Ironing out DISTORTION

As is often the case with articles on audio subjects, Douglas Self's[#] recent series on amplifier distortion caused a great deal of interest worldwide. Building on Doug's work, Edward Cherry offers an in-depth look at distortion, and discusses how to reduce it.

At the outset I should state that there appears to be something of a philosophical difference between our approaches to distortion inside the feedback loop. There is nothing wrong with tackling the distortion of various stages on an individual basis, but my approach to designing a high-quality amplifier is to choose a simple topology based on common-emitter amplifying stages and apply negative feedback to reduce distortion. Variations in circuit topology (other than push-pull operation) rarely give better than a ten-fold reduction in distortion on a production basis; feedback, however, can reduce distortion almost indefinitely[1].

The beautiful thing about feedback is that it reduces all distortions simultaneously. If there is enough feedback to fix the major sources of distortion, the minor sources will be taken care of automatically. However, as Self points out, feedback cannot correct distortions arising outside the feedback loop.

Common-emitter stages have theoretical advantages over common-collector amplifiers[3] and, in my opinion, have important practical advantages too. The theoretical basis for my positions regarding both feedback and common emitter stages rests ultimately on the work of Bode[4], who points out that a common-collector stage can be considered as a common emitter stage with a kind of local feedback, which rarely accomplishes anything for the stage within the local loop and usually makes matters worse for the stages outside it. In part, the complex behaviour of distortion in Self's amplifiers is attributable to the local feedback.

Incidentally, I take it as now being universally accepted that there is no basis for linking transient, interface and phase intermodulation distortions to large amount of feedback.

This commentary addresses audible distortions only; that is, nonlinearities which generate distortion products in the audible frequency range 20Hz to 20kHz. It is not concerned with nonlinearities that generate ultrasonic distortions, for I am not in the business of trying to '...please any passing bat...'![2]

Distortion products can arise as harmonics of a single input frequency, or from intermodulation between two or more simultaneous inputs, in which case the distortion products lie above and/or below the input frequencies. Ultrasonic distortion in an amplifier may, of course, be an indicator of trouble at audible frequencies, but not necessarily; what matters is the presence or absence of audible distortion products.

Indeed, I believe the 20kHz upper limit should be reduced, because practically no-one can hear distortions at even 15kHz. As an easy demonstration of this assertion, readers could compare the sounds of a 5kHz square wave with an accurate 1:1 mark-space ratio and a 5kHz sine wave of 1.273 times the peak-peak amplitude. The square wave contains a 5kHz component of the same amplitude as the sine wave, but it also contains a component at 15kHz of one-third the amplitude. Shift the frequency up or down to find your own frequency limit for audible distortion. Be honest, and remember that this is the equivalent of 33% third-harmonic distortion!

Distortion analysed

Figure 1 is Self's Fig. 1a from *EW&WW* of August 1993, with some minor changes. I analysed it in May 1982[5] and a main purpose of this commentary is to point out that many of Self's conclusions have a rigorous mathematical basis. In addition, given that an audio amplifier is to be of this basic topology (and it would not be my first choice), then the simple changes shown in **Fig. 2** give substantial improvement at little cost. Figures 1 and 2 can, of course, be flipped upside down, with n-p-n and p-n-p transistors interchanged; the analysis is identical. It is actually the flipped version that is considered in Reference 5.

Nonlinearity, sensitivity and distortion

Distortion can be considered as variation of incremental gain from point to point on the signal waveform. In other words distortion is caused by nonlinearity in parameters like ß and g_m. For example, when an amplifier is driven near the point of clipping, its incremental gain falls at the waveform peaks; these

[#] *Distortion in power amplifiers. Doubt Self. August, 1993 to March, 1994.* [*] *keith@keith-snook.info*

Definition of terms

The analysis in Reference 5 is in terms of the effective values of ß and g_m for each stage.

Current amplification factor ß of a transistor is defined formally as the gradient of a graph of collector current I_C versus base current I_B. Nonlinearity in ß is any departure of the graph from a straight line, from any physical mechanism whatsoever. In practice there are many such mechanisms, and the magnitude of ß falls at both large and small currents.

Mutual conductance g_m is similarly defined as the gradient of a graph of collector current I_C versus base-emitter voltage V_{BE}. Nonlinearity in g_m is any departure of the graph from a straight line. One physical mechanism for such a departure is the exponential I_C *versus* V_{BE} characteristic, inherent in bipolar junction transistors and which results in the well-known formula

$$g_m \approx qI_C / kT \tag{1}$$

where kT/q is approximately 25mV at room temperature.

Although these definitions of ß and g_m are most often applied to an intrinsic transistor (a transistor from which parasitic elements such as base-spreading resistance r_B have been removed), basic formulae like Eq. 1 can be adapted to a complete transistor:

$$g_{m(\text{eff})} \Rightarrow \frac{1}{r_B/\beta + kT/qI_C} \tag{2}$$

For the Darlington transistor in **Fig. 3**,

$$g_{m(\text{eff})} \Rightarrow \frac{1}{[r_{B(\text{a})}/\beta_{(\text{a})} + kT/qI_{C(\text{a})} + r_{B(\text{b})}]/\beta_{(\text{b})} + kT/qI_{C(\text{b})}} \tag{3}$$

where the (a) and (b) subscripts identify parameters of the individual members. The formulae can even be adapted to include local emitter degeneration

$$g_{m(\text{eff})} \Rightarrow \frac{1}{kT/qI_C + R_E} \tag{4}$$

or a resistance R_S between base and emitter:

$$\frac{1}{\beta_{\text{eff}}} = \frac{1}{i_C/i_S} = \frac{1}{\beta} + \frac{R_E}{R_S(1 + qI_C R_E/kT)} \tag{5}$$

neglecting r_B for simplicity. Equation 5 is a particularly useful analytical trick, since it allows the output resistance of one transistor or stage (a kind of source resistance) to be incorporated into the effective ß for the next.

In all such cases, 'base' and 'collector' currents and 'base-emitter' voltage are measured at the effective terminals of the device, and include the currents in shunt resistances or the voltage drops across series resistances. Nonlinearity in effective g_m then includes any nonlinear component of voltage drop across r_B, and therefore involves the nonlinearity in ß; note that ß occurs in Eqs. 2 and 3.

peaks are amplified less than the rest of the input waveform, and the output is 'squashed'.

