How do amplifiers treat big signals?

J M Woodgate BSc(Eng) C.Eng MIET SMIEEE Life FAES HonFInstSCE MIOA

jmw1937@btinternet.com www.jmwa.demon.co.uk

2017-09-22

1 Introduction

We know how amplifiers treat small signals; modern amplifiers deliver to the load a very accurate magnified copy of the input signal. But what happens with big signals? We need to understand this very well, especially for amplifiers in sound systems, that often have to work with very big signals for much of the time, and have to handle noise signals (which randomly vary from small to big and back again) during test procedures.

This study concentrates on non-switching amplifiers with single-voltage or 'split rail' DC supplies and no low signal-level processing, such as automatic gain control (AGC). Class D amplifiers are not discussed because different techniques need to be treated separately.

2 What controls what happens?

There are six things that control what happens:

a) the input signal amplitude and frequency (bandwidth if it is not sinusoidal);

b) the amplifier bandwidth, which may be a function of the input signal amplitude and frequency;

c) the no-signal DC supply voltage (voltages if the amplifier has a split supply);

d) the resistive component of the power supply source impedance;

e) the stored energy in the power supply;

f) the load impedance (which in real life is almost never a pure resistance).

The stored energy doesn't matter if the resistive component of the power supply source impedance is extremely low, because in that case, the very large stored energy of the AC electricity supply is accessible to the amplifier. But that is a rather rare condition.

'What happens' happens in the output stage, unless the amplifier is not well-designed. Note that 'RMS' doesn't appear anywhere, neither for the input signal or the output signal. It is relevant only to the power supply temperature.

3 Simple start

We begin with a sine-wave input signal, well within the amplifier bandwidth, a regulated DC power supply of negligible source resistance and a resistive load.

As we increase the input voltage, the output voltage increases in exact proportion until the peak-to-peak output voltage approaches the DC supply voltage (the total voltage if the supply is split). The output voltage cannot quite reach the DC supply voltage because there are residual voltage drops across the semiconductors and, for example, emitter series resistors in some configurations. If we apply a larger input voltage, the tips of the sine-wave signal are clipped off. If the amplifier is well-designed, this clipping is symmetrical on positive and negative peaks, and introduces only odd-harmonic distortion. (Don't panic! As we shall see, in moderation this is not important.) Figure 1 shows these effects. In this case, the total DC supply voltage is 20 V regulated, with 1 mΩ source resistance, and the 100 Hz input voltage is increased from 0.2 V RMS to 0.8 V RMS in steps of 0.2 V. The amplifier voltage gain is set to 10 by feedback.

4 Major complication

The change from a regulated to an unregulated power supply introduces a lot of troubles. Back in the days of vacuum valves, power supplies even in costly amplifiers were mostly not regulated, because it was expensive and increased the heat produced. With an unregulated supply, its output voltage rises when the output current is small, and falls when it is large. These effects take time, and the result is that for relatively brief periods (with normal component values), the clipping level

is higher than it is with a continuous signal. Typically, for a power supply with the usual full-wave rectifier, the DC output voltage drops to 0.7 of its low-load value, corresponding to a drop in clipping voltage of 3 dB. For the rare case of a half-wave rectifier, the drop os about 10 dB. The effect can be quite noticeable – the amplifier 'sounds louder', leading to the concept of 'music power' and others of even less probity. You can test this if you have an amplifier that can be operated from the mains or from a battery. Listen at high level (not for too long!) in each mode. Maybe the mains supply gives an initial voltage of 34 V, while the battery gives a constant 24 V. That means roughly 3 dB more SPL. (I say 'SPL' because I'm desperate to avoid mentioning the dreadfully misleading 'P-word' (starts with 'p' and ends in 'ower').

