MIT 6.02 **DRAFT** *Lecture Notes*
Last update: April 11, 2012
Comments, questions or bug reports?
Please contact {hari, verghese} at mit.edu

CHAPTER 14
Modulation and Demodulation

This chapter describes the essential principles behind *modulation* and *demodulation*, which we introduced briefly in Chapter 10. Recall that our goal is to transmit data over a communication link, which we achieve by mapping the *bit stream* we wish to transmit onto analog *signals* because most communication links, at the lowest layer, are able to transmit analog signals, not binary digits. The signals that most simply and directly represent the bit stream are called the *baseband signals*. We discussed in Chapter 10 why it is generally untenable to directly transmit baseband signals over communication links. We reiterate and elaborate on those reasons in Section 14.1, and discuss the motivations for **modulation** of a baseband signal. In Section 14.2, we describe a basic principle used in many modulation schemes, called the *heterodyne principle*. This principle is at the heart of *amplitude modulation* (AM), the scheme we study in detail. Sections 14.3 and 14.4 describe the "inverse" process of **demodulation**, to recover the original baseband signal from the received version. Finally, Section 14.5 provides a brief overview of more sophisticated modulation schemes.

■ 14.1 Why Modulation?

There are two principal motivating reasons for modulation. We described the first in Chapter 10: matching the transmission characteristics of the medium, and considerations of power and antenna size, which impact *portability*. The second is the desire to *multiplex*, or share, a communication medium among many concurrently active users.

■ 14.1.1 Portability

Mobile phones and other wireless devices send information across free space using electromagnetic waves. To send these electromagnetic waves across long distances in free space, the frequency of the transmitted signal must be quite high compared to the frequency of the information signal. For example, the signal in a cell phone is a voice signal with a bandwidth of about 4 kHz. The typical frequency of the transmitted and received signal is several hundreds of megahertz to a few gigahertz (for example, the popular WiFi standard is in the 2.4 GHz or 5+ GHz range).

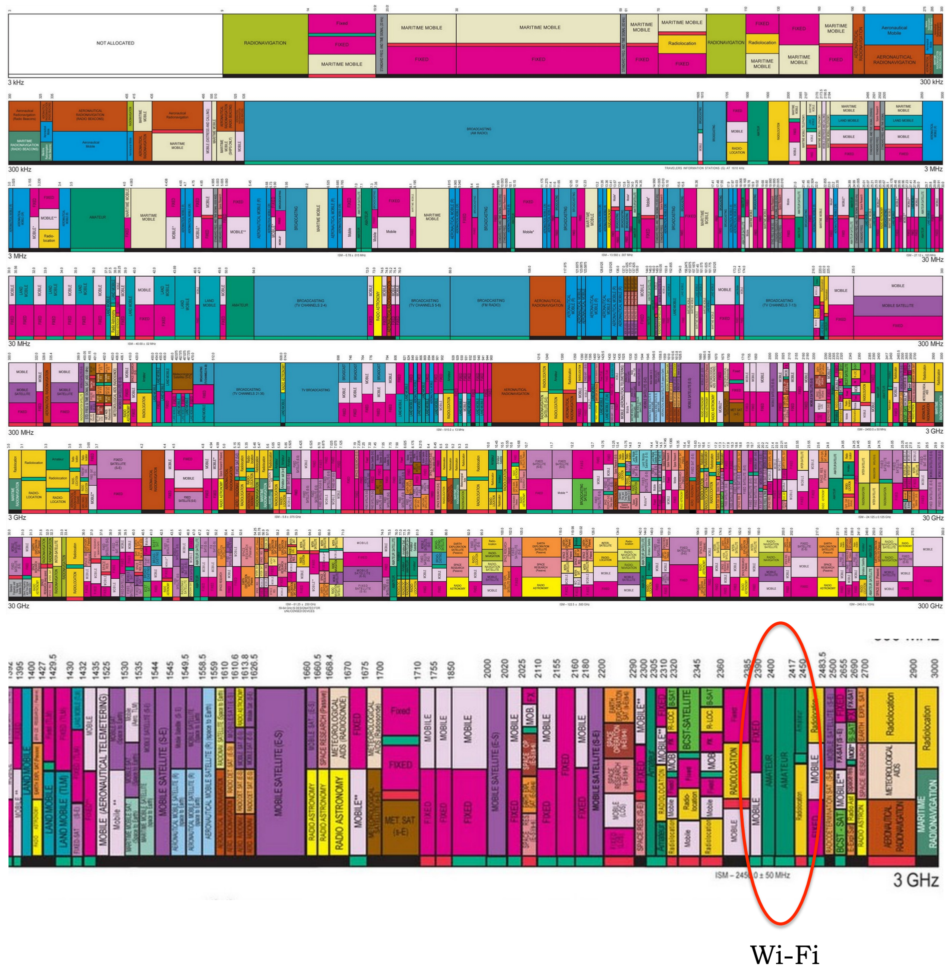


Figure 14-1: Top: Spectrum allocation in the United States (3 kHz to 300 GHz). Bottom: a portion of the total allocation, highlighting the 2.4 GHz ISM (Industrial, Scientific, and Medical) band, which is unlicensed spectrum that can be used for a variety of purposes, including 802.11b/g (WiFi), various cordless telephones, baby monitors, etc. From `http://www.wireless-technology.org/wp-content/uploads/2011/02/Wireless-Spectrum-Photo.jpg`

One important reason why high-frequency transmission is attractive is that the size of the antenna required for efficient transmission is roughly one-quarter the wavelength of the propagating wave, as discussed in Chapter 10. Since the wavelength of the (electromagnetic) wave is inversely proportional to the frequency, the higher the frequency, the smaller the antenna. For example, the wavelength of a 1 GHz electromagnetic wave in free space is 30 cm, whereas a 1 kHz electromagnetic wave is one million times larger, 300 km, which would make for an impractically huge antenna and transmitter power to transmit signals of that frequency!

■ 14.1.2 Sharing using Frequency-Division

Figure 14-1 shows the electromagnetic spectrum from 3 kHz to 300 GHz; it depicts how portions of spectrum have been allocated by the U.S. Federal Communications Commis-

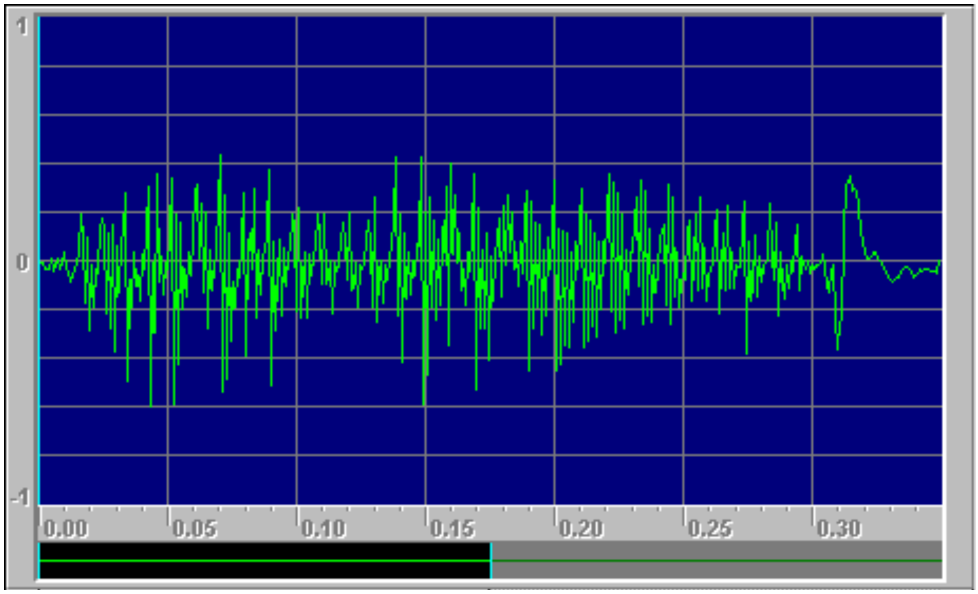


Figure 14-2: An analog waveform corresponding to someone saying "Hello". Picture from `http://electronics.howstuffworks.com/analog-digital2.htm`. The frequency content and spectrum of this waveform is inherently band-limited to a few kilohertz.

sion (FCC), which is the government agency that allocates this "public good" (spectrum). What does "allocation" mean? It means that the FCC has divided up frequency ranges and assigned them for different uses and to different entities, doing so because one can be assured that concurrent transmissions in different frequency ranges will not interfere with each other.

The reason why this approach works is that when a sinusoid of some frequency is sent through a linear, time-invariant (LTI) channel, the output is a sinusoid of the *same frequency*, as we discovered in Chapter 12. Hence, if two different users send pure sinusoids at different frequencies, their intended receivers can extract the transmitted sinusoid by simply applying the appropriate *filter*, using the principles explained in Chapter 12.

Of course, in practice one wants to communicate a baseband signal rather than a sinusoid over the channel. The baseband signal will often have been produced from a digital source. One can, as explained in Chapters 9 and 10, map each "1" to a voltage V_1 held for some interval of time, and each "0" to a voltage V_0 held for the same duration (let's assume for convenience that both V_1 and V_0 are non-negative). The result is some waveform that might look like the picture shown in Figure 10-2.[1] Alternatively, the baseband signal may come from an analog source, such as a microphone in an analog telephone, whose waveform might look like the picture shown in Figure 14-2; this signal is inherently "band-limited" to a few kilohertz, since it is produced from human voice. Regardless of the provenance of the input baseband signal, the process of modulation involves preparing the signal for transmission over a channel.

