

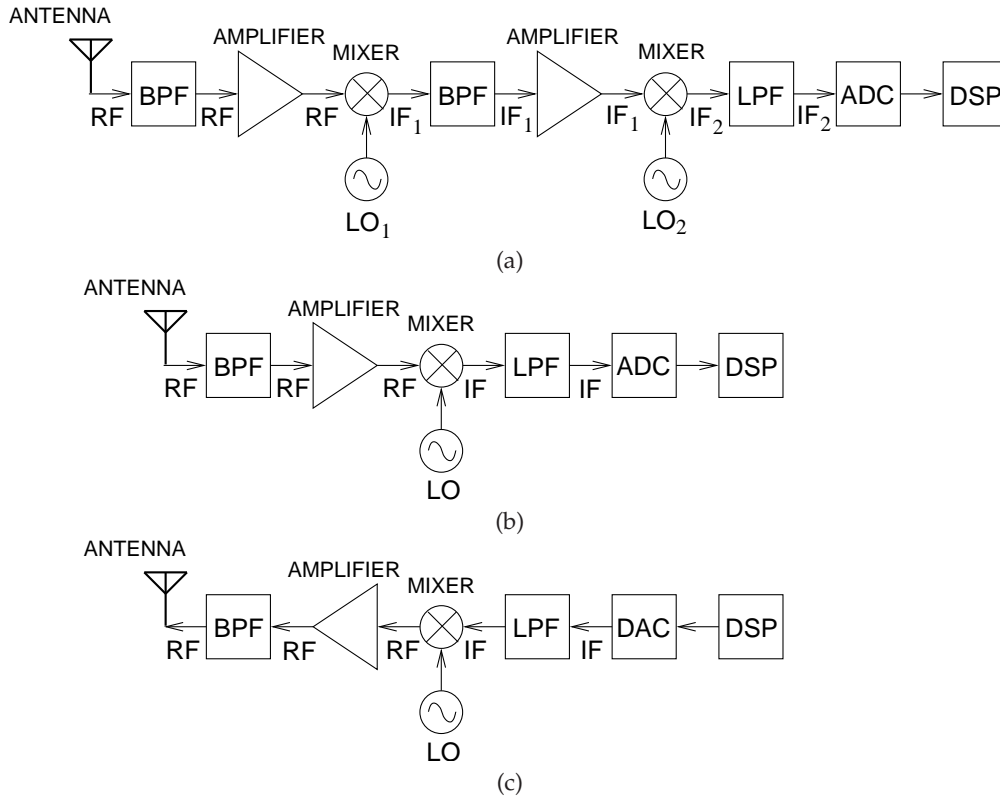
# Modulation, Transmitters and Receivers

1.1	Introduction .....	1
1.2	RF Signals .....	4
1.3	Analog Modulation .....	5
1.4	Digital Modulation .....	13
1.5	Amplifiers .....	30
1.6	Noise and Nonlinear Distortion .....	50
1.7	Active Switch .....	59
1.8	Mixers .....	62
1.9	Early Receiver Technology .....	66
1.10	Modern Transmitter Architectures .....	68
1.11	Modern Receiver Architectures .....	72
1.12	Summary .....	82

## 1.1 Introduction

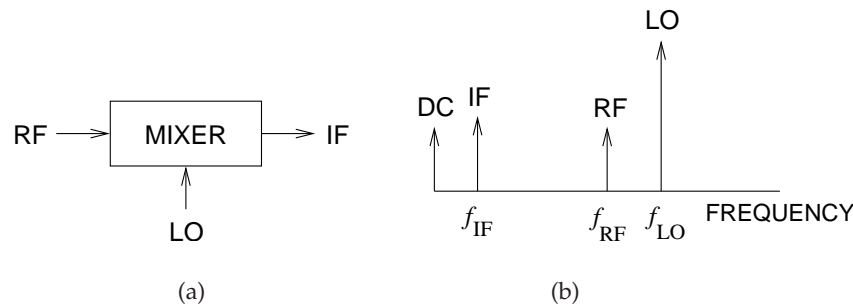
The frontend of a radio frequency (RF) communication receiver combines a number of subsystems in cascade to achieve several objectives. Filters and matching networks provide frequency selectivity to eliminate interfering signals. Amplifiers manage noise levels by boosting both received signals and signals to be transmitted. Mixers coupled with oscillators translate the modulated information from one frequency to another.

There are only a few types of receiver and transmitter architectures. In a receiver, the central idea is to take information superimposed on an RF signal or carrier and convert it to a lower frequency form which can be directly applied to a speaker or digitized. In a cellular communication system, the low-frequency signal, often called the baseband signal, could have a bandwidth of 30 kHz to 5 MHz and the carrier frequency could be 500 MHz to 2 GHz. A transmitter takes the baseband signal and superimposes it on an RF carrier which can be more easily radiated



**Figure 1-1** Unilateral RF frontend: (a) a receiver with two mixing stages; (b) a receiver with one heterodyne stage; and (c) a one-stage transmitter.

into space and propagates easily from one antenna to another. The essential receiver and transmitter architectures are shown in Figure 1-1. In a **receiver** mixers down-convert information superimposed on an RF carrier to a lower frequency that can be directly connected to speakers or digitized by an analog-to-digital converter (ADC). With a transmitter, the low-frequency information-bearing signal is translated to a frequency that can be more easily radiated. The most common receiver architecture is shown in Figure 1-1(a). First, an antenna collects a broad portion of the electromagnetic spectrum. Antennas have relatively low frequency selectivity (they have broad bandwidth) and unwanted signal levels can be large, so additional filtering by a bandpass filter (BPF) is required to reduce the range of voltages presented to the first amplifier. Eventually this signal is digitized by an ADC but to do this the frequency of the information-carrying part of the signal must be reduced. The stepping down of frequency is accomplished by a mixer stage. With the mixer driven by a large local oscillator (LO) signal, the output at the intermediate frequency (IF) is at the difference frequency of the RF and LO (see Figure 1-2). Thus  $f_{IF} = f_{RF} - f_{LO}$  (although sometimes the LO is above the RF so that  $f_{IF} = f_{LO} - f_{RF}$ ). LOs generally have noise close to the operating frequency so that there is a limit on how close the RF and LO can be in frequency without oscillator noise appearing at the IF. If there is a single mixer, then the IF

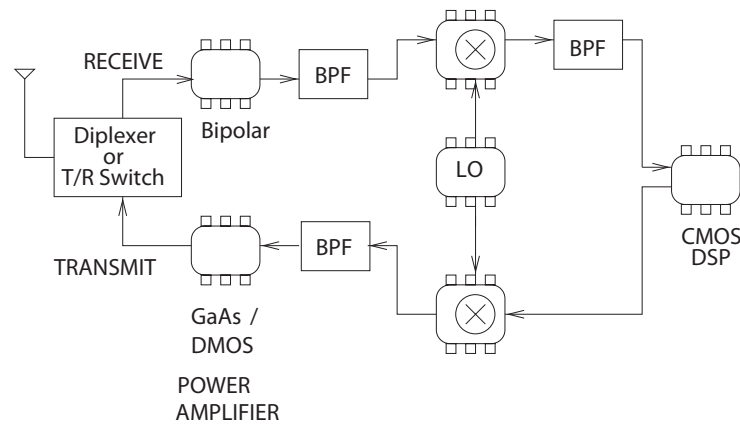


**Figure 1-2** Simple mixer circuit: (a) block diagram; and (b) spectrum.

may still be too high. A solution is to use two stages of mixing. A BPF between the mixing (or heterodyning) stages further blocks unwanted signals. Eventually a lowpass filter (LPF) allows only the final IF (here  $IF_2$ ) to be presented to the ADC. Once digitized, it is possible to further filter the intended signal which originally appeared as modulation at the RF. A one-stage receiver, see Figure 1-1(b), generally requires a better ADC, but the elimination of a mixing stage reduces cost and size. The architecture of a transmitter is similar, with a key difference being the digital-to-analog converter (**DAC**) (see Figure 1-1(c)).

The major **active elements** in the RF **frontend** of both the transmitter and receiver are the **amplifiers**, **mixers**, and **oscillators**. These subsystems have much in common using nonlinear devices to convert power at DC to power at RF. In the case of mixers, power at the **local oscillator (LO)** is also converted to power at RF. The frontend of a typical cellphone is shown in Figure 1-3. The components here are generally implemented in a module and use different technologies for the various elements, optimizing cost and performance. There are many variants of the architecture shown here. At one extreme a module is used with all of the components packed in a shielded structure perhaps 1 cm on a side and 2–3 mm thick. Another extreme is a single-chip implementation, usually in **BiCMOS** (**bipolar** with complementary metal oxide semiconductor, **CMOS**), **SiGe** (silicon germanium) technology, or high performance CMOS called **RF CMOS**. However, it is necessary to use a gallium arsenide GaAs device to efficiently achieve the hundreds of milliwatts typically required.

Return now to the mixer-based **transceiver** (for receiver and transmitter) architecture shown in a multichip form in Figure 1-3. Here, a single antenna is used, and either a **duplexer** (a combined lowpass and highpass filter) or a switch is used to separate the (frequency-spaced) transmit and receive paths. If the system protocol requires transmit and receive at the same time, a duplexer is required to separate the transmit and receive paths. This filter tends to be large, lossy, or costly (depending on the technology used). Consequently a transistor **switch** is preferred if the transmit and receive signals operate in different time slots. In the receive path, a CMOS or BiCMOS chip initially amplifies the low-level received signal, and so low noise is important. This amplifier is thus called a **low-noise amplifier (LNA)**. The amplified receive signal is then bandpass filtered and frequency down-converted by a mixer (indicated by a circle with a cross in it) to IF that can be sampled by an **ADC** to produce a digital signal that is further processed by



**Figure 1-3** RF frontend organized as multiple chips.

digital signal processing (**DSP**). Variants of this architecture include having two down-conversion stages, and a variant with no mixing that relies instead on direct conversion of the receive signal using a subsampling ADC. In the transmit path the architecture is reversed, with a DAC driven by the DSP chip that produces an information-bearing signal at the IF which is then frequency up-converted by a mixer, bandpass filtered, and amplified by what is called a power amplifier to generate the hundreds of milliwatts required. An alternative transmitter design is **direct digital synthesis (DDS)**, which bypasses the conversion stage. Direct conversion and DDS are difficult to implement, but are essential for the highly desired single or few chip solution.

This chapter describes the operation and design strategies for the RF frontend architecture of Figure 1-3, looking at amplifiers, mixers, switches, and oscillators. This architecture is used in most high-performance RF and microwave communication and radar systems. While the subsystems are preferably linear at RF, this can only be approximated, as the active devices used are intrinsically nonlinear. Performance is limited fundamentally by distortion, which is related to the characteristics of the RF signal, and this in turn is determined by the modulation scheme that impresses information on an RF carrier.

## 1.2 RF Signals

Radio frequency communication signals are engineered to trade off efficient use of the electromagnetic (**EM**) spectrum with the complexity and performance of the RF hardware required to process them. The process of converting baseband (or low-frequency) information to RF is called modulation of which there are two types: analog and digital modulation. In analog modulation, the RF signal has a continuous range of values; in digital modulation, the output has a number of prescribed discrete states. There are just a few modulation schemes that achieve the optimum trade-offs of spectral efficiency and ease of use with hardware complexity. The major modulation schemes are

Analog modulation	
AM	Amplitude modulation
FM	Frequency modulation
PM	Phase modulation
Digital modulation	
FSK	Frequency shift keying
PSK	Phase shift keying
MSK	Minimum shift keying (a form of FSK)
GMSK	Minimum shift keying using Gaussian filtered data
BPSK	Binary phase shift keying
QPSK	Quadrature PSK (QPSK is also referred to as quaternary PSK, quadriphase PSK, and quadra PSK )
$\pi/4$ -DQPSK	$\pi/4$ Differential encoded QPSK
OQPSK	Offset QPSK
8PSK	8-state phase shift keying
$3\pi/8$ -8PSK	$3\pi/8$ , 8-state phase shift keying
16PSK	16-state phase shift keying
QAM	Quadrature amplitude modulation

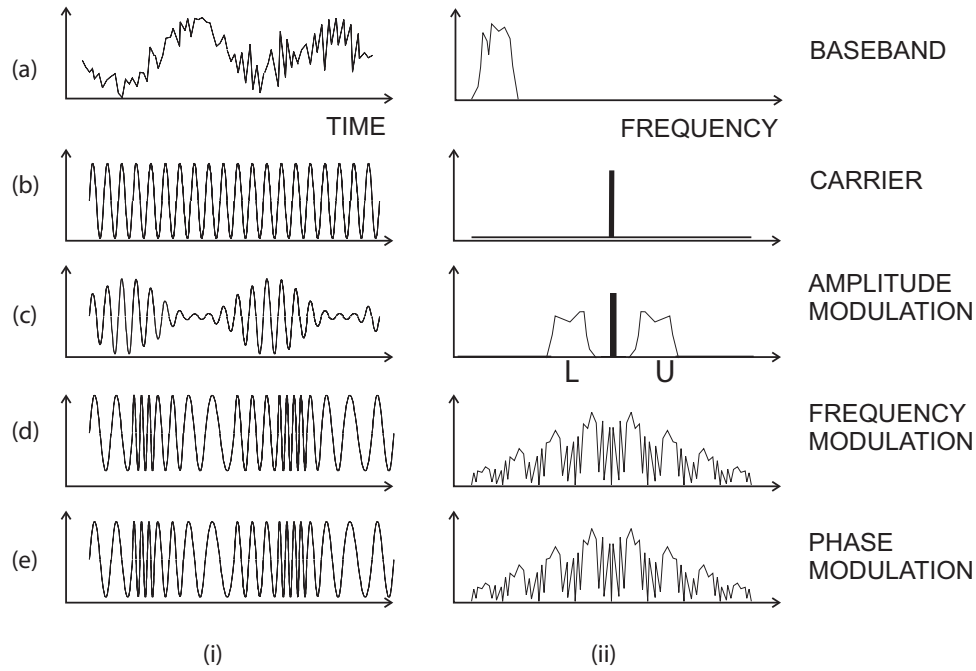
Frequency modulation, and the similar PM modulation schemes, are used in analog cellular radio. With the addition of legacy AM, the three schemes are the bases of analog radio. The other schemes are used in digital radio including digital cellular radio. GMSK is used in the GSM cellular system and is a form of FSK and produces a constant amplitude modulated signal. The FM, FSK, GMSK, and PM techniques produce constant RF envelopes, thus no information is contained in the amplitude of the signal. Therefore errors introduced into the amplitude of the system are of no significance and so efficient saturating-mode amplifiers such as class C can be used. So there is a trade-off in the complexity of RF design, choice of modulation format and battery life. In contrast, the MSK,  $\pi/4$ -DQPSK,  $3\pi/8$ -8PSK, and QAM techniques do not result in constant RF envelopes, so information is contained in the amplitude of the RF signal. Thus more sophisticated RF processing hardware is required.

## 1.3 Analog Modulation

Wireless modulation formats in conventional narrowband radio are based on modifying the properties of a carrier by slowly varying the amplitude and phase of the carrier. The waveforms and spectra of common analog modulation formats are shown in Figure 1-4.

### 1.3.1 Amplitude Modulation, AM

Amplitude modulation (AM) is the simplest analog modulation scheme to implement. Here a signal is used to slowly vary the amplitude of the carrier according to the level of the modulating signal. The modulating signal is generally referred to as the baseband signal and it contains all of the information to be transmitted or interpreted. The waveforms in Figure 1-4 are stylized as the



**Figure 1-4** Analog modulation showing (i) waveform and (ii) spectrum for (a) baseband signal; (b) carrier; (c) carrier modulated using amplitude modulation; (d) carrier modulated using frequency modulation; and (e) carrier modulated using phase modulation.

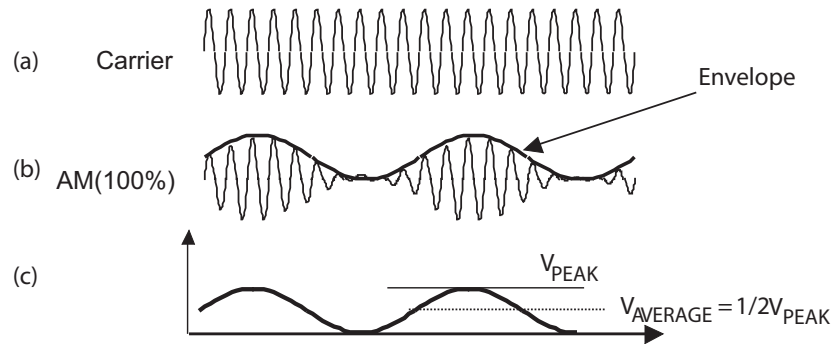
variation in the carrier is relatively fast. They are presented this way so that the effects of modulation can be more easily interpreted. The baseband signal (Figure 1-4(a)) is shown as having a period that is not too far away from the period of the carrier (Figure 1-4(b)). In reality, there would be hundreds or thousands of RF cycles for each cycle of the baseband signal so that the frequency of the baseband signal would have frequency components which are a tiny fraction of the frequency of the carrier.

With AM (Figure 1-4(c)) the amplitude of the carrier is modulated and this results in a broadening of the spectrum of the carrier, as shown in Figure 1-4(c)(ii). This spectrum contains the original carrier component and upper and lower sidebands designated as U and L, respectively. In AM, the two sidebands contain identical information, so all the information would be transmitted if the carrier and one of the sidebands were suppressed. With the carrier present, it is easy to receive a signal by bandpass filtering the incoming modulated signal, rectifying the result, and then lowpass filtering the rectified signal to remove harmonics of the baseband signal.

An AM signal  $x(t)$  has the form

$$x(t) = A_c [1 + my(t)] \cos \omega_c t, \quad (1.1)$$

where  $m$  is called the modulation index and  $y(t)$  is the baseband information-bearing signal that has frequency components which are below the carrier radian



**Figure 1-5** AM showing the relationship between the carrier and modulation envelope: (a) carrier; (b) 100% amplitude modulated carrier; and (c) modulating or baseband signal.

frequency  $\omega_c$ . Provided that  $y(t)$  varies slowly relative to the carrier, that is, the frequency components of  $y(t)$  are significantly below the carrier frequency,  $x(t)$  looks like a carrier whose amplitude varies slowly. To get an idea of how slowly the amplitude varies in actual systems, consider an AM radio that broadcasts at 1 MHz (which is in the middle of the AM broadcast band). The highest frequency component of the modulating signal corresponding to voice is about 4 kHz. Thus the amplitude of the carrier takes 250 carrier cycles to go through a complete amplitude variation. At all times a cycle of the carrier appears to be periodic, but in fact it is not quite. It is common to refer to the modulated carrier as being quasi-periodic and to the apparent carrier as being the pseudo-carrier.

The concept of the envelope of a modulated RF signal is introduced in Figure 1-5. Figure 1-5(a) is the carrier; the AM-modulated carrier is shown in Figure 1-5(b). The outline of the modulated carrier is called the envelope, and for AM this is identical to the modulating signal. Both the envelope and the modulating signal are shown in Figure 1-5(c). At the peak of the envelope, the RF signal has maximum power (considering the power of a single RF cycle). Since we are dealing with 100% AM modulation,  $m = 1$  in Equation (1.1) and there is no RF power when the envelope is at its minimum.

One of the characteristics of various modulation formats is the ratio of the power of the signal when the carrier is at its peak (i.e., the power in one cycle of the carrier when the envelope is at its maximum) relative to its average value (the power averaged over all time). This is called the **peak-to-average ratio (PAR)** and is a good indicator of how sensitive a modulation format is to the effects of nonlinearity of the RF hardware.

It is complex to determine the PAR for a general signal, but a good estimate can be obtained by considering that the modulating signal is a sinewave. Let  $y(t)$  ( $= \cos \omega_m t$ ) be a cosinusoidal modulating signal with radian frequency  $\omega_m$ . Then (for AM)

$$x(t) = A_c [1 + m \cos \omega_m t] \cos \omega_c t. \quad (1.2)$$

Thus if just one quasi-period of this signal is considered (i.e., one variation of the modulated signal at the carrier frequency), then the signal has a power that varies with time.

Consider a voltage  $v(t)$  across a resistor of conductance  $G$ . The power of the signal, or the average power, must be determined by integrating over all time, which is work, and dividing by the time period yields the average power:

$$P_{\text{avg}} = \lim_{\tau \rightarrow \infty} \int_{-\tau}^{\tau} \frac{1}{2\tau} G v^2(t) dt. \quad (1.3)$$

Now, if  $v(t)$  is a cosinusoid,  $v(t) = A \cos \omega t$ , then

$$\begin{aligned} P_{\text{avg}} &= \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} A_c^2 G \cos^2(\omega t) dt \\ &= \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} A_c^2 G \frac{1}{2} [1 + \cos(2\omega t)] dt \\ &= \frac{1}{2} A_c^2 G \left\{ \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1 dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega t) dt \right\} = \frac{1}{2} A_c^2 G \end{aligned} \quad (1.4)$$

In the above equation, a useful equivalence has been employed by observing that the infinite integral of a cosinusoid can be simplified to just integrating over one period,  $T = 2\pi/\omega$ :

$$\lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos^n(\omega t) dt = \frac{1}{T} \int_{-T/2}^{T/2} \cos^n(\omega t) dt \quad (1.5)$$

where  $n$  is a positive integer. In power calculations there are a number of other useful simplifying techniques based on **trigonometric identities**. Some of the ones that will be used are the following:

$$\begin{aligned} \cos A \cos B &= \frac{1}{2} [\cos(A - B) + \cos(A + B)] \\ \cos^2 A &= \frac{1}{2} [1 + \cos(2A)] \end{aligned} \quad (1.6)$$

$$\lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos \omega t dt = \frac{1}{T} \int_{-T/2}^{T/2} \cos(\omega t) dt = 0 \quad (1.7)$$

$$\frac{1}{T} \int_{-T/2}^{T/2} \cos^2(\omega t) dt = \frac{1}{T} \int_{-T/2}^{T/2} \frac{1}{2} [\cos(2\omega t) + \cos(0)] dt \quad (1.8)$$

$$\begin{aligned} &= \frac{1}{2T} \left[ \int_{-T/2}^{T/2} \cos(2\omega t) dt + \int_{-T/2}^{T/2} 1 dt \right] \\ &= \frac{1}{2T} (0 + T) = \frac{1}{2}. \end{aligned} \quad (1.9)$$

More trigonometric identities are given in appendix A.3 on page 576. Also, when cosinusoids  $\cos At$  and  $\cos Bt$  having different frequencies ( $A \neq B$ ) are multiplied together, then

$$\int_{-\tau}^{\tau} \cos At \cos Bt dt = \int_{-\tau}^{\tau} [\cos(A + B)t + \cos(A - B)t] dt = 0, \quad (1.10)$$



and, in general, if  $A \neq B \neq 0$ ,

$$\int_{-\infty}^{\infty} \cos At \cos^n Bt dt = 0. \quad (1.11)$$

Now the discussion returns to characterizing an AM signal by considering long-term average power and the short-term power of the signal. The maximum amplitude of the pseudo-carrier at its peak amplitude is, from Equation (1.2),

$$x_p(t) = A_c [1 + m] \cos \omega_c t. \quad (1.12)$$

Then the power ( $P_{\text{peak}}$ ) contained in the peak pseudo-carrier is obtained by integrating over one period:

$$\begin{aligned} P_{\text{peak}} &= \frac{1}{T} \int_{-T/2}^{T/2} Gx^2(t) dt = \frac{1}{T} \int_{-T/2}^{T/2} A_c^2 G (1 + m)^2 \cos^2(\omega_c t) dt \\ &= A_c^2 G (1 + m)^2 \frac{1}{T} \int_{-T/2}^{T/2} \cos^2(\omega_c t) dt = \frac{1}{2} A_c^2 G (1 + m)^2. \end{aligned} \quad (1.13)$$

The **average power** ( $P_{\text{avg}}$ ) of the modulated signal is obtained by integrating over all time, so

$$\begin{aligned} P_{\text{avg}} &= \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} Gx^2(t) dt \\ &= A_c^2 G \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \{[1 + m \cos(\omega_m t)] \cos(\omega_c t)\}^2 dt \\ &= A_c^2 G \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \{[1 + 2m \cos(\omega_m t) + m^2 \cos^2(\omega_m t)] \cos^2(\omega_c t)\} dt \\ &= A_c^2 G \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} [\cos^2(\omega_c t) + 2m \cos(\omega_m t) \cos^2(\omega_c t) \\ &\quad + m^2 \cos^2(\omega_m t) \cos^2(\omega_c t)] dt \\ &= A_c^2 G \left[ \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos^2(\omega_c t) dt \right. \\ &\quad + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 2m \cos(\omega_m t) \cos^2(\omega_c t) dt \\ &\quad \left. + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} m^2 \cos^2(\omega_m t) \cos^2(\omega_c t) dt \right] \\ &= A_c^2 G \left[ \frac{1}{2} + 0 + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} m^2 \cos^2(\omega_m t) \cos^2(\omega_c t) dt \right] \\ &= A_c^2 G \left\{ \frac{1}{2} + m^2 \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \frac{1}{4} [1 + \cos(2\omega_m t)] [1 + \cos(2\omega_c t)] dt \right\} \end{aligned}$$

$$\begin{aligned}
P_{\text{avg}} &= A_c^2 G \left\{ \frac{1}{2} + \right. \\
&\quad \left. \frac{m^2}{4} \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} [1 + \cos(2\omega_m t) + \cos(2\omega_c t) + \cos(2\omega_m t) \cos(2\omega_c t)] dt \right\} \\
&= A_c^2 G \left\{ \frac{1}{2} + \frac{m^2}{4} \left[ \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} 1 dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_m t) dt \right. \right. \\
&\quad \left. \left. + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_c t) dt + \lim_{\tau \rightarrow \infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} \cos(2\omega_m t) \cos(2\omega_c t) dt \right] \right\} \\
&= A_c^2 G [1/2 + m^2(1/4 + 0 + 0 + 0)] \\
&= \frac{1}{2} A_c^2 G (1 + m^2/2). \tag{1.14}
\end{aligned}$$

So the PAR of an AM signal (i.e.  $\text{PAR}_{\text{AM}}$ ) is

$$\text{PAR}_{\text{AM}} = \frac{P_{\text{peak}}}{P_{\text{avg}}} = \frac{\frac{1}{2} A_c^2 G (1 + m)^2}{\frac{1}{2} A_c^2 G (1 + m^2/2)} = \frac{(1 + m)^2}{1 + m^2/2}.$$

For 100% AM described by  $m = 1$ , the PAR is

$$\text{PAR}_{100\% \text{AM}} = \frac{(1 + 1)^2}{1 + 1^2/2} = \frac{4}{1.5} = 2.667 = 4.26 \text{ dB}. \tag{1.15}$$

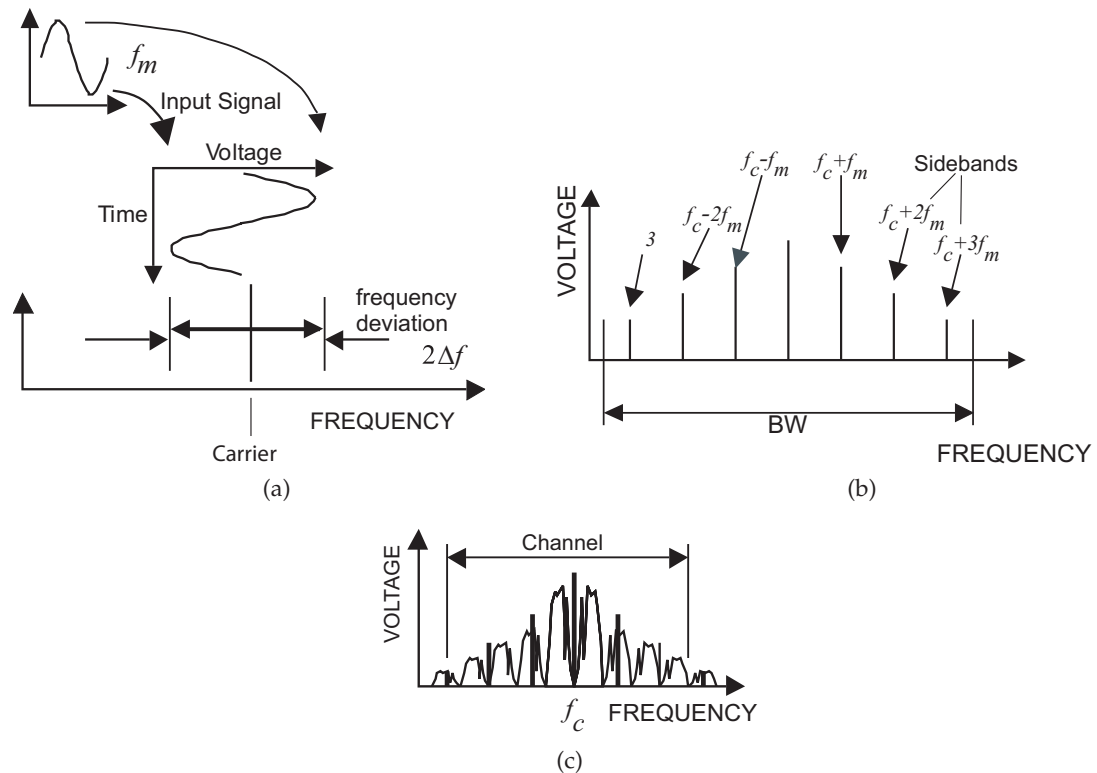
In expressing the PAR in decibels the formula  $\text{PAR}_{\text{dB}} = 10 \log_{10}(\text{PAR})$  was used, as the PAR is a power ratio. As an example, for 50% AM described by  $m = 0.5$ , the PAR is

$$\text{PAR}_{50\% \text{AM}} = \frac{(1 + 0.5)^2}{1 + 0.5^2/2} = \frac{2.25}{1.125} = 2 = 3 \text{ dB}. \tag{1.16}$$

The PAR is an important attribute of a modulation format and impacts the types of circuit designs that can be used. It is much more challenging to achieve low levels of distortion when the PAR is high.

It is tempting to consider if the lengthy integrations can be circumvented. Powers can be added if the signal components (the tones making up the signal) are uncorrelated. If they are **correlated**, then the complete integrations are required.<sup>1</sup> Consider two uncorrelated sinusoids of (average) powers  $P_1$  and  $P_2$  then the average power of the composite signal is  $P_{\text{avg}} = P_1 + P_2$ . However, in determining peak power, the RF cycle where the two sinusoids align is considered, and here the voltages add to produce a sinewave with a higher amplitude. So peak power applies to just one RF pseudo-cycle. Generally the voltage amplitude of the two sinewaves would be added and then the power calculated. If the uncorrelated carriers are modulated and the modulating signals (the baseband signals) are uncorrelated then the average power can be determined in the same way, but the peak power calculation is much more complicated. The integrations are the only calculations that can always be relied on. They can be used with all signals, including digitally modulated signals.

<sup>1</sup> For the purposes here, two signals are uncorrelated if the integral of their product over all time and all offsets is zero. That is,  $x(t)$  and  $y(t)$  are uncorrelated if  $C = \int_{-\infty}^{+\infty} x(t)y(t + \tau) dt = 0$  for all  $\tau$ , otherwise they are correlated (or partly correlated). For a more complete definition see reference [1].

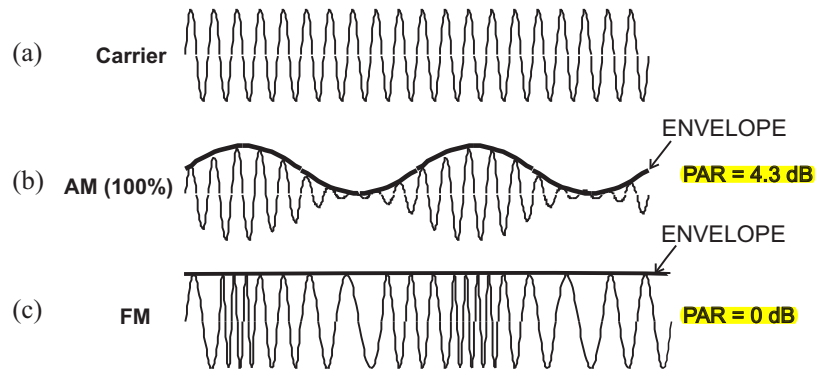


**Figure 1-6** Frequency modulation by a sinewave: (a) signal varying the frequency of carrier; (b) spectrum of the resulting waveform; and (c) spectrum when modulated by a continuous baseband signal.

