By H. Ward Silver, NØAX

About FM

If you learn its fundamentals, you can demystify FM's seemingly complex behavior.

The first practical frequency modulation (FM) systems were developed by Major Edwin Howard Armstrong, the prolific radio inventor, in the early 1930s. They offered immediate advantages over competing AM systems—higher fidelity and lower noise—advantages that are still true today.[1] As a result, FM systems are in wide use by people around the world, not only radio amateurs, as versatile communications and broadcast transmission mediums. It is the rare ham that has not used a portable or handheld transceiver on a repeater or a simplex frequency—many amateurs use it as their exclusive communications mode.

Even though FM is widely used, it is often poorly understood. The basic idea is straightforward—make the RF signal's frequency, rather than its amplitude, rise and fall with the modulating signal. One can quite easily imagine the result—as the amplitude of the message changes, the RF signal's frequency changes—tracking each peak and valley as shown in Figure 1. The creation of this signal requires an interesting and surprising bag of tricks, as we shall see.

Some Useful Concepts

Instantaneous Frequency and Deviation

Let's start by learning the appropriate terminology so that we can be precise in our discussion. The frequency of the FM signal at any instant in time is called the *instantaneous frequency*. The variations back and forth around the carrier frequency are known as *deviation*.

FM is one of two types of *angle modulation*. The other type is known as *phase modulation* (or PM). Both modulation types are in use by amateurs—they create similar on-the-air signals, and they can be received by the same equipment. The difference between FM and PM is that while the deviation of an FM signal depends *only* on the amplitude of the message signal, the deviation of a PM signal depends on *both* the amplitude and frequency of the message.[2]

MARVIN COLLINS, W6OQI

Modulation Index

Just as amplitude modulation (AM) has a *modulation index* that measures the degree to which the message is modulating the RF signal, so does angle modulation. For AM, the index measures the amplitude relationship between the single pair of sidebands and the carrier. Angle-modulated signals have more than one set of sidebands—two, three, five or more, depending on how the phase of the carrier is altered by the message signal.

But doesn't the modulation cause a frequency shift in an FM or PM signal? Yes, because changing a signal's phase is effectively the same as changing its frequency during that same period. And, vice versa, changing a signal's frequency can be thought of as shifting its phase. There is an excellent description of the relationship between frequency and phase in the reference cited in Note 2.

The modulation index, m, for angle modulation, is a measure of the maximum amount of phase change the message signal can cause—more phase change (a larger value of m) results in more sidebands. For our single-tone message:

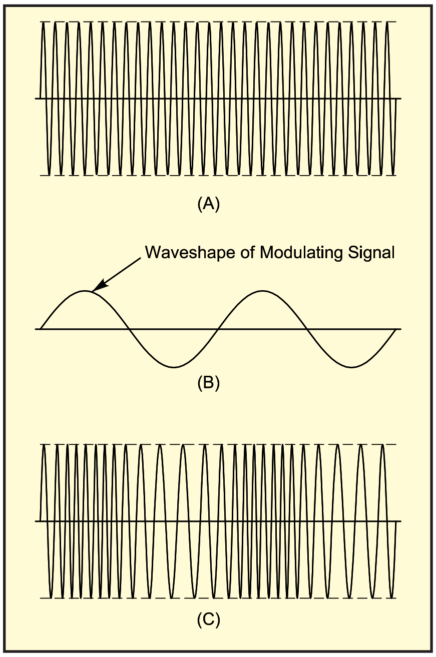

Figure 1—Graphical representation of frequency modulation. In the unmodulated carrier (A) each RF cycle occupies the same amount of time. When the modulating signal (B) is applied, the radio frequency is increased and decreased according to the amplitude and polarity of the modulating signal (C).

[1] Notes appear on page 42.

