Amplifier Nonlinearity: Dynamic Effects

Erik Margan
Experimental Particle Physics Dept., Jožef Stefan Institute, Ljubljana, Slovenia

Abstract

Amplifier feedback error correction introduces certain well defined dynamic effects, which can be investigated by employing a simple differential technique. We present and propose a preferred test circuit realization for DIYers, and show an example of error signal analysis and interpretation.

Introduction

Amplifier error correction techniques, both feedforward and feedback, have been invented by a Bell Labs engineer H.S. Black in 1925 and 1927. It seems odd that after all these years of use and development there is still a lot of misunderstanding, ungrounded fear, as well as prejudice about the technique itself and its consequences, particularly under dynamic conditions. Especially in audio a school of thought has developed which is deeply suspicious of any form of overall feedback application, relying instead on only local feedback linearization techniques. Whilst being a perfectly valid design approach, it only illustrates all the irrational distrust in feedback.

To be completely fair, it must be admitted that such a situation developed as a reaction to an exaggerated use and even abuse of feedback by many designers of cheap equipment, where the overall quality of performance was (mostly inadvertantly, but sometimes deliberately) sacrificed to a low manufacturing cost. Thus amongst audiphiles feedback became synonimous to cheap, low quality equipment.

In the way it is being tought at school by relying on oversimplified amplifier models, feedback appears to have only beneficial consequences to amplifier performance, lowering both the self generated noise and the distortion of the amplified signal. Yet it was Black himself to point out in his article of 1934 that feedback should be applied judiciously in regard to the Nyquist's stability criterion.

As Barrie Gilbert likes to point out in his lectures, a designer must always think of an amplifier as an open loop integrator with feedback. Since the feedback loop is negative, it represents a phase shift of 180°; a further 90° is used up by the dominant pole; the Nyquist's stability criterion requires at least 45° of physe margin at the unity gain frequency to prevent the negative feedback signal becoming effectively positive; therefore only 45° remain to the amplifier designer to play with. The greater the open loop gain, the greater will be the feedback correction factor, however with more gain the secondary amplifier poles tend to show up above the unity gain level, introducing some additional phase shift at high frequencies and compromising stability of the closed feedback loop, consequently requiring a lower dominant pole to suppress those secondary poles and restore stability.

But a lower dominant pole comes with a price: an increased value of the compensation capacitor loads the output of the first amplifying stage, demanding more driving current at high frequencies; this may lead to transient intermodulation distortion (one of the first to point this out was Matti Otala in mid 1970s). A lower dominant pole also limits the error correction bandwidth, increasing the feedback reaction time, thus also leading to an increased amount of error at high frequencies.

These inherent limitations of the classical amplifier configuration, including odrinary operational amplifiers, led in mid 1980s to the development of a fully complementary current feedback topology, providing current on demand to the compensation capacitor. But because of the high open loop bandwidth, and because all of the voltage gain is being provided by a single amplifying stage, secondary poles tend to emerge above the unity gain at a relatively low open loop gain, thus 60 dB is seldom surpased. With a closed loop gain between 20 and 30 dB, only some 30-40 dB remain for the feedback error correction. This may be adequate in applications where the main design goal is a high signal bandwidth, but not in audio, where 0.1% of distortion is generally considered worse than mediocre.

Crossover Distortion

Besides the high frequency distortion and transient intermodulation distortion, the most notorious linearity problem of audio power amplifiers used to be the class B output stage crossover distortion, particularly in the first generation of solid state amplifiers, and to a degree it is still so today. As has been shown by many authors, most notably by Douglas Self in his series of articles published in Electronics World and later in a book **[1]**, the nonlinearity of a complementary output stage can be minimized by a suitable choice of the quiscent current, but it can never be completely avoided, unless of course special local feedback and/or compensation circuitry is employed. Readers of Linear Audio will surely remember such a circuit presented by Kendall Castor-Perry in Volume 2 **[2]**.

I am very grateful to Kendall for his kind remarks on my own work on this subject, as well as quoting the article I published in Electronics World & Wireless World, first as a circuit idea in October 1985, and then as a somewhat more elaborate article in July 1987 **[3]**. Without going into a detailed description, I must mention that the motivation for my design came after I saw two other circuits, an active bias compensation by Nelson S. Pass (US patent of 1973, **[4]**), and an article by Susumu Tanaka published in the Journal of the Audio Engineering society of America in May 1981 **[5]**. The circuit by Pass was intended for class A operation, which I did not want, and the circuity by Tanaka involved some local positive feedback, wich I also did not want. I felt that a proper non-switching-off class B operation could be attained by using negative feedback only. A couple of months and three blocks of blue grid paper later, I finally got what I was looking for, a circuit effective and simple enough that could be added to any ordinary amplifier. This testifies to the old joke by G.B. Shaw saying that people start behaving rationally only after they have exhausted all other options.

After publishig that circuit idea, I realized that I needed a way to mesure the difference in performance between an ordinary amplifier and my improvement. Not having any precision low distortion signal source, I turned to the idea shown by Peter J. Baxandall in Wireless World, November 1977 **[6]** (the Distortion Magnifier by Robert C. Cordel in Linear Audio Volume 0, September 2010 **[7]** was still some 25 years into the future, and there was no time travel machine at hand). I thought that at audio frequencies a simple differential circuit of **Fig.1** should do the job nicely, so that by applying a music signal to the input, a dummy load to the tested amplifier output, and the difference signal inserted into the other amplifier of a stereo pair connected to a loudspeaker would allow me to adjust the potentiometer for a suitable signal cancellation and listen to the distortion only. I was wrong.