Sensitivity is the ratio of a percentage change in some parameter like ß or g_m to the resulting percentage change in incremental gain. At any point on a signal waveform the instantaneous voltage and current in, say, the output transistors of an amplifier can be found. Hence the fall in, say, ß at the signal peaks can found from the known nonlinearity. Then, if sensitivity to changes in ß is known, the gain compression can be calculated and ultimately distortion can be predicted quantitatively. For example, if the gain compressions at the positive and negative peaks of a signal waveform are γ' and γ'' respectively, the second and third harmonic distortions are:

$$D_2 \approx \frac{\gamma' - \gamma''}{8} \tag{12a}$$

$$D_3 \approx \frac{\gamma' + \gamma''}{24} \tag{12b}$$

Figure 4, reproduced from Reference 5, shows the sensitivity of the overall gain of Fig. 1 to changes in parameters, as functions of frequency on logarithmic scales. Many of the labelled points have the physical significance of quantities like mid-band loop gain considered above. Numerical values are calculated for,

- β_2 = 100 (typical);
- β_3 = 5000 (a typical Darlington);
- g_{m1} = 4mA/V (typical for bjts operating at a few hundred microampères without emitter

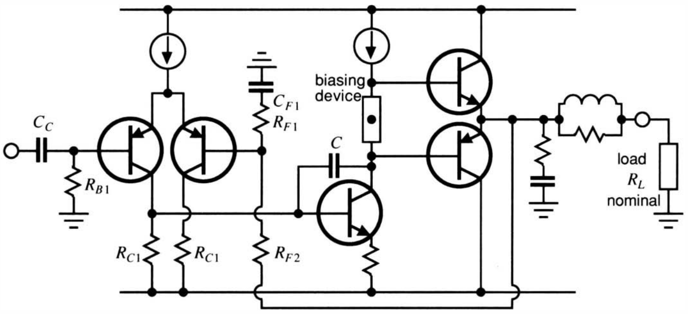

Fig.1. Outline circuit of many audio power amplifiers. This was Self's Fig. 1a.

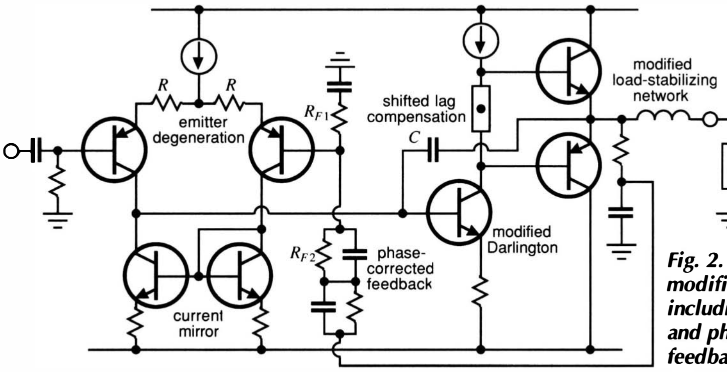

Fig. 2. Suggested modifications of Fig. 1, including a current-mirror and phase correction in the feedback loop are shown to possess advantages.

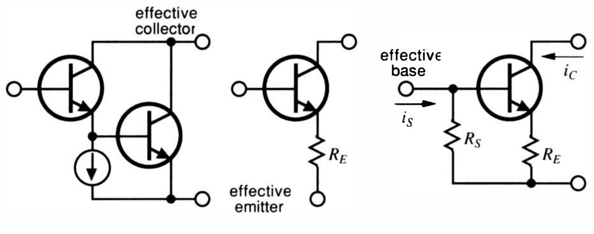

Fig. 3. Some examples of sub-circuits for which effective ß and gm can be defined: Darlington, emitter degeneration, and degeneration plus shunt resistance.

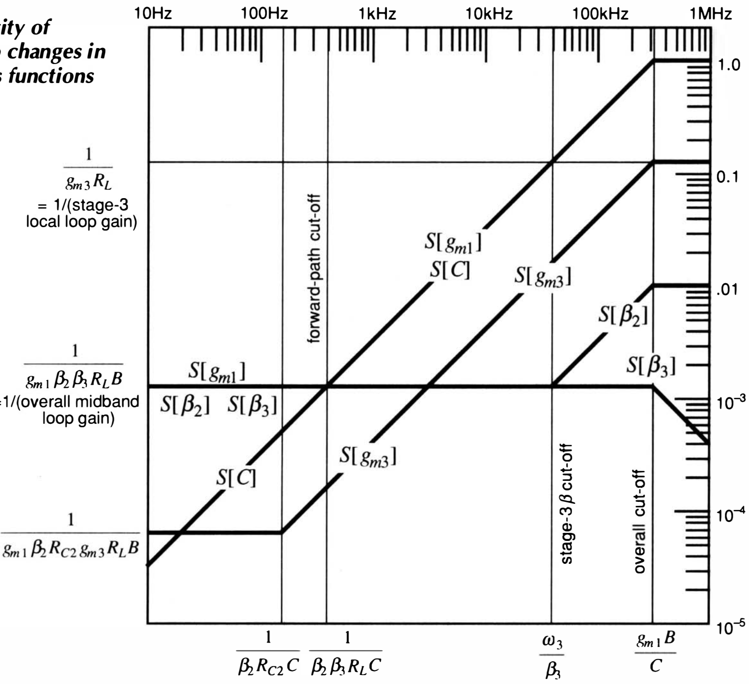

Fig. 4. Sensitivity of overall gain to changes in parameters, as functions of frequency.

degeneration, or at larger currents with degeneration);
- g_{m2} = 10mA/V (corresponding to about 68Ω in the second-stage emitter as part of the protection circuitry, Eq. 4);
- g_{m3} = 1A/V (corresponding to about 0.68Ω ballast in each of the third-stage emitters, Eq. 4);
-ω_3 = 1Gr/s (typical for a Darlington which consists of a reasonably fast first member and a slow second member);
- C = 100pF (to set the overall 3dB bandwidth at 300kHz, Eq. 8);
- B = 0.05 (corresponding to an overall mid-band gain around 20, perhaps R_{F1} = 2k2Ω and R_{F2} = 47kΩ);
- R_{C2} = 100kΩ (a guess, but it hardly affects the results)
- R_L = 8Ω nominal load.

First stage

As stated by Self, signal amplitude in the first stage increases in proportion to frequency above the forward-path cut-off $1/\beta_2\beta_3RLC$ where overall loop gain falls away. The nonlinearity of g_{m1} is therefore more strongly exercised as the frequency increases. Simultaneously, sensitivity to changes in g_{m1} increases with frequency as shown in Fig. 4. Therefore, overall distortion rises with frequency, either as the square or cube, depending on details.