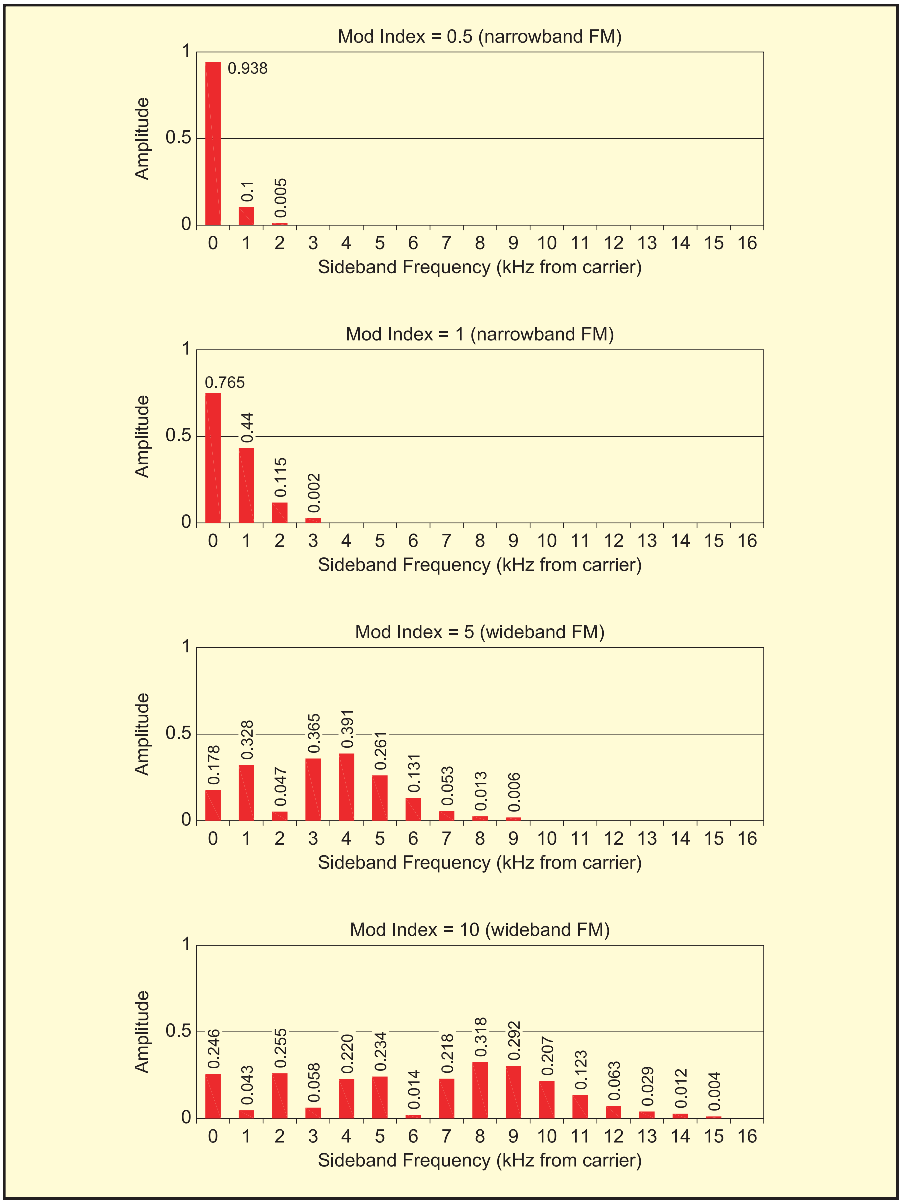

Figure 2—These graphs show the carrier and one side of an FM signal with a 1 kHz message signal at different modulation indexes. Note how the sidebands increase *and* decrease as modulation index changes. All of these spectrums have the *same total power*.