If multiple users concurrently transmitted their baseband signals over a shared

[1]We will see in the next section that we will typically remove its higher frequencies by lowpass filtering, to obtain a "band-limited" baseband signal.

medium, it would be difficult for their intended receivers to extract the signals reliably because of interference. One approach to reduce this interference, known as **frequency-division multiplexing**, allocates different **carrier frequencies** to different users (or for different uses, e.g., one might separate out the frequencies at which police radios or emergency responders communicate from the frequencies at which you make calls on your mobile phone). In fact, the US spectrum allocation map shown in Figure 14-1 is the result of such a frequency-division strategy. It enables users (or uses) that may end up with similar looking baseband signals (those that will interfere with each other) to be transmitted on different carrier frequencies, eliminating interference.

There are two reasons why frequency-division multiplexing works:

1. Any baseband signal can be broken up into a weighted sum of sinusoids using Fourier decomposition (Chapter 13). If the baseband signal is band-limited, then there is a finite maximum frequency of the corresponding sinusoids. One can take this sum and modulate it on a carrier signal of some other frequency in a simple way: by just multiplying the baseband and carrier signal (also called "mixing"). The result of modulating a band-limited baseband signal on to a carrier is a signal that is band-limited around the *carrier*, i.e., *limited to some maximum frequency deviation from the carrier frequency*.

2. When transmitted over a linear, time-invariant (LTI) channel, and if noise is negligible, each sinusoid shows up at the receiver as a sinusoid *of the same frequency*, as we saw in Chapter 12. The reason is that *an LTI system preserves the sinusoids*. If we were to send a baseband signal composed of a sum of sinusoids over the channel, the output will be the sum of sinuoids of the same frequencies. Each receiver can then apply a suitable filter to extract the baseband signal of interest to it. This insight is useful because the noise-free behavior of real-world communication channels is often well-characterized as an LTI system.

■ 14.2 Amplitude Modulation with the Heterodyne Principle

The **heterodyne principle** is the basic idea governing several different modulation schemes. The idea is simple, though the notion that it can be used to modulate signals for transmission was hardly obvious before its discovery!

> **Heterodyne principle: The multiplication of two sinusoidal waveforms may be written as the sum of two sinusoidal waveforms, whose frequencies are given by the sum and the difference of the frequencies of the sinusoids being multiplied.**

This result may be seen from standard high-school trigonometric identities, or by (perhaps more readily) writing the sinusoids as complex exponentials and performing the multiplication. For example, using trigonometry,

$$\cos(\Omega_s n) \cdot \cos(\Omega_c n) = \frac{1}{2}\Big(\cos(\Omega_s + \Omega_c)n + \cos(\Omega_s - \Omega_c)n\Big). \tag{14.1}$$

We apply the heterodyne principle by treating the baseband signal —think of it as periodic with period $\frac{2\pi}{\Omega_1}$ for now—as the sum of different sinusoids of frequencies $\Omega_{s1} = k_1\Omega_1, \Omega_{s2} = k_2\Omega_2, \ldots$ and treating the carrier as a sinusoid of frequency $\Omega_c = k_c\Omega_1$. Here, Ω_1 is the *fundamental frequency* of the baseband signal.

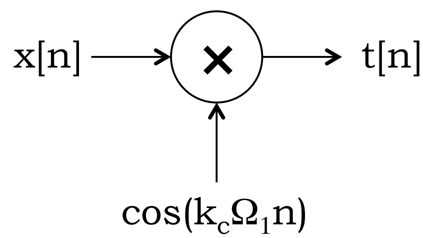


Figure 14-3: Modulation involved "mixing", or multiplying, the input signal $x[n]$ with a carrier signal ($\cos(\Omega_c n) = \cos(k_c\Omega_1 n)$ here) to produce $t[n]$, the transmitted signal.

The application of the heterodyne principle to modulation is shown schematically in Figure 14-3. Mathematically, we will find it convenient to use complex exponentials; with that notation, the process of modulation involves two important steps:

1. **Shape the input to band-limit it.** Take the input baseband signal and apply a low-pass filter to *band-limit* it. There are multiple good reasons for this input filter, but the main one is that we are interested in frequency division multiplexing and wish to make sure that there is no interference between concurrent transmissions. Hence, if we limit the discrete-time Fourier series (DTFS) coefficients to some range, call it $[-k_x, -k_x]$, then we can divide the frequency spectrum into non-overlapping ranges of size $2k_x$ to ensure that no two transmissions interfere. Without such a filter, the baseband could have arbitrarily high frequencies, making it hard to limit interference in general. Denote the result of shaping the original input by $x[n]$; in effect, that is the baseband signal we wish to transmit. An example of the original baseband signal and its shaped version is shown in Figure 14-4.

   We may express $x[n]$ in terms of its discrete-time Fourier series (DTFS) representation as follows, using what we learned in Chapter 13:

   $$x[n] = \sum_{k=-k_x}^{k_x} A_k e^{jk\Omega_1 n}. \tag{14.2}$$

   Notice how applying the input filter ensures that high-frequency components are zero; the frequency range of the baseband is now $[-k_x\Omega_1, k_x\Omega_1]$ radians/sample.

2. **Mixing step.** Multiply $x[n]$ (called the baseband modulating signal) by a carrier, $\cos(k_c\Omega_1 n)$, to produce the signal ready for transmission, $t[n]$. Using the DTFS form,

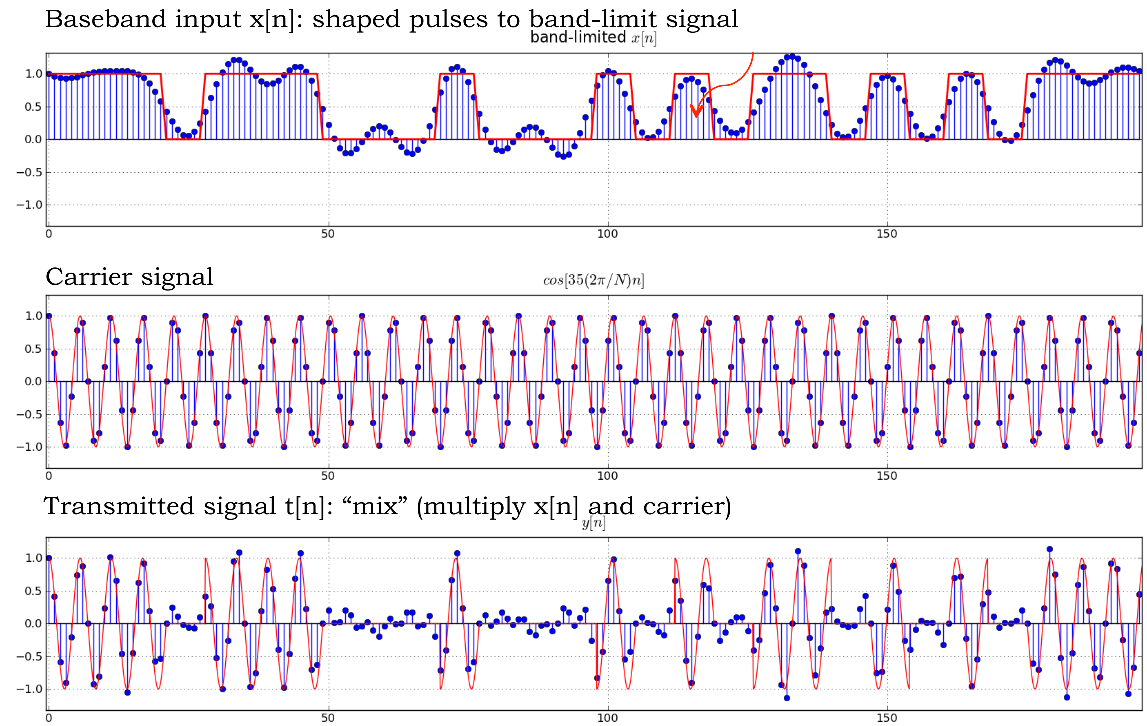


Figure 14-4: The two modulation steps, input filtering (shaping) and mixing, on an example signal.

we get

$$
\begin{aligned}
t[n] &= \left(\sum_{k=-k_x}^{k_x} A_k e^{jk\Omega_1 n}\right)\left(\frac{1}{2}(e^{jk_c\Omega_1 n} + e^{-jk_c\Omega_1 n})\right) \\
&= \frac{1}{2}\sum_{k=-k_x}^{k_x} A_k e^{j\underline{(k+k_c)}\Omega_1 n} + \frac{1}{2}\sum_{k=-k_x}^{k_x} A_k e^{j\underline{(k-k_c)}\Omega_1 n}. \qquad (14.3)
\end{aligned}
$$

Equation (14.3) makes it apparent (see the underlined terms) that the process of mixing produces, for each DTFS component, *two* frequencies of interest: *one at the sum* and *the other at the difference* of the mixed (multiplied) frequencies, each scaled to be one-half in amplitude compared to the original.

We transmit $t[n]$ over the channel. The heterodyne mixing step may be explained mathematically using Equation (14.3), but you will rarely need to work out the math from scratch in any given problem: all you need to know and appreciate is that the (shaped) baseband signal is simply *replicated* in the *frequency domain* at two different frequencies, $\pm k_c$, which are the nonzero DTFS coefficients of the *carrier sinusoidal signal*, and scaled by 1/2. We show this outcome schematically in Figure 14-5.

The time-domain representation shown in Figure 14-4 is not as instructive as the *frequency-domain* picture to gain intuition about what modulation does and why frequency-division multiplexing avoids interference. Figure 14-6 shows the same information as Figure 14-4, but in the frequency domain. The caption under that figure explains the key insights.