### 1.3.2 Phase and Frequency Modulation, PM and FM

The two other analog modulation schemes commonly used are phase modulation (PM) (Figure 1-4(e)) and frequency modulation (FM) (Figure 1-4(d)). The signals produced by the two schemes are identical; the difference is how the signals are generated. In PM, the phase of the carrier depends on the instantaneous level of the baseband signal. In FM, the amplitude of the baseband signal determines the frequency of the carrier. The result in both cases is that the bandwidth of the time-varying signal is spread out, as seen in Figure 1-6. A receiver must compress the spread-out information to recreate the original narrowband signal, and this can be thought of as processing gain, as the compression of correlated signals significantly increases the tolerance to noise. As will be seen, processing gain is essential in digital radio, which uses digital modulation. The peak amplitude of the RF phasor is equal to the average amplitude and so the PAR is 1 or 0 dB. A summary of the PAR of the primary analog modulated signals is given in Figure 1-7.

Frequency modulation was invented by Edwin H. Armstrong and patented in 1933. FM is virtually static free and clearly superior to AM radio. However, it was not immediately adopted largely because AM radio was established in the 1930's, and the adoption of FM would have resulted in the scrapping of a large installed



**Figure 1-7** Comparison of 100% AM and FM highlighting the envelopes of both: (a) carrier; (b) AM signal with envelope; and (c) FM or PM signal with the envelope being a straight line or constant.

infrastructure (seen as a commercial catastrophe) and so the introduction of FM was delayed by decades. The best technology does not always win immediately! Commercial interests and the interests of those heavily invested in an alternative technology have a great deal to do with the success of a technology.

### Carson's Rule

Frequency and phase modulated signals have unlimited bandwidth but the information content of the sidebands drops off rapidly. The bandwidth required to reliably transmit a PM or FM signal is subjective but the best accepted criterion is called Carson's bandwidth rule or just Carson's Rule [2,3]. It provides an estimate of the bandwidth capturing approximately 98% of the energy when a carrier is frequency or phase modulated by a continuous spectrum baseband signal. An FM signal is shown in Figure 1-6. In particular, Figures 1-6(a) and 1-6(b) show the FM function and then the spectrum that results when a single sinewave modulates the frequency of a carrier. As time passes, the carrier moves up and down in frequency synchronously with the level of the input baseband signal. The level (typically voltage) of the baseband signal determines the frequency deviation of the carrier from its unmodulated value. The frequency shift when the modulating signal is at its maximum amplitude is called the peak frequency deviation,  $\Delta f$ , and the maximum frequency of the modulating frequency is  $f_m$ . Figure 1-6(c) shows the spectrum that results when the modulating signal, or baseband signal, is continuous. There are multiple sidebands, with the relative strength of each dependent on a Bessel function of the highest modulation frequency,  $f_m$ , and the maximum frequency deviation,  $\Delta f$ . Carson's Rule, derived from these considerations, is

$$\text{Bandwidth required} = 2 \times (f_m + \Delta f). \quad (1.17)$$

## Narrowband and Wideband FM

The FM signal, as used in FM broadcast radio, is also called wideband FM, as the maximum frequency deviation is much greater than the highest frequency of the modulating or baseband signal, that is,  $\Delta f \gg f_m$ . A more spectrally efficient form of FM is called narrowband FM, where  $\Delta f \ll f_m$ . Narrowband FM was developed as a more bandwidth efficient form of FM, but of course digital radio has passed this now and narrowband FM is no longer an important modulation type. The trade-off is that narrowband FM, as opposed to wideband FM, requires more sophisticated demodulation and hence more complex circuits are required. It should also be noted that FM as used in conventional FM broadcast radio is being phased out so that spectrum can be used more efficiently.

### 1.3.3 Two-Tone Signal

A **two-tone signal** is a signal which is the sum of two cosinusoids. Thus

$$y(t) = X_A \cos(\omega_A t) + X_B \cos(\omega_B t) \quad (1.18)$$

is a two-tone signal. Generally the frequencies of the two tones are close with the concept being that the two tones both fit within the passband of a bandpass filter, so it would be reasonable to assume that the individual tones have frequencies that are within 1% of each other. A two-tone signal is not a form of modulation but is commonly used to characterize the performance of RF systems. The composite signal would then look like a slowly varying pseudo-carrier not unlike an AM signal. The tones are uncorrelated so that the average power of the composite signal,  $y(t)$ , is the sum of the powers of each of the individual tones. The peak power of the composite signal is the peak pseudo-carrier, so  $y(t)$  has a peak amplitude of  $X_A + X_B$ . Similar concepts apply to three-tone and  $n$ -tone signals.

#### Example 1.1 PAR of a Two-Tone Signal

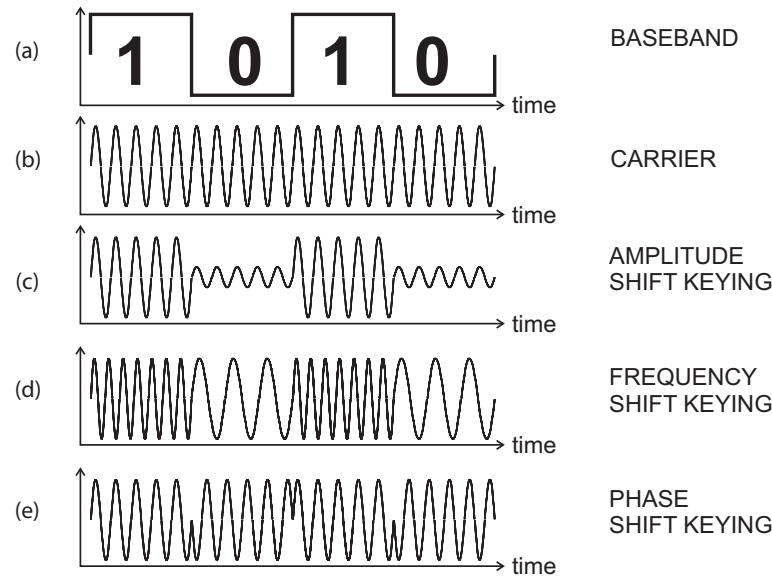
What is the PAR of a two-tone signal with both tones having equal amplitude?

**SOLUTION:** Let  $X_A = X_B = X$ , the peak pseudo-carrier has amplitude  $2X$ , and so the power of the peak RF carrier is  $(2X)^2 = 4X^2$ . The average power is proportional to  $X_B^2 + X_B^2 = X^2 + X^2 = 2X^2$ , as each one is independent of the other, and so the powers can be added.

$$\text{PAR} = \frac{4X^2}{2X^2} = 2 = 3 \text{ dB}. \quad (1.19)$$

## 1.4 Digital Modulation

Digital modulation was first employed in sending telegraph signals wirelessly in which a carrier was switched, or keyed, on and off to create pulses of the carrier signal. This modulation is now known as amplitude shift keying (**ASK**), but today

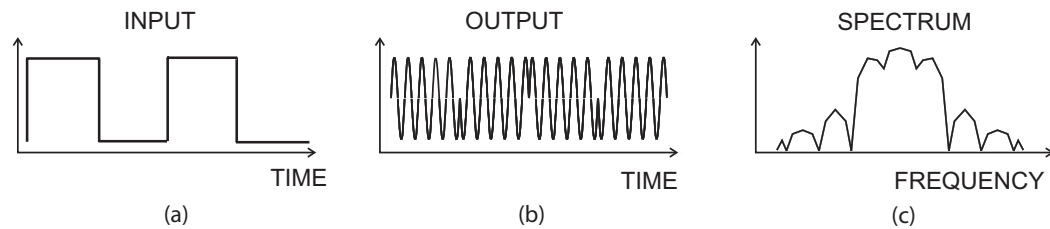


**Figure 1-8** Modes of digital modulation: (a) modulating bit stream; (b) carrier; (c) carrier modulated using amplitude shift keying (ASK); (d) carrier modulated using frequency shift keying (FSK); and (e) carrier modulated using binary phase shift keying (BPSK).

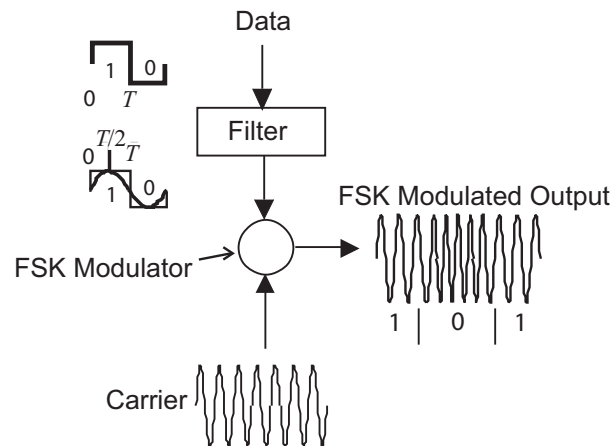
this scheme is little used. Several digital modulation formats are shown in Figure 1-8. The fundamental characteristic of **digital modulation** is that there are discrete states, each of which defines a symbol, with a symbol representing one or more bits. In Figure 1-8, there are only two states representing one of two values for a bit (0 or 1). With multiple states, groups of bits can be represented. There are many digital modulation formats that have proved successful and many of these are considered below. In modern communication schemes it is important to be able to recover the original carrier, so it is important that the amplitude of the carrier not be small for an extended period of time as it is in the ASK scheme illustrated in Figure 1-8(c)

### 1.4.1 Phase Shift Keying, PSK

The waveforms and spectrum of a PSK modulated signal are shown in Figure 1-9. The incoming baseband bit stream (Figure 1-9(a)) is lowpass filtered and used to modulate the phase of a **carrier** (Figure 1-9(b)). The spectrum of this signal is shown in Figure 1-9(c). The PSK modulation scheme is similar to that represented in Figure 1-10, with the FSK modulator replaced by a PSK modulator which shifts the phase of the carrier rather than its frequency. There are many variants of PSK, with the most fundamental characteristics being the number of phase states (e.g., with  $2^n$  phase states,  $n$  bits of information can be transmitted) and how the phasor of the RF signal transitions from one phase state to another. Generally PSK schemes shape the spectrum of the modulated signal to fit as much energy as possible within a spectral mask. This results in a modulated carrier whose amplitude varies



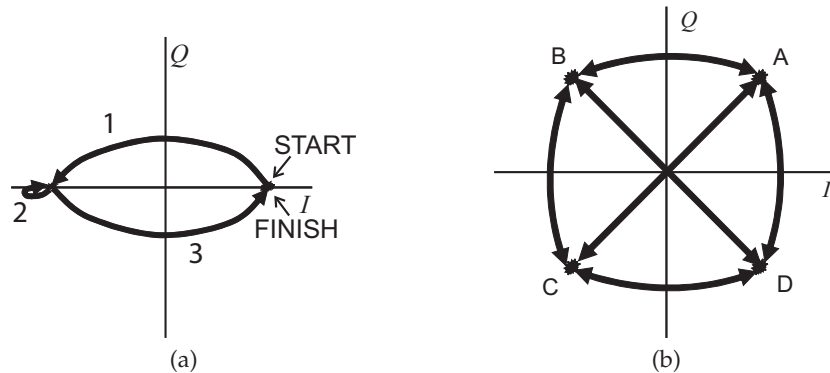
**Figure 1-9** Characteristics of phase shift keying (PSK) modulation: (a) modulating bit stream; (b) the waveform of the carrier modulated using PSK with the phase determined by the 1s and 0s of the modulating bit stream; and (c) the spectrum of the modulated signal.



**Figure 1-10** The frequency shift keying (FSK) modulation system.

(and thus a time-varying envelope). Such schemes require highly linear amplifiers to preserve the amplitude variations of the modulated RF signal. Other schemes orchestrate the phase transitions to achieve a constant envelope modulated RF signal but have lower spectral efficiency. Two approaches to achieving this are first to slow the transitions down, and, second, to eliminate transitions from a phase state to one which is rotated by  $180^\circ$  and so avoid the RF phasor traversing the origin. The result of both approaches is that relatively simple hardware can be used as amplitude distortion is not a problem. So system design affects RF hardware complexity, and the sophistication of available and affordable hardware impacts system design. There are only a few variants that achieve optimum properties and many of these will be considered later in this chapter.

The communication limit of 1 symbol per hertz of bandwidth, the **symbol rate**, comes from the **Nyquist signaling theorem**. Nyquist determined that the number of independent pulses that could be put through a telegraph channel per unit of time is limited to twice the bandwidth of the channel. With a modulated RF carrier, this translates to a pulse of information on the  $I$  (or cosine) component, and a pulse of information on the  $Q$  (or sine) component, in a unit of time equal  $1/\text{bandwidth}$ .



**Figure 1-11** Constellation diagrams with possible transitions: (a) a binary modulation scheme; and (b) QPSK, a four-state phase modulation scheme. Each state is a symbol.

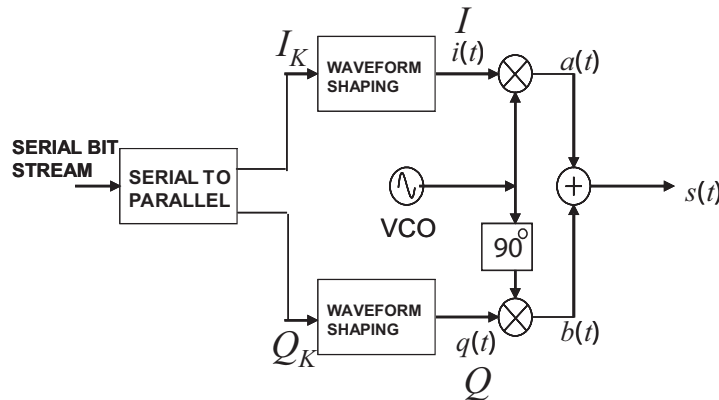
Combining the  $I$  and  $Q$  channels, the phasor can move from one value to another in a unit of time equal to  $1/\text{bandwidth}$ . The phasor transition identifies a symbol, and hence one symbol can be sent per hertz of bandwidth.

### 1.4.2 Binary Phase Shift Keying, BPSK

Phase shift keying demodulation requires more sophisticated signal processing than does FSK. PSK uses prescribed phase shifts to define symbols, each of which can represent one, two or more bits. Binary phase shift keying (BPSK), illustrated in Figure 1-8(e), has one bit per symbol and is relatively a spectrally inefficient scheme, with a maximum spectral efficiency of 1 bit/second/hertz ( $1 \text{ b}\cdot\text{s}^{-1}\cdot\text{Hz}^{-1}$ ). Although spectrally inefficient, it is ideally suited to low-power applications and single-chip implementations, perhaps with an off-chip reference resonator. The typical signal flow is from an antenna, through an RF-tuned amplifier, with quadrature mixing to produce  $I$  and  $Q$  channels which are then lowpass filtered. The filtered  $I$  and  $Q$  channels are then commonly integrated over the duration of a bit. In the most sensitive scheme the  $I$  and  $Q$  channels are oversampled (by an ADC) at a multiple of the bit rate and the signal correlated with the expected zero-crossing.

BPSK is commonly used in pagers and is used in Bluetooth. The operation of BPSK modulation can be described using the constellation diagram shown in Figure 1-11(a). The BPSK constellation diagram indicates that there are two states. These states can be interpreted as the values of  $i(t)$  and  $q(t)$  at the sampling points corresponding to the bit rate. The curves in Figure 1-11(a) indicate three transitions. The states are at the ends of the transitions. If a 1, in Figure 1-11(a), is assigned to the positive  $I$  value and 0 to a negative  $I$  value, then the bit sequence represented in Figure 1-11(a) is '1001'.





**Figure 1-12** Quadrature modulation block diagram indicating the role of pulse shaping.

### 1.4.3 Quadrature Phase Shift Keying, QPSK

QPSK modulation is usually referred to as quadrature PSK although it is also referred to as quaternary PSK, quadriphase PSK, and quadrature PSK. In QPSK wireless systems, good spectral efficiency is obtained by sending more than one bit of information per hertz of bandwidth. Information is encoded in four phase states. Thus referring to QPSK as quadrature phase shift keying is more precise but this is not the common usage. The higher order modulation schemes that achieve more than two states require that the characteristics of the channel be taken into account. The dominant characteristic of the wireless channel are deep fades resulting from destructive interference of multiple reflections. Fades can be viewed as deep amplitude modulation and so it is difficult to transfer information in the amplitude of a carrier. Consequently phase modulation schemes falling in the class of M-ary phase shift keying (MPSK) are most appropriate in the mobile context. In mobile environments there are just a few modulation formats that have been found acceptable. These all fall in the class of either FSK-like schemes or quadrature phase shift keying (QPSK) (also called quadrature phase shift keying as the modulation can be viewed as the superposition of two modulated quadrature carriers). The characteristic of QPSK modulation is that there are four allowable phase states per symbol period, so two bits of information are transmitted per change in the characteristic of the modulated signal. There are many other four-state PSK schemes and there are schemes that have more than four phase states.

Quadrature phase shift keying modulation can be implemented using the quadrature modulator shown in Figure 1-12. The constellation diagram of QPSK is the result of plotting  $i(t)$  and  $q(t)$  on a rectangular graph (or equivalently  $A(t)$  and  $\phi(t)$  on a polar plot) in the generalized modulation circuit of Figure 1-12. More commonly these quantities are referred to as  $I$  and  $Q$ . In Figure 1-12, the input bit stream is first converted into two parallel bit streams. Thus a two-bit sequence in the serial bit stream becomes one  $I_K$  bit and one  $Q_K$  bit. The  $(I_K, Q_K)$  pair constitutes the  $K$ th symbol. A modulation scheme with four allowable states A, B, C, D—is shown in Figure 1-11(b). In the absence of wave-shaping circuits,  $i(t)$  and  $q(t)$  have very sharp transitions and the paths shown in Figure 1-

11 are almost instantaneous. This leads to large spectral spreads in the modulated waveform,  $s(t)$ . So to limit the spectrum of the RF signal  $s(t)$ , the shape of  $i(t)$  and  $q(t)$  is controlled; the waveform is shaped, usually by lowpass filtering. So a pulse-shaping circuit changes binary information into a more smoothly varying signal. Each transition or path in Figure 1-11 represents the transfer of a symbol or minimum piece of information. The best efficiency that can be obtained in point-to-point communication is one symbol per hertz of bandwidth. In the QPSK modulation scheme shown, there are three possible transitions from each point in the constellation in addition to the possibility of no transition. Thus each symbol contains two bits. So the maximum efficiency of this type of modulation scheme is  $2 \text{ b} \cdot \text{s}^{-1} \cdot \text{Hz}^{-1}$  (2 bits/second/hertz of bandwidth). What is actually achieved depends on the pulse shaping circuits and on the criteria used to establish the bandwidth of  $s(t)$ . Various modulation schemes have relative merits in terms of spectral efficiency, tolerance to fading (due to destructive interference), carrier recovery, spectral spreading in nonlinear circuitry, and many other issues that are the realm of communication system theorists.

The waveforms corresponding to the state transitions shown in Figure 1-11(b) most immediately affect the bandwidth of  $s(t)$  and the ability to demodulate the signal. The constellation diagram is analogous to a phase diagram of the carrier signal.<sup>2</sup> Thus, signal trajectories through the origin indicate that the amplitude of the carrier is very small for many RF cycles, and this is particularly troublesome as it is difficult to track the carrier in the presence of noise. The ability to demodulate signals is equivalent to being able to reconstruct the original constellation diagram of the modulation signal. Also, transitions through the origin indicate that there is significant amplitude variation of the RF signal and so this has high PAR.

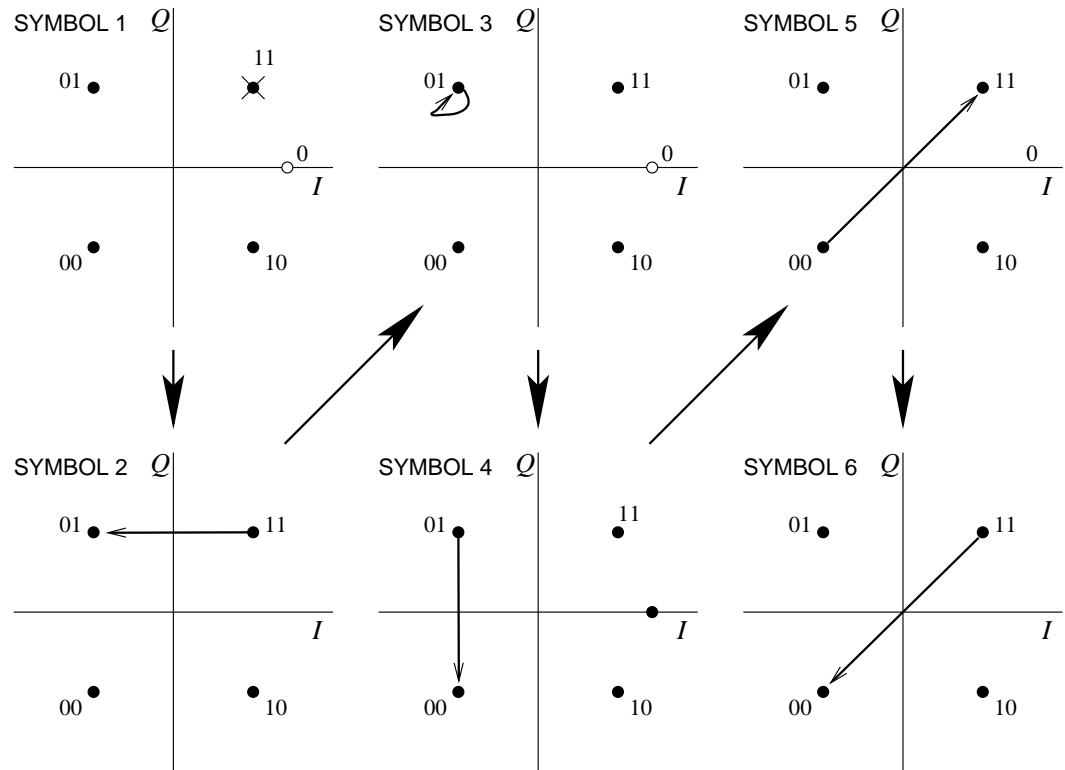
### Example 1.2 QPSK Modulation and Constellation

The bit sequence 110101001100 is to be transmitted using QPSK modulation. Show the transitions on a constellation diagram.

**SOLUTION:** The bit sequence 110101001100 must be converted to a two-bit wide parallel stream of symbols resulting in the sequence of symbols 11 01 01 00 11 00. The symbol 11 transitions to the symbol 01 and then the symbol 01 and so on. The states (or symbols) and the transitions from one symbol to the next required to send the bitstream 110101001100 are shown in Figure 1-13.

QPSK modulation results in the phasor of the carrier transitioning through the origin so that the average power is lower and the PAR is high. A more significant problem is that the phasor will fall below the noise floor, making carrier recovery almost impossible.

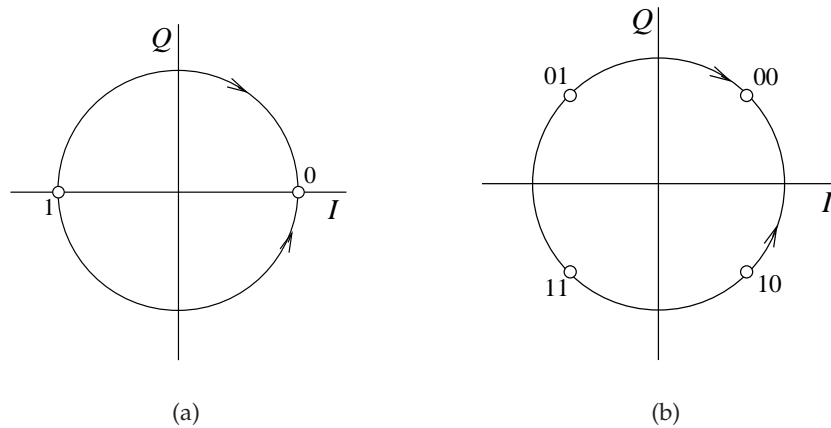
<sup>2</sup> This is true here but not in all cases. The constellation diagram is a way of representing symbols first and is not simply a phasor diagram.



**Figure 1-13** Constellation diagram states and transitions for the bit sequence 110101001100 sent as the set of symbols 11 01 01 00 11 00 using QPSK. Note that Symbols 2 and 3 are identical, so there is no transition and this is shown as a self-loop, whereas there will be no transition in going from Symbol 2 to Symbol 3. The SYMBOL numbers indicated reference the symbol at the end of the transition (the end of the arrowhead).

#### 1.4.4 Frequency Shift Keying, FSK

Frequency Shift Keying is one of the simplest forms of digital modulation, with the frequency of the transmitted signal indicating a symbol, usually either one or two bits. It was the first form of digital modulation employed. FSK is illustrated in Figure 1-8(d). The schematic of an FSK modulation system is shown in Figure 1-10. Here, a binary bit stream is lowpass filtered and used to drive an FSK modulator, one implementation of which shifts the frequency of an oscillator according to the voltage of the baseband signal. This function can be achieved using a phase locked loop (PLL) with considerably less sophistication than PSK-based schemes which require digital signal processing to demodulate a modulated signal. With FSK, an FM demodulator can be used to receive the signal. A characteristic feature of FSK is that the amplitude of the modulated signal is constant so efficient saturating (and hence nonlinear) amplifiers can be used without introducing much distortion of concern. Not surprisingly FSK was the first form of digital modulation used in mobile digital radio. Prior to digital radio FSK was used in analog radio to transmit bits. A particular form of FSK is minimum shift keying (MSK), which

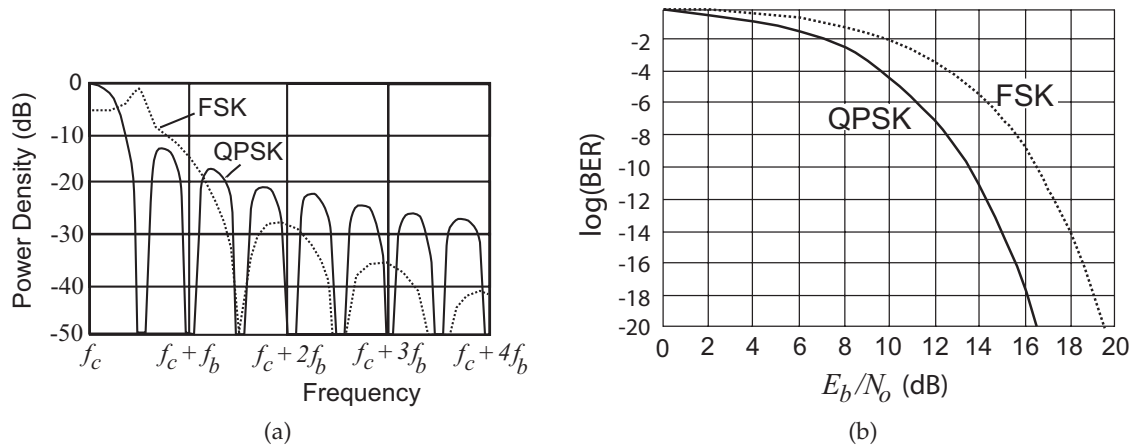


**Figure 1-14** Constellation diagrams of FSK modulation: (a) two-state FSK; and (b) four-state FSK.

uses a baseband lowpass filter so that the transitions from one state to another are smooth in time and limiting the bandwidth of the modulated signal.

In the preceding sections the **constellation diagram** was introduced as a phasor diagram of the (modulated) carrier. However, the equivalence is only approximate and the similarity is most distinct with FSK modulation. Strictly speaking, a phasor diagram describes a phasor that is fixed in frequency. Still, if the phasor is very slowly phase modulated then this approximation is good. That is, the frequency of the modulated carrier is considered to be fixed and the phase changes over time. FSK modulation cannot be represented on a phasor diagram, as the information is in the frequency transition rather than a phase transition. However, the discrete states must be represented and the constellation diagram is used to graphically represent them and the transitions. The departure from the phasor diagram can no longer be ignored. In reality, with FSK modulation, the frequency of the modulated carrier changes slowly if the baseband signal is lowpass filtered. For example, consider an FSK modulated signal with a bandwidth of 200 kHz and a carrier at 1 GHz. This is a 0.02% bandwidth, so the phasor changes very slowly. So going from one FSK state to another takes a very long time, about 5000 cycles. In trying to represent FSK modulation on a pseudo-phasor diagram, the frequency is approximated as being fixed and the maximum real frequency shift is arbitrarily taken as being a  $180^\circ$  shift of the phasor.

In FSK, the states are on a circle on the constellation diagram (see Figure 1-14). Note that the constellation diagram indicates that the amplitude of the phasor is constant, as FSK is a form of FM. In four-state FSK modulation (see Figure 1-14(b)), transitions between states take twice as long for states that are on opposite sides of the constellation diagram compared to states that are only separated by  $90^\circ$ . Filtering of the baseband modulating signal is required to minimize the bandwidth of the modulated four-state FSK signal. This reduces spectral efficiency to less than the theoretical maximum of 2 bits per hertz. In summary there are slight inconsistencies and arbitrariness in using a phasor diagram for FSK but FSK does have a defined constellation diagram which is closely related to a phasor diagram.



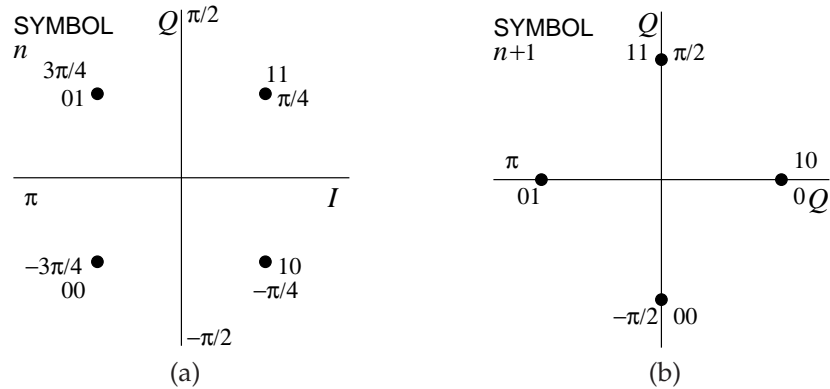
**Figure 1-15** Comparison of FSK and QPSK: (a) power spectral density as a function of frequency deviation from the carrier; and (b) BER versus signal-to-noise ratio (SNR) as  $E_b/N_o$  (or energy per bit divided by noise per bit).

### 1.4.5 Comparison of FSK and QPSK Modulation

The modulation format used impacts the choice of circuitry, battery life, and the tolerance of the system to noise. Figure 1-15 contrasts two types of digital modulation: FSK as used in the Global System for Mobile Communications (GSM) cellular system, and QPSK used in the Digital Advanced Mobile Phone System (DAMPS) cellular system. In Figure 1-15(a),  $f_b$  is the bit frequency and it is seen that FSK and QPSK have different spectral shapes. Most of the energy is contained within the bandwidth defined by the bit frequency. At multiples of the bit frequency, the power density with FSK is much lower than with QPSK, resulting in less interference (adjacent channel interference [ACI]) with neighboring radios in adjacent channels. This is an important metric with radios that is captured by the adjacent channel power ratio (ACPR), the ratio of the power in the adjacent channel to the power in the main channel. Another important metric is the bit error rate (BER). Different modulation formats differ in their susceptibility to noise. The level of noise is captured by the ratio of the power in a bit,  $E_b$ , to the noise power,  $N_o$ , in the time interval of a bit. This ratio,  $E_b/N_o$  (often referred to as E B N O), is also the signal-to-noise ratio (SNR). In particular, consider the plot of the BER against the SNR shown in Figure 1-15(b). QPSK is less susceptible to noise than is FSK.

### 1.4.6 $\pi/4$ Quadrature Phase Shift Keying, $\pi/4$ -QPSK

A major objective in digital modulation is to ensure that the RF trajectory from one phase state to another does not go through the origin. The transition is slow so that if the trajectory goes through the origin, the amplitude of the carrier will be below the noise floor for a considerable time and it will not be possible to recover its frequency. One of the solutions developed to address this problem is the  $\pi/4$  quadrature phase shift keying ( $\pi/4$ -QPSK) modulation scheme. In this scheme the



**Figure 1-16** Constellation diagram of  $\pi/4$ -QPSK modulation: (a) constellation diagram at one symbol; and (b) the constellation diagram at the next symbol.