$m = A \times f_d / f_M$ (for FM)

$m = (\phi_{MAX}) = k_p \times A$ (for PM)

where:

- m is calculated in radians and there are 180/π radians in the 360° of one complete cycle of a sine wave. (1 radian ≈ 57.3°)
- A is the amplitude of the message signal in volts.
- f_M is the frequency of the message signal in herts.
- f_d is the *frequency deviation constant* that represents the sensitivity of the modulator in hertz of deviation per volt of the message signal.
- $A \times f_d$ is called the *peak deviation*.
- ϕ_{MAX} is the maximum value of phase change caused by the message signal.
- k_p is the *phase deviation constant* and is similar to f_d in that it specifies the sensitivity of the phase modulator in radians of phase change per volt of the message signal.

For PM, m doesn't depend on message frequency at all. For an FM signal, m will be larger if the peak deviation gets larger or if f_M gets smaller. For example, loud low-frequency signals can cause m to become quite large unless the transmitter limits deviation and microphone gain or frequency response.

Bandwidth

FM signals are classified as narrowband (m≤1) or wideband (m>1). Amateur and commercial mobile services use narrow-band FM to preserve power and spectrum space, sacrificing fidelity and message signal bandwidth. Commercial broadcast FM is wideband, delivering high-fidelity entertainment-quality signals with high power transmitters in 150-kHz-wide channels.

Note that nothing changes the power of the signal: *Angle-modulated signals are constant power signals*—regardless of the amplitude or frequency of the message signal. That's why your power meter doesn't change whether you're speaking softly, loudly or not at all! Power amplifiers for FM don't have to be linear, either, since there are no amplitude variations to preserve. The frequency of the signal is all that matters. The amplifier can be designed for optimum efficiency instead of linearity. That's what the SSB/FM switch on a VHF amplifier does—it changes the amplifier's operating point from one that optimizes linearity to another that optimizes efficiency.

Bessel Functions

The modulation index, m, is the key to determining the overall shape of an FM or PM signal. This shape—which specifies the amplitude of each sideband—is described mathematically by *Bessel* functions. Technically, there are an infinite number of sidebands in an angle-modulated signal, but the amplitudes of sidebands far from the carrier are insignificant and can be ignored. Narrowband angle-modulated signals are usually considered to have up to three sets of sidebands with useful power.

As m increases, so does the number of sidebands, but their relative amplitudes do not always increase. In fact, at certain values of m, some of the sidebands or the carrier disappear completely. For example, the carrier disappears when m equals 2.405, 5.52 and 8.654 (and other values). The first sideband will disappear when m equals 3.85, and so on.

For the simple case of a message consisting of a single tone, all of the sidebands are separated from the carrier by integral multiples of the message signal frequency. Figure 2 shows what the sidebands on *one* side of the carrier look like for different values of m and a 1 kHz message signal. (There is a symmetric set on the other side.) The component at 0 kHz offset represents the carrier. It's important to remember that all four examples have the same amount of *total* power—it's just distributed differently as m changes.[3]

Obviously, one can't be looking at tables of Bessel functions to determine signal bandwidth on the air. Certain conventions and calculation shortcuts are used instead. The Amateur and Land Mobile services have settled on 5 kHz of deviation as providing a good compromise between bandwidth, fidelity, noise and power requirements.

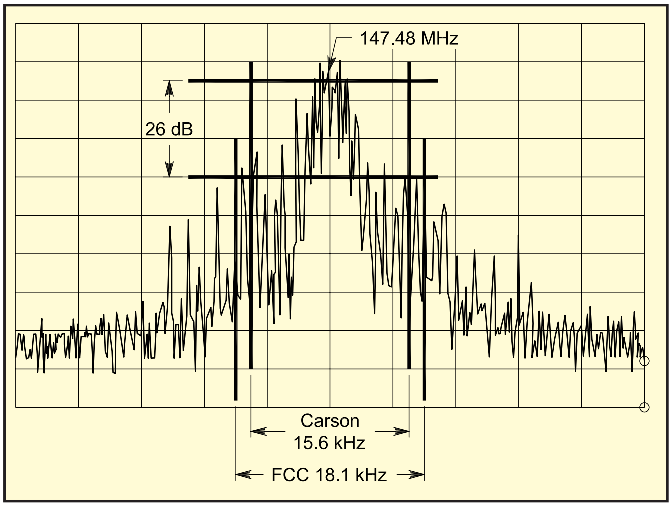

Figure 3—This is a spectrum analyzer display of a typical 2 meter VHF FM repeater. The bandwidths and amplitudes for Carson's Rule and the FCC Part 97.307 bandwidth definition are overlaid on the spectrum.

