Amplifier Dynamic Range Requirements for Digital Communications Systems

Daniel M. Dobkin, Director, Technical Marketing
Walter Strifler, Staff Scientist
Gleb Klimovitch, Senior Member of the Technical Staff
Greg Fung, Senior Design Engineer

Introduction

In modern data communications systems, system designers are often concerned with measurements of nonlinearity and distortion. Such quantities as input or output third-order intercept points, spurious-free dynamic range, or composite second order distortion, composite triple beat distortion, and cross-modulation, become critical specifications. At first glance, this is a surprising development. A digital data stream is fundamentally a string of 1's and 0's, and one would at first assume that transmission of digital data would place very few demands on system linearity beyond the normal variation in signal strength encountered in radio design. This tutorial will examine why dynamic range plays an important role in digital communications networks, and what steps can be taken to obtain high dynamic range in a cost-effective fashion.

Digital systems are not just binary

Binary data: what range?

The purpose of a digital transmission is to move a sequence of 1's and 0's from here to there. Accurate representation of intermediate states – that is, the linearity of the system – seems completely irrelevant. Why then is linearity important in practical communications systems?

The importance of linearity derives from the limited slice of frequency spectrum allocated to any particular signal in wireless communications. Nonlinear systems transmit distorted signals. The distortions are equivalent to changes in the Fourier transform of the signal; that is, unintended frequencies are radiated which may interfere with neighboring channels. Thus radio designers need to avoid operating at amplitudes at which such distortion is significant.

Accomplishing this design objective would seem very simple if a stream of 1's and 0's is all that is to be transmitted, but practical systems are more complex because of the large ***peak-to-average ratio*** of many digital signals. Communications systems must be designed for minimum distortion not just at amplitude levels which are typical of the signal, but also for rare excursions to much higher voltage or power.

In the remainder of this tutorial we'll examine how this situation comes about, and how a system designer can solve the problems that result through appropriate design procedures and the use of semiconductor components intended for these applications.

 ©2001 WJ Communications Inc.

Spectral Efficiency and Filtering

Wireless transmissions must fit into a specific frequency band, typically allocated by a regulatory agency. As the reader will recall from basic Fourier transform theory, the spectrum of an ideal rectangular pulse (i.e. an isolated "1") has spectral components out to infinitely high frequency. If this baseband data is mixed up to the carrier frequency, one would have to reduce the bit rate of the signal to an intolerably low rate to avoid having the sidebands of the carrier extend past the allowed bandwidth.

The conventional approach to limit signal bandwidths within allocated channels is to filter the waveform and tolerate less abrupt transitions of the signal state. If one arbitrarily filters the spectrum of a datastream, the resulting time domain waveform may suffer from interference between neighboring pulses, since the rise- and fall-times of the pulses may have become comparable to the separation between bits. Nyquist [1] showed the optimal way of filtering a pulse stream results in waveforms in the time domain with the very convenient property that each pulse, though it may be much wider than a single bit time, goes to zero at ***every*** neighboring sampling time. At evenly spaced sample times, only the voltage from the current bit is being sampled, even though at other times considerable interference from other bits is present. Figure 1 shows an example of original discrete data and the resulting filtered data: note that the filtered data intersects the discrete data at each sampling time, though there is considerable disagreement between the two at other times.

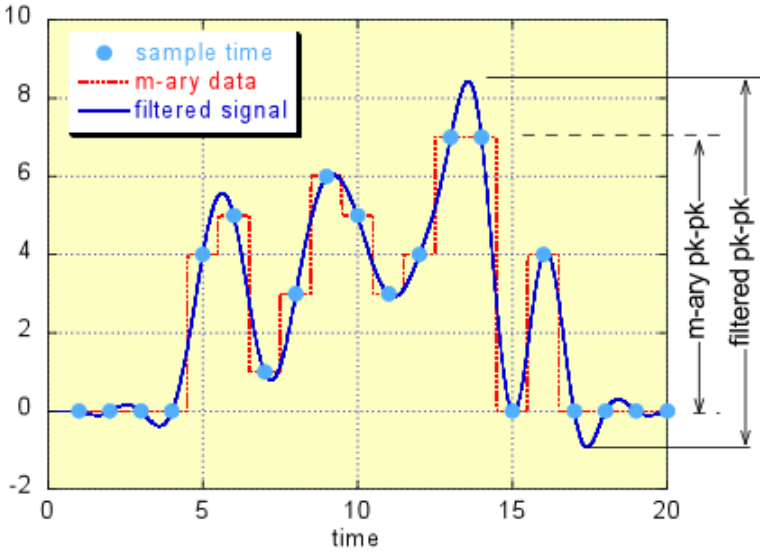


Figure 1: Comparison of ideal "m-ary" data and Nyquist-filtered version of the same data, using less bandwidth. [A raised-cosine filter with rolloff of 0.3 was used.]

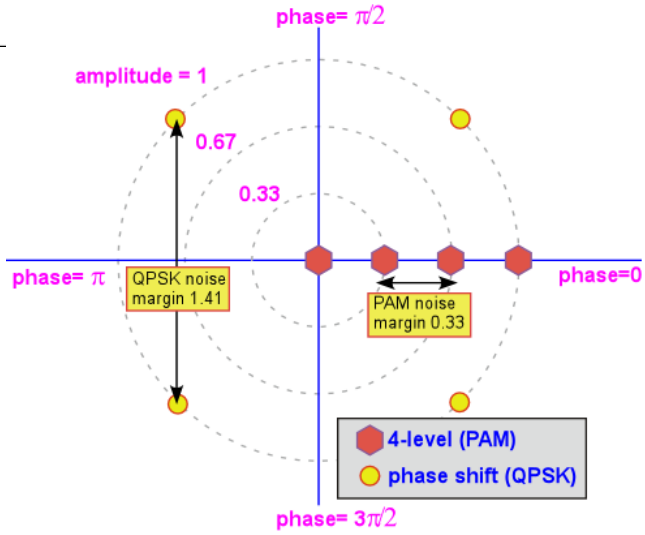
Such filtering serves to maximize the data rate for a given bandwidth, but at the cost of distortion of the waveform in the time domain. Bandwidth-limited distortion leads to excursions of the time-domain waveform past the nominal data levels; the peak-to-average ratio of the signal is increased beyond that of the unfiltered signal. Filtering the datastream leads to an increase of around 30-40% in the peak amplitude, depending on the exact filter used.

Spectrally Efficient Modulations

Transmission of a 1 or 0 in a given time slot is equivalent to sending one bit per sample period. One can increase the rate of data transmission in bits per second by decreasing the sample period. However,even when the data has been filtered as discussed above, the spectral width of the data grows larger, since the filter width (roughly B/2 where B is the bit rate) is inversely proportional to the sample period.

 ©2001 WJ Communications Inc.

An alternative solution is to use the same rate B, but send more information in each sample time. It is useful to denote the signal that is sent in a given sample time as a ***symbol***. Then if a symbol can be made to represent more than one bit, the bit rate may be increased without decreasing the sample time. One might use several differing amplitudes of the signal: a symbol which can take four amplitude levels transmits 2 bits. However, the noise immunity of such a multilevel approach is inferior to that obtained by exploiting the ***phase*** of the carrier. Figure 2 shows a ***signal constellation*** (depiction of the phase and amplitude of the desired levels) for both the 4-amplitude ***pulse-amplitude modulated*** (PAM) signal and an equivalent signal exploiting changes in the phase: ***quaternary phase shift keying*** **(QPSK)**. The QPSK signal points are spaced farther from each other than the PAM signal points for the same number of bits per symbol, and thus have better immunity to noise [2].



