Chapter 8
FM and PM

Now that we understand the concept of a carrier and sidebands, let us extend that to the more complex FM, or Frequency Modulation.

Today, there are very few AM-only or FM-only radios made; most radios cover both AM and FM broadcast bands. So most of us listen to both AM and FM broadcast stations. Unfortunately, this often leads us to the wrong conclusions about their relative features.

Most people know that FM sounds better — it has better frequency response and less noise and interference, and can be heard in underpasses and tunnels, where AM fades out. They also know that you can hear AM broadcast stations farther away than FM stations, especially at night. And so they become convinced that this is always true for all AM and FM. In reality, much of what we think about FM comes not because it is FM, but because commercial FM radio stations use different carrier frequencies from AM stations, and they are allowed to use wider bandwidths.

AM broadcast station carrier frequencies are in the range from 0.55 to 1.6 MHz, and are limited to 20 kHz bandwidth. On the other hand, FM broadcast carriers lie between 88 and 108 MHz, and are assigned 200 kHz bandwidth. The fact that AM signals travel farther is because of the lower frequencies; the fact that FM signals can be heard better in underpasses and tunnels is because of the higher frequencies. And if AM was allowed wider bandwidths, it too could have a better frequency response. Just about the only inherent advantage that FM has over AM is that it resists noise and interference better.

Frequency Modulation (FM)

Just as Amplitude Modulation changes the amplitude of a carrier, Frequency Modulation changes the frequency of the carrier. There is also Phase Modulation which — you guessed it — changes the phase; we will discuss it later in this chapter.

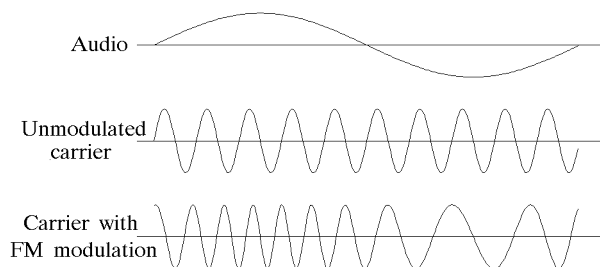

Fig. 8-1. Frequency Modulation

Fig. 8-1 shows what we mean by FM. At the top, we see the audio waveform that we want to transmit (although we're talking about sending audio at this point, in reality you could send video, data, or any other kind of information via FM.) In the middle is the carrier, as it would look before modulation. At the bottom is the resulting FM wave. As you can see, the frequency goes up (cycles are closer together) when the audio voltage is high, and the frequency goes down (cycles farther apart) when the audio signal goes down (it could also be the other way — one goes up when the other goes down.)

When there is no modulation (such as when an FM broadcast station transmits silence) the carrier frequency rests at what is called the *center frequency*; if you hear an announcer say "This is WXYZ on 96.3 MHz," that tells you what the center frequency is. Once the music starts, the frequency starts to vary above and below that center frequency; we say that it is *deviating* from the center frequency. *Deviation* is the word that describes how far it deviates from the center frequency.

DETOUR

Here is a short Basic program that generates the picture in Fig. 8-1. You can experiment with it to see what happens as you change values.

```
10 'Program to generate FM
20 TWOPI=2*3.14159265
30 FAUDIO=1
40 FCARRIER=10
50 MODINDEX=5
60 SCREEN 2
70 FOR X=0 TO 639
80   T = X/640
90   AUDIO = SIN (TWOPI * FAUDIO * T)
100  CARRIER = SIN(TWOPI * FCARRIER
* T)
110  FM = SIN(TWOPI * FCARRIER * T -
MODINDEX * COS(TWOPI*FAUDIO*T))
120  PSET(X,50-20*AUDIO)
130  PSET(X,100-20*CARRIER)
140  PSET(X,150-20*FM)
150 NEXT X
160 IF INKEY$="" THEN GOTO 160
170 SCREEN 0
```

At 640 different values of X going across the screen, the program calculates the voltage of the AUDIO signal and the CARRIER signal (whose frequencies are FAUDIO and FCARRIER, respectively), and then combines them into an FM signal. The three PSET instructions in lines 120, 130, and 140, then plot these values on the screen.

Copyright © 2002 by Peter A. Stark - All Rights Reserved

END OF DETOUR

FM differs from AM in many ways. Its major advantage over AM is its resistance to noise and interference. Noise and interference usually corrupt signals by changing their voltage — by adding spikes, or in some other way changing the shape of the signal; it cannot change its frequency. But since FM signals only change their frequency, not their amplitude, it's possible to design a receiver so it ignores amplitude changes (for instance, an FM radio often contains a *limiter* circuit which clips the tops off the signal to make the signal the same height no matter what level it comes in.) The receiver will then ignore the noise altogether. This won't work for very weak signals, of course, since then there's nothing to clip, but for any reasonably strong signal, FM can be remarkably free of noise.

Producing FM

There are many ways to produce an FM signal; one way is to build an oscillator, but then wiggle its frequency back and forth in step with the modulation signal using a special kind of diode called a *varicap*.

DETOUR

Most semiconductor diodes are made of silicon. pure silicon is a semiconductor — it conducts electricity, but not well, because its electrons are not very free to travel, and so they cannot carry any substantial current. But it can be made into a better conductor by adding other materials to it. These make it less pure, and so are called *impurities*; the resulting impure silicon is said to be *doped*.

One type of impurity has *more* conduction electrons than silicon; since electrons have a negative charge, the resulting doped silicon is called N-type. It has an excess of electrons, and the new electrons which were added by the N-type impurity are now freer to move and therefore make the silicon into a better conductor.

Another type of impurity has *fewer* conduction electrons than silicon; this produces a lack of electrons, and leaves "holes" — places where the silicon would like to have an electron, if there were any around. This sort of doped silicon is called P-type, because these holes behave sort of like positive charges.

Think of this as a game of musical chairs. In the normal musical chairs game, there is a ring of chairs and a ring of people, but there is always one less chair than there are people. When the music plays, the people move from chair to chair, and when the music suddenly stops, everyone sits down. But of course there is always one person without a chair, so he is "out", and so on. But suppose we reverse the game — instead of having one chair too few, put in one chair too many. Now there is always a chair free. As people move from chair to chair, ignore the people and instead focus your eyes on whichever chair is empty at that instant. As the people are moving in one direction, the empty chair seems to be moving in the opposite direction.

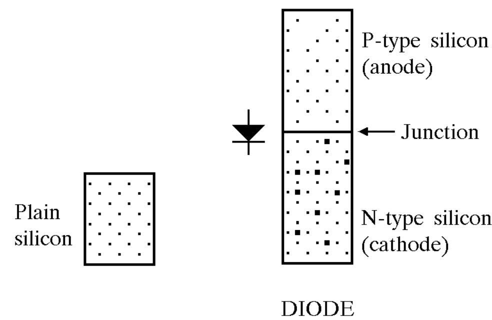

Fig. 8-2. Plain silicon and a silicon diode

P material behaves the same way, except that the electrons are like the people, and the holes — places where there should be an electron, but isn't — are the empty chairs. As a nearby electron jumps into a hole, it leaves another hole behind it. And so it looks as though the hole moved in a direction opposite to that of the electron.