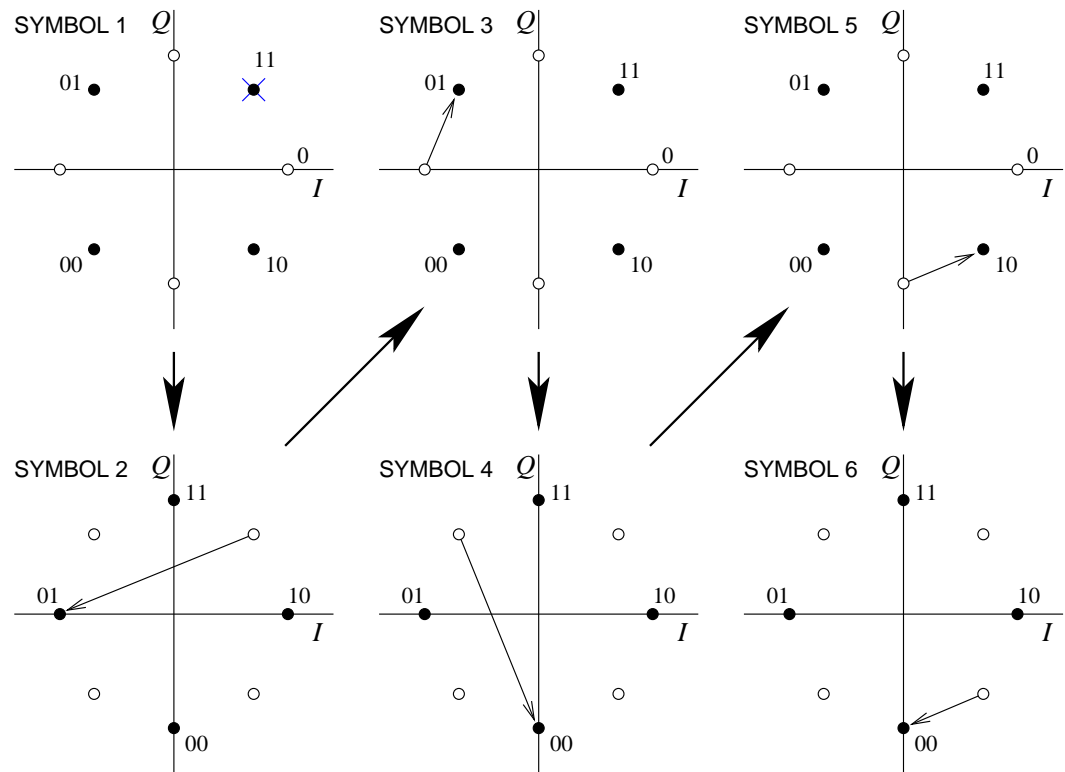
constellation at each symbol is rotated  $\pi/4$  radians from the previous symbol, as shown in Figure 1-16.

### Example 1.3 $\pi/4$ -QPSK Modulation and Constellation

The bit sequence 110101001000 is to be transmitted using  $\pi/4$ -QPSK modulation. Show the transitions on a constellation diagram.

**SOLUTION:** The bit sequence 110101001000 must be converted to a two-bit-wide parallel stream of symbols resulting in the sequence of symbols 11 01 01 00 10 00. The symbol 11 transitions to the symbol 01 and then the symbol 01 and so on. The constellation diagram of  $\pi/4$ -QPSK modulation really consists of two QPSK constellation diagrams that are shifted by  $\pi/4$  radians, as shown in Figure 1-16. At one symbol (or time) the constellation diagram is that shown in Figure 1-16(a) and at the next symbol it is that shown in Figure 1-16(b). The next symbol uses the constellation diagram of Figure 1-16(a) and the process repeats. The states (or symbols) and the transitions from one symbol to the next required to send the bitstream 110101001000 are shown in Figure 1-17.

One of the unique characteristics of  $\pi/4$ -QPSK modulation is that there is always a change, even if a symbol is repeated. This helps with recovering the carrier frequency, which is an important function in a demodulator. Also, the carrier phasor does not go through the origin and so the PAR is lower than if QPSK modulation were used, as this would result in transitions through the origin. If the binary bitstream itself (with sharp transitions in time) is the modulation signal, then the transition from one symbol to the next occurs instantaneously and hence the modulated signal has a broad spectrum around the carrier frequency. The transition, however, is slower if the bitstream is filtered, and so the bandwidth of the modulated signal will be less. Ideally the transmission of one symbol per hertz would be obtained. However, in  $\pi/4$ -QPSK modulation the change from one symbol to the next has a variable distance (and so takes different times) so that the ideal spectral efficiency of one symbol per hertz (or two bits/Hz) is not obtained. In practice, with realistic filters and allowing for the longer transitions,  $\pi/4$ -QPSK modulation achieves 1.62 bits/Hz.

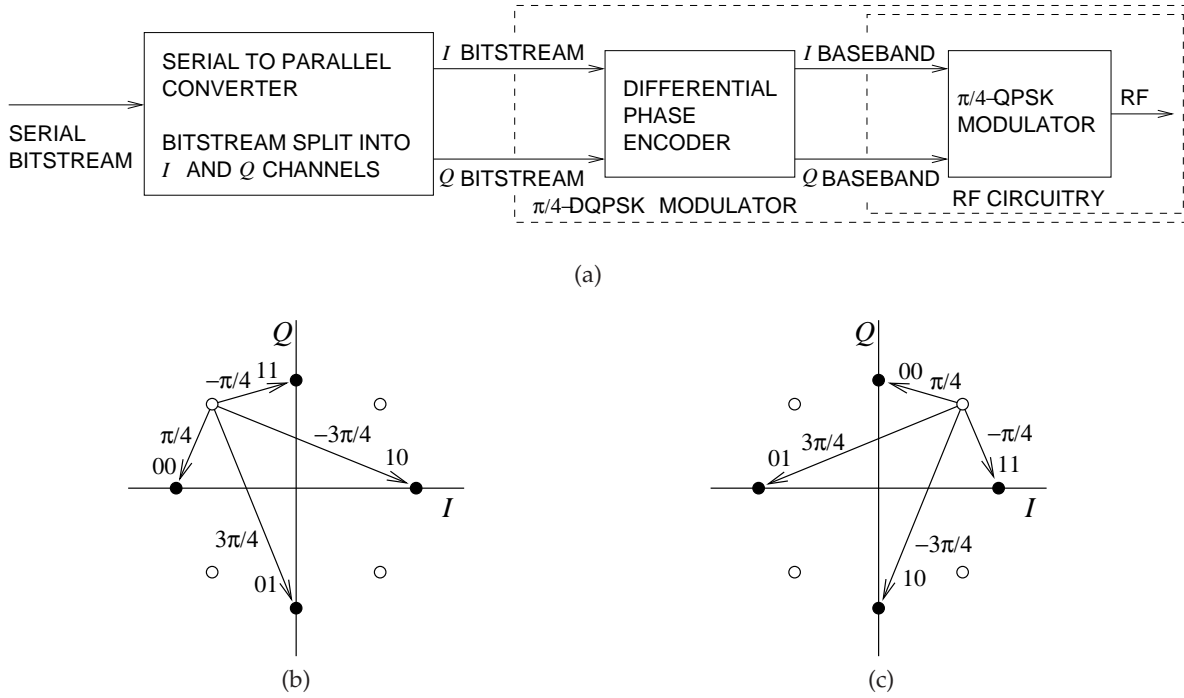


**Figure 1-17** Constellation diagram states and transitions for the bit sequence 110101001000 sent as the set of symbols 11 01 01 00 10 00 using  $\pi/4$ QPSK modulation.

### 1.4.7 Differential Quadrature Phase Shift Keying, DQPSK

Multiple transmission paths, or multipaths, result in constructive and destructive interference and can result in rapid additional phase rotations. As a result, relying on the phase of a phasor at the symbol sample time to determine the symbol transmitted is prone to error. When an error results at one symbol, this error accumulates when subsequent symbols are extracted. The solution is to use encoding, and one of the simplest encoding schemes is differential phase encoding. In this scheme the information of the modulated signal is contained in changes (differences) in phase rather than in the absolute phase.

The  $\pi/4$ -DQPSK modulation scheme is a differentially encoded form of  $\pi/4$ -QPSK. The  $\pi/4$ -DQPSK scheme incorporates the  $\pi/4$ -QPSK modulator and an encoding scheme, as shown in Figure 1-18(a). The scheme is defined with respect to its constellation diagram, shown in Figure 1-18(b) and repeated in Figure 1-18(c) for clarity. The D indicates **differential coding**, while the  $\pi/4$  denotes the rotation of the constellation by  $\pi/4$  radians or  $45^\circ$  from one interval to the next. This can be explained by considering Figure 1-18(a). A four-bit stream is divided into two quadrature nibbles of two bits each. These nibbles independently control the I and



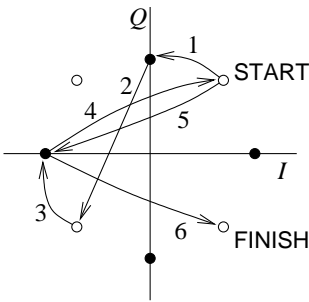
**Figure 1-18** A  $\pi/4$ -DQPSK modulator consisting of (a) a differential phase encoder and a  $\pi/4$ -QPSK modulator; (b) constellation diagram of  $\pi/4$ -DQPSK; and (c) a second example clarifying the information is in the phase change rather than the phase state.

**Table 1-1** Phase changes in a  $\pi/4$ -DQPSK modulation scheme.

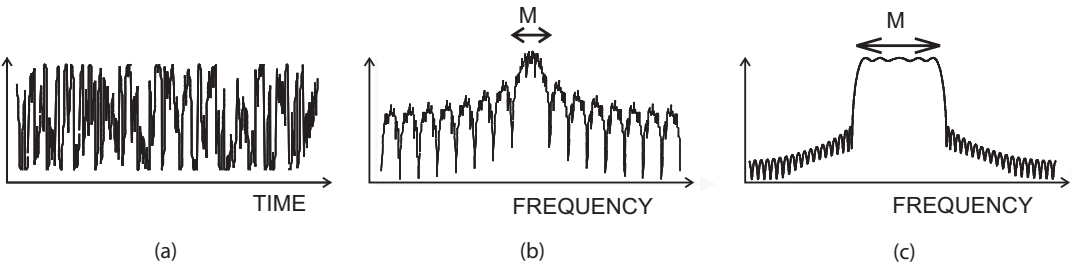
$O_k$	$E_k$	$\Delta\pi$
1	1	$-3\pi/4$
0	1	$3\pi/4$
0	0	$\pi/4$
1	0	$-\pi/4$

$Q$  encoding, respectively, so that the allowable transitions rotate according to the last transition. The information or data is in the phase transitions rather than the constellation points themselves. The relationship between the symbol value and the transition is given in Table 1-1. For example, the transitions shown in Figure 1-19 for six successive time intervals describes the input bit sequence 000110110101. Its waveform and spectrum are shown in Figure 1-20. More detail of the spectrum is shown in Figure 1-21. In practice with realistic filters and allowing for the longer transitions,  $\pi/4$ -DQPSK modulation achieves 1.62 bits/Hz, the same as  $\pi/4$ -QPSK, but of course with greater resilience to changes in the transmission path. Sometimes a distinction is made between the transmitted symbols and the encoded symbols. The encoded symbols already have the data represented as transitions

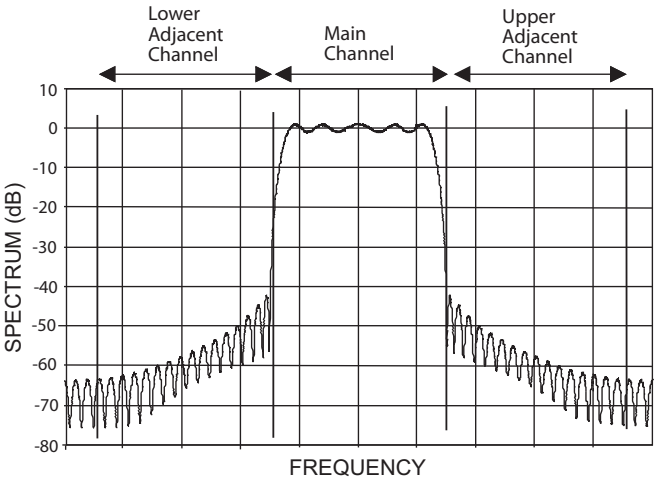




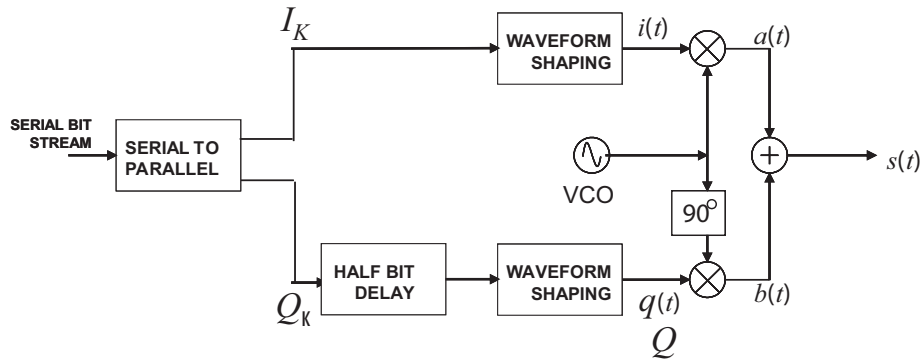
**Figure 1-19** Constellation diagram of  $\pi/4$ -DQPSK modulation showing six symbol intervals coding the bit sequence 000110110101.



**Figure 1-20** Details of digital modulation obtained using differential phase shift keying ( $\pi/4$ -DQPSK): (a) modulating waveform; (b) spectrum of the modulated carrier, with M denoting the main channel; and (c) details of the spectrum of the modulated carrier focusing on the main channel.



**Figure 1-21** Detailed spectrum of a  $\pi/4$ -DQPSK signal showing the main channel and lower and upper adjacent channels.



**Figure 1-22** Block diagram of an OQPSK modulator.

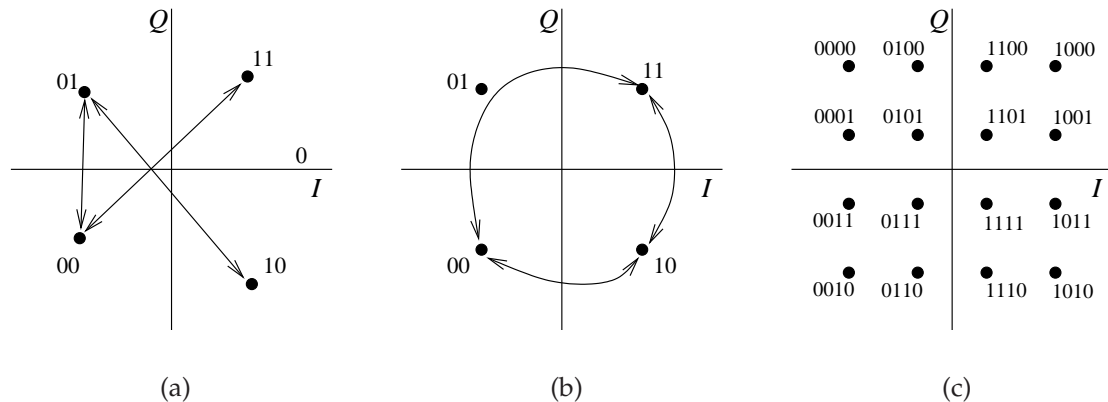
from one transmitted symbol to the next. Similar reference is made to received symbols and decoded symbols. The received symbols are the output of the  $\pi/4$ -QPSK demodulator, while the decoded symbols are the actual data extracted by comparing one received symbol with the previous received symbol. The decoded symbol is extracted in the DSP unit. In a differential scheme, the data transmitted are determined by comparing a symbol with the previously received symbol, so the data are determined from the change in phase of the carrier rather than the actual phase of the carrier. This process of inferring the data actually sent from the received symbols is called decoding. When  $\pi/4$ -DQPSK encoding was introduced in the early 1990s the DSP available for a mobile handset had only just reached sufficient complexity. Today, encoding is used with all digital radio systems and is more sophisticated than just the differential scheme of DQPSK. There are new ways to handle carrier phase ambiguity. The sophistication of modern coding schemes is beyond the hardware-centric theme of this book.

#### 1.4.8 Offset Quadrature Shift Keying, OQPSK

The offset quadrature phase shift keying (OQPSK) modulation scheme avoids  $IQ$  transitions passing through the origin on the constellation diagram (see Figure 1-23(a)). As in all QPSK schemes, there are two bits per symbol, but now one bit is used to directly modulate the RF signal, whereas the other bit is delayed by half a symbol period as shown in Figure 1-22. The maximum phase change for a bit transition is  $90^\circ$ , and as the  $I_K$  and  $Q_K$  are delayed, a total phase change of approximately  $180^\circ$  is possible during one symbol. The constellation diagram is shown in Figure 1-23(a).

#### 1.4.9 Gaussian Minimum Shift Keying, GMSK

Gaussian minimum shift keying (GMSK) is the modulation scheme used in the GSM cellular wireless system and is a variant of MSK with waveform shaping coming from a Gaussian low pass filter. It is very similar to FSK modulation and can be implemented with the same hardware, generally using a PLL.



**Figure 1-23** Constellation diagram for various formats: (a) OQPSK; (b) GMSK; and 16QAM.

GMSK modulation is also used in the Digital European Cordless Telephone (DECT) standard. The spectral efficiency of GMSK as implemented in the GSM system (it depends slightly on the Gaussian filter parameters) is 1.35 bits/s/Hz ( $1.35 \text{ b}\cdot\text{s}^{-1}\cdot\text{Hz}^{-1}$ ). Unfiltered MSK has a constant RF envelope and so the linear amplification requirement is reduced. Filtering is required to limit spectral spreading—in GMSK this results in amplitude variations of about 30%. However, this is still very good, so one of the fundamental advantages of this modulation scheme is that nonlinear, power-efficient amplification can be used. GMSK is essentially a digital implementation of FM with a binary change in the frequency of modulation. The switch from one modulation frequency to the other is timed to occur at zero phase. Put another way, the input bitstream is shaped to form half sinusoids for each bit of the input stream. The phase of the modulating signal is always continuous, but at the zero crossings the half sinusoid continues as a positive or negative half sinusoid depending on the next bit in the input stream. The constellation diagram for GMSK (Figure 1-23(b)) is similar to that for OQPSK, but on decoding, the information is not in the phase but the frequency. So GMSK is an FSK scheme and can be implemented using traditional frequency modulation and demodulation methods. While QPSK schemes can transmit more data in a given channel bandwidth, GMSK (and other FSK techniques) have the advantage that implementation of the baseband and RF hardware is simpler. A GMSK transmitter can use conventional frequency modulation. On receive, an FM discriminator can be used, avoiding the more complex  $I$  and  $Q$  demodulation. In GMSK modulation, a data stream is passed through a Gaussian filter and the filtered response drives an FM modulator with the FM deviation set to one-half of the data rate. For example, an 8000bits/s GMSK data stream is modulated onto an RF carrier with a peak deviation of 4 kHz or  $\pm 2$  kHz. One type of MSK and GMSK modulator is shown in Figure 1-24.

Most GSM phones input the baseband signal to a PLL to implement frequency modulation. The output of the PLL is input to a power amplifier. This amplifier can be quite efficient, as amplitude distortion is not a concern.

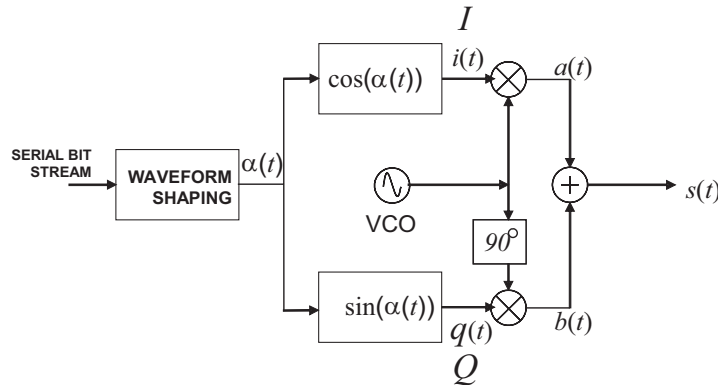


Figure 1-24 Block diagram of an MSK modulator.

#### 1.4.10 $3\pi/8$ -8PSK, Rotating Eight-State Phase Shift Keying

The  $3\pi/8$ -8PSK modulation scheme is similar to  $\pi/4$ -DQPSK in the sense that rotation of the constellation occurs from one time interval to the next. This time however the rotation of the constellation from one symbol to the next is  $3\pi/8$ . This modulation scheme is used in the EDGE system, and provides 3 bits per symbol (ideally) compared to GMSK used in GSM which has 2 bits per symbol. GSM/Edge provides data transmission of up to 128 kbps, and faster than the 48 kbps possible with GSM.

Quadrature modulation schemes with four states, such as QPSK, have two  $I$  states and two  $Q$  states that can be established by lowpass filtering the  $I$  and  $Q$  bitstreams. For higher order modulation schemes such as 8PSK this approach will not work. Instead,  $I(t)$  and  $Q(t)$  are established in the DSP unit and then converted using a DAC to generate the analog signals applied to the hardware modulator. Alternatively the modulated signal is created directly in the DSP and a DAC converts this to an IF and a hardware mixer up-converts this to RF. This approach is required in multimode phones supporting multiple standards.

#### 1.4.11 Quadrature Amplitude Modulation, QAM

The digital modulation schemes described so far modulate the phase or frequency of a carrier to convey binary data and the constellation points lie on a circle of constant amplitude. The effect of this is to provide some immunity to amplitude changes to the signal. However, much more information can be transmitted if the amplitude is varied as well as the phase. With sophisticated signal processing it is possible to reliably use quadrature amplitude modulation (QAM). In particular, it is necessary to characterize the channel. In wired and line-of-sight (LOS) systems, the channel changes slowly. However, in non-LOS wireless systems it is necessary to incorporate a pilot code or pilot signal with the data so that the characteristics of the channel can be continually updated.

**Table 1-2** Spectral efficiencies of various modulation formats.

Modulation	bits/s/Hz
BPSK	1
FSK	1
QPSK	2
GMSK	1.35
$\pi/4$ -DQPSK	1.63
$3\pi/8$ -8PSK	2.7
64-QAM	4
256-QAM	6

A sixteen-state rectangular QAM constellation is shown in Figure 1-23(c). This constellation can be produced by separately amplitude modulating an  $I$  carrier and a  $Q$  carrier. Both carriers have the same frequency but are  $90^\circ$  out of phase. The two carriers are then combined, with the result that the fixed carrier is suppressed. The most common form of QAM is square QAM, or rectangular QAM with an equal number of  $I$  and  $Q$  states. The most common forms are 16-QAM, 64-QAM, 128-QAM, and 256-QAM. The constellation points are closer together with high-order QAM and so are more susceptible to noise and other interference. Thus high-order QAM can deliver more data, but less reliably, than can lower order QAM.

The constellation in QAM can be constructed many ways and while rectangular QAM is the most common form, nonrectangular schemes exist; for example, having two PSK schemes at two different amplitude levels. While there are minor advantages to such schemes, square QAM is generally preferred as it requires simpler modulation and demodulation. The rapid fading in a mobile environment has a bigger impact on amplitude than on phase. As a result, PSK schemes have fewer errors than QAM schemes in mobile use.

#### 1.4.12 Digital Modulation Summary

The spectral efficiencies of various digital modulation schemes are summarized in Table 1-2. For example, in 1 kHz of bandwidth the  $3\pi/8$ -8PSK scheme (supported in third generation cellular) transmits 2700 bits. Digital transmission requires greater bandwidth than does analog modulation for transmission of the same amount of information that was originally analog (e.g., voice). However, digital modulation is essential for data, and digital modulation is also advantageous for voice. Direct digitization of an audio waveform for high-quality reproduction requires 8 bits of resolution captured at a sample rate of 8000 samples/s for a total 64 kbps (64000 bits/s). The appeal of digital modulation for audio is directly related to the reduction of bit rate accomplished by speech coding algorithms. Acceptable speech is achieved with bit rates of 3.8 kbps and higher. (The measure of speech quality is purely subjective.) The speech coding algorithms achieve bit rate reduction by utilizing the characteristics of human hearing. There is a lot of redundancy in speech, but this is not used specifically. Human

hearing responds to time-varying spectral content, and the unique characteristic is that the statistics of the signal, the autocorrelation functions and higher order moments, can be captured at low resolution. The spectrum of the signal can be adequately reconstructed from these statistics.<sup>3</sup> In a typical speech coding algorithm implemented in a codec, units of 160 samples are characterized by just a few autocorrelation and related parameters. These parameters generally do not require many bits to enable fully intelligible speech to be synthesized so that a reduction factor of 8 or even 16 in the bit rate is achieved.

The spectral efficiencies shown in Table 1-2 are sometimes less than the ideal. QPSK ideally conveys 2 bits per symbol, but in a communication system, lowpass filtering at baseband is required to constrain the spectrum of the RF modulated signal. It is clear that an ideal filter cannot be realized, and this in part limits the achievable spectral efficiency. A more significant limitation can be understood by considering the constellation diagram of actual schemes such as those shown in Figure 1-23. Now consider that the constellation diagrams are equivalent to phasor diagrams (which they are in first- and second- generation radio). With the same baseband bandwidth it will take different times for the phasor to make the transition from one symbol to the next; longer transitions require more bandwidth than do shorter transitions. As a result, the spectrum efficiency will be less than the ideal. So in a QPSK-like scheme 2 bits per symbol are achievable, but in practice a symbol requires more than the minimum of one Hz of bandwidth. Also, with QAM some of the outlying constellation points at the corners of the constellation cube (high  $I$  and  $Q$ ) are not used.

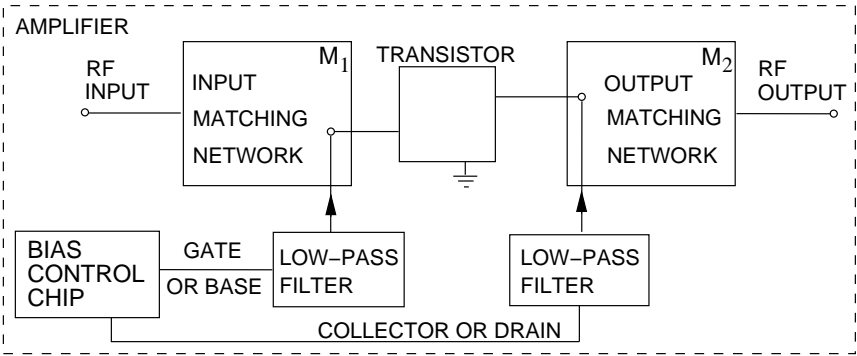
## 1.5 Amplifiers

Amplifiers can be optimized for low noise, moderate to high gain, or substantial power output. Using GaAs metal epitaxy semiconductor field effect transistors (MESFETs), pseudomorphic high electron mobility transistors (pHEMTs) or heterojunction bipolar transistors (HBTs), high power amplification can be obtained at frequencies extending well into the millimeter-wave regions (above 30 GHz). silicon (Si) complementary metal oxide semiconductor (CMOS), and Si BiCMOS (combining Bipolar Junction Transistors (BJTs), and CMOS), and the higher-performance SiGe BiCMOS, provide gain above 30 GHz but with limited power capabilities. At lower frequencies Si laterally diffused metal oxide semiconductor transistors (LDMOS) dominate in basestation high-power transistor amplifiers at cellular frequencies (2 GHz and lower). A special high frequency version of Si CMOS called RF CMOS, supporting microwave elements and having a lower loss substrate, is beginning to dominate for low power RF applications. SiGe BiCMOS has had an 8 GHz performance advantage over RF CMOS for many years. However, RF CMOS is now capable of supporting applications at 10 GHz covering most commercial applications. A late-2000s comparison of these technologies is given in Table 1-3.

<sup>3</sup> Moments themselves are not transmitted, however. Residuals are transmitted from which the time-varying moments and spectra can be reconstructed. Note that the first moment of a signal is its mean, the second-order moment is the signal's standard deviation, etc.

**Table 1-3** Comparison of basic process technologies. Patterning uses optical lithography except that e-beam lithography is required for lithography below 250  $\mu\text{m}$  with GaAs. Costs do not include one-off, costs (i.e. non recurrent engineering [NRE] costs) such as mask and design costs.

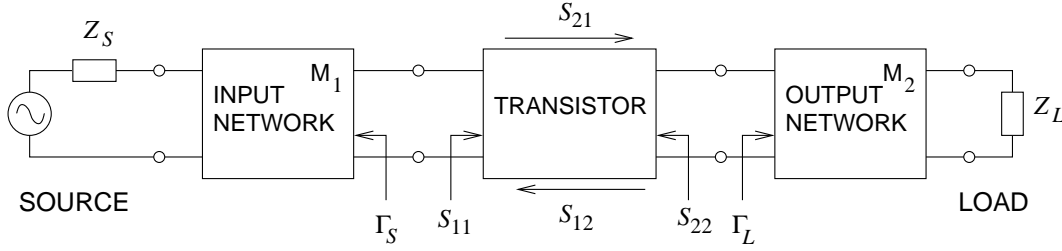
Substrate	GaAs MESFET HBT, pHEMT	Si BIPOLAR	Si/SiGe BiCMOS	Si RF CMOS
Processing	150 mm	300 mm	300 mm	300 mm
Metal layers	semi-insulating	semiconducting	semiconducting	semiconducting
Mask layers	Implant/anneal	epitaxial growth	epitaxial growth	epitaxial growth
Cost (optical)	2–3	2–4	2–4	4–11
Cost (e-beam)	13	27–33	40–50	35–45
	$\approx 10$ cents/ $\text{mm}^2$	$\approx 3$ cents/ $\text{mm}^2$	$\approx 3$ cents/ $\text{mm}^2$	$\approx 3$ cents/ $\text{mm}^2$
	$\approx 100$ cents/ $\text{mm}^2$			



**Figure 1-25** Block diagram of an RF amplifier including biasing networks.

An RF amplifier requires the circuit arrangement indicated in the general block form of Figure 1-25. The DC biasing circuit is fairly standard; it does not involve any microwave constraints and it will not be discussed here. The lowpass filters (the bias circuits) can have one of several forms and are often integrated into the input and output network design sometimes through the choice of appropriate alternative topologies. Synthesis of the input, output, and (occasionally) feedback networks are the primary design objectives of any amplifier. The essence of RF amplifier design is synthesis of the network, shown in Figure 1-26.

RF transistors used to amplify small signals should have high maximum available gain and low noise characteristics. For transistors used in transmitters, where the efficient generation of power is important, it is important that the transistor characteristics be close to linear in the central region of the current-voltage characteristics so that distortion is minimized when the RF voltage variations range over a large portion of the current-voltage characteristics. The ultimate limit on output power is determined by the breakdown voltage at high drain-source voltages and also by the maximum current density. Finally, for efficient amplification of large signals the knee voltage (where the current-voltage curves bend over and start to flatten at low drain-source voltages) should be low.



**Figure 1-26** The scattering parameter ( $S_{nm}$ ) and reflection coefficients ( $\Gamma$ ) associated with a microwave transistor amplifier.

