Modulation for non-linear systems

3.1	The effect of non-linear systems	51
3.2	Linear modulation schemes for non-linear channels	55
3.3	Non-linear modulation	63

The object of this chapter is primarily to describe the class of non-linear, or exponential, modulation schemes. However, we also consider the reasons why these schemes are important, namely for use with non-linear systems such as RF power amplifiers, and accordingly we also describe the effects of these systems on linear modulation. This also leads to a study of methods for the amelioration of non-linear effects in linear modulation, and a comparison of linear and non-linear schemes. Since non-linear amplifiers are most often used where power consumption is of primary concern, and particularly in personal and cordless communication systems, the schemes described in this chapter include those that are most important in a wide range of current mobile and cordless radio standards, including GSM, DECT, the TETRA private mobile (or business) radio (PMR) standard, the North American standards IS-54 and IS-136 and PACS-UB [3.1, ch. 10], and many others.

3.1 The effect of non-linear systems

In most power-limited radio systems, the important issue is to optimize the efficiency of the RF power amplifier (commonly known as the 'high power amplifier', or HPA). This is either to minimize overall power consumption in a battery operated transmitter such as a mobile handset, or to make best use of the investment made in the HPA. This latter is especially true of satellite systems, where it is desirable to operate the HPA as near saturation as possible.

RF power amplifiers [3.2, 3.3] are classified according to the proportion of the carrier cycle for which the output device conducts, which determines both power efficiency and linearity. The most linear (and much the least efficient) are *class A*, in which the device conducts throughout the cycle. In *class B* conduction is for nominally half the cycle, which is much more efficient,

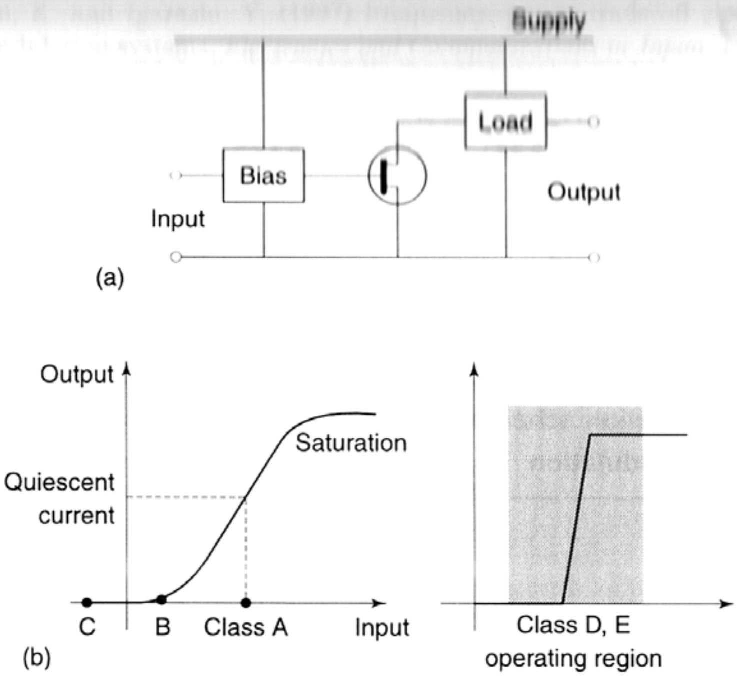

Figure 3.1 HPA classes (a) structure; (b) operating points

since it avoids large quiescent currents. ('Class AB' audio amplifiers usually have two devices in the familiar 'push–pull' configuration, each of which conducts for slightly more than half the time.) *Class C* denotes conduction for less than half the cycle. *Classes D* and *E* use the output device only as a switch, and as such may approach 100% efficiency. Figure 3.1(a) shows the general structure of these circuits, while Fig. 3.1(b) shows the device characteristics with the operating points or regions of the different configurations. Note that even the most linear class A amplifier is subject to saturation around its maximum output power.

The effect will be to distort the carrier sine wave, as shown in Fig. 3.2. Note that the amplifier will in nearly all cases be followed by a band-pass filter, which will pass only frequencies close to the fundamental. Thus the harmonic distortion that results is not in itself a severe problem. However, it will result in an amplitude-dependent phase and amplitude distortion, as shown in Fig. 3.3.