Fig. 8-2 shows how this can be used to make a diode. On the left we see a piece of plain silicon. The dots show the positions of the electrons which might carry current if they were more free to move. The silicon is a crystal, and the atoms and their electrons form a neat pattern throughout the material.

On the right in Fig. 8-2 we see how a semiconductor diode is made up out of two pieces of silicon. The anode is a piece of P-type silicon, while the cathode is made of N-type silicon. The place where they abut against each other is called the *junction*.

If you compare the P-type anode with the plain silicon, you will see that there are white areas where there are missing electrons; these are the holes. On the other hand, we drew some larger dots in the N-material to show the extra electrons added by the doping.

Now imagine that the diode is forward biased; that is, the anode is connected to a positive voltage, while the cathode connects to negative. Since the negative electrons are repelled by the negative voltage at the cathode, and attracted by the positive voltage at the anode, they start trying to move upward. Every now and then (quite often, actually), one of the extra electrons in the cathode will look across the junction, spot a hole on the other side, and jump across the junction into that

Copyright © 2002 by Peter A. Stark - All Rights Reserved

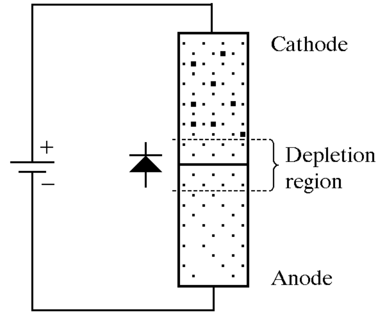

Fig. 8-3. Reverse-biased diode

hole. Of course, new electrons coming in from the negative terminal replace the lost electrons in the cathode, while the extra electrons in the anode jump from hole to hole and eventually go out the positive terminal. This gives us a current through the diode.

But now look at Fig. 8-3, which shows the diode reverse biased — that is, the anode terminal is connected to the negative terminal of a battery, while the cathode is connected to the positive terminal. Under these conditions, there is no current flow through the diode, simply because all the electrons up in the cathode are being attracted by the positive terminal of the battery, while all the holes at the bottom (which behave like positive charges) are attracted to the negative terminal of the battery.

Because all the electrons and holes are being pulled away from the junction, there is a region around the junction (sort of like a demilitarized zone) where there are no holes and no extra electrons. In electronics, this zone is called a *depletion region*, and you can see in Fig. 8-3 that this region is just plain silicon — no extra electrons, and no holes.

The size of this depletion region depends on how much voltage there is. If the battery voltage is small, there isn't much pull on the holes and electrons, and the region is fairly thin. If the voltage is high, then the region becomes much larger. The depletion region acts like an insulator, and prevents current from flowing through the diode.

So the diode acts like a conductive region at the top, an insulator in the middle, and another conductive region at the bottom. But that is the exact description of a capacitor — two conductors separated by an insulator. The diode therefore acts like a capacitor. But the capacity depends on the spacing between the conductors, which in turn depends on the applied voltage. Hence this is a variable capacitor, or varicap.

Varicap diodes are specially made to exhibit large changes of capacitance with voltage, but all diodes exhibit this effect to some degree. For example, the common 1N4001 rectifier diode makes a dandy varicap for many purposes.

So now back to FM modulation.

To produce FM, we need to produce a sine wave carrier whose frequency varies with the modulation we want to put on it. One way to do this is to build an oscillator which will produce the carrier, and then vary its frequency in step with the modulation. If we use an LC circuit (a tuned circuit with an inductor L and a capacitor C) to set the frequency of the oscillator, then we can add a varicap across the LC circuit. The audio (or other modulation) is sent to the diode, changing its capacitance. This changes the resonant frequency of the tuned circuit, which in turn changes the oscillation frequency. Voilà — FM.

An oscillator is a circuit which produces an ac wave (the word *oscillate* means to swing back and forth.) A typical oscillator consists of an amplifier, with part of its output sent back to the input by a *feedback* circuit.

Fig. 8-4 shows a circuit sometimes called a *feedback loop* or *servo loop*. An input signal V_{in} is applied to a combiner circuit (the circle with an X inside), and then to an amplifier which has a gain G. The output from the amplifier is V_{out}.

A portion of V_{out}, however, is sent back through a box whose gain is H, and combined with the input in the combiner; the combiner is also sometimes called a *summer* or *comparator*, depending on the application.

The normal equation for the overall gain of a servo loop circuit is

$$A = \frac{V_{out}}{V_{in}} = \frac{G}{1 - G\,H}$$

There are two possible things that can happen at this point, depending on whether we have *negative feedback* or *positive feedback*.

(A) Negative feedback. As we can see in Fig. 8-4, the actual input into the amplifier is a combination of the input voltage V_{in}, and the feedback voltage. If these two signals are out of phase, then the actual input would

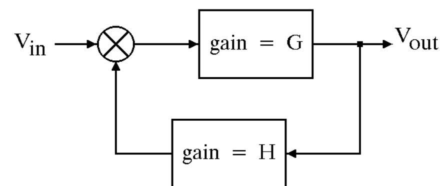

Fig. 8-4. A servo loop

Copyright © 2002 by Peter A. Stark - All Rights Reserved

be less than V_{in}, and therefore the output voltage from the amplifier would be less than if there was no feedback. In other words, the gain of the servo circuit is now smaller than the gain of the amplifier all by itself.

We can see this in the above equation. If the product of G times H is negative, then the denominator of the equation is larger than 1 (since subtracting a negative number is the same as adding a positive number). Dividing G by a number larger than 1 gives us a smaller number; in other words, the gain A of the entire circuit is smaller than the gain G of the amplifier by itself.

Since negative feedback reduces the gain of an amplifier, you may be tempted to think that it is a bad thing. Actually, though, negative feedback is used a lot because, while it reduces the gain, it also reduces the distortion of the amplifier and improves other factors. Almost every hi-fi amplifier uses it to improve operation.

(B) Positive feedback, however, is what we need to make an oscillator. With positive feedback, the signal being fed back into the summer is *in* phase with the input signal V_{in}, and so it adds to the input. This gives the amplifier more input than it would otherwise have, and so there is a larger output voltage V_{out} than before.

Returning to our equation, suppose that the product GH was a positive number such as +0.5 or +0.8; it would then subtract from the 1 in the denominator, making the denominator smaller than 1. From the equation, we see that dividing the amplifier gain G by a number smaller than 1 would make the quotient larger. This increases the gain, but it also reduces the amplifier quality. If the amplifier produces some distortion, then some of that distortion is sent back to the amplifier and amplified again; the output therefore has even more distortion than before. Positive feedback is therefore not generally used in plain amplifiers, unless you need more gain than the amplifier can produce by itself.

But suppose that the gain G of the feedback circuit were carefully adjusted so that the product GH were *exactly* equal to +1. Then the denominator of the equation would be 1 – 1, or exactly zero! Since dividing the amplifier gain G by zero gives infinity as the answer, the gain of this servo loop would be infinite. In other words, there could be an output from the circuit even though there is no input. As it happens, that is exactly what we want from an oscillator!

An oscillator is therefore an amplifier with a positive feedback loop around it. The feedback loop is adjusted so that its gain H is exactly equal to $1/G$ at whatever frequency we want it to oscillate at (so that the product of $G \times H$ is exactly equal to +1.)