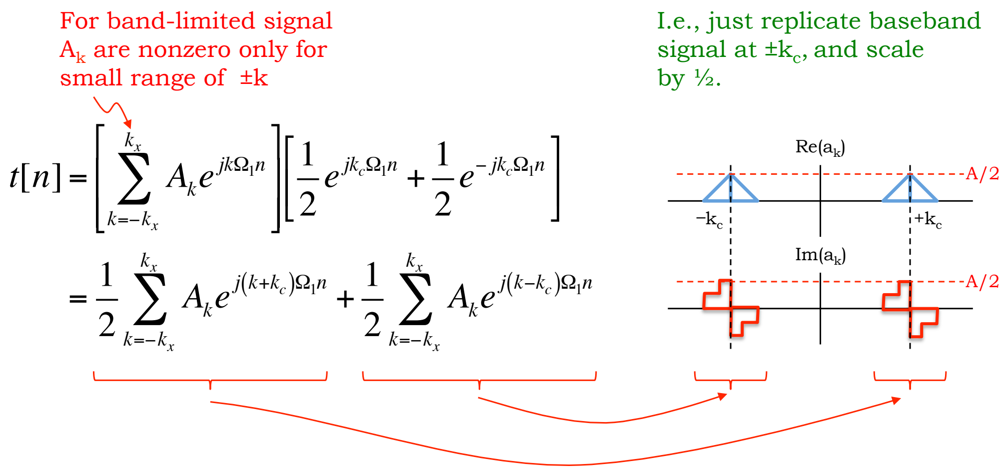


Figure 14-5: Illustrating the heterodyne principle.

This completes our discussion of the modulation process, at least for now (we'll revisit it in Section 14.5), bringing us to the question of how to extract the (shaped) baseband signal at the receiver. We turn to this question next.

■ 14.3 Demodulation: The Simple No-Delay Case

Assume for simplicity that the receiver captures the transmitted signal, $t[n]$, with no distortion, noise, or delay; that's about as perfect as things can get. Let's see how to demodulate the received signal, $r[n] = t[n]$, to extract $x[n]$, the shaped baseband signal.

The trick is to apply the heterodyne principle once again: *multiply the received signal by a local sinusoidal signal that is identical to the carrier*! An elegant way to see what would happen is to start with Figure 14-6, rather than the time-domain representation. We now can pretend that we have a "baseband" signal whose frequency components are as shown in Figure 14-6, and what we're doing now is to "mix" (i.e., multiply) that with the carrier. We can accordingly take each of the two (i.e., real and imaginary) pieces in the right-most column of Figure 14-6 and treat each in turn.

The result is shown in Figure 14-7. The left column shows the frequency components of the original (shaped) baseband signal, $x[n]$. The middle column shows the frequency components of the modulated signal, $t[n]$, which is the same as the right-most column of Figure 14-6. The carrier ($\cos(35\Omega_1 n)$, so the DTFS coefficients of $t[n]$ are centered around $k = -35$ and $k = 35$ in the middle column. Now, when we mix that with a local signal identical to the carrier, we will shift each of these two groups of coefficients by ± 35 once again, to see a cluster of coefficients at -70 and 0 (from the -35 group) and at 0 and $+70$ (from the $+35$ group). Each piece will be scaled by a *further* factor of $1/2$, so the left and right clusters on the right-most column in Figure 14-7 will be $1/4$ as large as the original baseband components, while the middle cluster centered at 0, with the *same spectrum as the original baseband signal*, will be scaled by $1/2$.

What we are interested in recovering is precisely this *middle portion*, centered at 0, because in the absence of any distortion, it is *exactly the same as the original (shaped) baseband,*

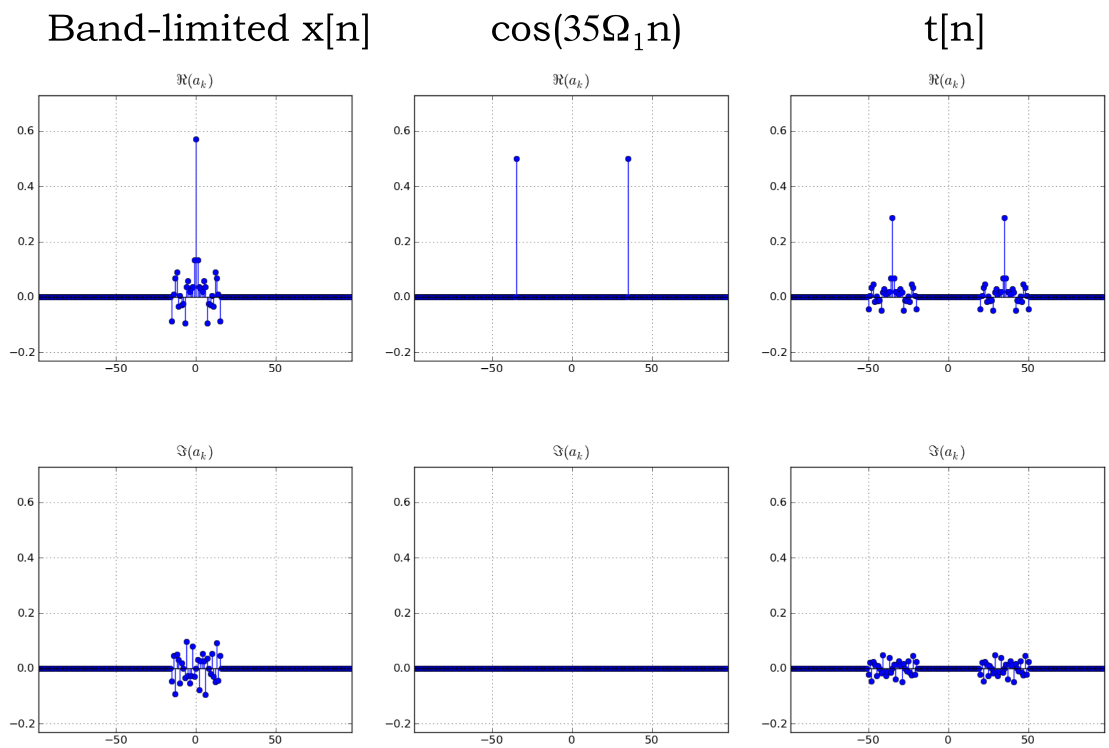


Figure 14-6: Frequency-domain representation of Figure 14-4, showing how the DTFS components (real and imaginary) of the real-valued band-limited signal $x[n]$ after input filtering to produce shaped pulses (left), the purely cosine sinusoidal carrier signal (middle), and the heterodyned (mixed) baseband and carrier at two frequency ranges whose widths are the same as the baseband signal, but that have been shifted $\pm k_c$ in frequency, and scaled by $1/2$ each (right). We can avoid interference with another signal whose baseband overlaps in frequency, by using a carrier for the other signal sufficiently far away in frequency from k_c.

except that is scaled by $1/2$.

How would we recover this middle piece alone and ignore the left and right clusters, which are centered at frequencies that are at twice the carrier frequency in the positive and negative directions? We have already studied a technique in Chapter 12: a *low-pass filter*. By applying a low-pass filter whose cut-off frequency lies between k_x and $2k_c - k_x$, we can recover the original signal faithfully.

We can reach the same conclusions by doing a more painstaking calculation, similar to the calculations we did for the modulation, leading to Equation (14.3). Let $z[n]$ be the signal obtained by multiplying (mixing) the local replica of the carrier $\cos(k_c\Omega_1 n)$ and the received signal, $r[n] = t[n]$, which is of course equal to $x[n]\cos(k_c\Omega_1 n)$. Using Equation 14.3, we can express $z[n]$ in terms of its DTFS coefficients as follows:

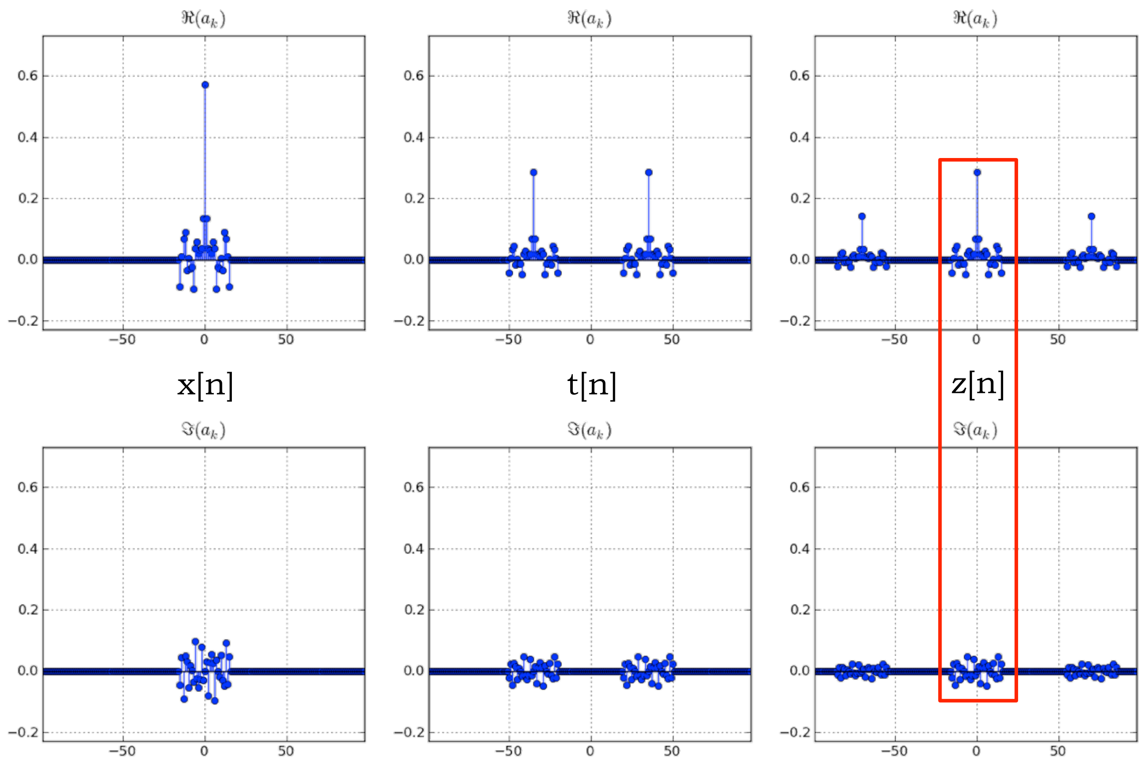


Figure 14-7: Applying the heterodyne principle in demodulation: frequency-domain explanation. The left column is the (shaped) baseband signal spectrum, and the middle column is the spectrum of the modulated signal that is transmitted and received. The portion shown in the vertical rectangle in the right-most column has the DTFS coefficients of the (shaped) baseband signal, $x[n]$, scaled by a factor of $1/2$, and may be recovered faithfully using a low-pass filter. This picture shows the simplified ideal case when there is no channel distortion or delay between the sender and receiver.

