Non-linear distortion in audio amplifiers

Why do some amplifiers pass static distortion tests but fail listening tests?

by M. Otala, *Technical Research Centre, Oulu, Finland*

The debate about amplifier distortion and especially its audibility has always been an interesting subject. Most of us still remember the battle over triodes and pentodes, and a few years ago such epithets as "transistor sound" were discussed intensely. Right now we are in the middle of "operational amplifier sound", and although these negative attributes may seem ridiculous at first glance, there really seems to be some clearly audible differences. These differences must be "distortion", whatever that may then mean.

It is a commonplace to divide distortion in amplifiers into two classes: linear distortions, i.e. linear departures from straight frequency or phase characteristic, and non-linear distortions, i.e. distortions caused by non-linear amplitude relationship between the input and output signals. This article concentrates on the last-mentioned form of distortion and divides it into two groups according to their dependence on the signal

- static non-linear distortion, dependent solely on the amplitude of the signal, and
- dynamic non-linear distortion, dependent not only on the amplitude but also on the time properties or frequency composition of the signal.

Historical perspective

In the early valve era the cost of gain was high. This led to the use of few active devices and careful design to yield acceptable harmonic and intermodulation distortion figures. When the benefits of feedback were discovered, it was applied mostly locally. The presence of an output transformer with its stray reactances made the amplifier transfer function so complicated and dependent on the momentary signal and load conditions at high frequencies that heavy overall feedback could not be used without loss of stability. The average overall feedback varied between 15 and 30dB, and the static harmonic and intermodulation distortion were the primary sources of audible amplifier quality impairment.

The introduction of transistors and especially the transformerless amplifier circuits permitted the use of heavy overall feedback. This led to the unwarranted myth of the amplifier being the better, the higher the feedback. The following advantages were attributed to the use of feedback

- static distortions decreased to practically zero
- bandwidth of the amplifier increased
- output impedance of the amplifier decreased and hence the damping factor increased

The decreasing cost of components and the trend toward monolithic integration made possible the use of almost-unlimited gain resources, and consequently the main trend in the design philosophy has been the use of very high open-loop gain and high values of feedback.

This trend has been further intensified by the use of operational amplifiers, which more and more are finding their way into audio equipment as low-level amplifiers and power amplifier drivers. The need to minimize the size, weight and power dissipation of amplifiers also led to another trend: the minimization of the class A operation region of an amplifier. The result is cross-over distortion, which sounds ghastly and is difficult to eliminate with feedback or any circuit tricks.

Those two effects, the overdose of feedback, causing dynamic non-linear distortion, and the almost class B operation causing near-incurable cross-over distortion, seem to be the main distortion problems of present-day audio amplifiers.

Static non-linear distortion

Every stage of an amplifier has a more or less non-linear transfer function. Fig. 1 shows the typical static non-linearities usually encountered in audio amplifiers, namely s-type, cross-over and clipping distortions.

S-type non-linearity. There are numerous reasons for the s-type non-linearity. In the case of transistors it may, for instance, be caused by the non-linear dependence of current gain, *versus* collector current and voltage, by the non-linear base-emitter voltage characteristic, or by possible avalanche-type

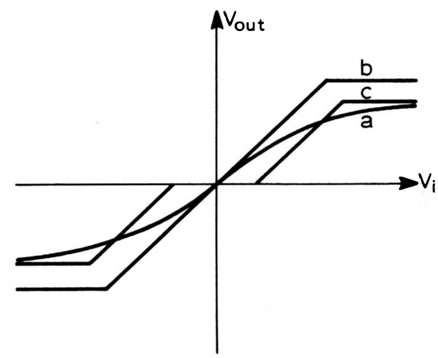

Fig. 1. Different kinds of static non-linear distortions (a) s-type, (b) clipping and (c) cross-over.

collector current non-linearity due to collector-emitter voltage. In the case of vacuum tubes, the list of sources for non-linearity includes the space-charge effects around the control grid, the change of mutual conductance and anode resistance as function of voltage, the possible negative impedance contribution of screen grid in beam tetrodes and pentodes, etc.

On the circuit side the most notable method of minimizing the non-linearity is the choice of interstage resistors to ensure that the stage interface transfer function is as linear as possible. If transformers are used, their non-linearities are important too. All of these sources of s-type non-linearity are well understood and design rules exist for their minimization. The effects are, however, too numerous to be considered here. Furthermore, the remaining s-type non-linearities can easily be decreased with the use of local or overall feedback.