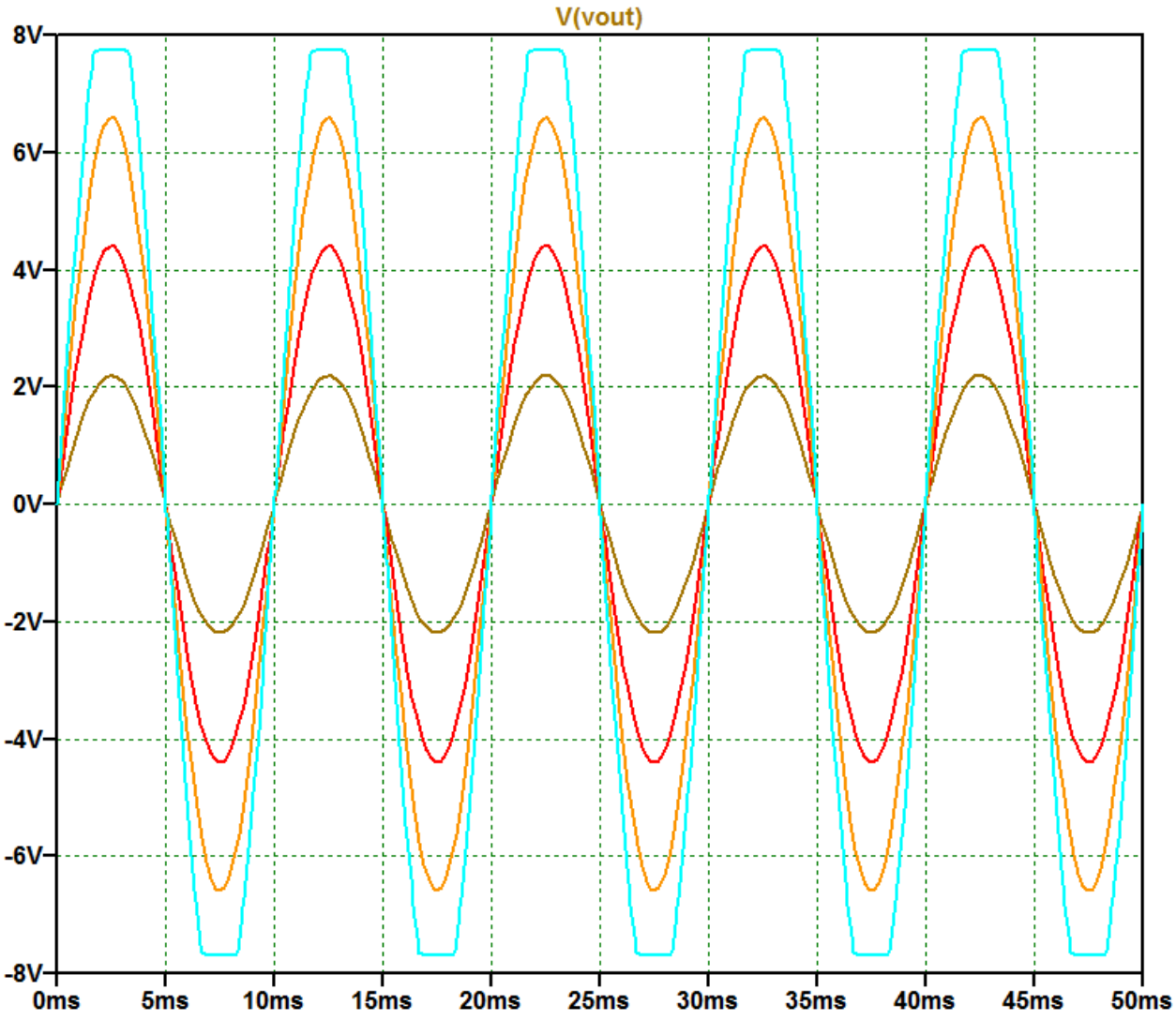


Figure 1 Effects of increasing input signal amplitude

Figure 2 shows what is going on. The initial peak-to-peak output voltage after a silence is 14.5 V, but after 1 s this has fallen to 11 V. That's a 2.4 dB drop. In this case, the stored energy in the power supply is sufficient (large filter capacitor) but the source resistance is not negligible. The opposite can occur, rarely, but we will leave few stones unturned, and there is an important point involved. Figure 3 shows what happens if the power supply has insufficient energy storage – too small a filter capacitor. This effect can be disguised if the source resistance is low enough. Not only is the peak-to-peak voltage initially reduced, but it actually decreases further during the clipping time. This introduces high-frequency components into the signal that are modulated at the signal frequency, so sound rough. We do not want this. The results are different if the load on the amplifier is not a pure resistance, but more on that later.

