Negative feedback in audio amplifiers: Why there is no such thing as too much (Part 2)

Bruno Putzeys - August 26, 2013

This article originally appeared in Linear Audio, a book-format audio magazine published half-yearly by Jan Didden.

[Part 1 began with a look at the theory behind feedback loops.]

The controversy

Control theory is a vital and standard part of engineering studies: the effectiveness of negative feedback to keep jet-fighters airborne, ferries afloat and nuclear power plants a-non-exploding has somehow never become the subject of much controversy.

Only in audio does the usefulness of feedback draw heated debate, with detractors saying that reasonably good measured performance obtained without feedback sounds better than excellent performance obtained with feedback. Proponents opine that there is nothing wrong with taking a good amplifier and making it better by putting feedback around it, so long as it's not overdone.

What is "an amplifier with a lot of feedback?" What is "slow" and "fast?"

Peculiar to the discussion is a mixing of terms that stews technical terms and lay words like speed, open-loop bandwidth, GBW, amount of feedback and loop gain into a dissonant cassoulet.

Open-loop bandwidth as a measure of speed

Most informative is the phrase "a slow amplifier with a lot of feedback." Translated into technical language, this means an amplifier - let's call it amp X - with the kind of open and closed loop gains as shown in Figure 7.

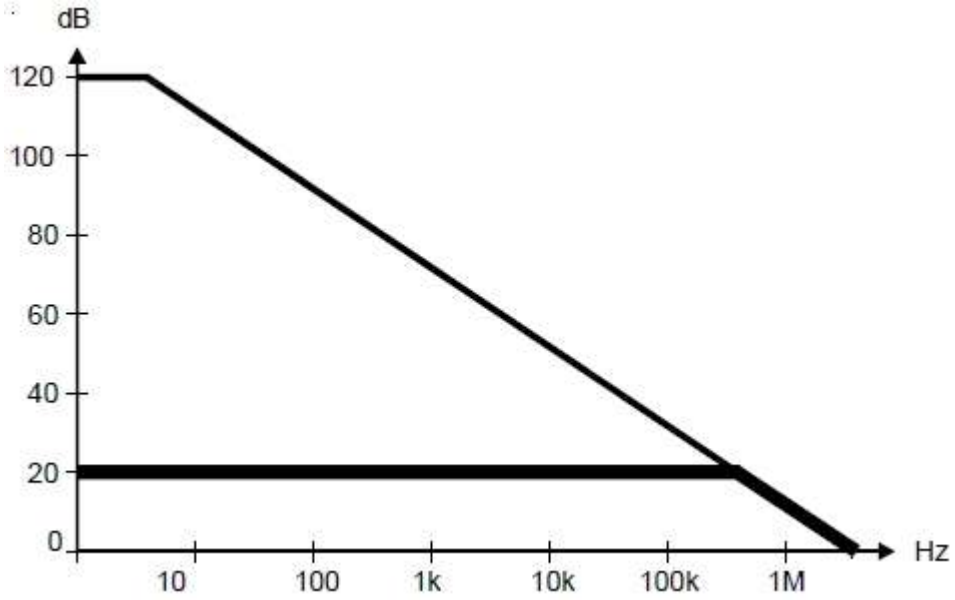

Figure 7: Amplifier X is "a slow amp with a lot of feedback."

The thinner trace in Figure 7 is the open-loop gain plot, A(s). It is the frequency response that the amplifier would have if no feedback were used. Bandwidth would be less than 10Hz, hence "slow." "A lot of feedback" refers to the fact that the feedback network whacks the gain down a full 100dB. Closed-loop gain is very nearly 1/B(s). B(s)=0.1, so here's an amp with a gain of ten. The difference between open-loop gain and closed-loop gain largely corresponds with loop gain A(s) • B(s).

It is very easy to understand how a semantic mode of thinking makes many feel uneasy about this. Intuitively it sounds very much like turning a donkey into a racehorse by putting feedback around it.

By contrast, amplifier Y (Figure 8) is a "fast amplifier with little feedback around it." Even without feedback it manages the full 20kHz so it "doesn't need much feedback to have a wide frequency response."

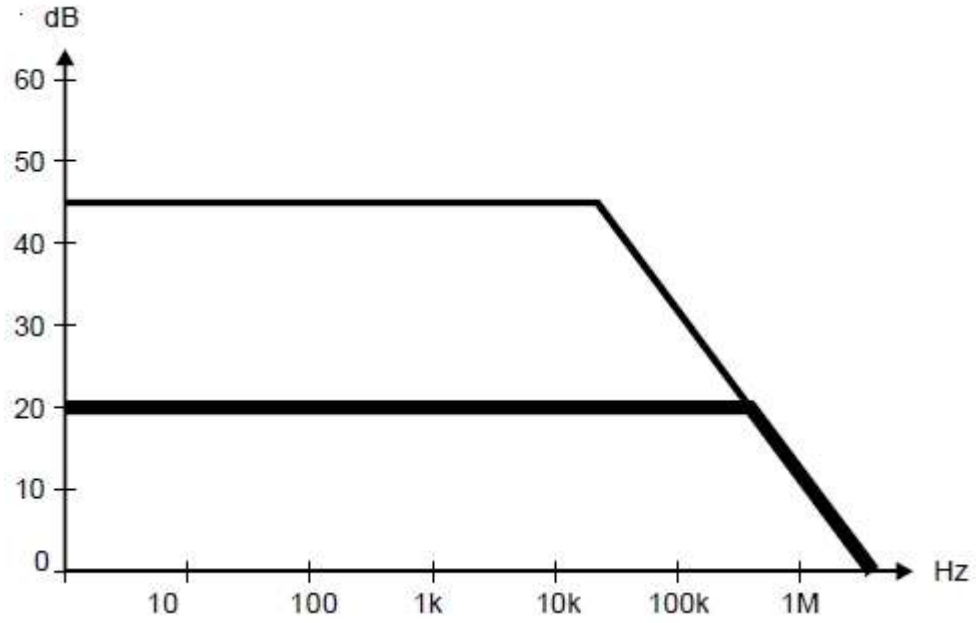

Figure 8: Amplifier Y is a "fast amplifier with little feedback around it."