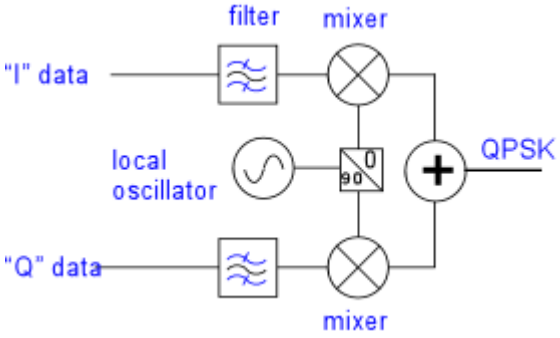


Figure 3: I/Q Modulator for QPSK and other QAM modulations

What happens to this simplistic picture when we include the transition from one symbol to another? In practice, QPSK is often implemented by combining two binary bit streams, one modulating the carrier and the second modulating a 90-degree-phase-shifted version of the carrier: these are the ***in-phase (I)*** and ***quadrature*** **(Q)** parts of the signal (Figure 3).

Just as in the case of a simple one-bit symbol, the incoming bit streams are filtered to make efficient use of the allocated spectrum.. The resulting signal does not in general follow a simple path from one constellation point to the next, but a complex data-dependent gyration. An example is shown in Figure 4.

This complex path gives rise to significantly larger peak values of voltage and signal power than the nominal values associated with the constellation points. For the particular trajectories shown in Figure 4, the peak instantaneous power is about 2.7 times larger than the average value, or 4.3 dB in logarithmic terms.

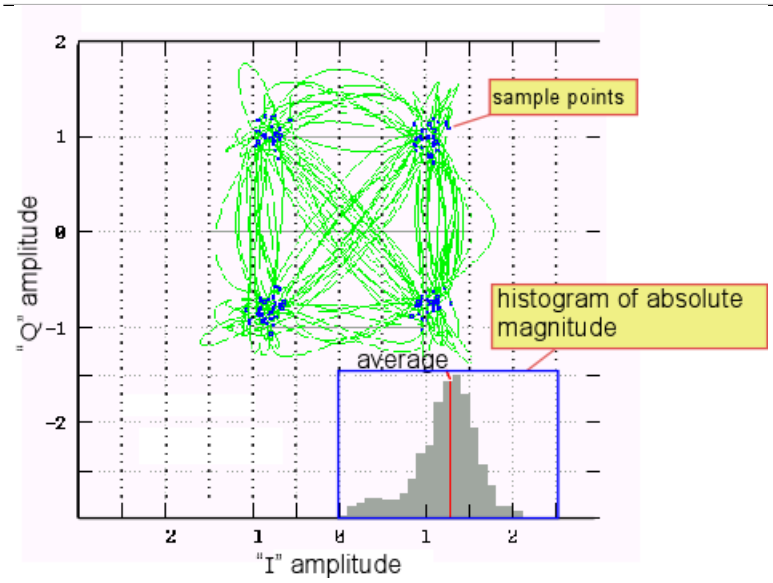


Figure 4: Trajectory in phase-amplitude space for a filtered QPSK signal corresponding to 256 bits of pseudo-random data; inset contains histogram of instantaneous magnitudes (16 points per data bit)

Multilevel Signals

In addition to the challenges of filtered complex constellation paths, standard communications protocols often involve the transmission of more than one signal simultaneously, especially in the "downstream" direction, from a basestation to a number of handset users. Multiple simultaneous signals inevitably lead to rare peak levels much higher than the average signal power.

An example for binary signals is shown in Figure 5: the addition of multiple uncorrelated bit streams produces a final signal which can have many possible levels. The probability of each "voltage" level is proportional to the number of ways in which it can result. For example, in the case in which two bilevel signals are combined, there is only one way to make a level of +2 and only one way to make a level of –2, but there are two ways to make a level of 0. Thus, a zero is twice as likely to occur as either of the extreme cases. In the general case, if we assume equal likelihoods of 1 or 0 in the incoming data streams, the probability of obtaining a level m from n streams is described by the binomial coefficient:

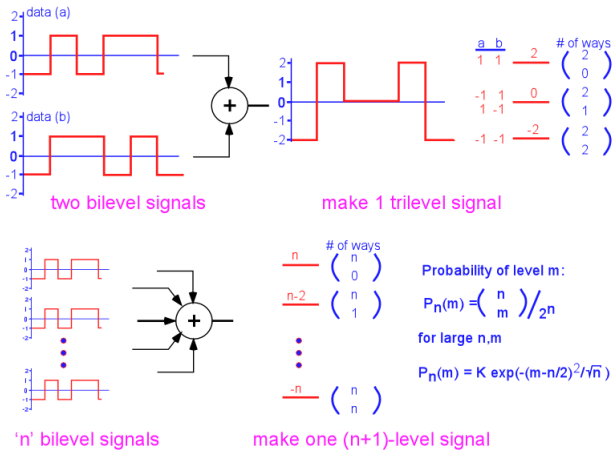


Figure 5: The sum of a large number of binary signals gives rise to a normal distribution.

$$\binom{n}{m} = \frac{n!}{m!\,(n-m)!}$$

As the number of signals n grows large, the distribution of voltages closely approximates a normal or Gaussian distribution, with standard deviation of √N / 2. The distribution of signal power is a chi-squared distribution of order 1 for the case where each signal is either on or off, or a Rayleigh distribution in the case where there are two orthogonal components ("I" and "Q") combining to form the final signal. The corresponding peak to average ratios, where the peak is defined to be a signal with probability of 10^{-5}, are 13.8 and 11.4 dB, respectively.

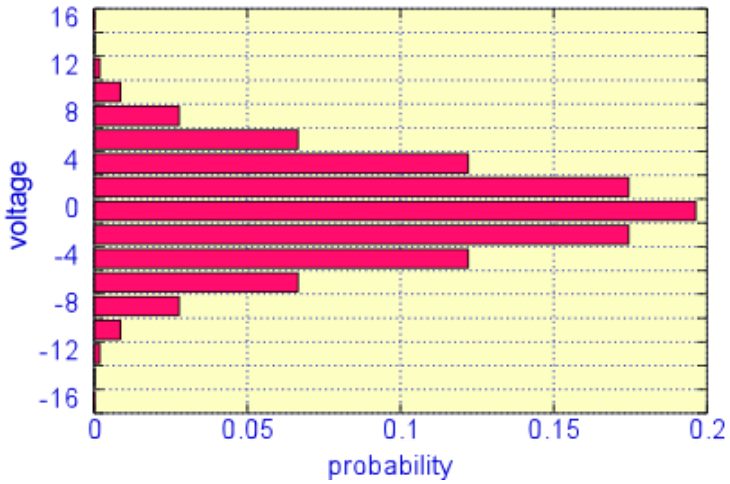


Figure 6: Probability of positive voltage values for the sum of 16 uncorrelated (+1/-1) streams

Consider a finite example of 16 combined binary streams, each of which can take a value of either +1 or –1 volt (chosen to make the average value = 0). The possible levels vary from +16 to –16 volts in steps of 2. The probability distribution for obtaining each possible voltage level is shown in Figure 6.