$$\begin{aligned} z[n] &= t[n]\Big(\frac{1}{2}e^{jk_c\Omega_1 n} + \frac{1}{2}e^{-jk_c\Omega_1 n}\Big) \\ &= \Big(\frac{1}{2}\sum_{k=-k_x}^{k_x} A_k e^{j(k+k_c)\Omega_1 n} + \frac{1}{2}\sum_{k=-k_x}^{k_x} A_k e^{j(k-k_c)\Omega_1 n}\Big)\Big(\frac{1}{2}e^{jk_c\Omega_1 n} + \frac{1}{2}e^{-jk_c\Omega_1 n}\Big) \\ &= \frac{1}{4}\sum_{k=-k_x}^{k_x} A_k e^{j(k+2k_c)\Omega_1 n} + \underline{\frac{1}{2}\sum_{k=-k_x}^{k_x} A_k e^{jk\Omega_1 n}} + \frac{1}{4}\sum_{k=-k_x}^{k_x} A_k e^{j(2k_c+k)\Omega_1 n} \end{aligned} \tag{14.4}$$

The middle term, underlined, is what we want to extract. The first term is at twice the carrier frequency above the baseband, while the third term is at twice the carrier frequency below the baseband; both of those need to be filtered out by the demodulator.

■ 14.3.1 Handling Channel Distortions

Thus far, we have considered the ideal case of no channel distortions or delays. We relax this idealization and consider channel distortions now. If the channel is LTI (which is very

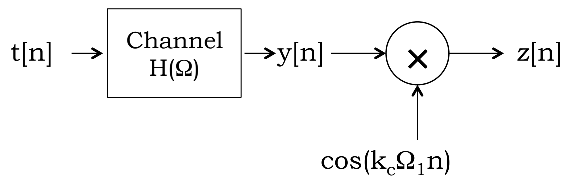


Figure 14-8: Demodulation in the presence of channel distortion characterized by the frequency response of the channel.

often the case), then one can extend the approach described above. The difference is that each of the A_k terms in Equation (14.4), as well as Figure 14-7, will be multiplied by the frequency response of the channel, $H(\Omega)$, evaluated at a frequency of $k\Omega_1$. So each DTFS coefficient will be scaled further by the value of this frequency response at the relevant frequency.

Figure 14-8 shows the model of the system now. The modulated input, $t[n]$, traverses the channel en route to the demodulator at the receiver. The result, $z[n]$, may be written as follows:

$$\begin{aligned} z[n] &= y[n]\cos(k_c\Omega_1 n) \\ &= y[n]\Big(\frac{1}{2}e^{jk_c\Omega_1 n} + \frac{1}{2}e^{-jk_c\Omega_1 n}\Big) \\ &= \Big(\frac{1}{2}\sum_{k=-k_x}^{k_x} H((k+k_c)\Omega_1)A_k e^{j(k+k_c)\Omega_1 n} + \frac{1}{2}\sum_{k=-k_x}^{k_x} H((k-k_c)\Omega_1)A_k e^{j(k-k_c)\Omega_1 n}\Big)\cdot \\ &\quad \Big(\frac{1}{2}e^{jk_c\Omega_1 n} + \frac{1}{2}e^{-jk_c\Omega_1 n}\Big) \\ &= \underline{\frac{1}{4}\sum_{k=-k_x}^{k_x} A_k e^{jk\Omega_1 n}\Big(H((k+k_c)\Omega_1) + H((k-k_c)\Omega_1)\Big)} + \\ &\quad \frac{1}{4}\sum_{k=-k_x}^{k_x} A_k e^{j(k+2k_c)\Omega_1 n}\Big(H((k+k_c)\Omega_1) + H((k-k_c)\Omega_1)\Big) + \\ &\quad \frac{1}{4}\sum_{k=-k_x}^{k_x} A_k e^{j(k-2k_c)\Omega_1 n}\Big(H((k+k_c)\Omega_1) + H((k-k_c)\Omega_1)\Big) \end{aligned} \tag{14.5}$$

Of these three terms in the RHS of Equation (14.5), the first term contains the baseband signal that we want to extract. We can do that as before by applying a lowpass filter to get rid of the $\pm 2k_c$ components. To then recover each A_k, we need to pass the output of the lowpass filter to another LTI filter that undoes the distortion by multiplying the k^{th} Fourier coefficient by the *inverse* of $H((k+k_c)\Omega_1) + H((k-k_c)\Omega_1)$. Doing so, however, will also amplify any noise at frequencies where the channel attenuated the input signal $t[n]$, so a better solution is obtained by omitting the inversion at such frequencies.

For this procedure to work, the channel must be relatively low-noise, and the receiver needs to know the frequency response, $H(\Omega)$, at all the frequencies of interest in Equation (14.5); i.e., in the range $[-k_c - k_x, -k_c + k_x]$ and $[k_c - k_x, k_c + k_x]$. To estimate $H(\Omega)$, a common approach is to send a known preamble at the beginning of each packet (or frame)

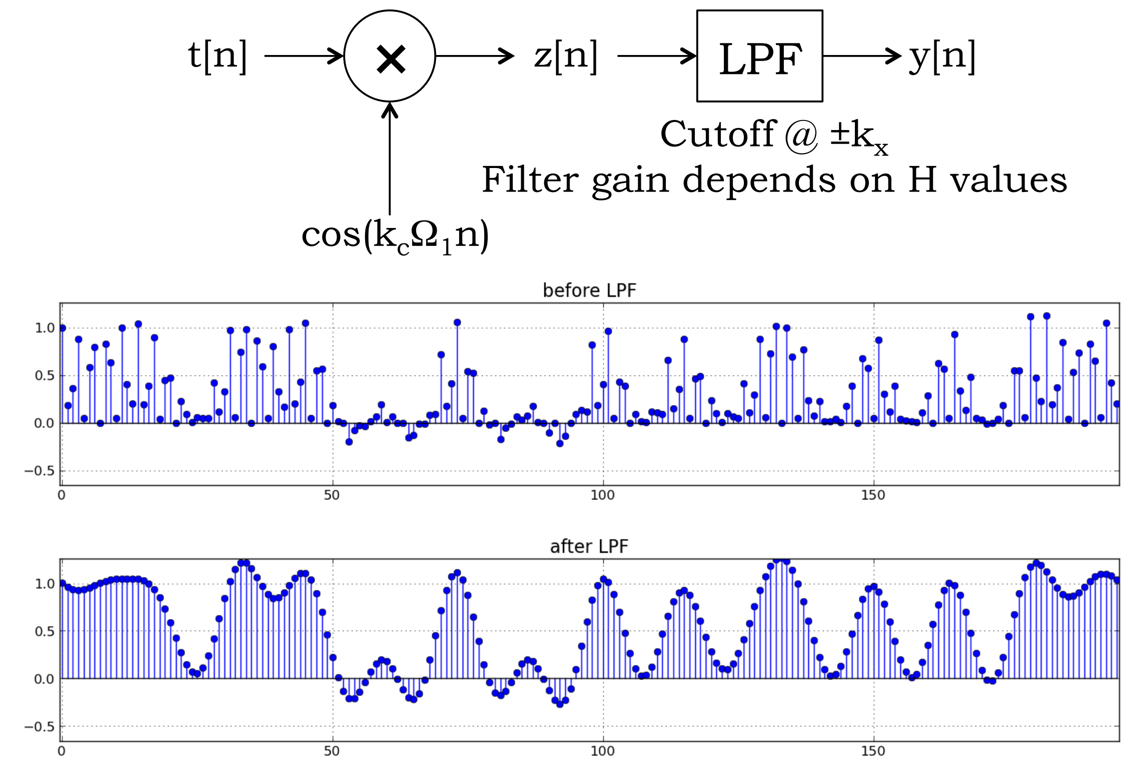


Figure 14-9: Demodulation steps: the no-delay case (top). LPF is a lowpass filter. The graphs show the time-domain representations before and after the LPF.

of transmission. The receiver looks for this known preamble to synchronize the start of reception, and because the transmitted signal pattern is known, the receiver can deduce channel's the unit sample response, $h[\cdot]$, from it, using an approach similar to the one outlined in Chapter 11. One can then apply the frequency response equation from Chapter 12, Equation (2.2), to estimate $H(\Omega)$ and use it to approximately undo the distortion introduced by the channel.