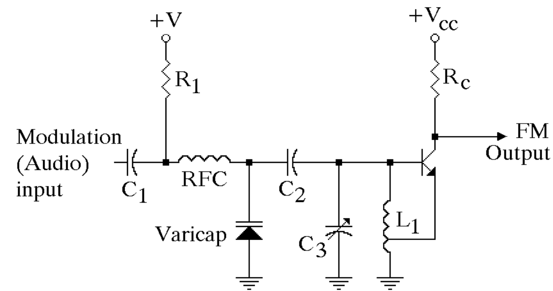

Fig. 8-5. A frequency-modulated Hartley oscillator

Fig. 8-5 shows a Hartley oscillator circuit. The circuit is very simplified, and needs some more components to make it work — for example, there is no biasing for the transistor — but the important components are there. It works like this:

First, note the tuned circuit, consisting of inductor L_1, and capacitor C_3. This circuit sets the frequency of operation. But at the same time, inductor L_1 acts as an *autotransformer.* (An autotransformer acts very much like a normal transformer, but to save money, it combines both the primary and the secondary windings into one. In this case, the entire inductor is the secondary, while the portion from the tap down to ground is the primary.) Since the secondary has more windings than the primary, the output voltage from the autotransformer is larger than the input voltage; hence the gain (which is actually the H in the above equations) is greater than 1.

The transistor is connected both as an emitter follower, and as a common-emitter amplifier. As usual, the emitter follower circuit has a gain slightly less than 1; this gain is the G in our previous equations. The output from the transistor goes into the tap of the inductor (the primary), while the output from the top of the inductor (the secondary) is the input into the transistor's base.

So the transistor provides the gain G (which is smaller than 1), while the tuned circuit provides the feedback H (which is larger than 1). Multiply the two out, and you get the product $G \times H = 1$, just what is needed for oscillation. Once the transistor oscillates, the output is taken from its collector.

Now let's look at capacitor C_2, and the components to its left. C_2 is relatively large, so that at the frequency of oscillation it behaves almost like an ac short circuit. Hence, for ac purposes, the varicap diode is connected directly across the tuned circuit. The exact oscillation frequency therefore depends not just on L_1 and C_3, but also on the capacitance of the varicap. (Notice the symbol for the varicap — a diode whose cathode side consists of two lines, like a capacitor.)

Copyright © 2002 by Peter A. Stark - All Rights Reserved

In order to produce the depletion region in the varicap, the varicap is reverse-biased by connecting its cathode to +V through resistor R_1. But note how the audio voltage comes in through capacitor C_1 on the left. This audio voltage adds to, or subtracts from, the voltage +V, so the width of the depletion region varies with the audio. That changes the capacitance, which in turn changes the frequency of oscillation, giving us FM.

Finally, note the inductor labelled RFC. This is an *RF Choke* — a small inductor whose purpose is to keep radio-frequency signals from flowing through it. It keeps the high-frequency signal at the tuned circuit from getting back into the audio or +V bias circuit. Capacitor C_2, on the other hand does the opposite — it keeps the audio and dc voltage across the varicap from getting to (and being short-circuited by) inductor L_1. We can't connect the varicap diode directly across the tuned circuit since it has to be reverse biased, and the inductor would short the dc bias voltage.

There are other ways to generate FM, but the varicap circuit is the simplest and most obvious. Its disadvantage is that it produces some distortion; hence it is most often used for voice or data communications, and seldom for high quality music.

Detecting FM

Just as there are various ways of producing FM, so there are many ways of detecting (or demodulating) it. Let's begin with the very simple circuit of Fig. 8-6.

Let's assume that the FM carrier has a center frequency of 10.7 MHz, and that it varies above and below that value. Let us tune the two resonant circuits so the top one, consisting of L_1 and C_1, is resonant at 10.8 MHz, while the bottom one, consisting of L_2 and C_2, resonates at 10.6 MHz. Further, make R_1 and R_2 the same, R_3 and R_4 the same, and C_3 and C_4 the same.

Start with the FM input at 10.7 MHz. The input signal is split by resistors R_1 and R_2 into two paths; part goes to the top tuned circuit, part to the bottom. The top resonant circuit is tuned 100 kHz too high, while the bottom one is tuned 100 kHz too low, so the input signal appears across both tuned circuits, slightly attenuated (since the signal is not at resonance), but almost exactly the same amplitude.

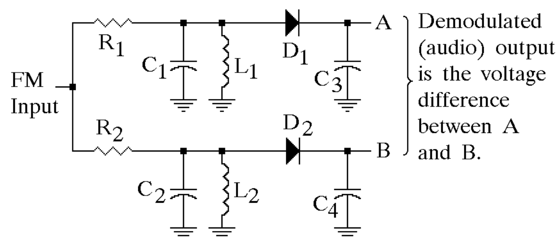

Fig. 8-6. A simple FM Demodulator

Like the diode and filter capacitor in a power supply, diode D_1 and capacitor C_3 rectify the ac signal from the top tuned circuit into a positive dc signal, and filter it. Likewise, diode D_2 and capacitor C_4 rectify the ac signal from the bottom tuned circuit.

Since the ac signals across the two tuned circuits are about the same amplitude, the dc voltages across C_3 and C_4 are also equal. Points A and B therefore have the same voltage. The voltage between A and B is the actual output, and this is zero.

Now suppose the frequency increases above 10.7 MHz. Since the signal is now closer to the 10.8 MHz resonant frequency of the top tuned circuit, the ac signal across that circuit increases; likewise, the ac signal across the bottom tuned circuit decreases because the input frequency is now farther away from the 10.6 MHz resonant frequency of that circuit. Point A is now more positive than point B, and so the net output voltage is positive. In fact, the closer the input frequency gets to 10.8 MHz, the bigger the positive output.

In exactly the same way, if the input frequency drops toward 10.6 MHz, point B becomes more positive than A, and so the difference between A and B becomes negative. And the more it drops, the more negative the output gets.

In other words, the output voltage is proportional to the frequency of the input signal — when the frequency goes up, the output voltage goes positive; when the frequency goes down, the output voltage becomes negative. This converts the frequency changes of the FM signal back into the original modulation.

While the circuit of Fig. 8-6 would work, the fact that it has two tuned circuits makes it too difficult to adjust; it also has a bit too much distortion. Most receivers therefore use a slightly different circuit.

Fig. 8-7 shows one common circuit, called the *Foster-Seeley Discriminator*. (Don't confuse this with two companies, one called "Stearns-Foster", and the other called "Sealy" — they make mattresses, which have nothing to do with FM!) By relying on the phase differences between several signals, the circuit produces voltages at the two diodes which depend on the frequency.

