MAS.160 / MAS.510 / MAS.511 Signals, Systems and Information for Media Technology
Fall 2007

MAS 160/510 Additional Notes: Modulation

From Amplitude Modulation to Frequency Modulation

As usually implemented, FM uses much more bandwidth than AM. You'll note, for instance, that FM radio stations in the US are spaced 200KHz apart, while AM stations are spaced only 10KHz apart. So why would one want to use FM? Among other interesting features, it allows signal-to-noise ratio, or SNR, to be traded off for bandwidth.

We've already seen that amplitude modulation simply requires us to multiply our input signal $f(t)$ by a "carrier" sinusoid $c(t) = \cos \omega_c t$, where ω_c (or more precisely $\omega_c/2\pi$ Hertz) is the frequency to which you tune your radio dial in order to receive this signal. In order to simplify receiver design, actual AM broadcasts are of the form

$$s(t) = K[1 + mf(t)] \cos \omega_c t,$$

where m is chosen so that $|mf(t)| < 1$. This causes the "envelope" of the signal to follow the shape of the input, preventing a negative $f(t)$ from flipping the phase of the sinusoid. This permits demodulating the signal very simply: a narrow bandpass filter is tuned to the frequency of the desired station, its output goes to a nonlinear device called a rectifier which removes the negative portion of the signal, and a lowpass filter then essentially connects the peaks to recover the envelope, which is $(1 + mf(t))$.[1] Such a receiver requires less precision than synchronous demodulation, or multiplying the received signal by a phase-locked sinusoid to shift a spectral replica back down so that it centers on zero frequency (this method is used, however, in some of the digital modulation methods we will examine later).

We can rethink the formulation of our sinusoidal carrier $c(t)$. Let $\theta(t) = \omega_c t$. In this case, $\theta(t)$ is a linear function of time, and ω_c is its derivative. But in other kinds of modulation $\theta(t)$ won't be linear with time, and we can't think about frequency as we're accustomed to do. Thus we need to define something called *instantaneous frequency* ω_i as the derivative of the angle:

$$c(t) = \cos \theta(t), \quad \omega_i = \frac{d\theta}{dt}.$$

In FM, we want ω_i to vary linearly with the modulating signal $f(t)$. Therefore,

$$\omega_i = \omega_c + Kf(t),$$

which implies that

$$\theta(t) = \int \omega_i dt = \omega_c t + K \int f(t) dt.$$

The analysis of FM is far harder than that for AM, as superposition doesn't hold. It's typical, nevertheless, to consider what happens when our $f(t)$ is a sinusoid:

$$f(t) = a \cos \omega_m t.$$

Now

$$\omega_i = \omega_c + \Delta\omega \cos \omega_m t, \quad \Delta\omega \ll \omega_c.$$

[1] If you've ever built an AM receiver, you will justifiably charge us with gross oversimplification, but the basic idea is correct.

Then

$$s(t) = \cos(\omega_c t + \beta \sin \omega_m t),$$

where we call β the *modulation index* and define it as the ratio of the maximum frequency deviation to the bandwidth of $f(t)$:

$$\beta \equiv \frac{\Delta\omega}{\Delta\omega_m}.$$

What we're going to investigate is how the bandwidth of the signal $s(t)$ depends on β.

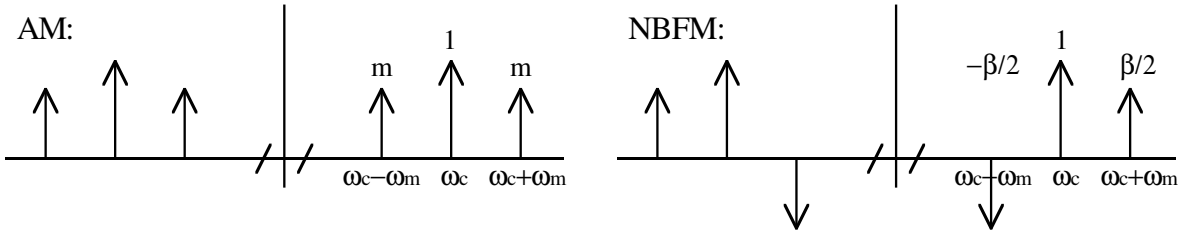

Figure 1: Spectra for AM and NBFM, given a modulating signal that is a single sinusoid.