Hence the often encountered (mis)perception that the speed of an amplifier is indicated by its open-loop bandwidth.

Semantics can be treacherous. "Speed," if anything, is a time-domain thing. If we want to determine which of those amplifiers is the faster one, we are better off looking at a time-domain plot - step response for instance. The four traces in Figure 9 show the output of both amplifiers in response to a 10mV step, once with feedback, once without.

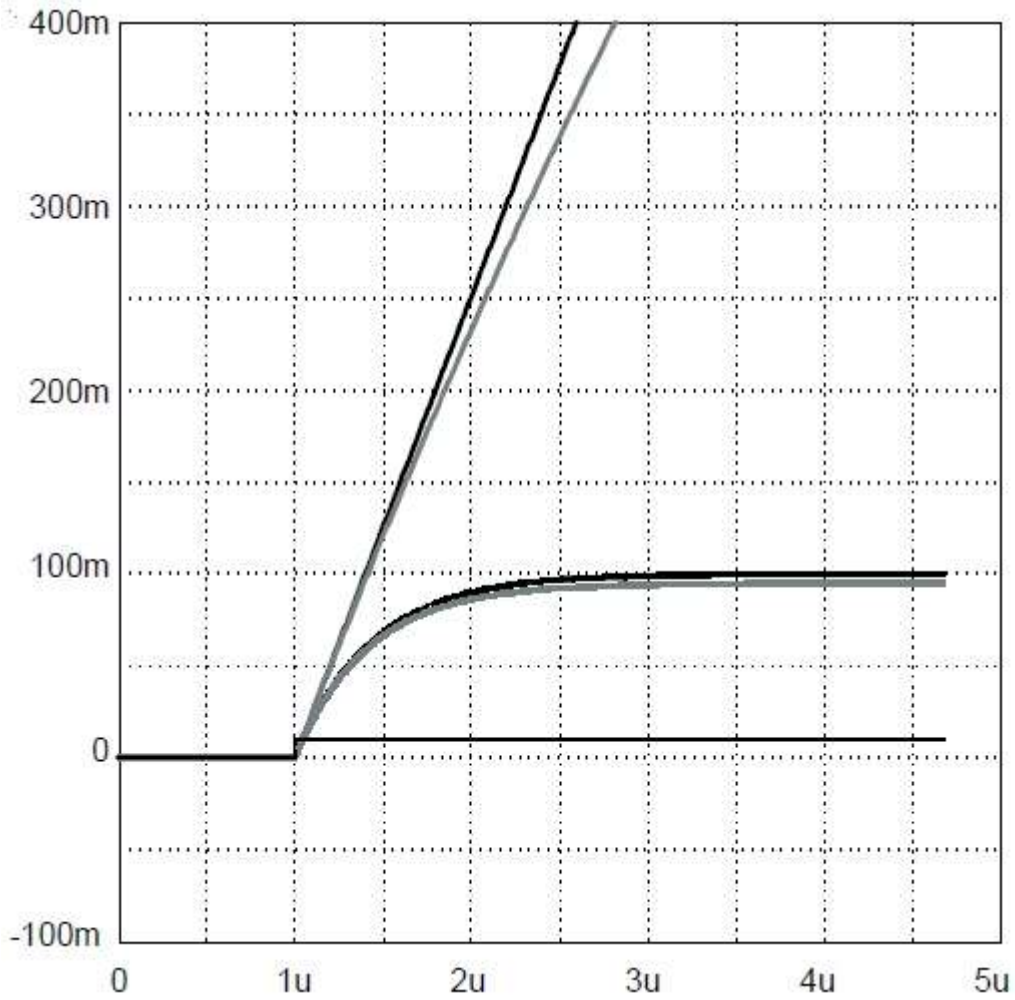

Figure 9: time domain step responses of amplifiers X (dark curves) and Y (lighter curves) with (lower curves) and without (upper curves) feedback.

In open loop (upper curves), amplifier X produces a straight line in response to the constant input voltage. Amplifier Y starts off with the same slope, except that it starts levelling off toward some voltage outside the graph's scale (around 1.7V) after 100µs or so. Actually amplifier X does the same but it'll only start levelling out when it reaches about 10kV after almost a second, or it would if its supply voltage allowed it to.

Looking at the closed-loop responses (lower curves), they are nearly identical, and also start exactly as fast as the open-loop responses. The only difference is that while Y at some point decides that it's "close enough," X adjusts a smidgen further. Both amplifiers are just as "fast," with or without a feedback loop in place.

To update the metaphor: we didn't turn a donkey into a racehorse. It was a racehorse to begin with. The only thing feedback does is to put a jockey in the saddle to tell the horse when to stop running.

"Fast" really means how quickly it'll get there. DC open-loop gain only tells you how far the horse is prepared to run without being ordered to stop.

The same is borne out in an overlaid gain plot of amps X and Y (Figure 10). If you're used to juggling bode diagrams and scope pictures like me you'll probably be in the habit of looking to the right-hand side of one to see what the left hand side of the other will do and vice versa. Just like the four traces overlap in the left of the time domain plot one can expect them to overlap on the right side of the frequency domain plot.