### 1.5.1 Linear Amplifier

The linear amplifier is generally known as a Class A amplifier and is defined by its ability to amplify small to medium and even large signals with minimal distortion. This is achieved by biasing a transistor in the middle of its  $I$ - $V$  (or current-voltage) characteristics. Figure 1-27 shows the  $I$ - $V$  characteristics of the FET and bipolar transistors shown in Figures 1-28(a) and 1-28(b), together with the DC loadline. The loadline is the locus of the output current and voltage, established by the amplifier configurations shown in Figures 1-28(c) and 1-28(d). For the Class A amplifiers in Figures 1-28(c) and 1-28(d) the loadlines are defined by

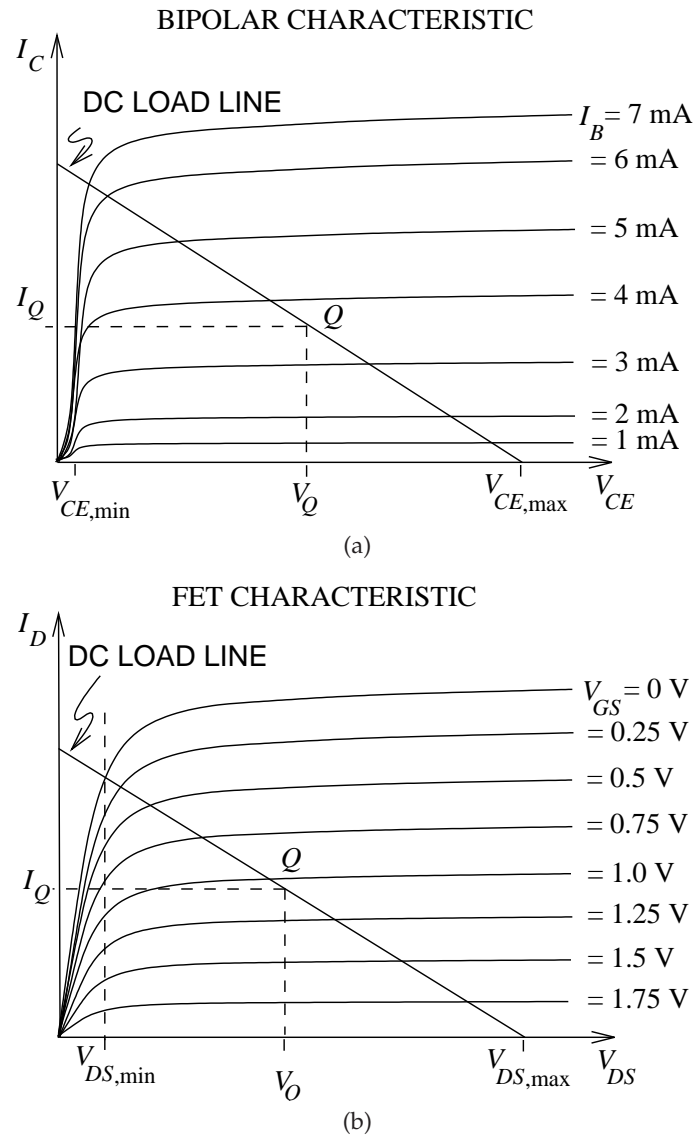
$$I_C = (V_{CC} - V_{CE}) / R_L \quad \text{and} \quad I_D = (V_{DD} - V_{DS}) / R_L. \quad (1.20)$$

These are called single-ended amplifiers as the input and output voltages are referred to ground. The opposite type of amplifier is the differential amplifier configurations to be considered shortly. An amplifier using a bipolar transistor (either a BJT, or an HBT) is shown in Figure 1-28(c), with the transistor terminals labeled in Figure 1-28(a). Referring to Figure 1-27(a), the output voltage of the bipolar amplifier is  $V_{CE}$  and this swings from a maximum value of  $V_{CE,\max}$  to a minimum of  $V_{CE,\min}$ . For a bipolar transistor  $V_{CE,\min}$  is approximately 0.2 V, while  $V_{CE,\max}$  for a resistively-biased circuit is just the supply voltage  $V_{CC}$ . The quiescent or bias point is shown with collector-emitter voltage  $V_Q$  and quiescent current  $I_Q$ . For a Class A amplifier, the quiescent point is just the bias point and this is in the middle of the output voltage swing and the slope of the loadline is established by the load resistor  $R_L$ .

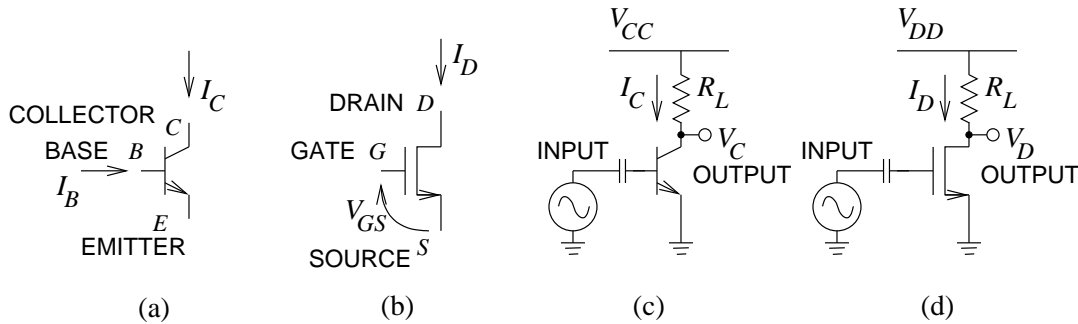
The  $I$ - $V$  characteristics of a FET amplifier are shown in Figure 1-27(b). The notable difference between these characteristics and those of the bipolar transistor is that the curves are less abrupt at low output voltage ( $V_C$  or  $V_D$ ). This results in the minimum output voltage ( $V_{DS,\min}$ ) being larger than  $V_{CE,\min}$ . For a typical RF FET amplifier, as shown in Figure 1-28(d), the supply voltage ( $V_{DD}$ ) is 3 V, while  $V_{DS,\min}$  is 0.5 V. So for the same supply voltage, the output voltage swing with a FET amplifier will be smaller than for a BJT amplifier.

The bipolar and FET amplifiers of Figure 1-28 use resistive biasing so that the maximum output voltage swing is limited. As well, the bias resistor is also the load resistor. Various alternative topologies have been developed yielding a range of output voltage swings. The common variations are shown in Figure 1-29 for an FET amplifier. Figure 1-29(a) is a resistively biased Class A amplifier with





**Figure 1-27** Current-voltage characteristics of transistor amplifiers shown with a Class A loadline: (a) FET amplifier and bipolar amplifier.

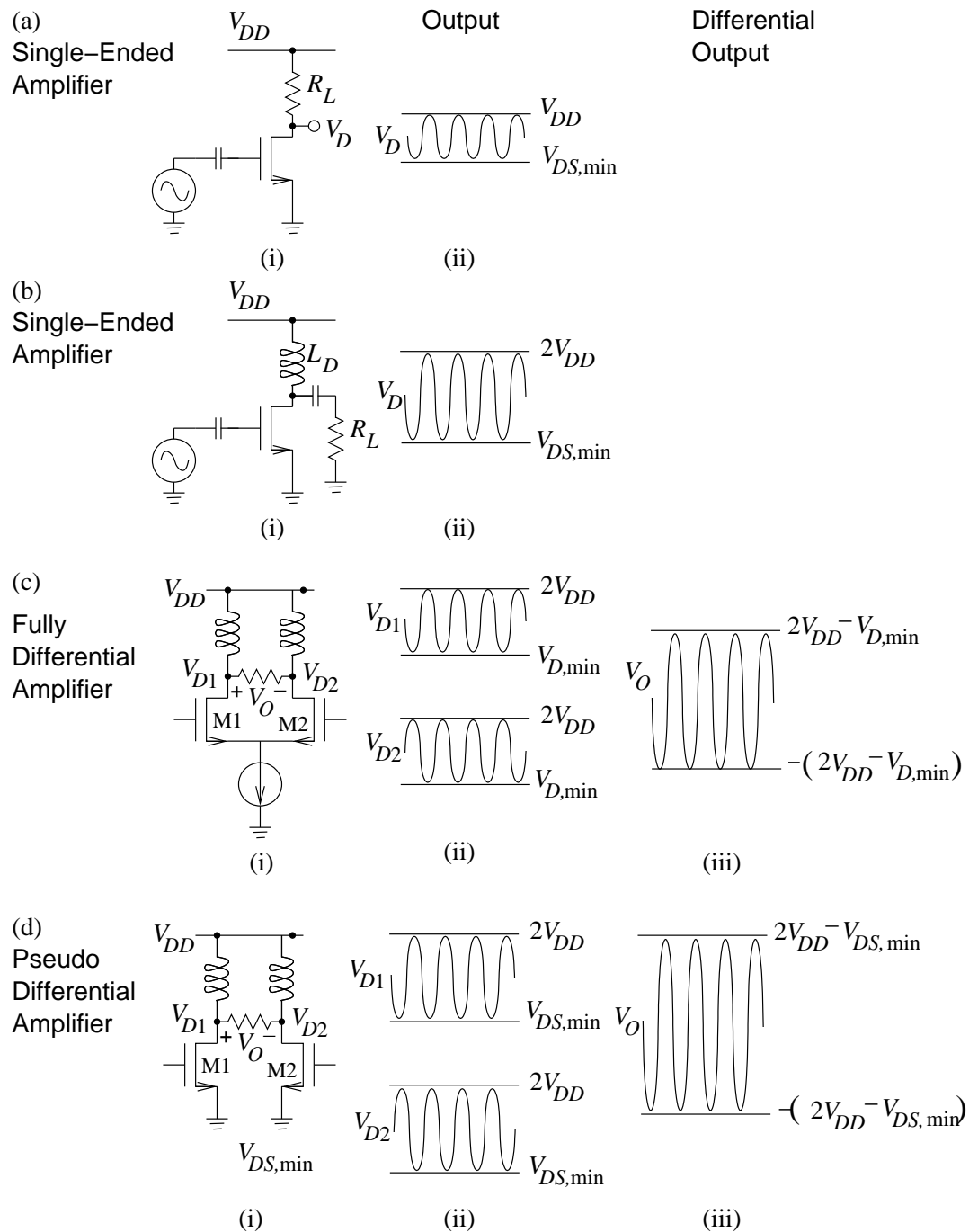


**Figure 1-28** Class A single-ended amplifiers: (a) BJT transistor with B for base terminal, C for collector terminal, and E for emitter terminal; (b) MOSFET transistor with G for gate terminal, D for drain terminal, and S for source terminal; (c) single-ended BJT Class A amplifier with resistive bias; and (d) single-ended MOSFET Class A amplifier with resistive bias.

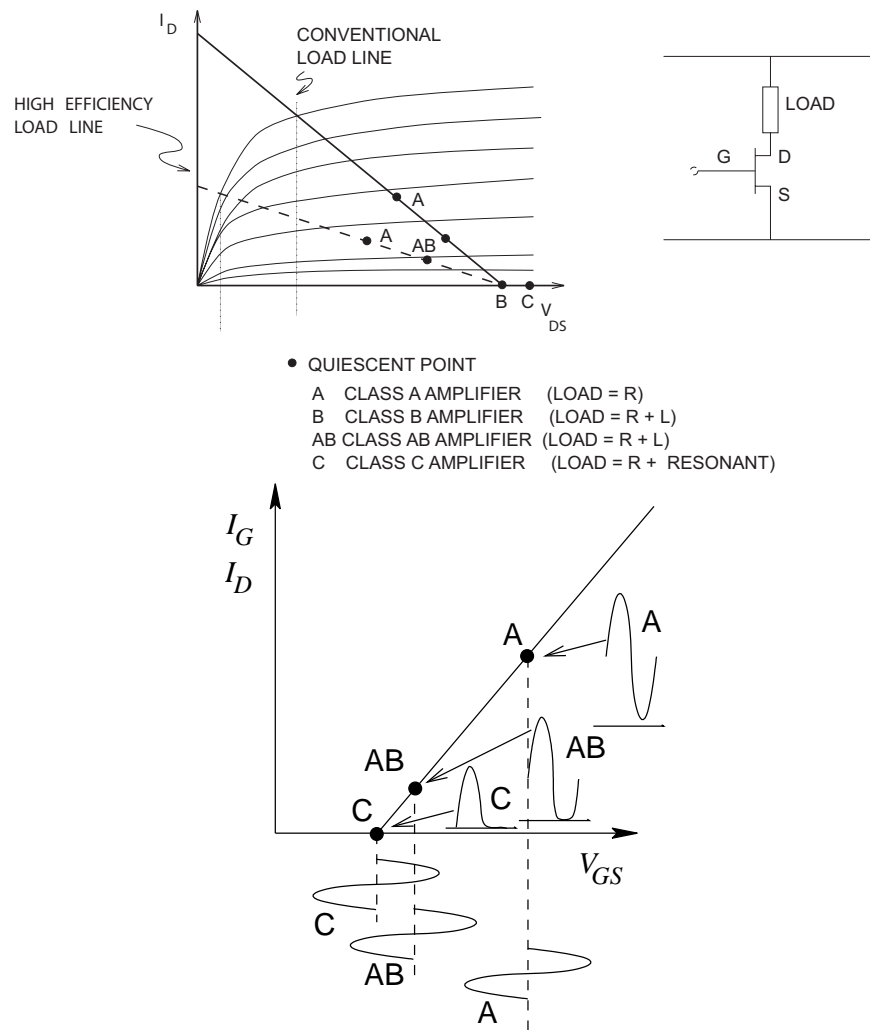
the output voltage swing between  $V_{DS,min}$  and  $V_{DD}$ . The quiescent drain-source voltage is halfway between these extremes. The load  $R_L$  also provides correct biasing. This amplifier is also called a single-ended amplifier to differentiate it from a differential amplifier. A more efficient Class A amplifier uses inductive biasing as shown in Figure 1-29(b). Bias current is now provided via the drain inductor and the load  $R_L$  is not part of the bias circuit. With the inductively loaded Class A amplifier, the quiescent voltage is  $V_{DD}$  and the output voltage swing is between  $V_{DS,min}$  and  $2V_{DD}$ , slightly more than twice the voltage swing of the resistively loaded amplifier. Another topology that provides enhanced voltage swing is the differential resistively biased amplifier shown in Figure 1-29(c). This amplifier topology is also called a fully differential amplifier (FDA). This is the topology commonly found in silicon **radio frequency integrated circuits (RFICs)**, where the current source common to the sources of the FETs results in good differential-mode gain (when the inputs to the two FET gates is  $180^\circ$  out of phase) and the common-mode gain is low. This is important in RFICs, as noise in the substrate will affect the inputs and outputs of both transistors equally. As can be seen in Figure 1-29(c)(ii), the differential voltage swing will be approximately  $4V_{DD}$  less the voltage drop across the current source. The supply voltage of RFICs is limited so at the final output stage the current source is sometimes sacrificed so that larger voltage swings can be obtained. The resulting amplifier is the **pseudo-differential amplifier (PDA)**, shown in Figure 1-29(d). The maximum output voltage swing is now  $4V_{DD} - 2V_{DS,min}$ —almost four times the voltage swing, (or 16 times the power into the same load) of a single-ended resistively biased Class A amplifier.

### 1.5.2 Classes of Amplifiers

The class A amplifier has limited efficiency mainly because there is always substantial quiescent current flowing whether or not RF current is flowing. Higher order classes of amplifiers achieve higher efficiency, but distort the RF signal. The current and voltage locus of Class A, B, AB, and C amplifiers have a similar



**Figure 1-29** Class A MOSFET amplifiers with output voltage waveforms: (a) single-ended amplifier with resistive biasing; (b) single-ended amplifier with inductive biasing; (c) fully differential amplifier with inductive biasing; and (d) pseudo-differential amplifier. Schematic is shown in (i), drain voltage waveforms in (ii); and differential output in (iii).



**Figure 1-30** Current-voltage characteristics of the transistor used in an amplifier showing the quiescent point and output waveforms of various amplifier classes.

trajectory on the output current-voltage characteristics of a transistor. The output characteristics of a transistor are shown in Figure 1-30, showing what is called the linear loadline and the bias points for the various amplifier classes. The loadline is the locus of the DC current and voltage as the DC input voltage is varied. With the Class A amplifier, the transistor is biased in the middle of the transistor characteristics where the response has the highest linearity. That is, if the gate voltage varies from an applied signal the output voltage and current variations are nearly linearly proportional to the applied input. The drawback is that there is always considerable DC current flowing, even when the input signal is very small. That is there is DC power consumption whether or not RF power is being generated at the output of the transistor. This is not of concern if small RF signals

**Table 1-4** Comparison of efficiency metrics for an amplifier producing 1 W RF output power and consuming 2 W of DC power with various power gains. The industry standard power-added efficiency used by RF and microwave engineers is  $\eta_{\text{PAE}}$ .

Power gain (dB)	$\eta_{\text{TOTAL}}$	$\eta_{\text{PAE}}$	$\eta$
3	40%	25%	50%
6	44%	37%	50%
10	48%	45%	50%
15	49%	48%	50%
20	50%	50%	50%
40	50%	50%	50%

are to be amplified, as then a small transistor can be chosen so that the DC current levels are small. It is a problem if an amplifier must handle both large and small signals.

This leads to a discussion of efficiency. The efficiency of a circuit is the useful output power divided by the input power. Such a measure of efficiency is called power-added efficiency (PAE). Two different definitions of PAE are used [4]. The more universal definition that can be used with any two-port network is also called the total power added efficiency or transmit chain efficiency and is denoted here as  $\eta_{\text{TOTAL}}$  defined as

$$\eta_{\text{TOTAL}} = \frac{P_{\text{RF,out}}}{P_{\text{DC}} + P_{\text{RF,in}}} \quad (1.21)$$

At RF and microwave frequencies, the most common definition of PAE used with power amplifiers focuses on the additional RF power divided by the DC input power. This is designated as  $\eta_{\text{PAE}}$  and is defined as

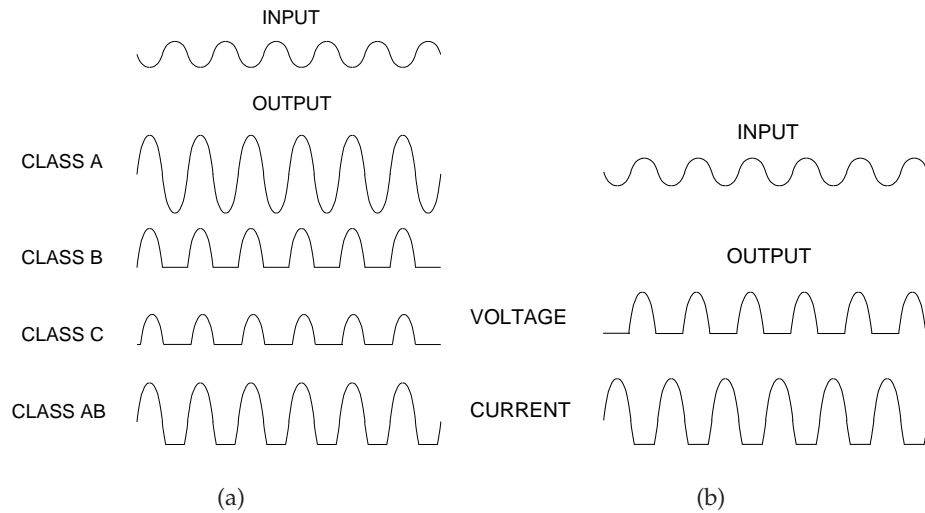
$$\eta_{\text{PAE}} = \frac{P_{\text{RF,out}} - P_{\text{RF,in}}}{P_{\text{DC}}} \quad (1.22)$$

For high-gain amplifiers,  $P_{\text{RF,in}} \ll P_{\text{DC}}$ , and both  $\eta_{\text{TOTAL}}$  and  $\eta_{\text{PAE}}$  reduce to the efficiency  $\eta$  of the amplifier:

$$\eta = \frac{P_{\text{RF,out}}}{P_{\text{DC}}} \approx \eta_{\text{PAE}} \approx \eta_{\text{TOTAL}} \quad (\text{high gain}) \quad (1.23)$$

These efficiency metrics are compared in Table 1-4 for an amplifier with 1 W RF output power. The first amplifier has a power gain of 3 dB, which is commonly the gain of the final amplifier stage producing the maximum output power available from a particular transistor technology.

Returning now to a discussion of the efficiency of the various classes of amplifiers, since the Class A amplifier is always drawing DC current the efficiency of Class A amplifiers is near zero when the input signal is very small. The maximum efficiency of Class A is 25% if resistive biasing is used and 50% when inductive biasing is used. Efficiency is improved by reducing the DC power and this is achieved by moving the bias point further down the DC loadline, as in the Class B, AB, and C amplifiers shown in Figure 1-30. Reducing the bias results in signal

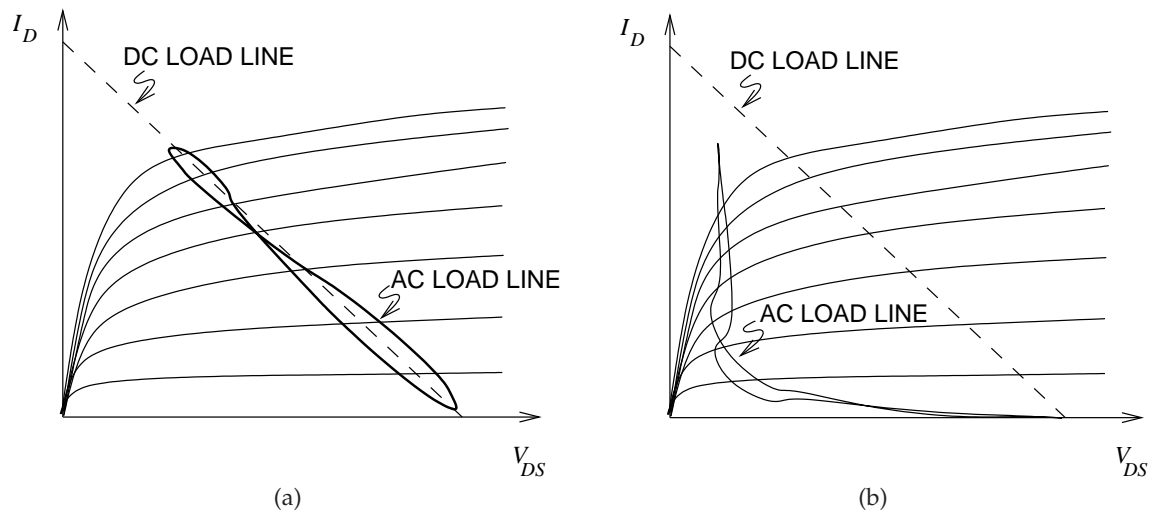


**Figure 1-31** Input and output waveforms for various classes of amplifier: (a) Class A, B, C, and AB amplifiers; and (b) switching amplifiers.

distortion for large RF signals. This can be seen in the various output waveforms shown in Figure 1-31(a).

Class A amplifiers have the highest linearity and Class B and C amplifiers result in considerable distortion. As a compromise, class AB amplifiers are used in many cellular applications, although Class C amplifiers are used with constant envelope modulation schemes, as in GSM. Nearly all small-signal amplifiers are Class A. This is also true for most broadband amplifiers, as amplifier stability is more certain. The class A amplifier presents impedances that are almost independent of the level of the signal. However, a Class B, AB, or C amplifier presents an impedance whose value varies depending on the level of the RF signal. Thus design requires more care, as the chances of instability are higher and it is more likely that an oscillation condition will be met. Also, Class B, AB, and C amplifiers are generally not used in broadband applications or at high frequencies mainly because of the problem of maintaining stability. Class A amplifiers are the preferred solution for amplifiers at 10 GHz and above and for broadband amplifiers, again mainly because it is easy to ensure stability, and thus design is much simpler and more tolerant to parasitic effects and variations.

The effect of parasitic capacitances and delay effects (such as those due to the time it takes carriers to move across a base for a BJT or under the gate for an FET) result in the current-voltage locus for RF signals differing from the DC situation. This effect is captured by the dynamic or AC loadline which is shown in Figure 1-32(a). The **Class A** amplifier is biased in the middle of the  $I$ - $V$  characteristics and the output from this amplifier has the least distortion, as seen in Figure 1-31. This seems very good, but the drawback is that the Class A amplifier draws DC current even when the input signal is negligible. This is a low efficiency, but highly linear class. When designers refer to a “linear amplifier” they are referring to a Class A amplifier. The other amplifier classes shown in Figure 1-30 have higher efficiencies but varying degrees of distortion. The outputs are shown in Figure 1-



**Figure 1-32** DC and RF loadlines: (a) loadlines of Class A, B and C amplifiers; and (b) loadlines of switching amplifiers.

31. The output of the Class B amplifier contains an amplified version of only half of the input signal but draws just a small leakage current when no signal is applied. With the Class C amplifier there must be some positive RF input signal before there is an output: there is more distortion but no current flows, not even leakage current, when there is no RF input signal. The **Class AB** amplifier is a compromise between Class A and Class B amplifiers. Less DC current flows than with Class A when there is negligible input signal and the distortion is less than with Class B. There are higher classes of amplifier, Class D, E etc., and these rely on resonant circuits to change the shape of the loadline to result in better trade-offs between efficiency and distortion than can be achieved with Class AB.

**Class C** amplifiers are biased so that there is almost no drain-source (or collector-emitter) current when no RF signal is applied, so the output waveform has considerable distortion, as shown in Figure 1-31. This distortion is important only if there is information in the amplitude of the signal. FM, GMSK, and PM modulation schemes result in signals with constant RF envelopes, thus there is no information contained in the amplitude of the signal. Therefore errors introduced into the amplitude of a signal are of no significance and efficient saturating mode amplifiers such as a Class C amplifier can be used. In contrast, MSK,  $\pi/4$ -DQPSK and  $3\pi/8$ -8PSK modulation schemes do not result in signals with constant RF envelopes and so information is contained in the amplitude of the RF signal. For these modulation techniques reasonably linear amplifiers are required.

Switching amplifiers are a conceptual departure from Class A, AB, B, and C amplifiers, as can be seen in the typical AC loadline of a switching amplifier shown in Figure 1-32(b). This loadline is achieved by presenting the appropriate harmonic impedances to the transistor amplifier. The particular scheme of harmonic termination (e.g., short or open circuits at the even and odd harmonics) leads to the designation of a switching amplifier as Class D, E, F, etc. The key characteristic of a switching amplifier is that when there is current through the

**Table 1-5** Theoretical maximum efficiencies of amplifier classes.

Amplifier Class	Maximum Efficiency
Class A (resistive bias)	25%
Class A (inductive bias)	50%
Class B	78.53%
Class C	100%
Class E	96%
Class F	88.36%

**Table 1-6** Efficiency reductions due to signal type. The class A amplifier uses inductive drain biasing.

Signal	PAR (dB)	Efficiency Reduction Factor	Class A (L bias) PAE	Class E PAE
FSK (MSK, GMSK)	0	1.0	50%	96%
QPSK	3.6	0.437	21.9%	42%
$\pi/4$ DQPSK	3.0	0.501	25.1%	48.1%
OQPSK	3.3	0.467	23.4%	44.8%
8PSK	3.3	0.467	23.4%	44.8%
64QAM	7.8	0.166	8.3%	15.9%

transistor, there is negligible voltage across the output. Also, when there is voltage across the transistor, there is little current through it (see Figure 1-32(b)). The power dissipated by the transistor is approximately the product of the current through it and the voltage across the output. Thus the switching amplifier consumes very little DC power, transferring nearly all of the DC power to the RF signal. Bandpass filtering of the output of the amplifier results in a final RF output with little distortion. Switching amplifiers are emerging as the preferred linear amplifier in both handsets and base stations of cellular systems.

The theoretical maximum power-added efficiencies achieved by the various amplifier classes are given in Table 1-5. With modulated signals, the maximum efficiencies cannot be achieved, so that typically the average input power of the amplifier must be backed off by the PAR of the signal so that the peak carrier portion of the signal has limited distortion. Generally the acceptable distortion of the peak signal occurs at the 1 dB compression point of the amplifier. This is only an approximate guide, but useful. The PARs of several digitally modulated signals are given in Table 1-6 together with their impact on efficiency. If there are two carriers, then the PAR of the combined signal will be higher, requiring greater amplifier back-off. In practice, the efficiencies will differ from these theoretical values because of loss in the amplifier and the trade-off between efficiency and distortion. This is because the PAR does not fully capture the statistical nature of signals, and because of coding and other technologies that can be used to reduce the PAR.

Two microwave amplifiers are shown in Figures 1-33 and 1-34. They are



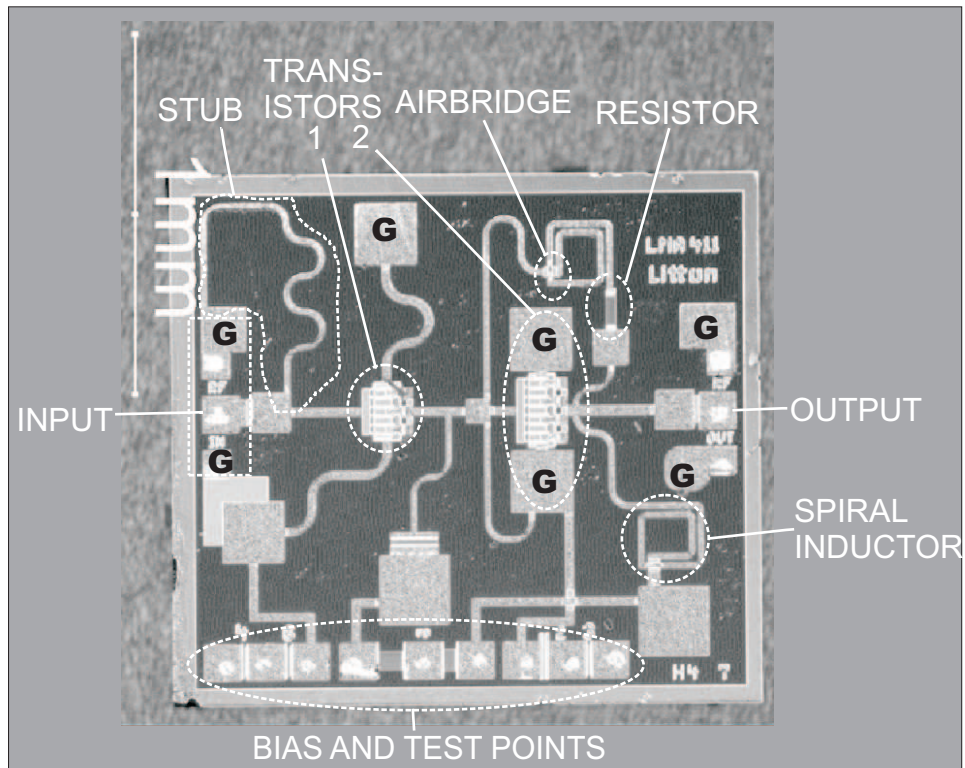
called **microwave monolithic integrated circuits (MMICs)** and are fabricated on compound semiconductor substrates. Both are broadband high-frequency amplifiers covering 8 to 12 GHz, which is known as X-band. Note that high-power RF and microwave transistors are put in parallel yielding the required power.

### 1.5.3 Differential Amplifier

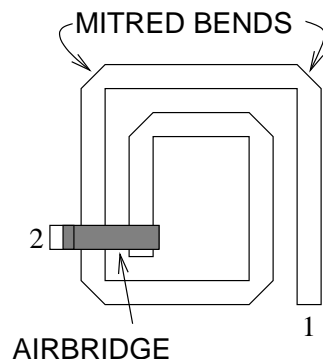
Differential amplifiers are the preferred amplifier topology with silicon monolithically integrated circuits including RFICs. Figure 1-35(a) shows a fully **differential amplifier (FDA)** with resistive biasing in the drain legs. As well as providing biasing current, the resistors are also the loads of the circuit. The supply voltage of an RFIC can be quite low (a few volts or less), so choosing circuit topologies that provide for large voltage swings is important, particularly for an output amplifier. Differential topologies lead to an almost doubling of the output voltage swing compared to the output voltage swing of a single-ended amplifier. An FDA includes a common current source (see Figures 1-35(a) and 1-35(b)). The circuit of Figure 1-35(a) has a higher voltage swing, as previously described. Commonly the schematic in Figure 1-35(c) is used. The current source at the common source point of the FDA in Figure 1-35(a) limits the voltage swing, when larger output voltage swings are required, the current source is eliminated and the resulting amplifier is called a pseudo-differential amplifier (PDA), as shown in Figure 1-35(d). Again, inductive biasing as shown in Figure 1-35(e) almost doubles the possible voltage swing. The schematic representation of the PDA is shown in Figure 1-35(f).

