.7: Simulated Minimum Shift Keying (MSK) signals. At the top is "unfiltered" MSK. At the bottom is "filtered" GMSK.

The "two-sincs" picture of the power spectrum breaks down when Δf gets less than about $1/T_b$ because the sincs overlap. The theoretical, unfiltered MSK power spectrum is

$$|S(f)|^2 = \frac{16}{\pi^2}\left[\frac{\cos(2\pi(f-f_c)T)}{(4(f-f_c)T)^2-1}\right]^2 \tag{12.17}$$

This is shown in Fig. 21.8 (boxes) along with numerical results for the waveform of Fig. 21.7.

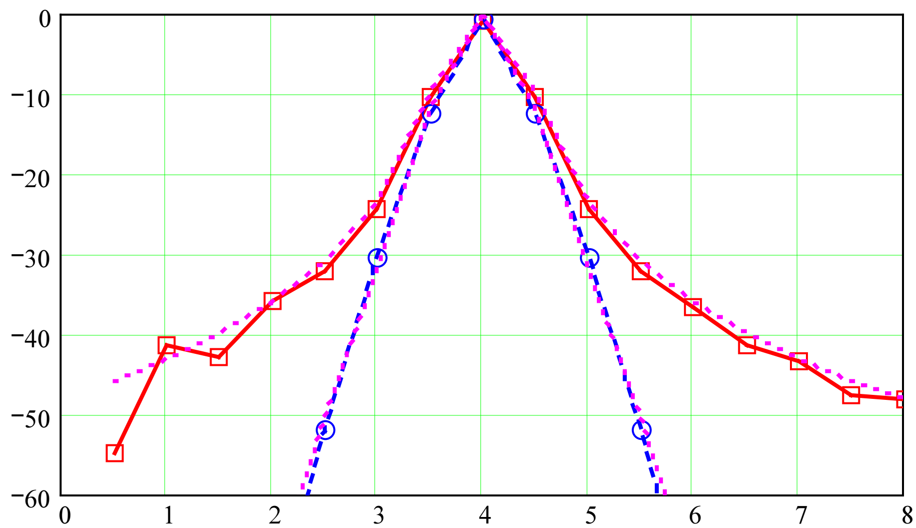

Figure 21.8: MSK and GMSK power spectra, simulation and theory. Boxes/red curve is unfiltered MSK spectrum. Circles/blue curve is GMSK. Dotted magenta curves are theoretical spectra.

As can be seen in Fig. 21.3, in an FSK waveform there is a sharp transition from one frequency to the next. Rapid transitions generally require a wider bandwidth than smooth transitions. As a result, filtering the modulation $m(t)$ can reduce the MSK bandwidth because the transmitter will smoothly vary between the two frequencies. This is illustrated at the bottom of Fig. 21.7. The corresponding spectrum is shown in Fig. 21.8.

Fitlering $m(t)$ with a Gaussian impulse response results in multiplying the MSK spectrum by a Gaussian frequency response. If the impulse reponse is

$$h(t) = \frac{\sqrt{\pi}}{\alpha} e^{-\left(\frac{\pi}{\alpha}t\right)^2} \tag{21.18}$$

the spectrum is multiplied by

$$H(f) = e^{-[\alpha(f-f_c)]^2} \tag{21.19}$$

The result is called *Gaussian Minimum Shift Keying*, or GMSK. The parameter α determines the bandwidth – a larger α value results in smaller bandwidth. The case shown in Fig. 21.8 has $\alpha \approx 1$. The GSM digital cellular standard uses $\alpha \approx 2$. Although GMSK reduces the bandwidth, it also increases the BER somewhat because the signal no longer spends a full T_b at either of the two frequencies f_0 or f_1.

References

1. Anderson, J. B., *Digital Transmission Engineering*, IEEE Press, 1999, ISBN 0-13-082961-7.

2. Burr, A., *Modulation and Coding for Wireless Communications,* Prentice Hall, 2001, ISBN 0-201-39857-5.

Appendix

The following is the Matcad code used to generate most of the figures in the text.

FSK Simulation

Bit period, carrier frequency, and frequency deviation constant:

$$T := 1 \qquad f_c := \frac{4}{T} \qquad k_f := \frac{4}{4 \cdot T}$$

unfiltered square-wave modulation: $m_0(t) := \mathrm{sign}\left(\cos\left(\pi \cdot \frac{t}{T}\right)\right)$

"filtered" modulation: $m_1(t) := \cos\left(\pi \cdot \frac{t}{T}\right)$

FM signals:

$$s_0(t) := \cos\left(2\pi \cdot f_c \cdot t + 2\pi k_f \cdot \int_0^t m_0(x)\,dx\right)$$

$$s_1(t) := \cos\left(2\pi \cdot f_c \cdot t + 2\pi k_f \cdot \int_0^t m_1(x)\,dx\right)$$

These are periodic, even functions with period 2T, Fourier coefficients are:

$n := 1 .. 16$

$$a0_n := \frac{2}{2T} \cdot \int_{-T}^{T} s_0(t) \cdot \cos\left(\frac{2\pi}{2T} \cdot n \cdot t\right) dt \qquad a1_n := \frac{2}{2T} \cdot \int_{-T}^{T} s_1(t) \cdot \cos\left(\frac{2\pi}{2T} \cdot n \cdot t\right) dt$$