Note that the probability of the extreme values is so low that no bar is shown on the graph; the actual value is (1/65,536) = 1.5×10^{-5} for +16 or –16. Each voltage value is associated with a power dissipation; for simplicity, assume that the voltages modulate the level of a carrier signal in a 50 Ω system, so that the power is $V^2/50$. We can then obtain a similar distribution for the probability of a given output power level, shown in Figure 7

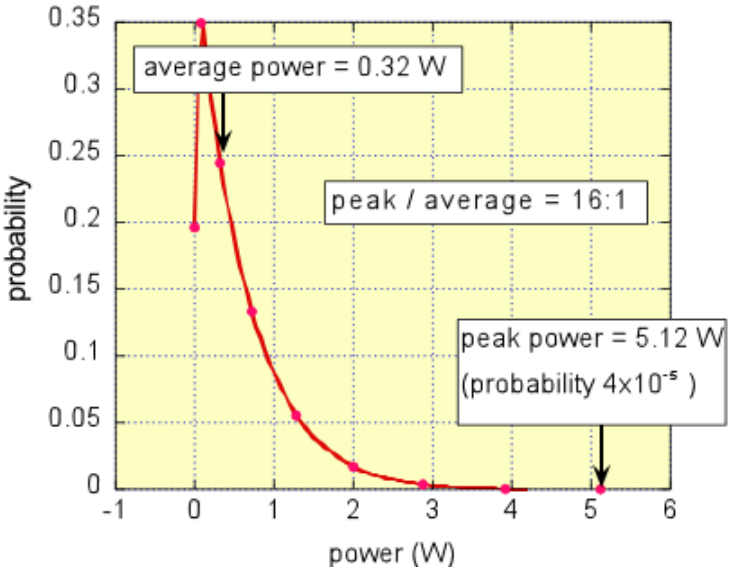


Figure 7: Probability of a given power level for the 16 combined binary streams of figure 6.

The peak power occurs when all the signals are either +1 or –1: i.e. $(16^2)/50 = 5.12$ W. However, the average power is much lower, because in the vast majority of cases the signals oppose each other and average out to a small transmitted power. The ratio of the peak to average is 16:1 or 12 dB.

Such a huge ratio has significant practical consequences. If one were to design a radio transmitter to handle twice the average power (0.5 W) without distortion, the transmitter would work well 90% of the time. However, in order to achieve a reasonable bit error rate of 10^{-5}, it is necessary to provide a much larger power handling capability of about 5 Watts, with associated increases in cost, size, and DC power consumption. The extent to which the average power must be backed off from the peak power will be discussed in more detail later in this tutorial.

 ©2001 WJ Communications Inc.

CDMA

Let's examine a practical example in which filtering, efficient modulation, and multiple simultaneous signals combine to generate a high peak-to-average ratio signal.

Code-division multiple access (CDMA) is a spread-spectrum technique employed in cellular communications. In this approach, many subscribers can use the same frequency band at the same time without significant interference. Each bit of a user's binary bit stream is multiplied by a code consisting of a sequence of positive or negative values, each of which is much shorter in duration than the data bit it multiplies. The resulting signal has a higher effective bit rate than the original signal, and is thus spread out in frequency. If individual codes are chosen to be orthogonal to each other or nearly so, multiple signals can be sent using the same frequency. Each user multiplies the total signal by their individual code, thereby extracting only their data stream.

In actual practice, the individual bit streams are passed through a Nyquist filter and encoded in a variant of QPSK, then added together. In the IS-95 CDMA standard for cellular telephony, as many as 20 individual user data streams may be simultaneously encoded onto a single downstream signal, leading to a situation very much like that shown in Figures 6 and 7 above.

An example of the resulting complex trajectory is shown in Figure 8, which depicts the path of a signal representing 10 CDMA channels, each transmitting 256 symbols, in phase space. The trajectory spends most of its time near the center of the phase-amplitude plane, at low signal amplitudes. However, occasional excursions occur to the constellation points at the outer edges of the range, giving rise to infrequent but large peak signal voltage and power. The amplitude of this 10-channel signal could grow as large as $\sqrt{2}$ (the corners of the dotted box in Figure 8) if all the signals were in phase, but the extreme is sufficiently unlikely that it doesn't appear in the limited simulation shown in the figure. The estimate of the average amplitude (about 0.4) obtained from the simulation should be quite accurate, so that we can estimate the ratio of peak to average power at a reasonable probability level (e.g. 10^{-5}) as

$$\left[P/A\right] \approx \left(\frac{\sqrt{2}}{0.4}\right)^2 = \frac{2}{.16} \approx 11 dB$$

This result is very close to the 12 dB peak-average ratio obtained from the simplified Gaussian signal we looked at above.

 ©2001 WJ Communications Inc.

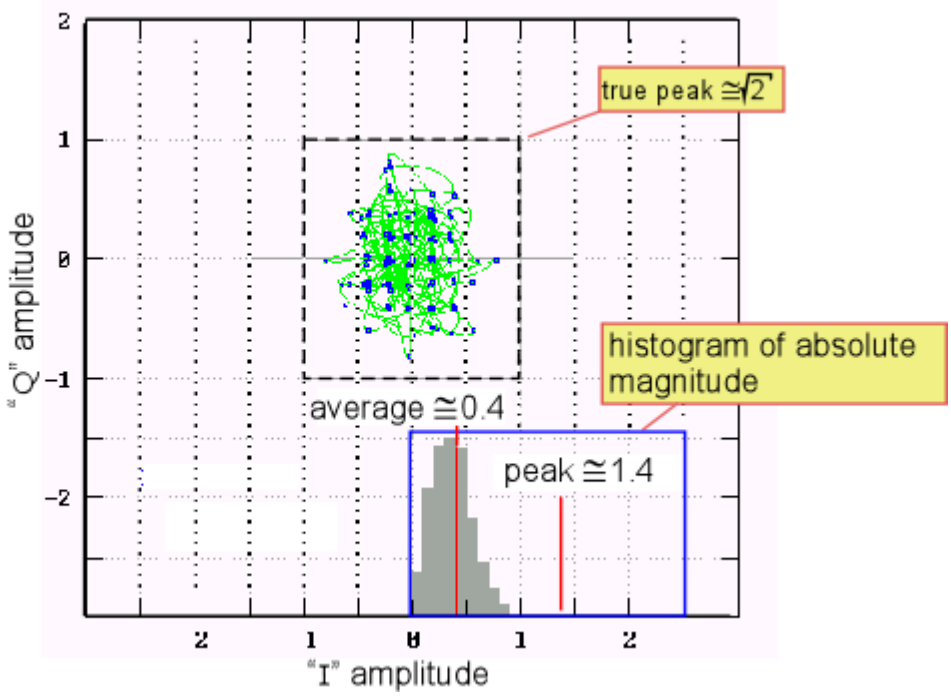


Figure 8: Simulated constellation diagram and amplitude distribution for 10 superimposed QPSK-modulated CDMA signal streams.

Other High-Dynamic-Range Signals

Orthogonal frequency-division multiplexing (OFDM) is a modulation technique that is used in fixed-wireless communications systems and in some current and proposed wireless LAN standards. In this technique, a high-rate bitstream is separated into a set of lower-rate streams, each of which is transmitted on a separate carrier frequency. The low bit rate means that the individual carriers are relatively immune to time-dependent impairments such as multipath distortion, a phenomenon resulting from the differing time delays of the various paths an electromagnetic wave can follow as it travels from transmitter to receiver. The individual carrier frequencies are chosen to be orthogonal to each other, so that they can be spaced very close together and still allow accurate extraction of each signal. In practical systems QPSK or a similar efficient modulation scheme, with filtering, is employed for each carrier, and the final signal is actually generated and demodulated digitally using advanced signal processing techniques, rather than employing separate modulation of a large number of analog carrier signals. The OFDM signal is thus also a sum of a large number of uncorrelated bitstreams; the amplitude varies widely, with a characteristic time of roughly 1/BW where BW is the bandwidth of the signal. A realistic OFDM signal, such as that specified in the IEEE 802.11a standard, employs 52 subcarriers and has a roughly Gaussian distribution of amplitudes, with a peak-average ratio of about 10 dB at a probability of 10^{-5}.