Consider first the expansion of the above $s(t)$:

$$s(t) = \cos \omega_c t \cos(\beta \sin \omega_m t) - \sin \omega_c t \sin(\beta \sin \omega_m t).$$

If $\beta \ll \pi/2$, which is called *narrowband FM*,

$$\cos(\beta \sin \omega_m t) \approx 1,$$

and

$$\sin(\beta \sin \omega_m t) \approx \beta \sin \omega_m t$$

so

$$s(t) \approx \cos \omega_c t - \beta \sin \omega_m t \sin \omega_c t.$$

If you consider how we got here, you should be able to see that for small β, for *any* $f(t)$,

$$s_{NBFM}(t) \approx \cos \omega_c t - K \int f(t)dt \sin \omega_c t.$$

If we recall that

$$s_{AM}(t) = \cos \omega_c t + m f(t) \cos \omega_c t,$$

we can see that narrowband FM of a sinusoidal $f(t)$ is very similar to AM except that the sidebands are $\pi/2$ radians out of phase with the carrier. The bandwidth is essentially the same. If $f(t) = \cos \omega_m t$ the spectra look like the illustration in Figure 1. Possible systems for generating each are shown in Figure 2.

But NBFM is more complicated and doesn't appear to offer us any real advantages. Let's now consider *wideband FM* ($\beta > \pi/2$).

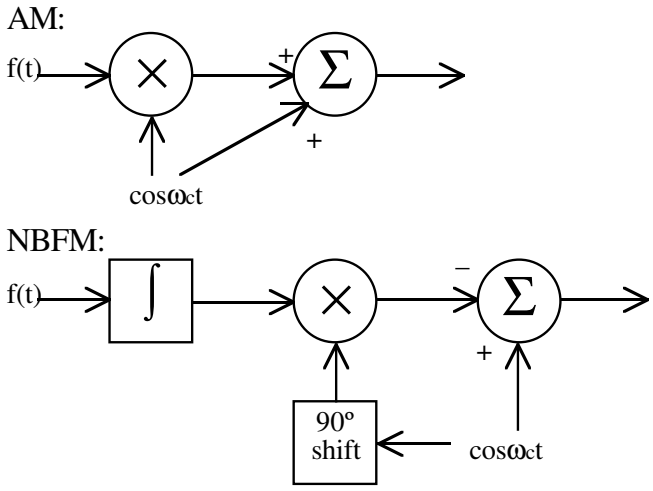

Figure 2: Modulators for AM and NBFM.

When $\beta \not\ll \pi/2$ the approximation we did above doesn't hold. To understand what happens as β increases, it's usual to expand $s(t)$ into a power series and to retain all the significant terms. See Schwartz's book, referenced at the end of these notes, for more details. We'll let an illustration suffice. For a sinusoidal modulating signal again, with $\beta \approx 2$, we get a spectrum as in Figure 3.

The bandwidth of FM is, strictly speaking, infinite. But since the terms far away from the carrier are very, very small, they can usually be ignored. A common rule of thumb is to say that if the maximum frequency in $f(t)$ is B, then the approximate FM bandwidth is $2B(1+\beta)$. For broadcast FM radio, $\beta = 5$ and B is 15KHz, giving us a bandwidth of 180KHz, which corresponds well with the 200KHz channel spacing.

Given that an integrator was used in the generation of the signal, you shouldn't be surprised that we use a differentiator to recover it. Since the gain of a differentiator varies linearly with frequency, the output is the input signal with its amplitude (or envelope) varying as the modulating input $f(t)$. Then we can just use an envelope detector (as in AM) to get back $f(t)$, as in Figure 4.

Incidentally, there is an important theorem called Logan's Theorem[2] that applies to FM. It states that if the bandwidth of a signal is less than an octave, the signal may be recovered exactly (except for a multiplicative constant) from its zero-crossings.[3] Analog laser videodiscs work in this fashion, as the video signal is FM modulated and the spacing of the pits on the disc records the position of the zero-crossings of the FM signal.