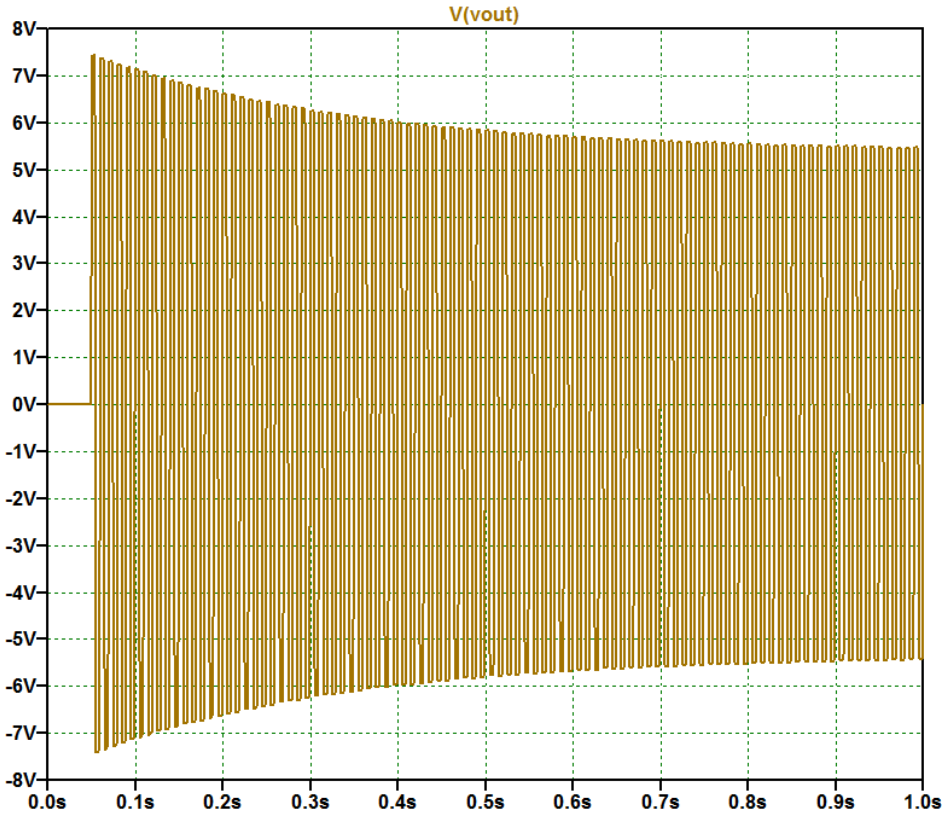


Figure 2 Variation of output voltage due to power supply impedance

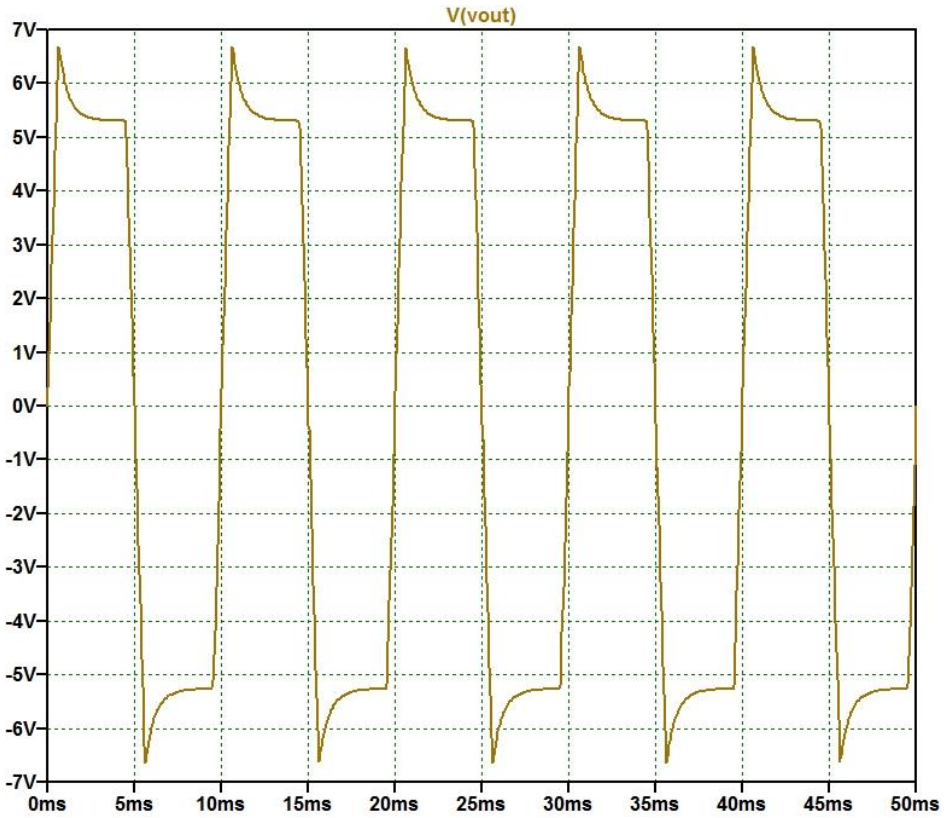


Figure 3 Low stored energy

5 Real-life loads

Audio amplifiers do not normally feed pure resistive loads, except during testing. We need to take this into account, as it throws a lot of necessary light on the misleading P-word. Figure 4 shows the impedance/frequency curves (magnitude and phase) of a simulated single driver loudspeaker. The curves for vented boxes and multiple-driver boxes are even more different from the straight horizontal lines produced by a pure resistive load.