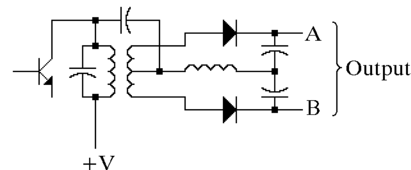

Fig. 8-7. Foster-Seeley Discriminator

Copyright © 2002 by Peter A. Stark - All Rights Reserved

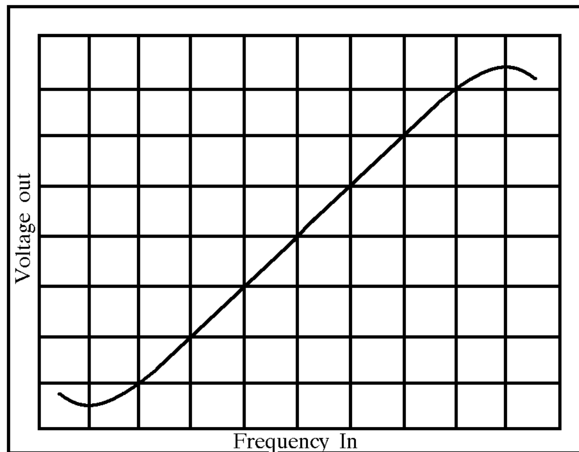

Fig. 8-8. The FM Discriminator "S" Curve

By the time the signals are rectified and filtered, we again have a demodulated output which depends on frequency.

The circuits of Fig. 8-6 and Fig. 8-7 both have one disadvantage — their output voltage also depends on amplitude. The stronger the FM input signal, the greater the output voltage. This means that noise (which changes the input voltage) would get through the detector circuit. As we pointed out earlier in this chapter, one way to avoid this problem is to put a limiter in front of the detector; the limiter keeps the input signal at a constant height, and so avoids the problem. But this means we need to add one more circuit to the radio, so many designers try to cut costs by using a slightly different circuit, called a *ratio detector*. Instead of taking the difference between the two diode outputs in a pair of resistors, the ratio detector looks at the *ratio* of the two diode voltages. Regardless of how strong or weak the signal is, at any given frequency the ratio of the two signals should be the same, and so this circuit ignores any amplitude changes in the signal. In many cases (especially in commercial FM broadcast receivers) that eliminates the need for a limiter, and cuts cost.

Fig. 8-8 shows the output vs. input curve of a typical discriminator or ratio detector; you can see why it is often called the "S" Curve. It shows how the output voltage (the vertical axis) depends on the input frequency (the horizontal axis). If the circuit is well designed, then the center of the curve should be as straight as possible, to make the output voltage exactly proportional to the frequency change, and produce the lowest distortion. The ends of the "S" curve are curved, but the FM signal usually doesn't change frequency that far up or down, so it doesn't matter.

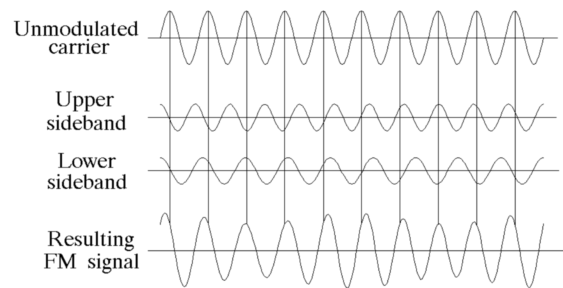

Fig. 8-9. FM from one set of sidebands

FM Sidebands

Like AM, transmitting an FM signal results in sidebands, and those sidebands increase the bandwidth of the FM signal.

Let's start by looking at Fig. 8-9. At the top, we have 10 cycles of an unmodulated carrier with a constant frequency. Note the vertical lines, which cross the carrier signal exactly at the top of each cycle. If you carefully measure, you will see that the distances between all the lines are exactly the same, which proves that each cycle is exactly the same length.

Below the carrier we see two sidebands, the upper sideband with a slightly higher frequency (it shows 11 cycles instead of 10), while the lower sideband has a slightly lower frequency (it shows only 9 cycles.)

In the bottom waveform, we used a computer to sum the carrier and sidebands to produce a resulting waveform. Like the unmodulated carrier at the top, it shows 10 cycles, but these cycles are not all the same width — you can see that the vertical lines cross the bottom wave at different places in each cycle. From this we conclude that the FM signal at the bottom is changing frequency (because some cycles are longer and some are shorter), but that the average frequency is the same as the unmodulated carrier. In other words, modulating the carrier to produce FM varies the frequency back and forth, above and below the "resting" or unmodulated value.

We therefore see that, like an AM signal, an FM signal consists of a carrier plus sidebands. But there is one catch — if you examine Fig. 8-9 carefully, you can see that the cycles of the FM signal at the bottom are not all the same height. In other words, the frequency is varying, but so is the amplitude. But this can't be — there is something wrong here! FM is supposed to stay the same amplitude all the time, not change! What's wrong?

The problem is that Fig. 8-9 doesn't tell us the whole story. Fig. 8-9 would be good enough to explain FM *if the two sidebands were much smaller.* When the amount of FM modulation is very small (for example,

Copyright © 2002 by Peter A. Stark - All Rights Reserved

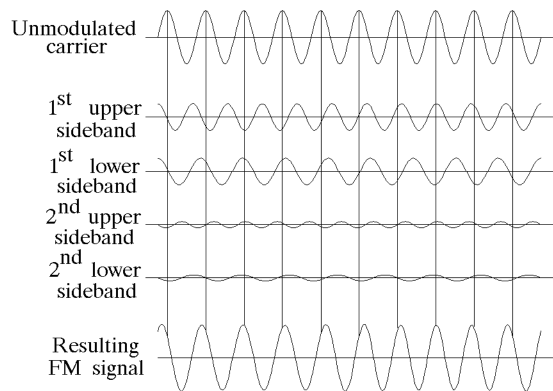

Fig. 8-10. Multiple FM sidebands

when the music on an FM station is very soft), one tiny set of sidebands, much smaller than those shown in Fig. 8-9, is enough to swing the resulting FM carrier back and forth in frequency a small amount without changing the amplitude.

But when an FM station transmits loud music (which produces much more modulation), the simple picture of Fig. 8-9 needs something else — let's call it a "finagle factor" — something which will keep the amplitude constant.

This "finagle factor" is another set of sidebands. A real FM signal usually contains more than one pair of sidebands; there could be dozens or even hundreds of sidebands on each side of the carrier when there is a LOT of modulation.

Fig. 8-10 shows how adding just one more set of sidebands can fix the problem of Fig. 8-9. Here we see the following:

- The original unmodulated carrier with 10 cycles,
- One pair of 1st sidebands, an upper one with 11 cycles, and a lower one with 9 cycle,
- A pair of 2nd sidebands, an upper one with 12 cycles, and a lower one with 8 cycles,
- The resulting FM signal with 10 cycles of varying frequency, but constant amplitude.

So we see how adding just one additional set of sidebands (with frequencies different from the first set) can fix the amplitude variations.

As we mentioned earlier, with more modulation, even more sidebands might be required to keep the amplitude constant. So we need to learn how to measure the amount of modulation. This can be done in two ways – using the deviation, or using a modulation index.

Deviation

Deviation measures how far the carrier swings from its center or unmodulated value. For example, commercial FM broadcast stations are allowed to deviate up to ±75 kHz from their center frequency. Thus a station on 96.3 MHz could swing its carrier down to 95.225 MHz (96.3 MHz minus 75 kHz) or up to 96.375 MHz. Police, fire, or ham FM radios, on the other hand, generally use ±5 kHz deviation.