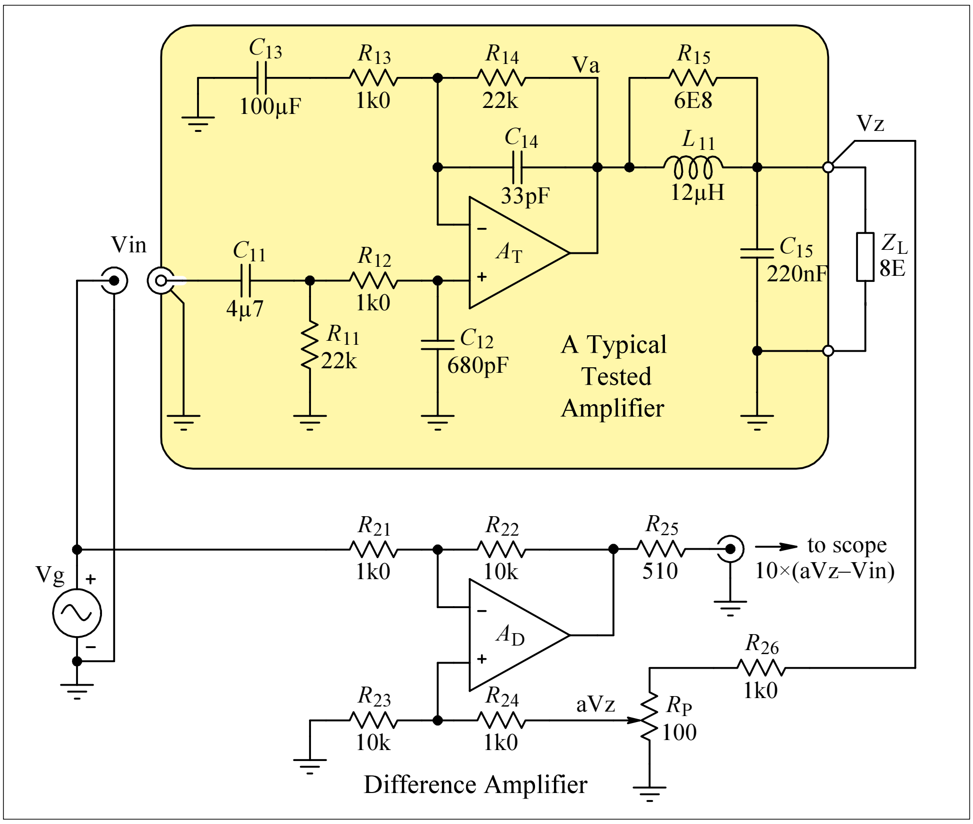

Fig.1: My first unsuccessful attempt to measure distortion by a differential method.

Whichever way and however carefully I was turning the potentiometer, I coul not achieve good signal nulling, a small amount of signal was always present, completely masking the distortion. An experienced older colleague once told me that the main difference between a good and an average engineer is not so much in how good one is in predicting possible mistakes (that's the stuff of a genius), but rather in how quickly one realizes what was his own mistake. I am not sure if there is any truth in this, but when I make a mistake I instantly have a very bad feeling, and often I can also pinpoint the cause.

Circuit Description

Obviously, the output of an RC network will be only 5×10^5 down in amplitude at a frequency two decades away from the cutoff, or sqrt{1/[1+(1/100)^2]}, but the phase difference goes with the arctangent of the imaginary to the real part impedance ratio, so arctan(1/100) = 0.01 radian, and the difference of sin(π) – sin(π+0.01) = 0.01. So one cannot ignore small phase differences.

As shown in **Fig.1**, a typical amplifier under test will usually have an input band limiting filter, as well as a feedback filter, and either a Zobel or a Thiele network type at the load. The load can also be AC coupled occasionally, if not transformer coupled as are tube/valve amplifiers. It is therefore necessary to properly account for those phase differences before attempting to null the signal. So the circuit complexity grows quckly. In principle, it is enough to provide for two or three high-pass filters with a cutoff adjustable somewhere between 1 and 10 Hz, and two low-pass filters with a cutoff adjustable between 20 and 200 kHz.

It would be very difficult to set all the filters to match the tested amplifier response, but that would be necessary only if your test signal were music or other wideband signal. If you limit the choice of input signal to one or two frequencies, the filter system can be simplified to two fixed and one variable section for each band extreme, with the bonus of relatively easy adjustment. In any case the amplifier distortion always increases with frequency, so it makes sense to test it at, say, 2 kHz and 20 kHz. Of course nobody will hear the harmonics of the 20 kHz test tone, but seeing it on the oscilloscope or the screen of a PC running some audio card readout and analysis program will give you a good indication of the amplifier's quality.

One important point I became aware of when working with the prototype was that there is a difference in response when the signal output current from the tested amplifier was larger than the quiscent current, and I wanted to see the transition dynamics. So instead of a constant amplitude sine wave generator, I rather used a band-pass filter, driven by a very low frequency square wave (between 1 and 10 Hz). At every transition of the sqare wave the filter produces a long exponentially decaying ringing at its resonant frequency. To be able to adjust the decay rate a variable Q-factor filter was employed. This resulted in a typical musical transient simulated, like a guitar string or a *pizzicato* played violine.

Such a choice of an input signal turned out to have an important consequence for the interpretation of the amplifier distortion "mechanism", as will be discussed later.

Fig.2 shows the block diagramme of the test circuit. A low frequency square wave generator drives directly a 2 kHz bandpass filter, and a 20 kHz bandpass filter via a small delay. Each filter's ringing signal is fed to a potentiometer and a summing stage, which in turn drives the amplifier under test and the internal reference amplifier. The reference amplifier has an adjustable highpass and lowpass filters to match the bandwidth and the passband phase of the amplifier under test. The two outputs are being fed to a passive attenuator, so they must be of opposite polarity to obtain a signal difference. A signal polarity switch provides the possibility to adapt the test system to either inverting or noninverting amplifiers, even if inverting amplifiers are very rare in the audio market. The difference output is buffered and sent to the output connector. Another amplifier with a gain of 10× is used to amplify the distortion.