Digital Modulation – PSK, QAM

[2]B. F. Logan, "Information in the Zero Crossings of Bandpass Signals," *Bell Sys. Tech. J., 56,* pp. 487-510, April 1977.

[3]There is also a requirement that the signal have no zeros in common with its Hilbert transform, among other provisos not important here.

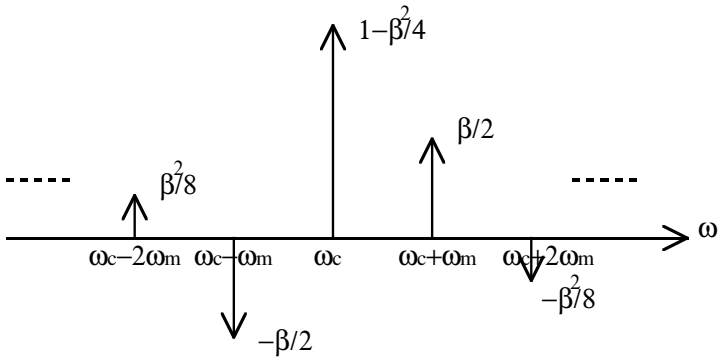

Figure 3: Spectrum for FM, with a modulating signal of a single sinusoid.

We have only a brief amount of time and space to dedicate to the important topic of digital modulation, so we will concentrate on some basic methods. Readers with greater interest in this topic should consult the additional references cited below.

The simplest method we will examine is Phase-Shift Keying, or PSK.[4] Consider a system in which we have two carriers out of phase by $\pi/2$ radians, in other words *in quadrature* with each other. We could then send two bits at a time by converting our pair of bits to two signals $i(t)$ and $q(t)$ which multiply the cosine and sine components respectively:

bit pair	i	q
(0,0)	1	0
(0,1)	0	1
(1,0)	-1	0
(1,1)	0	-1

We call the group of bits transmitted at one time a "symbol." See Figure 5. You can also think about this as a complex multiplication, and the results are often illustrated that way.

A plot of the possible signals (called a *constellation*) is shown in Figure 6.

This system is called QPSK (for Quaternary PSK), or 4-PSK. More bits can be sent at once by selecting more closely-spaced points on the unit circle and multiplying the carriers by the sines and cosines of those angles, though the noise immunity decreases as points are added. Decoding a PSK signal requires a stable phase reference, which may be difficult to generate, so a variant is Differential QPSK, or DQPSK. Here the phase reference is the just-transmitted bit pair, and each bit pair *shifts* the phase from the preceding one. So (0,0) is sent as a shift of $\pi/4$ radians, (0,1) shifts by $3\pi/4$, et cetera (the $\pi/4$ spacing assures that

[4]The word "keying" hearkens back to the days when the only binary information on the radio was Morse Code from radio telegraph keys! Who says there is no respect for tradition in this fast-moving field?

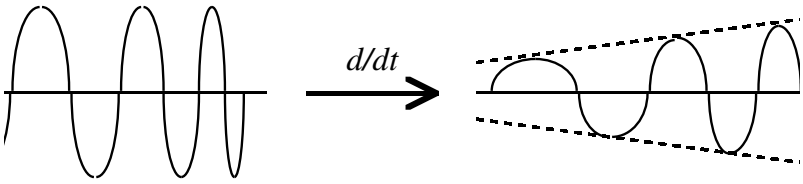

Figure 4: Effect of a differentiator on an FM signal.

there is *always* a phase shift from bit pair to bit pair).

Since maximum noise immunity requires as much distance as possible (for a given amount of signal power) between adjacent points in the constellation, we should be able to do better if we change both amplitude and phase of the carriers. This is called APK, or Amplitude-Phase Keying; QAM, or Quadrature Amplitude Modulation is a very common special case.[5]

Suppose we take the preceding modulator and stipulate that we send four bits at once with two each going to generate an i and a q that can each take one of four values. Then our 16-QAM constellation looks like a square grid, as in Figure 7.

In practice, we need to lowpass filter $i(t)$ and $q(t)$, since if the bit stream has square pulses, the bandwidth is essentially infinite no matter what the bit rate. With proper bandlimiting 16-QAM can carry up to four bits per second per Hertz of channel bandwidth (so we could send four megabits per second in a 1MHz channel).