Note how we said that commercial FM broadcast stations *are allowed* to deviate up to ±75 kHz. This is not a technical limit — it would be easy to build equipment which deviates a lot more. Rather, 75 kHz is a maximum limit imposed by the FCC to prevent interference with adjacent stations. Likewise, 5 kHz deviation is used by fire, police, ham, and other communications equipment to avoid interference with other users. So, unlike the case of AM where 100% modulation is a limit which you simply cannot exceed for technical reasons, the maximum amount of FM deviation is more of a "good neighbor" policy, usually enforced by the government.

Modulation Index

The modulation index is the second way of measuring the amount of FM modulation. The modulation index (we will abbreviate it as *MI*, but there are several other symbols also often used) is defined as

$$MI = \frac{\text{deviation}}{\text{modulation frequency causing that deviation}}$$

For example, an FM station transmitting a 5 kHz audio signal (this is the modulation frequency) at a 15 kHz deviation would have a modulation index of 15/5, or 3.

You can see that, even if the deviation stays the same, the modulation index can vary all over the place. For example, if that FM station keeps the same 15 kHz deviation, but the audio frequency changes from 5 kHz to 50 Hz, the modulation index suddenly changes to 15,000 Hz divided by 50 Hz, or 300. (One must be careful to use the same units on both the top and bottom of the equation.)

It's the modulation index that determines how many sidebands are needed to keep the amplitude of the FM signal constant. For example, if the modulation index is 0.2 or 0.3, then just one set of sidebands is needed; if the modulation index is 10, then 13 sidebands would be needed on each side of the carrier to keep the amplitude constant.

The calculations to determine how many sidebands are needed at any given modulation index are fairly complex, and involve something called Bessel Functions. To avoid all this work, most people use a table

Copyright © 2002 by Peter A. Stark - All Rights Reserved

Table 8-1. Carrier and sideband amplitudes (from Bessel functions)														
		Sideband												
MI	Carrier	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
0.0	1.00												,	
0.25	0.98	0.12												
0.5	0.94	0.24	0.03											
1.0	0.77	0.44	0.11	0.02										
1.5	0.51	0.56	0.23	0.06	0.01									
2.0	0.22	0.58	0.35	0.13	0.03									
2.5	-0.05	0.50	0.45	0.22	0.07	0.02								
3.0	-0.26	0.34	0.49	0.31	0.13	0.04	0.01						,	
4.0	-0.40	-0.07	0.36	0.43	0.28	0.13	0.05	0.02						
5.0	-0.18	-0.33	0.05	0.36	0.39	0.26	0.13	0.05	0.02					
6.0	0.15	-0.28	-0.24	0.11	0.36	0.36	0.25	0.13	0.06	0.02				
7.0	0.30	0.00	-0.30	-0.17	0.16	0.35	0.34	0.23	0.13	0.06	0.02			
8.0	0.17	0.23	-0.11	-0.29	-0.10	0.19	0.34	0.32	0.22	0.13	0.06	0.03		
9.0	-0.09	0.24	0.14	-0.18	-0.27	-0.06	0.20	0.33	0.30	0.21	0.12	0.06	0.03	0.01
10.0	-0.25	0.04	0.25	0.06	-0.22	-0.23	-0.01	0.22	0.31	0.29	0.20	0.12	0.06	0.03

like Table 8-1 on the next page, which actually provides more information than just the number of sidebands.

Let's look at an example. Let's assume we have an FM broadcast station which transmits a 4 kHz tone at a deviation of 2 kHz, giving us a modulation index of 0.5. For this modulation index, Table 8-1 gives us the following information:

MI	Carrier	Sideband	
		1st	2nd
0.5	0.94	0.24	0.03

This tells us that

- There are two sets of sidebands on each side of the carrier,
- The carrier amplitude is 94% (0.94) of the unmodulated value,
- The first set of sidebands has an amplitude of 24% (0.24) of the unmodulated carrier,
- The second set of sidebands has an amplitude of 3% (0.03) of the unmodulated carrier.

Notice that the carrier is now smaller than it would be if left unmodulated. Unlike in AM, where the carrier stays the same size no matter what the sidebands do, in FM the carrier changes its size as you modulate it. This makes sense when you realize that the total FM signal voltage and power (which consists of the carrier plus all the sidebands) has to stay the same all the time. When the sidebands appear, something else has to decrease to keep the total voltage and power the same.

Let's look at this in more detail. Suppose the original carrier (before modulation) was 10 volts rms, and was sent into 50 ohms. The total power would then be found from $P = V^2/R$ to be 2 watts.

Let's now see what happens with the modulation index of 0.5:

The carrier voltage is now $10 \times 0.94 = 9.4$ volts, and its power (still assuming a 50-ohm load) is $9.4^2/50$ or 1.7672 watts, less than before.

The first sideband is $10 \times .24 = 2.4$ volts, and its power is 2.4^2 / 50, or .1152 watt. But remember — there are two of these sidebands (one upper and one lower), so their total power is twice that, or 0.2304 watt.

The second sideband is $10 \times .03 = 0.3$ volts, and its power is 0.3^2 / 50, or .0018 watt. But there are again two of these sidebands, so their total power is twice that, or 0.0036 watt.

When we add all of these powers together, the 1.7672, 0.2304, and 0.0036 watt add to almost exactly 2 watts (The slight error is because Table 8-1 only gives the coefficients to two decimal places.) So the modulated carrier has the same power as the unmodulated carrier.

Let's try another example: Suppose the unmodulated carrier is again 10 volts, the load is still 50 ohms, the modulating frequency (which would be the audio in an FM broadcast station) is 1 kHz,and the deviation is 4 kHz, giving a modulation index of 4. Table 8-1 tells us that the carrier has an amplitude of –0.40, and there are seven sidebands, of which the first also has a negative

Copyright © 2002 by Peter A. Stark - All Rights Reserved

amplitude. Since amplitudes can't be smaller than zero, what do the minus signs mean?

The negative amplitude here simply means that the phase has reversed, so the carrier and the first sideband both have opposite phases from what they were at lower values of the modulation index. As far as the power is concerned, this doesn't matter, so we can still add up the total powers as follows:

	Voltage (volts)	Power each (watts)	Total Power (watts)
Carrier	–4	.3200	.3200
1st SB	0.7	.0098	.0196
2nd SB	3.6	.2592	.5184
3rd SB	4.3	.3698	.7396
4th SB	2.8	.1568	.3136
5th SB	1.3	.0338	.0676
6th SB	0.5	.0050	.0100
7th SB	0.2	.0008	.0016
Total power, carrier plus sidebands:			1.9904

Again, the total power is about 2 watts, with a slight error because the Bessel values are only accurate to two decimal places. So the total power of the FM signal doesn't change with modulation, although the power gets shifted around between the carrier and the sidebands.

Let's look at the negative numbers in Table 8-1 one more time. At a modulation index of 2, the carrier had a value of 0.22, while at an index of 2.5 the carrier became -.05. Obviously, somewhere between 2 and 2.5,the carrier value went from plus, through zero, to minus. At the zero (which occurs when the modulation index is 2.405) the carrier completely disappears, and all the power is in the sidebands. This point is useful because it lets us calibrate a deviation meter — we observe the signal on a spectrum analyzer and increase the deviation until the carrier disappears. Of course, we have to be careful because the carrier also disappears at other values, such as at a modulation index of 5.5 and 8.65.