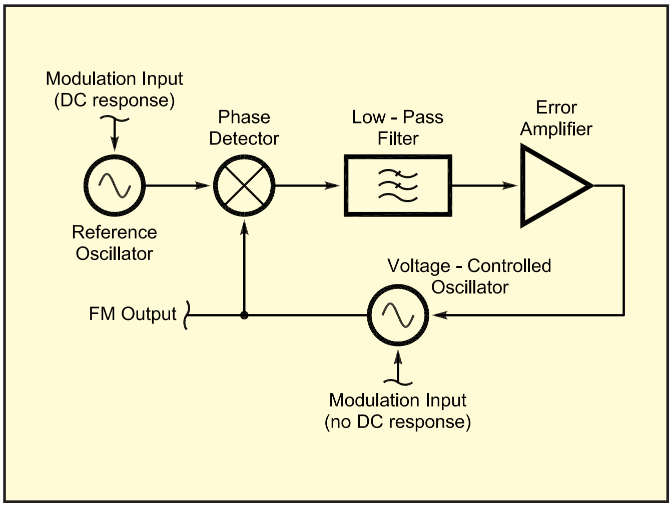

Figure 4—A phase-locked loop (PLL) can be used to generate FM by modulating either the reference oscillator or the voltage-controlled oscillator (VCO). To preserve a dc level in the message signal, the reference must be modulated.

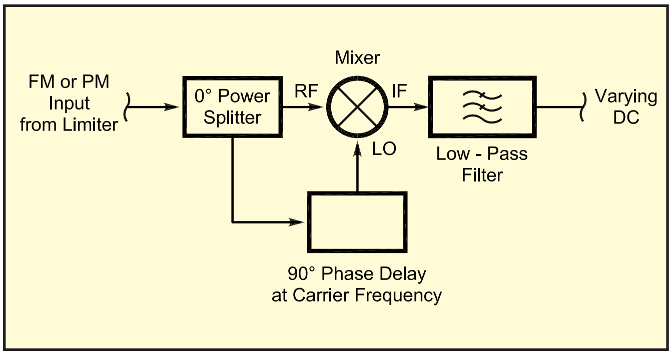

Figure 5—The quadrature detector produces a dc voltage output that varies with amplitude and polarity of phase differences from those at the carrier frequency.

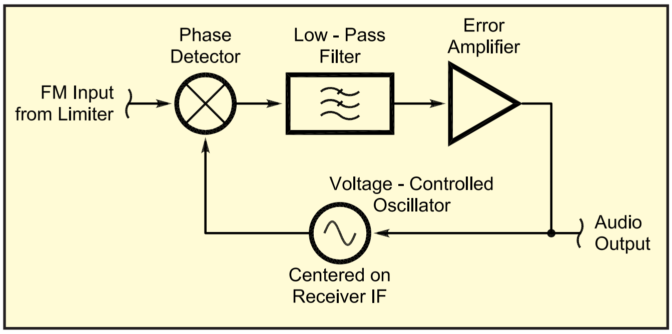

Figure 6—A PLL can be used to demodulate FM by letting the loop track the incoming signal. The error signal that causes the VCO to track is a replica of the message signal used for modulation.