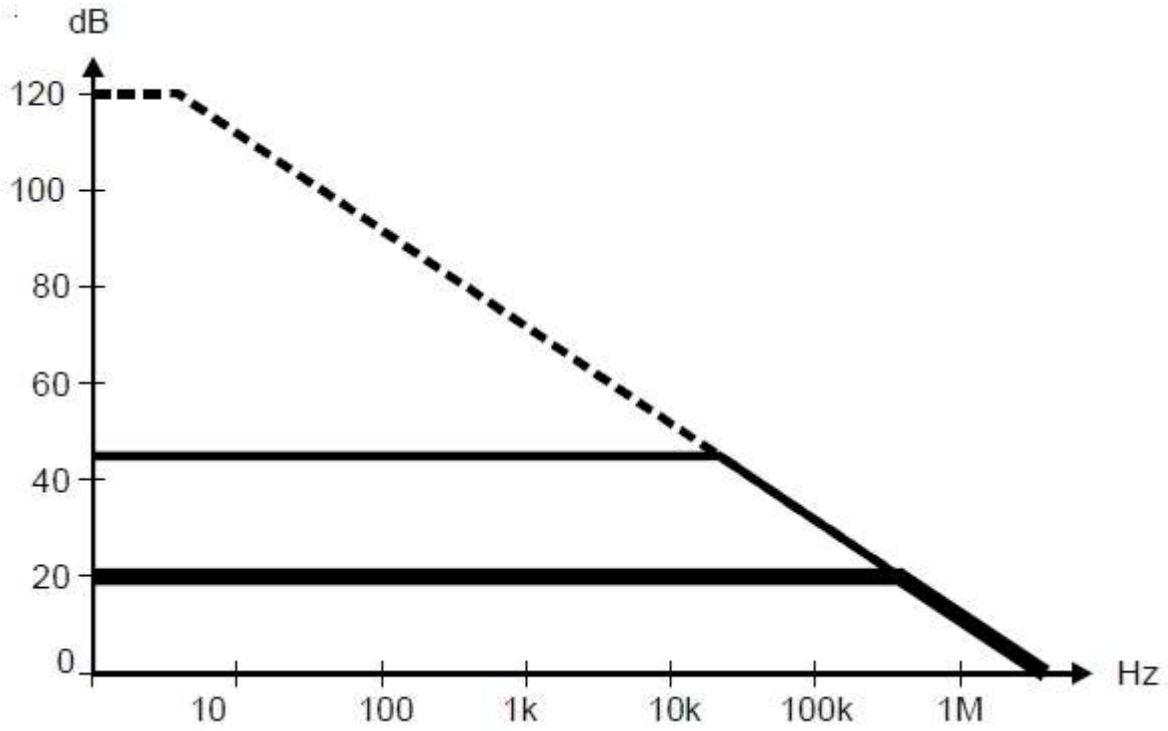

Figure 10: Amplifiers' X and Y gain plots.

Open-loop bandwidth is no indication of speed and tells us nothing about the qualitative behaviour of the feedback system. The distortion products that we hope to attenuate using feedback are all at audio frequencies, so what matters is loop gain at audio frequencies. If at some impossibly low frequency it is much higher, we don't care. If at the other end of the audio spectrum it's not enough, we do.

Gain-Bandwidth product as measure of speed

All three-stage amplifiers share one obvious characteristic: loop gain drops 20dB per decade. Referring to box 1 ("Quick refresher on the three-stage amplifier"), this is the transimpedance stage's job. If we look at the step response graphs in Figure 9: what would make those amplifiers respond more quickly? Shifting the 20dB/decade asymptote to the right of course! But isn't that the same as shifting it upward? Indeed, my dear Watson. Making the amplifier faster increases loop gain across the entire asymptote.

A faster amplifier is simply one with more loop gain at high frequencies, which in most cases means the entire audio range. Conversely, the only way of getting "more feedback" from a three-stage amplifier is to make it faster.

Physically this is done by decreasing the compensation capacitor. Unfortunately we can't just do that. Both input and output stages are low-pass filters. Control theory tells us that the additional

phase shift will cause the loop to go unstable if there's too much loop gain left, so we set loop gain to become less than 1 well below the corner frequencies of either the input or output stages. The bandwidth of the input and output stages limits the unity gain frequency and hence loop gain.

I'm deliberately cutting corners here – refer to one of the excellent power amplifier books by the two usual suspects or to any good textbook on control theory for more details. In short: all other things being equal, faster means more loop gain.

Why it happened

Why it happened
I don't know exactly since when negative feedback in audio amplifiers has become the issue it currently is but several storylines converge.

Storyline 1: TIM
In 1970, Otala identified his infamous "Transient Intermodulation Distortion," a fancily named manifestation of Slew Induced Distortion. SID is the Chihuahua to slew rate limiting's Bullmastiff so let's make sure we understand the latter first.

Slew rate limiting means that an amplifier is trying to reproduce a very fast rate-of-change signal but can't. The maximum rate of change it can produce is determined by the current available to charge or discharge the compensation capacitor. This current is provided by the input stage and the maximum is the tail current Ib. When a faster rate of change is demanded, the input stage is overloaded and becomes completely unresponsive to any further change. At this point, the feedback loop stops working and is no longer able to control the amplifier.

Something similar, although much subtler, happens at lower slew rates. Box 2 ("What is Slew Induced Distortion?") shows the details and notes that the distortion shows up as an error voltage across the differential input stage. Again we find that the feedback loop is powerless. Add the input stage distortion to our previous diagram and we get figure 11. With some reflection we realise that the nonlinear portion of Vdiff is just an error source in series with the differential input.

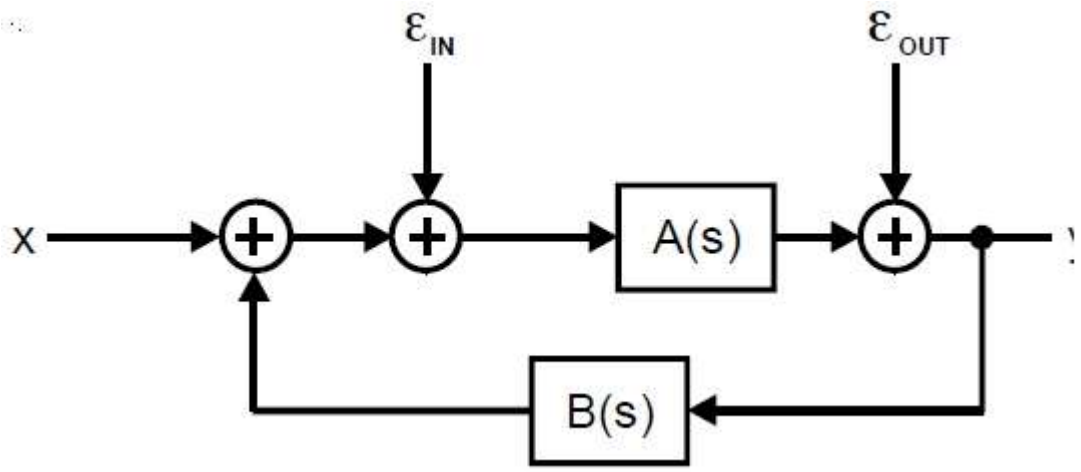

Figure 11: Adding the differential input stage's error signal.