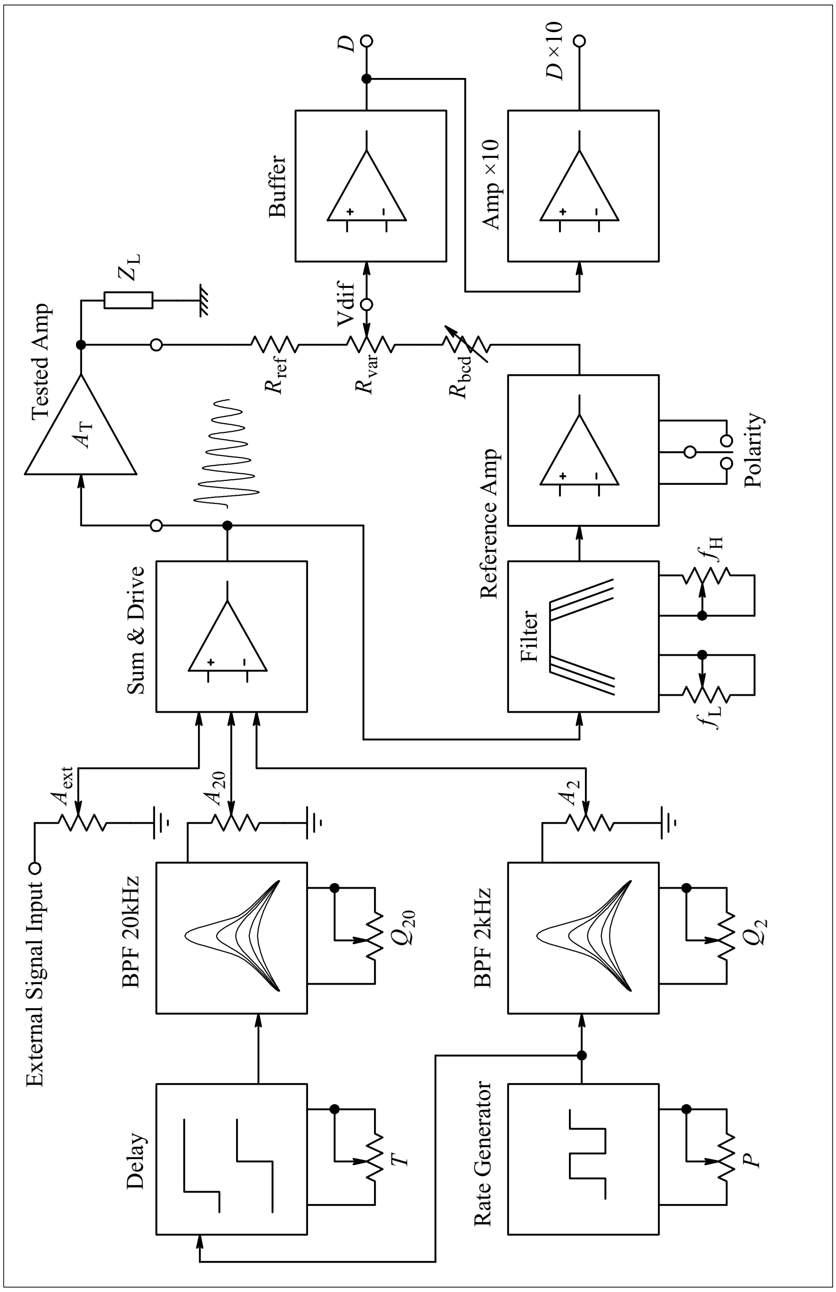

Fig.2: Block circuit diagramme. A low frequency generator drives a 2 kHz, and via a small delay a 20 kHz bandpass filter. Each output ringing signal is level adjusted and fed to a summing amplifier which drives both the amplifier under test and the reference amplifier. The reference amplifier is adjusted in bandwidth and passband phase, and is passively subtracted from the output of the amplifier under test. The difference signal is buffered and amplified by 10× before sending it to the oscilloscope, or a PC sound card. There is also a possibility to use an external test signal or a music source.

My original intention, besides having a separate 2 kHz and a 20 kHz signal source, was to also have a possibility of combining both signals to check for any intermodulation effects in the amplifier under test. For this, a small delay is needed to synchronize the startup of both signals. However this feature was found to be not very useful, so if only a single signal source at a time is used, the delay can be omitted.

Fig.3 shows the low frequency rate generator and the delay generator. A simple CMOS Schmitt trigger NAND gate 4093 with an *RC* feedback network generates the system clock, which is divided by 2 by a 4013 D-type flip-flop to make the repetition rate with a symmetrical duty cycle. The 4538 monostable with the other half of the 4013 D flip-flop generate the delay necessary to synchronize the 20 kHz signal startup with the slower 2 kHz signal, but this function can be omitted if only a single signal will be used at a time. A hex inverting buffer connected in parallel drives a 50 Ohm line for triggering the oscilloscope. The circuit is powered by a symmetrical +/–5V supply in order to generate a +/–5V square wave for the bandpass filters.

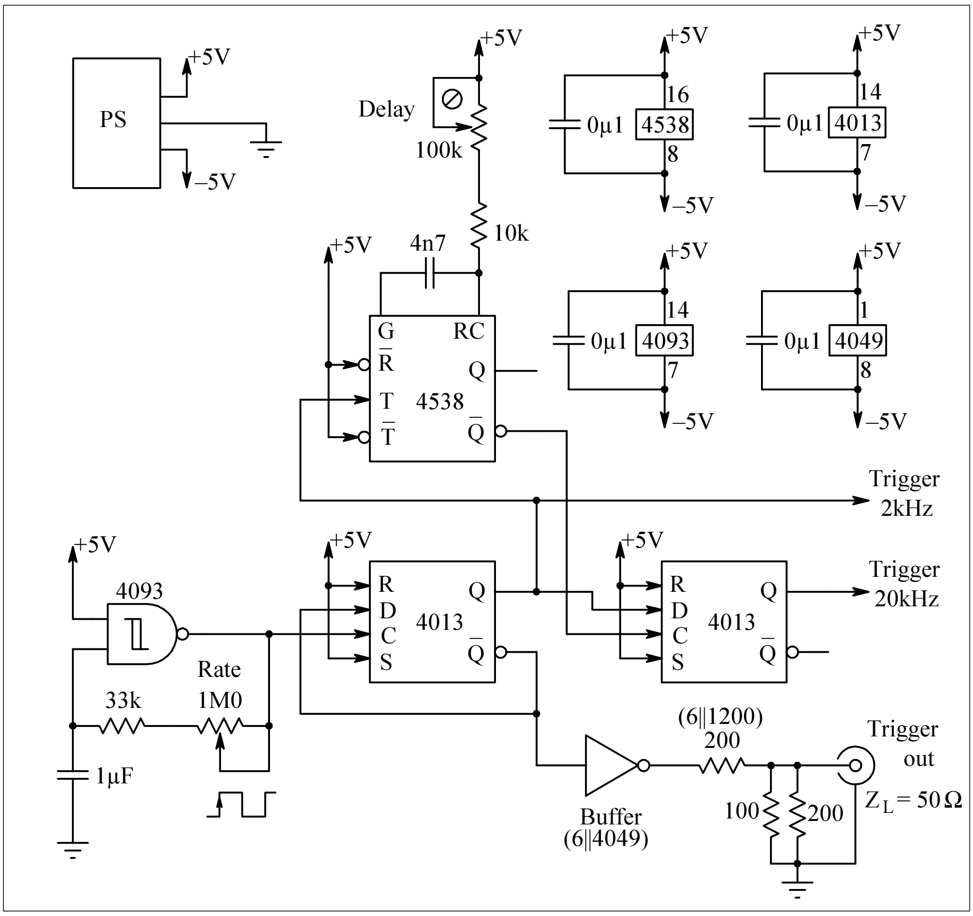

Fig.3: The Low frequency repetition rate generator. Note the symmetrical +/–5V power.