### 1.5.4 Distortion

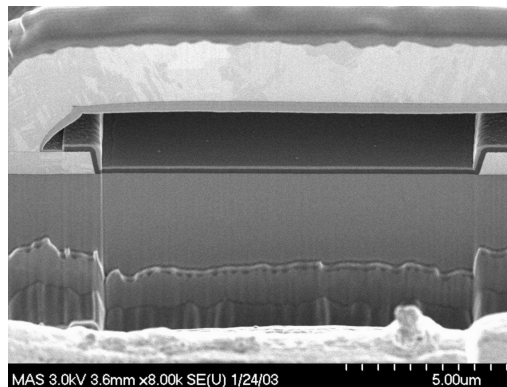
Distortion imposes a fundamental limit to the practical efficiency that can be realized by an amplifier. Distortion originates when the output signal from an amplifier approaches the extremes of the loadline so that the output is not an exact amplified replication of the input signal. For a one-tone signal, the amplitude gain of the signal rollsoff as the input power increases, as shown in Figure 1-36(a). This figure plots the amplitude of the output sinewave against the amplitude of the input sinewave, putting both amplitudes in terms of power. The plot is called the AM-to-AM (**AM-AM**) characteristic of the amplifier. The ideal amplifier would follow the linear relationship between the output and input powers. The AM-AM characteristic is linear at low input powers, but eventually the gain compresses and the output power drops in proportion to the input power. At large powers, the parasitic capacitances of the transistors in the amplifier vary the signal phase, and hence phase distortion results. Figure 1-36(b) shows what is called the AM-to-PM (**AM-PM**) characteristic. The AM-AM distortion is generally more significant, and considerable departure from the linear response occurs before the output phase varies appreciably. In Figure 1-36(a), the 1 dB gain compression point is at the point where the difference between the extrapolated linear response exceeds the actual gain by 1 dB.  $P_{1\text{dB}}$  is the output power at the 1 dB gain compression point and is the single most important metric of distortion, and amplifier designs use  $P_{1\text{dB}}$  as a point of reference. Phase distortion generally occurs at higher powers (see Figure 1-36(b)). While Figure 1-36 presents the distortion characteristics for



(a)

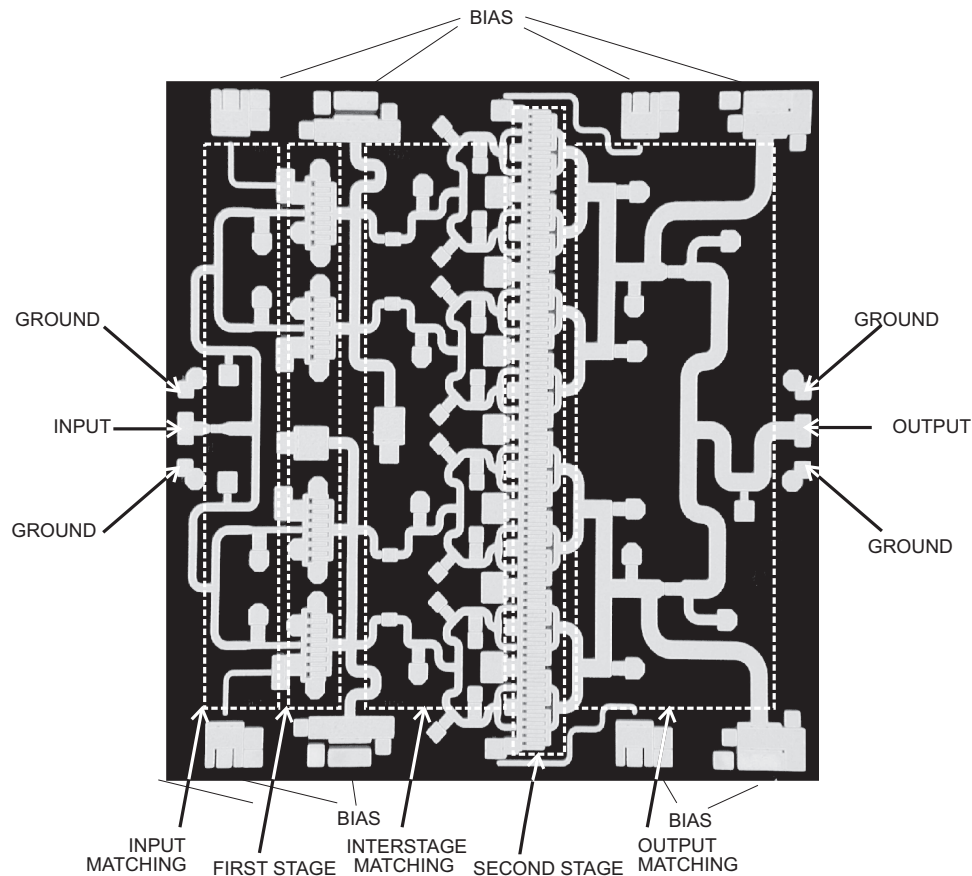


(b)



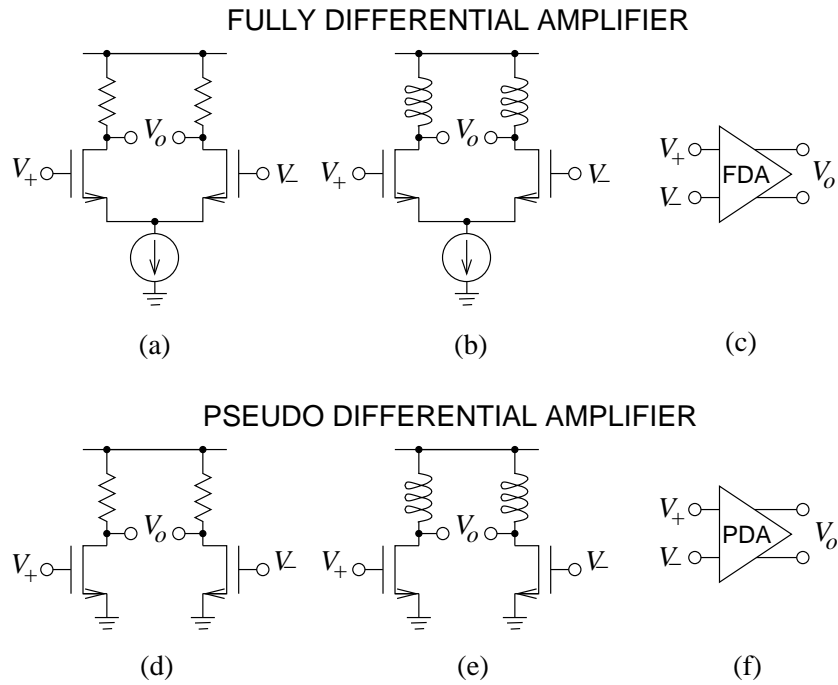
(c)

**Figure 1-33** A two-stage, two-transistor X-band (8–12 GHz) MMIC amplifier producing 100 mW of power: (a) photomicrograph with key networks identified, **G** indicates ground; (b) layout of the top spiral inductor; and (c) scanning electron microscope image of the crosssection of the **airbridge**.

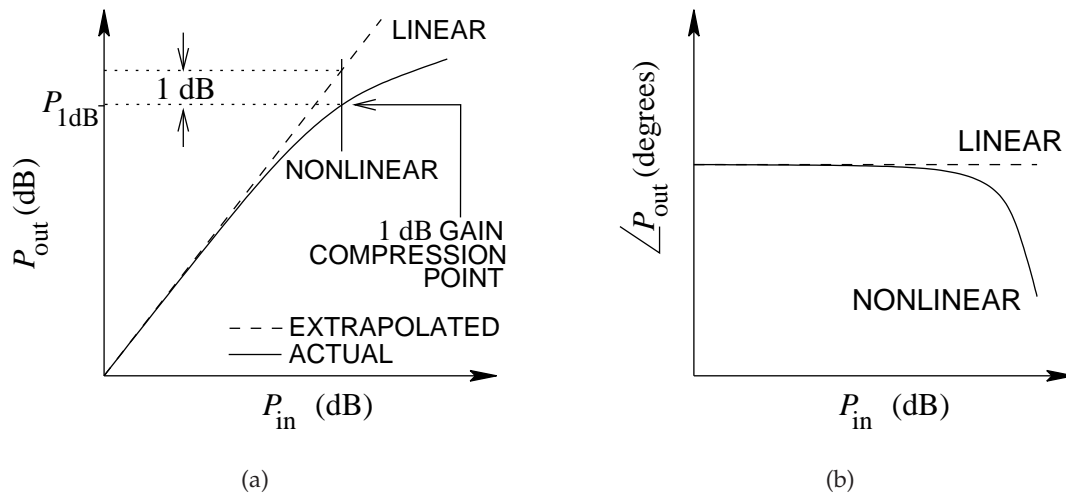


**Figure 1-34** An 8–12 GHz MMIC amplifier producing approximately 1 W of output power with key networks identified. (Courtesy Filtronic, PLC, used with permission.)

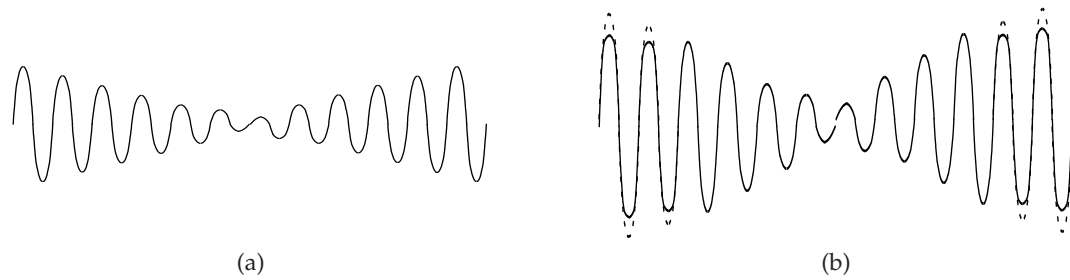
a single sinewave, it has proved to be a good indicator of performance with modulated signals. A two-tone signal consisting of two sinusoidal signals is a better representation of modulated signals. A signal linearly combining (adding) two sinusoidal signals of equal amplitude is shown in Figure 1-37(a). When the input signal to a Class A amplifier is large the extremes of the signal on the loadline (see Figure 1-27), are compressed when the signal reaches its extremities. This results in the saturated output waveform shown in Figure 1-37(b). In the frequency domain this distortion shows up as additional tones so that this distortion is said to produce inter-modulation products (**IMPs**) as shown in Figure 1-38. Here  $f_1$  and  $f_2$  components have the frequencies of the two tones comprising the two-tone signal. The extra ones in the output,  $f_3$  and  $f_4$ , are the intermodulation tones. In Figure 1-38, the tone at  $f_3 = 2f_1 - f_2$  is known as the lower **IM3** (or lower intermod) and  $f_4 = 2f_2 - f_1$  is known as the upper **intermod**. As well as the amplitude distortion resulting in additional tones, there is phase distortion as captured by the AM-PM characteristic. Amplifiers, however, introduce less phase distortion



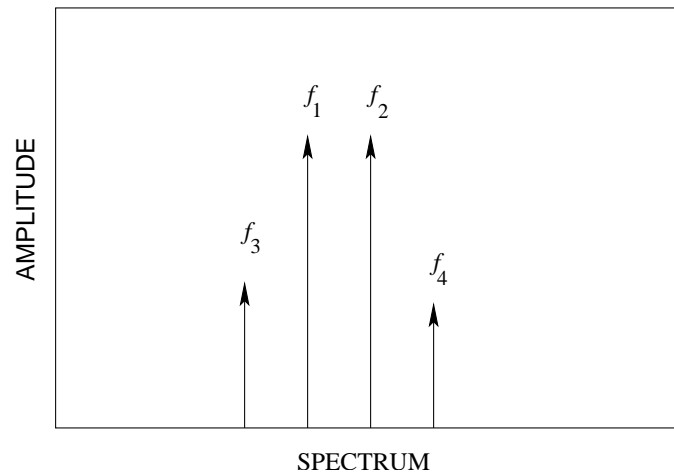
**Figure 1-35** Differential amplifiers: (a) fully differential amplifier (FDA); (b) FDA with inductive biasing; (c) schematic representation; (d) pseudo-differential amplifier (PDA); (e) PDA with inductive biasing; (f) schematic representation.



**Figure 1-36** Nonlinear effects introduced by RF hardware: (a) amplitude (AM-AM) distortion; and (b) phase (AM-PM) distortion.



**Figure 1-37** A two-tone signal: (a) a two-tone input waveform; and (b) distorted output showing compression (dashed waveform is undistorted).



**Figure 1-38** Spectrum at the output of a nonlinear amplifier with a two-tone input signal.

than amplitude distortion, which is fortuitous since most communication systems encode information in the phase or frequency rather than in the amplitude.

Amplifier design consists of both design for good low-power linear operation requiring maximum power transfer at the input and output of the amplifier, and a trade-off of acceptable distortion and efficiency. In practice a certain level of distortion must be tolerated, and what is acceptable is embedded in the specifications for the various wireless systems. Further discussion on the design of amplifiers is best undertaken after scattering parameters, or  $S$ -parameters, are introduced.  $S$ -parameters describe power flow, and the design of RF and microwave amplifiers is intricately tied to power considerations, both noise power and signal power. Linear amplifier design is described in section 6.5.1 on page 316. For low distortion, the peaks of the RF signal must be amplified linearly, however, the DC power consumed depends on the amplifier class. With Class A amplifiers, the DC power must be sufficient to provide undistorted amplification of the largest RF signal so that the DC power is proportional to the peak AC power. The situation is similar for Class AB amplifiers, with the difference being that the intent is to live with some distortion of the peak signal so that the relationship between peak

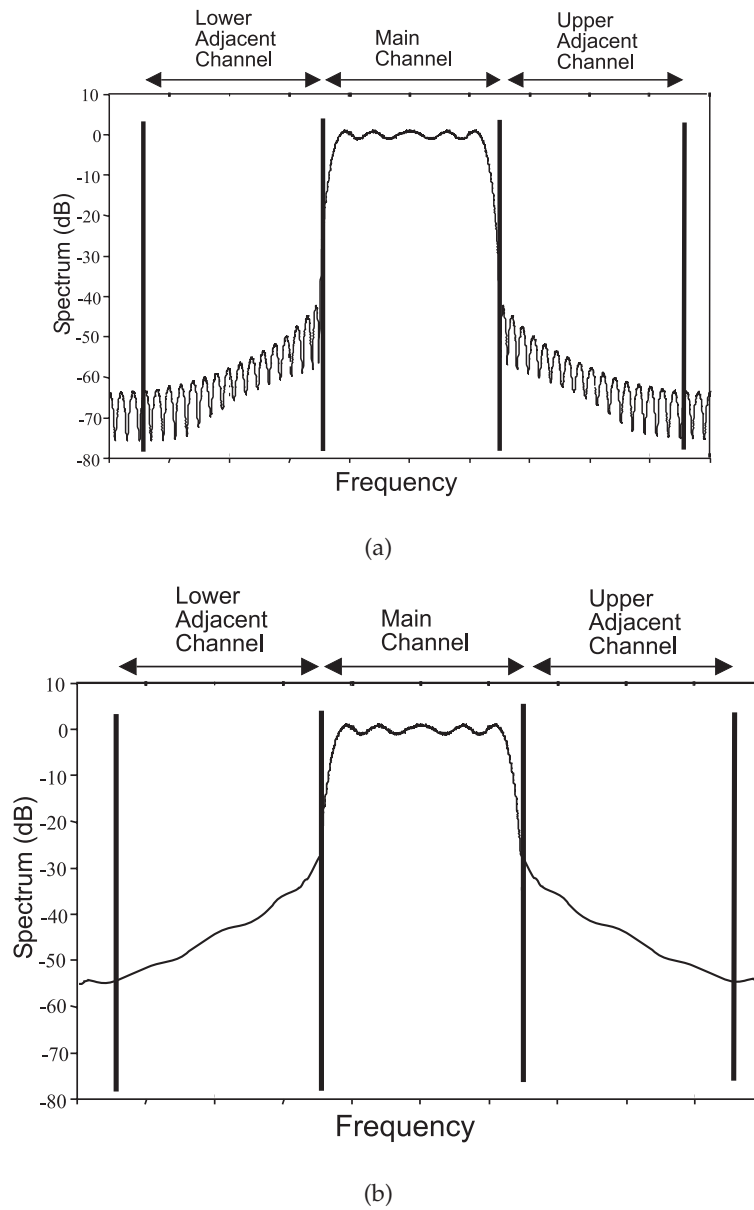
power and DC power still exists, but the direct proportionality no longer holds. For Class C and higher class amplifiers, the DC power is mostly proportional to the average RF power. So for Class A and Class AB amplifiers, the average operating point must be “backedoff” to allow for manageable distortion of the peaks of a signal, with the level of back-off required being proportional to the PAR. For Class C and higher classes, the back-off required comes from experience and experimentation. The characteristics of the signal also determine how much distortion can be tolerated.

The PAR of the signal is an indication of the type and amount of distortion that can be tolerated. The PAR of the two-tone signal is 6 dB, and digitally modulated signals can have PARs ranging from 0 dB to 10 dB or more. A signal with a higher PAR results in lower efficiency, as more back-off is required. Putting this another way, the DC bias must be set so that there is minimum distortion when the signal is at its peak, but the average RF power produced can be much less than the peak RF power. (It is approximately an amount PAR below.) Thus for a high-PAR signal, generally a higher DC power is required to produce the same RF power. This is especially true for Class A amplifiers. Efficiency can be increased by using switching amplifiers. The PAR of a modulated signal is an indication of how much information is being transferred in the amplitude of the signal. For example, a GMSK signal has a PAR of 0 dB and there is no information in the amplitude of a signal so that a highly efficient Class C amplifier can be used as any amplitude introduced does not matter. Signals that have higher PARs contain increasing amounts of information in the amplitude of the signals.

AM-AM and AM-PM distortion, and two-tone distortion also provide indications as to the distortion that occurs with digitally modulated signals. Distortion with digitally modulated signals consists of in-band and out-of-band distortion. Out-of-band distortion is represented in Figure 1-39, where the spectra at the input and output of a nonlinear system are shown. The process that results in increased levels in the adjacent sidebands is called spectral regrowth. This distortion is approximately captured by the intermodulation distortion with a two-tone signal. The generation of signal in the adjacent channel affects the function of other radios and the level of signal is contained in system specifications. Distortion generated in-band affects the ability to interpret the constellation of the signal and hence difficulties in demodulating a signal. This type of distortion will be considered in the next section.

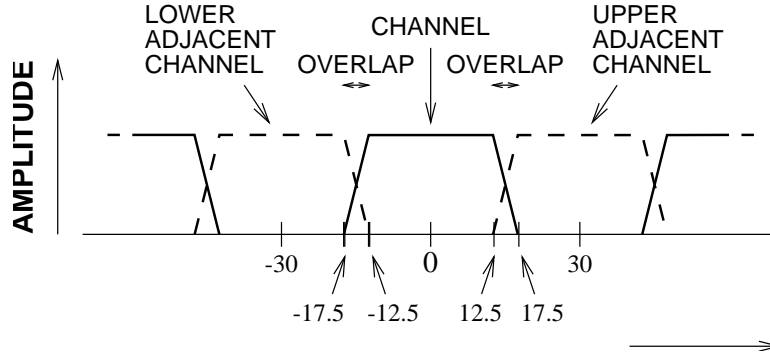
### ***1.5.5 Cochannel Interference***

The minimum signal detectable in conventional wireless systems is determined by the signal-to-interference ratio at the input. The noise is due to background noise sources, including galactic noise and thermal noise. In cellular wireless systems, the minimum signal detectable is also determined by the signal-to-interference ratio (**SIR**), but now the dominant interference is due to other transmitters in the cell and adjacent cells. The noise that is produced in the signal band from other transmitters operating at the same frequency is called cochannel interference. The level of cochannel interference is dependent on cell placement and frequency reuse patterns. The degree to which cochannel interference can be controlled has a large



**Figure 1-39** Input and output spectra of a digitally modulated signal: (a) a digitally modulated signal at the input of a nonlinear amplifier; and (b) at the output of the amplifier.





**Figure 1-40** Adjacent channels and overlap in the AMPS and DAMPS cellular systems.

effect on system capacity.

Control of cochannel interference is largely achieved by controlling the power levels at the basestation and at the mobile units. Factors affecting interference are

- The signal power falls off quickly with distance.
- Transmitted power is reduced to the minimum acceptable signal-to-interference ratio

Cochannel interference is not a nonlinear affect and is addressed using cell placement.

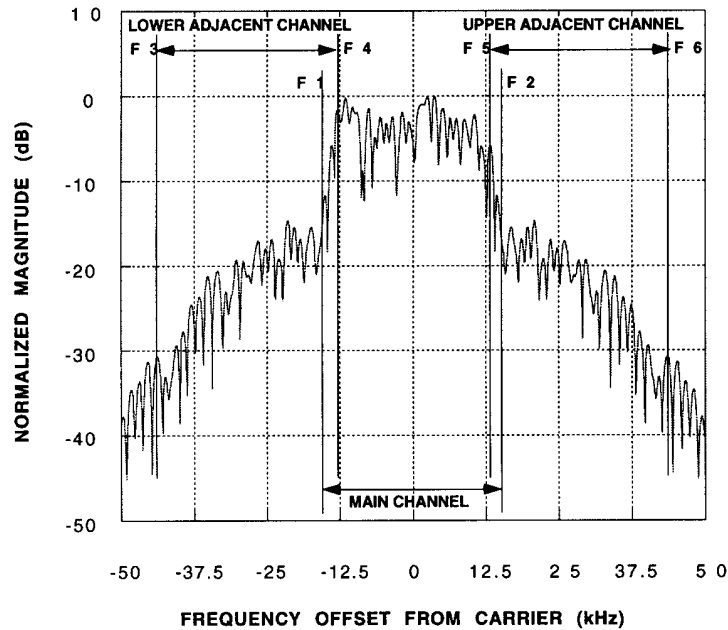
### 1.5.6 Adjacent Channel Interference

Adjacent channel interference is the result of several factors. Since ideal filtering cannot be achieved, there is inherent overlap of neighboring channels (Figure 1-40). For this reason, adjacent channels are assigned to different cells. The nonlinear behavior of transmitters also contributes to adjacent channel interference. Thus characterization of nonlinear phenomena is important in RF design. Adjacent channel interference occurs with both digitally modulated and analog modulated RF signals. It turns out that conventional design approaches can be used to control and predict adjacent channel interference for analog modulated signals but there is as yet not a good design practice for digitally modulated signals.

The spectrum of a DAMPS signal is shown in Figure 1-41. The signal between frequencies  $f_1$  and  $f_2$  is due to the digital modulation scheme and filtering. Most of the signal outside this region is due to nonlinear effects which result in what is called spectral regrowth, a process similar to third- and fifth-order intermodulation in two-tone systems. Using the frequency limits defined in Figure 1-41, the lower channel ACPR is defined as

$$\begin{aligned}
 \text{ACPR}_{\text{ADJ,LOWER}} &= \frac{\text{Power in Lower Adjacent Channel}}{\text{Power in Main Channel}} \\
 &= \frac{\int_{f_3}^{f_4} X(f)df}{\int_{f_1}^{f_2} X(f)df}, \quad (1.24)
 \end{aligned}$$





**Figure 1-41** Definition of adjacent channel and main channel integration limits using a typical DAMPS spectrum as an example.

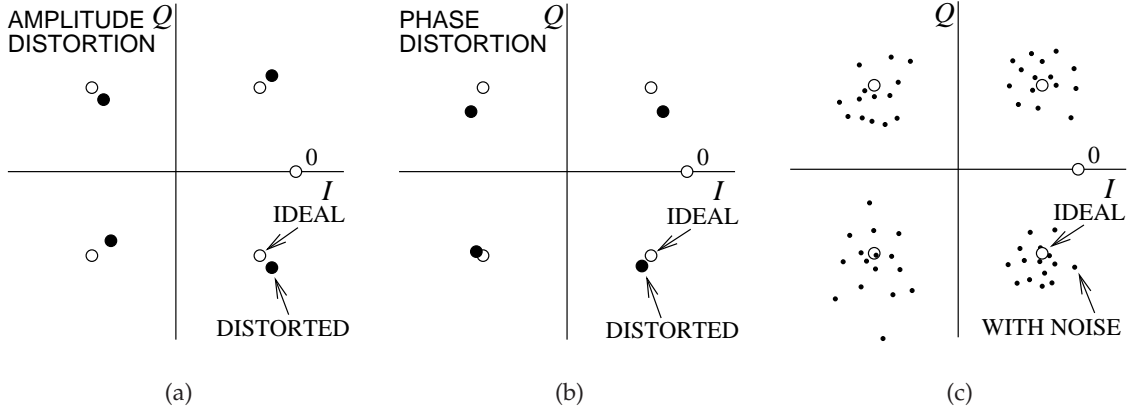
where  $X(f)$  is the RF signal spectral power density.

### 1.5.7 Noise, Distortion, and Constellation Diagrams

Noise and nonlinear distortion affect the received constellation diagram and the ability to demodulate signals. Noise is mostly introduced from the environment, particularly from other radios, but the noise introduced by the RF hardware itself is significant when the signal received is close to the minimum detectable signal. These distortion effects can be described in part by their effect on constellation diagrams (see Figure 1-42). An additional impact is impairment introduced in adjacent channels. It should be emphasized that the constellation diagram shows the state of the system at the sampling instant that is determined by the recovered clock. Errors in recovering the clock further distort the constellation diagram.

### 1.5.8 Error Vector Magnitude

The error vector magnitude (EVM) is a measure of the departure of a sampled phasor from the ideal phasor located at the constellation point. Introducing an error vector,  $X_{\text{error}}$ , and a reference vector,  $X_{\text{reference}}$ , which points to the ideal constellation point, the EVM is defined as the ratio of the power of the error vector



**Figure 1-42** Impact of signal impairments on the constellation diagram of QPSK: (a) amplitude distortion; (b) phase distortion; and (c) noise.

root mean square (RMS) power to the reference power, so that

$$\text{EVM} = \frac{|X_{\text{error}}|^2}{|X_{\text{reference}}|^2} . \quad (1.25)$$

Expressing the error and reference vector in terms of the powers  $P_{\text{error}}$  and  $P_{\text{reference}}$  respectively, enables EVM to be expressed in terms of a power ratio:

$$\text{EVM} = \frac{P_{\text{error}}}{P_{\text{reference}}} . \quad (1.26)$$

In decibels:

$$\text{EVM(dB)} = 10 \log_{10} \frac{P_{\text{error}}}{P_{\text{reference}}} . \quad (1.27)$$

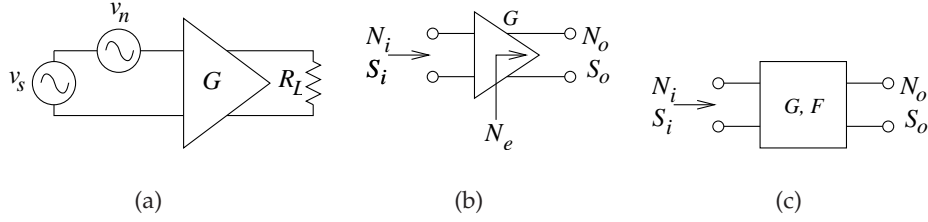
or as a percentage,

$$\text{EVM(\%)} = \frac{P_{\text{error}}}{P_{\text{reference}}} \times 100\% . \quad (1.28)$$

If the modulation format results in constellation points having different powers, the constellation point with the highest power is used as the reference.

## 1.6 Noise and Nonlinear Distortion

Noise and nonlinear distortion set the bounds on the signals that can be processed in an RF circuit. Noise establishes the minimum detectable signal while nonlinear distortion, by introducing distortion of the constellation diagram, sets the level of the largest signal from which information can be reliably extracted. The range is referred to as dynamic range and is one of the performance limits characterizing analog circuits. In this section noise is considered first and then expressions for dynamic range developed.



**Figure 1-43** Noise and two-ports: (a) amplifier; (b) amplifier with excess noise; and (c) noisy two-port network.

### 1.6.1 Noise

Amplifiers, filters and mixers in an RF frontend process (e.g., amplify, filter, and mix) input noise the same way as an input signal. In addition, these components can contribute what is called excess noise of their own. Without loss of generality, the following discussion will consider noise with respect to the amplifier shown in Figure 1-43(a), where  $v_s$  is the input signal. The noise signal with source designated by  $v_n$  is uncorrelated and random and must be described as an RMS voltage or by its noise power. The most important noise-related metric is the signal-to-noise ratio (SNR). With the noise power input to the amplifier being  $N_i$  and the signal power input to the amplifier  $S_i$ , the input SNR is  $\text{SNR}_i = S_i/N_i$ . If the amplifier is noise free then the input noise and signal powers are amplified by the power gain of the amplifier,  $G$ . Thus the output noise power is  $N_o = GN_i$  and the output signal power is  $S_o = GS_i$  and the output SNR is  $\text{SNR}_o = S_o/N_o = \text{SNR}_i$ .

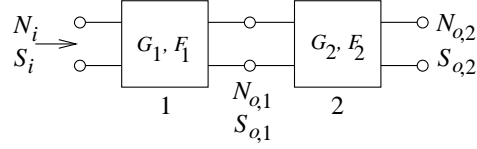
In practice, an amplifier is noisy, with the addition of excess noise,  $N_e$ , indicated in Figure 1-43(b). The excess noise originates in different components in the amplifier and is either referenced to the input or the output of the amplifier. Most commonly it is referenced to the output so that the total output noise power is  $N_o = GN_i + N_e$ . In the absence of a qualifier, the **excess noise** is referred to the output.  $N_e$  is not measured directly. Instead the ratio of the SNR at the input to that at the output is called the **noise factor**,  $F$ :

$$F = \frac{\text{SNR}_i}{\text{SNR}_o} \quad (1.29)$$

and this is the way it is normally measured. If the system circuit is noise free then  $\text{SNR}_o = \text{SNR}_i$  and  $F = 1$ . If the circuit is not noise free, then  $\text{SNR}_o < \text{SNR}_i$  and  $F > 1$ .  $F$  can be related to the excess noise produced in the circuit and from this relationship the noise of a cascaded system can be calculated. With the excess noise,  $N_e$ , referred to the circuit output,

$$\begin{aligned} F &= \frac{\text{SNR}_i}{\text{SNR}_o} = \frac{\text{SNR}_i}{1} \frac{1}{\text{SNR}_o} \\ &= \frac{S_i}{N_i} \frac{N_o}{S_o} = \frac{S_i}{N_i} \frac{GN_i + N_e}{GS_i} \\ &= 1 + \frac{N_e}{GN_i} \end{aligned} \quad (1.30)$$

One of the conclusions that can be drawn from this is that the noise factor,  $F$ , depends on the available noise power at the input of the circuit. As a standard



**Figure 1-44** Cascaded noisy two-ports.

reference the available noise power,  $N_R$ , from a resistor at **standard temperature**  $T_0$  (290 K), and over a bandwidth  $B$  is used:

$$N_i = N_R = kT_0B, \quad (1.31)$$

where  $k$  is **Boltzmann's constant**. If the input of an amplifier is connected to this resistor and all of the noise power is delivered to the amplifier, then

$$F = 1 + \frac{N_e}{GN_i} = 1 + \frac{N_e}{GkT_0B}. \quad (1.32)$$

Several random physical processes inside a circuit contribute to excess noise, and not all of these process vary linearly with temperature. Consequently  $F$  is a function of temperature although usually a weak one. It is also a function of bandwidth and there is a problem in using  $F$  with cascaded systems in which bandwidths vary for different subsystems. Even with all these problems,  $F$  is the most important measure used to characterize noise. It can be used to determine the noise performance of a cascade, when the noise factors and gains of the subsystem constituents are known.  $F$  is the ratio of powers, and when expressed in decibels, **noise figure (NF)** is used:

$$\text{NF} = 10 \log_{10} F = \text{SNR}_i(\text{dB}) - \text{SNR}_o(\text{dB}). \quad (1.33)$$

Development of the noise factor of a cascade begins by considering the noise contributions of the first system, and then the next cascaded system, and so on. The majority of RF and microwave systems are organized as cascades of two-port networks with an input port and an output port, as shown in Figure 1-44.

If the excess noise contribution of an amplifier is ignored, the output noise power will be

$$N_o = GkT_0B. \quad (1.34)$$

With excess noise,  $N_e$ , from the amplifier included, the output noise power is

$$N_o = GkT_0B + N_e = GkT_0B(1 + N_e/(GkT_0B)) = FGkT_0B. \quad (1.35)$$

Rearranging this equation the excess noise power can be written as

$$N_e = (F - 1)GkT_0B. \quad (1.36)$$

This result can be generalized for a system. Considering the second stage of the cascade, the excess noise at the output of the second stage, due solely to the noise generated internally in the second stage, is

$$N_{e,2} = (F_2 - 1)kT_0BG_2. \quad (1.37)$$

Then the total noise power at the output of a two-stage cascade is

$$N_{o,2} = (F_2 - 1)kT_0BG_2 + N_{o,1}G_2 \quad (1.38)$$

$$= (F_2 - 1)kT_0BG_2 + F_1kT_0BG_1G_2 . \quad (1.39)$$

The second term above is the noise output from the first stage amplified by the second gain.

Generalizing the above result yields the total noise power at the output of the  $m$ th stage:

$$N_{o,m} = \sum_{n=2}^m \left[ (F_n - 1)kT_0B \prod_{i=2}^n G_i \right] + F_1kT_0B \prod_{n=1}^m G_n . \quad (1.40)$$

Thus an  $m$ -stage cascade has a total cascaded system noise factor  $F^T = N_{o,m}/(G^T N_{i,1})$ , with  $G^T$  being the total cascaded available gain and  $N_{i,1}$  is the noise power input to the first stage. In terms of the parameters of individual stages the total system noise factor is

$$F^T = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1G_2} + \frac{F_4 - 1}{G_1G_2G_3} + \cdots ; \quad (1.41)$$

that is,

$$F^T = F_1 + \sum_{n=2}^m \frac{F_n - 1}{\prod_{i=1}^{n-1} G_i} . \quad (1.42)$$

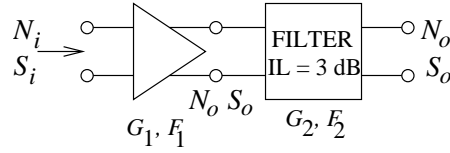
This equation is known as **Friis' formula**.