We can see this a bit better if we graph the data from Table 8-1, as in Fig. 8-11. Here you can see how the carrier starts out with a magnitude of 1 when the modulation index is 0; it then decreases until at a modulation index of 2.405 it crosses the zero line and becomes negative. It then reaches a negative maximum at a modulation index of about 4, and then goes up again. It keeps oscillating, but generally gets smaller and smaller, while more and more sidebands appear at the right.

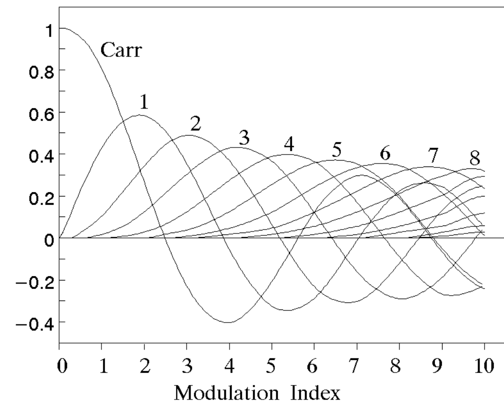

Fig. 8-11. FM carrier and sideband amplitudes

Bandwidth

The more sidebands there are, the greater the bandwidth of the signal. At a very low modulation index (below 0.5), there is only one set of sidebands, which makes the total bandwidth about the same as AM. But with larger modulation indexes, the number of sidebands increases and the bandwidth goes up.

Fortunately, something else is happening which keeps the total bandwidth reasonable. Remember how we defined the modulation index:

$$MI = \frac{\text{deviation}}{\text{modulation frequency causing that deviation}}$$

To get a large modulation index, you need a large deviation, but a small modulation frequency. The modulation frequency, however, determines the spacing between sidebands. So at a high modulation index, you may have many sidebands, but they will be close together, so the total occupied bandwidth will not be as large as you might think.

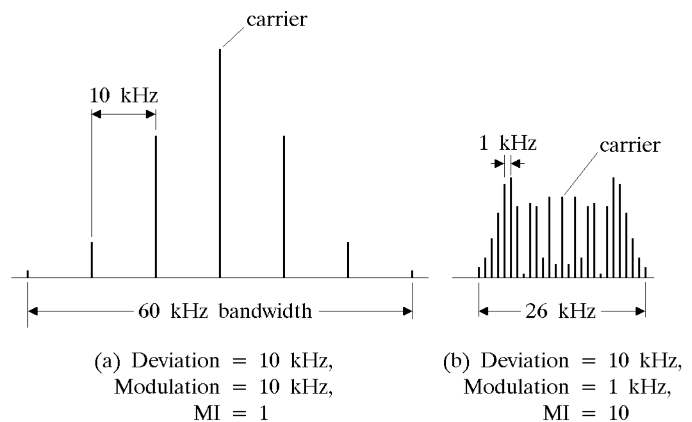

Fig. 8-12. FM, same deviation, different modulation index

Copyright © 2002 by Peter A. Stark - All Rights Reserved

Fig. 8-12 shows an example. At left, we see the spectrum when the deviation is 10 kHz and the modulation frequency is also 10 kHz; the resulting modulation index is 1, and there are three sidebands on each side. Since they are 10 kHz apart, the total bandwidth is 60 kHz.

On the right, the deviation is also 10 kHz but the modulation frequency is only 1 kHz. The modulation index is now 10 and there are 13 sidebands. But the sidebands are only 1 kHz apart, so the bandwidth is only 26 kHz.

In the real world, things are somewhat more complicated than what we've described so far. Our discussion so far has assumed that the FM transmitter is sending out pure tones — much like an announcer whistling a single note into the mike. In practice, regardless of whether we deal with voice, music, or data, the modulation signal consists of many different frequencies at the same time. The Bessel function analysis then becomes much more complicated, because the interactions between all these tones themselves produce additional sidebands.

Then too, even when the modulation is fairly simple, different frequencies exist at different voltages. In music or voice, for example, the high frequencies tend to be much weaker than the lows. Even though FM stations often boost the highs (this is called *pre-emphasis*) to overcome noise, the highs are still generally weaker than the lows, and so produce lower deviation.

Commercial FM broadcast stations therefore carefully adjust their audio response to utilize their assigned spectrum as fully as possible.

Channels

As we mentioned earlier, commercial FM stations are allowed to deviate up to ±75 kHz. Although the frequency response of their broadcast sound only extends to 15,000 Hz, they often transmit modulation up to almost 100 kHz (more on this later.) The high-frequency material they may transmit up near 100 kHz is at a modulation index below 0.5, so that it generates only one set of sidebands, but even this one single sideband set means their bandwidth may extend out to 100 kHz on each side of their center frequency. Hence commercial FM broadcast stations are assigned a *channel* which is 200 kHz wide.

The U.S. FM broadcast band extends from 88 to 108 MHz; the lowest FM station on the band would be assigned a 200-kHz-wide channel from 88.0 to 88.2 MHz, with its center frequency at 88.1. The next station up could be centered at 88.3 MHz, with its channel extending from 88.2 to 88.4 MHz. This explains why all the FM stations in your area are at odd tenths of MHz: .1, .3, .5, .7, or .9 on the dial.

In practice, though, two stations in the same area cannot be assigned adjacent channels, because the bandwidth of the tuned circuits in FM radios is not narrow enough to separate two such adjacent stations. So rather than being spaced 200 kHz apart, FM stations in the same city are often 400 or even 600 kHz apart. The inbetween channels are used by stations farther away.

TV channels are assigned in the same way. Each TV channel is 6 MHz wide; starting at the low end, their frequencies are:

Channel 2: 54 to 60 MHz
Channel 3: 60 to 66 MHz
Channel 4: 66 to 72 MHz
Channel 5: 76 to 82 MHz
Channel 6: 82 to 88 MHz
Channel 7: 174 to 180 MHz
Channel 8: 180 to 186 MHz

and so on. Note several interesting things:

- There is no channel 1.
- Channels are usually adjacent, but there is a 4 MHz break between channels 4 and 5
- Channel 6 is just under the FM band, which starts at 88 MHz
- There is a big break between the "low VHF" channels, channels 2 through 6, and the "high VHF" channels, numbered 7 through 13. The space between them contains the FM broadcast band, aircraft frequencies, the 2-meter amateur band, marine channels, and commercial two-way frequencies.
- The UHF channels begin at 470 MHz and extend up to 806 MHz.

In any given area, to avoid interference between them, two TV stations will never be assigned adjacent channels. In New York City, for example, there are TV stations on channels 2, 4, 5, 7, 9, 11, and 13. How can there be stations on both channels 4 and 5? Only because there is a 4-MHz-wide space between them — they are not really adjacent.

If you have cable TV, you may be wondering why there can be stations on adjacent cable channels. The reason is that most modern TV sets can (just barely!) separate two adjacent TV stations under the condition that they are both the same strength. The cable network has full control over the strength of all its channels, and can satisfy this condition. On-the-air signals, on the other hand, vary in strength depending on how far you are from the transmitter, and there can often be tremendous differences in signal strength between stations.