Ultimately, however, our interest is not in accurately recovering $x[n]$, but rather the underlying *bit stream*. For this task, what is required is typically not an inverse filtering operation. We instead require a filtering that produces a signal whose samples, obtained at the bit rate, allow reliable decisions regarding the corresponding bits, despite the presence of noise. The optimal filter for this task is called the **matched filter**. We leave the discussion of the matched filter to more advanced courses in communication.

■ 14.4 Handling Channel Delay: Quadrature Demodulation

We now turn to the case of channel delays between the sender and receiver. This delay matters in demodulation because we have thus far assumed that the sender and receiver have no phase difference with respect to each other. That assumption is, of course, not true, and one needs to somehow account for the phase delays.

Let us first consider the illustrative case when there is a *phase error* between the sender

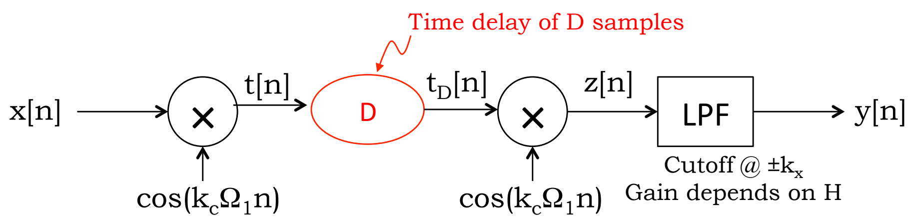


Figure 14-10: Model of channel with a delay of D samples.

and receiver. We will then show that a non-zero delay on the channel may be modeled exactly like a phase error. By "phase error", we mean that the demodulator, instead of multiplying (heterodyning) by $\cos(k_c\Omega_1 n)$, multiplies instead by $\cos(k_c\Omega_1 n - \varphi)$, where φ is some constant value. Let us understand what happens to the demodulated output in this case.

Working out the algebra, we can write

$$\begin{aligned} z[n] &= t[n]\cos(k_c\Omega_1 n - \varphi) \\ &= x[n]\cos(k_c\Omega_1 n)\cos(k_c\Omega_1 n - \varphi) \end{aligned} \tag{14.6}$$

But noting that

$$\cos(k_c\Omega_1 n)\cos(k_c\Omega_1 n - \varphi) = \frac{1}{2}\Big(\cos(2k_c\Omega_1 n - \varphi) + \cos\varphi\Big),$$

it follows that the demodulated output, after the LPF step with the suitable gains, is

$$y[n] = x[n]\cos\varphi.$$

Hence, a phase error of φ radians results in the demodulated amplitude being scaled by $\cos\varphi$. This scaling is problematic: if we were unlucky enough to have the error close to $\pi/2$, then we would see almost no output at all! And if $x[n]$ could take on both positive and negative values, then $\cos\varphi$ going negative would cause further confusion.

A channel delay between sender and receiver manifests itself as a phase error using the demodulation strategy we presented in Section 14.3. To see why, consider Figure 14-10, where we have inserted a delay of D *samples* between sender and receiver. The algebra is very similar to the phase error case: with a sample delay of D samples, we find that

$$y[n] = x[n-D]\cos(k_c\Omega_1 n) = \cos(k_c\Omega_1(n-D) + k_c\Omega_1 D).$$

The cos factor in effect looks like it has a phase error of $k_c\Omega_1 D$, so the output is *attenuated* by $\cos(k_c\Omega_1 D)$.

So how do we combat phase errors? One approach is to observe that in situations where $\cos\varphi$ is 0, $\sin\varphi$ is close to 1. So, in those cases, multiplying (heterodyning) at the demodulator by $\sin(k_c\Omega_1 n) = \cos(\frac{\pi}{2} - k_c\Omega_1 n)$ corrects for the phase difference. Notice, however, that if the phase error were non-existent, then multiplying by $\sin(k_c\Omega_1 n)$ would

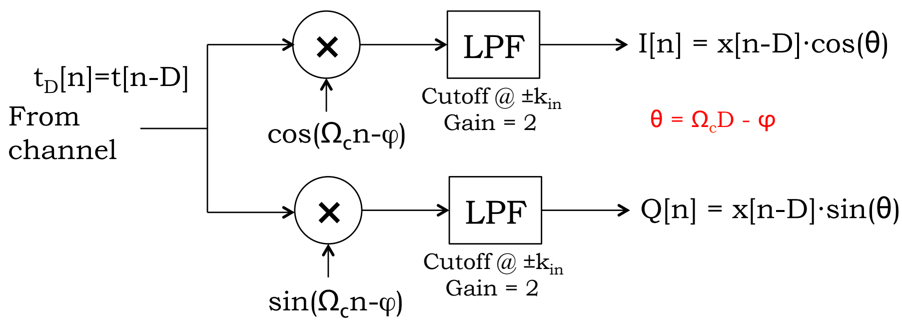


Figure 14-11: Quadrature demodulation to handle D-sample channel delay.

lead to *no baseband signal*—you should verify this fact by writing

$$z[n] = t[n]\sin(k_c\Omega_1 n) = t[n]\Big(\frac{-j}{2}e^{jk_c\Omega_1 n} + \frac{j}{2}e^{-jk_c\Omega_1 n}\Big),$$

and expanding $t[n]$ using its DTFS. Hence, multiplying by the sin when the carrier is a cos will not always work; it will work only when the phase error is a fortunate value ($\approx \pi/2$).

This observation leads us to a solution to this problem, called **quadrature demodulation**, depicted in Figure 14-11 for the case of channel delay but no channel distortion (so we can apply a gain of 2 on the LPFs rather than factors dependent on $H(\Omega)$). The idea is to multiply the received signal by *both* $\cos(\Omega_c n)$ (where $\Omega_c = k_c\Omega_1$ is the carrier frequency), *and* $\sin(\Omega_c n)$. This method is a way of "hedging" our bet: we cannot be sure which term, cos or sin would work, but we *can* be sure that they will not be 0 at the same time! We can use this fact to recover the signal reliably always, as explained below.

For simplicity (and convenience), suppose that $x[n] \geq 0$ always (at the input). Then, using the notation from Figure 14-11, define $w[n] = I[n] + jQ[n]$ (the I term is generally called the *in-phase* term and the Q term is generally called the *quadrature* term). Then,

$$\begin{aligned} |w[n]| &= \sqrt{(I[n])^2 + (Q[n])^2} \\ &= |x[n-D]|\sqrt{(cos^2\theta + \sin^2\theta)} \\ &= |x[n-D]| && (14.7) \\ &= x[n-D] \,\text{because} x[n] \geq 0 && (14.8) \end{aligned}$$

Hence, the quadrature demodulator performs the following step, in addition to the ones for the no-delay case explained before: compute $I[n]$ and $Q[n]$, and calculate $|w[n]|$ using Equation (14.8). Return this value, thresholding (to combat noise) at the mid-point between the voltage levels corresponding to a "0" and a "1". With quadrature demodulation, suppose the sender sends 0 volts for a "0" and 1 volt for a "1", the receiver would, in general, demodulate a *rotated* version in the complex plane, as shown in Figure 14-12. However, the magnitude will still be 1, and quadrature demodulation can successfully recover the input.

Figure 14-13 summarizes the various steps of the quadrature demodulator that we described in this section.

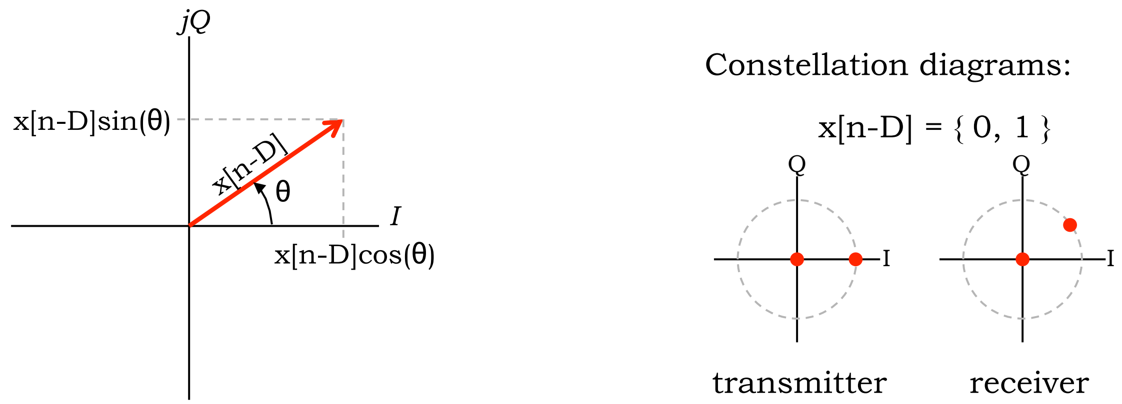


Figure 14-12: Quadrature demodulation. The term "constellation diagram" refers to the values that the sender can send, in this case just 0 and 1 volts. The receiver's steps are shown in the picture.

This concludes our discussion of the basics of demodulation. We turn next to briefly survey more sophisticated modulation/demodulation schemes.

■ 14.5 More Sophisticated (De)Modulation Schemes

We conclude this chapter by briefly outlining three more sophisticated (de)modulation schemes.

■ 14.5.1 Binary Phase Shift Keying (BPSK)

In BPSK, as shown in Figure 14-14, the transmitter selects one of two *phases* for the carrier, e.g. $-\pi/2$ for "0" and $\pi/2$ for "1". The transmitter does the same mixing with a sinusoid as explained earlier. The receiver computes the I and Q components from its received waveform, as before. This approach "almost" works, but in the presence of channel delays or phase errors, the previous strategy to recover the input does not work because we had assumed that $x[n] \geq 0$. With BPSK, $x[n]$ is either $+1$ or -1, and the two levels we wish to distinguish have the same magnitude on the complex plane after quadrature demodulation!