Bad news. The error simply adds itself to the input signal. The feedback loop does not see any difference between the wanted signal x and the unwanted signal εin. The physical source of SID is outside the feedback loop! This belies the idea that "TIM is caused by the use of feedback."

If there is a relationship, it has to be through the behaviour of the specific circuit implementation. For a given GBW, slew rate is a fixed quantity and slew induced distortion can be predicted exactly from the ratio of actual to maximum slew rate. Nothing will reduce SID other than improving slew rate unless we modify the circuit.

Unfortunately, at the time, many pundits missed the conditional clause and started a slew rate rat race whose followers to this day believe that a fast slew rate is the most important specification predicting sound quality, and can't get high enough.

It is ironic that the received opinion in those days (and for some still today) is that "more feedback" means "more TIM." SID is inversely proportional to the third power of slew rate and hence GBW. Precisely that which increases loop gain yields an immediate and large decrease in measured TIM.

What oversight could explain that intelligent people come to a belief that is the exact opposite of reality?

I think two things had to come together. First, the confusion of what "a lot" or "less" feedback actually meant (see earlier). Second, the results of actual practical experiments that led to lower TIM after measures taken to lower loop gain. Indeed, those thrifty experimenters of the 70s reported just that. The key lies in how it was done. One experiment that I could trace consisted of degenerating the input pair; that is, placing a resistor in series with each emitter of the input pair. This is effectively a form of local feedback and a sure-fire way of reducing the transconductance of the input stage.

In Figure 12, trace (1) shows the original loop gain. After degenerating the input pair by, say, a factor 10, something like trace 2 would result. It would, however, be quite pointless to let the unity-gain frequency drop by a factor of ten since the amplifier will now remain just as stable with a compensation C that's ten times smaller. The whole curve shifts a decade to the right, which effectively jacks HF loop gain back up 20dB (trace 3). In the end, only DC loop gain is 20dB down but little else has changed. For most of the audio band, traces 1 and 3 overlap.

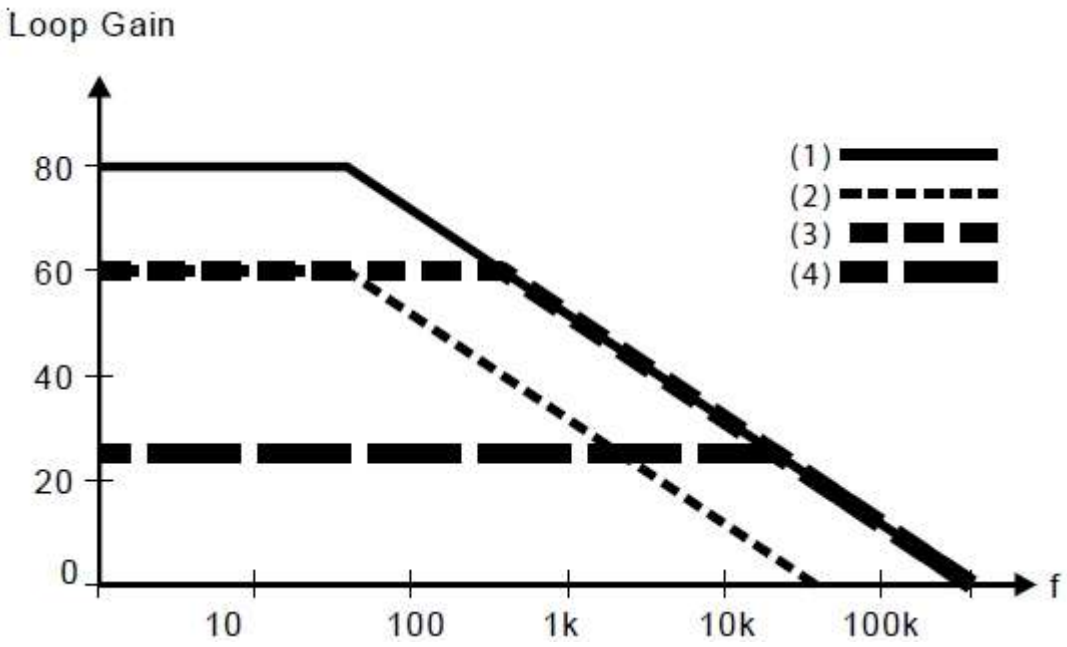

Figure 12: Degenerating the input differential pair.

What else has happened though? Degenerating the input pair has a second consequence: *the non-*

linear contribution of the input transistors is reduced (Figure 13)!