Another practical example of large peak-average ratio signals are those employed in cable TV systems. In the United States, such signals are composed of a multitude of 6 MHz wide channels (up to 110). Each channel contains either an analog NTSB television signal or a digital signal using a generalized version of

QPSK that allows multiple phases and amplitudes: ***quadrature amplitude modulation*** (QAM). A filled band contains signals from roughly 50 to 850 MHz. The peak-to-average ratio is roughly 12 dB, much as we observed in the ideal Gaussian-distributed signal we examined in Figures 6 and 7. The extremely broadband nature of the CATV signal (> 4 octaves wide) means that both second and third order distortions are in-band, as will be discussed in the next section, so the question of system linearity is particularly important for cable TV analog design.

Effects of Nonlinearity

We have shown that many digital wireless communications requirements lead to signals with a significant range of amplitudes. What consequences result? Why do we care?

Third-order distortion: spurious frequencies and ACPR

A perfect linear system can change the relative intensities and phases of the frequencies in its input, but doesn't generate any new frequencies. If the input is bandlimited to a specific frequency channel (in order to meet regulatory specifications and avoid interference with other wireless communications channels), the output will be bandlimited too.

However, any nonlinearity in the transmitter can generate additional frequencies not present in the input signal. The origin of these additional frequencies can be understood very simply from the trigonometric identities:

$$[\cos(x)]^2 = \tfrac{1}{2} + \tfrac{1}{2}\cos(2x)$$
$$[\cos(x)]^3 = \tfrac{1}{4}[3\cos(x) + \cos(3x)]$$

with similar identities for sin(x). We can see that a nonlinear response that looks like a quadratic function generates a new frequency component at twice the frequency of the input signal, and that a distortion that looks cubic not only generates a new frequency at three times the original, but also a distortion term that changes the amplitude of the original input frequency. These changes are depicted in Figures 9 (a) and (b).

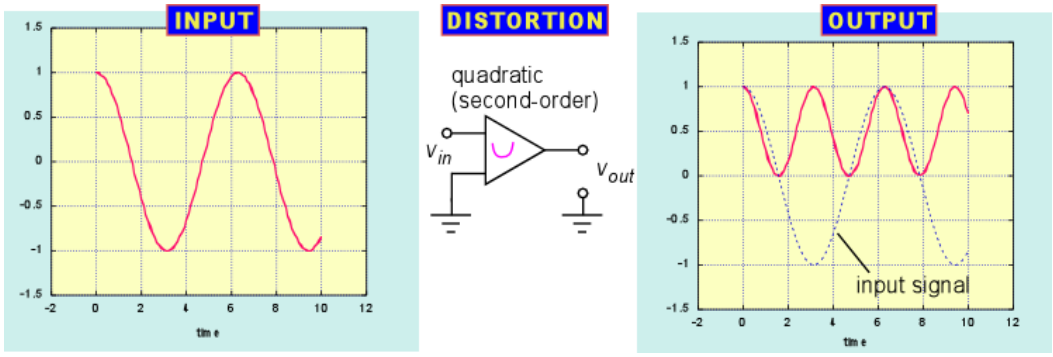


Figure 9(a): Second-order distortion of a signal results in an offset, frequency-doubled output

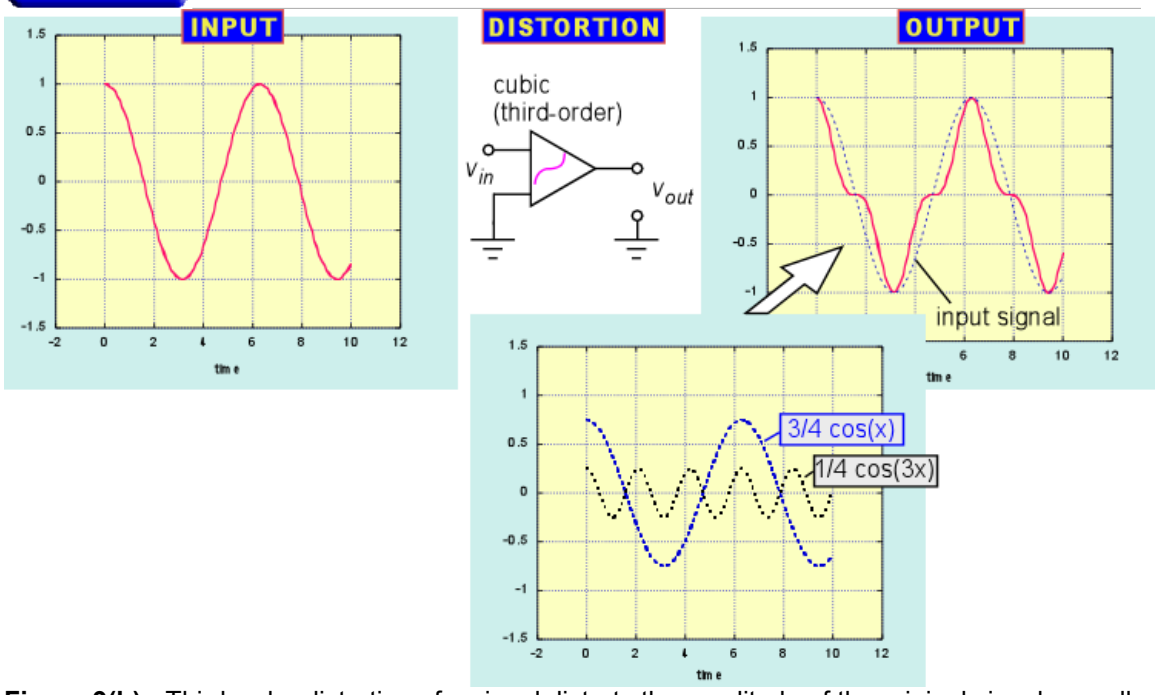


Figure 9(b): Third-order distortion of a signal distorts the amplitude of the original signal as well as adding a component at triple the original frequency

These simple cases illustrate the general rule that even-order distortion creates DC offsets and even harmonics, whereas odd-order distortions create odd harmonics and signals ***at or near the input frequency***. The even-order harmonics are usually easy to filter out (except in very broadband systems like cable TV); it is thus odd-order distortion that mostly concern us in digital communication systems.