The bandpass filter circuit which provides the ringing signal is shown in **Fig.4**. The component values shown are for the 2 kHz filter. The 20 kHz filter is identical, but for the values of all the capacitors being 10× smaller. In order to achieve a high Q-factor, and thus a long ringing, the bandpass filter uses a two integrator loop configuration, with one integrator phase compensated by an inverting amplifier A2. In order to have a nearly constant initial amplitude for every Q setting, the driving square wave must be applied to the summing point of A2, and so is the Q setting feedback.

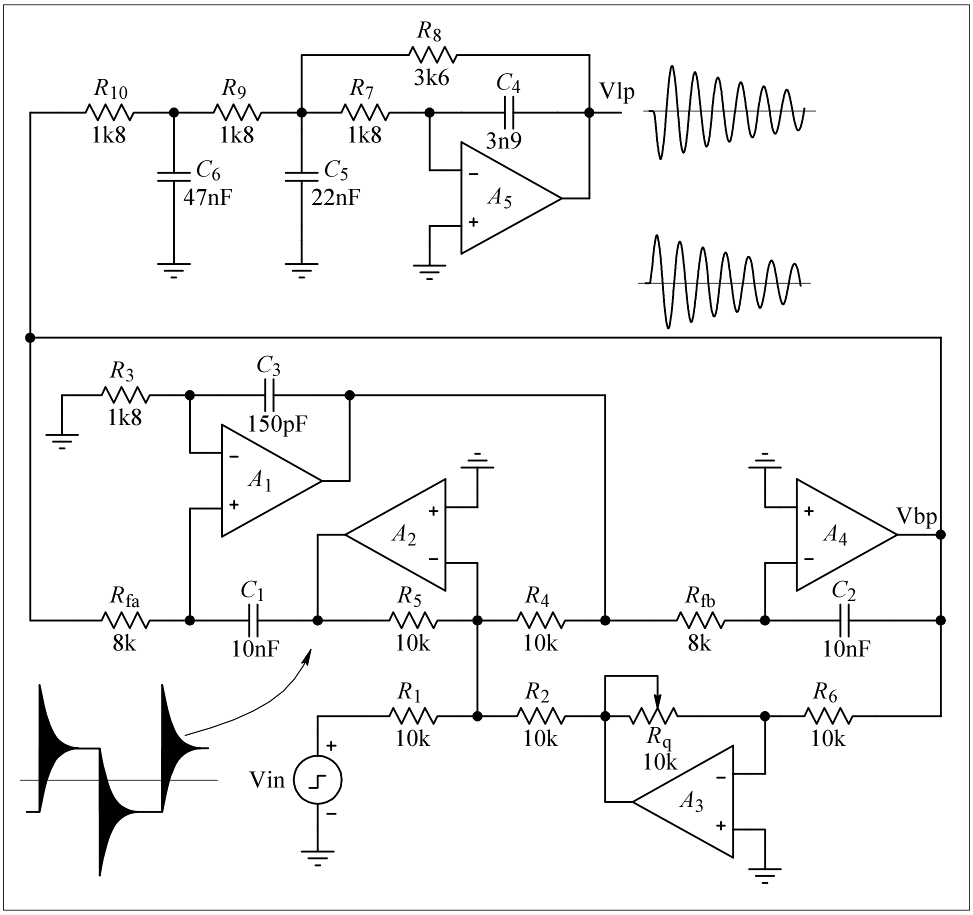

Fig.4: The two integrator loop bandpass filter generates a long ringing at each transition edge of the input squarewave. To achieve the necessary high Q-factor for long ringing, one of the integrators must be phase compensated by an inverting amplifier (A2). The bandpass output from the output of A4 is fed to a 3-pole inverting Bessel lowpass filter, set at 5 kHz (50 kHz for the 20 kHz section) to provide a smooth initial signal transition, and filter out any residues of the digital driving input spikes. $Q = R6/R\text{q}$.

Note that the output of A2 will have a bandpass ringing superimposed on the driving squarewave. The first amplitude of the filter response will be some 90% of the driving signal. Thus for a 10Vpp square wave, the total peak to peak signal will be about 18V (+/–9V). The circuit power supply must be high enough to offer the headroom needed for the combined signal amplitude. A +/–12V power supply should be adequate for the filters, as well for the rest of the circuit.

The bandpass output ringing from the output of A4 is fed to a 3-pole inverting Bessel lowpass filter, whose cutoff frequency was chosen to be 5 kHz (50 kHz for the 20 kHz section). The purpose of this filter is to provide a smooth initial signal transition, instead of an abrupt sinewave startup, as well to filter out any residual spikes of the digital input drive signal. This is important, because the initial abrupt signal transition and the digital signal spike may overdrive the difference section, which may mask the amplifier distortion within the first wave period.

In **Fig.5** the magnitude and the phase response of the 2 kHz bandpass filter and the subsequent 5 kHz lowpass filter for a Q-factor setting of about 20 (Rq = 500 Ohm). Note that the filter gain at resonance is proportional to the Q-factor. With an increasing Q, the filter bandwidth is narrowed, and if the gain were kept constant the total spectral energy would decrease, and so would the initial signal amplitude. By inserting both the input signal and the Q feedback into the summing input of A2, the configuration achieves an almost constant inital signal amplidtude at any Q.