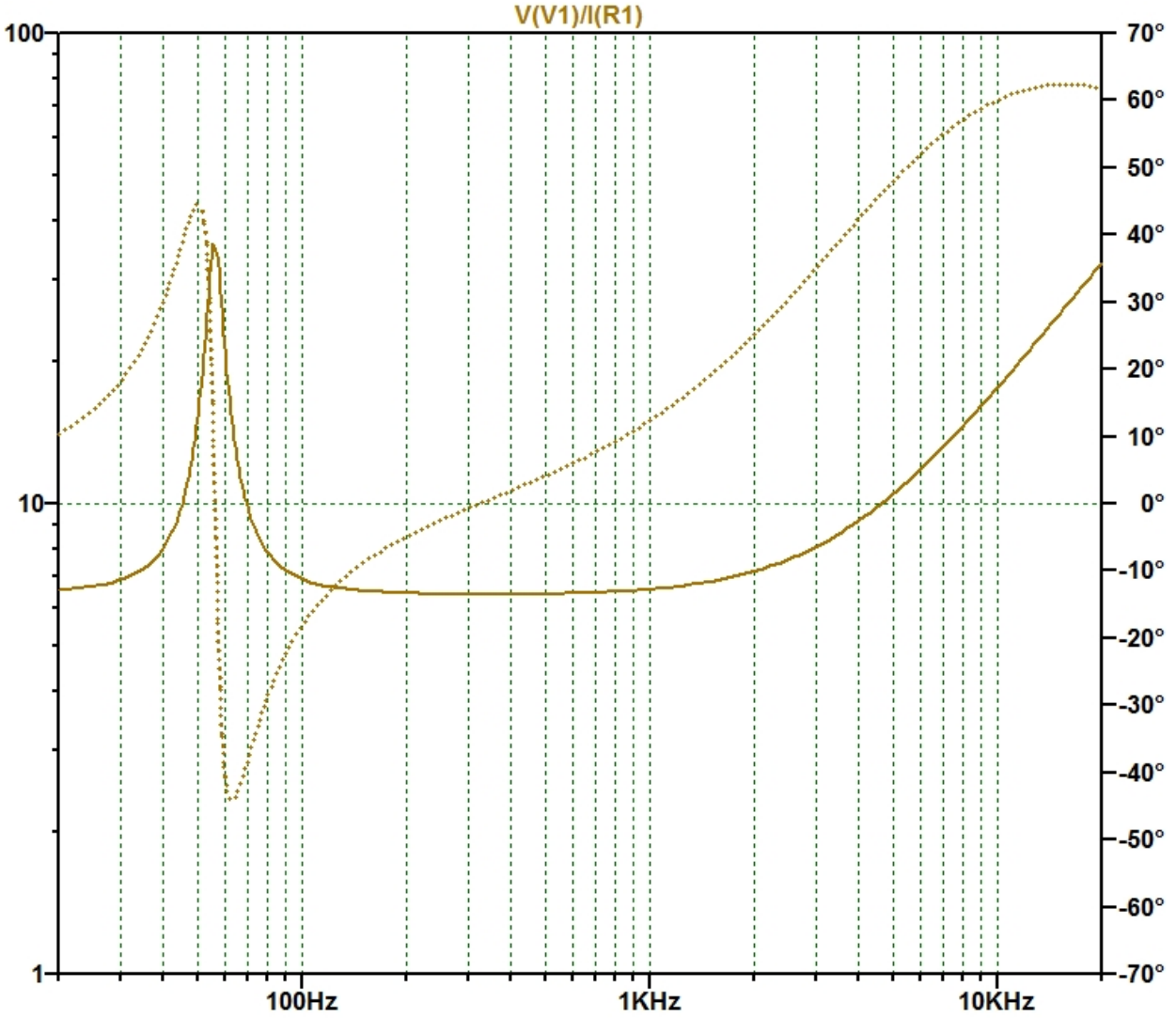


Figure 4 Magnitude and phase impedance curves of a single driver

The impedance peaks at 36 Ω at 56 Hz and is resistive (zero phase angle) at that frequency and at just one other, 302 Hz, where the resistance is 6.4 Ω, as permitted for an 8 Ω driver in IEC 60268-5. The lower frequency is the bass resonance frequency, so varies a lot between different drivers, but the higher frequency is usually between 250 Hz and 450 Hz. It is due to the voice-coil inductance resonating with the effective capacitance of the electromechanical part.

We can find the actual power input by multiplying the applied voltage by the real part of the complex input current (the fraction that is in-phase with the voltage). Figure 5 shows the result, where the applied voltage has been set so as to expect an input power of 100 W. We see that in fact that voltage produces over 100 W at any frequency between about 90 Hz and 2 kHz, while at 55 Hz and 10 kHz it is only about 22 W. While the frequency response of the driver may not be very flat, it undoubtedly bears no resemblance to the power curve. Loudspeakers are designed to have a flat frequency response with constant *voltage* input, not constant power.

What does this input power do anyway? Almost all of it just heats up the voice coil; only a tiny percentage is radiated as sound power. It is clear that 'power' is a very misleading concept in the context of loudspeakers. So why was it ever introduced? Even I am not old enough to remember,

but I suspect it dates from before the Rice-Kellogg moving-coil loudspeaker. The much earlier Siemens device was more like a pressure-driver or an earphone. That and the later moving-iron loudspeakers have quite different impedance/frequency characteristics, more resistive, so that input power is a little more meaningful. Even so, these devices were meant to operate from a constant-voltage source.

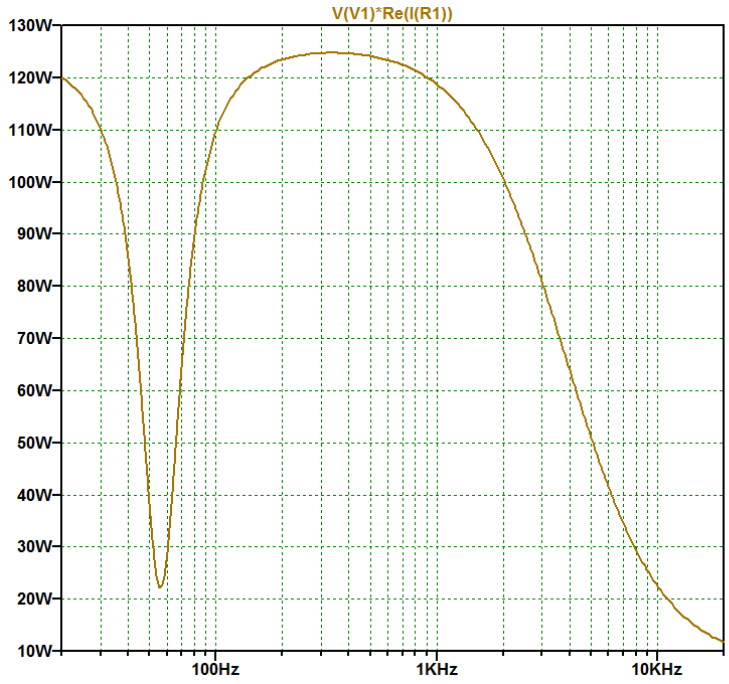


Figure 5 Actual input power to the loudspeaker

What effect does this wild impedance have on clipping? Figure 6 shows that the answer is 'not much'. We can see an uptilt of the flat top at 100 Hz and a downtilt at 10 kHz, but the effects are minor. Note that this is with a good, low impedance power supply.