Carson's Rule

The bandwidth, B, of an angle-modulated signal is often estimated by an approximation known as *Carson's Rule*. If one adds up the power in all the sidebands, it can be shown that to include 98% of the signal power the following formula applies:

$B \approx 2\ (m + 1)\ f_M$

So far, we've only discussed the cases when the message signal is a simple tone. For complex signals, such as voice or data, m changes constantly as the amplitude and frequencies of the message signal change. Luckily, convenient substitutes for m and f_M are available.[4] Taking the place of m in the approximation is the ratio of the peak deviation to the bandwidth of the message signal, called the *deviation ratio,* D. This is a constant for any given FM transmitter.

$D = f_d \times A / W$

For f_M, we can substitute W, the bandwidth of the message signal, and the approximation becomes:

$B \approx 2\ (D + 1)\ W$ (Carson's Rule)

If we use typical values for deviation (5 kHz) and message bandwidth (3 kHz for voice), D is 1.6 and B is 15.6 kHz. This combination appears to fit comfortably in a 20 kHz channel, the most common amateur FM channel spacing, with some spare room (called a *guard band*) to accommodate extra sideband energy and off-frequency stations.

Be careful! Carson's Rule is just a method of estimating the bandwidth in which a certain amount (98%) of the signal power is contained. As an example of another method, the FCC defines signal bandwidth (for any type of signal) in Part 97.3(a)(8) as "The width of a frequency band outside of which the mean power of the transmitted signal is attenuated at least 26 dB below the mean transmitted power within the band." In addition, the modulation index of angle-modulated signals is limited by Part 97.305(f)(1) to a maximum value of 1.[5] These two rules define a single "channel" and how much energy is allowed to be present outside that channel.

As an example, an FM signal carrying 3 kHz bandwidth speech from a 100 W transmitter, with an average output of 50 W across the channel, just meeting the FCC's output level and modulation index rules, could contain sidebands with a power of 0.125 W, 9 kHz from the carrier.[6] Including these low-power sidebands increases the bandwidth to 18 kHz. Figure 3 shows the spectrum of a typical 2 meter repeater output showing the limits for both Carson's Rule and the FCC bandwidth definition.

Given the sensitivity of modern receivers and the power of modern transmitters, using either Carson's Rule or the FCC definitions will likely be insufficient to guarantee that there is no interference to adjacent channels close to the minimum spacing.

Transmitting FM

Direct and Indirect FM

Generating an angle-modulated signal requires some method of varying either the frequency or phase of an RF signal. The first method that might come to mind is varying the reactance of the frequency-determining elements of an oscillator. This method, called *direct FM*, causes the frequency of oscillation to change.

An unmodulated carrier can also be passed through a tuned circuit whose reactances are modulated by a message signal, altering the signal's phase. The result is a PM signal whose deviation is directly proportional to both the amplitude and frequency of the message signal. If the message signal is filtered, so that its amplitude is reduced in half as its frequency doubles, the amplitude and frequency effects on the deviation are balanced, and an FM signal known as *indirect FM* results. Examples of both direct and indirect FM modulators can be found in the references.[7,8]

A phase-locked loop (or PLL) can be used to generate FM by modulating the voltage-controlled oscillator (VCO) or reference oscillator as shown in Figure 4. The resulting frequency error between the VCO and reference causes the phase detector to generate a signal that attempts to force the VCO back to the reference frequency. If the message signal needs to include dc, such as for transmitting a pulse train or a video signal, then the reference oscillator must be modulated or else the loop's frequency-correcting mechanics will "correct" the dc level out.

Modulation Quality

If high speed data at 9600 baud or greater is to be transmitted using FM, then the method used to generate FM and the quality of the modulator are significant. Minor distortion that makes no difference for voice transmission can slow or prevent high-speed data transmission. Most older radios were designed for voice, and while suitable for low-speed 1200 and 2400 baud data, they may not be usable at data rates of 9600 baud and higher.

Modems used to encode and recover data from transmitted signals are particularly sensitive to the linearity of the modulator and to the transitions between the different data symbols. Nonlinearity results in a blurring of the message signals—just as if noise had been added. Transitions that are too abrupt or not smooth also make it more difficult for the modem to distinguish the new symbol properly. When you purchase a radio for data applications, make sure it is rated for that service.