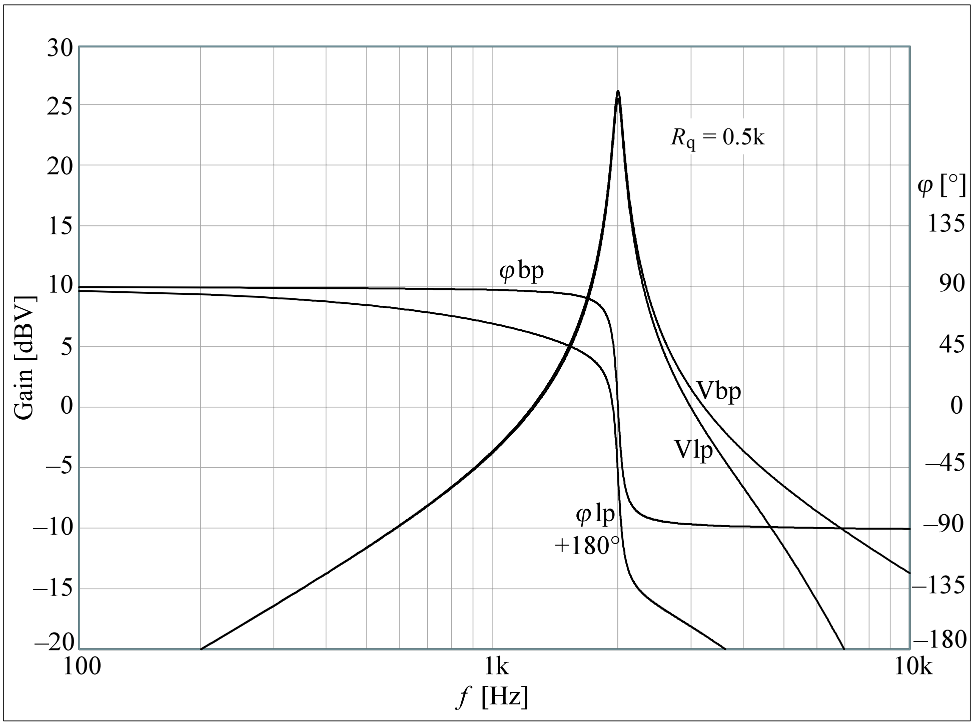

Fig.5: The magnitude and phase angle of the 2 kHz bandpass and the subsequent 5kHz lowpass filter. A similar response is obtained tor the 20 kHz bandpass and 50 kHz lowpass section. Because of the inverting lowpass section, the appropriate phase graph is increased by 180° for comparison. The filter gain at resonance and the sharpness of the phase transition are proportional to the Q-factor setting (about 20, or Rq = 500 Ohm in this graph).

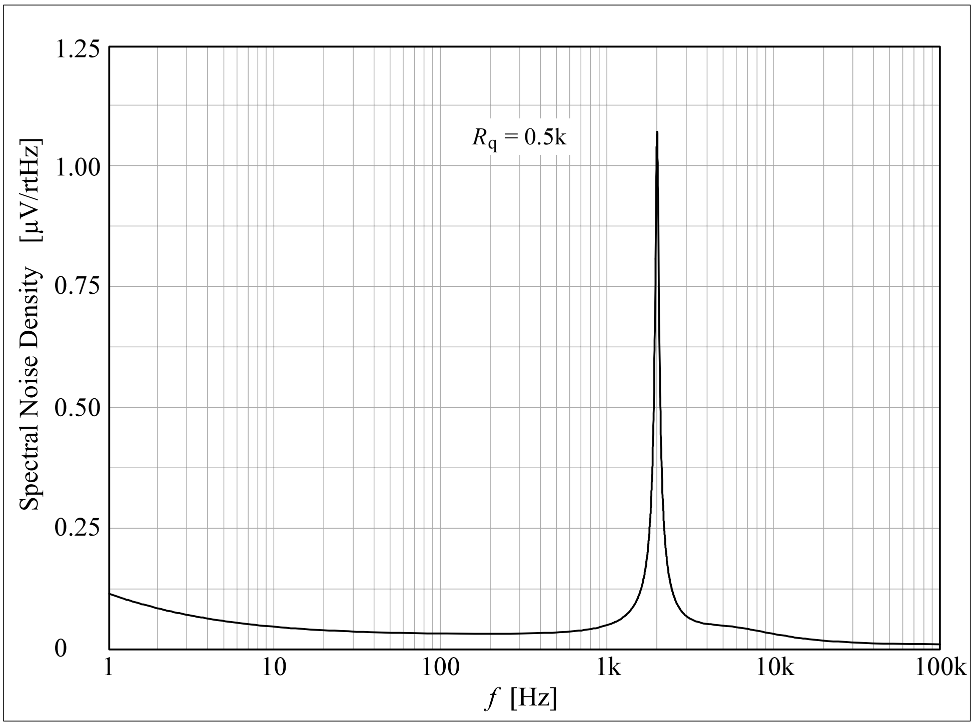

Fig.6: The self noise spectral density of the BP/LP filter for the same Q setting.

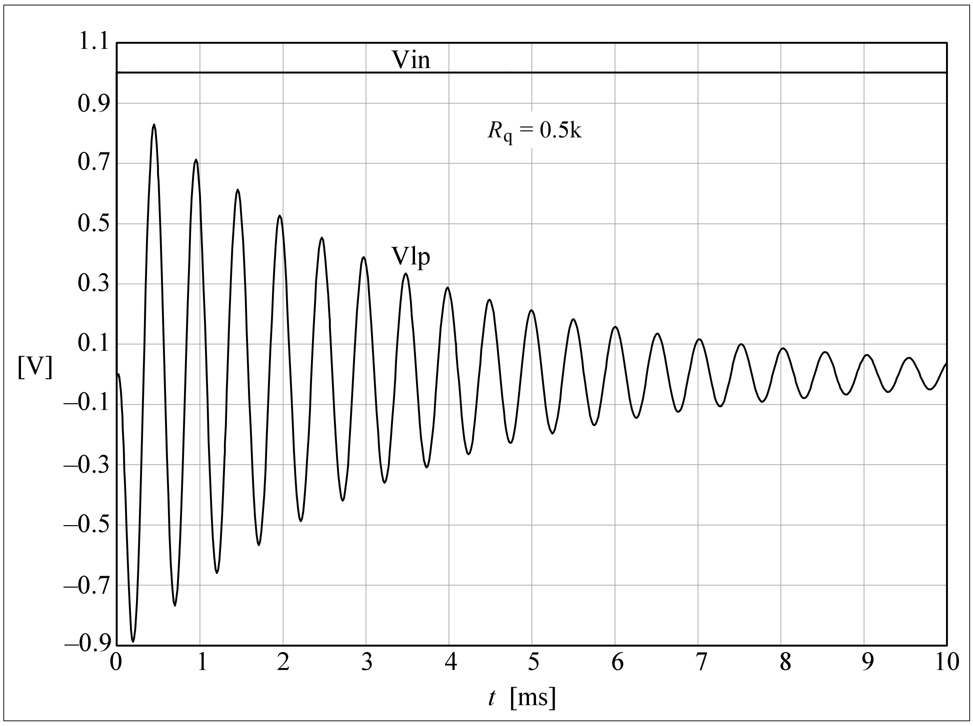

Fig.7: The step response (normalized to 1V) of the 2 kHz bandpass and 5 kHz lowpass section for the same *Q*-factor setting of about 20. For a driving signal of 10V total, an approximately 9V initial amplitude of the ringing signal is achieved.