Cross-over distortion. The operation of power amplifiers in class B presents some important special problems. The first is cross-over distortion, and the second the time asymmetry of the amplifier halves, Fig. 2. Both occur around the class B transition from one circuit half to another. The source of these distortions is the decrease of the gain of each half to almost zero at almost zero collector current, and the different transition frequency behaviour of each half. In the cross-over region, therefore, the open-loop gain of the amplifier drops drastically. Feedback has little effect on this type of distortion, as there is no open-loop gain available for the feedback. The only possibility is to allow sufficient quiescent current to ensure the full gain at all times. These two forms of distortion are very clearly audible, probably because they generate harmonic and intermodulation products of high odd order. In the case of harmonic products, the high order components are non-musical and therefore annoying. In the case of intermodulation products, a high order means a multiplicity of products falling within the audio band. Being non-musical, the musical masking of these kinds of products is small. However, the sensitivity of the ear may also stem from the strong phase modulation they introduce in heavily feedbacked amplifiers. The details of this effect are outlined later in the section on dynamic non-linear distortion.

Clipping occurs when an amplifier is overloaded. Therefore it is not an operational non-linearity in the proper sense of the definition. However, as overloading peaks do exist in usual programme material, the amplifier overload performance becomes important. The audibility of clipping is dependent on the clipping mechanism, soft s-type clipping being less audible than hard limiting, which may be aggravated further by saturation recovery effects. This increased audibility depends again on the generation of higher-order harmonic and intermodulation distortion products.

It would be desirable to "soften" the clipping. The problem is, however, that the overall feedback effectively linearizes the clipping, making it hard, and may also cause an internal excess drive signal within the feedback loop during the clipping, thus aggravating the saturation problems and delaying recovery. The desire for a soft clipping and the present use of feedback are therefore incompatible, and it remains to be seen which one will be considered more important in the future.

Static distortion versus feedback

Suppose that in a given circuit all the possible means for minimizing distortion *in situ* have been used by selecting linear active devices, by choosing optimum load and generator impedances for all stages, and by careful selection of the working points. Suppose further that so far no feedback has been used. The interesting question then arises: whether one should use local feedback stage by stage, or overall feedback around the whole amplifier to reduce remaining static distortion. Most present-day amplifiers seem to be constructed according to the last men-

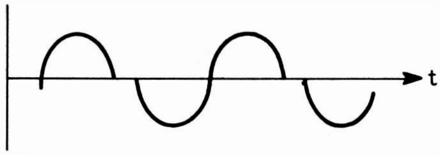

Fig. 2. Cross-over distortion caused by time asymmetry of the class B amplifier halves.

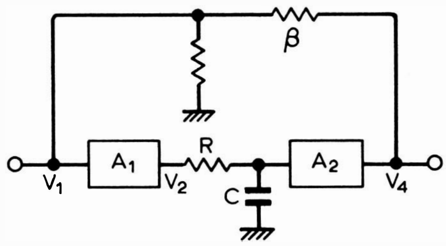

Fig. 3. Division of a feedback amplifier incorporating the driver A_1, the output stage A_2, the compensation network RC and the feedback network β.

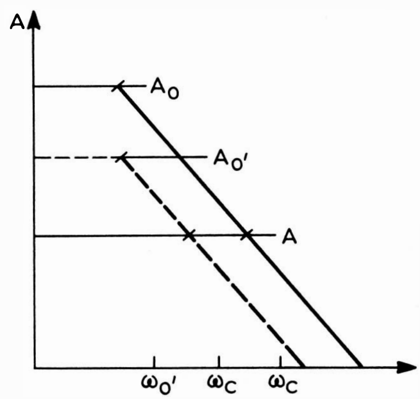

Fig. 4. Bode plot of the feedback amplifier.

tioned principle, i.e. the main design objective has been to realize as high (and often very non-linear) a gain as possible and to rely on overall feedback to make the amplifier behave correctly.