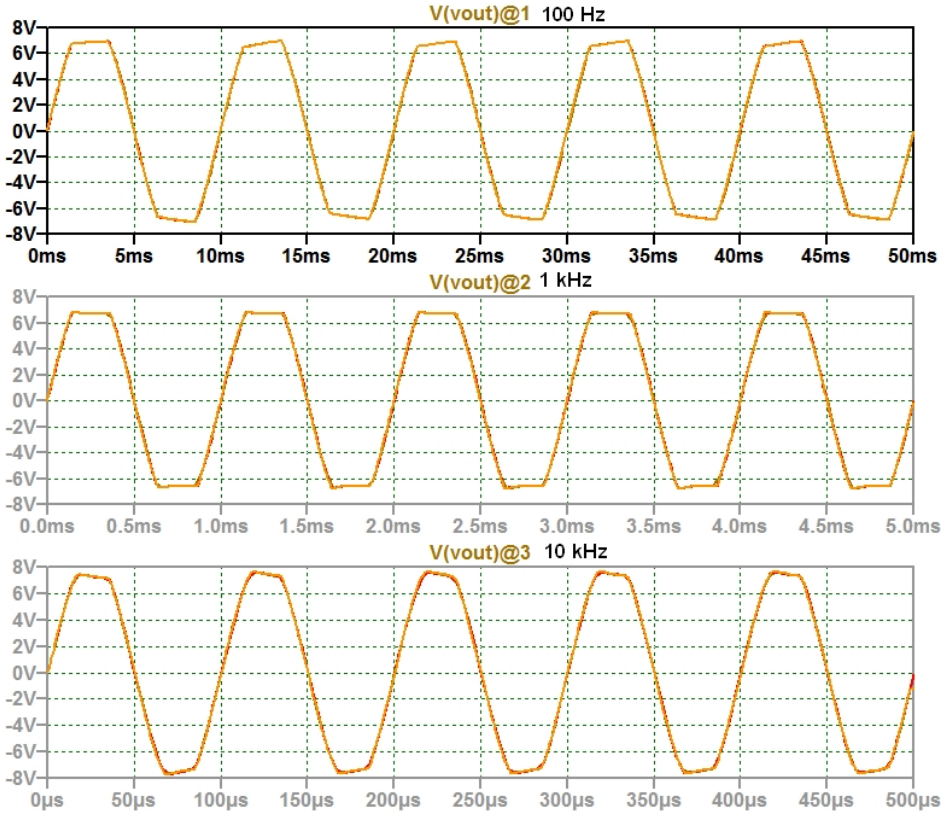


Figure 6 Effect of loudspeaker impedance on clipped waveforms

6 Noise

One could wish that testing with noise signals had never been invented. There are just too many ways of getting a wrong answer, especially one that isn't wrong enough to be easily detected. First, we need to study the nature of the beast, and that isn't so easy. There's white noise and pink noise and even more artistic descriptions, such as 'reddish-brown with blue lines', which was the description of the spectrum of noise emitted by a rather poor loudspeaker.

We have to start with some basic physics. The random motion of atoms in an electrically-conducting material produces a very small random voltage, whose RMS value is given by $V = \sqrt{(4kTB)}$, where k is Boltzmann's constant, 1.38×10^{-23}J/K, T is the absolute temperature in K and B is the bandwidth in Hz. However, the RMS value hides the interesting stuff. What does the randomness do? Well, the amplitude varies according to a formula about probability derived by the mathematician and physicist Gauss. The resulting curve of amplitude probability density (APD) versus amplitude is called the Gaussian curve or the 'normal' curve, because it arises from natural processes.

Probably few people in the audio industry have a way of calculating with the Gauss formula; I had to write a Mathcad sheet that accepts a .wav file as the input and calculates the APD curve and the cumulative probability curve of the signal. The latter is useful for highlighting asymmetry in the signal, and it's surprising how often this exists.

We can plot the probability of a particular (voltage) amplitude occurring against the amplitude scaled in multiples of the 'standard deviation σ (sigma)'. 'What's that?' you ask. Luckily, it's just another name for the RMS value. The way the curve is usually plotted, it has a bell shape, so it's often called 'the bell curve'. But it's far more useful to plot the probability density on a log scale, which not only turns the curve into a relatively familiar parabola, but also shows very clearly how often extreme amplitudes occur. The formula doesn't have an upper limit of amplitude, but that doesn't matter because anything above 5σ is so rare that we can forget it.

Figure 7 shows the formula curve (Reference) and a practical example (Sample), which is the difference between the two noise signals produced by a pair of 12 V avalanche ('Zener') diodes, a technique invented by Bruel & Kjaer. This amplitude-probability distribution (APD) is a very important characteristic of a noise signal. The two curves are nearly coincident. The 'hash' at high amplitudes is due to the short duration (10 s) of the signal. There wasn't enough running time of the .wav file to get a continuous curve of these rare events.