In my opinion,* Self's discussion of input stages is over-kill. Despite the rapid increase of distortion with frequency, the simple long-tailed pair with emitter degeneration shown in **Fig. 5** contributes vanishingly small audible distortions. Fancy topologies are simply not required.

Emitter degeneration. If a feedback amplifier is fed with a fast-rise mid-frequency square wave, the peak-to-peak input to the first stage is twice as large as the square wave itself. Therefore, if an amplifier is not to go into slew-rate limiting (alternatively, is not to generate hard transient intermodulation distortion) when fed with a full-amplitude fast-rise square wave, its input stage must be designed not to clip on a signal twice the amplitude of rated mid-band sinusoidal input to the complete amplifier[6]. This result is independent of the overall bandwidth and slewing rate.

Taking the numerical values in Fig. 5 as an example, g_m is about 4mA/V and maximum output is about 8mApk-pk; the stage clips at about 2Vpk-pk input. Therefore, this circuit should not be used as first stage in an amplifier rated at more than 1V pk-pk input (350mV rms) for full output. Suppose this input is at 6.67kHz, so that the third harmonic is at 20kHz (our upper limit for audible distortion), and suppose that other parameters are as in Fig. 4. Then:

- Overall loop gain at 6.67kHz is 46.
- Therefore, the differential component of the input to the first stage is overall input/loop gain, which is 22mVpk-pk.
- Therefore, signal current in each transistor is 1/2×g_{m1}×(differential input) = 47µApk-pk.
- Compression of incremental gain at this peak current is 0.0075%, found from the formula for effective g_m of a degenerated long-tailed pair:

$$g_{m(\text{eff})} \Rightarrow \frac{2}{kT/qI_{C(\text{left})} + kT/qI_{C(\text{right})} + 2R_E} \qquad (13)$$

- But, from Fig. 4, sensitivity towards changes in g_{m1} at 6.67kHz is 0.022.
- Hence the (equal) compressions γ' and γ'' of overall gain at the signal peaks = sensitivity × first-stage compression = 0.00016%.
- Hence overall third harmonic at 20kHz associated with first-stage nonlinearity is 0.000013%, from Eq. 12b. This is at least a factor of ten smaller than the wildest suggestion I have ever seen as a target figure for an 'ideal' amplifier.

An analytical approach allows one to classify for certain the nonlinearities in an ampli-

Basic equations

With the notation shown in the Definitions panel, the main features of the small-signal response of Figs 1 and 2 are determined by just four components:

- the overall feedback resistors R_{F1} and R_{F2} via the overall feedback factor B:

$$B = \frac{R_{F1}}{R_{F1} + R_{F2}} \qquad (6)$$

- the first-stage mutual conductance g_{m1}
- the second-stage lag-compensating capacitor C.

Overall mid-band gain

$$A_{\text{mid}} \approx \frac{1}{B} \qquad (7)$$

Overall high-frequency 3dB cut-off

$$\omega_{3\text{dB}} \approx \frac{g_{m1}B}{C} \qquad (8)$$

Forward-path mid-band gain

$$G_0 \approx g_{m1}\beta_2\beta_3R_L \qquad (9)$$

Forward-path high-frequency 3dB cut-off

$$\omega_0 \approx \frac{1}{\beta_2\beta_3R_LC} \qquad (10)$$

Mid-band loop gain

$$A_L \approx g_{m1}\beta_2\beta_3R_LB \qquad (11)$$

where,
- ß$_1$ (not used here), ß$_2$ and ß$_3$ are the effective current amplification factors of the transistors, including the effect of any series or shunt resistors, in the first, second and third stages respectively;
- g_{m1}, g_{m2} and g_{m3} are the effective mutual conductances of the transistors (including the effect of resistors);
-ω_1, ω_2 (neither actually used here, but used in the JAES paper) and ω_3 are the projected gain-bandwidth products of the transistors;
- for later use, R_{C2} is the equivalent resistance (ideally infinite) of the second-stage current-source load.

The substance of these results is the same as given by Self.

fier as significant or insignificant contributors to audible distortion. Repeating the above calculation using a 20kHz input (the popular 20kHz thd test beloved of spec. men), gives the third harmonic as 0.0004% – not nearly so impressive, and even casting doubt on the intermodulation performance with real programme material. But this harmonic is, of course, at 60kHz and of itself has nothing to do with audible distortions.

In one of the more savage forms of the IEC total-difference-frequency intermodulation test, the input consists of two equal-amplitude sine waves at approximately 10kHz and 15kHz. For Fig. 5, the total of the audible intermodulation products near 5kHz is 0.000008%; the inaudible products near 25, 30, 35, 40 and 45kHz are larger, but it is the audible distortion that matters. To chase anything better than Fig. 5 would be folly.

The theoretical requirement that a complete amplifier should be able to accept a full-amplitude fast-rise square-wave input is unnecessarily severe; real programme material (even the output from a digital synthesiser) is subject to some form of band limiting. Typically I relax the requirement by about a factor of two. This increases 20kHz third-harmonic distortion of 6.67kHz by four, to 0.00005%.

However, I do regard some form of emitter degeneration as mandatory in a bjt first stage (perhaps not with fets). If there is none, a bjt long-tailed pair clips at about 100mV pk-pk input. Ideally, therefore, an amplifier that uses an undegenerated first stage should be designed to operate with no more than 50mVpk-pk or 18mV rms input. Self's Fig.1a of August 1993 would certainly generate hard transient intermodulation distortion in the square-wave test.

Including adequate emitter degeneration in the first stage carries the penalty of a slight increase in noise. The numerical situation for Fig. 5 is confused because total noise is dominated by the current mirror. However, if mirror noise could be eliminated (it can, but not in an amplifier of the Fig. 1 type), then the thermal and shot noise noise referred to the input of Fig. 5 would be 5.6nV/√Hz. After removing the emitter degeneration and adjusting the quiescent current to give the same gain, the noise drops about 1dB to 5.0nV/√Hz.

Current mirror. Self correctly points out that a current mirror in the first stage, rather than a simple resistance load: doubles the first-stage gain g_{m1} and therefore the overall feedback if nothing else is changed; doubles the available output current and hence the slewing rate; and improves the common-mode rejection.

Far more importantly, it raises the source impedance seen by the second stage. I shall show that distortion associated with output-stage ß nonlinearity is inversely proportional to effective ß of the second stage. Raising the source resistance for the second stage by using a current mirror in the first stage has the potential for increasing $ß_2$ (Eq. 5) and reducing this distortion by orders of magnitude.

