A perceptual approach on clipping and saturation

Stefania Barbati
stefania@simulanalog.org

Thomas Serafini
serafini.thomas@unimo.it
http://www.simulanalog.org

1 Introduction

This article is about the analysis of the most common form of nonlinearities and distortions in the acoustical field. Both negative aspects and useful application will be shown. But the main topic of the article is to disprove many wrong ideas about distortions, that we usually think true without discussing them.

1.1 Definitions

Let's give some definitions used in this article. A **transfer function** is a function $f : R \longrightarrow R$ where R is the domain of the signal. A transfer function is an object that alters a signal by a relation between the instant input value and the instant output value. If the function is linear, it is called **amplifier**; if it is not linear, it is called **non linear function** or **nonlinearity** or **waveshaper**. A non linear function $f :]-\infty, +\infty[\longrightarrow [a, b]$ is called **saturation** or **softclip** if satisfies these conditions:

1. f is surjectiv
2. $\lim_{x\to-\infty} f(x) = a$ and $\lim_{x\to+\infty} f(x) = b$
3. $\lim_{x\to\pm\infty} f'(x) = 0$
4. $\begin{cases} f''(x) > 0 & for \quad x < 0 \\ f''(x) = 0 & for \quad x = 0 \\ f''(x) < 0 & for \quad x > 0 \end{cases}$

A saturation is called clipping if it is defined like:

$$f(x) = \begin{cases} a & for \quad x \leq a \\ x & for \quad a < x < b \\ b & for \quad x \geq b \end{cases}$$

so, it is perfectly linear in the range $[a, b]$, and clips (cuts) the signal if it is out of the range.

1.2 Harmonic analysis

We usually study a transfer function analyzing its behaviour on a sinusoidal signal: in fact, feeding a nonlinearity with a sine source, results in a signal made by an harmonic series whose frequency is integer multiple of the original sine. If we want to calculate the amplitude of the harmonics, we only need to perform the Fourier analysis over a period of the waveform out of the nonlinearity. In the next section we are going to study the relation between the shape of the nonlinearity and its sonic result. Here is one of the most common error: many of us suppose that these rules are valid for all kind of signals, and a transfer function is often choosen according to the spectral content that we want to obtain; for example, if we want second harmonic into a mix, we may suppose to choose an even transfer function. But this is true only for a sinusoidal signal, with no other partials than the foundamental. For example, if we have a signal with two components, one at 1KHz and the other at 1.2KHz, we may suppose that a nonlinearity produces two harmonic series, the first with partials multiple of 1KHz, and the other with partials multiple of 1.2KHz, so the result is a signal with component at 1KHz, 1.2KHz, 2KHz, 2.4KHz, 3KHz, 3.6KHz, etc But this is absolutely false! The real signal has also the intermodulation products, so there are inharmonic components and partials extended into all the spectrum, also under 1KHz. A formal demonstration of this fact is quite complex and requires a lot of mathematics, so I'm only giving an intuitive explanation. Take the example above: two sinusoids, one at 1KHz and the other at 1.2KHz, make a signal whose period is a common multiple of the two periods; so, in this particular case, the first sinusoid completes 5 cycles in the same time the second one completes 6 cycles, so the resulting signal has a frequency of $\frac{1000}{5} = \frac{1200}{6} = 200Hz$. So, if you apply a nonlinearity to this signal, you should expect partials at multiples of 200Hz. Now think of a signal like a complete mix, that is much more complex than two sinusoids; a mix is not a periodic signal, that is the period has infinite time or the base frequency is 0, so a nonlinearity creates harmon-

ics on all frequencies. This is not a formal explanation of intermodulation, but can give a rough idea of this effect.

2 Hard and soft saturations

Let's begin with one of the most important characteristic of saturations, the sharpness of the curve, or what many people call hard or soft. Clipping is certainly the hardest saturation, because the transition from the linear zone to the nonlinear one is not progressive; the function is perfectly linear under the threshold, but cannot exceed it. A saturation is called soft if the transition from the linear to the nonlinear zone occours gradually or, in other words, if the function is already nonlinear before the threshold. Here is an example of a soft saturation:

$$\begin{cases} \frac{1}{2} & for \quad x \geq 1 \\ -\frac{1}{2}x^2 + x & for \quad 0 \leq x < 1 \\ \frac{1}{2}x^2 + x & for \quad -1 < x < 0 \\ -\frac{1}{2} & for \quad x \leq -1 \end{cases}$$

Here, the function reaches the threshold level very gradually, with a parabolic law.

Now let's compare hard and soft saturations from a perceptual point of view. We are especially interested in finding a creative use of them. It is quite easy to understand that the clipping...

1. creates more evident effects on the sound, because of its sharp shape;
2. does not create harmonics for low level signals (inside the range [a, b]), but when the signal exceeds the threshold it immediately begins to produce a lot of harmonics.
3. The more the saturation is hard, the more the series of harmonic's amplitudes decrease slower.
4. The more the saturation is soft, the more it produces harmonics also for low level signal.