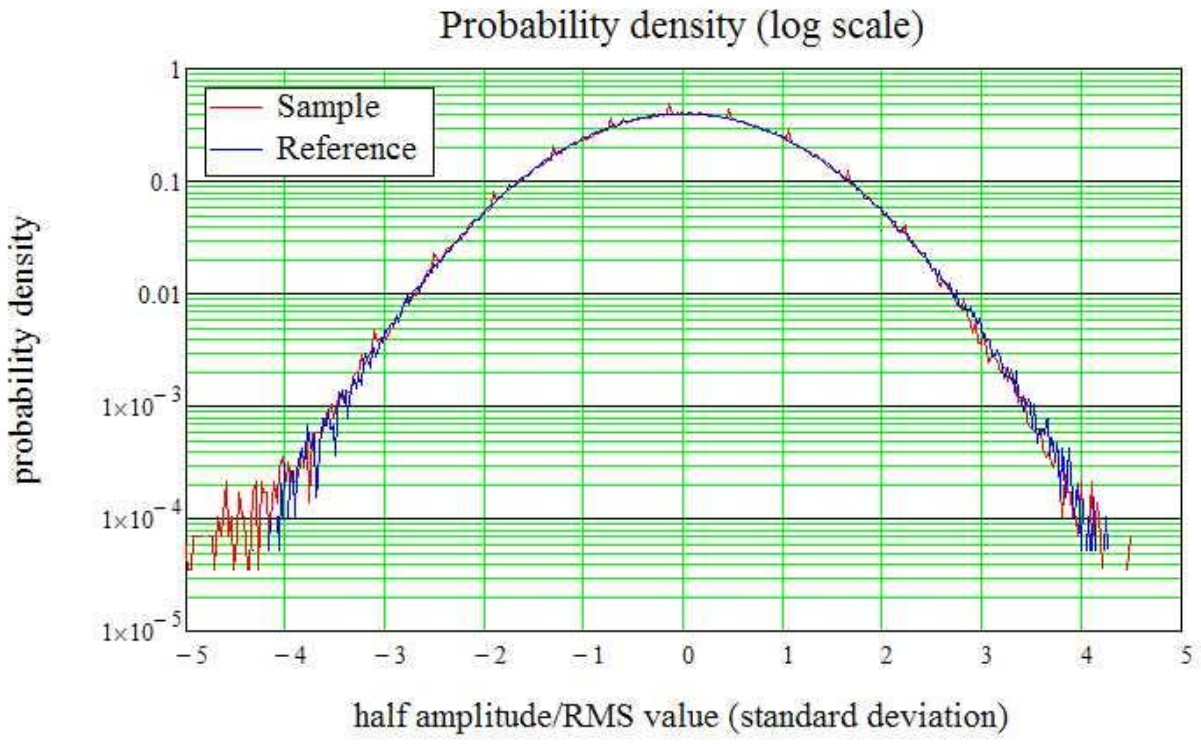


Figure 7 Probability density of Gaussian noise, theoretical and practical

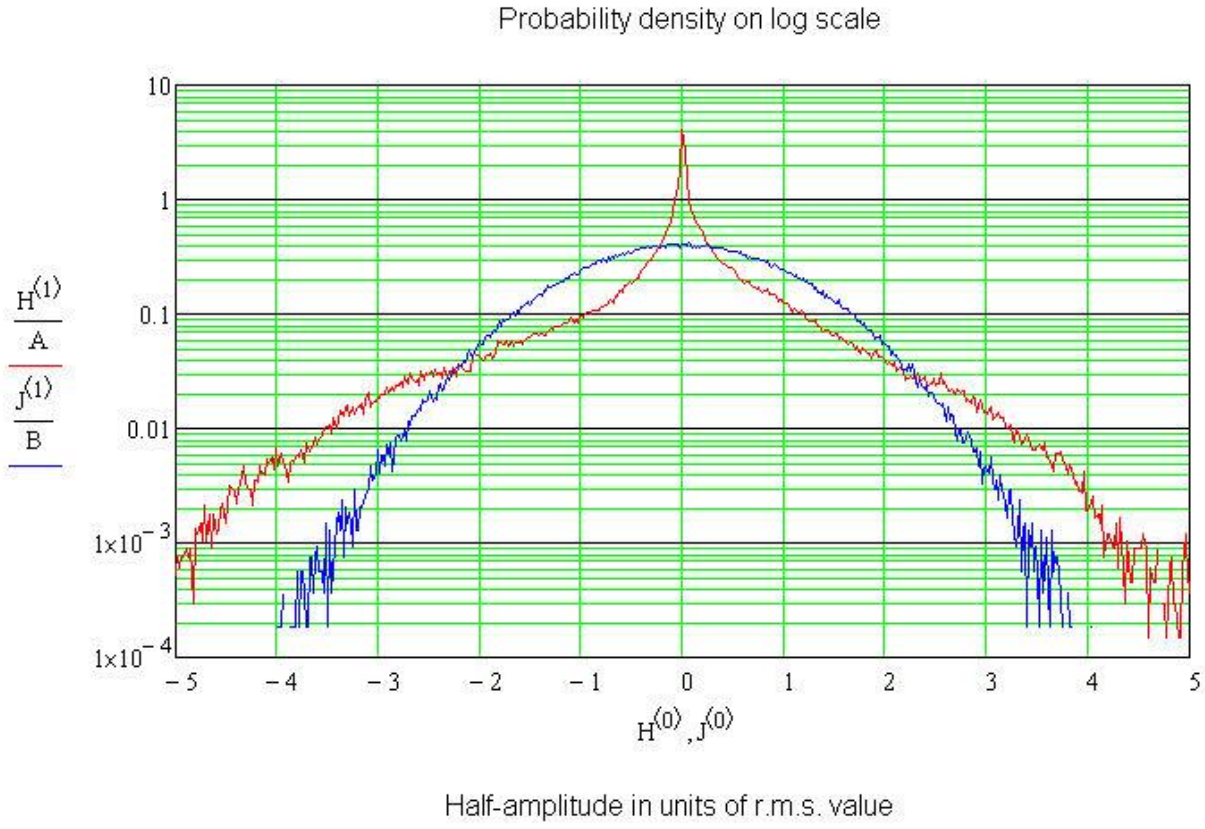
We can see that a half-amplitude (peak value) of 4 times the RMS value has a probability density of around 1 in 10 000. The probability density of a zero value is about 0.4, because the area under the curve (with a *linear* probability axis) – the total probability – must be 1 if the signal exists at all. Another way of looking at this is that the signal spends 99.73 % of the time between -3σ and +3σ.

I measured several noise sources and calculated their APDs for an Audio Engineering Society project. A number of sources showed APDs that differ considerably from the Gauss curve. This is likely to affect the results of measurements, but how much is impossible to say. One source produced 'triangular noise', which is used for dithering digital signals. This would produce very strange results if used for a measurement that expects a Gaussian signal.