Increased first-stage gain is a mixed blessing because it may provoke high-frequency instability. Self's suggested remedies are to double the first-stage emitter degeneration resistors, which is wasteful if these are already adequate, and it increases the noise; and to double the compensating capacitor C, which loses the slewing-rate improvement.

However, there is a third solution, which I strongly recommend: halve the value overall feedback factor B (halve R_{F1}, for example), thereby doubling the overall mid-band gain and halving the signal input voltage required to produce full output from the amplifier. This has two advantages in that it halves the common-mode voltage present in the first stage, and thereby halves a second-harmonic distortion mechanism associated with finite common-mode rejection; and it reduces the likelihood of clipping in the first stage, thereby reducing the incidence of hard transient intermodulation distortion.

Halving the input voltage required for full output, from something like the typical 0.6-1.0V of modern transistor amplifiers to 0.3-0.5V, makes the transistor amplifier more like the earlier Leak/Mullard/Quad vacuum-tube amplifiers. Why was this 'standard' ever changed? My best amplifiers are designed with 300mV sensitivity.

Second stage

Distortion in the second stage originates from three quite distinct types of nonlinearity:

- distortion associated with variation of effective $ß_2$ from point to point on the signal waveform, as the instantaneous current and voltage in the transistor change;
- distortion associated with similar variation of effective g_{m2} (this turns out to be very small);
- distortion associated with variation of the collector-base capacitance.

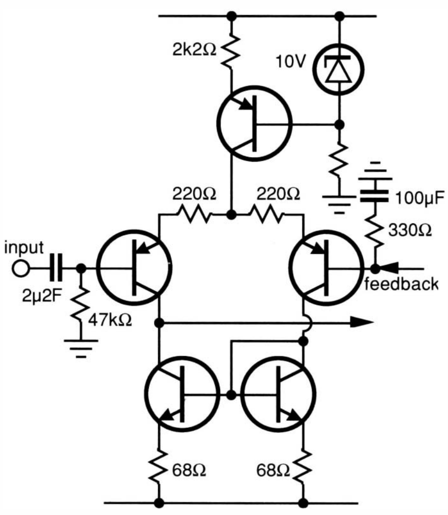

Fig. 5. First stage, with emitter degeneration, current source and current mirror, and with typical numerical values as used by the author.

Distortion associated with $ß_2$. Nonlinearity in $ß_2$ models the changes in current gain at high and low collector voltages and currents. Figure 4 shows that sensitivity to these changes is constant over most of the audible band of frequencies, but increases somewhere near the top of the band at the ß cut-off frequency $\omega_3/ß_3$ of the output transistors (40kHz for the assumed data). However, sensitivity is inversely proportional to $ß_2$; distortion from this nonlinearity can be reduced simply by making $ß_2$ large – for example, by using a Darlington. Sensitivity and distortion are not affected by the choice of lag compensation C, and hence are independent of both the forward-path and overall high-frequency cut-off.

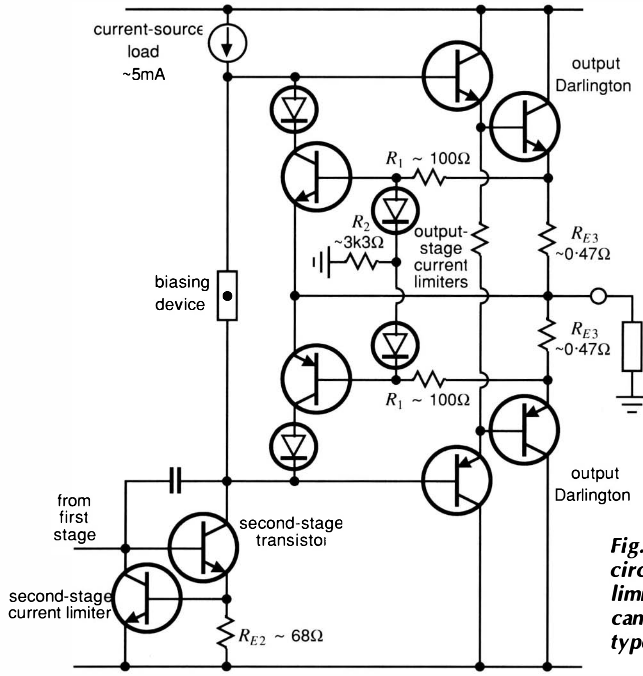

Fig. 6. Suggested protection circuitry for Fig. 1. Current-limiting transistors and diodes can be ordinary small-signal types.

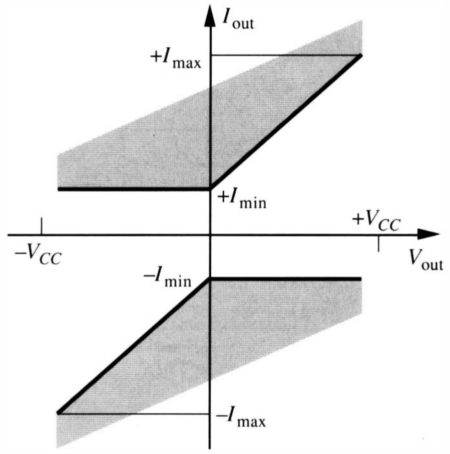

Fig. 7. Combinations of output voltage and current available with nonlinear fold-back protection of the output stage. For a nominal 50W, 8Ω amplifier, I_{min} should be around 1.5A and I_{max} around 6A.

Distortion associated with g_{m2}. Nonlinearity in g_{m2} models the exponential I_C versus V_{BE} characteristic intrinsic to a bjt, and the voltage drop across r_B. Sensitivity to changes in g_{m2} is extremely small (it is not even shown in Fig. 4) so the distortion associated with this nonlinearity is small too. Additionally, g_{m2} does not appear in the sensitivities of the other parameters. Therefore, emitter degeneration in the second stage, which reduces effective g_{m2}, has no effect on overall distortion and might at first appear pointless.

However, emitter degeneration can improve high-frequency stability. A significant amount of such degeneration is normally provided inadvertently, as part of the current-limiting protection circuitry. Second-stage degeneration therefore costs nothing in components, it does no harm, and it may do some incidental good.

Collector-base capacitance. Collector-base capacitance c_{CB} of the second-stage transistor is basically in parallel with the lag-compensating capacitor C and adds to its value. Collector-base capacitance is inevitably nonlinear, and has something like an inverse-square-root dependence on collector voltage.