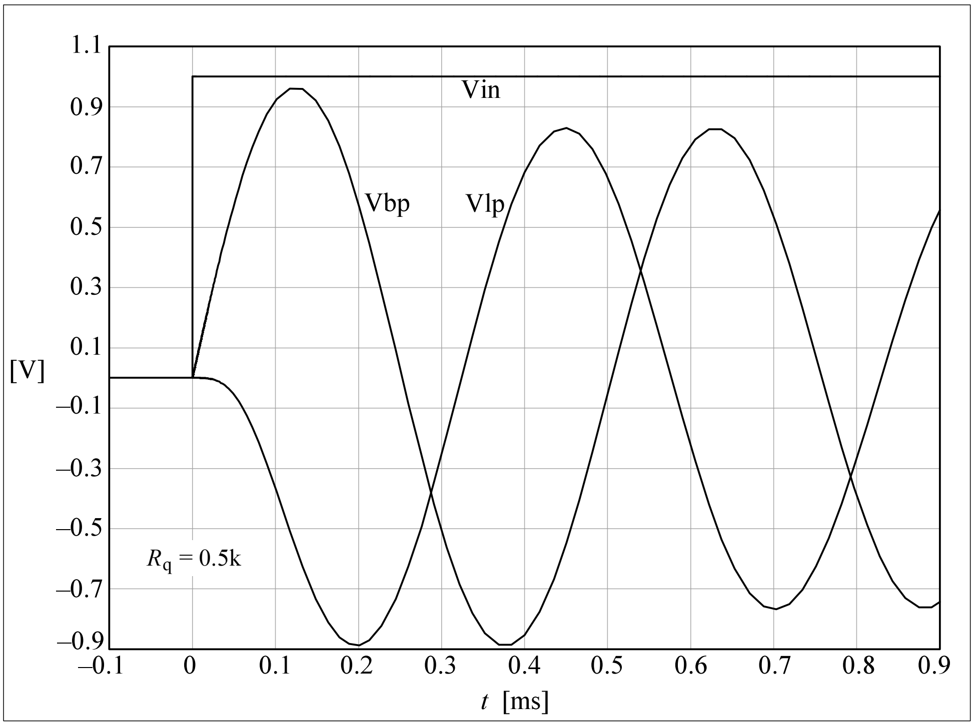

Fig.8: The detail of the initial couple of periods of the step response in Fig.7. Note the abrupt sine wave startup immediately after the input step, and the smoothed transition of the lowpass filter.

Fig.6 shows the filter self noise spectral density, which is less than 50 nV/rtHz out of band. The noise within the filter passband is proportional to the *Q*-factor setting (as the gain), but for very high *Q* the bandwidth is narrowed accordingly, so any in band noise will be indistinguishable from the ringing signal. We are interested in low noise mostly in the region where the ringing decays to <10% of the initial amplitude. With a 9Vpp initial amplitude, the decay region of interest will be about 0.9Vpp, so the spectral noise density of ~1 µV/rtHz withion a ~100 Hz bandwidth about the nominal 2 kHz is going to be down by almost 100 dBV. This is a signal f adequate purity for power amplifier testing.

Fig.7 shows the exponential decay of the 2 kHz signal at a $Q = 20$ during the first 10 ms. In **Fig.8** the first 900 µs of that same signal are shown, together with the input step (normalized to 1V), and the bandpass response, which starts immediately with the input step, whilst the 5 kHz lowpass filter response is much more gradual and inverted.

Fig.9 shows the signal summing and output buffer A6 driving the amplifier being tested (AT), as well as the reference amplifier and passband matching filter A7. The passband matching filter has a coarse and a fine setting potentiometer at each band extreme for good phase matching. The amplifier A8 serves as the signal inversion stage for noninvetring amplifiers under test. For inverting amplifiers the reference output is taken from A7. The output from the tested amplifier is fed to a fixed 60 kOhm resistor (two 120k 1% resistors in parallel), whilst the reference output is fed to a 3 decade adjustable resistor, comprised of three 1-2-4-8 scaling values selected by binary coded decimal (BCD) switches. This arrangement allows for precise difference setting, and in addition the BCD switches code shows the gain of the amplifier under test. The resistor values are selected for the gain set to a decade digit, a unit digit, and one decimal digit. In between there is a potentiometer covering a range of about +/–2 decimal digits to allow for fine tuning. The output Vdif from this potetniometer is sent to the output buffer and another gain of 10× stage to expose the difference signal, as shown in **Fig.10**.

Back in 1986 the best reasonably priced opamp was the SE/NE5534 or the dual 5532, and I therefore used it in the critical driving amplifier, and the reference and polarity amplifiers. In the rest of the circuitry the dual LF353 were used. Today of course there are many low noise and low distortion amplifiers from ADI, TI (BB & National), LT, and others. The choice is yours.

One important thing to watch out is not to take the amplifier under test speaker ground for the ground return of the output cable. Because of the high output current, this ground point might not be at the same level as the input signal ground, thus a ground loop with some of the output current circulating through the distortion measurement system ground and to the input ground may impair proper signal cancellation and the distortion measurement. It is a better choice to have both cable shields tied to the signal input ground of the amplifier under test, as indicated in **Fig.9**.

The use of the system is straightforward, the only thing to care about is to provide a suitable input signal level to the amplifier under test. The +/–4.5V signal provided by the driving amplifier should be more than enough for any audio power amplifier, but for many it may be too much and will drive the tested amplifier into clipping. Pay attention to your neighbors and use a dummy load instead of loudspeakers. You can use two 55W automobile head lamps connected in series, which will give you visual feedback of the amplifier's output, but the impedance of those lamps will vary widely with temperature, and so will the amount of output distortion. It is thus beter to use a large number of ceramic coated power resistors of a suitable value in parallel/serial combination to give you the required nominal load impedance for your amplifier. If you must have audio indication of the test signal, use a suitably attenuated output to drive a loudspeaker.