Not necessary these are negative characteristics: for some application, hard clipping is much more appropriate than soft saturation, which can produce many unwanted effects. There are two different kind of applications where saturations are useful: limiting/mastering and compression/warming. Next sections are about these classes of application.

2.1 Limiting / Mastering

In these applications, soft saturation presents many unwanted effects; on the other side, hard clipping has many interesting charatteristics that make it very useful. In fact, limiting a mix does not mean distorting it: on the contrary, it means that we should keep it as cleaner and similar to the original as possible (that is linear), and on the same time it cannot exceed a given threshold. That's why hard clipping is perfect for this task. First of all, clipping is perfectly linear in all its range: this means that in the greater part of the mix, the signal is unaltered and perfectly equal to the original. Then, its action is quite strong on the peaks of the signal that exceed the threshold. Consider that these peaks usually correspond with fast transients: consider also that a typical effect of analog (especially tube) circuits is a bigger harmonic production on the transients. This effect is perceived as an increase of the punch in the mix and is usually wanted by the recording or mastering engineers on some musical styles, like rock or dance. We can find similar processes in the studios: exciters can be a first example, because their working principle is the production of harmonics on the transients. Another example is when percussive sounds like drums are first compressed, in order to emphatize the attacks, and then saturated to get more harmonics on the attacks. On the contrary, limiting a mix with soft saturation generates too much harmonics of the wrong type and the acoustical result is a dirty and confused sound. Let's see why this happens: a complete mix is a very complex sound source, because harmonics extend on all the spectrum and there are many uncorrelated and inharmonic components. Applying a nonlinearity on a mix creates a great amount of intermodulation products, that are uncorrelated partials; these partials are perceived more like noise than harmonic extension of the mix. If a soft saturation is used, also low levels signals are in the nonlinear zone, and so the harmonic production is on the whole mix and not only on the transients; from a perceptual point of view, this is like a constant noise that removes crispness and definition from the mix.

2.2 Compression / Warming

This is the opposite situation than the previous section. Here, the nonlinearity is not applied on a mix (i.e. a complex signal), but usually on a single instrument or on a simpler and more regular source. Most of the harmonics in the signal are well correlated, so a nonlinearity produces many correlated harmonics, that give the sound a warmer feeling. It is important to remember that warming does not mean distortion, so the harmonic production should stop before the feeling of distortion. Usually on real instrument, a louder note has a wider spectrum. Thanks to its soft curve, a soft saturation produces gradually more harmonics when the level of the sound increases, just like real instrument. On the contrary, a hard clipping cannot reproduce this effect, because the production of harmonics begins suddenly when the level exceeds the threshold.

3 Symmetric / Asymmetric

A transfer function f is called symmetric if:

$$f(-x) = -f(x)$$

This means that a symmetric transfer functions has the same effects both on the positive and on the negative halfwaves of the signal. Feeding a symmetric transfer function with a sine wave, only odd order harmonics are produced: for example, with a 100Hz sinusoid, the produced harmonics are on 300Hz, 500Hz, 700Hz, 900Hz, etc On the other side, an asymmetric transfer function produces harmonics of all orders, on a sinusoid. Many people think that "even order harmonics are more musical correlated than odd order ones, because they represent an octave, two octave, two octave and a fifth, three octave... and so on. So, asymmetric transfer functions are better because they produces these 'musical' harmonics". But this is not completely true, and now we should try to understand why. First of all, they produces a greater number of harmonics than symmetric transfer functions: in fact, symmetric ones generates only odd order harmonics, but asymmetric ones generates both odd and even order harmonics, that is a bigger number of harmonics. We have already seen that a high harmonic production on complex sounds, like mixes, creates many intermodulation products that are uncorrelated from the original harmonic content, so the resulting sound is more dirty and confused. After that, an asymmetrical function is not suitable for limiting at all. The aim of the limiting is containing the signal into a given range (usually the symmetric digital 0dB). With an asymmetric transfer function, the less compressed side have to remain into its fixed limit, but doing so the more compressed side is limited below its threshold and not all the available dynamic is used. The consequence is that loudness is not maximized because not all the potential headroom is used. Finally, an asymmetric transfer function produces a lot of DC component; this is obvious thinking that one side of the signal is more compressed than the other, so there is more power on one side. There are some applications where asymmetrical functions are useful, especially if used very carefully on simple sounds in warming applications. But my personal opinion is that thinking asymmetric = tube like = good sound, is not true in a general case.