While we are discussing TV transmission, this might be a good place to again mention what kind of modula-

Copyright © 2002 by Peter A. Stark - All Rights Reserved

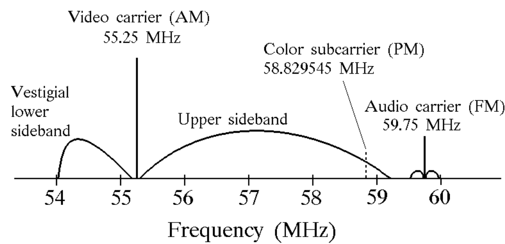

Fig. 8-13. TV channel 2 spectrum

tion is used for TV, and how the 6 MHz channel is divided up.

Fig. 8-13 shows a typical TV channel spectrum; we show channel 2, but all the VHF and UHF TV channels are laid out exactly the same way.

Each TV channel takes exactly 6 MHz of bandwidth; channel 2 uses 54 to 60 MHz. The picture is sent using AM vestigial sideband modulation (which was discussed at the end of Chapter 7), with the AM carrier exactly 1.25 MHz from the bottom frequency of the channel; in channel 2, this places it at 55.25 MHz.

The sound is sent on an FM carrier, whose frequency is exactly 4.5 MHz above the picture carrier; this places it at 59.75 MHz. (If you live in an area with a channel 6, and if you have an FM radio which will tune slightly below the 88 MHz end of the FM band, you can hear the channel 6 TV sound at 87.75 MHz.)

3.579545 MHz above the picture carrier (which is 58.829545 MHz in this case) is a third carrier, which carries the color or *chroma* information. Since this carrier is sent as part of the picture signal, it is called a *subcarrier*. This signal uses phase modulation.

Phase Modulation

PM or *Phase Modulation* is so similar to frequency modulation that equipment designers often use a PM modulator to produce FM. Because of their similarity, it is difficult to clearly explain the difference between FM and PM, so read this section very carefully.

First, let's make sure you fully understand FM, so let's talk about a commercial FM broadcast station at 96.3 MHz. The frequency of its unmodulated FM carrier is its center frequency (96.3 MHz). When the audio signal starts, the frequency varies (deviates) from its center value. At any given instant of time, the frequency change at that instant is proportional to the audio voltage at that instant. For example, suppose the transmitter is designed to produce the full ±75 kHz deviation with an input voltage of ±7.5 volts. The signal frequency might then obey the following table:

Audio voltage (volts)	Frequency change (kHz)	Signal frequency (Mhz)
+7.5	+75	96.375
+1.0	+10	96.310
0	0	96.300
–7.5	–75	96.225

Thus whenever the audio voltage is +1 volt, the transmitter frequency at that instant will be 96.310 MHz. It doesn't matter how the audio voltage got there — the signal could be coming from a 1-volt DC battery, or it could be the peak of a 1-volt sine wave, or it could just be passing through the +1 volt value on its way up or down some strange-looking waveform. At any instant of time, the FM signal frequency depends only on the audio voltage at that same instant.

Another important concept to understand is that, if the audio voltage remains constant for a while, then the signal frequency is also constant (though it may not be at the center frequency.) When the audio signal changes, the signal frequency changes. And the faster the audio changes, the faster the signal frequency changes.

Next, let us discuss phase. Phase measurements are used to compare two or more signals to see if they are "in step" with each other, and if not, how much they differ. Let's consider a real simple-minded example, where someone talks about a parade and says "that marcher is 180 degrees out of step with the others". What is meant is that this particular marcher is doing the exact opposite of everyone else — when everyone else steps forward on their left foot, this marcher steps with his right foot, and so on. Zero degrees phase difference means being in step; 180 degrees difference means doing the exact opposite; 90 degrees difference would mean being "halfway out of step", and so on.

Although phase measurements usually compare two or more signals with each other, we can also use them compare one signal to itself. Let's return to that parade marcher. Suppose you see him (or her) take two steps; based on this, you can predict when the third step should come. If the marcher takes that third step a little sooner than you expected, then that step occurs earlier in phase. But to do that, the steps are coming closer together, so the frequency has also increased. So phase changes go with frequency changes.

Hence frequency modulating a signal produces phase modulation, and phase modulation produces frequency modulation. So what is the difference?

Remember the definition of modulation index:

$$MI = \frac{\text{deviation}}{\text{modulation frequency causing that deviation}}$$

Copyright © 2002 by Peter A. Stark - All Rights Reserved

The difference between FM and PM is this: suppose an announcer steps up to the microphone and whistles a low note (for instance 500 Hz), followed by a high note (for instance 1000 Hz) at the same volume (imagine a sound like "boop-beep".) In both FM and PM, these two notes produce both frequency and phase modulation. But in FM, the deviation for both notes would be the same (meaning that the modulation index is different), whereas in PM, the modulation index is the same (meaning that the deviation is different.)

In practice, things get a bit confused here, because radio stations routinely boost the treble to reduce the noise, so the deviation would change slightly even in FM, but let's ignore that little complication. So, for example, FM might produce 2000 Hz deviation for both notes; the modulation index would then be 2000/500, or 4 for the 500 Hz note, but only 2000/1000, or 2 for the 1000 Hz note.

In PM, on the other hand, the 500 Hz note might produce 1000 Hz deviation (for a modulation index of 1000/500, or 2), while the 1000 Hz note would produce 2000 Hz deviation (for a modulation index of 2000/1000, also equal to 2.)

This greatly affects the bandwidth of the signal. In FM, lower frequencies produce a higher modulation index; this gives more sidebands, but they are closer together, so the total bandwidth stays pretty constant and doesn't depend on the audio frequency. In PM, on the other hand, the modulation index is independent of audio frequency, so the number of sidebands stays the same regardless of audio frequency. But with higher audio frequencies, the sidebands are farther apart, so the bandwidth increases.

The bottom line is that PM is useful for communications-quality audio, whose audio content doesn't go much above 3000 Hz; it would not be good for real hi-fi audio, because the high frequencies in hi-fi audio would require too much bandwidth.

To finish up, here are two more points. First, you can generate FM with a PM transmitter, and vice versa. All you have to do is to process the audio going into the transmitter. For example, an FM transmitter produces too little deviation for high frequencies, so sending the audio through a high-pass filter (which increases the high-frequency treble) will increase the deviation to that of a PM transmitter. Doing the opposite – cutting the treble with a low-pass filter — will make a PM transmitter send out frequency modulation. In fact, many FM transmitter designs do just that; it results in a more stable signal.

Finally, as we said earlier, things get a bit more complicated because of FM stations' efforts to reduce noise. Since most audio signals contain mostly mid-range signals and relatively little energy in the high frequencies, most commercial FM transmitters boost the treble; this process is called *pre-emphasis.* This makes the signal a bit more like PM, but it's OK because of the low amount of high frequencies to begin with. A *de-emphasis* circuit in the receiver then reduces the treble back to the normal level. In the process, the de-emphasis circuit also reduces any hiss which crept into the signal.

Differences between AM, FM, and PM

FM and PM are similar to each other, but fundamentally different from AM. All three have a carrier which is modulated in some way, and that modulation causes sidebands, but how those sidebands interact with the carrier is very different.