### Example 1.4 Noise Figure of an Attenuator

What is the noise figure of a 20 dB attenuator in a 50  $\Omega$  system?

**SOLUTION:** Denoting the attenuator as being in a 50  $\Omega$  system indicates that an appropriate circuit model to use in the analysis consists of the attenuator driven by a generator, with a 50  $\Omega$  source impedance, and the attenuator drives a 50  $\Omega$  load. Also, the input impedances of the terminated attenuator is 50  $\Omega$ , as is the impedance looking into the output of the attenuator when it is connected to the source. The key point is that the noise coming from the source is the noise thermally generated in the 50  $\Omega$  source impedance, and this noise is equal to the noise that is delivered to the load, as the impedance presented to the load is also 50  $\Omega$ . So the input noise,  $N_i$ , is equal to the output noise:

$$N_o = N_i . \quad (1.43)$$



**Figure 1-45** Amplifier and filter combination for which the total system noise figure is to be calculated.

The input signal is attenuated by 20 dB (= 100). So

$$S_o = S_i/100 \quad (1.44)$$

and thus the noise factor is

$$F = \frac{SNR_i}{SNR_o} = \frac{S_i N_o}{N_i S_o} = \frac{S_i}{N_i} \frac{N_i}{S_i/100} = 100 \quad (1.45)$$

and the noise figure is

$$NF = 20 \text{ dB} . \quad (1.46)$$

So the noise figure of an attenuator (or filter) is just the loss of the component. This is not true for amplifiers of course, as there are other sources of noise, and the output impedance of a transistor is not a thermal resistance.

### Example 1.5 Noise Figure of Cascaded Stages

The cascaded two-port network in Figure 1-45 consists of a noisy amplifier with a noise figure of 2 dB and a gain of 20 dB followed by a filter with an insertion loss of 3 dB. Determine the total gain and noise figure of the cascaded system.

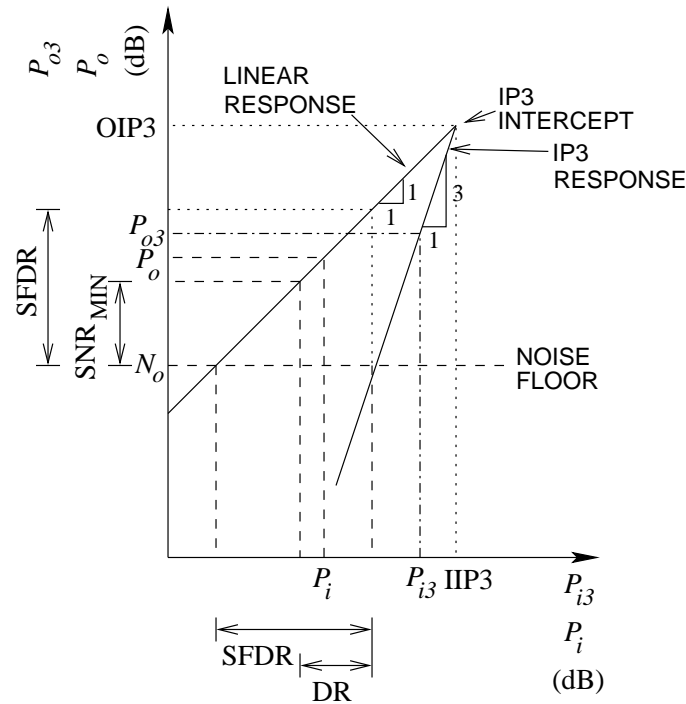
**SOLUTION:** Friis' formula can be used to calculate the total noise factor of the cascaded system, from which the total noise figure can be determined. Since the insertion loss of the filter is 3 dB, the gain of the filter is  $-3 \text{ dB}$  and its noise figure is 3 dB, thus  $G_2 = -3 \text{ dB} = 0.5$  and  $NF_2 = 3 \text{ dB}$ , so that  $F_2 = 10^{NF_2/10} = 1.995$ . Also, for the first stage,  $G_1 = 20 \text{ dB} = 100$  and  $F_1 = 10^{NF_1/10} = 1.585$ . The total gain of the cascaded system is

$$G^T(\text{dB}) = G_1(\text{dB}) + G_2(\text{dB}) = (20 \text{ dB}) + (-3 \text{ dB}) = 17 \text{ dB} . \quad (1.47)$$

The total system noise factor is, from Equation (1.42),

$$F^T = F_1 + (F_2 - 1)/G_1 = 1.585 + (1.995 - 1)/100 = 1.595 . \quad (1.48)$$

Thus the total noise figure of the system is  $NF^T = 10 \times \log_{10}(1.595) = 2.03 \text{ dB}$ . Note the importance of the first stage in determining the overall noise figure. If the gain of the first stage is sufficiently large, subsequent stages have much less of an effect.



**Figure 1-46** Output power versus input power of a stage or system plotted as output power in decibels versus input power in decibels. The IP3 response is a result of two-tone intermodulation, and the input power is the combined power of the two signals that have equal amplitude. Extrapolations of the 1:1 linear response and the 3:1 third-order intermodulation response intersect at the IP3 point.

### 1.6.2 Dynamic Range

While modern communication and radar systems use digitally modulated signals, two-tone signals are used to both approximately characterize nonlinearity, and in manual calculations. At low powers before compression becomes a factor, the fundamental response has a 1:1 slope with respect to the input, as shown in Figure 1-46. The IP3 response varies as the cube of the level of input tones when both tones vary by the same amount, as is common in a two-tone test. Thus IP3 has a 3:1 logarithmic slope with respect to the input. Since the relations are linear in a log-log sense it is possible to describe the nonlinear performance of an amplifier by a single quantity called the dynamic range (**DR**) or by the spurious free dynamic range (**SFDR**). **SFDR** and **DR** also capture noise properties.

In the following, an expression for **SFDR** is developed in terms of input-referenced quantities—the input referred **SFDR**, ( $\text{SFDR}_i$ ). A similarly referenced dynamic range **DR** ( $\text{DR}_i$ ) is also developed. **SFDR** describes the difference between a signal and the noise floor, whereas **DR** incorporates the Minimum Detectable Signal (**MDS**) which is the noise level plus a minimum acceptable SNR ( $\text{SNR}_{\text{MIN}}$ ), expressed here in decibels.

Figure 1-46 illustrates the input-output relationship of a one-tone signal to the IP3

response of a subsystem and also graphically defines the dynamic ranges. The point of intersection of the extrapolated linear output (of power  $P_o$ ) and third-order (IP3) output (of power  $P_{IP_3}$ ) is called the third-order intercept point (**IP3 intercept**). The point is identified by the output-referred intercept power (OIP3) or by the input referred IP3 intercept power (IIP3), and these are key parameters in describing the linearity of nonlinear subsystems.

In the linear gain region,  $P_o$  versus  $P_i$  has a slope of 1:1 so that

$$P_{dBm,i} = P_{dBm,o} - G_{dB} , \quad (1.49)$$

where  $G_{dB}$  is the power gain in decibels.  $P_o$  is used here as the output power, with  $P_{dBm,o}$  indicating the output power in dBm.  $P_i$  and  $P_{dBm,i}$  are similarly defined. In terms of input quantities

$$IIP3_{dBm} = OIP3_{dBm} - G_{dB} , \quad (1.50)$$

where again the dBm subscript indicates that the quantity is expressed in decibels referred to 1 milliwatt. The nonlinearity of RF active components results in harmonics and intermodulation components. With the narrowband amplifiers of communication and radar systems, output filters conveniently filter out harmonics. However, intermodulation distortion cannot be filtered out, as these components are within the main passband. The intermodulation components are therefore spurious tones. Generally just one of these defines the maximum spurious tone and nearly always it is the third-order intermodulation tone with a two-tone input. Consideration of the maximum spurious tone and the noise floor defines the SFDR.

Examining Figure 1-46 leads to the following inequality describing the linear gain of the amplifier:

$$\frac{OIP3_{dBm} - P_{dBm,o}}{IIP3_{dBm} - P_{dBm,i}} = \frac{OIP3_{dBm} - P_{dBm,o}}{(OIP3_{dBm} - G_{dB}) - P_{dBm,i}} = 1 . \quad (1.51)$$

Here,  $P_o$ ,  $P_i$ , OIP3, and IIP3 are the output power, the input power, the input-referred IP3 intercept, and the output-referred IP3 intercept. The third-order response is characterized by first introducing an equivalent input power,  $P_{dBm,i3}$  ( $P_{i3}$  expressed in dBm), defined as the average power of the two-tone signal that generates an IP3 of power  $P_{dBm,o3}$ . Noting that  $P_{dBm,o3}$  varies with a 3:1 logarithmic slope with respect to  $P_{dBm,i3}$ , then

$$\frac{OIP3_{dBm} - P_{dBm,o3}}{(OIP3_{dBm} - G_{dB}) - P_{dBm,i3}} = 3 \quad (1.52)$$

or

$$P_{dBm,i3} = \frac{1}{3} (2 \times OIP3_{dBm} + P_{dBm,o3} - 3G_{dB}) . \quad (1.53)$$

The SFDR can now be defined when the third-order intermodulation product of two-tone excitation is the dominant spurious tone. The SFDR is defined as the difference between  $P_{i3}$  and  $P_i$  when they produce IP3 and linear output respectively that are both equal to the output noise power  $N_o$  (see Figure 1-46); that is, when  $P_o = P_{o3} = N_o$ . Replacing  $P_{dBm,o}$  in Equation 1.53 with  $N_o$  gives

$$P_{dBm,i3} = \frac{1}{3} (2 \times OIP3_{dBm} + N_{dBm,o} - 3G_{dB}) \quad (1.54)$$



and

$$P_{\text{dBm},i} = N_{\text{dBm},o} - G_{\text{dB}} . \quad (1.55)$$

Note that the difference between the linear output and the third-order intermodulation reduces as the input power increases above  $P_{i3}$ . Thus the output-referred SFDR is

$$\begin{aligned} \text{SFDR}_{\text{dB},o} &= P_{\text{dBm},i3} - P_{\text{dBm},i} \\ &= \frac{2}{3}\text{OIP3}_{\text{dBm}} + \frac{1}{3}N_{\text{dBm},o} - G_{\text{dB}} - N_{\text{dBm},o} + G_{\text{dB}} \\ &= \frac{2}{3}(\text{OIP3}_{\text{dBm}} - N_{\text{dBm},o}) . \end{aligned} \quad (1.56)$$

A similar development can be used to define the input-referred SFDR:

$$\text{SFDR}_{\text{dB},i} = \frac{2}{3}(\text{IIP3}_{\text{dBm}} - N_{\text{dBm},i}) . \quad (1.57)$$

Note that  $N_i$  is the input-referred noise and includes noise applied to the subsystem as well as the noise produced internally in the subsystem and referred to the input. The SFDR provides a combined measure of distortion and noise. However, for usable dynamic range the minimum acceptable SNR must be considered. The minimum SNR ( $\text{SNR}_{\text{MIN}}$ ) required is determined by the communication or radar modulation format, error coding, and acceptable BER. So in defining DR the input power of the desired signal must increase sufficiently to produce an SNR of at least  $\text{SNR}_{\text{MIN}}$ . Since the desired spurious level is still at the noise floor, this implies a direct subtraction in decibels of the desired SNR. Therefore the input-referred third-order dynamic range, preferred for use with receivers, is

$$\text{DR}_i = \frac{2}{3}(\text{IIP3} - N_{\text{dBm},i}) - \text{SNR}_{\text{MIN}} \quad (1.58)$$

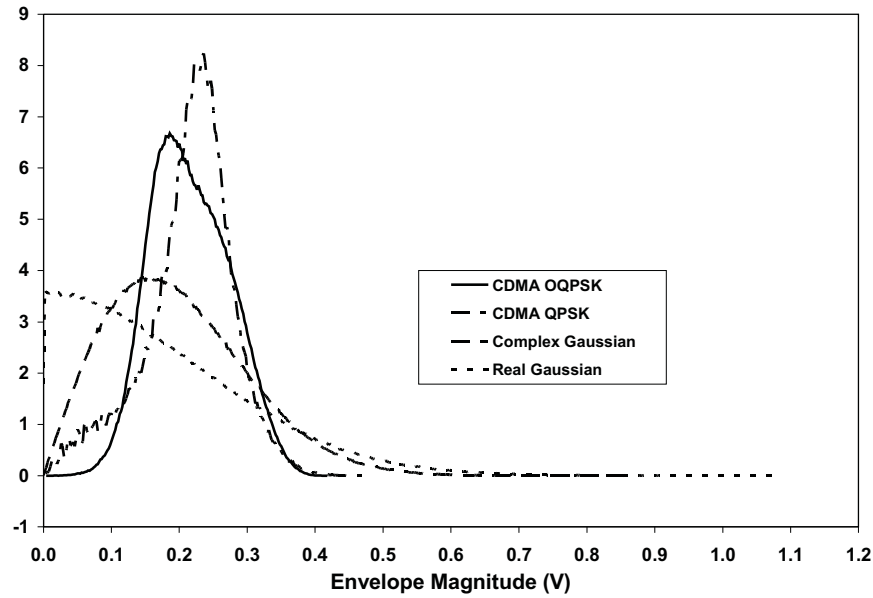
and the output-referred dynamic range, preferred for use with transmitters, is

$$\text{DR}_o = \frac{2}{3}(\text{OIP3} - N_{\text{dBm},o}) - \text{SNR}_{\text{MIN}} . \quad (1.59)$$

### 1.6.3 Probability Density Function and Distortion

Amplitude modulation is not inherently required for information transmission with PSK modulation schemes. For instance, a QPSK signal consists of two digital data streams, equal in amplitude, modulated in quadrature onto a carrier signal. The resulting signal would have a constant envelope; however, the occupied bandwidth is quite large, as the spectrum of a pulse train is  $\sin(x)/x$ , the sinc function. The first **sidelobe** of the sinc spectrum is only 13 dB down from the carrier level and is in the middle of the adjacent channel. To reduce the spectrum, a low-pass filter is typically applied to each digital data stream to minimize the out-of-band spectrum of the modulated signal. This comes with a drawback: the filters cause a finite memory effect resulting in amplitude variations as the ringing energy from a previous data pulse adds to the current filtered data pulse.

Amplitude variations of the modulated signal are characterized by measured waveform statistics such as the PAR. A signal with a high PAR requires that the RF system have high linearity to handle the average power requirements



**Figure 1-47** Amplitude PDF for CDMA and Gaussian modulation signals. After [6].

and the peak amplitude excursions without generating excessive out-of-band distortion. However, it is possible for a signal with a higher PAR to exhibit less nonlinear distortion than a signal with lower PAR [5]. The reason for this apparent inconsistency is because the signal peak is a singular point measurement with, typically, a low probability of occurrence. Thus PAR is an incomplete statistic for determining the linearity requirements for a transmitter to carry a signal.

The amplitude probability density function (APDF) is a more complete statistical description of the amplitude variations of a modulated signal. The APDF defines the maximum and minimum variation along with the relative probability of occurrence of amplitudes within the variation. The APDF is typically estimated from a histogram of amplitudes, with a uniform bin size, by

$$f(A) = \frac{N}{\Delta A \times N_c}, \quad (1.60)$$

where  $N$  is the number of counts per bin,  $\Delta A$  is the bin amplitude width, and  $N_c$  is the total number of samples. The shape of the amplitude density between the mean and peak amplitude influences the sensitivity of a particular signal to spectral regrowth due to nonlinear gain compression or expansion. For example, Figure 1-47 shows the APDF for a **CDMA** mobile transmitter using OQPSK modulation, the same signal using QPSK modulation, a real **Gaussian signal**, and a complex Gaussian QPSK signal (with  $I$  and  $Q$  each having Gaussian distribution) where the average power of each signal is set to 0 dBm. Gaussian signals are of particular interest as their simple statistics lends them, and their interaction with nonlinearities, to quasi-analytic treatment. The PAR for each signal is shown in Table 1-7. The shape of the amplitude density after the mean differs for both signals where it can be seen that it is difficult to determine, a priori, which signal will be

**Table 1-7** Peak-to-average ratios in decibels for OQPSK and QPSK used in CDMA compared to the PAR's of Gaussian signals.

Signal Modulation	PAR (dB)
OQPSK CDMA	5.4
QPSK CDMA	6.6
Real Gaussian	13.5
Complex Gaussian	11.8

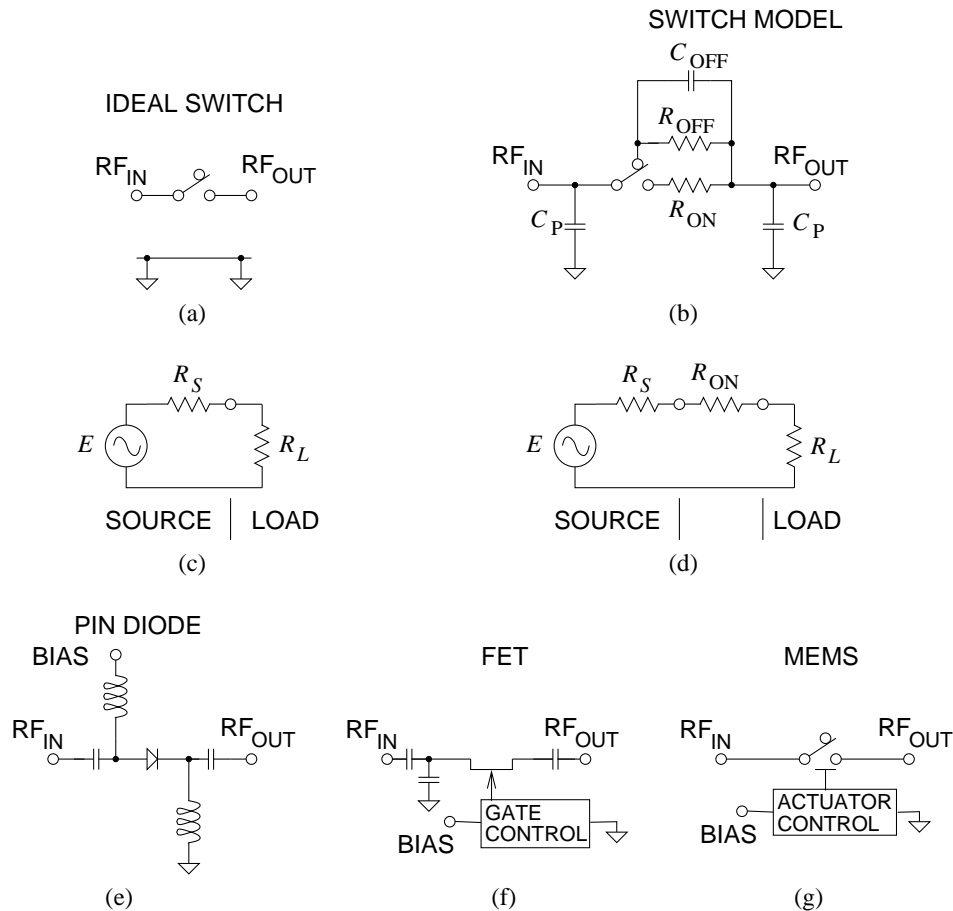
more sensitive to nonlinear gain compression. For example, even though QPSK has a higher PAR than OQPSK, the probability that the OQPSK signal is near the peak is higher than for the QPSK signal. It is not surprising then that the measured spectral regrowth of a OQPSK signal is higher than for a QPSK signal. So PAR is only a rough guide to the distortion that is produced.

## 1.7 Switches

Microwave switches are commonly used to alternately connect an antenna to a transmitter or a receiver. In some communication systems, such as GSM, a phone does not transmit and receive simultaneously. Consequently a switch can be used to separate the transmitted and received signals. In multi-band phones, a switch is used to connect the right transmitter and receiver, which are band specific, to the antenna. In radar systems, switches are used to steer an antenna beam by changing the phase of the microwave signal delivered to each antenna in an array of antennas. An ideal microwave switch is shown in Figure 1-48(a) where an input port, designated  $RF_{IN}$ , and an output port, identified as  $RF_{OUT}$ , are shown. For maximum power transfer between the ports, the switch should have little loss and so small on resistance. At microwave frequencies, switches must be modeled with parasitics and have finite on and off resistance. A realistic model applicable to many switch types is shown in Figure 1-48(b). The capacitive parasitics, the  $C_P$ 's, limit the frequency of operation of the switches, and the on resistance,  $R_{ON}$ , impacts the switch loss. Ideally the off resistance,  $R_{OFF}$ , is very large however the parasitic shunt capacitance,  $C_{OFF}$ , is nearly always more significant. The result is that at high frequencies, there is an alternative capacitive connection between the input and output through  $C_{OFF}$ . The on resistance of the switch introduces voltage division which can be seen by comparing the ideal connection shown in Figure 1-48(c), and the more realistic connection shown in Figure 1-48(d). From the voltage division ratio the loss of the filter can be calculated.

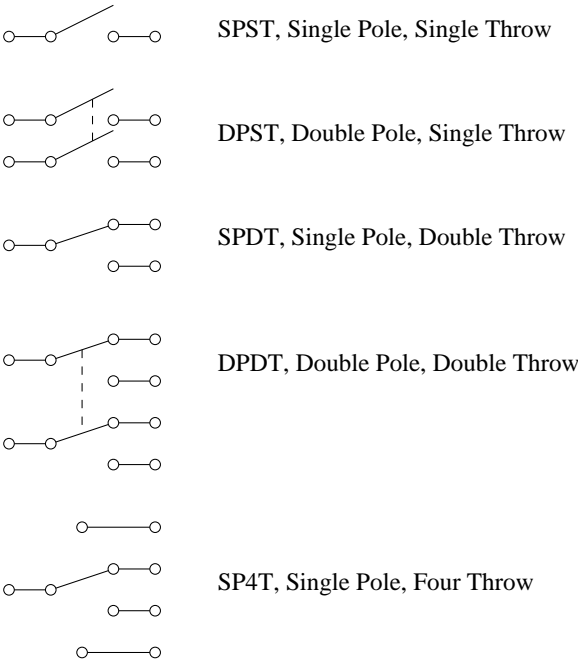
Switches are configured to provide connection from one or more inputs to one or more outputs. The configuration of a switch is indicated by poles and throws and several configurations are shown in Figure 1-49. Most commonly in microwave applications single pole switches are used and this input is connected to an antenna. The throws would be connected to different bands of a multi-band phones for example.

There are four main types of microwave switches: mechanical, **PIN diode**, FET,



**Figure 1-48** Microwave switches: (a) ideal switch connecting  $RF_{IN}$  and  $RF_{OUT}$  ports; (b) model of a microwave switch; (c) ideal circuit model with switch on and with source and load; (d) realistic low-frequency circuit model with switch on and with source and load; (e) switch realized using a PIN diode; (f) switch realized using a FET; and (g) switch realized using a MEMS switch; .

and **MEMS** switches. Mechanical switches are nearly ideal but tend to be large, relatively expensive, and mostly used in laboratory settings. The other switches are of most interest for use in systems. The PIN diode, FET, and MEMS switches are shown in Figures 1-48(e), 1-48(f), and 1-48(g), respectively. With these technologies, most higher order switches are based on interconnections of **SPST switches**. The attributes of these switches is summarized in Table 1-8 for switches that are suitable for cell phone applications. PIN diode switches are the most robust, handling the most RF power, and operating at higher frequencies, than either FET- or MEMS-based switches. However this comes at a price. The PIN diode used is similar to a PN junction diode with the addition of an intrinsic layer between the p- and n-type materials. With applied forward bias the diode has low series resistance. In reverse bias the diode resistance is large. Forward bias requires DC current and voltage, so control power is consumed when a PIN diode switch is on. The circuit



**Figure 1-49** Switch configurations.

**Table 1-8** Typical properties of small microwave switches. (Sources: <sup>1</sup> Radant MEMS, <sup>2</sup> RF Micro Devices, and <sup>3</sup> Tyco Electronics.)

Switch Type	Configuration	Power Handling	Loss at 2 GHz	Operating Frequency	Actuation Voltage	Response Time
MEMS <sup>1</sup>	SPST	1 W	0.15 dB	to 12 GHz	40–120 V	5 $\mu$ s
MEMS <sup>1</sup>	SPST	0.5 W	0.27 dB	to 40 GHz	40–120 V	5 $\mu$ s
pHEMT <sup>2</sup>	SPDT	2 W	0.5 dB	to 2.5 GHz	6 V	0.5 $\mu$ s
pHEMT <sup>2</sup>	SPDT	10 W	0.5 dB	to 6 GHz	6 V	0.5 $\mu$ s
PIN <sup>3</sup>	SPDT	13 W	0.35 dB	to 2 GHz	12 V	0.5 $\mu$ s
PIN <sup>3</sup>	SPDT	10 W	0.4 dB	to 6 GHz	12 V	0.5 $\mu$ s

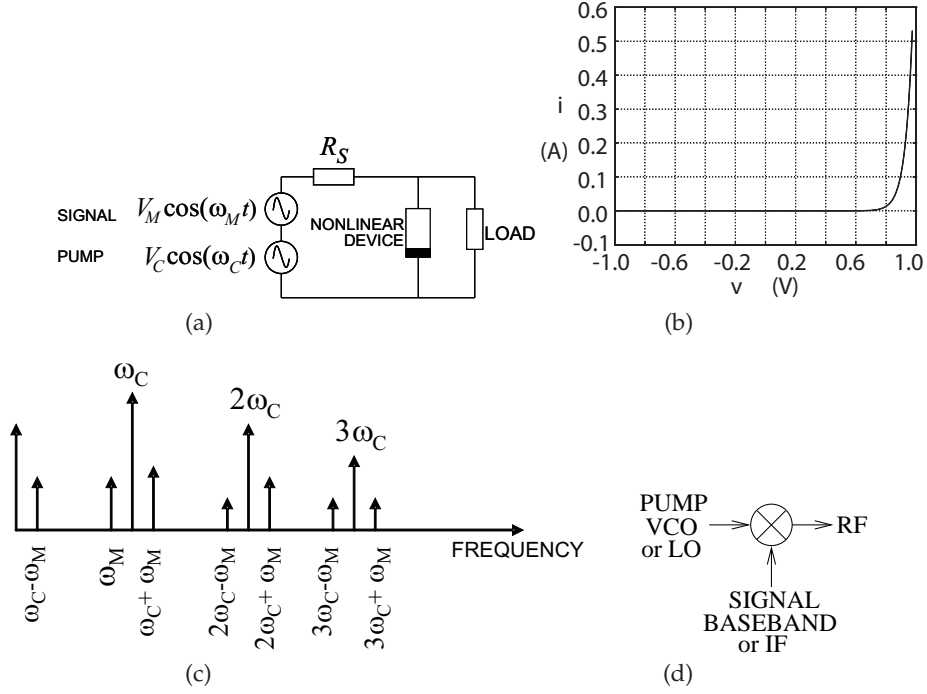
configuration for an SPST PIN diode switch is shown in Figure 1-48(e). Series bias decoupling capacitors are required at the RF ports.

A field effect transistors (FET) makes a good electronic switches; with the correct bias applied to the gate, the drain-source connection looks like a small resistance. Changing the bias to the other extreme removes free carriers from the channel between the drain and source, and a large resistance is the result. Both Si and GaAs switches are used at cellular frequencies with GaAs switches operating at extended frequencies approaching 6 GHz. The operation of an FET can be described as a variable drain-source resistance with the gate-source voltage controlling the crosssection of the channel. The circuit for a FET-based SPST switch is shown in Figure 1-48(f). Series bias blocking capacitors are required at the RF ports. Control power is only required to change the state of the switch; negligible power is required to maintain the switch state.

A microelectromechanical system (MEMS) switch is fabricated using photolithographic techniques similar to those used in semiconductor manufacturing. They are essentially miniature mechanical switches with a voltage used to control the position of a shorting arm which is usually a cantilever or a membrane. As there is no direct connection between the RF signal path and the control circuitry, MEMS switches have inherently high operating frequencies. Power is required to change the switch but once switching has been accomplished negligible DC power is required to maintain the connection.

## 1.8 Mixers

Frequency conversion, mixing or heterodyning, is the process of converting information at one frequency (present in the form of a modulated carrier) to another frequency. The second frequency is either higher, in the case of frequency **up-conversion**, where it is more easily transmitted, or lower, when mixing is called frequency **down-conversion**, where it is more easily captured. Capture of the down-converted signal is nearly always by an Analog to Digital Converter (ADC). Frequency conversion occurs with any nonlinear element. In Figure 1-50(a) a nonlinear device is driven by two signals at  $\omega_m$  and  $\omega_c$ . The larger signal, the LO, is called the **pump** and the other signal is called the RF. The spectrum of the signals present in the circuit is shown in Figure 1-50(c). The aim here is to produce a signal at the difference frequency (or intermediate frequency (IF)) with the same modulation, and hence the same information, as the original RF signal. Two mixers based on transistors are shown in Figure 1-51. The transistor mixer shown in Figure 1-51(a) uses filtering to separate the RF, LO, and IF components. Filters can be large, so one of the particular advantages of the **Gilbert mixer** shown in Figure 1-51(b) is that it is a balanced mixer and filtering is not required to separate the various signals. The symmetrical (or balanced) nature of this circuit means that only differential mode signals at the input of the common source differential pairs can appear at the output. Thus the largest signal present, the LO, is suppressed. A balanced mixer can also be realized using diodes arranged in a ring to form the **diode ring double-balanced mixer** shown in Figure 1-52. This mixer has the advantage of being bidirectional, whereas the transistor mixer circuits of Figure 1-51 are unidirectional.



**Figure 1-50** Diode mixer: (a) circuit; (b) diode current-voltage characteristic; (c) spectrum across the nonlinear device; and (d) schematic symbol for a mixer.

### 1.8.1 Mixer Analysis

A two-tone input

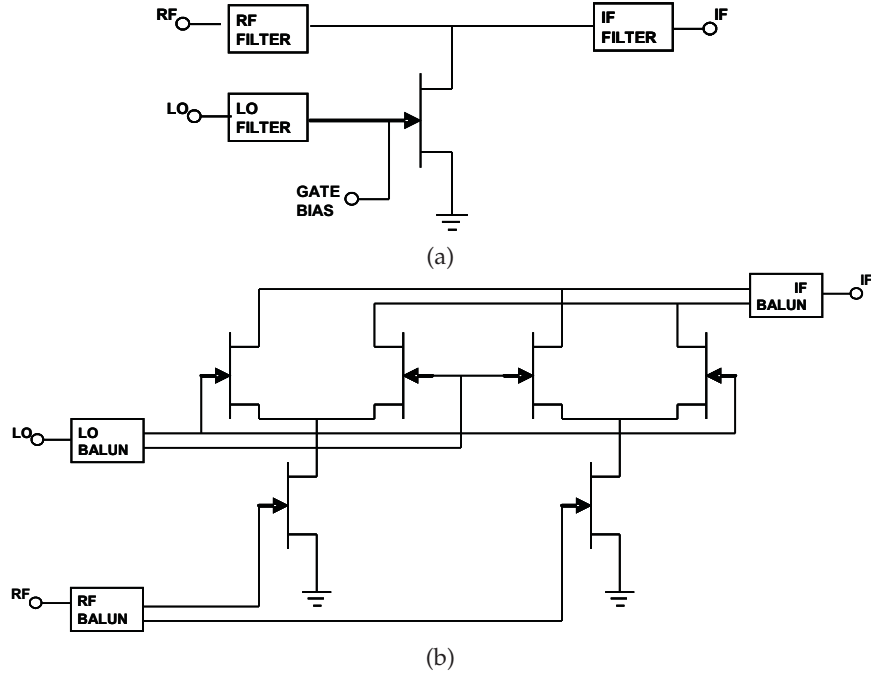
$$x(t) = |X_1| \cos(\omega_1 t + \phi_1) + |X_2| \cos(\omega_2 t + \phi_2)$$

can be written using complex notation as

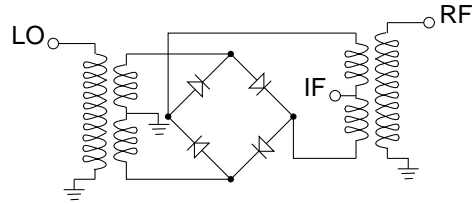
$$x(t) = \frac{1}{2} \left[ \tilde{X}_1 e^{j\omega_1 t} + \tilde{X}_1^* e^{-j\omega_1 t} + \tilde{X}_2 e^{j\omega_2 t} + \tilde{X}_2^* e^{-j\omega_2 t} \right].$$

Note that the coefficient of the positive exponential frequency component is one-half that of the phasor. Thus the phasor of the  $\omega_1$  component is  $\hat{X}_1 = |X_1| e^{j\phi} = 2\tilde{X}_1$ , and the phasor of the  $\omega_2$  component is  $\hat{X}_2 = |X_2| e^{j\phi} = 2\tilde{X}_2$ . So the first three powers of  $x$  can be easily expanded manually, for example, expanding  $x^2$  gives

$$\begin{aligned} x^2(t) = & \left( \frac{1}{2} \right)^2 \left[ X_1^2 e^{j2\omega_1 t} + 2X_1 X_1^* + 2X_1 X_2 e^{j(\omega_1 + \omega_2)t} + 2X_1 X_2^* e^{j(\omega_1 - \omega_2)t} \right. \\ & + (X_1^*)^2 e^{-j2\omega_1 t} + 2X_1^* X_2 e^{j(\omega_2 - \omega_1)t} + 2X_1^* X_2^* e^{-j(\omega_1 + \omega_2)t} + X_2^2 e^{j2\omega_2 t} \\ & \left. + 2X_2 X_2^* + (X_2^*)^2 e^{-j2\omega_2 t} \right], \end{aligned} \quad (1.61)$$



**Figure 1-51** Transistor-based mixer circuits: (a) single-ended **FET mixer** with LO, RF and IF bandpass filters; (b) Gilbert mixer with baluns producing differential LO and RF signals.