Figure 4 shows that sensitivity to changes in C (hence c_{CB}) increases in proportion to frequency over the whole of the amplifier passband, and reaches unity at ω_{3dB}. The only way of reducing sensitivity towards C while retaining the basic amplifier topology is to increase overall cut-off frequency. Contrary to intuition, it does not help to use a larger value of C (the idea being that the nonlinear transistor capacitance would represent a smaller part of the total), nor does it help to increase ß (by using a Darlington); it does help to use a cascode for the second stage (Self's Fig. 4d of October 1993) or his modified Darlington (Fig. 4c same ref). Either removes signal voltage from the collector of the first member.

Output stage

In a push-pull class-B stage, the values of $ß_3$ and g_{m3} for the n-p-n and p-n-p transistors individually apply well into each half of the signal waveform, where only one transistor is conducting. In the overlap region near the middle where both transistors are conducting, or in a class-A stage, the values of $ß_3$ and g_{m3} are appropriately-defined averages.

Distortion associated with $ß_3$. Sensitivity to changes in $ß_3$ is constant throughout the amplifier passband. This rather surprising result is confirmed by experiment. Distortion associated with nonlinearity in $ß_3$ does not increase above the forward-path cut-off frequency ω_0 as loop gain falls away, nor does it increase above the ß cut-off frequency $\omega_3/ß_3$ of the output transistors.

Sensitivity to changes in $ß_3$ is inversely proportional to both $ß_2$ and $ß_3$. Increasing either reduces distortion without jeopardising stability. Notice particularly that $ß_2$ corresponds to the effective current gain of the second stage. Even in the ideal situation of very high transistor ß (such as a Darlington) and very large quiescent current I_C, $ß_{eff}$ given by Eq. 5 cannot exceed R_S/R_E. For example, in Self's Fig. 1a of August 1993, the first-stage collector-load resistors are 2.2kΩ and second-stage quiescent current is about 6mA. In a real circuit there would almost certainly be a resistor of 50-100Ω in Tr_4 emitter, as part of the protection circuitry. Hence $ß_{eff}$ of the second stage cannot exceed about 30. In contrast, if there is a current mirror in the first stage, the source resistance as seen by the second stage is large and $ß_{eff}$ approaches ß of the transistor. Herein lies the greatest advantage of the first-stage current mirror.

Self considers nonlinearity associated with $ß_3$ as a nonlinearity in the input resistance of the third stage and hence as a nonlinear loading on the second stage. He goes on to consider the benefits of an emitter-follower buffer between the second and third stages. This is perfectly valid, but I prefer to consider such a buffer as an extra member in the third-stage Darlington where it increases $ß_3$. Note that the 220Ω resistor in Self's Fig. 4f of October 1993 reduces effective ß (Eq. 5 again!); his Fig. 4e would be my preferred option. However, because sensitivity to changes in $ß_3$ is inversely proportional to both $ß_2$ and $ß_3$, it would do just as much good (and probably be simpler) to put the extra transistor into a second-stage Darlington.

Distortion associated with g_{m3}. Nonlinearity in g_{m3} models the exponential I_C versus V_{BE} intrinsic to transistors (twice over, because the transistor is usually a Darlington, Eq. 3); it models the nonlinear voltage drop across r_B associated with ß (also twice over in Eq. 3); and it models cross-over distortion. Effective g_{m3} includes the local emitter degeneration that is associated with emitter ballast resistors (Eq. 4).

Sensitivity to changes in g_{m3} increases in proportion to frequency, starting from a very small value which depends of all things on the equivalent resistance R_{C2} of the second-stage current-source load. The only way of reducing sensitivity towards g_{m3} while retaining the basic amplifier topology is to increase the overall cut-off frequency. Changing the emitter degeneration in any stage does not help, nor does increasing ß of any transistor.

Notice that cross-over distortion is predominantly associated with nonlinearity in g_{m3}. Self considers the cross-over region in detail; he points out the near impossibility of eliminating cross-over nonlinearity, stresses that the overall feedback is relatively ineffective in an amplifier of the topology of Fig. 1, and concludes that cross-over nonlinearity is the greatest source of distortion in a 'blameless' amplifier. In short, his observations confirm the theoretical prediction.

However, a great improvement can be achieved by slightly changing the amplifier topology: move the second-stage compensating capacitor C so that it encloses the third stage as shown in Fig. 27. Sensitivity to changes in g_{m3} becomes constant throughout

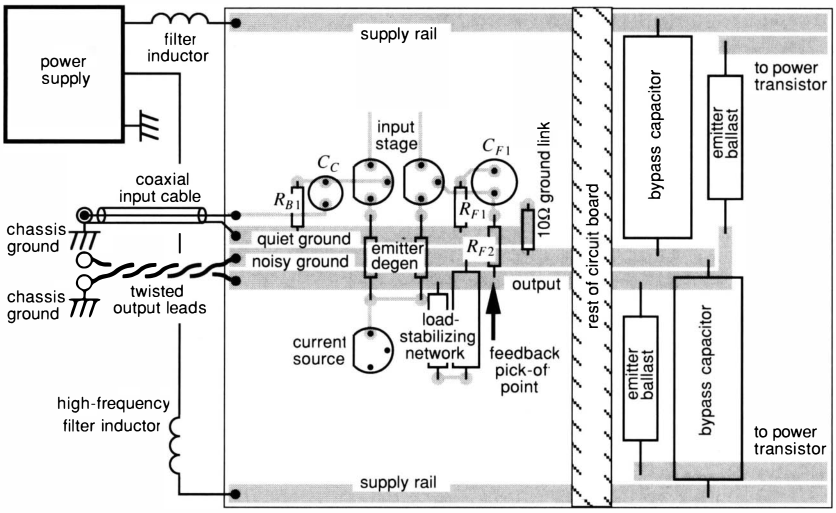

Fig. 8. Circuit-board and chassis layout for low distortion, showing separate tracks for noisy and quiet circuitry, separately grounded.

the passband, instead of increasing with frequency.

Many people believe that moving *C* provokes high-frequency oscillation, but this is not my experience and I strongly recommend the change. My amplifiers always incorporate a judicious amount of emitter degeneration in the second stage and a properly-designed load-stabilising network. If an amplifier oscillates when *C* is moved, it usually oscillates at several megahertz (far above the frequency of unity overall loop gain) and will usually continue to oscillate if the overall feedback can somehow be removed. The oscillation is a local parasitic. Try adding capacitors of around 50pF between collector and base of the first member of the output Darlingtons, using the shortest possible leads. Try shortening all leads to the output transistors. Try a small resistor in series with *C*, in theory about 20% larger than the second-stage emitter-degeneration resistor.