In the most general case, APK needn't be on a square array. Triangular packing maximizes Euclidean distance between points, and is how 64-APK is usually implemented.

Digital Modulation – CDMA, OFDM

Spread-spectrum techniques, just like FM, modulate a signal in a way that occupies more bandwidth than the minimum needed to transmit the signal. They do this by shifting the carrier frequency as a function of time. Advantages include increased security, less sensitivity to interference at a fixed frequency, and less sensitivity to transmission channels that may have reduced frequency response in a particular narrow frequency range.

[5]QAM shows up in the analog world as well, where it is used for the color-difference signals in NTSC and PAL color television. It is no mere coincidence that the NTSC color-difference signals are called i and q!

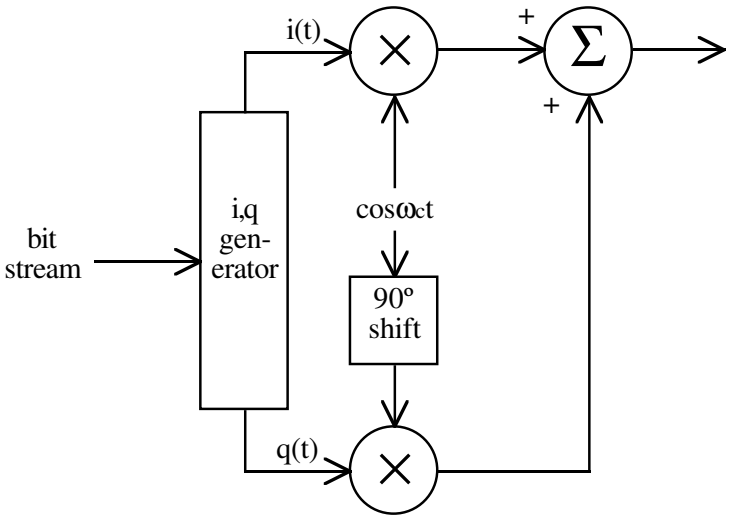

Figure 5: Digital quadrature modulator.

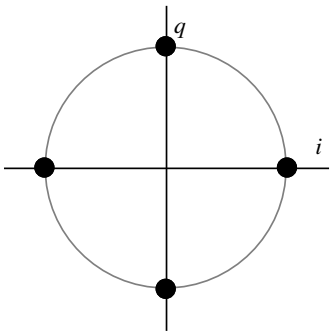

Figure 6: Constellation for 4-PSK.

The first such method was invented and patented in 1942 by actress Hedy Lamarr and composer George Antheil, and is nowadays known as frequency-hopping code division multiple access, or FH-CDMA. A transmitter and receiver are synchronized, and the "hoppping" from frequency to frequency takes place many times a second under control of an algorithm such as a pseudo-random sequence.

A related technique, direct sequence CDMA (or DS-CDMA), combines the data with a spreading sequence (or "chipping code") which divides the data among the carriers with a degree of redundancy (similarly to the error-correcting codes we examine in this class). This adds an additional layer of robustness against data loss.

It's usually the case that an RF signal will be received with several trailing echoes, as the signal bounces off buildings or large geographical features. This effect, called multipath, is the source of "ghosts" on an analog television picture, and can also be thought of as the

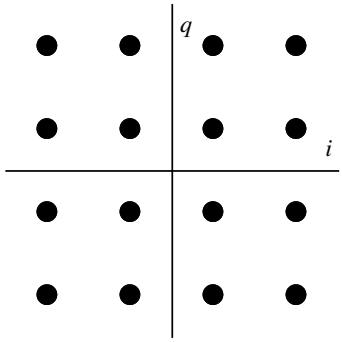

Figure 7: Constellation for 16-QAM.