When one adjusts an FM transmitter—either for repair or during construction—it is important to set the deviation properly to avoid interference to adjacent channels and distortion in the received signals. This can be done without sophisticated test equipment if one has access to a receiver with a CW filter that can tune to a carrier frequency of the equipment being adjusted. A single audio tone is used to modulate the transmitter while the test receiver listens on the carrier frequency. The audio tone's frequency is set to:

f = maximum desired deviation / 2.405

where the constant 2.405 represents the modulation index at which the carrier goes to zero. The deviation is slowly increased from zero (an unmodulated carrier) until the carrier received on the test receiver reaches a minimum value. This and other methods are more fully described in the reference material.[9]

Receiving FM

Once the FM or PM signal has been generated and transmitted, it is then necessary to demodulate it and recover the message signal—voice or data. Both FM and PM signals are received and demodulated using the same methods.

AM signals are demodulated by circuits that react to the amplitude of the signal. For FM, the message is encoded in the frequencies of the received signal, so any amplitude variations, such as noise or static, can be discarded. This is why FM reception can be largely free of atmospheric and man-made noise. These effects cause amplitude variations in the signal—very few external processes make unwanted changes in a signal's frequency.

Limiting, Quieting and Capture Ratio

Amplitude variations are removed from the received signal by amplifying it until it essentially becomes a square wave of constant voltage. An AM signal subjected to this treatment would be horribly distorted, but for FM, the information is still there in the form of frequency variations. This process is called *limiting*. When a signal is strong enough to drive an FM receiver's amplifiers all the way into saturation or *hard limiting*, so that no further increase in output is possible, then all AM noise is eliminated. This is the origin of the term *full quieting*—the signal is strong enough for the receiver to have completely eliminated all AM noise. As the signal strength drops, the limiters can't remove all of the noise so it starts to reappear in the output audio.

The ability of an FM receiver to amplify small signals and suppress noise is a measure of its sensitivity. A receiver spec sheet will state how much noise is suppressed for a certain input signal level. An example might be "0.2 µV for 20 dB of quieting." This means that, compared to the receiver output with no signal present (imagine the receiver output noise with the squelch opened), the audio level will be 20 dB quieter if a 0.2 µV unmod-ulated signal is present. The more sensitive the receiver, the lower the input level requirement will be for an equivalent amount of quieting.

A similar effect occurs when two FM signals are present at the receiver input. Because of the high gain of the limiters, the stronger of the two signals will dominate as if the weaker is noise. The weaker signal effectively disappears. This property of FM receivers is called the *capture effect* and the measurement of that effect results in a specification called *capture ratio*. The capture ratio represents the difference in dB required between two signals such that the stronger signal will suppress the weaker one. Lower numbers are better and a capture ratio of 1.5 dB is considered good.

It doesn't take much for the stronger signal to suppress the weaker one completely—as little as a 1 dB difference in signal strength could achieve that. Capture effect can be observed when two stations try to access a repeater receiver at the same time. The weaker signal will be suppressed until the stronger signal stops transmitting. If you've ever taken an out of town drive while listening to your favorite FM station and notice another station on the same frequency suddenly "popping" into the channel, you've experienced capture effect.

Detectors

How do FM receivers change the frequency variations back into a message signal whose amplitude varies? There are several methods that can turn the frequency variations into voltage variations of received audio.

Slope detection is the simplest (and oldest) method. Imagine trying to tune past a carrier with an SSB or CW receiver. When the carrier is centered in the receiver passband, it will be at its loudest. As the receiver is tuned past the carrier, the filter's rolloff will cause the carrier amplitude to diminish. If the receiver was tuned back and forth rapidly with the carrier on the slope of the filter's response, the result would be an ac signal that varies with the tuning. Slope detection works similarly except that the tuning is fixed and the signal frequency varies, sliding up and down the slope of the detector's response curve.