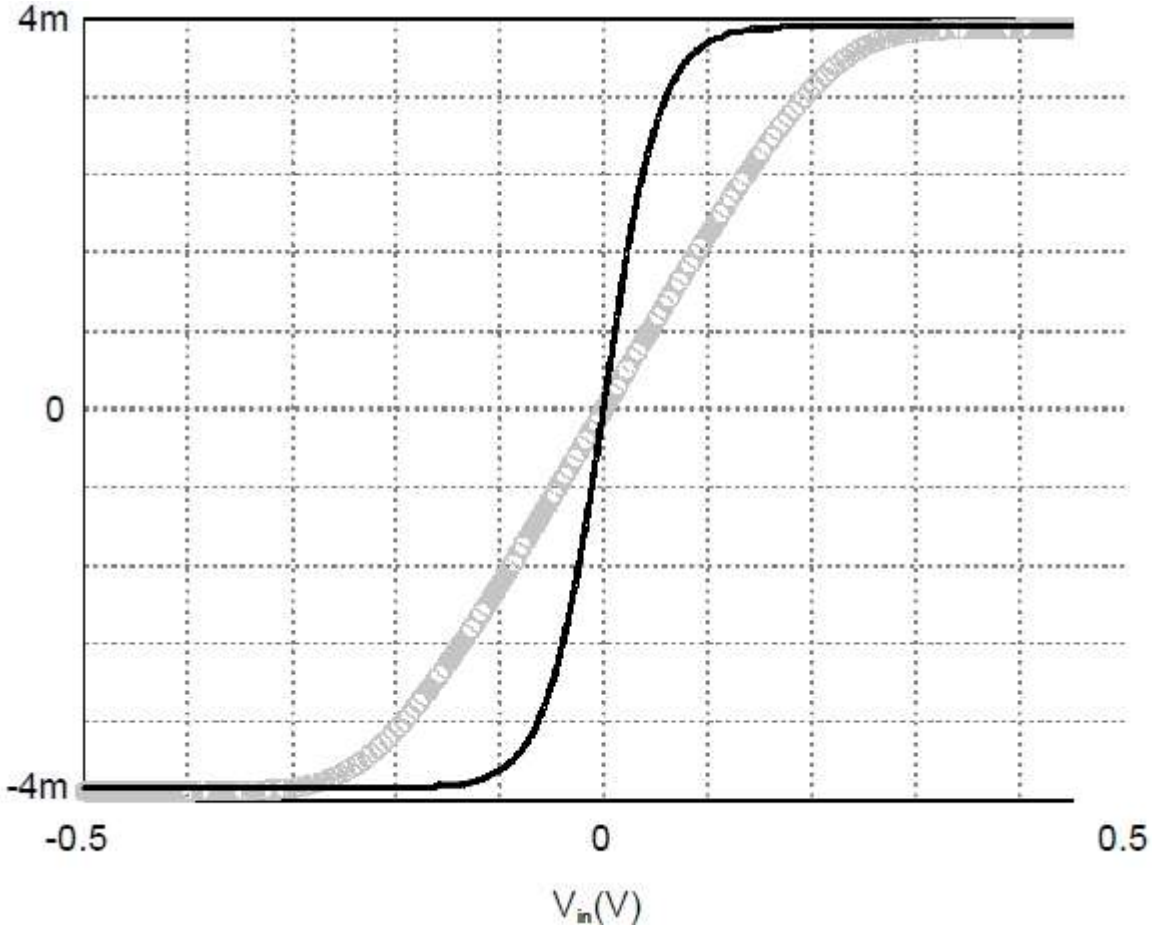

Figure 13: differential input stage transfer function without (dark) and with (light) emitter degeneration.

By how much? Well, after reducing C by about a factor 10, the same rate-of-change demands only one tenth the current. Reducing I by a factor 10 will reduce the 3rd harmonic in Vdiff by a factor of 1000 and the 5th by a staggering one hundred thousand times. Loop gain hasn't fundamentally changed but SID is all but eliminated entirely. The sonic improvement must have been spectacular, but clearly this cannot be attributed to reduced feedback, as feedback hasn't been reduced!

Most people now know that degenerating the input stage is the recognized method of getting rid of SID. Since SID can be eliminated independently from feedback, there are no grounds for claiming a causal relation between the two. The experiment was a resounding success, but the conclusions drawn from it were incorrect. The correct conclusion is: SID can be eliminated without affecting loop gain, so loop gain does not cause SID. SID is a circuit flaw pure and simple, and degeneration of the input stage fixes it.

A second step in the experiment consisted of placing a resistor across compensation capacitor C that reduced DC gain to the same value as that at 20kHz (Figure 12, trace 4). The test amplifier was of the folded-cascode persuasion, which allowed this. At this stage, loop gain has indeed been reduced across the full audio range. I surmise that since the amplifier's distortion was never negligible, making it constant across the audio band makes it fly under the psychoacoustic radar more easily. My own subjective experience would support this.

To my ears, amplifiers with the normal 20dB/decade behaviour but whose distortion is not negligible

at the end of the audio range have glassy mid-highs, a "superglue stereo image" as KK once put it, and the illusion of spectacularly, unnaturally tight and impossibly controlled bass. Some love this, and seceded into a subculture of ultra-beefy amplifiers. I don't and when forced to make a choice I'll take higher but consistent distortion across the band.

The fourth step is a very important one to our story: to eliminate the feedback loop altogether and set the gain using only emitter degeneration and a resistor en lieu of C. Several commercially available class A amps are still made in this manner.

I heard that this last change was a bit of an epiphany. There is no discussion here: loop gain was reduced from about 26dB (at 20kHz) to none at all. Unlike the first step, this one does point at a valid relationship between loop gain and perceived sound quality. Something interesting is happening here.

Storyline 2: Re-entrant distortion

Storyline 2: Re-entrant distortion
I said earlier that the error is signal dependent. In 1978, Baxandall noted that negative feedback around simple nonlinearities creates distortion components that weren't there before (Figure 14).