The solution is to think of the phase encoding as a *differential*, not absolute: a change in phase corresponds to a change in bit value. Assume that every message starts with a "0" bit. Then, the first phase change represents a $0 \rightarrow 1$ transition, the second phase change a $1 \rightarrow 0$ transition, and so on. One can then recover all the bits correctly in the demodulator using this idea, assuming no intermediate glithces (we will not worry about such glitches here, which do occur in practice and must be dealt with).

■ 14.5.2 Quadrature Phase Shift Keying (QPSK)

Quadrature Phase Shift Keying is a clever idea to add a "degree of freedom" to the system (and thereby extracting higher performance). This method, shown in Figure 14-15, uses a quadrature scheme at both the transmitter and the receiver. When mapping bits to voltage values in QPSK, we would choose the values so that the amplitude of $t[n]$ is constant.

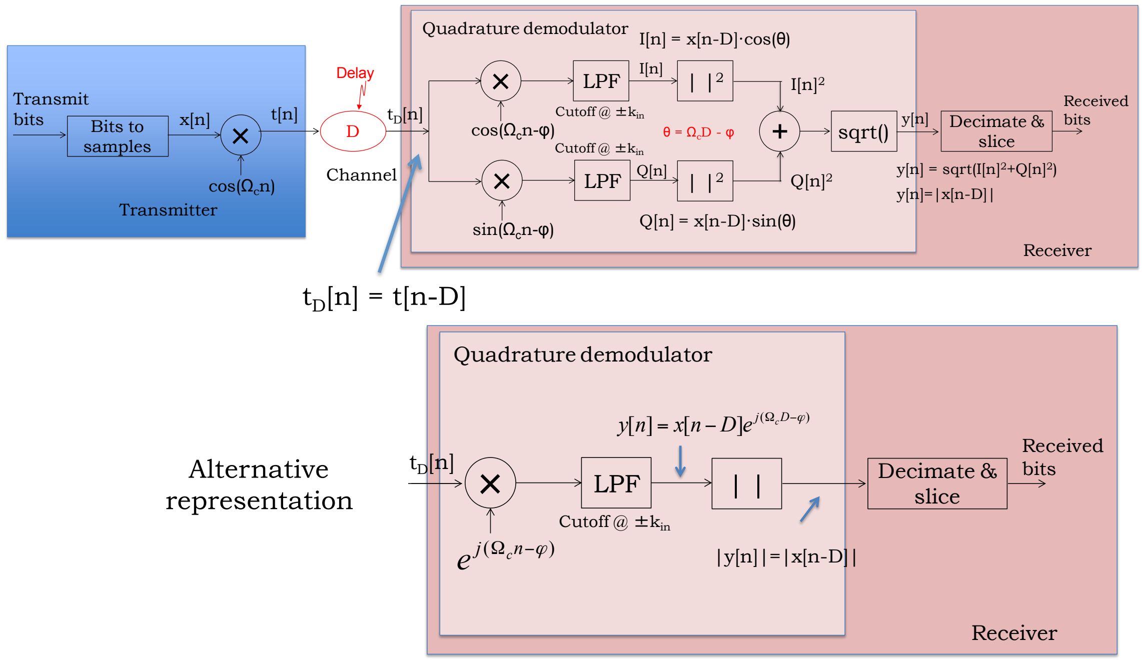


Figure 14-13: Quadrature demodulation: overall system view. The "alternative representation" shown implements the quadrature demodulator using a single complex exponential multiplication, which is a more compact representation and description.

Moreover, because the constellation now involves *four* symbols, we map two bits to each symbol. So 00 might map to (A, A), 01 to $(-A, A)$, 11 to $(-A, -A)$, and 10 to $(A, -A)$ (the amplitude is therefore $\sqrt{2}A$). There is some flexibility in this mapping, but it is not completely arbitrary; for example, we were careful here to not map 11 to $(A, -A)$ and 00 to (A, A). The reason is that any noise is more likely to cause (A, A) to be confused fo $(A, -A)$, compared to $(-A, -A)$, so we would like a symbol error to corrupt as few bits as possible.

■ 14.5.3 Quadrature Amplitude Modulation (QAM)

QAM may be viewed as a generalization of QPSK (in fact, QPSK is sometimes called QAM-4). One picks additional points in the constellation, varying both the amplitude and the phase. In QAM-16 (Figure 14-16), we map four bits per symbol. Denser QAM constellations are also possible; practical systems today use QAM-4 (QPSK), QAM-16, and QAM-64. Quadrature demodulation with the adjustment for phase is the demodulation scheme used at the receiver with QAM.

For a given transmitter power, the signal levels corresponding to different bits at the input get squeezed closer together in amplitude as one goes to constellations with more points. The resilience to noise reduces because of this reduced separation, but sophisticated coding and signal processing techniques may be brought to bear to deal with the effects of noise to achieve higher communication bit rates. In many real-world communication systems, the physical layer provides multiple possible constellations and choice of codes; for any given set of channel conditions (e.g., the noise variance, if the channel is well-described using the AWGN model), there is some combination of constellation,

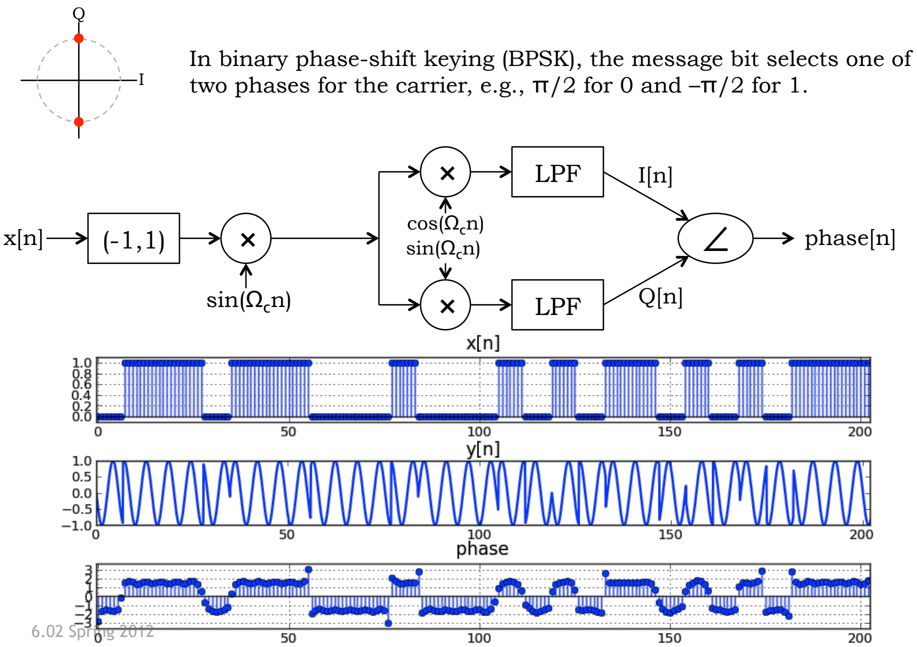


Figure 14-14: Binary Phase Shift Keying (BPSK).

coding scheme, and code rate, which maximizes the rate at which bits can be received and decoded reliably. Higher-layer "bit rate selection" protocols use information about the channel quality (signal-to-noise ratio, packet loss rate, or bit error rate) to make this decision.

■ Acknowledgments

Thanks to Mike Perrot, Charlie Sodini, Vladimir Stojanovic, and Jacob White for lectures that helped shape the topics discussed in this chapter, and to Patricia Saylor for bug fixes.

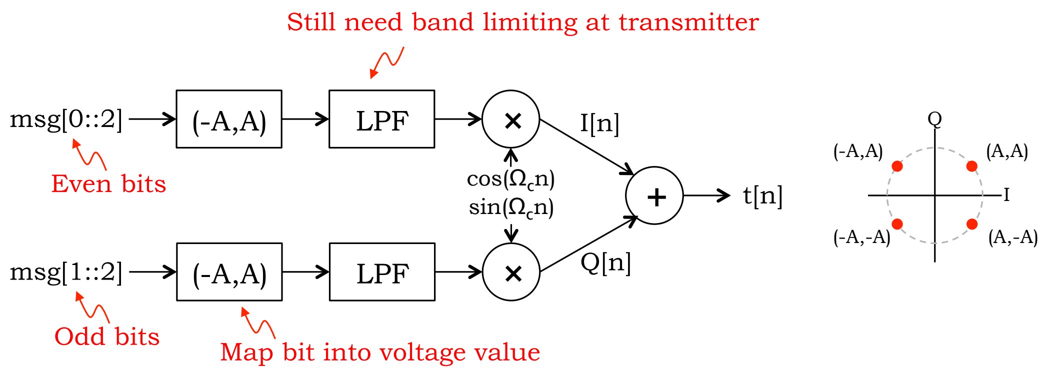


Figure 14-15: Quadrature Phase Shift Keying (QPSK).