A real amplifier operating in the small-signal regime will be nearly linear with some small second- and third-order curvature in its transfer characteristic. The relative amount of third-order distortion in the output will be proportional to the square of the signal amplitude: distortion grows as the amplitude cubed, and the signal itself grows linearly, so their ratio is $x^3/x = x^2$. This relationship is depicted schematically in Figure 10. To the extent that the slope of the distortion line is really equal to 2, one can completely describe the amount of third-order distortion at any input power simply by specifying one point on this line. Typically, the point that is specified is the power at which the third-order distortion becomes equal in magnitude to the output signal : this point defines the input and output ***third-order intercept points*** (IIP3 and OIP3). The dotted portion of the distortion line shown in figure 10 emphasizes the fact that the OIP3 or IIP3 cannot be directly measured: for input powers close to IIP3, higher-order distortions will become large, and the distortion term will deviate strongly from a simple line of slope=3. Reported IIP3 values are

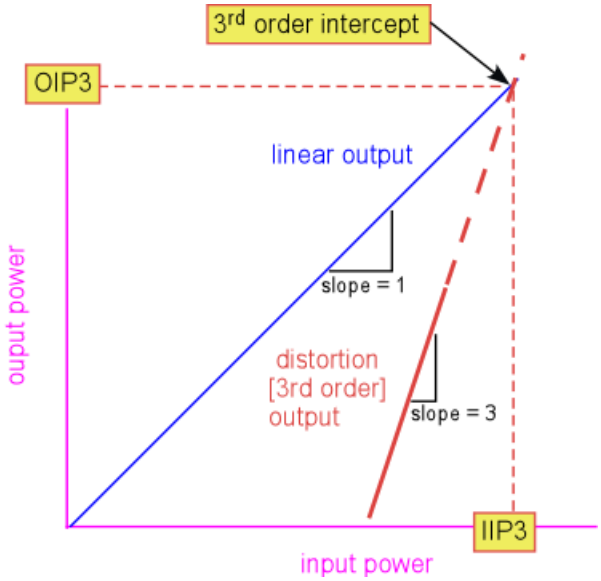


Figure 10: linear output power and third-order distortion (logarithmic scales for input and output power)

extrapolated from measurements at very low power where the distortion is small (typically 10-30dB below the power at which significant gain compression is observed).

The amount of distortion is very sensitive to signal amplitude. Signals that have a large peak-average ratio, like the digital signals we discussed above, will encounter significant distortion even when a constant-amplitude signal of the same average power would undergo essentially no distortion at all.

Third-order nonlinearities produce distortion of the input signal: ***in-band*** distortion. However, when more than one frequency is present in the input signal, third-order distortion will also result in ***adjacent channel interference***. This effect arises from the interaction of the frequencies in the signal: ***intermodulation*** distortion. A simple example will clarify the mechanism. Assume there are two nearby frequencies in the input signal, and examine what happens when a third-order distortion acts on them:

$$[\cos(x) + \cos(x-\delta)]^3 = [\cos(x)]^3 + [\cos(x-\delta)]^3 + 3[\cos(x)]^2 \cos(x-\delta) + 3\cos(x)[\cos(x-\delta)]^2$$

Using the trigonometric identity $\cos(x)\cos(y) = \frac{1}{2}\cos(x-y) + \frac{1}{2}\cos(x+y)$, and remembering from our previous discussion that squaring a cosine doubles the frequency, we see that the third and fourth terms contribute frequencies of

$$[2x - (x-\delta)] = x+\delta$$

and $$[2(x-\delta) - x] = x - 2\delta$$

(as well as the in-band distortions and higher harmonics we've examined before). If the two original frequencies were at opposite ends of the allowed spectrum, then the two new frequencies are ***outside*** the boundaries of the initial channel but too close to it to be easily filtered out. This effect is shown in Figure 11.

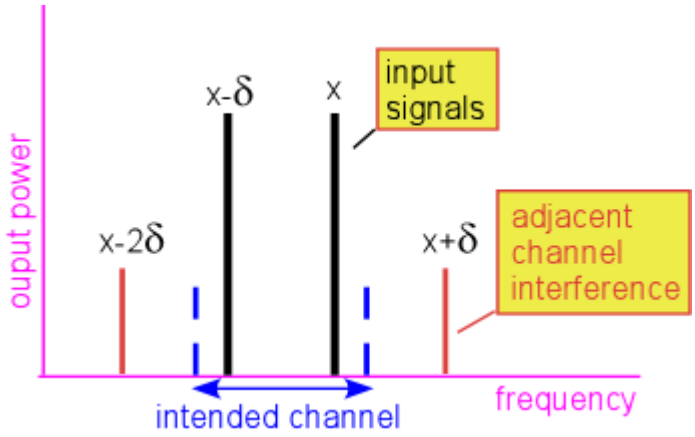


Figure 11: Third-order distortion leads to intermixing of neighboring frequencies, creating interference in adjacent channels

The Fourier transforms of most digital signals generated from pseudo-random input bitstreams have roughly constant amplitude out to the first zero – that is, digital signals contain lots of frequencies out to the edges of the channel. Each possible pair of frequencies will contribute to the adjacent channel interference. The overall effect is to produce a pair of shoulders on either end of the intended spectrum as the input power (and thus distortion) increase: this phenomenon is known as ***spectral regrowth***. An example is shown in Figure 12, for the same 10-channel CDMA signal described previously.

 ©2001 WJ Communications Inc.

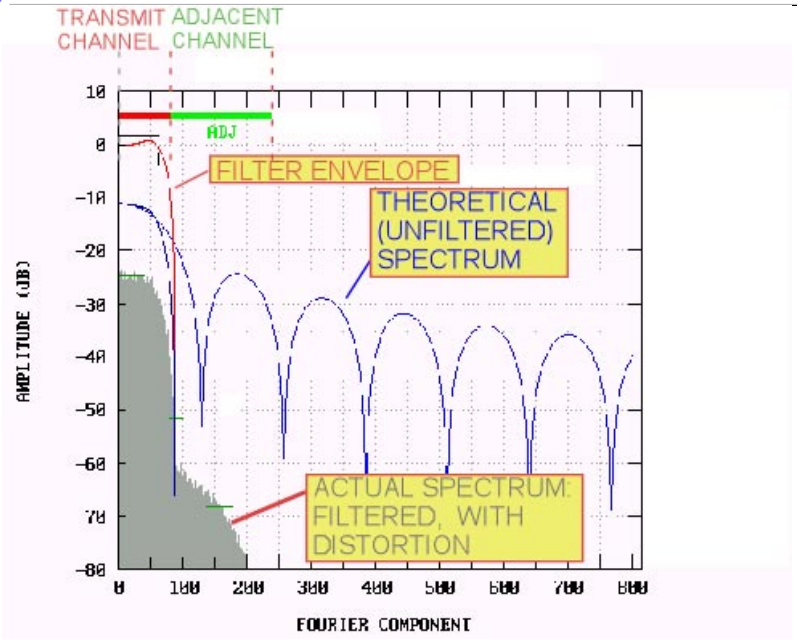


Figure 12: Adjacent–channel interference due to nonlinear distortion of a multichannel CDMA signal

Clipping

Any real amplifier can only supply a finite output voltage: even if the device is ideal, the output can't exceed the supply voltage. For large signals, this clipping of the output signal becomes the dominant nonlinear effect. For multichannel digital signals with large peak-to-average ratios, clipping will first occur at the rare excursions to high amplitude, as shown schematically in Figure 13.

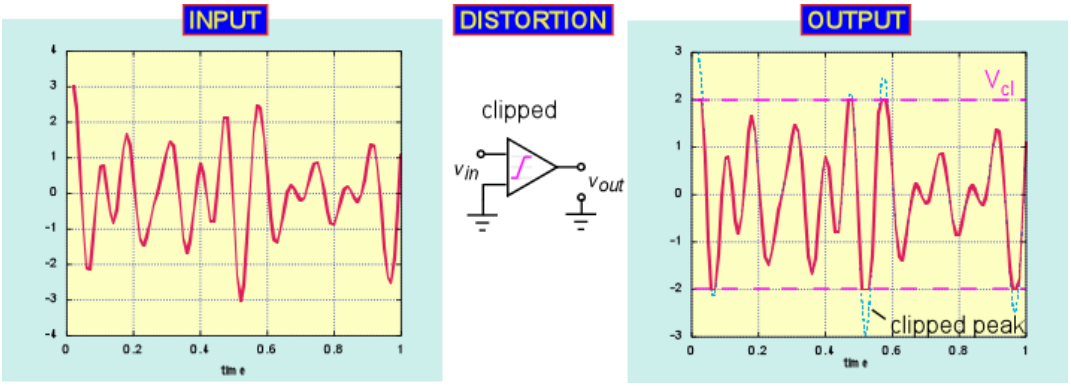


Figure 13: Clipping of signal with large peak-average ratio

Unlike 3rd-order distortion, clipping distortion results in a relatively broad spectrum typically extending well beyond the adjacent channel. The magnitude of a complex Gaussian signal has a Raleigh distribution, so that in the limit when clipping is rare, the probability of clipping is exponential in the ratio of the average power to the clipping level:

$$P(clip) \propto \exp\left(-\left[\frac{P_{clip}}{2P_{in}}\right]\right)$$

To analytically estimate the adjacent-channel interference due to this infrequent but abrupt distortion, one would need to calculate the bandwidth of the distortion, which turns out to be quite difficult to do in general. A semi-empirical approach using a slightly modified functional form give an excellent fit to data from full numerical simulations (Figures 14 and 16 below):

$$ACPR = -17 - \exp\left(\frac{6-\left(P_{out}-P_{clip}\right)}{0.9+0.26\left[P/A\right]}\right) \qquad \{1\}$$