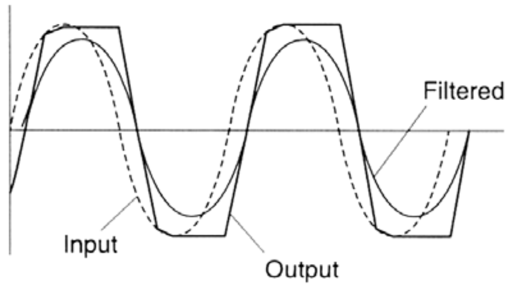

Figure 3.2 Amplitude-phase distortion in non-linear HPA: waveforms

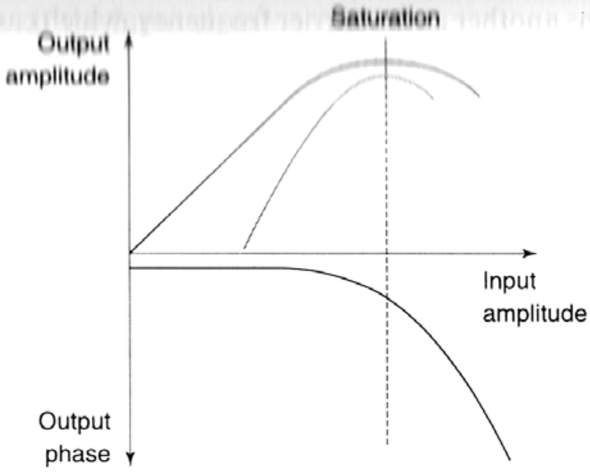

Figure 3.3 Amplitude-phase distortion in non-linear HPA: amplitude-phase characteristic

An amplitude-dependent phase delay due, for example, to the switching delay of saturated semiconductor elements, will result in the amplitude-dependent phase shift. For class B or higher, the characteristic will be more severe still, similar to the grey line shown in Fig. 3.3, and will virtually preclude any operation in the linear part. Even for more linear amplifiers, however, Fig. 3.3 shows that any amplitude variation on the input will result in unwanted modulation of both amplitude and phase of the output. These effects are known as *AM–AM* and *AM–PM conversions*. Note that if the input amplitude is constant neither of these occurs, and there is effectively no distortion.

We may express the transfer function $y(x)$ in polynomial notation:

$$y = \alpha x + \beta x^2 + \gamma x^3 + \cdots \tag{3.1}$$

Using the complex baseband representation of the modulated signal from Equation (2.8):

$$\left.\begin{aligned} y(t) = {} & \frac{\alpha}{2}[b(t)\exp(j\omega_c t) + b^*(t)\exp(-j\omega_c t)] \\ & + \frac{\beta}{4}[b^2(t)\exp(2j\omega_c t) + 2b(t)b^*(t) + b^{*2}(t)\exp(-2j\omega_c t)] \\ & + \frac{\gamma}{8}[b^3(t)\exp(3j\omega_c t) + 3b^2(t)b^*(t)\exp(j\omega_c t) \\ & + 3b(t)b^{*2}(t)\exp(-j\omega_c t) + b^{*3}(t)\exp(-3j\omega_c t)] \\ & + \cdots \end{aligned}\right\} \tag{3.2}$$

Of these terms, the first-order term (in α) is of course undistorted. The second order (in β) has a term at twice the carrier frequency and another at baseband, neither of which will be passed by the band-pass filter. The third-order term contains a term at three times the carrier frequency, which will also be rejected, but it also contains another at the carrier frequency, which cannot be separated from the modulated signal. Expanding this term:

$$\frac{\gamma}{8}[3b^2(t)b^*(t)\exp(j\omega_c t) + 3b(t)b^{*2}(t)\exp(-j\omega_c t)]$$

$$= \frac{3\gamma[b_p^2(t) + b_q^2(t)]}{8}[b(t)\exp(j\omega_c t) + b^*(t)\exp(-j\omega_c t)] = \frac{3\gamma}{4}A^2(t)a(t) \tag{3.3}$$

In general, the even-order terms in the transfer function will not generate interfering products at the carrier frequency, while odd-order terms will. However, in most cases it is the third-order term that will dominate. From Equation (3.3) we note that if the amplitude $A(t)$ is constant, then there is no distortion, which agrees with the conclusion reached from our discussion of Fig. 3.3.