Because noise basically affects amplitude, you can remove it from an FM or PM signal by clipping (limiting) the signal to a fixed amplitude; this does not change the frequency or phase and so doesn't affect the modulation. However, you cannot remove this noise from an AM signal without affecting the amplitude, and therefore the modulation itself. FM and PM can therefore provide the better signal-to-noise ratio that listeners want for good music.

But to get the best signal-to-noise ratio requires also a wide deviation; hence FM signals used for music also have fairly much wider bandwidth than most AM signals.

In those communications applications where real hi-fi sound is not required, narrow-band FM (NBFM) with fairly small deviation and a small modulation index can produce bandwidths not much larger than AM; this is the mode that is used for most mobile communications such as amateur repeaters, police and fire communications, business users, and cellular telephones.

FM and PM receivers also experience a "capture effect", which affects how interfering signals affect each other.

For example, suppose that within reach of a receiver there are two transmitters (let's call them A and B), both of which transmit on the same frequency, but A is close to you, while B is farther away and weaker.

If the two transmitters are AM, then their signals will add to each other, and interfere with each other. If the two carrier frequencies are *exactly* the same, then the receiver might hear the audio from both; more likely, the frequencies would be just a bit different, in which case the receiver would get a combination of both signals, plus a *beat note* — an audio tone equal to the difference between the two carrier frequencies. The result would be most unpleasant, and possibly useless.

Copyright © 2002 by Peter A. Stark - All Rights Reserved

If, on the other hand, the two signals are FM or PM, the stronger signal "captures" the receiver, and the weaker signal doesn't interfere at all.

In a typical receiver, the stronger signal must be about 2 to 3 dB stronger than the weaker one to completely capture the receiver; this number is called the *capture ratio*, and depends on the receiver design. If you have an FM receiver in your car, you may have noticed the capture effect when you drive in a region somewhere between two FM stations (call them A and B again) which are on the same frequency. Since signals vary in strength in different places, the relative strengths of A and B also vary. In one spot, A may be stronger and thus capture the receiver; move over a small distance and suddenly B is stronger and thus captures the receiver. Thus your receiver rapidly switches back and forth between stations A and B as you drive.

Continuous Tone Squelch

Capture effect is one reason why walkie-talkies and mobile communication equipment usually use FM rather than AM. Because there is a limited number of available frequencies, many users often share the same frequency. For example, two taxi companies in two nearby towns may share the same frequency. If both transmit at the same time, the capture effect causes each town's taxi radios to hear the local signals, and ignore the distant signals.

Nevertheless, the two taxi companies could still hear each other's signals when only one transmits. There may be times when this is desirable (in fire or police radios, for instance), but it is a nuisance at other times. The usual solution is to add a subaudible tone (near 100 Hz, low enough not to be heard) to the normal voice signals, one frequency for one company, another frequency for the other. Their receivers then use a tone decoder which will "unsquelch" the radio (i.e., turn on the speaker) only when the correct tone is received. Each taxi company's radios will therefore ignore signals from the other company because they have the wrong tone frequency.

There are various names for this *CTS* or *Continuous Tone Squelch* system, but most two-way radio people refer to is as *PL* or *Private Line*, which is Motorola's trademark for the system.

Repeaters and Trunking

On the subject of mobile two-way radio, let's cover several more concepts.

A few decades ago, mobile two-way radio systems mostly used the 30–50 MHz and 150–170 MHz radio bands. Later, 450–470 MHz became popular, and many modern systems operate at frequencies of 800 MHz or higher. At higher frequencies, signals generally carry for shorter distances, and tend not to go around obstructions. Providing good coverage in a large area, or behind and around hills and tall buildings becomes difficult. A common solution is to use a repeater.

A *repeater* is a receiver/transmitter combination located on top of a hill, a tall building, or even a tall tower. Mobile units and walkie-talkies talk to the repeater, which receives the signal and immediately retransmits it on a different frequency; the mobile units and walkie-talkies listen not to the original signal, but to the one coming from the repeater. Since the repeater is located high up (and may have an extra sensitive receiver and a high power transmitter), it can cover a large area.

This can tremendously increase the range. For example, out in flat country, the direct range between two walkie-talkies might only be a mile or two; in hilly terrain or in a city, that might shrink to a fraction of a mile. But with a repeater between them, a walkie-talkie 30 or 40 miles on one direction from the repeater might be able talk to another one the same distance in the opposite direction, for a total distance of 60 or 80 miles between them.

The problem with a repeater is that it uses up two frequencies — one to receive, the other to transmit. A moderately busy police department in a big city, or an organization far out in "the sticks" might be able to justify this, but in most populated areas frequencies are scarce. One solution is to share a *community repeater* among many users. Someone puts up a repeater, and then rents time on it to local taxi companies, oil companies, and other businesses in the area. Each uses a different CTS frequency to keep their transmissions separate.

Many users now share the same frequencies, but every now and then two (or more) users will want to use the repeater at the same time. The modern approach to solving the problem in two-way radio is trunking.

With *trunking*, there are several interlocked community repeaters in an area. Each user that subscribes to the system has a receiver which continuously scans all the repeaters, looking for a signal. When he wants to transmit, his transceiver automatically finds an unused repeater, and transmits on its frequency. This makes much better use of the repeaters, and allows many more users.

Appendix D discusses a similar problem faced by telephone companies, which also reuse a limited number of channels for a large number of users, with the result that occasionally a user gets a busy signal or has to wait his turn. We can get an idea of how trunking im-

Copyright © 2002 by Peter A. Stark - All Rights Reserved

proves efficiency by using some of the results from that discussion.

Suppose we have one repeater and five users. Assume that, on the average, each user makes a 5-second transmission about once every 500 seconds. With five users, there will be a 5-second transmission about every 100 seconds. (If you've read Appendix D, that makes $5/100 = 0.05$ erlangs.) According to Appendix D, the blocking is 5%, meaning that 5% of the time a user wanting to transmit will find the repeater busy, and will have to wait.

Now assume that business is good, and you get 5 more users. If you put all ten users on the same repeater, the blocking rate would rise to about 9%, which might make your customers very unhappy. So (not using trunking) you build a new repeater and put the 5 new users on that system. It also now has 5% blocking. You now have two repeaters and 10 users, with each one getting a "busy" 5% of the time.

Let's use trunking instead: Put all 10 users on two trunked repeaters, calling them A and B. Whenever a user wants to transmit, his equipment tries channel A first. 91% of the time he gets through, while 9% of the time channel A is busy. But his equipment immediately switches to channel B and tries that. Since B only gets an occasional call, it is blocked about 5% of the time, so 5% of the 9% of the calls are blocked. As a result, the overall blockage is only about $0.05 \times 0.09 = 0.0045$ of the time, which is about 0.45% of the time. So more than 99.5% of the calls now get through. Another way to look at it is that adding the second trunked repeater has reduced the blocking percentage by more than a factor of 10.

Still another way of analyzing the situation is to assume that the original 5% blocking would have been acceptable — how many users could be put on the two trunked repeaters and still get only 5% blocking? This requires more reference to Appendix D, but we find that we can now handle almost 38 users. Thus two trunked repeaters can handle almost four times as many users as two untrunked repeaters, with the same 5% blocking percentage.

Copyright © 2002 by Peter A. Stark - All Rights Reserved