Nested feedback loops

Self makes brief reference to multiple feedback loops, nested one inside the another, but this of course is to depart from the basic amplifier topology under consideration. He also mentions multi-pole roll-off.

Nested feedback loops in general, and my own nested differentiating feedback loops in particular, offer a very great improvement in amplifier performance. However, the designer of a nested-loop amplifier needs to understand what he is about: time constants must be in correct ratios or the whole becomes impossible to stabilise. This is not to say that nested feedback circuits become more critical towards component tolerances – far from it – but the nominal values do need to be right.

Interested readers might refer to Reference 7, which describes how two nested differentiating feedback loops can be added to an amplifier of Self's basic topology, leading to an order-of-magnitude reduction in distortion. Loop roll-off is at a three-pole rate.

Protection

Self's class-B amplifier (February 1994) includes no protection – no doubt the circuit as printed was never intended to be a complete design. This amplifier would almost certainly be destroyed by even a momentary short-circuit of the output terminals; it requires current limiting in the output stage (probably of the fold-back variety) and also in the second stage, as in **Fig. 6**.

Despite what I may have published in the past, I have in recent years become an advocate of nonlinear fold-back limiting for the output stage. The circuit is not complicated, and it gives better protection than either simple limiting or linear fold-back limiting without restricting an amplifier's ability to drive reasonable reactive loads, so much so that it may even be possible to dispense with fuses in the supply rails.

Figure 7 shows the accessible regions of the load *VI* plane; the applicable design equations are,

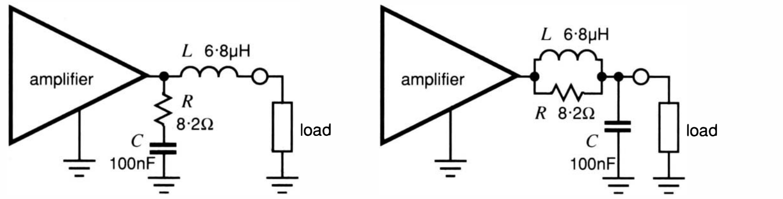

Fig. 9. Two forms of Thiele's load-stabilising network, with component values for 8Ω nominal load and 200kHz cut-off, where the output transistors see a nominally constant load.

$$I_{\min} \approx \frac{V_{BE}}{R_{E3}} \qquad (14a)$$

$$I_{\max} \approx \frac{1}{R_{E3}}\left[V_{BE} + V_{CC}\left(\frac{R_1}{R_2}\right)\right] \qquad (14b)$$

where V_{BE} is about 0.7V and R_1 should be somewhere around 100Ω.

I regard current limiting as mandatory in the second stage. If the load is short circuited and the input signal goes negative, the second stage is turned hard on, fighting against the limiter for the p-n-p half of the output stage. A simple current limit is sufficient for the second stage, and I set this at rather more than twice the quiescent current

$$I_{\text{limit}} \approx \frac{V_{BE}}{R_{E2}} \qquad (15)$$

Typically, this quiescent current is a few milliampères, so R_{E2} becomes 50-100Ω. This provides just about the optimum level of emitter degeneration for high-frequency stability, as referred to above.

Distortions outside the feedback loop

Hum and distortion currents. Correct layout of an amplifier pcb is essential, to isolate hum and distortion currents in the output stage from the low-level wiring. **Figure 8** shows the approach I adopt[8] to reduce both conductive and inductive coupling.

Note first the use of separate quiet and noisy ground tracks on the pcb, connected to chassis ground at separate points. Power-supply ground is connected to the chassis at yet another point. I don't believe in single-point grounding! Within the power supply, the transformer centre-tap and filter-capacitor grounds are all joined together as described by Self, and then a single lead comes out from this junction.

Quiet and noisy ground tracks run parallel to each other on the pcb, and as close as possible to minimise the area between them. Magnetic fields associated with the large currents in the output stage induce voltages between these tracks, proportional to this area.

Connect the quiet ground track to chassis ground at the input socket, via the shield on the input coaxial cable. This track carries the input-resistor and feedback-resistor ground currents and, depending on circuit details, may carry the ground currents from intermediate-level stages. Also, the vector area of the loop formed by the input lead/coupling capacitor/input transistors/local emitter-degeneration resistors/feedback capacitor/feedback resistor is zero; follow this loop, and note how the

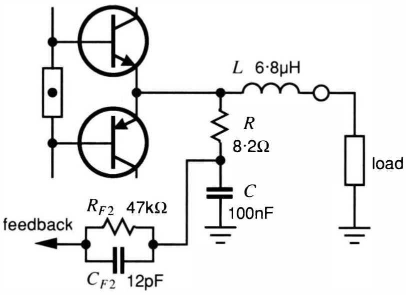

Fig. 10. Modified load-stabilising network which incorporates the overall feedback network. Values are for 8Ω nominal load and 200kHz cut-off.

areas enclosed on its left and right sides are equal. All these components hug the quiet ground track.

Connect the noisy ground track to chassis ground at the output terminal, via the twisted output leads. This track carries the ground currents from the supply bypass capacitors, the load-stabilising network (if any), and it may also carry ground currents from intermediate-level stages. The ground ends of the bypass capacitors are connected to this track as close together as possible, and no connections are made to the track between these two capacitors.

Similarly, the emitter ballast resistors in the output stage are connected to the output track as close together as possible, and no connections are made to this track between these two resistors. The output track runs parallel and close to the noisy ground track. The feedback pick-off point from this track is located between the ballast-resistor connections and the output connection, as is any load-stabilising network.

Mutual inductance should be zero between signal wiring and the loop formed by the bypass capacitors, emitter ballast resistors, power transistors and associated wiring. This corresponds approximately to setting zero vector area for the loop; in Fig. 8 the ballast resistors and bypass capacitors form a figure-of-eight, and the tracks to collector and emitter are spaced as closely as possible. Self's recommendation of twisted power-supply leads (February 1994) is really not enough; harmonic currents flow in all components of the loop, including the wiring to the power transistors. I recommend small filter inductors of a few microhenries in the positive and negative supply rails, to confine high-frequency components of supply current to the figure-of-eight loop on the pcb where the wiring layout is well defined[8]; these inductors should be

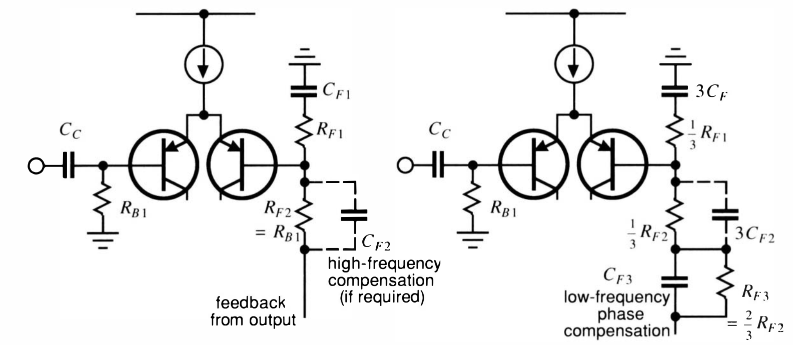

***Fig. 11.** Simple and phase-compensated low-frequency cut-off: as important at low frequencies as at high frequencies.*

mounted well away from the pcb itself, so that their magnetic fields do not interact with the input stage.