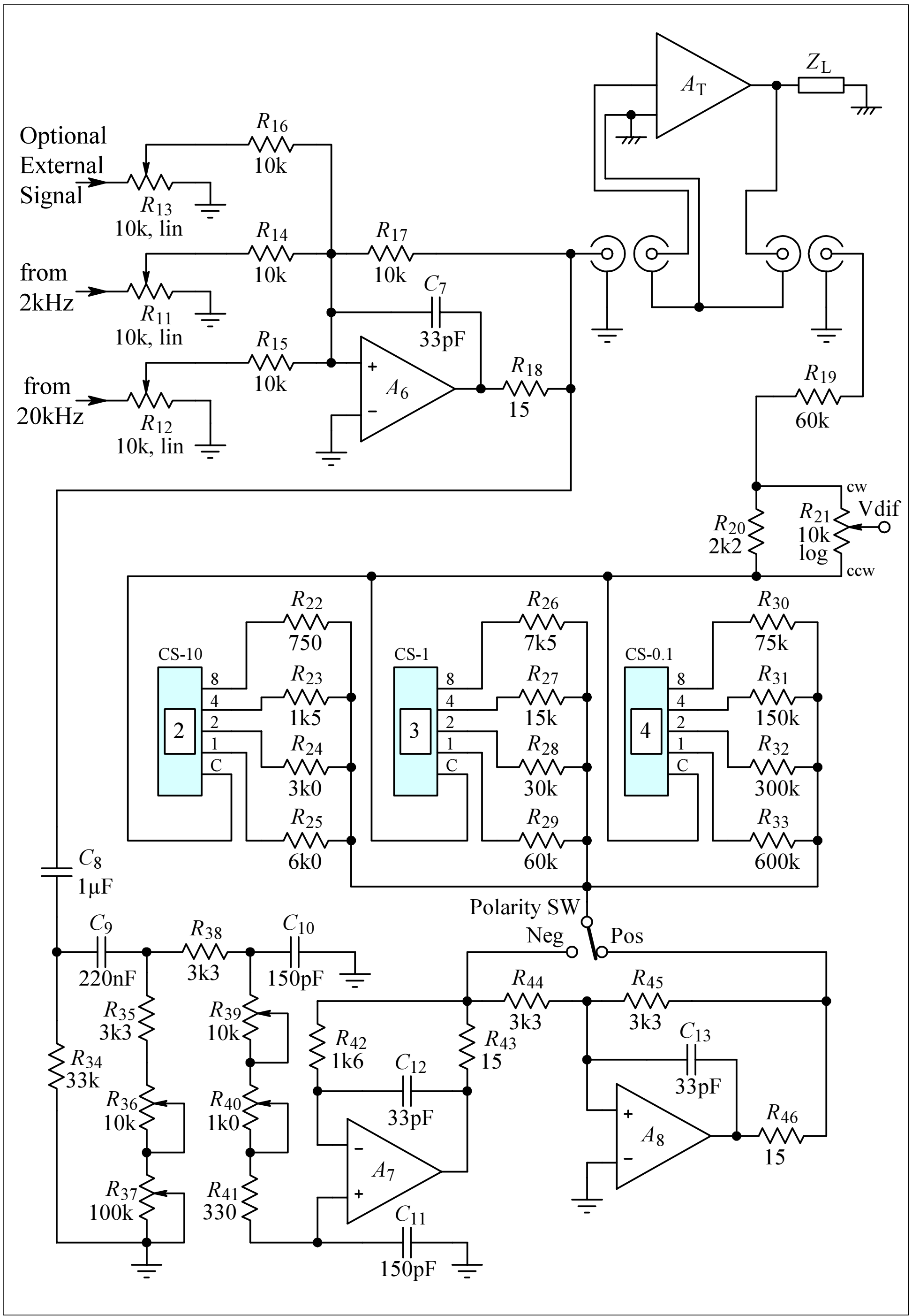

Fig.9: the driving and the difference stage. The BCD switches tell the gain of the amplifier under test. The output potentiometer serves for fine tuning the signal difference within a +/–2 decimal digits range. The lowpass and highpass filters in the referenc amplifier input must be set for good phase matching with the amplifier under test. The polarity switch must be set so that the output of the tested amplifier and the reference amplifier are of opposite phase in order to allow signal cancellation by the digital attenuator and the fine tuning potentiometer.

If you are planing to use a PC or a laptop Sound Card line input for distortion viewing and analysis, ground noise from the computer power supply can be a big problem, so it will be a good idea to decouple the ground noise by a differential driver shown in **Fig.11**. Since the ground decoupling amplifier inverts the signal, one more signal inverter is necessary to preserve the signal polarity.

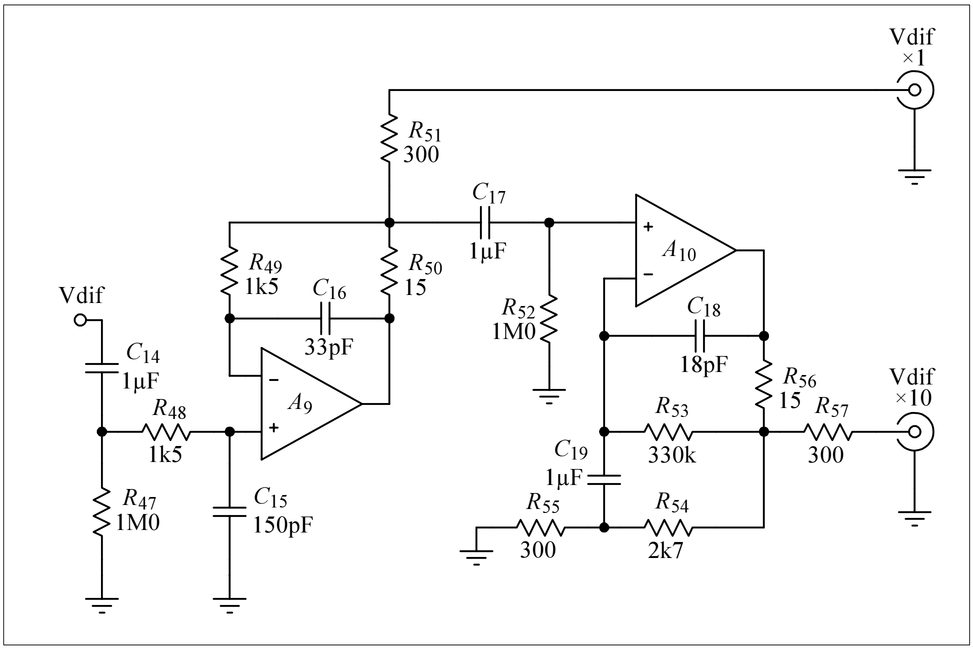

Fig.10: Difference signal buffer and a gain of 10× stage.

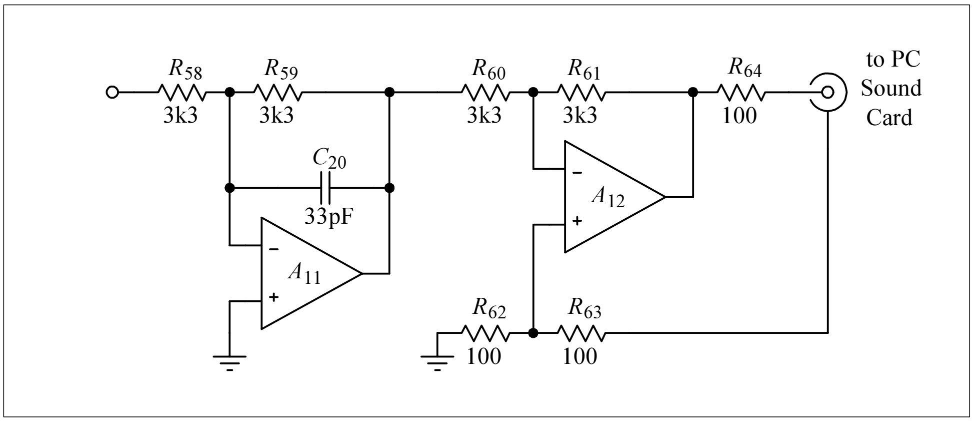

Fig.11: PC or laptop Sound Card output ground decoupling.