In a log-log plot (i.e. dB on both axes) typically used to depict distortion power or adjacent-channel interference, the clipping distortion will appear as an exponential in the input power, increasing rapidly from a small value as the input power approaches within (roughly) the clipping power divided by the peak-average ratio.

Combined Effects of Clipping and Third-Order Distortion

In real devices, both these distortion effects act simultaneously, in addition to higher-order amplitude distortion and phase-distortion effects that we have ignored here for simplicity. The net result is that clipping dominates adjacent-channel interference at high input power, whereas third-order distortion dominates when the input power is backed off from the clipping level by more than the peak-average ratio. Thus, the clipping-dominated regime expands for signals with increasing peak-average ratio.

To demonstrate these phenomena, we have modeled a simple system in which the transfer characteristic has a small third-order term for amplitudes less than a clipping level, and then is completely clipped to a constant output amplitude for inputs larger than the clipping level. The result is shown in Figure 14 for two different values of the third-order distortion and two different composite CDMA-format signals. We can see that the interference behavior is separated into distortion-dominated and clipping-dominated regions. The clipping-dominated region extends to lower input powers when more channels are superposed, as one would expect from the increase in peak-average ratio that results. The two-tone 3rd order distortion is almost equal to that generated by the 10-channel signal in the low-power regime. This fortunate agreement explains in part the utility of simple two-tone measurements [3].

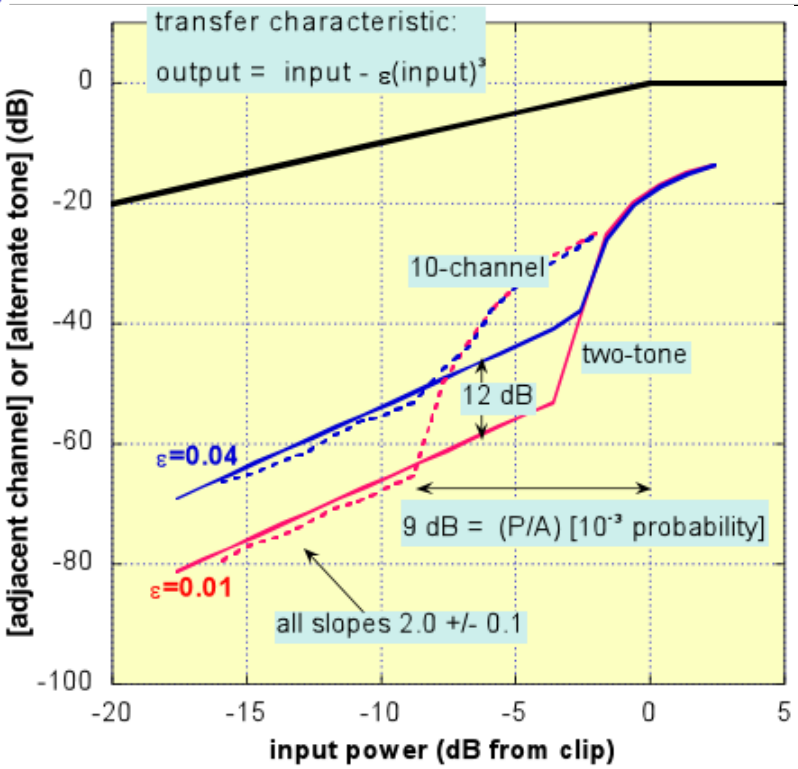


Figure 14: Combined third-order / clipping distortion characteristic for piecewise-cubic transfer characteristic. SOLID: two-tone signal. DASHED: 10-channel CDMA signal. BLUE (top): high cubic distortion. RED (bottom): low cubic distortion.

Design of High Dynamic Range Systems

Component Specifications

We have established that the designer must account for nonlinear effects and distortion in specifying components for digital wireless communications systems. Let us consider how a designer, when constrained by a minimum output power requirement, maximum adjacent channel interference specification, and signal of known peak-average ratio, might select an appropriate amplifier. Assume in the discussion below that the designer seeks to satisfy the specifications with the smallest acceptable values of clipping power and intercept. As discussed previously, interference due to clipping falls rapidly as the input signal is reduced. To a first approximation, one might simply ensure that the device clipping level is larger than the desired output signal by at least the peak-average ratio to ensure that clipping is negligible:

$$P_{clip} = P_{out} + \left[P/A \right] \quad \text{(where all quantities are in dB or dBm)} \qquad \{2\}$$

That is, we back off the signal level from the clipping level by the peak to average ratio.

It is then necessary to ensure that the third-order distortion is sufficiently small to meet the ACPR specification. Since the third-order power is proportional to the cube of the signal, its ratio to the signal power goes as the square of the input power; that is, ACPR changes by 2 dB for every dB change in signal power. We therefore have as our second condition:

$$OIP3 - \frac{|ACPR|}{2} = P_{out} \qquad \{3\}$$

where ACPR is the specification requirement for adjacent channel interference. These relationships are shown schematically in figure 15.

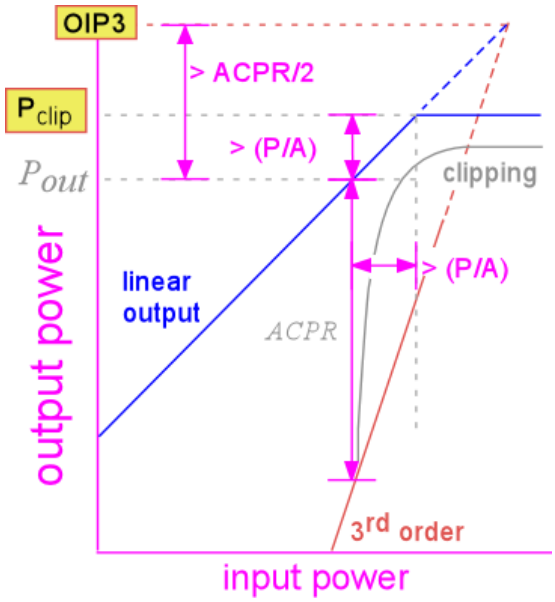


Figure 15: Simplified scheme for determining component performance from specifications

By subtracting equation {2} from equation {3} we obtain a requirement on the amplifier which is independent of the output power level:

$$OIP3 - P_{clip} = \frac{|ACPR|}{2} - \left[P/A \right] \qquad \{4\}$$

Equantion {4} shows that demanding ACPR specifications will require large values of ***linear efficiency*** , $\eta = OIP3 - P_{clip}$.