The two ground tracks on the pcb are linked via a 10Ω resistor, so that the bypass capacitors are effective for the low-level stages. This resistance is practically short-circuit in comparison with the impedances in typical low-level stages (it is smaller than the reactance of a 10nF capacitor at all frequencies up to 1MHz), but is open-circuit in comparison with the impedances in high-level stages. Signal components of current in the low-level stages can cross into the noisy ground track, but currents in the high-level stages cannot cross into the quiet ground.

Further thoughts on distortion

Here are a few ideas, unrelated to Self's articles.

First, there was another outstanding series on audio amplifiers by Peter Baxandall in *Wireless World*, beginning in January 1978[9]. Sixteen years on, these articles are still well worth reading.

Load-stabilising networks. Thiele[10] has proposed an *LRC* network to be connected between an amplifier and its load, to reduce the problem of high-frequency instability when the load is capacitive and also to reduce the problem of radio-frequency pickup on the loudspeaker leads. The output transistors are in principle presented with a constant-resistance load, and in practice are protected from the worst excesses of high-frequency variation in loudspeaker impedance.

Figure 9 shows two forms of Thiele's circuit. Parameter inter-relations are:

$$R = R_{L(\text{nominal})} \qquad (16)$$

$$\frac{1}{RC} = \frac{R}{L} = \omega_X \qquad (17)$$

where ω_x is the network cut-off frequency, usually corresponding to 100-300kHz. Figure 9b, with 100nF connected directly across the load looks crazy – more like an unstabilising network – but it is correct and has some advantages.

It is amazing how few published circuits are correctly designed (Self's are not). Usually they appear to be based on Fig. 9a, but they include a resistor in parallel with *L* as shown in Fig. 1, and the values are all wrong anyway!

Figure 10 is a modified load-stabilising network[2,7] which incorporates the overall feedback network and has two advantages over Fig. 9 in that the network does not introduce a 3dB loss at the cut-off frequency ω_x; and that radio-frequency interference picked up on the loudspeaker leads is isolated from the feedback point by a two-pole filter (isolation in Fig. 9 is single-pole).

If the amplifier without feedback has just one dominant pole, and if the overall loop gain without the network falls through unity at ω_x, then the overall response is made phase-linear by choosing

$$R_{F2}C_{F2} = \frac{(\sqrt{3}-1)}{\omega_X} \approx \frac{0{\cdot}7}{\omega_X} \qquad (18)$$

In practice, where the amplifier has second-order poles, C_{F2} is selected around this value to give the best square-wave response. The inductor should be air cored and mounted with nylon or other non-conducting screws, well away from any metalwork to avoid nonlinear eddy-current losses. I usually mount it on the pcb, in the under-populated area near the first-stage tail current source.

Low-frequency phase compensation. There is a simple modification to the overall feedback network which linearises the phase of the low-frequency cut-off and improves the square-wave response.

In **Fig. 11a** the low-frequency cut-off associated with the feedback network is

$$\omega_{\text{low}} = \frac{1}{R_{F1}C_{F1}} \qquad (19)$$

There is an additional fall-off associated with R_{B1} and C_C, usually small in comparison. If ω_{low} is chosen corresponding to 5Hz, a 20Hz square wave is reproduced with about 40% tilt and looks nothing like a square wave. This is the result of phase nonlinearity; all the Fourier components in the waveform are reproduced within 3% of their correct amplitudes.

However, if C_{F3} in Fig. 11b is chosen,

$$R_{F3}C_{F3} = \frac{2}{\omega_{\text{low}}} \qquad (20)$$

the phase is linearised and a 20Hz square wave is reproduced with essentially zero tilt. In practice, C_{F3} should be somewhat smaller than this theoretical value, to compensate also for the phase associated with C_C. The nearest preferred-value resistors are close enough.

Maximal flatness of frequency response is incompatible with phase linearity, at both low and high frequencies. Linearising the high-frequency phase inevitably results in a small drop in gain. Linearising the low-frequency phase inevitably results in a small peak (1dB at 1.6Hz for the values given). Given the choice between flat frequency response and phase linearity, everybody opts for the latter at high frequencies (that is, for best square-wave response). Why not at low frequencies?

Capacitor types. Most readers will know that polyethylene-terephthalate (Mylar) capacitors exhibit nonlinear effects at audio frequencies, associated with the dielectric relaxation time, and should be avoided in high-quality amplifiers. Polycarbonate capacitors are recommended for values up to a few microfarads. A problem is C_{F1}, of the order of 100μF. Nonlinearity in C_{F1} results in distortion that increases at low frequencies.

Ten years ago I made a study of the capacitors available in Australia. The surprising result was that ordinary cheap aluminium electrolytic capacitors were remarkably linear, far better than most tantalum types. I have made no measurements on more modern components, but the underlying chemistry has not changed, so it is unlikely that the situation has changed. Be sure to use the capacitors in the correct polarity – the positive side of both C_C and C_{F1} towards the transistor bases if amplifier polarities are as in Figs 1 and 2. ■