The *discriminator* and *ratio detector* rely on the phase relationships of voltage and current in the primary and secondary windings of a transformer tuned to the IF of the receiver. Because the balance of the signals in the transformer changes above and below the center frequency of the transformer, the rectified signals from each side of the secondary winding can be combined in an external circuit to create a voltage that varies with the frequency of the signal at the transformer's primary winding. This varying voltage is amplified to create the receiver's output audio. A more detailed discussion of these detectors is available in the references.[7,10]

A *quadrature detector* utilizes an interesting property of angle-modulated signals that allows the message signal to be recovered when the RF signal is multiplied by a time-delayed version of itself, as shown in Figure 5. Quadrature detectors are currently the most popular FM detector, as they are a simple circuit to implement on ICs. The time delay function is usually provided by a tuned circuit external to the IC.[7]

FM can also be detected by a PLL. As shown in Figure 6, the PLL's natural function of tracking a changing input frequency can be employed to generate a voltage that varies as the input frequency changes. The PLL phase detector compares the output from the receiver's limiter to the VCO frequency. Any errors are fed back to the VCO so that it follows the input signal. The error signal shifts up and down as the input signal frequency varies, creating a replica of the original message signal.

Conclusions

After reading this article, I hope that perhaps some of FM's fundamentals will have been made clearer and that the jargon commonly tossed around on the repeater will have a little more meaning. Modulation techniques and theory are fascinating areas of radio theory for the interested amateur and are a key to cutting-edge wireless technology.

The author would like to thank John Davidson, KC6TFS, ARRL Orange Section Official Observer Coordinator for his communications regarding FM signal bandwidth and repeater channel spacing.

Notes

[1] M. Eisenberg, K3DG, "The Father of Modern Radio," *QST*, May 1991, pp 49-51.

[2] H. Hyder, W7IV, "Phase Versus Frequency Modulation," *QST*, Jul 1981, pp 33-34.

[3] If your Web browser can display Java applets, an excellent graphic demonstration is available at **www.algomusic.com/jmsl/tutorial/FMSpectrumApplet.html**. Using this program, you can adjust both the amplitude and frequency of a single-tone message signal and watch the sideband amplitudes change while the value of m is displayed.

[4] Ziemer and Tranter, *Principles of Communications: Systems, Modulation and Noise*, Boston: Houghton-Mifflin Company, 1976, section 3.2.

[5] CFR Title 47, Volume 5, Part 97.

[6] If sideband power is calculated as 20 log [$J_n(1)/0.5$], then the 3rd sideband of a 3 kHz signal would be 27.9 dB below the average power in the channel, represented by a sideband amplitude of 0.5.

[7] *The ARRL Handbook for Radio Communications*, Newington: ARRL, 2003, Chapter 15, "Mixers, Modulators and Demodulators."

[8] D. DeMaw, W1FB, "First Steps in Radio, Part 17—Understanding FM Transmitters," *QST*, May 1985, pp 23-25, **www.arrl.org/tis/info/pdf/8505023.pdf**.

[9] *The ARRL Handbook for Radio Communications*, Newington: ARRL, 2003, Chapter 12, "Modulation Sources."

[10] D. DeMaw, W1FB, "First Steps in Radio, Part 18—Understanding FM Receivers," *QST*, Jun 1985, pp 25-27, **www.arrl.org/tis/info/pdf/8506025.pdf**.

First licensed as WN∅GQP in 1972, H. Ward Silver, N∅AX, is a frequent contributor to QST. *The author of the current* QST *column, "Hands-On Radio," he is an engineer, an author and a teacher. Ward enjoys contesting and DXing; he has helped new and prospective hams of all ages and is the author of the newly published book* Ham Radio for Dummies. *He can be contacted at 22916 107th Ave SW, Vashon, WA 98070 or at* **n0ax@arrl.org**.