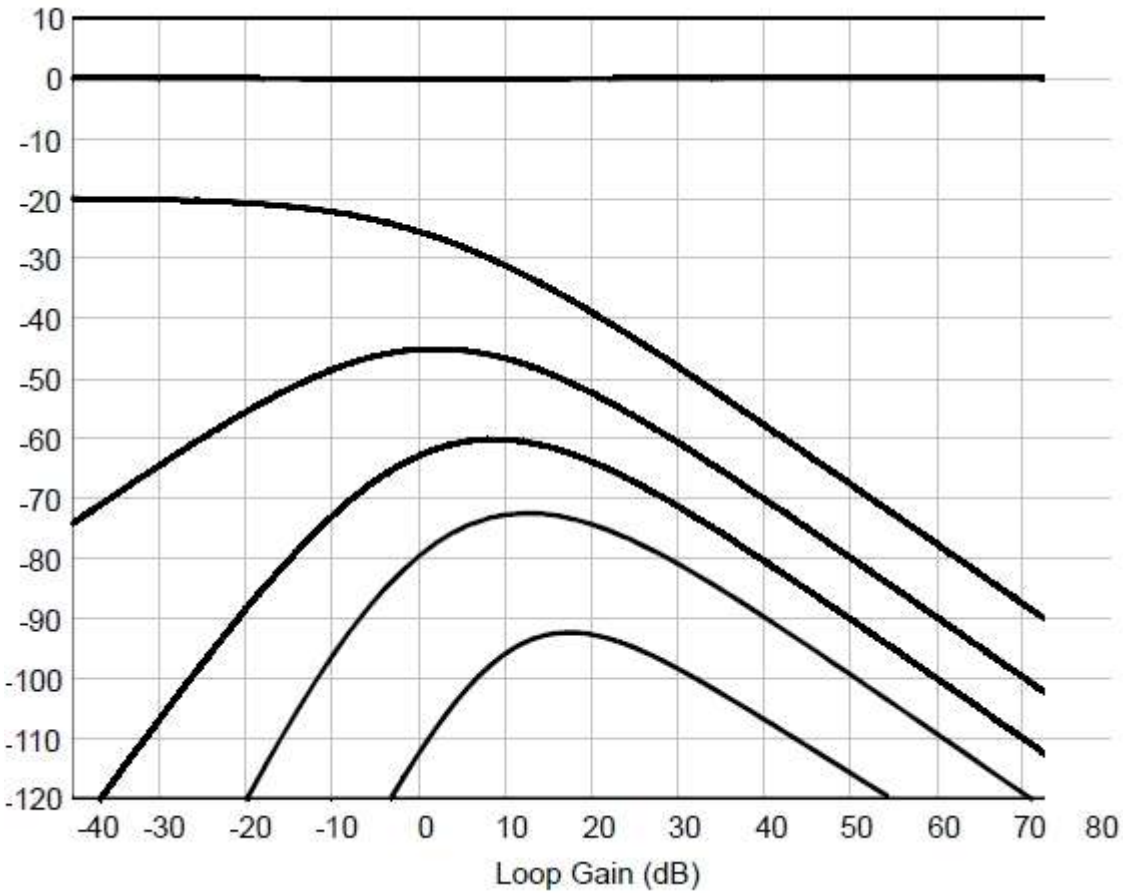

Figure 14: Spectral decomposition of the output of Figure 15 as a function of A (dB). Top to bottom: fundamental, 2nd harmonic, 3rd harmonic, etc.

Trying to solve analytically an integrating control loop that has a dependent error in it is impossible. Differential equations with nonlinearities in them do not have algebraic solutions. Luckily we can get a qualitative understanding of the issue by looking only at very low frequencies where gain is constant.

Imagine a hefty but purely second-order nonlinearity with a loop wrapped around it (Figure 15).

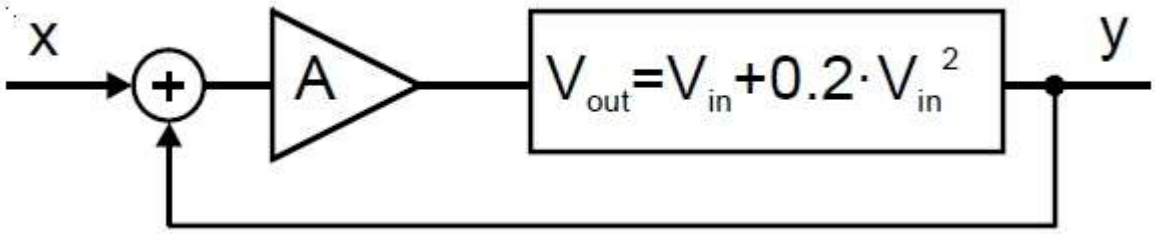

Figure 15: Purely 2nd-order non-linearity within a feedback loop.

The nonlinearity is parabolic. As soon as feedback is added, it will become something that is neither a straight line nor a parabola. In fact, the answer has a square root in it. That has an infinite series expansion which includes both even and odd terms. Plotting the harmonic distribution against loop gain, we get Baxandall's familiar picture of Figure 14.

Forget the old pot-boiler that second-harmonic distortion is inaudible or uncritical, it's not. Even-order terms produce IMD products that clog up the whole bottom end. Nevertheless there is little doubt that higher harmonics are both increasingly detectable and increasingly annoying. This mathematically derived graph predicts exactly what experimenters found: if you start with a decently sounding zero-feedback amplifier and you add some feedback, it sounds somewhat opener in the bottom end but otherwise more unpleasant. The trend continues at least for the first 10dB to 20dB so one finds oneself compromising between musicality and something we'll call "accuracy" for want of a better term.

The trouble is that such listening experiments are by necessity done on amplifiers that are flat all the way to 20kHz, even in open loop, usually valve amplifiers. A valve amplifier that has a flat response from 20Hz-20kHz is very likely to have an uncontrolled outgrowth of poles just outside this range and will become unstable well before loop gain hits 15dB. You get worsening musicality right up to the point when the whole thing no longer works.

Of course this experiment gives the impression that more feedback is worse. You have to get past that bump. Hardly anybody who has ever tried it like this has actually heard the inevitable (and frankly magical) improvement that happens once you do get beyond, say 20 or 30dB. From there on you get an unambiguous net improvement that goes on forever.

In a flight of fancy I set a friendly audio company in the south of the Netherlands on this by suggesting a method of wrapping almost 60dB of loop gain over the full audio range around a valve amplifier using a third-order loop. Whenever it was stable it sounded immaculate. Measured rather well too.

The story of re-entrant distortion has an unambiguous conclusion: there simply is no such thing as "too much" feedback. There is only something as not enough feedback and it happens to be exactly what so-called moderate audio designers call "modest amounts" of feedback.

At this stage we understand how at a certain juncture enthusiastic, knowledgeable audiophiles were getting confused about feedback, and drew tentative, negative conclusions about it. Normally though, science and engineering works by making mistakes and rectifying them as work continues. Therefore, a third factor is needed to explain how a tentative conclusion suddenly got cemented into an immovable orthodoxy.