4 Static / Dynamic

Many people think that distortion = waveshaping; in their opinion, each type of distortions can be simulated by choosing an appropriate transfer function. But usually these people ask themselves why digital distortions do not sound like analog ones! The answer is that waveshaping is only one kind of distortions, probably the more simple and intuitive. But analog world is much more complex and there are many other type of distortion that we should consider. The way a circuit distorts a signal is not invariant over time and dynamic levels; many circuits have a "memory effect" which discriminates transients and dynamic changes. According to them, the circuit alters it type of distortion. This is not a feature intentionally implemented in the circuit design, but in an intrinsic characteristic of some components. As an example, we may consider the coupling between two tube stages: when the grid to cathode voltage is above 0V and the triode saturates the positive halfwave, the grid resistance get lower. So, only in the positive halfwave, the coupling capacitor charges faster because the RC product lowers; in the negative halfwave the RC time constant returns to its original value, so the capacitor discharges slower. The result is a slow variable DC component on the grid, moudulated by the dynamic level of the signal. This dynamically changes the operation point and the transfer function of the tube according to the level of the sigal. My opinion is that this is one of the most important feature that should be considered when working on saturations. The acoustical result is a greater production of harmonics on the transients, and this is usually perceived as more punch on the sound. We should note an important fact: punch is not the production of harmonics on transients. Instead, punch is a psychoacoustic phenomena, is a musical term that describes the perception of the sound and there is no mathematical definition of "punch". The only thing we know is that listening tests have proved that a production of harmonics on the transients is perceived as punch, but this is not the definition of punch.

5 DC filtering

A nonlinear process on a signal usually creates DC component. Asymmetrical transfer functions produce a big amount of DC, but also symmetrical ones generates a little amount of DC. Usually this component should be filtered out, using a high pass tuned around 20Hz. But there is a situation where this operation seems impossible: in limiting application. In these applications, the transfer function have to keep the signal inside a given range; but high pass filtering the saturated signal produces peaks over the threshold, due to the Gibbs effect, so the signal will be no more limited into the range. Many listening tests have suggested a particular chain of processing that do not produce any negative artifacts on the musical signal: the signal is initially limited, then high pass filtered around 25Hz and finally clipped a second time to push it back into the limit range (usually of the digital 0dB). One may suppose that this double distortion process creates acoustical artifacts, but listening tests have proved the opposite: in fact, if the extremely low frequencies won't be filtered out, the sound will loose definition and presence in its low end.

6 Emphasis

Very often we would like that the action of a nonlinearity is more evident on a particular range of frequencies but, in the same time, that the overall frequency response of the system is flat. For example, in mastering applications high frequencies can be distorted more than the low; that's because the harmonics created on the high frequencies lays mainly out of the audible range. Instead, a little distortion on the bass frequencies is immediately perceptible, because it falls into the most sensible frequency zone of the ear. We can obtain frequency dependent distortions by applying, before the saturation, an equalizer that emphatizes the frequencies we want to distort more; finally, after the saturation, the opposite equalization have to be performed. Using this chain, the frequency response of the system is perfectly flat when the signal is in the linear range. For example, if we'd like 6dB more distortion over 5KHz, we could place a high shelf filter with +6dB at 5KHz before the non linear function, and a high shelf with -6dB at 5KHz after the nonlinearity. We can find a similar chain into analog tape recorders. The high frequencies of the signal are emphatized before writing on the tape: then the tape saturate (quite symmetrically) the emphatized signal and finally, during the playback, de-emphasis is performed to bring the signal to its original spectral balance. Many recording engineers like this kind of processing, and some of them talk about warming.

7 Aknowledgments

I'd like to tank Davide Barbi and Sascha Eversmeier for our interesting talks about this paper.

References

[1] Charles Rydel, "Simulation of Electron Tubes with Spice", in Preprint no.3965 of AES 98th Convention 1995

[2] Menno van der Veen, "Modeling Power Tubes and Their Interaction with Output Transformers", in Preprint 4643 of AES 104th Convention 1998

[3] Foti Frank, "Aliasing Distortion in Digital Dynamics Processing, the Cause, Effect, and Method for Measuring It: The Story of 'Digital Grunge!' ", in Preprint no.4971 of AES 106th Convention 1999

[4] H. Yahiro, A. Kameoka, M. Kuriyagawa, "Psychological Evaluation of Nonlinear Distortion", in Preprint no.665 of AES 37th Convention

[5] S. Ohashi, T. Sampei, E. Okuda, "Investigations of Various Forms of Distortion Inherent in Transistor Amplifiers", in Preprint no.766 of AES 39th Convention

[6] E. Leinonen, M. Otala, J. Curl, "Method for Measuring Transient Intermodulation Distortion (TIM)", AES Journal Vol 25, Issue 4, Page 170

[7] D. Preis, P. J. Bloom, "Perception of Phase Distortion in Anti-Alias Filters", AES Journal 1984 Vol 32, Issue 11, Page 842

[8] Richard C. Cabot, "Perception of Nonlinear Distortion", in AES Preprints 1984