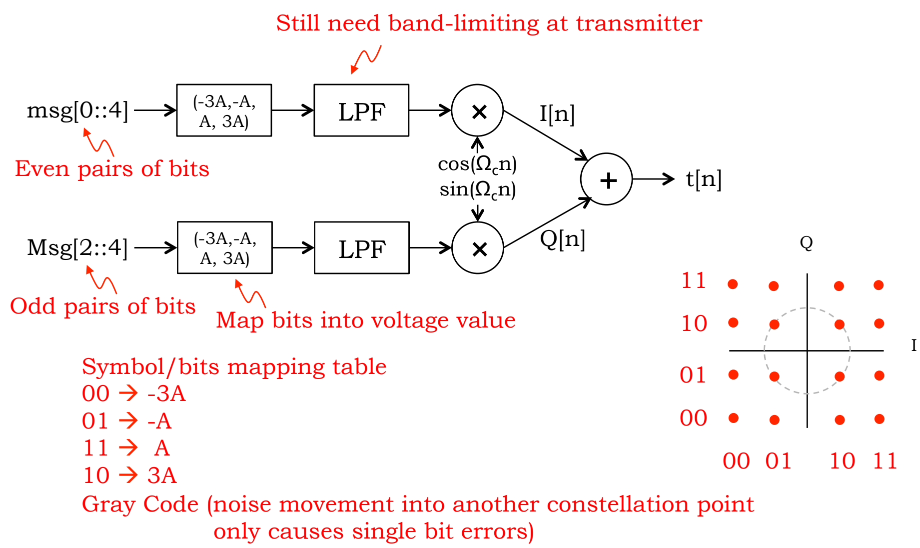


Figure 14-16: Quadrature Amplitude Modulation (QAM).

■ Problems and Questions

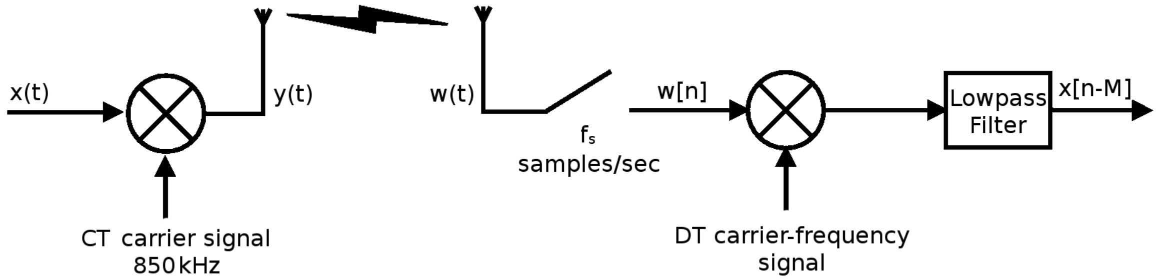
Please solve the problems at
`http://mit.edu/6.02/www/s2012/handouts/tutprobs/modulation.html`

1. The Boston sports radio station WEEI AM ("amplitude modulation") broadcasts on a carrier frequency of 850 kHz, so its continuous-time (CT) carrier signal can be taken to be $\cos(2\pi \times 850 \times 10^3 t)$, where t is measured in seconds. Denote the CT audio signal that's modulated onto this carrier by $x(t)$, so that the CT signal transmitted by the radio station is

$$y(t) = x(t)\cos(2\pi \times 850 \times 10^3 t)\ , \qquad (14.9)$$

as indicated schematically on the left side of the figure below.

We use the symbols $y[n]$ and $x[n]$ to denote the discrete-time (DT) signals that would



have been obtained by respectively sampling $y(t)$ and $x(t)$ in Equation (14.9) at f_s samples/sec; more specifically, the signals are sampled at the discrete time instants $t = n(1/f_s)$. Thus

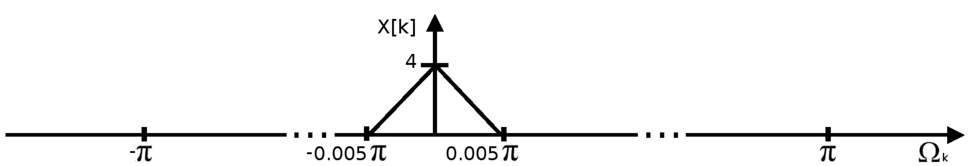
$$y[n] = x[n]\cos(\Omega_c n) \tag{14.10}$$

for an appropriately chosen value of the angular frequency Ω_c. Assume that $x[n]$ is periodic with some period N, and that $f_s = 2 \times 10^6$ samples/sec.

Answer the following questions, explaining your answers in the space provided.

(a) Determine the value of Ω_c in Equation (14.10), restricting your answer to a value in the range $[-\pi, \pi]$. (You can assume in what follows that the period N of $x[n]$ is such that $\Omega_c = 2k_c\pi/N$ for some integer k_c; this is a detail, and needn't concern you unduly.)

(b) Suppose the Fourier series coefficients $X[k]$ of the DT signal $x[n]$ in Equation (14.10) are *purely real*, and are as shown in the figure below, plotted as a function of $\Omega_k = 2k\pi/N$. (Note that the figure is not drawn to scale. Also, the different values of Ω_k are so close to each other that we have just interpolated adjacent values of $X[k]$ with a straight line, rather than showing you a discrete "stem" plot.) Observe that the Fourier series coefficients are non-zero for frequencies Ω_k in the interval $[-.005\pi, .005\pi]$, and 0 at all other Ω_k in the interval $[-\pi, \pi]$.



Draw a carefully labeled sketch below (though not necessarily to scale) to show the Fourier series coefficients of the DT modulated signal $y[n]$. However, rather than labeling your horizontal axis with the Ω_k, as we have done above, you should label the axis with the appropriate frequency f_k in Hz.

Assume now that the receiver detects the CT signal $w(t) = 10^{-3}y(t - t_0)$, where $t_0 = 3 \times 10^{-6}$ sec, and that it samples this signal at f_s samples/sec, thereby obtaining the

DT signal

$$w[n] = 10^{-3}y[n-M] = 10^{-3}x[n-M]\cos\Big(\Omega_c(n-M)\Big) \qquad (14.11)$$

for an appropriately chosen integer M.

C. Determine the value of M in Equation (14.11).

D. Noting your answer from part **B**, determine for precisely which intervals of the frequency axis the Fourier series coefficients of the signal $y[n-M]$ in Equation (14.11) are non-zero. You **need not find the actual coefficients**, only the frequency range over which these coefficients will be non-zero. Also **state whether or not the Fourier coefficients will be real**. Explain your answer.

E. The demodulation step to obtain the DT signal $x[n-M]$ from the received signal $w[n]$ now involves multiplying $w[n]$ by a DT carrier-frequency signal, followed by appropriate low-pass filtering (with the gain of the low-pass filter in its passband being chosen to scale the signal to whatever amplitude is desired). Which one of the following six DT carrier-frequency signals would you choose to multiply the received signal by? Circle your choice and give a brief explanation.

i. $\cos\Big(\Omega_c n\Big)$.

ii. $\cos\Big(\Omega_c(n-M)\Big)$.

iii. $\cos\Big(\Omega_c(n+M)\Big)$.

iv. $\sin\Big(\Omega_c n\Big)$.

v. $\sin\Big(\Omega_c(n-M)\Big)$.

vi. $\sin\Big(\Omega_c(n+M)\Big)$.

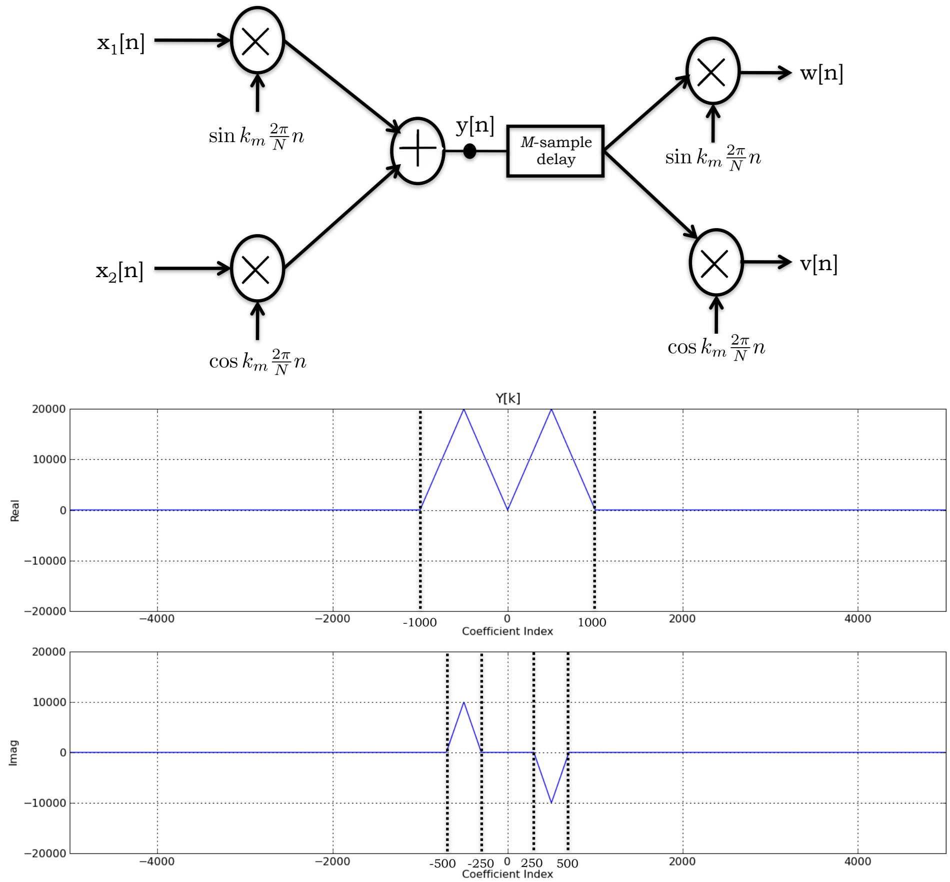


Figure 14-17: System for problem 2.

2. All parts of this question pertain to the following modulation-demodulation system shown in Figure 14-17, where all signals are periodic with period $P = 10000$. Please also assume that the sample rate associated with this system is 10000 samples per second, so that k is both a coefficient index and a frequency. In the diagram, the modulation frequency, k_m, is 500.