**Figure 1-52** Diode ring double-balanced mixer.

and similarly expanding  $x^3$  yields

$$\begin{aligned}
 x^3(t) = & \left(\frac{1}{2}\right)^3 \left[ X_1^3 e^{j3\omega_1 t} + 3X_1^2 X_1^* e^{j\omega_1 t} + 3X_1^2 X_2 e^{j(2\omega_1 + \omega_2)t} \right. \\
 & + 3X_1^2 X_2^* e^{j(2\omega_1 - \omega_2)t} + 3X_1 (X_1^*)^2 e^{-j\omega_1 t} + 6X_1 X_1^* X_2 e^{j\omega_2 t} \\
 & + 6X_1 X_1^* X_2^* e^{-j\omega_2 t} + 3X_1 X_2^2 e^{j(\omega_1 + 2\omega_2)t} \\
 & + 6X_1 X_2 X_2^* e^{j\omega_1 t} + 3X_1 (X_2^*)^2 e^{j(\omega_1 - 2\omega_2)t} + (X_1^*)^3 e^{-j3\omega_1 t} \\
 & + 3(X_1^*)^2 X_2 e^{j(\omega_2 - 2\omega_1)t} + 3(X_1^*)^2 X_2^* e^{-j(2\omega_1 + \omega_2)t} + 3X_1^* X_2^2 e^{j(2\omega_2 - \omega_1)t} \\
 & + 3X_1^* (X_2^*)^2 e^{-j(\omega_1 + 2\omega_2)t} + X_2^3 e^{j3\omega_2 t} + 6X_1^* X_2^* X_2 e^{-j\omega_1 t} + 3X_2^2 X_2^* e^{j\omega_2 t} \\
 & \left. + 3X_2 (X_2^*)^2 e^{-j\omega_2 t} + (X_2^*)^3 e^{-j3\omega_2 t} \right], \quad (1.62)
 \end{aligned}$$



**Table 1-9** The intermodulation products resulting from  $x$ ,  $x^2$ , and  $x^3$ , where  $x$  is a two-tone signal, showing only the positive frequencies. The first column gives the complex amplitudes of the frequency components.

Intermodulation Product	Frequency	Order
$\frac{1}{2}X_1X_1^*$	0	2
$\frac{1}{2}X_2X_2^*$	0	2
$2\frac{1}{2}X_1$	$\omega_1$	1
$2(\frac{1}{2})^33X_1^2X_1^*$	$\omega_1$	3
$2(\frac{1}{2})^36X_1X_2X_2^*$	$\omega_1$	3
$2\frac{1}{2}X_2$	$\omega_2$	1
$2(\frac{1}{2})^33X_2^2X_2^*$	$\omega_2$	3
$2(\frac{1}{2})^36X_1X_1^*X_2$	$\omega_2$	3
$2(\frac{1}{2})^2X_1^2$	$2\omega_1$	2
$2(\frac{1}{2})^2X_2^2$	$2\omega_2$	2
$2(\frac{1}{2})^3X_1^3$	$3\omega_1$	3
$2(\frac{1}{2})^3X_2^3$	$3\omega_2$	3
$2\frac{1}{2}X_1X_2$	$\omega_1 + \omega_2$	2
$2\frac{1}{2}X_1X_2^*$	$\omega_1 - \omega_2$	2
$2(\frac{1}{2})^33X_1^2X_2$	$2\omega_1 + \omega_2$	3
$2(\frac{1}{2})^33X_1^2X_2^*$	$2\omega_1 - \omega_2$	3
$2(\frac{1}{2})^33X_1X_2^2$	$\omega_1 + 2\omega_2$	3
$2(\frac{1}{2})^33X_1^*X_2^2$	$2\omega_2 - \omega_1$	3

so that the output of the cubic equation

$$y(t) = a_0 + a_1x(t) + a_2x^2(t) + a_3x^3(t)$$

can be calculated for a two-tone input. Table 1-9 lists these phasors and groups them according to frequency. The phasors of the various intermodulation products resulting from  $x$ ,  $x^2$ , and  $x^3$  can be taken as the coefficients of the positive exponential frequency components after the factor of two correction for terms other than DC. Terms of the same frequency are summed to obtain the output at a particular frequency. For example, the phasor output at  $\omega_1$  is given by the sum of three intermodulation products:

$$Y_{\omega_1} = a_1 \left(\frac{1}{2}\right) X_1 + 3a_3 \left(\frac{1}{2}\right)^3 X_1^2X_1^* + 6a_3 \left(\frac{1}{2}\right)^3 X_1X_2X_2^* . \quad (1.63)$$

## 1.9 Early Receiver Technology

In this section, historical receivers are considered first, in part because the terms associated with the early receivers are still used, but also because the early trade-offs influence the architectures used today. Today receivers use DSP technology, very stable LOs and sophisticated clock recovery schemes. This was not always so. One of the early problems was using an LO to demodulate a signal when transmitter oscillators drifted by many kilohertz. Radio at first used AM and the carrier was sent with the information-carrying sidebands. With this signal, a simple rectifier circuit connected to a bandpass filter could be used to rectify, but the reception was poor. A crystal rectifier consists of a single diode with filters. To improve performance it was necessary to lock an oscillator to the carrier and then amplify the received signal. Here some of the early schemes that addressed some of the problems are discussed. There were many more variants, but the discussion covers the essential ideas.

### 1.9.1 *Heterodyne Receiver*

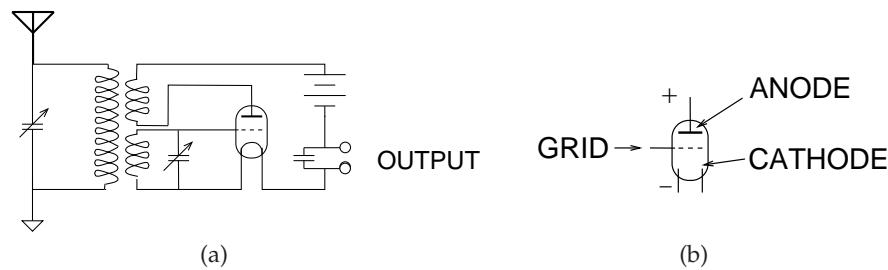
The heterodyning principle mixes a single-tone signal, the LO, with a finite bandwidth signal to produce a lower frequency version of the information-bearing signal. With the LO frequency set appropriately, the low-frequency signal would be in the audio range. If the information-bearing signal is an AM signal, then the low-frequency version of the signal is the original audio signal which is the envelope of the AM signal. This type of receiver is called a **tuned radio frequency (TRF) receiver**, and performance is critically dependent on the stability of the LO and the selectivity of the receive filter. The TRF receiver required the user to adjust a tunable capacitor so that, with a fixed inductor, a tunable bandpass filter was created. Such a filter has limited  $Q^4$  and a bandwidth that is wider than the bandwidth of the radio channel. Even worse, a user had to adjust both the frequency of the bandpass filter and the frequency of the LO. The initial radios based on this principle were called **audions**, used a triode vacuum tube, and have been in use since 1906. They were an improvement on the **crystal detectors**, but there was a need for something better.

### 1.9.2 *Homodyne Receiver*

The homodyne [7], syncrodyne (for synchronous heterodyne) [8], and autodyne (for automatic heterodyne) circuits were the needed improvements on the audion and are based on the regenerative circuit invented by Edwin **Armstrong** in 1912 while he was an electrical engineering student at New York City's Columbia University [9]. Armstrong's circuit fed the input signal into an amplifying circuit and a portion of this signal was coupled back into the input circuit so that the signal was amplified over and over again. This is a positive feedback amplifier. A

---

<sup>4</sup>  $Q$  is the **quality factor** and is the ratio of the energy stored to the energy resistively lost each cycle. Good frequency selectivity in a filter requires high  $Q$  components. Tunable components have lower  $Q$  than fixed components.



**Figure 1-53** The Colebrook's original homodyne receiver: (a) circuit with an antenna, tunable bandpass filter and triode amplifier; and (b) triode vacuum tube.

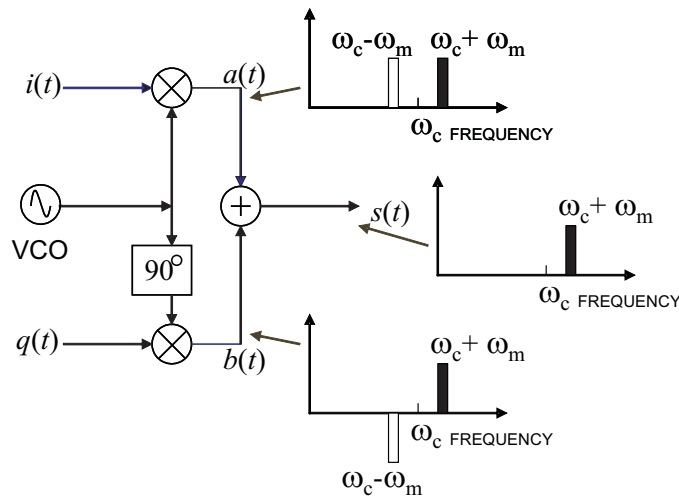
small input RF signal was amplified to such a large extent that it resulted in the amplifying circuit becoming nonlinear and consequently it rectified the amplitude modulated RF signal. **Colebrook** used this principle and developed the original homodyne receiver shown in Figure 1-53(a). This serves to illustrate the operation of the family of regenerative receivers. The antenna shown on the left-hand side is part of a resonant circuit that is in the feedback path of a triode oscillator. The triode vacuum tube is annotated in Figure 1-53(b). Here the grid coils (which control the flow of carriers between the bottom cathode<sup>5</sup> and top anode) are weakly coupled to the anode circuit. When an AC signal appears at the top anode, the part within the passband of the tuned circuit is fed back to the grid and the signal reinforced. The radio signals of the day were AM and had a relatively large carrier and so the oscillator tended to lock on to the carrier. The AM sidebands were then successfully heterodyned down to the desired audio frequencies.

The **autodyne** worked on a slightly different principle in that the oscillation frequency was tuned to a slightly different frequency from the carrier. Still, the autodyne combined the functions of an oscillator and detector in the same circuit.

### 1.9.3 Superheterodyne Receiver

The superheterodyne receiver was invented by Edwin Armstrong in 1918 [10]. The key concept was to **heterodyne** down in two stages and to use fixed filters and use a tunable LO. The receiving antenna was connected to a bandpass filter that allowed several channels to pass. This relaxed the demands on the receive filter, but also filters with higher selectivity could be constructed if they did not need to be tuned. Today we use high-order filters that are manually or machine tuned, as manufacturing tolerances do not allow high- $Q$  high-order filters to be manufactured unmodified. The filtered received signal is then mixed with an offset LO to produce what was called a **supersonic** signal—a signal above the audio range—and hence the name of this architecture. The performance of the superheterodyne (or super HET) receive architecture has only recently been achieved at cellular frequencies using direct conversion architecture requiring large-scale integrated (and hence silicon) circuits. However, the superheterodyne

<sup>5</sup> The **cathode** is heated (the heater circuit is not shown) and electrons are spontaneously emitted in a process called **thermionic emission**.



**Figure 1-54** Quadrature modulator showing intermediate spectra.

architecture is still superior above about 6 GHz.

## 1.10 Modern Transmitter Architectures

Modern transmitters maximize both spectral efficiency and electrical efficiency. Spectral efficiency is achieved by suppressing the carrier on transmit and transmitting a single sideband. The classic technique for achieving this is quadrature modulation, described in the next section. Electrical efficiency must be achieved with tight specifications on allowable distortion and designs must achieve this with minimum manual adjustments. Electrical efficiency has resulted in compound semiconductor transistors, including GaAs HBTs and pHEMTs, mostly preferred for cellular handsets. For basestation and point-to-point applications Si LDMOS is the dominant technology below a few gigahertz, with high-breakdown gallium nitride (GaN) FETs being introduced. Another trend is the development of universal amplifier concepts so that the same RF frontend can be used for a number of different applications. Multifunctional capability is a cost driving transmitter architectures to minimize the RF analog hardware. The discussion here focuses on narrowband communications when the modulated RF carrier can be considered as a slowly varying RF phasor.

### 1.10.1 Quadrature Modulation

Quadrature modulation describes the frequency conversion process in that the real and imaginary parts of the RF phasor are varied separately. A subsystem which implements quadrature modulation is shown in Figure 1-54. This is quite an ingenious circuit. The operation of this subsystem is described by what is known

as the generalized quadrature modulation equation:

$$s(t) = i(t) \cos [\omega_c t + \varphi_i(t)] + q(t) \sin [\omega_c t + \varphi_q(t)] . \quad (1.64)$$

Here,  $i(t)$  and  $q(t)$  embody the particular modulation rule for amplitude,  $\varphi_i(t)$  and  $\varphi_q(t)$  embody the particular modulation rule for phase, and  $\varphi_c$  is the carrier radian frequency. In terms of the signals identified in Figure 1-54, the quadrature modulation equation can be written as

$$s(t) = a(t) + b(t) \quad (1.65)$$

$$a(t) = i(t) \cos [\omega_c t + \varphi_i(t)] \quad (1.66)$$

$$b(t) = q(t) \sin [\omega_c t + \varphi_q(t)] , \quad (1.67)$$

where  $a(t)$  describes the output of the mixer at the top and  $b(t)$  describes the output of the mixer on the bottom. The spectrum of  $a(t)$  as shown in Figure 1-54 has two bands above and below the frequency of the carrier,  $\omega_c$ . Similarly the spectrum of  $b(t)$  has two bands above and below the frequency of the carrier. However, there is a difference. The LO (here designated as the **voltage-controlled oscillator (VCO)**) is shifted  $90^\circ$  (perhaps using an RC delay line) so that the frequency components of  $b(t)$  have a different phase relationship to the carrier than those of  $a(t)$ . When  $a(t)$  and  $b(t)$  are combined the carrier content is canceled, as is one of the sidebands. This is exactly what is desired: the carrier should not be transmitted, as it contains no information. Also it is desirable to transmit only one sideband, as it contains all of the information in the modulating signal. This type of modulation is called **suppressed carrier single-sideband (SCSS)** modulation. The actual characteristics depend on the particular forms of  $i(t)$ ,  $q(t)$ ,  $\varphi_i(t)$  and  $\varphi_q(t)$ , and these define the modulation schemes. Only a few have the optimum properties of a well-defined spectrum with steep sidewalls so that adjacent channels can be closely packed. In the next section frequency modulation is used to demonstrate the SCSS operation. In digital modulation,  $i(t)$  and  $q(t)$  are each derived from a bitstream, perhaps simply by filtering a binary waveform.

### 1.10.2 Frequency Modulation

Frequency modulation is considered here to demonstrate suppressed carrier single-sideband operation. Let  $i(t)$  and  $q(t)$  be finite bandwidth signals centered at radian frequency  $\omega_m$  with  $(\phi_q(t) - \phi_i(t))$  being  $90^\circ$  on average. This is shown in Figure 1-54, where  $\omega_m$  represents the frequency components of  $i(t)$  and  $q(t)$ . With reference to Figure 1-54, and setting  $\varphi_i(t) = 0 = \varphi_q(t)$ ,

$$i(t) = \cos(\omega_m t) \quad \text{and} \quad q(t) = -\sin(\omega_m t) \quad (1.68)$$

and the general quadrature modulation equation, Equation (1.64), becomes

$$s(t) = i(t) \cos (\omega_c t) + q(t) \sin (\omega_c t) = a(t) + b(t)$$

where

$$a(t) = \frac{1}{2} \{ \cos[(\omega_c - \omega_m)t] + \cos[(\omega_c + \omega_m)t] \} \quad (1.69)$$

and

$$b(t) = \frac{1}{2} \{ \cos[(\omega_c + \omega_m)t] - \cos[(\omega_c - \omega_m)t] \} . \quad (1.70)$$

So the combined frequency modulated signal at the output is

$$s(t) = a(t) + b(t) = \cos[(\omega_c + \omega_m)t] , \quad (1.71)$$

and as well as the carrier being suppressed, so is the lower sideband. This lower sideband is also referred to as the image. In modulators it is important to suppress this image and in demodulators it is important that undesired signals at the image frequency not be converted along with desired signals .

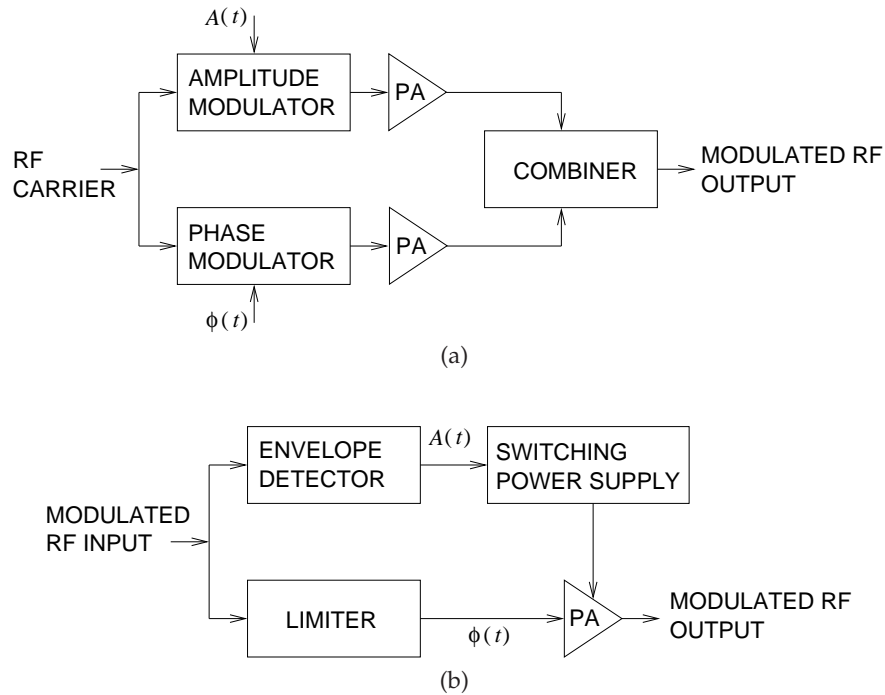
### 1.10.3 Polar Modulation

Polar modulation is a relatively new modulation scheme for impressing information on a carrier. In polar modulation, the  $i(t)$  and  $q(t)$  quadrature signals are converted to polar form as amplitude  $A(t)$  and phase  $\phi(t)$  components. This is either done in the DSP unit or if a modulated RF carrier is all that is provided, using an envelope detector to extract  $A(t)$  and a limiter to extract the phase information corresponding to  $\phi(t)$ . Two polar modulator architectures are shown in Figure 1-55. In the first architecture, Figure 1-55(a),  $A(t)$  and  $\phi(t)$  are available and  $A(t)$  is used to amplitude modulate the RF carrier which is then amplified by a Power Amplifier (PA). The phase signal,  $\phi(t)$  is the input to a phase modulator implemented as a PLL. The output of the PLL is fed to an efficient (i.e., nonlinear) amplifier operating near saturation (also called a **saturating amplifier**). The outputs of the two amplifiers are combined to obtain the large modulated RF signal to be transmitted.

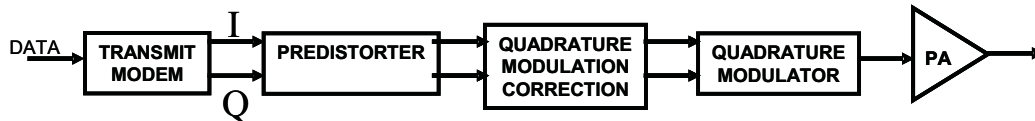
In the second polar modulation architecture, Figure 1-55(b), a low-power modulated RF signal is decomposed into its amplitude and phase modulated components. The phase component,  $\phi(t)$ , is extracted using a limiter which produces a pulse-like waveform with the same zero crossings as the modulated RF signal. Thus the phase of the RF signal is captured. This is then fed to a saturating amplifier whose gain is controlled by the carrier envelope or  $A(t)$ . Specifically  $A(t)$  is extracted using an envelope detector, with a simple implementation being a rectifier followed by a lowpass filter with a corner frequency equal to the bandwidth of the modulation.  $A(t)$  then drives a switching (and hence efficient) power supply that drives the **saturating power amplifier**. Polar modulation is finding application in the Universal Mobile Telecommunications System (UMTS), as the 8PSK modulation used in UMTS results in low efficiency if direct power amplification of a modulated carrier is used.

### 1.10.4 Direct Conversion Modulation

The transmitted signal has a controlled spectral content, being just a single channel. Issues such as image rejection and interferers are not problems. The overwhelming majority of wireless transmitters are based on the heterodyne principle where baseband  $I$  and  $Q$  signals are used in a vector modulator to produce a modulated IF that is upconverted in a mixer to an RF signal. The final



**Figure 1-55** Polar modulator architectures: (a) amplitude and phase modulated components amplified separately and combined; and (b) the amplitude used to modulate a power supply driving a saturating amplifier with phase modulated input.



**Figure 1-56** Architecture of a direct conversion transmitter.

RF signal is then amplified by a power amplifier. A direct conversion transmitter generates the RF signal directly without an IF stage using the architecture shown in Figure 1-56. Here, the transmit modem first produces  $I$  and  $Q$  baseband signals from the data. This is then translated directly to RF via a quadrature modulator and then amplified by a PA. Practically, the nonlinearities of the PA must be linearized using predistortion, and quadrature modulation errors must be accounted for in a quadrature modulation corrector. The transmit modem, the predistorter and the quadrature modulation corrector can be combined in a DSP, so considerable complexity is shifted to the DSP chip. One of the major problems in this architecture is the noise introduced by errors in step-size mismatch. Noise-shaping techniques implemented in a DSP have been developed to shift this noise outside the bandwidth of the generated signal. Similar errors are associated with mismatches of the ADC and of the analog circuit paths of the separate  $I$  and  $Q$

paths. This distortion is also pushed out of band by the noise-shaping algorithm.

## 1.11 Modern Receiver Architectures

Communication receivers most commonly use mixing of the RF signal with a fixed signal called an LO to produce a lower frequency replica of the modulated RF signal. Some receiver architectures use one stage of mixing, while others use two stages of mixing. In cellular systems, the receiver must be sensitive enough to detect signals of 1 pW or less. Some of the architectures used in modern receivers are shown in Figure 1-57. Figure 1-57(a) is the superheterodyne architecture in much the same form it has been used for almost a century. Key features of this architecture are that there are two stages of mixing, and filtering is required to suppress spurious mixing products. Each mixing stage has its own VCO. The receiver progressively reduces the frequency of the information bearing signal. The image rejection mixer in the dashed box achieves rejection of the image frequency to produce an IF (or baseband frequency) that can be directly sampled. It is, however, difficult to achieve the amplitude and phase balance, especially when the image reject filter is realized on an integrated circuit. Instead the architecture shown in Figure 1-57(b) is used. The filter between the two mixers can be quite large. For example, if the incoming signal is 1 GHz, the frequency of the signal after the first mixer could be 100 MHz. Filters become smaller at higher frequencies for the same performance, so the dual-conversion receiver shown in Figure 1-57(c). This is very similar to the traditional superheterodyne architecture except that the intermediate frequency between the two mixers is high. In the previous example it could be 3 GHz. The required bandpass filter can be realized as an integrated circuit. The Low-IF or Zero-IF receiver shown in Figure 1-57(d) uses less hardware and is used in less demanding communication applications.

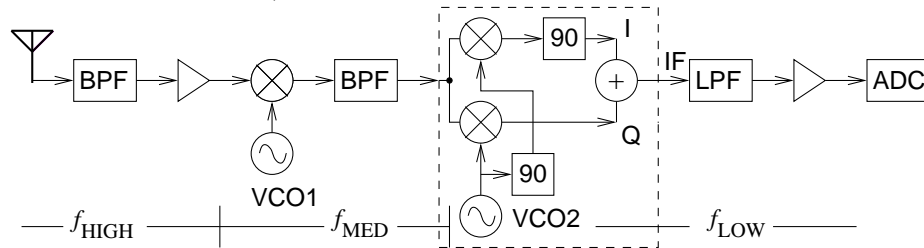
### 1.11.1 Homodyne Frequency Conversion

Homodyne mixing and detection is one of the earliest wireless receiver technologies and is used in AM radio. Homodyne mixing can be used for detecting modulation formats other than AM, including digitally modulated signals, and is a particularly attractive architecture for monolithically integrated circuits. In homodyne mixing, the carrier of a modulated signal is regenerated and synchronized in phase with the incoming carrier frequency. Mixing the carrier with the RF signal results in an IF signal centered around zero frequency. The only simple way to ensure that the pump is in phase is to transmit the carrier with the RF signal. AM transmission does just this but at the cost of transmitting large carrier power, as well as the additional prospect of interference that goes along with this. Homodyne mixing can be used with digitally modulated signals.

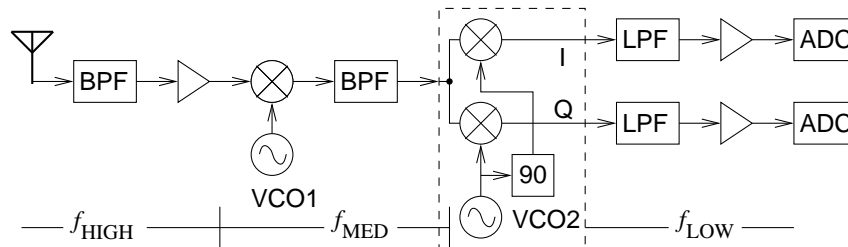
Signal spectra that result in homodyne mixing are shown in Figure 1-58. In Figure 1-58(a), the RF signals are shown on the right-hand side and the baseband signals are shown on the left hand side. It is usual to show both positive and negative frequencies at the lower frequencies so that the conversion process is more easily illustrated. The characteristic of homodyne mixing is that the pump corresponds to the carrier and is in the middle of the desired RF channel. RF



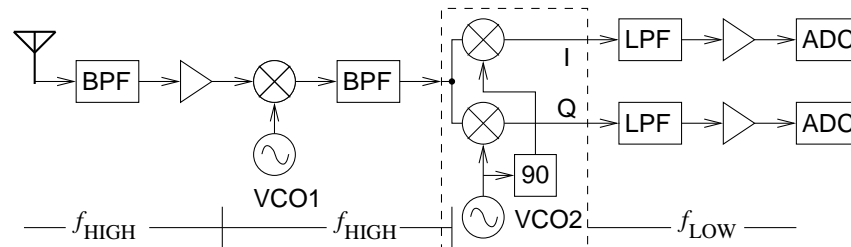
## (a) SUPERHETERODYNE, HARTLEY IMAGE REJECTION



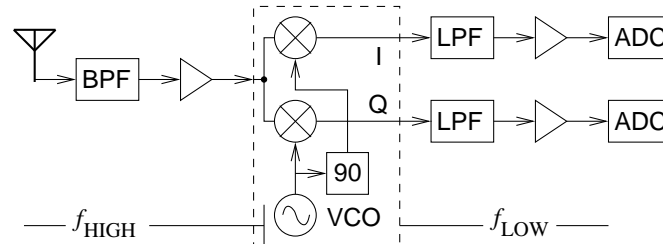
## (b) SUPERHETERODYNE RECEIVER



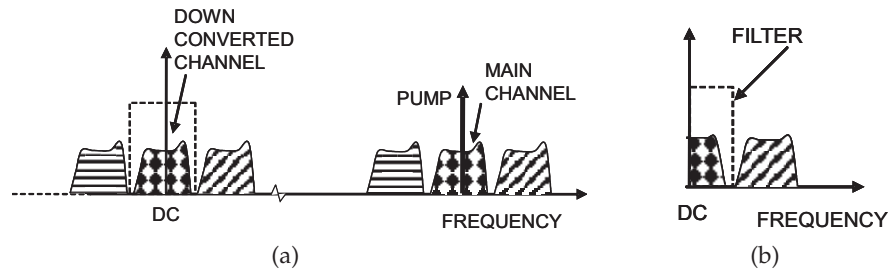
## (c) DUAL CONVERSION RECEIVER



## (d) LOW-IF OR ZERO-IF RECEIVER



**Figure 1-57** Architecture of modern receivers: (a) superheterodyne receiver using the **Hartley** architecture for image rejection; (b) superheterodyne receiver; (c) dual-conversion receiver; low-IF or zero-IF receiver. PBF, bandpass filter; LBF, LowPass Filter; ADC, analog to digital converter; VCO, voltage controlled oscillator; 90, 90° phase shifter; I, in-phase component, Q, Quadrature component;  $f_{HIGH}$ ,  $f_{MED}$  and  $f_{LOW}$  indicate relatively high, medium and low frequencies in the corresponding section of the receiver.

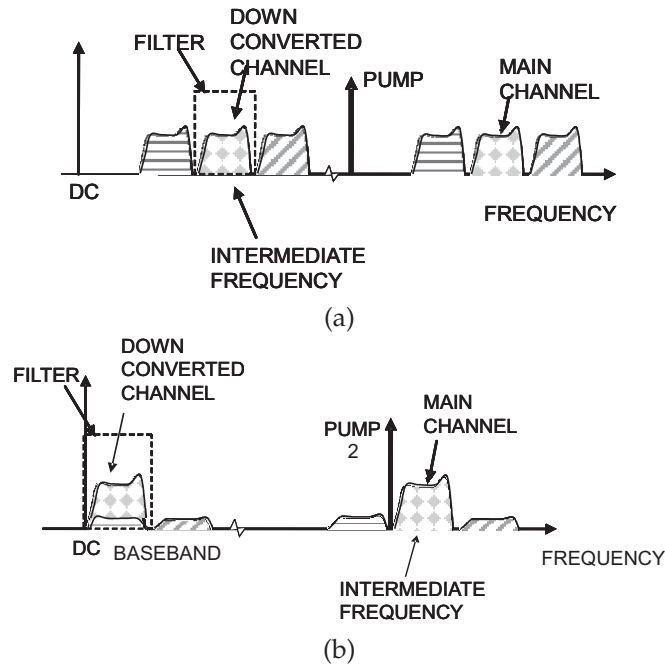


**Figure 1-58** Frequency conversion using homodyne mixing: (a) the spectrum with a large local oscillator or pump; and (b) the baseband spectrum showing only positive frequencies.

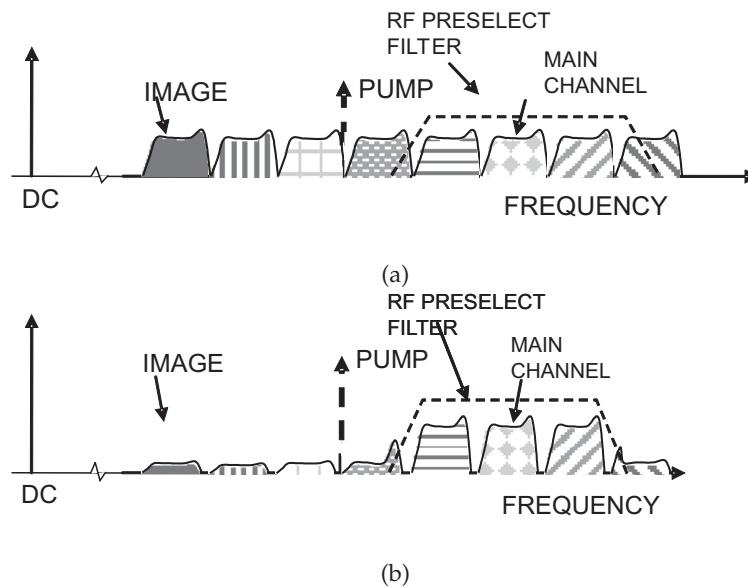
signal components mix with the pump and it appears that the entire RF spectrum is down-shifted around DC. Of course, the actual baseband spectrum (the term for the lowest frequency signals) is only defined for positive frequencies, so the negative frequency baseband signals and the positive frequency baseband signals add to yield the baseband spectrum shown in Figure 1-58. With other modulation schemes, this loss of information is avoided using quadrature demodulation discussed later in the discussion of zero IF conversion. An amplitude modulated signal has identical modulation sidebands and so the collapsing of positive and negative frequencies at baseband results in no loss of information. Then simple amplitude detection circuitry, such as a rectifier, is used and the rectified signal is passed directly to a speaker.