The use of local feedback has some drawbacks which make its use unpopular

- it increases the number of parts in the amplifier
- if the amplifier uses i.cs, linear unbypassed emitter resistors may be difficult to manufacture
- local feedback often limits the available voltage swing of the stage (Crucial at driver stages and may necessitate separate power supplies for them)
- large unbypassed resistors at the output transistor emitters may severely limit output power

However, local feedback has some advantages:

- it linearizes and stabilizes each stage separately, eliminating certain difficult cross-coupling linearity and stability troubles between stages.
- it decreases the effect of individual device tolerances, which may cause some working point problems, especially in d.c.-coupled multi-stage amplifiers.
- it increases the cut-off frequency of the stage

The last remark is important. For the same total gain, the use of overall feedback alone yields the same distortion figures as the use of local feedback alone but with one significant exception: whereas local feedback increases the usable frequency range of the amplifier, the overall feedback usually decreases it. This apparent contradiction may be explained as follows:

To ensure stability, the amplifier open-loop frequency response must have a −6dB/octave roll off. For heavy overall feedback, the amplifier must then be frequency compensated to eliminate the influence of the second, third, etc. poles of the transfer function[1]. If overall feedback is increased, this compensation must be made proportionally heavier, resulting in the closed-loop small-signal frequency response remaining the same. The generally held belief that overall feedback increases the small-signal frequency range is thus invalid in the case of multiple-stage amplifiers. However, the large-signal frequency range usually decreases with increasing feedback. This is caused by the heavier frequency compensation requiring more error signal headroom from the driver stages. If there is not much of this headroom available, and such is usually the case, the driver stages will clip at proportionally lower frequency as the compensation is made heavier. High overall feedback therefore has the tendency of decreasing the power-bandwidth of an amplifier.

The optimum choice with present-day components is probably to use all the possible local linearization methods available, and thereafter to use local feedback until the open-loop large-signal total harmonic distortion is around 0.2 to 2%. Moderate overall feedback is then added, the optimum value being around 20 to 40dB. It seems possible with this kind of technique to obtain harmonic distortion figures as low as 0.05% without increased risk to dynamic non-linear distortions.

Dynamic non-linear distortions

If the frequency content or the time properties of the input signal affect the transfer function of the amplifier, the resulting non-linearities may be called dynamic. We know at present of at least one dynamic distortion of this kind, namely the transient intermodulation distortion (t.i.m.) which has been described in detail elsewhere[2]. It stems from overall feedback in the following way.

Consider an amplifier with heavy feedback, and consequently heavy compensation, shown in Fig.3, having the Bode plot of Fig.4. The raw, open-loop gain is A_0 and the corresponding open-loop upper cut-off frequency is ω_0, typically 5 to 500Hz. The open-loop transfer function of A_0 is shown in Fig.5.

Now consider an input signal consisting of a transient and a sinusoid. The error voltage V_2 is proportional in amplitude to the frequency of V_1 (Fig.6) due to the compensation network RC. Suppose that the input transient has sufficiently low rise time to let V_2 excurse to V_2'. The incremental open-loop gain now drops to A_0', also shown in Fig.4 with a dashed line. If the feedback is large, the closed-loop gain A is not affected, but the closed-loop upper cut-off frequency ω_c (typically 20 to 200kHz) drops momentarily one or two decades to ω_c' during the rise of the transient. This causes phase modulation of the sinusoid if it is smaller in frequency than ω_c', and combined amplitude and phase modulation of the sinusoid if it is between ω_c' and ω_c in frequency. In both cases, the phase and amplitude modulations give rise to interference components between the transient and the sinusoid, thereby creating non-harmonic audible components in V_4, the output signal[3]. In an extreme case, driver A_1 is driven into saturation and A_0 drops to zero. This corresponds to momentary 100% intermodulation distortion of the sinusoid.

This effect is phenomenologically equivalent to intermodulation distortion caused by rapidly sweeping the upper cut-off frequency of the amplifier in synchronism with the frequency content of the input signal. Whereas t.i.m. is principally caused by the overall feedback, similar effects occur with the so-called dynamic noise limiters, although there the speed of the sweep is limited. A similar effect occurs in power output transistors, where the cut-off frequency f_β depends on the instantaneous collector current and collector-emitter voltage.

Heavy cross-over distortion causes almost identical phase and/or amplitude modulation effects to those produced by t.i.m. although in principle it is a static non-linearity. This is due to the fact that it causes the same kind of momentary variation in the open-loop gain.

Amplifier distortion budget

The distortion compromise that a designer must make in designing an amplifier consists of at least the following parts:

1. The smooth, s-type non-linearity of the transfer function caused by device and circuit non-linearities. These are easy to correct to a certain extent by local feedback, optimum load and generator impedances and by overall feedback. Usually this type of distortion is neither difficult to handle nor severely audible, the only prerequisite being the necessity of a few extra stages to compensate for the losses of gain caused by the corrections mentioned above.