   (a) Suppose the DFT coefficients for the signal $y[n]$ in the modulation/demodulation diagram are as plotted in Figure 14-17.

   Assuming that $M = 0$ for the M-sample delay (no delay), plot the coefficients for the signals **w** and **v** in the modulation/demodulation diagram. Be sure to label key features such as values and coefficient indices for peaks.

   (b) Assuming the coefficients for the signal $y[n]$ are the same as in part (a), please plot the DTFS coefficients for the signal x_1 in the modulation/demodulation diagram. Be sure to label key features such as values and coefficient indices for peaks.

(c) If the M-sample delay in the modulation/demodulation diagram has the right number of samples of delay, then it will be possible to nearly perfectly recover $x_2[n]$ by low-pass filtering $w[n]$. Determine the smallest positive number of samples of delay that are needed and the cut-off frequency for the low-pass filter. Explain your answer, using pictures if appropriate.

3. Figure 14-18 shows a standard modulation/demodulation scheme where $N = 100$.

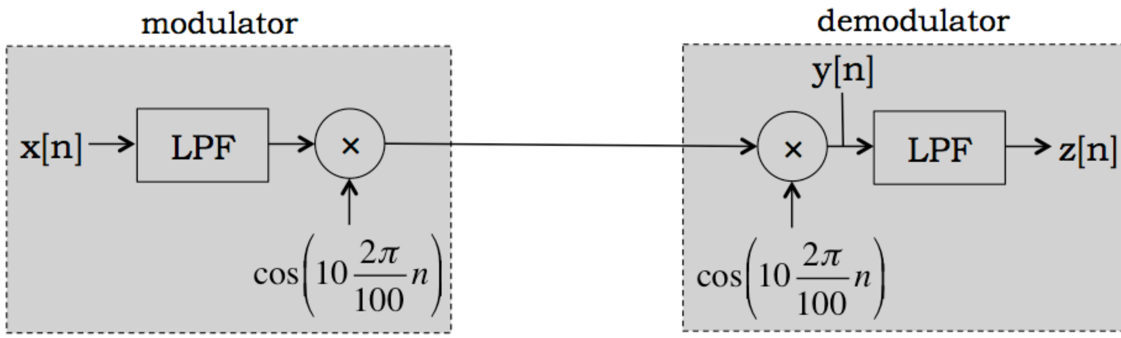


Figure 14-18: System for problem 3.

(a) Figure 14-19 shows a plot of the input, $x[n]$. Please draw the approximate time-domain waveform for $y[n]$, the signal that is the input to the low-pass filter in the demodulator. Don't bother drawing dots for each sample, just use a line plot to indicate the important timing characteristics of the waveform.

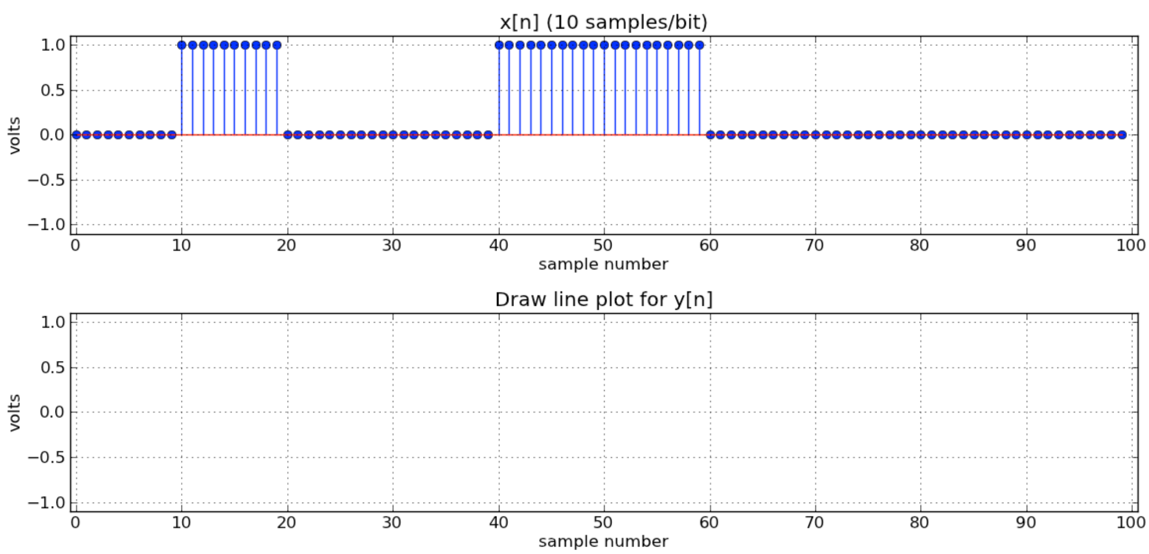


Figure 14-19: Plot for problem 3(a).

(b) Building on the scheme shown in Part (a), suppose there are multiple modulators and demodulators all connected to a single shared channel, with each modulator given a different modulation frequency. If the low-pass filter in each modulator is eliminated, briefly describe what the effect will be on signal $z[n]$, the output of a demodulator tuned to the frequency of a particular transmitter.

4. The plot on the left of Figure 14-20 shows a_k, the DTFS coefficients of the signal at the output of a transmitter with $N = 36$. If the channel introduces a 3-sample delay, please plot the Fourier series coefficients of the signal entering the receiver.

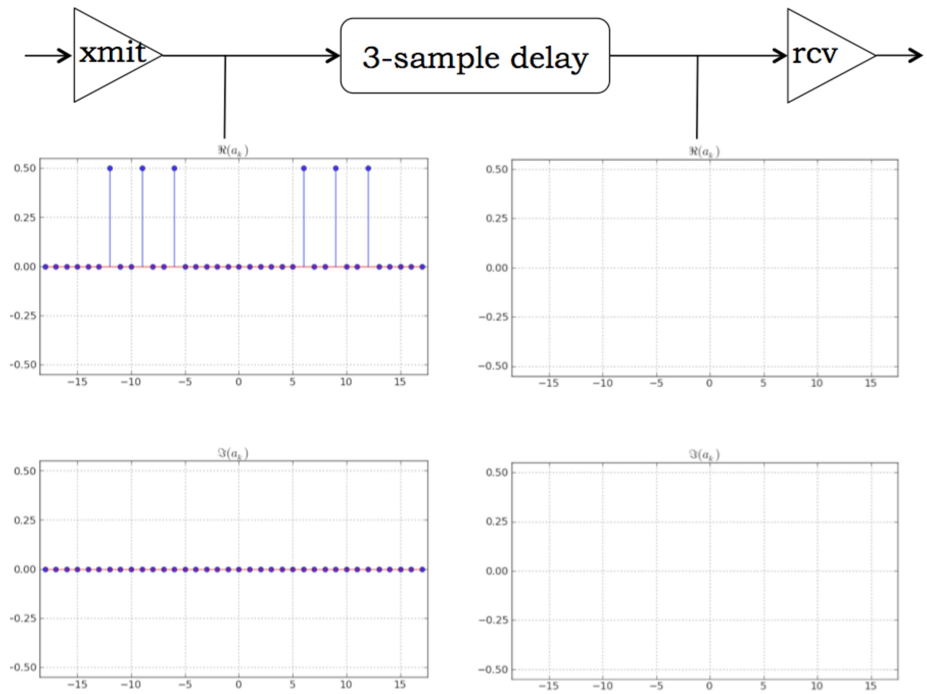


Figure 14-20: System for problem 4.

5. Figure 14-21 shows an *image rejection mixer*. The frequency responses of the two filter components (the 90-degree phase shift and the low-pass filter) are as shown. The spectral plot to the left in figure above shows the spectrum of the input signal, x[n]. Using the same icon representation of a spectrum, draw the spectrum for signals $p[n], q[n], r[n]$, and $s[n]$ below, taking care to label the center frequency and magnitude of each spectral component. If two different icons overlap, simply draw them on top of one another. If identical icons overlap, perform the indicated addition/subtraction, showing the net result with a bold line.

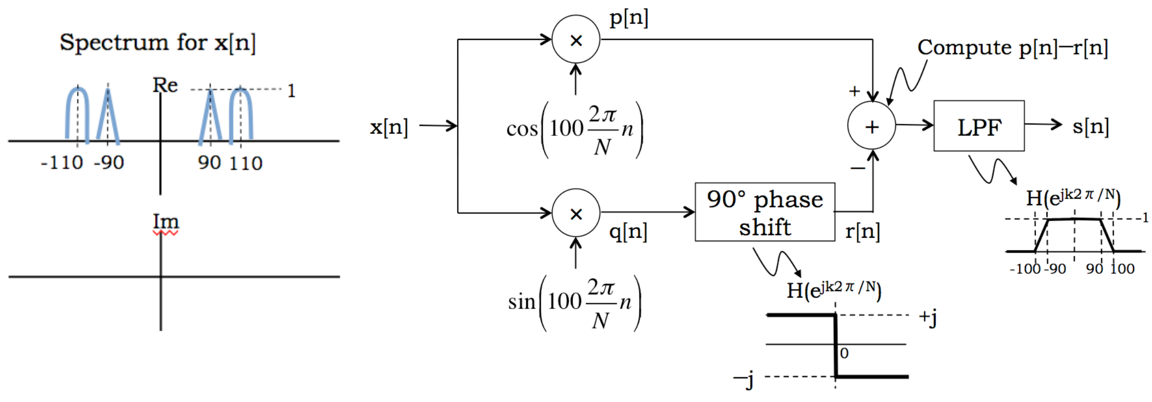


Figure 14-21: Problem 5: image rejection mixer.