Note that this doesn't say anything about the spectrum of the noise signal. That's a separate matter. White noise (like white light) has a flat spectrum if we plot with a constant measuring bandwidth. Pink noise has a flat spectrum if we plot with a constant relative bandwidth, such as 1/12 octave. That poor-quality loudspeaker had a hump in the bass region ('reddish-brown')and several sharp peaks in the treble ('blue lines').

Suppose we have a good noise signal like that shown in Figure 7. We look at its spectrum and we find it goes from below 10 Hz to above 50 kHz. We shouldn't apply to an amplifier signals well outside its rated frequency range, so we apply a 22.5 Hz to 22.5 kHz filter (this is a standard audio band' filter). Unfortunately, we now find that the APD isn't Gaussian any more. Does it matter? It probably matters more if we don't realise it than if we do. It depends what we hope to measure with this noise signal. If we applied a clever process that restored the Gaussian APD, we would find that the bandwidth is no longer 22.5 Hz to 22.5 kHz.

Now for the BIG catch. The APD of real programme signals especially speech, is nothing like Gaussian. This affects important things like the average current drawn from the mains; it's much higher with Gaussian noise than with speech producing the same SPL.



Blue curve – Gaussian noise Red curve – typical speech

Figure 8 Probability densities of Gaussian noise and typical speech

We see that the speech spends a great deal of time at low and even zero amplitudes. On the other hand, it spends more time at high amplitudes than the noise signal does. We have to conclude that Gaussian noise is not a good signal for investigating the dynamic characteristics of audio amplifiers. But that won't stop people using it, unfortunately. There is a better signal – the EHIMA simulated speech signal (not the dreadful ITU one). EHIMA is the European Hearing Instrument Manufacturers Association and the way the signal was developed is well-documented at http://www.ehima.com.

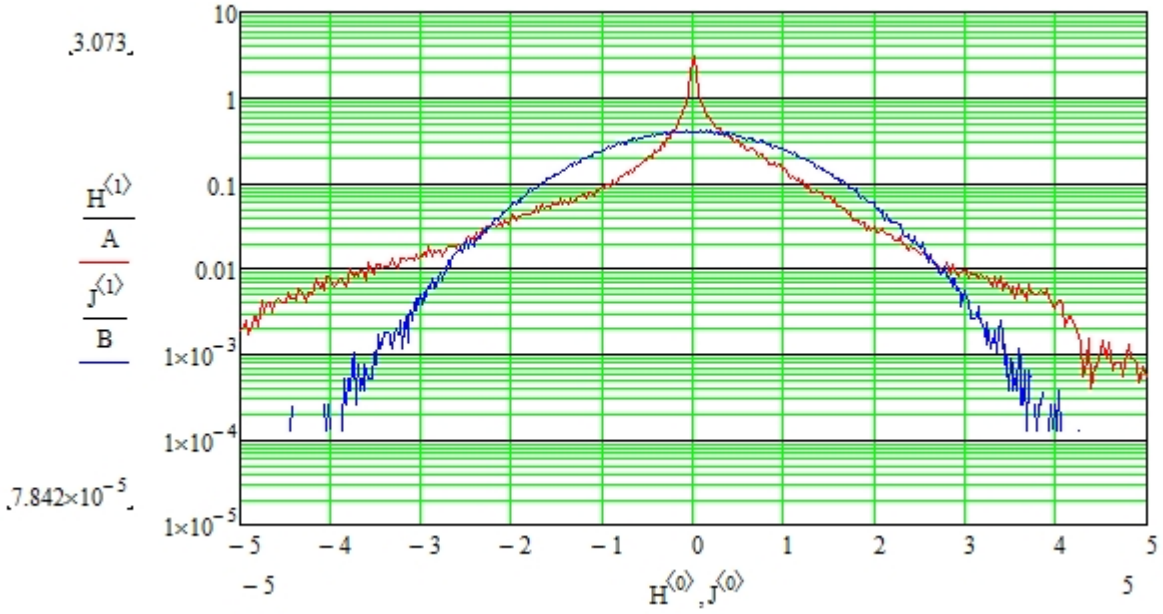


Figure 9 Probability densities of Gaussian noise and EHIMA simulated speech

7 Effects of signals with different APDs

Considering that what limits the output voltage of a well-designed amplifier is the clipping, we can see that a signal whose APD shows that it spends much time at low voltages loses less energy from clipping that one that spends much time at high voltages, and a most extreme case of this is a sine wave, whose APD has two big spikes at amplitudes $\pm\sqrt{2}\sigma$ and not much in between. Thus if we measure the RMS output voltage, we can put in a much larger signal of the first type, such as speech, before the RMS output voltage levels off due the amount of clipping. If we are experimenting in this area, we need to look at input and output waveforms as well as RMS voltages, which simply don't tell the whole story.

8 Crest factor

Crest factor is the ratio of the peak value of a signal to its RMS value. It may be given as a number or in decibels. Not too difficult if the amplitude doesn't fluctuate, as, for example, the diode current in a rectifier with capacitive filter and a constant load. But we are dealing with fluctuating signals as well as sine waves. For fluctuating signals, time becomes important. Over what period do we average, in order to get the root-*mean*-square value? How long do we wait to see if an improbable extra-high peak occurs? Without controlling, and even specifying, these times, the value we adopt as the crest factor is not well determined.

For noise signals, we can average for a long time so as to get a stable, and thus dependable, RMS value, but if we try that with speech signals, the average tends to fall continuously as the averaging time increases. For real electrically-generated signals, there is an upper limit to the peak value, determined (directly or indirectly) by the supply voltage of whatever is producing the signal, but the signal may not in fact ever get as large as that. On the other hand, it might be clipped at the supply voltage, which produces a spike on the APD curve (a pair of spikes if both polarities are clipped). In that case, the APD curve gives us a reliable peak value, but only if we have the resources to plot it. In other circumstances, any stated value must be regarded as approximate, and we shouldn't use it to make critical deductions or decisions.