### 1.11.2 Heterodyne Frequency Conversion

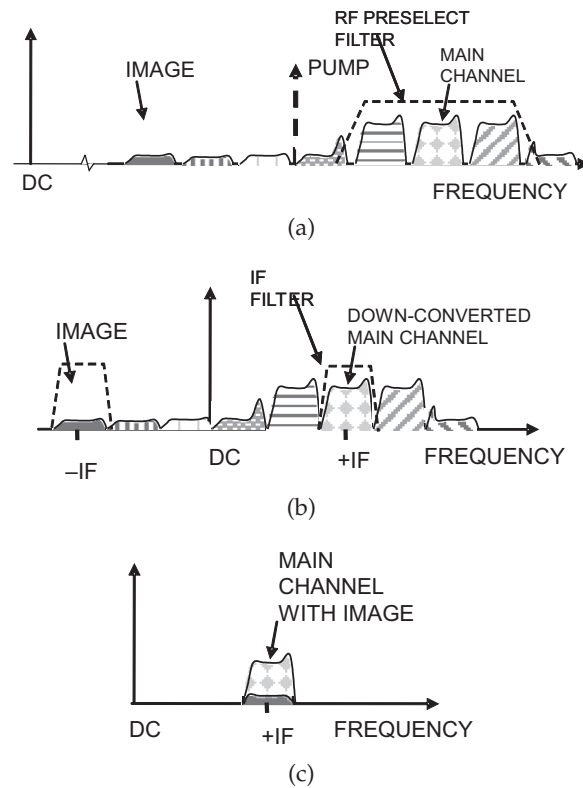
In heterodyne mixing, the pump and the main RF channel are separated in frequency, as shown in Figure 1-59. In this figure the RF signals (shown as three discrete channels on the right-hand side of the spectrum) mix with the pump signal to produce signals at a lower frequency. This lower frequency is usually not the final baseband frequency desired and so is called the IF. The intermediate frequency of the main channel is at the difference frequency of the RF signal and the pump. The pump must be locally generated and so is called the LO signal. There are several important refinements to this. The first of these is concerned with limiting the number of signals that can mix with the pump signal. This is done using an RF preselect filter, resulting in the spectrum shown in Figure 1-60(b). The important characteristic is that the signals at frequencies below the pump have been suppressed, however, the image channel is still of concern. To see the difficulties introduced by the image channel consider the frequency conversion-to-baseband process described in Figure 1-61. The RF spectrum after RF **preselect filtering** is shown in Figure 1-61(a) and the baseband (or IF) spectrum is shown in Figure 1-61(b). Again, positive and negative frequencies are used to better illustrate the down-conversion process. Note the down-converted image and the main channel are equidistant from DC. So referring the signals to a positive frequency-only spectrum, Figure 1-61(c), it can be seen that the image channel interferes with the main channel. In the worst-case scenario, the IF image could be larger than that of the desired channel. Fortunately there is a circuit fix that compensates for this.



**Figure 1-59** Frequency conversion using superheterodyne mixing: (a) the pump is offset from the main channel producing a down-converted channel at the intermediate frequency; and (b) a second pump down-converts the main channel, now at the intermediate frequency, to the baseband frequency.



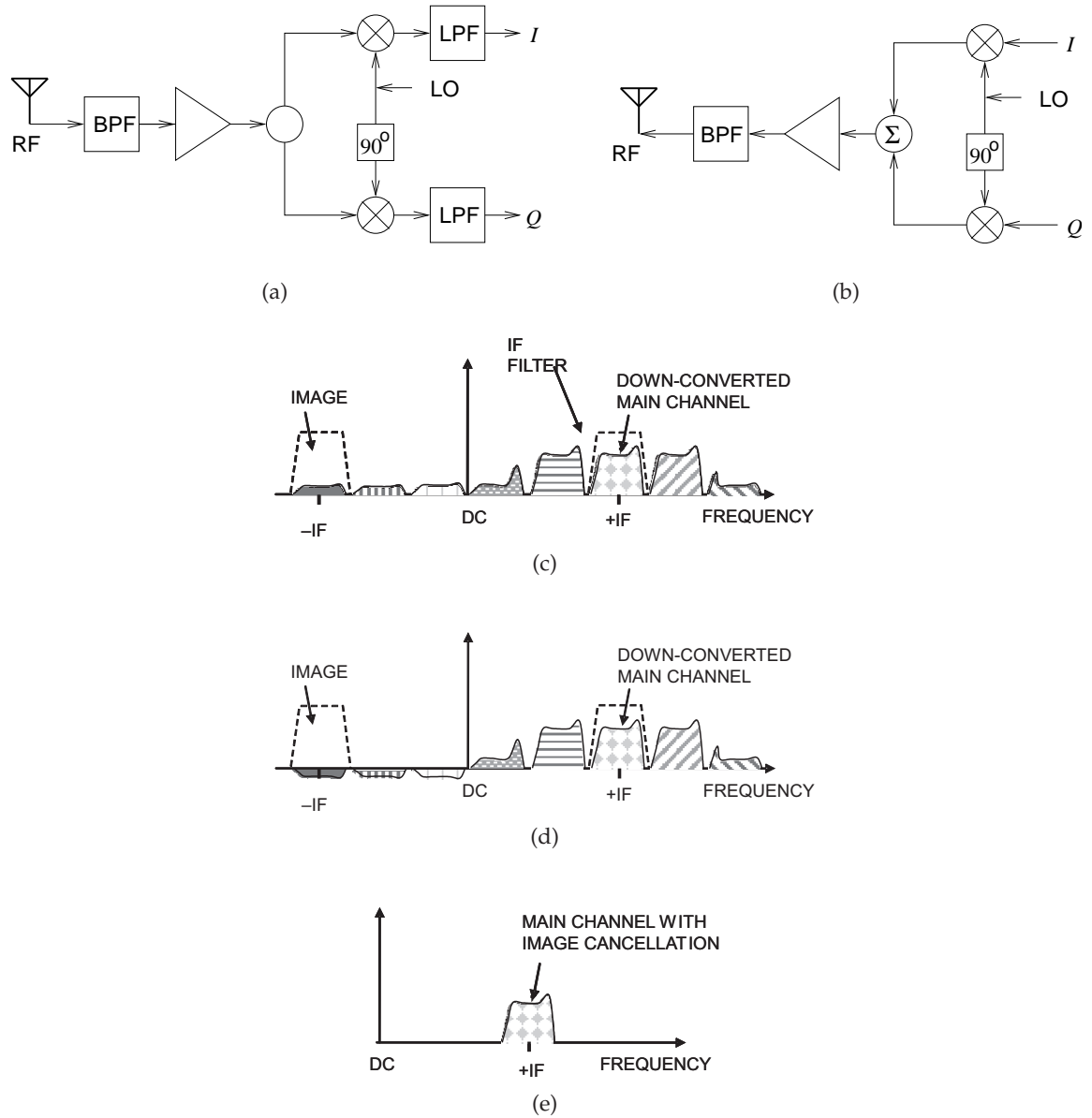
**Figure 1-60** Frequency conversion using heterodyne mixing showing the use of an RF preselect filter to reduce the image signal.



**Figure 1-61** Frequency conversion using heterodyne mixing showing the effect of image distortion: (a) the RF spectrum following filtering using an RF preselect filter; (b) the baseband down-converted signal showing positive and negative frequencies; and (c) the single-sided baseband spectrum following IF filtering showing the contamination of the final signal by the image signal.

The block diagrams of the circuits that correct this image problem are shown in 1-62(a) and Figure 1-62(b) for both transmit and receive functions. For now, consider the **quadrature receiver** shown in Figure 1-62(a). Shown here is an antenna that takes in the RF signal, and the RF preselect function is performed by a bandpass filter (BPF). This signal is initially amplified (usually), split and applied equally to two mixers. In an ideal mixer, the two signals are multiplied together. The mixers have pump (LO) signals that are  $90^\circ$  out of phase (i.e., in quadrature) with each other and contain different information. The phase relationships at the outputs of the two mixers have very particular relationships with each other such that when  $I$  and  $Q$  are added the image is eliminated, as shown in Figure 1-62(e). With digitally modulated signals, however, the  $I$  and  $Q$  waveforms are individually sampled by ADCs.

Figure 1-62(c) shows the baseband spectrum at the  $I$  output of the heterodyne receiver. Figure 1-62(d) shows the baseband spectrum at the  $Q$  output of the heterodyne receiver. It also shows the negative frequency components inverted. This is a shorthand way of saying that the negative (image) frequency components



**Figure 1-62** Quadrature mixing: (a) receive modulator; and (b) transmit modulator. Frequency conversion using heterodyne mixing and quadrature mixing; (c) the baseband spectrum at the  $I$  output of the heterodyne receiver; (d) the baseband spectrum at the  $Q$  output; and (e) the positive spectrum following the summation of the  $I$  and  $Q$  channels at the output of the heterodyne receiver.

are  $180^\circ$  out of phase with the down-converted image components of the in-phase signal. However, the down-converted main channel and neighbors are in phase in both spectra. Thus the combination of the  $I$  and  $Q$  outputs suppresses the image signals from baseband. Figure 1-62(e) illustrates the positive spectrum following the summation of the  $I$  and  $Q$  channels at the output of the heterodyne receiver. Image rejection is key to all commonly used wireless communications systems today.

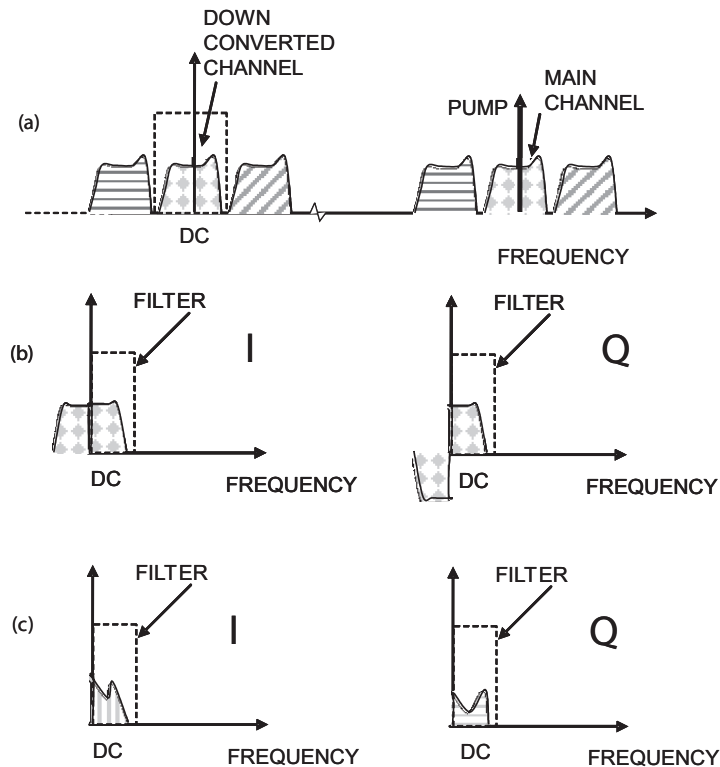
The **quadrature transmitter** shown in Figure 1-62(b) works almost identically, but in reverse. The  $I$  and  $Q$  waveforms are each applied to mixers, with one pumped by an LO and the other by the LO shifted by  $90^\circ$ . The outputs of each mixer contain upper and lower sidebands on either side of the LO (or carrier) frequency. When these are combined at the summing node, one of the sidebands (called the image) is canceled and only a single carrier is presented to the bandpass filter (BPF) and then radiated by the antenna. In QPSK digitally modulated systems, the  $I$  and  $Q$  waveforms are lowpass filtered digital bitstreams.

Heterodyne systems implement signal processing such as filtering, modulation and demodulation, and image rejection at RF and IF using hard to integrate discrete components leading to expense and limitations on size reductions. Heterodyne architectures are regarded as approaching their limit in size, integration, and fabrication cost. The primary issue in mixer design is limited image rejection resulting from gain and phase mismatches of the  $I$  (in-phase) and  $Q$  (quadrature) paths.

### 1.11.3 Direct Conversion Receiver

**Zero-IF** direct conversion receivers are very similar to quadrature homodyne receivers with an LO signal placed at the center of the RF channel. The difference is that in homodyne receivers the phase of the carrier (i.e., is the phase of the LO) is precisely known as the carrier is transmitted with the signal. In virtually all RF transmission schemes (above a few megahertz) the carrier is not transmitted. Thus in zero-IF direct conversion schemes the LO signal has inherent phase error with the original carrier. The important characteristic is that there is only one level of mixing. The conversion process is described in Figure 1-63. A particular advantage of direct conversion is that the relatively large IF filters are eliminated. Thus the  $I$  and  $Q$  mixer outputs are necessary, as the two sides of the RF spectrum contain different information, and there would be irreversible corruption if a scheme was not available to extract the information in each of the sidebands.

The main nonideality of this design is the DC offset in the down-converted spectrum. DC offset results mostly from self-mixing, or rectification, of the LO. This DC offset can be much larger than the down-converted signal itself, and because of the nonlinearities of baseband amplification stages, either severely limits the dynamic range of the receiver or places limitations on the modulation format that can be used. One way of coping with the DC offset is to highpass filter the down-converted signal, but highpass filtering requires a large passive component (e.g., a series capacitor) at least to avoid dynamic range problems with active filters. Highpass filtering the down-converted signal necessarily throws away information in the signal spectrum and it is only satisfactory to do this if



**Figure 1-63** Frequency conversion using homodyne mixing and quadrature mixing: (a) the baseband spectrum at the I output of the homodyne receiver; (b) the baseband spectrum at the Q output of the homodyne receiver; and (c) the positive spectrum following the summation of the I and Q channels at the output of the homodyne receiver.

there is very little information around DC to begin with.

The primary effort in zero-IF converters is overcoming the DC offset problem, and to a lesser extent coping with jitter of the LO. The primary LO noise of concern is close-in phase noise, which can be at appreciable levels 100 kHz from the carrier. This noise is commonly referred to as flicker noise and increases rapidly as the offset from the carrier reduces. This is of concern in all conversion processes. However, one of the properties of heterodyne mixing is that the RF signal is considerably offset from the large phase noise region. Consequently the LO phase noise at the frequency of the RF reduces the impact on the resulting offset IF signal. For these reasons, heterodyne mixing provides higher performance than direct conversion. Also, direct conversion receivers are difficult to implement for 8-state (8PSK etc.) and higher order modulation.

In cellular wireless, the radio signals are spectrally efficient and the spectrum is fairly constant across the channel. So the near-DC signals that result from direct conversion have appreciable information content and cannot be discarded so easily without significant distortion.

The main problems of zero-IF conversion in cellular radio applications are

1. Spurious LO leakage. Retransmission of the LO is possible because the LO is tuned precisely to the RF signal frequency and reverse leakage through the RF path will radiate from the antenna. Spurious LO transmission is severely regulated. The limit on this in-band LO radiation is between  $-50$  and  $-80$  dBm. The problem is reduced by using differential LOs and using multiple RF amplifier stages to increase the reverse isolation between the mixer and the antenna.
2. Interferer leakage. A large RF interferer can leak through the RF amplifier and enter the mixer through both the LO port as well as the RF port. Mixing of these components results in DC offset.
3. Distortion. Direct conversion receivers are more sensitive to undesired signals than are heterodyne receivers. The nonlinearities of the input mixers will rectify strong spurious RF signals to produce output components around DC. This is the result of second-order nonlinearities and so this effect can be suppressed through the use of balanced RF circuits which will have only odd-order nonlinearities. This reduces the baseband signals that can result from large RF interferers, but it is still possible to produce baseband distortions if the RF interferer is large enough to produce third harmonics in the mixer. The effect of this distortion can be largely eliminated through the use of highpass filtering at the baseband, but this is not acceptable for cellular radio signals. Heterodyne conversion is susceptible to distortions resulting from odd-order nonlinearities, but in zero-IF converters second order distortion is also a problem.
4. LO self-mixing. Mixing of the LO with itself will produce a DC signal in the mixer output. This DC level may be many orders of magnitude larger than the baseband signal itself, so it can desensitize or saturate the baseband amplifier. DC offsets can also result from circuit mismatch problems. The DC offset that results primarily from LO self-mixing is the most significant problem in the use of zero-IF architectures in cellular wireless. The DC offset can be reduced through the use of balanced designs, but circuit mismatch errors still result in very large DC offsets.
5. LO frequency error. A difference between the LO and the carrier will cause the RF signal to be asymmetrically converted around DC.
6. Second-order distortion. Because of second-order distortion, second harmonics of the signal can appear in the baseband. This is a problem if the RF signal is large to begin with. This problem can be circumvented by using designs that utilize differential signals.
7.  $I/Q$  mismatches. Mismatches of the  $I$  and  $Q$  paths also results in DC offsets. These offsets, however, vary negligibly with time, and analog or digital calibration techniques can be used to remove their effect.

The problem of DC offset is made worse because the DC level can vary with time as the amplitude of the interferer varies, or the LO that leaks from the antenna reflects off moving objects and is received as a time-varying interferer itself.

#### **1.11.4 Low-IF Receiver**

In a low-IF receiver, single-stage heterodyne mixing is used to down-convert the modulated RF carrier to a frequency just above DC, perhaps to a few hundred



kilohertz or a few megahertz depending on the bandwidth of the RF channel. In doing this, the DC offset problem of a direct conversion receiver is avoided. The particular advantage of a low-IF receiver is that it can be used with higher order modulation formats (8PSK and higher). It does, however, require a higher performance ADC than does a direct conversion receiver.

### *1.11.5 Subsampling Analog-to-Digital Conversion*

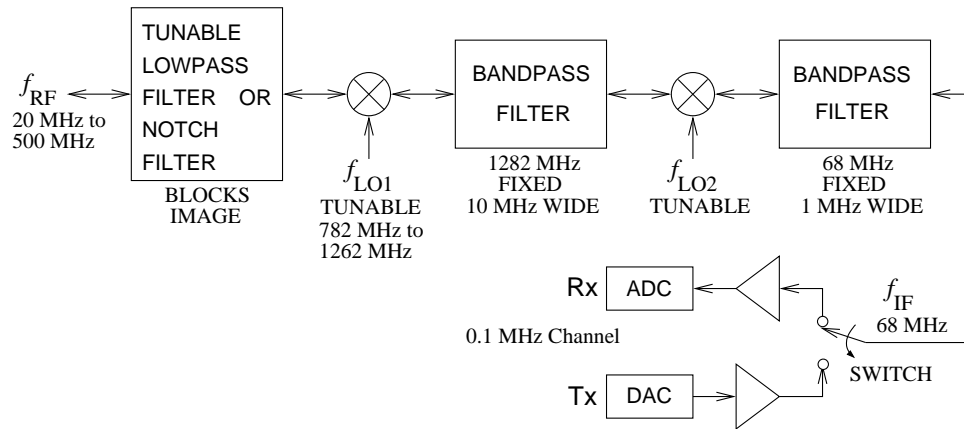
Subsampling receivers overcome the DC offset problem typical of other direct conversion receivers. The idea is to sample the modulated RF signal using an exact subharmonic of the carrier of the RF channel to be converted. The sampling rate must be at least twice the bandwidth of the RF signal and the track-mode bandwidth must be greater than the carrier frequency. Thus the sampling aperture is the critical parameter and must be several times smaller than the period of the carrier. Fortunately the aperture times of CMOS tracking circuits are adequate. It is critical that an RF preselect filter be used to eliminate unwanted interferers and noise outside the communication band. Aliasing of signals outside the **Nyquist** bandwidth onto the baseband signal is a consequence of subsampling. Adjacent channel signals are converted without aliasing, but these will lie outside the bandwidth of the baseband signal. Flicker noise on the sampling clock is multiplied by the subsampling ratio and appears as additional noise in baseband.

### *1.11.6 First IF-to-Baseband Conversion*

In a superheterodyne conversion architecture there are two heterodyne stages, with the IF of the first stage in the range of 20 to 200 MHz. The assignment of frequencies is known as frequency planning and this is treated as proprietary by the major radio vendors. This IF is then converted to a much lower IF, typically around 100 kHz or higher. This frequency is generally called baseband, but strictly it is not because the signal is still offset in frequency from DC. Some direct conversion architectures leave the first heterodyne mixing stage in place and use direct conversion of the first IF to baseband (true baseband—around DC).

### *1.11.7 Bilateral Double-Conversion Receiver*

The receivers considered so far are suitable for narrowband communications typical of point-to-point and consumer mobile radio. There are many situations where the received or transmitted RF signal covers a very wide bandwidth, such as with emergency radios, television, and military communications. If narrowband RF frontend architectures are used, a switchable filter bank would be required and this would result in an impractically large radio. Tunable bandpass filters are one option being explored. The current preferred solution to covering wide bandwidths is the double-conversion transceiver architecture shown in Figure 1-64. The frequency plan of a typical radio using 0.1 MHz channels between 20 MHz and 500 MHz is shown. The key feature of this radio is that bidirectional mixers are used, such as the **diode ring mixer** of Figure 1-52 on page 64. Following the



**Figure 1-64** Bilateral double conversion transceiver for wideband operation.

RF chain from left to right, the RF is first mixed up in frequency, bandpass filtered using a high- $Q$  distributed filter, and then down-converted to a lower frequency which can be sampled directly by an ADC. A much higher performance passive (and hence bidirectional) filter can be realized at gigahertz frequencies than at a few tens of megahertz. On transmit, the function is similar, with the mixers and LO source reused. As a receiver, the notch filter or lowpass filter is used to block the image frequency of the first mixer so that only the upper sideband IF is presented to the first bandpass filter. The lowpass or notch filter may be fixed, although with the plan shown there must be at least two states of the filters. On transmit, the lowpass or notch filters prevent the image frequency from being radiated.

## 1.12 Summary

This chapter presented the RF frontend architectures used from the beginnings of wireless communications up to those used in modern systems. Similar architectures are used in the frontends of radar and sensor systems. Wireless systems proliferate, and even in established domains such as cellphones, architectures are evolving to achieve greater efficiency, greater multifunctionality, and lower cost primarily by monolithically integrating and digitizing as much as possible of the RF frontend. Size drives the replacement of superheterodyne architecture by eliminating large intermediate filters.

## 1.13 Exercises

1. An amplifier has a power gain of 1200. What is the power gain in decibels?
2. The PAR of a signal is an important parameter in determining the efficiency that can be achieved by an amplifier with an allowable amount of distortion. The following questions are about determining the PAR of various signals. Note that the average power level is not necessarily the average of the

minimum and maximum power levels. A full power calculation and integral should be performed. [WS3]

- (a) Write down a formula for the power of a signal  $x(t)$ . You can consider  $x(t)$  to be a voltage across a  $1\ \Omega$  resistor.
  - (b) What is the PAR of an FM signal at 1 GHz with a maximum modulated frequency deviation of  $\pm 10$  kHz?
  - (c) What is the PAR of a two-tone signal (consisting of two sinewaves at different frequencies that are say 1% apart)? First, use a symbolic expression and then consider the special case when the two amplitudes are equal. Consider that the two tones are close in frequency.
  - (d) What is the PAR of a three-tone signal (consisting of three sinewaves say 1% apart in frequency) when the amplitude of each sinewave is the same?
  - (e) What is the PAR of an AM signal with 75% amplitude modulation?
3. Short answer questions on gain calculations.
- (a) An amplifier with  $50\ \Omega$  input impedance and  $50\ \Omega$  load impedance has a voltage gain of 100. What is the gain in decibels?
  - (b) An attenuator reduces the power level of a signal by 75%. What is the gain of the attenuator in decibels?
  - (c) What is the wavelength in free space of a signal at 4.5 GHz?
4. An amplifier consists of three cascaded stages with the following characteristics"

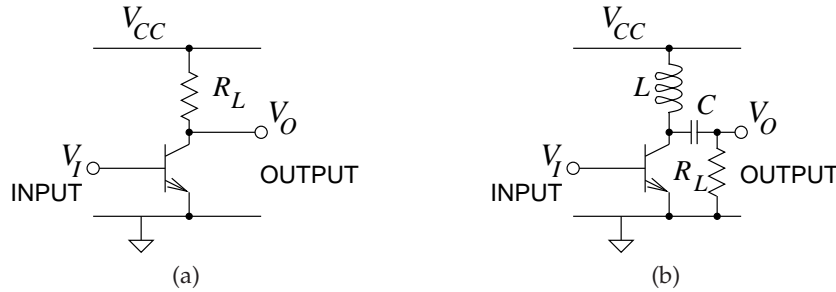
Characteristic	Stage 1	Stage 2	Stage 3
Gain	10 dB	15 dB	30 dB
NF	0.8 dB	2 dB	2 dB

What is the noise figure (NF) and gain of the cascaded amplifier?

5. The frontend of a receiver for a cellular phone has a bandpass filter with a 25 MHz passband and loss of 2 dB and is followed by two amplifier stages. The first stage has a gain of 20 dB and a noise figure of 0.5 dB and the second stage has a gain of 60 dB and a noise figure of 2 dB.
- (a) Sketch the block diagram of the system as described.
  - (b) What is the gain of the system?
  - (c) What is the noise figure of the bandpass filter?
  - (d) What is the noise figure of the system?
  - (e) The system is now connected to an antenna with an effective noise temperature of 30 K and which delivers a signal of 10 pW to the bandpass filter. Determine the noise temperature at the output of the system and hence the noise power in the 25 MHz bandwidth. Determine the signal-to-noise ratio at the output of the frontend system.
6. An FM signal has a maximum frequency deviation of 20 kHz and a modulating signal between 300 Hz and 5 kHz. What is the bandwidth required to transmit the modulated RF signal when the carrier is 200 MHz? Is this considered to be narrowband FM or wideband FM?
7. Consider two analog signals combined together. One signal is denoted  $x(t)$  and the other  $y(t)$  where  $x(t) = 0.1 \sin(10^9 t)$  and  $y(t) = 0.05 \sin(1.01 \cdot 10^9 t)$ .

What is the PAR of this combined signal? Express PAR in decibels.

8. A high-fidelity stereo audio signal has frequency content ranging from 50 Hz to 20 kHz. If the signal is to be modulated on an FM carrier at 100 MHz, what is the bandwidth required for the modulated RF signal? The maximum frequency deviation is 5 kHz when the modulating signal is at its peak value.
9. The following sequence of bits 0100110111 is to be transmitted using QPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the  $I$  and  $Q$  channels. Show the transitions on a constellation diagram.
10. The following sequence of bits 0100110111 is to be transmitted using OQPSK modulation. Take these data in pairs, that is as 01 00 11 01 11. These pairs, one bit at a time, drive the  $I$  and  $Q$  channels. Show the transitions on a constellation diagram. [WS24]
11. The following sequence of bits 0100110111 is to be transmitted using  $\pi/4$ -DQPSK modulation. Take these data in pairs, that is, as 01 00 11 01 11. These pairs, one bit at a time, drive the  $I$  and  $Q$  channels. Use four constellation diagrams, with each diagram showing one transition. On each diagram, indicate the initial symbols and the next symbols (rotated of course by  $45^\circ$ ). [WS25]
12. The cascaded two-port network in Figure 1-45 consists of a filter with an insertion loss of 3 dB followed by an amplifier with a noise figure of 2 dB and a gain of 20 dB. Calculate the total gain and noise figure. This problem parallels example 1.4 on page 237.
13. A Class A HBT amplifier is biased with a collector emitter quiescent voltage of 5 V and a quiescent collector-emitter current of 100 mA. When operated at the 1 dB compression point, the input RF power is 10 mW and the output power is 100 mW.
  - (a) What is the quiescent DC power consumed? Express your answer in milliwatts.
  - (b) What is the output power in dBm?
  - (c) What is the efficiency of the amplifier? Note that the efficiency of a class A amplifier can be more than 25% if distortion can be tolerated.
  - (d) What is the power-added efficiency of the amplifier?
  - (e) If the input power is reduced by 10 dB so that the amplifier is no longer in compression, will the DC quiescent point change? Explain your answer.
  - (f) If the input power is reduced 10 dB so that the amplifier is no longer in compression, what is the output power in dBm? Ignore any change in the quiescent point.
  - (g) With 1 mW input power, what is the power-added efficiency of the amplifier if the quiescent point does not change?
14. The BJT amplifier in Figure 1-65(a) has a load,  $R_L$ , and a maximum undistorted efficiency of 25%. Derive this efficiency.



**Figure 1-65** Class A HBT amplifier with load  $R_L$ : (a) with the collector bias also provided through  $R_L$ ; and (b) collector supplied through an inductor  $L$  which is an open circuit at AC frequencies.

15. The BJT amplifier in Figure 1-65(b) has an RF choke providing collector current and acting as an open circuit at RF. The load,  $R_L$ , is driven through a capacitor  $C$  which is effectively a short circuit at RF. The maximum undistorted efficiency of this circuit is 50%. Derive this efficiency. Ignore the base-emitter voltage drop,  $V_{CE, \min}$ , and note that the maximum of  $V_O$  is  $V_{CC}$ , allowing a voltage swing of  $\pm V_{CC}$  around the collector quiescent operating voltage.
16. A  $75 \Omega$  attenuator has a loss of 16 dB and is between a source with a Thevenin impedance of  $75 \Omega$  and a load of  $75 \Omega$ .
  - (a) What is the noise power,  $N_i$ , available from the  $75 \Omega$  source resistor at standard temperature (270 K) in a 1 MHz bandwidth?
  - (b) Now consider that the attenuator is connected to the attenuator which is also connected to the load. If the source generates a modulated signal that is 1 MHz wide and has an available power,  $S_i$ , of 1 fW, what is  $\text{SNR}_i$  at the input to the attenuator at standard temperature?
  - (c) With the attenuator connected to the source, what is the Thevenin equivalent impedance looking into the output of the attenuator?
  - (d) Calculate the noise power  $N_o$  available from the attenuator with the source attached at standard temperature (270 K) in a 1 MHz bandwidth?
  - (e) What is the signal power,  $S_o$ , delivered to the load?
  - (f) What is the SNR at the load,  $\text{SNR}_o$ ?
  - (g) What is the noise factor,  $F$ , of the attenuator?
  - (h) What is the noise figure, NF, of the attenuator?
17. A superheterodyne receiver has in order an antenna, a low-noise amplifier, a bandpass filter, a mixer, a second bandpass filter, a second mixer, a lowpass filter, an ADC, and a DSP which will implement quadrature demodulation. Develop the frequency plan of the receiver if the input RF signal is at 2 GHz and has a 200 kHz single-channel bandwidth. The final signal applied to the ADC must be between DC and 400 kHz so that  $I/Q$  demodulation can be done in the DSP unit. Noise considerations mandate that the LO of the first mixer must be at least 10 MHz from the input RF. Also, for a minimum size bandpass filter between the two mixers, the filter should be at as high a frequency as possible and it has been determined that a 100 MHz filter is available at an acceptable cost, so it has been decided that it will be used.

- (a) Draw a block diagram of the receiver and annotate it with symbols for the frequencies of the LOs and the RF and IF signals.
  - (b) What is the LO frequency  $f_{LO1}$  of the first mixer?
  - (c) What is the LO frequency  $f_{LO2}$  of the second mixer?
  - (d) Specify the cutoff frequency of the lowpass filter following the second mixer.
  - (e) Briefly discuss in less than half-page other design considerations as they relate to the frequency plan, filter size, and filter specification.
18. Short answer questions. Each part requires a short paragraph of about five lines and a figure where appropriate to illustrate your understanding.
- (a) Consider a two-tone signal and describe intermodulation distortion.
  - (b) Describe the effect of a lossy filter on the SNR. Consider signals at the input and output of the filter?
  - (c) What is meant by 1 dB gain compression?
  - (d) Consider a digitally modulated signal and describe the impact of a nonlinear amplifier on the signal. You must include several (at least) negative effects.
19. Short answer questions. Each part requires a short paragraph of about five lines and a figure where appropriate to illustrate your understanding.
- (a) Explain the operation of a superheterodyne receiver?
  - (b) Compare zero-IF and low-IF receivers.