In order to provide a more complete treatment of the problem at this level of approximation, we must take into account two effects ignored in the simplified treatment of equations {2}-{4}. First, the actual shape of the clipping distortion characteristic should be modeled more accurately, instead of assuming it to fall to zero for backoff $[P_{clip} - P_{out}] >$ (P/A) (see Figure 16). Using the empirical fit from eqn. {1} with all quantities expressed in dB or dBm, we obtain:

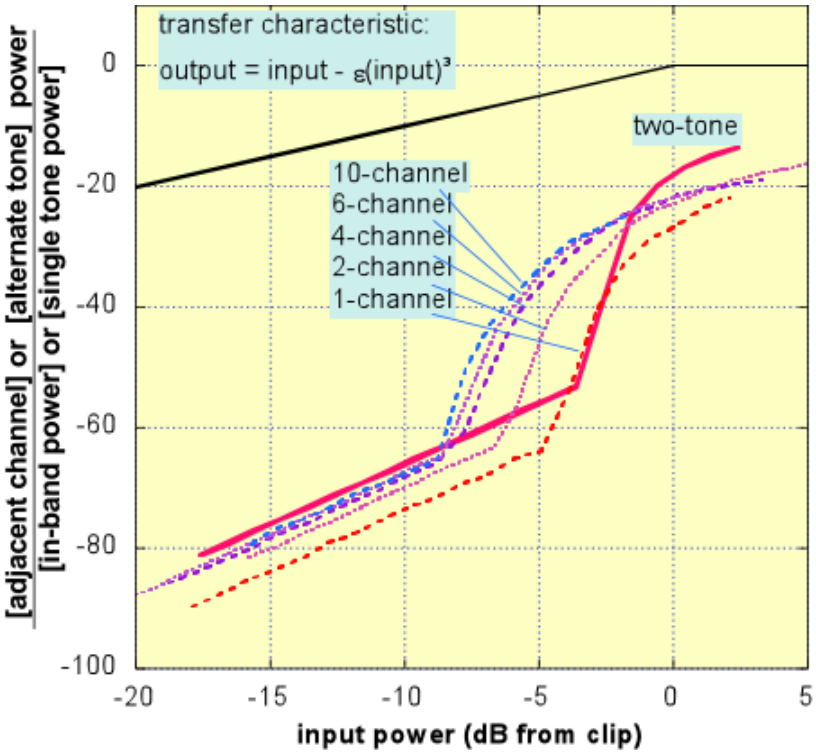


Figure 16: modeled distortion for overlaid CDMA-like QPSK channels, showing the effects of changes in the number of channels (i.e. in [P/A] ratio) on clipping and third-order distortion

 ©2001 WJ Communications Inc.

$$P_{clip} - P_{out} = \left(0.9 + .26\left[P/A\right]\right)\ln(|ACPR| - 17) - 6 \qquad \{2'\}$$

The resulting backoff is nearly linear in the peak-average ratio, as equation {2}, but is somewhat offset, so that less backoff is needed for smaller values of |ACPR|.

The second effect neglected in {3}is the change in third-order as the peak-average ratio changes, since we must average over a distribution of signal powers to obtain the expected distortion. This effect is also shown in figure 16. A convenient empirical fit to the model, assuming that the (P/A) ratio at a probability of 10^{-5} is roughly 11.5 dB, gives a correction factor to equation {3}:

$$OIP3 - \frac{|ACPR|}{2} = P_{out} - 11.5 + \left[P/A\right] \qquad \{3'\}$$

The expression for linear efficiency accounting for clipping shape and third-order effects is then:

$$OIP3 - P_{clip} = \frac{|ACPR|}{2} - 11.5 + \left[P/A\right] - \left(0.9 + .26\left[P/A\right]\right)\ln(|ACPR| - 17) + 6 \qquad \{4'\}$$

The results are summarized as contour charts in Figures 17 and 18. Linear efficiency is mostly determined by ACPR; stringent interference specifications require high values of OIP3 relative to the clipping power. Backoff is determined jointly by the peak-average ratio of the signal and the ACPR spec; modest interference requirements (e.g. –30 dBc) are easily met for output power close to clipping, whereas stringent ACPR specs cause the backoff to become roughly equal to the peak-average ratio.

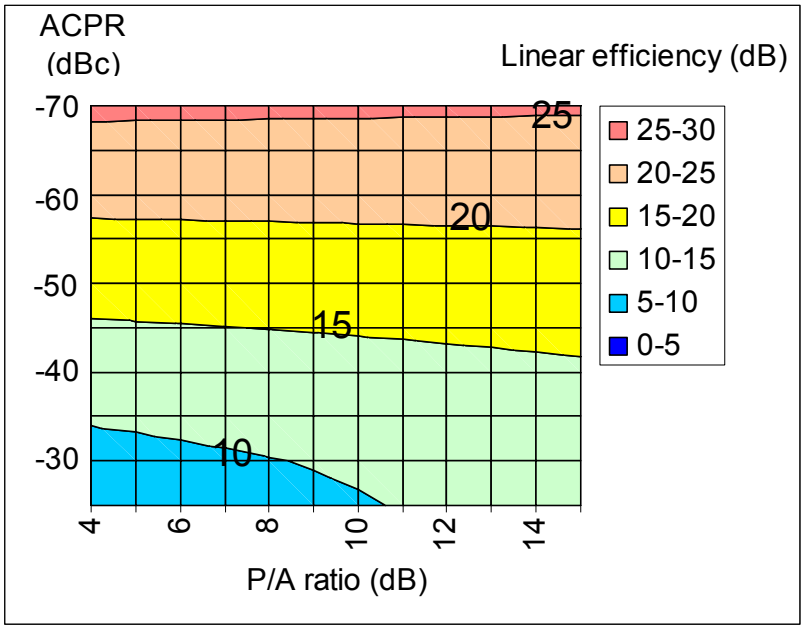


Figure 17: Linear efficiency (η = OIP3 – P_{clip}) requirement vs. signal (P/A) ratio and ACPR requirement, based on simplified model of clipped third-order distortion for CDMA-like signals

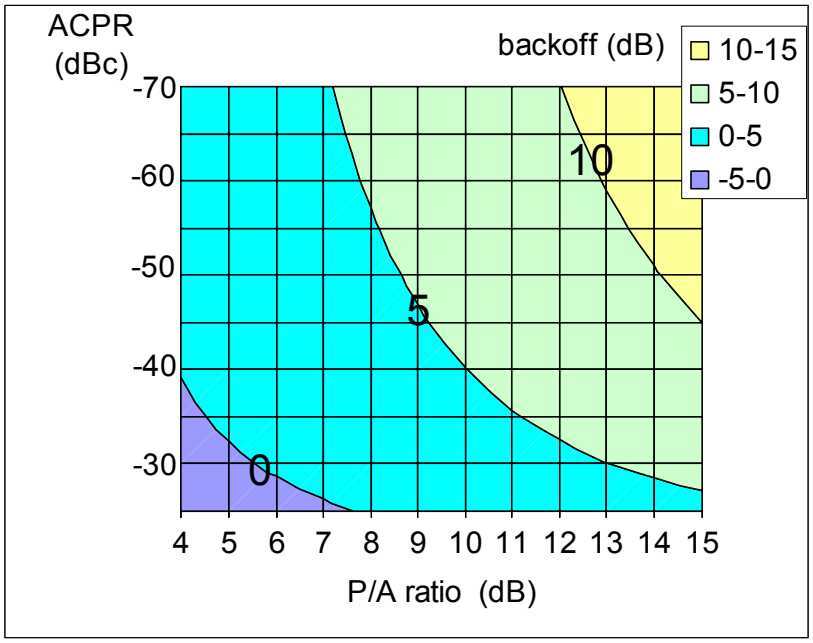


Figure 18: Backoff requirement vs. signal (P/A) ratio and ACPR requirement, based on simplified model of clipped third-order distortion for CDMA-like signals

It is important to note that these design graphs are based on simplified models of real devices, in order to enable selection using only the parameters typically available on datasheets. Important trends displayed in the graphs are representative of real devices, but for any specific circuit application it is imperative that the designer confirm these predictions with measured data. Factors not accounted for include higher-order distortions, phase distortion (AM-PM conversion), and the actual shape of the clipping region of the transfer function.