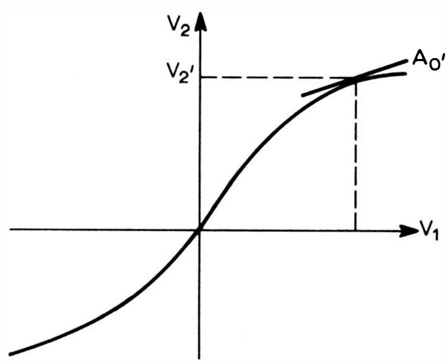

Fig. 5. Open-loop transfer function of the amplifier A_0 is the incremental gain.

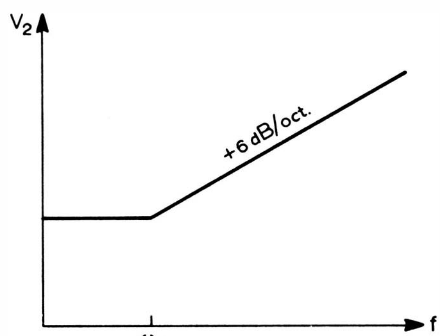

Fig. 6. Error voltage V_2 as function of frequency. www.keith-snook.info

2. The abrupt distortion such as cross-over distortion. These are difficult to cure, sound very bad and usually overall feedback has little effect on them. The possibility is to allow operation deeply enough in class A, a practical target specification being 14 to 20dB below maximum output power[4]. As compared to many present designs, this leads to higher quiescent power losses and consequently a larger heatsink.
3. The dynamic non-linear distortions. As the dynamic distortions are principally effects caused by poor frequency behaviour of an amplifier, they can be cured completely by following certain simple rules in the design[1, 5], and not by using too much overall feedback.
4. Some presently unknown dynamic distortion mechanisms such as the clear effect of loudspeaker load on the audible sound quality of some amplifiers.
–phase modulation effect, probably caused by power transistor cut-off frequency sweeping with the output power
–possible importance of reproducing faithfully the higher derivatives of the signal.

Of these distortions, cases 1 and 2 may be made very small with good design of the amplifier, and by a readiness to meet the cost of added components and a larger heatsink. Case 3 is easy to eliminate totally by proper design with practically no increase in parts cost. Case 4 remains to be studied but at least until it has been solved, the final sound quality measuring instrument must be the ear.

Conclusion

Dynamic distortions were unknown until recently. There seems to be some correlation with the phenomenology presented above and subjective listening tests. It is commonplace to find an amplifier having a good harmonic and SMPTE intermodulation distortion specification (and thus probably high overall feedback) which fails in the listening tests. It has also been shown that irrespective of unmeasurable harmonic and SMPTE intermodulation distortion, an amplifier may produce dynamic intermodulation products having amplitudes of tens of percent[3]. The t.i.m. seems to explain a part of this dilemma but, certainly, there must be other similar effects.

With the static non-linearity measurements, we have only stated that an amplifier must be capable of reproducing the absolute value of the signal correctly. What the dynamic non-linearity considerations show is that the amplifier must in addition be capable of reproducing faithfully the first and the higher-order derivatives of the signal as well. The t.i.m. is part of the non-linearity of first derivative reproduction. What the other parts are and what requirements the higher-order derivatives of the signal impose on the amplifier remains to be discovered.

At this moment we are living through a very exciting phase in electro-acoustics, the challenge of explaining the clear contradiction between our measurements and our subjective sound quality sensation. I forecast lively activity in this field in the near future.

References

1. Otala, M., Lohstroh, J., Audio power amplifier for ultimate quality requirements, *IEEE Transactions*, vol. AU-21, 1973, pp.545-51.
2. Otala, M., Transient distortion in transistorized audio power amplifiers. *IEEE Transactions*, vol. AU-18, 1970, pp.234-9.
Otala, M. & Leinonen, E., Theory of the transient intermodulation distortion. Monitor – *Proc. IREE*, vol.37, March 1976, pp.53-9.
3. Otala, M. & Leinonen, E., Possible methods for the measurement of transient intermodulation distortion. 53rd AES Convention, New York, October 1976.
4. Daugherty, D. G. & Greiner, R. A., Some design objectives for audio power amplifiers. *IEEE Transactions*, vol.AU-14, 1966, pp.43-8.
5. Otala, M., Circuit design modifications for minimizing transient intermodulation distortion in audio amplifiers. *Journal of the Audio Engineering Society*, vol.20, 1972, pp.396-9.