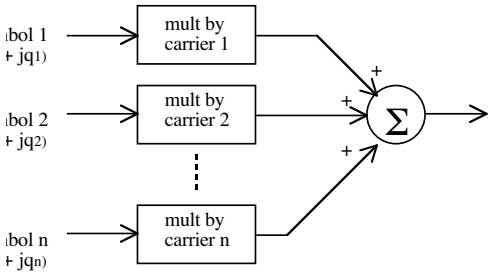

Figure 8: Orthogonal frequency-division multiplexing.

convolution of the signal with an impulse response having a series of weighted delays. In a digital modulation system, if the delays are on the order of the length of time it takes to transmit a symbol, we can get intersymbol interference, or ISI, in which a delayed symbol arrives at the same time as a symbol on the direct path. While it's possible to detect the echoes and equalize for them by deconvolving at the receiver, greater immunity to ISI comes at the cost of reducing the symbol rate. Orthogonal frequency-division multiplexing (OFDM) methods do this by performing simultaneous digital modulation onto a large number of closely-spaced carriers, each sending one symbol at a very low symbol rate (and thus a low bandwidth) as in Figure 8. Thus if there are n carriers, each modulator takes every n-th symbol. As shown in Figure 9, the digital signals are rectangular pulses (not bandlimited) in the time domain and thus sinc-shaped in the frequency domain. Appropriate spacing of the carriers makes them orthogonal and thus there is no interference between them (like FM, the bandwidth is strictly speaking infinite). Note that multiplying a set of complex numbers by a set of orthogonally-spaced sinusoids and summing their products is the same as an inverse DFT, and indeed this system is generally implemented by performing an IFFT (of as many points as the number of carriers) in the modulator and then converting from digital to analog just before transmitting. Similarly the demodulation is done using

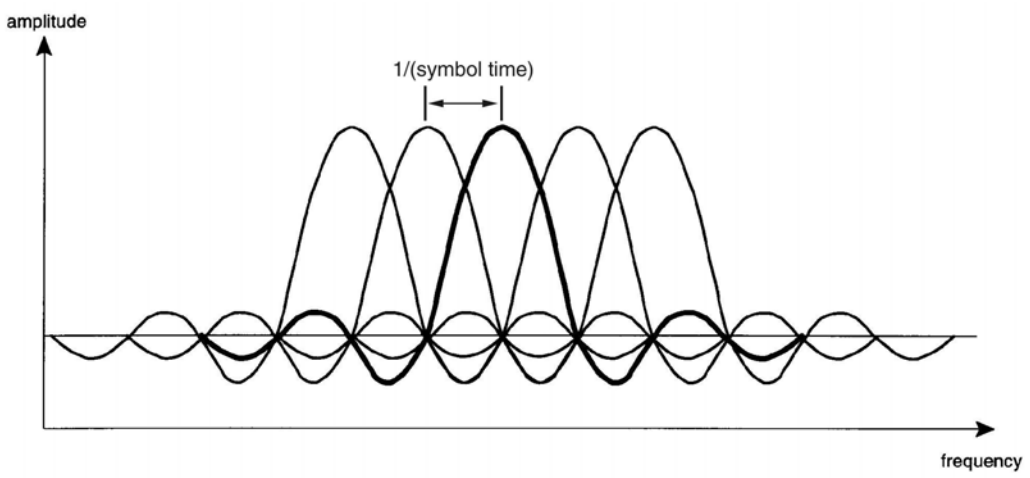

Figure 9: If COFDM signals are rectangular pulses in time, they are sinc-shaped in frequency. If the carriers are spaced orthogonally, there is no interference from one carrier to another. (Adapted from Maddocks, 1993)

an FFT. The result is a system with the same spectrum efficiency as the basic modulation method used on each carrier, but with much more immunity to multipath and interference. If the data have an error correcting code applied before modulation, the system is called coded OFDM, or COFDM.

Additional Reading

B. E. Keiser, *Broadband Coding, Modulation, and Transmission Engineering*, Prentice-Hall, 1989.

M. C. D. Maddocks, *An Introduction to Digital Modulation and OFDM Techniques,* BBC Research Department Report BBC RD 1993/10, 1993.
(http://www.bbc.co.uk/rd/pubs/reports/1993-10.pdf)

H. K. Markey and G. Antheil, US Patent 2,292,387, Aug. 11, 1942.

M. Schwartz, *Information Transmission, Modulation, and Noise*, McGraw-Hill, 1990.