Clipping a signal reduces the peak value and therefore reduces the crest factor. Another way of looking at this is that different RMS output voltages are produced by the same RMS input voltage of signals with different crest factors, if clipping occurs. The gain of the amplifier appears to depend on the type of input signal, but that is not at all what is happening. Observation of output waveforms with an oscilloscope is indicated.

9 Class G and H amplifiers

These amplifiers have controlled-voltage DC supplies, so that the voltage increases to accommodate high-amplitude signals. But clearly there is a limit to this increase, so clipping can still occur. With good design, the processing has no detectable effect on the output waveform up to the clipping point, so these amplifiers handle large signals in the same way as conventional Class AB amplifiers.

10 Mains power consumption

To show how the nature of the signal affects the amplifier power consumption from the mains supply, a series of measurements were made, with the following results.

Table 1 Power consumption of an amplifier with different signals

Amplifier output voltage at clipping, 1 kHz sine wave: 20.2 V. This amplifier has an unregulated power supply, but its impedance is very low.

Ratio of peak voltage to RMS voltage (using oscilloscope): pink noise: 3.3, (this particular) speech 4.0.

Headroom is the ratio, expressed in decibels, of the output voltage at clipping to the actual output voltage.

Amplifier output voltage V (RMS)	Test signal	Mains voltage V	Current (RMS) A	True power W	Notes
20.2	Sine	242	0.53	107	
	Pink noise	242	0.65	130	Very clipped
	Speech	242			Cannot reach 20.2 V: too heavily clipped
5.05 12 dB headroom	Sine	242	0.21	44	
	Pink noise	242	0.21	44	
	Speech	242	0.21 (0.16 to 0.33)	26 to 52	
10.1 6 dB headroom	Sine	242	0.31	64	
	Pink noise	243	0.31	64 (59 to 74)	Clipping visible
	Speech	243	0.14 to 0.46	27 to 100	Clipping visible
14.3 3 dB headroom	Sine	243	0.39	80	
	Pink noise	242	0.39 (0.30 to 0.49)	83 (76 to 91)	Clipping visible
	Speech	243	0.38 (0.15 to 0.65)	27 to 116	Very clipped

We can see from the table that pink noise and a sine-wave signal demand about the same mains power. The power fluctuates a lot with speech signals, but the average is quite near the lower figure reported, half to one-third of the power required for the other signals. This is quite an important result, and it's logical, considering that the APD curve shows the high proportion of the time that the signal spends at near-zero amplitude.

Time for another very surprising result. 'Clean' clipping (free from effects such as temporary DC shift due to bad power supply design) has only a very limited effect on the intelligibility of speech, especially if accompanied by a judicious attenuation of low frequencies. This applies even for large reductions, such as 15 dB (5.6 times) of the peak voltage. This can't be predicted, it can only be shown by experiment.

11 How much headroom do we need?

If we wanted to completely avoid clipping, we would need to accommodate peaks of the speech signal at least four times the RMS value, i.e. 12 dB of headroom. Since we measure the target SPL of a system (that we need to reach for the system to be acceptable) with an RMS meter, that correlates with the RMS output voltage, but the amplifier would have a maximum output voltage four times greater, giving an SPL 12 dB higher and implying that the rated power of the amplifier has to be sixteen times that which we need to get our target SPL. Can you hear the budget groaning?

The resolution is that sound systems operate with the amplifier clipping (unless the amplifier has an anti-clipping circuit, which just depresses the peaks of the signal more gently than clipping does). How much clipping can we tolerate? As always, it depends. If we are dealing only with speech, and the amplifier is well-behaved in clipping, 3 dB of headroom may well be enough. It's easy to tell, if you just listen to the system. If the headroom is insufficient, the sound will not be nice.

Annex
The Gauss formula

The general Gauss formula is:

$$P = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{(x-\mu)}{\sigma\sqrt{2}})^2}$$

where:

P is the probability density (see Note below)
σ is the standard deviation
x is the amplitude
μ is the mean value.

For our purposes we can set the mean value to zero (no DC component in the signal) and $\sigma = 1$, which simplifies the formula a good deal:

$$P = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

If we now take logs (to base e):

$$\ln P = \ln\frac{1}{\sqrt{2\pi}} - \frac{x^2}{2}$$

That's an upside-down parabola, as shown in Figure 7.

Note on probability density

This is an explanation, not a rigorous mathematical statement.

The probability function is rather special, but so is the more familiar spectrum function. In each case, the function is defined over small intervals, not at contiguous discrete points as in the usual $y = f(x)$ curves. For the spectrum, the interval is the measurement bandwidth, which might be 10 Hz or 1/12 octave. The spectrum level for a zero bandwidth (i.e. a single frequency) is

obviously zero, and you would have to wait for an infinite time to prove it. The total signal voltage is the *area under the spectrum curve, plotted on linear axes.* So the voltage between any two frequencies is the area under that slice of the spectrum curve. Spectrum curves should really be called 'spectral density curves', but they rarely are.

In the same way, the probability function is defined over small intervals of amplitudes, and the probability of a specific amplitude is actually zero, even though you can read a value from the curve. This sounds crazy, but it comes from the fact that the *area under the curve* (with linear axes) represents the total probability of 1. The probability that x is between a and b is the area under that slice of the curve, and obviously if $a = b$, the probability is zero. That is why these curves are called probability density curves.