Storyline 3: Marketing hype, specmanship and "lots of zeros"
During my tender years, "Japanese Transistor Amps" were held up as prime examples of things that measured well and sounded terrible. Dare, even today, to extol the virtues of an amplifier as having really low distortion and some know-it-all will stand up and say "you know measurements don't say it all; remember the 80s when we were flooded with amps that had 0.00001% distortion and sounded all screechy," bystanders nodding vigorously. One of them will go on, inevitably, to make the same point in another audio meeting, adding an extra zero. Such figures were never claimed by anyone at the time.

That doesn't mitigate that the leaflets were misleading. Sometimes subtly by stating only THD at 1kHz/1W, often more brutally through a technique called lying. These amps didn't measure at all well and they sounded the part.

Superb specs were claimed for the cheaper amps in the full expectation that no reviewer would bother to measure them, while more modest figures were given for top-of-the range products in the reasonable assumption that they would. Had the marketeers then realised what they were setting the industry up with, they would have committed seppuku.

The Backlash

The Backlash
The effect on the audio trade has been profound. Not only did negative feedback become the object of suspicion and ridicule, so did really good measured performance because good numbers were tainted by association with "large amounts of feedback." If you had to use so much feedback to get those numbers, it couldn't be good. Measurement reports became superfluous, opening the floodgates to claims of sonic excellence that had no support in physics whatsoever.

The whole high-end trade standardized on making amplifiers with low amounts of loop gain and hence significant distortion. Of course, these amps all have a "character" of their own, adding another layer to the relativism by making any amplifier that sounded different from others worthy of consideration. Some designers became world famous because they managed, more by trial and error than by forethought, to make amplifiers that, in spite of having little or no error control and distortion levels in the 0.1% - 1% range, did not colour the sound excessively.

The painstaking sculpting of the distortion spectrum of an amp by mixing and matching different makes of active devices and load conditions became a Zen-like activity that guaranteed guruship. Furthermore, with low or no loop gain came bad power supply ripple rejection: it became a mark of competence to overdesign the power supply, even regulating it, and to mix and match esoteric supply capacitors to mitigate their impact on sound. Having to contend with the ones that are actually in the signal path isn't enough for some.

Designers who should know better pay lip service to the ideal of fast amps with moderate feedback, or "local feedback only." The avoidance of feedback, specifically global feedback, also meant that longer signal chains quickly accumulated distortion products. A relentless drive for minimalist design ensued. If everything one adds to the signal path detracts from the result, only the smallest number of components will do. This resulted in the ludicrous situation where fantastic sounding recordings were made with signal chains numbering up to a hundred amplifying stages and replayed on audiophile systems where even a transparent buffer proved an impossibility.

Hi-fi review is a complete shambles. The few magazines that do measure are capable of reprinting the most frightening distortion spectra from amplifiers and actually call them good. "Objectivity" got downgraded from "independent of who's doing the observing" to "not favouring particular brands."

For me personally, the affair hit rock bottom when in 2009 two reviewers, one Dutch, one British, independently remarked of the same amplifier (a reasonably priced product with exemplary performance) that it sounded surprisingly musical for an amp with such low distortion.

In the 21st century audio engineers build equipment while actively avoiding two of the most powerful tools available to the whole of science and engineering: measurement and error control. The damage to the audio industry and its reputation in the wider engineering world will remain immeasurable until we decide to take control.

Finally, some stuff to remember

- Beware of error sources outside of the feedback loop.
- TIM is not a special type of distortion; it is a method to test for Slew Induced Distortion.
- SID can be eliminated without changing loop gain. Therefore, SID is not caused by negative feedback.
- Improving loop gain improves TIM. There is no horse trading between "ordinary"distortion and TIM.
- Otala's work neither implies nor proves that valve amplifiers are better than solid state.
- DC open-loop gain is no measure of how much feedback an amplifier has. Loop gain at 20kHz is.
- Slew Rate is a bad predictor of audio performance.
- Open-loop bandwidth is no measure of how fast an amplifier is. Gain-bandwidth product is.
- Make sure you have actually heard an amplifier with proven negligible distortion before having opinions re sound vs. measurements.
- Make sure you have actually heard an amplifier with large loop gains before having opinions re sound vs. feedback.
- Various proposed alternative error correction schemes are functionally equivalent to feedback.
- Nested feedback is functionally equivalent to global feedback.
- Higher-order loops make it possible for slower amplifiers to attain top-notch audio performance.
- There are only advantages and no disadvantages to applying stratospheric amounts of negative feedback in an amplifier. The only hard part is figuring out how to do it.
- The more feedback, the better it sounds provided that it's never less than 30dB at any audio frequency.

Acknowledgments
Many thanks go to Peter van Willenswaard, without whose memory and willingness to act as a sounding board this article would have missed some very important steps. Thanks also to Guido Tent and Bart van der Laan for bearing the brunt of my "flight of fancy."

About the author
An electrical engineer, Bruno Putzeys has been active since 1995 designing class D power amplifiers for high-end and consumer audio applications, digital and analogue delta-sigma and PWM modulators, AD/DA converters and discrete solid-state analogue audio circuits. He has further experience with vacuum tube electronics and loudspeaker design. Now working for Hypex, he was formerly employed at Philips where he invented, among others, the "UcD" power amplifier circuit.

Also see:
Negative feedback in audio amplifiers: Why there is no such thing as too much
Op amp myths – by Barrie Gilbert
Designing a low-distortion audio output stage - Part 1: Introduction, the problem with push-pull outputs

The Class i low-distortion audio output stage (Part 4)
New approaches to switched-mode audio power amplifiers (Part 1)

Quick refresher on the three-stage amplifier

Box 1: Quick refresher on the three-stage amplifier

The vast majority of class A/AB/B amplifiers follow the classic 3-stage op-amp structure.