Approaches to High Dynamic Range Devices

To fabricate amplifiers with good linear efficiency, the component designer can employ processes and devices designed specifically for enhanced dynamic range, or seek circuit topologies which minimize the deleterious effects of distortion, or both.

At WJ Communications, we have optimized the design of our GaAs MESFETs for low third-order distortion. This is accomplished by careful adjustment of the channel doping and geometry of the gate and gate recess, ensuring that signal-dependent variations in device transconductance are almost completely nullified by the signal dependence of the device output conductance. The result is a device with very low third-order distortion over a wide range of input power, with little compromise in noise figure, or equivalently extremely wide dynamic range.

Table I: GaAs FET Dynamic Range

Device Type	Output 3rd order Intercept (dBm)	Linear efficiency (dB)	Noise figure (dB)	Spurious-free dynamic range (dB)*
Typical MESFET	33	12	2.5	86
WJ AH-1	41	20	2.7	94

*Spurious-free dynamic range assumes a 1.25 MHz bandwidth (IS-95), 14 dB small signal gain

The device structure must be optimized for a given device operating condition. Most current WJ Communications high-dynamic range MESFETs are designed for operation at zero gate bias ($I_{ds} = I_{dss}$) , and thus require only a single positive voltage supply.

One of the advantages of employing MESFET technology is that the dominant nonlinear elements in the device are conductances, determined by doping concentration and electron mobility. The distortion behavior of MESFET amplifiers is thus relatively insensitive to variations in operating frequency and ambient temperature. However, MESFET nonlinear modeling is poorly understood in comparison to the nonlinear behavior of bipolar junction transistors (BJTs), so structural optimization is challenging.

To obtain very high gain in a small area, one may employ bipolar junction transistors instead of FETs. The transconductance of a BJT is to a good approximation just $I_c/(kT/q) = I_c/40$ at room temperature. For reasonable current densities, bipolar transistors can provide much higher transconductance per unit chip area than comparable MESFETs. The combination of high specific gain with a circuit configuration such as a Darlington pair allows the use of copious amounts of negative feedback while still preserving acceptable overall amplifier gain, achieving low third-order distortion and good dynamic range. Heterostructure bipolar transistors, with their heavily-doped base regions, also have very low output conductance, making the output matching design simpler than for a corresponding MESFET circuit. Bipolar devices have very low 1/f noise, making them suitable for low-phase-noise oscillators and low-frequency applications.

However, there are some disadvantages to using bipolar transistors in high-frequency, high-dynamic range amplifiers. The dominant nonlinear elements of a BJT are the transconductance and the parasitic capacitances, which have complex dependences on bias conditions [4]. Bipolar amplifiers require a linearizing resistance to remove the exponential dependence of the transconductance on operating current; this resistor dissipates a significant added power and reduces headroom available for a given supply voltage. The presence of a significant nonlinear capacitance and the strong frequency dependence of the intrinsic gain cause the distortion behavior of BJT circuits to be more frequency-dependent than that of their MESFET counterparts (Figure 19). Bipolar devices often have higher high-frequency noise figures than MESFETs, dominated by junction shot noise. The use of minority carriers in bipolar devices means that their characteristics are also more dependent on temperature than is the case for MESFETs. Finally, MESFETs tend to be more robust than BJT's when operated at high channel temperatures. FET channel current decreases with increasing channel temperature due to reduced electron mobility, whereas BJT collector current increases with increasing temperature due to increased injection from the emitter. Bipolar devices are thus subject to thermal runaway, requiring careful design and packaging to ensure thermal stability.

 ©2001 WJ Communications Inc.

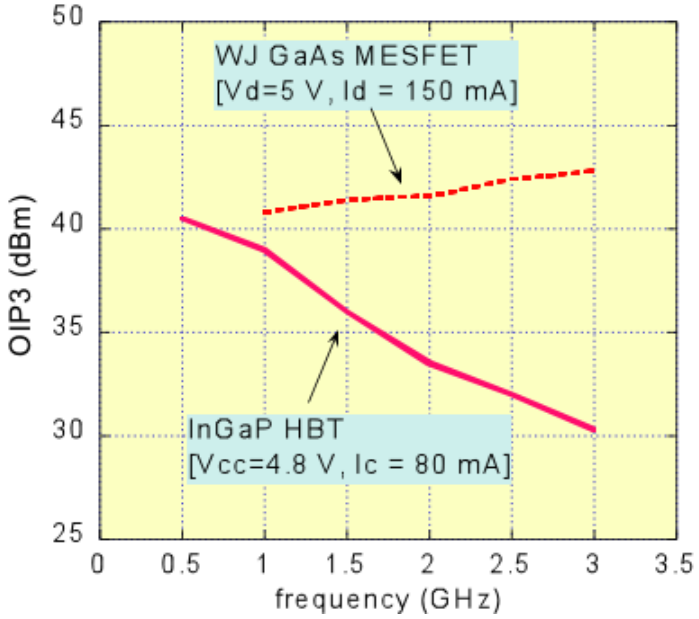


Figure 19: Third-order intercept point vs. operating frequency for generally similar HBT and MESFET amplifiers

Conclusion

The need for linearity in digital wireless communications arises primarily from the requirement of minimal interference with adjacent channels. To meet the specifications that result, the designer should understand how spectral regrowth is affected by device nonlinearities measured by intercept point, clipping measured by compressed or saturated output power, and the properties of the input signal, often summarized in the peak-to-average ratio at a given probability. Fortunately for the designer, the distortion properties of any signal composed of more than a few uncorrelated binary inputs are close to those of a Gaussian-distributed signal.

Device designers have several options to produce the highly linear semiconductor amplifiers required by modern communications systems. GaAs MESFETs and heterostructure bipolar transistors have both demonstrated excellent linearity; the circuit designer should choose the most suitable technology based on requirements for gain, reliability, and ease of use.

Acknowledgements

The authors would like to thank colleagues including John Bellantoni, Titus Wandinger, Brent Ostermann, Andrew Manzi, and Mark Kelly for reviewing various versions of this manuscript and providing data and thoughtful suggestions.

References

1. "Certain Topics in Telegraph Transmission Theory", H. Nyquist, Trans. AIEE, Vol 47, p. 617, April 1928.
2. **Wireless Multimedia Communications**, E. Wesel, Addison-Wesley 1998 ISBN 0-201-63394-9, chapter 4
3. "On the Use of Multitone Techniques for Assessing RF Component's Intermodulation Distortion", J. Pedro and N. de Carvalho, IEEE Trans MTT vol 47 p. 2393 (1999), esp. figure 5 therein.
4. "Influence of Collector Design on InGaP/GaAs HBT Linearity", M. Iwamoto, T. Low, C. Hutchinson, J. Scott, A. Cognata, X. Qin, L. Camnitz, P. Asbeck and D. D'Avanzo, Proc. IEEE MTT-S International Microwave Symposium Digest, volume 2 p. 757 (2000)