Obviously the signal cancellation will be an iterative process. First the coarse digital attenuation must be set to about the required amplifier gain, then the coarse low frequency and high frequency cutoff of the reference amplifier will be adjusted, then maybe a little more last digit attenuator adjustmet, and then the fine phase and level controls will be played with to achieve a minimum of output signal. This last procedure will probably need to be repeated many times. But let us see more of this in an actual example.

Achieving the Correct Nulling

How can we possibly know when a correct signal nulling (both in amplitude and phase) has been achieved? The usual method is to simply minimize the RMS value of the distorted signal, and this is done using a constant amplitude sinewave, with the amplitude high enough that the output current becomes considerably larger than the quiscent current in the output transistors of the teste amplifier. This condition ensures that under extreme signal excursion the transistor which does not supply the current to the load is effectively switched off in class B operation, with the consequence of being too late to catch up with the signal when it is required to conduct again. The indication of this is the appearing of "spikes" in the distorted signal whenever the output signal crosses the zero level.

Fig.12 shows one such example in the middle trace. However, with an exponentially deaying signal it becomes immediately apparent that as the signal amplitude falls below the quiscent threshold a relatively large uncompensated phase error appears in the undistorted signal region. That the error is a phase error, and not just an amplitude nulling error, can be verified by turning the appropriate potentiometers; and anyway, the oscillations of the middle trace in the undistorted region lags the top trace in phase by 90°!

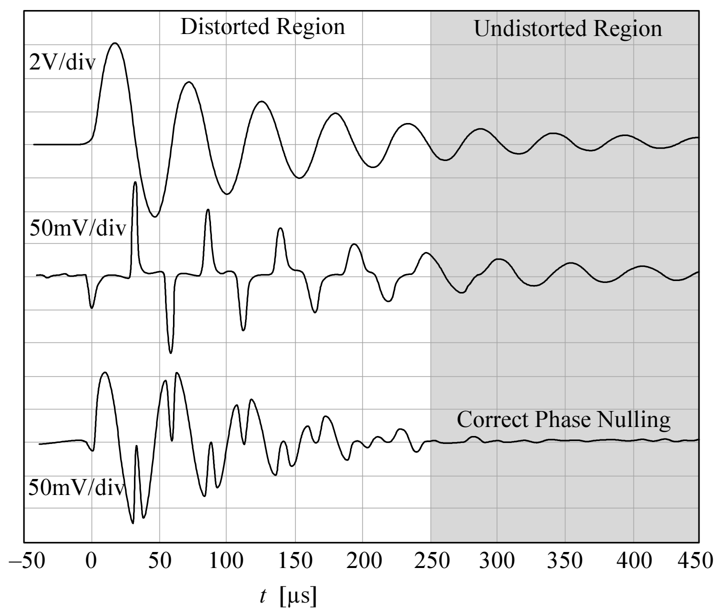

Fig.12: This is a reproduction of a graph from my 1987 article, showing in the top trace the actual output signal from a typical class B amplifier, the middle trace shows the crossover distortion "spikes" with the usual signal nulling technique, and the bottom trace shows the crossover distortion with the phase error using a correct phase nulling. The signal frequency was ~18 kHz, probably because I was using 5% tolerance resistors and 20% tolerance capacitors, but this does not matter here. See text for more details.

Now if the phase is correctly nulled in the undistorted region, a large phase error appears in the distorted signal region, as shown in the bottom trace of **Fig.12**. There is only a single logical explanation for such a behaviour: as the amplifier distorts the signal, the feedback loop must react accordingly, but it can do so only at its own bandwidth, which is the amplifier's open loop bandwidth.

This is because under distortion the feedback signal is momentarily severely reduced or lost. Because the open loop bandwidth of an ordinary amplifier is somewhere between 10 and 100 Hz, the reaction time is slow, and at high signal frequencies a large difference between the input and output signal appears before the feedback loop is reestablished. Effectively this means that as the amplifier distorts, its bandwidth is reduced and its phase lag is increased, but once the signal amplitude falls below the quiscent threshold, the distortion is reduced considerably, and so is the phase lag.

What we have is actually a **switching phase modulation** mechanism. Note that the amplitude of the distorted signal follows the output signal amplitude, and the phase angle is practically constant (but within the first quarter of the first period, where the initial phase modulation occurs), until it suddenly vanishes.

How large is this pahse error? It is possible to calculate the phase angle from the arctangent of the imaginary to real component ratio, the real part being the output signal, and the imaginary part being the distorted signal in the bottom trace of **Fig.12**. By observing the second period of the signal, where the phase relation is firmly established, we have about 0.1V in the distorted signal and about 4V in the output. The angle is then (180/pi)×arctan(0.1/4) = 1.5°. Not much, you will say. But bear in mind that the human ear is insensitive to static phase shift, but very sensitive to quick phase changes.

Note also that the amount of phase error is load impedance dependent, a dominantly capacitive load increases both the crossover distortion and the phase lag, a dominantly inductive load reduces both. Also, the effect is increasing with signal frequency. With musical signals at moderate levels, this phase switching always occurs during the initial 10-50 ms or so at every start of a played note. It may therefore be conjectured that this is the main cause for the often complained "stereo image instability" experienced during subjective evaluation, even if the of crossover distortion itself is unnoticeable.

This is also the reason for having always a small current in the transistor which does not supply the current to the load. There are several ways of doing it described in the literature, so pick your own choice for your next amplifier project.

References:

[1] *D. Self*: Audio Power Amplifier Design Handbook, Newnes, 2006

[2] *K. Castor-Perry*: Look, Ma, No Distortion! The Class i output stage, Linear Audio, Vol.2, Sept. 2011

[3] *E. Margan*: Crossover Distortion in Class B Amplifiers, Electronics & Wireless World, July 1987 <http://www-f9.ijs.si/~margan/Articles/Class_B_Dist.pdf>

[4] *N.S. Pass*: Active bias circuit for operating push-pull amplifiers in class A mode, US Patent 3 995 228, 1976, Nov. 30.

[5] *S. Tanaka*: New biasing circuit for class B operation, Journal of the Audio Engineering Society, 1981, March, p.148.

[6] *P.J. Baxandall*: Audible amplifier distortion is not a mystery, Wireless World, 1977, Nov., p.63.

[7] *R.S. Cordell*: The distortion Magnifier, Linear Audio, Vol.0, Sept. 2010