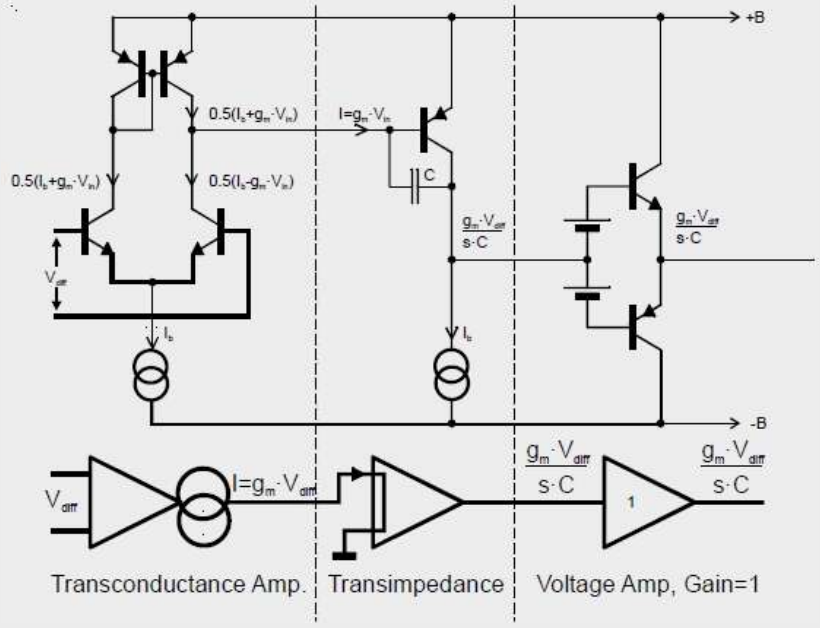

Figure 1-1: The basic three-stage audio amplifier.

In Figure 1-1, the first stage is a transconductance amplifier converting the voltage difference between two input nodes into a current. Transconductance of a long-tailed pair with no emitter degeneration is

$$g_m = \frac{I_b}{2 \cdot V_T}$$

where VT is the thermal voltage of 26mV (at room temperature). The second stage is a transimpedance amplifier that takes its input current from the first stage and outputs a voltage equal to

$$V_o(t) = \int \frac{-I}{C} dt$$

Under no circumstances should the second stage ever be referred to as the "voltage amplifier stage." It has a current input and a voltage output. Trying to understand the three-stage amplifier as the cascade of three voltage amplifying stages is impossible and meaningless. I urge my two more famous colleagues to cease this VAS rubbish forthwith. You're confusing your readers.

The compensation capacitor C has three functions, in this order:

• Reducing the output impedance of the middle stage so that the input capacitance of the output stage does not add significant roll-off and phase shift.

• Reducing the input impedance of the middle stage so that the impact of the miller capacitance of the input stage transistors is minimized.

• Insuring that loop gain is less than unity before the previous two introduce too much phase shift.

The third stage, the power stage, has a gain of essentially 1. C is chosen such that this is valid over the entire working bandwidth. Thus, the approximate transfer function becomes

$$H_{OL}(f) = \frac{g_m}{2 \cdot j \cdot \pi \cdot f \cdot C}$$

Without emitter degeneration this becomes

$$H_{OL}(f) = \frac{I_b}{4 \cdot j \cdot \pi \cdot f \cdot C \cdot V_T}$$

From this we learn that the gain-bandwidth product is

$$GBW = \frac{I_b}{4 \cdot \pi \cdot C \cdot V_T}$$

Maximum slew rate is also easily derived from the fact that the input stage can put out no more than its tail current in either direction.

$$SR = \frac{I_b}{C}$$

If no emitter degeneration is used, maximum slew rate and gain-bandwidth product are directly related:

$$SR = GBW \cdot 4 \cdot \pi \cdot V_T \approx GBW \cdot 0.32V$$

What is Slew Induced Distortion?

Box 2: What is Slew Induced Distortion?

In the section "Quick refresher on the three-stage amplifier" we saw that the transimpedance stage produces an output voltage equal to

$$V_O(t) = \int \frac{-I}{C} dt$$

That's as much as saying that the input current is proportional to the rate-of-change (slew rate) of the output voltage:

$$I = -C \cdot \frac{dV_o(t)}{dt}$$

For small-signal purposes we can model the input stage as a simple transconductance but for fast signals this is no longer the case. For large input signals the behaviour is rather non-linear:

$$I = I_b \cdot \tanh\left(\frac{V_{diff}}{2 \cdot V_T}\right)$$

This is a function with a slope equal to gm at Vdiff=0 and levelling off at +/-Ib. Solving for V_{diff} we get

$$V_{diff} = 2 \cdot V_T \cdot \operatorname{artanh} \frac{-C}{I_b} \cdot \frac{dV_O}{dt}$$

For some reason I now feel this sudden urge to rid myself of C and Ib in this formula:

$$V_{diff} = 2 \cdot V_T \cdot \operatorname{artanh} \frac{-1}{4 \cdot \pi \cdot GBW \cdot V_T} \cdot \frac{dV_O}{dt}$$

Of course I should now go on and derive a series expansion and work out exact distortion figures but I shan't. The main take-home point here is that the function is symmetrical and therefore consists of only odd harmonics of which the lowest, the third, increases proportionally to the third power of the amplitude of Vdiff, and hence of the third power of both the output amplitude and frequency. So we throw out all the constants and write that proportionality as:

$$D_{in} \sim \left(\frac{f \cdot |V_O|}{GBW \cdot}\right)^3$$

Now I remember why I wanted GBW in there! Not only is slew induced distortion proportional to the third powers of frequency and signal level, it is also inversely proportional to the third power of GBW. Improving GBW by a factor of two, causes slew induced distortion to drop by a factor of eight.

There is a common misconception that only TIM tests will lay bare slew induced distortion. This is not at all the case. On a THD vs. frequency sweep, slew induced distortion looks like a sharp increase in THD towards the highest frequencies. The only thing TIM does special is stress the amplifier harder and create distortion products inside the audio band, making the problem more easily visible.

A more modern approach that does the same thing is a simple IMD test with a 1:1 18.5kHz+19.5kHz two-tone stimulus. Anything a TIM test can tell us is exposed with equal ease with this test. In addition it does not rely on slew rates that no audio signal actually has.