We may often characterize the behaviour of a non-linear system in terms of the third-order term, by considering the third-order intermodulation product of two closely spaced sine wave signals within the pass-band. For example two signals, one at ω_c and another at $\omega_c + \omega_d$, generate an intermodulation product at $\omega_c + 2\omega_d$. If the amplitude of the signal at ω_c is held constant, while that at $\omega_c + \omega_d$ is increased, we will observe a linear increase in the output term at $\omega_c + \omega_d$ (dominated by the first-order term of the transfer function), while the term at $\omega_c + 2\omega_d$ starts much smaller, but increases much more rapidly. Figure 3.4 shows a logarithmic plot of both terms. At low amplitudes (in the linear region of the characteristic), both plots are linear, although the third-order term tends to be approximately twice as steep, showing that it varies as the square of the input. At higher amplitudes both terms cease to be linear, but if we extrapolate the linear portions of the plots, the point at which they cross is called the *third-order intercept* (TOI) [3.3, p. 52; 3.4, p. 186], and gives a gauge of the non-linearity of the amplifier, which can readily be measured in practice.

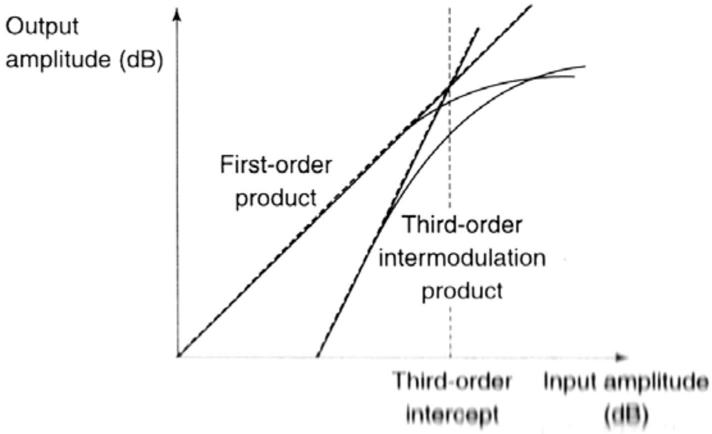

Figure 3.4 The third-order intercept

$$x(t) = \cos(\omega_c t) + a\cos[(\omega_c + \omega_d)t] \tag{3.4}$$

then the resultant (neglecting all but the first- and third-order terms) is:

$$\begin{aligned}
y(t) &= \alpha x + \gamma x^3 \\
&= \left(\alpha + \frac{3\gamma}{4} + \frac{3a^2\gamma}{2}\right)\cos(\omega_c t) + \left(a\alpha + \frac{3a\gamma}{2} + \frac{3a^3\gamma}{4}\right)\cos[(\omega_c + \omega_d)t] \\
&\quad + \frac{3a^2\gamma}{4}\cos[(\omega_c + 2\omega_d)t] + \frac{3a\gamma}{4}\cos[(\omega_c - \omega_d)t] + \frac{3a\gamma}{4}\cos[(3\omega_c + \omega_d)t] \\
&\quad + \frac{3a^2\gamma}{4}\cos[(3\omega_c + 2\omega_d)t] + \frac{a^3\gamma}{4}\cos[(3\omega_c + 3\omega_d)t]
\end{aligned} \tag{3.5}$$

Of these, the last three terms are out-of-band, and the first two are the terms at the two input frequencies. The third is the third-order term of interest to us. Observe that it varies with the square of the input amplitude a. The second is the linear term in Fig. 3.4, whose magnitude is dominated by the term $a\alpha$. The third-order intercept thus occurs where:

$$\frac{3a^2\gamma}{4} = a\alpha \Rightarrow a = \frac{4\alpha}{3\gamma} \tag{3.6}$$

where a is the relative amplitude of the second input signal.

A non-linear system also distorts the spectrum of a signal with time-varying amplitude. In general, the effect on a signal that has been filtered to reduce its bandwidth, for example by means of Nyquist filtering, is to regenerate the sidelobes, which had been eliminated. The effect is called *spectral regrowth*. Manipulating Equation (3.3) in the same way as (2.9), we obtain:

$$\begin{aligned}
&A(\omega) = \tfrac{1}{2}[C(\omega - \omega_c) + C^*(-\omega + \omega_c)] \\
\text{where}\quad &C(\omega) = \mathbf{F}[A^2(t)b(t)] = \mathbf{F}[A^2(t)] * B(\omega)
\end{aligned} \tag{3.7}$$

which shows that the output equivalent baseband spectrum is the convolution of the input baseband spectrum with the spectrum of the squared magnitude of the signal. Again, we note that if there is no amplitude variation, there is no distortion, and no spectral regrowth. It also shows that the smaller the variation in amplitude, the smaller the spectral regrowth, and therefore that it may be worthwhile minimizing amplitude variation, if constant amplitude is not feasible.

3.2 Linear modulation schemes for non-linear channels

In this section we describe techniques that may be applied to linear modulation to improve its performance on a channel subject to some degree of