References

1. E.M. Cherry. A high-quality audio power amplifier. *Proc. I.R.E.E. Australia*, 39, pp.1-8, Jan/Feb. 1978.
2. M. Flanders and D. Swann. A song of reproduction, At The Drop of a Hat, Parlophone Record PMCO1033, London, 1958.
3. E.M. Cherry and G.K. Cambrell. Output resistance and intermodulation distortion of feedback amplifiers. *J. Audio Eng. Society*, 30, pp. 178-191, April 1982.
4. H.W. Bode. Network Analysis and Feedback Amplifier Design. van Nostrand, Princeton N.J., 1945. See also a number of papers in the *Bell System Technical Journal* for the decade preceding 1945.
5. E.M. Cherry. Feedback, sensitivity and stability of audio power amplifiers. *J. Audio Eng. Society*, 30, pp. 282-294, May 1982. See also *ibid*, 31, pp. 854-857, November 1983
6. E.M. Cherry. Transient intermodulation distortion - Part 1: hard nonlinearity. *I.E.E.E. Trans.*, ASSP-29, pp. 137-146, April 1981.
7. E.M. Cherry. Nested differentiating feedback loops in simple audio power amplifiers. *J. Audio Eng. Society*, 30, pp. 295-305, April 1982.
8. E.M. Cherry. A new distortion mechanism in class-B amplifiers. *J. Audio Eng. Society*, 29, pp. 327-328, May 1981.
9. P.J. Baxendall. Audio power amplifier design. *Wireless World*, 55, pp. 53-57, January 1978, and subsequent issues.
10. A.N. Thiele. Load stabilising network for audio amplifiers. *Proc. I.R.E.E. Australia.*, 36, pp. 297-300, September 1975.

* keith@snook·eu

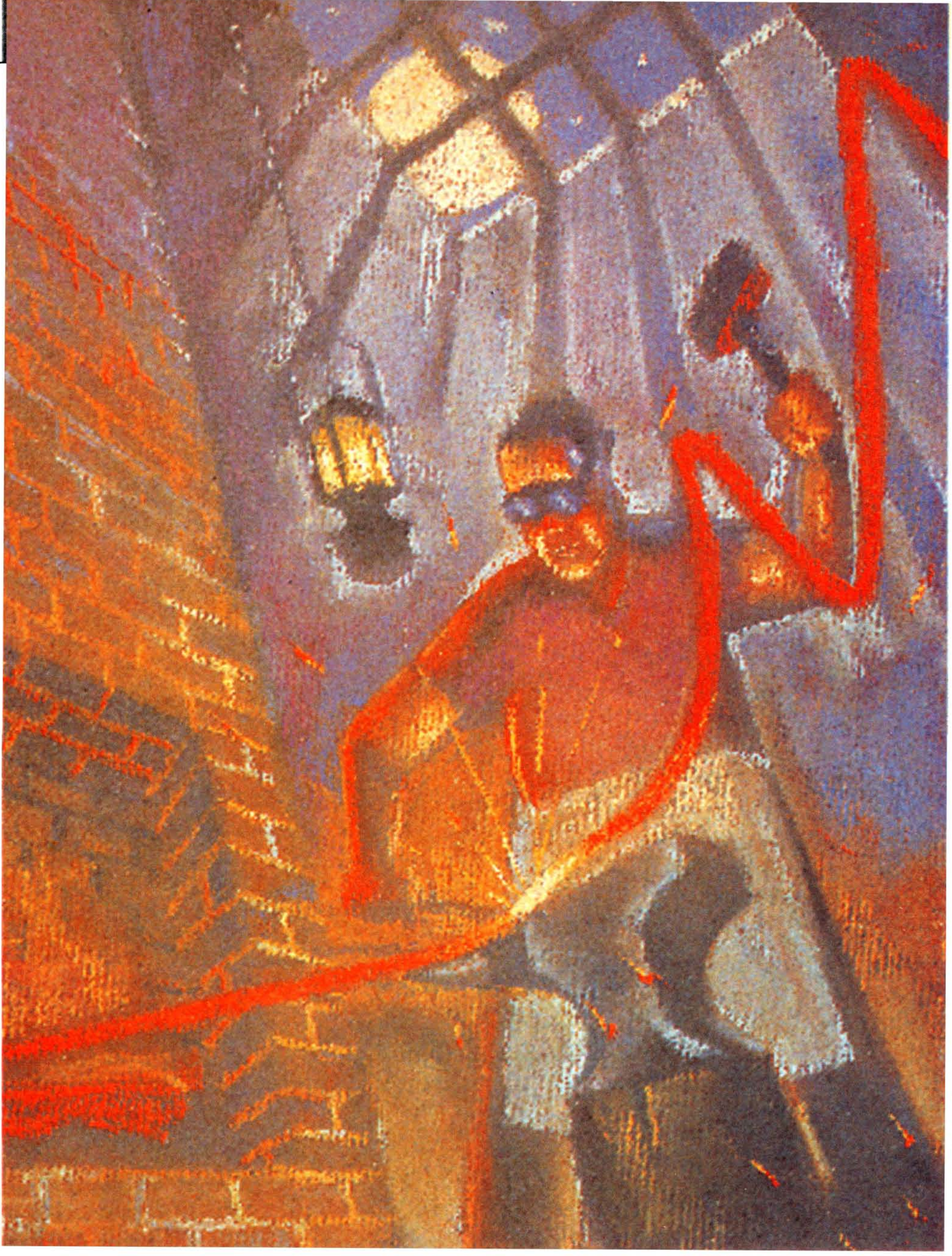

ILLUSTRATION HASHIM AKIB

Ed Cherry looks at distortion in audio power amplifiers and presents a critique of some of the novel attempts to try to reduce it published in recent years.

Ironing out distortion

There is an experiment on audio power amplifiers in the undergraduate teaching laboratory here at Monash University, in which students routinely observe that distortion, output resistance and slewing rate do conform to theoretical predictions. Amplifier design is not a mystery.

This article is a sequel to 'Ironing out distortion' in which I set out some of the basis for predicting amplifier distortions[1]. Between the time I submitted the final manuscript and when it was printed in January 1995, Douglas Self published two more articles on audio power amplifiers[2,3], and since then there have been other contributions from Self[4], Giovanni Stochino[5] and Bengt Olsson[6],

Common-emitter output stages

The common-emitter or common-source output stage is my preferred choice. Self refers[2] to my paper with Dr Greg Cambrell[7] in the *Journal of the Audio Engineering Society*. He gives a good account of some of the pros and cons of common-emitter and common-collector stages in the first part of his article.

Notably, he says that output resistance of a common-emitter amplifier with overall negative feedback is equal to that of a common-collector amplifier with overall feedback. Therefore, loudspeaker damping is the same for both. Self did not mention the principal conclusion of the paper, that intermodulation distortion is less for a common-emitter output stage than common-collector.

In my opinion the relation between common-emitter and common-collector stages could have been explained better. **Figure 1** herewith is Self's Fig. 9 re-drawn as I think it should have been.

Figure **1a)** is the starting point, a basic complementary common-emitter stage in which the collector currents are combined the load. Notice that the bias and drive for the p-n-p and n-p-n sides must be referenced to the positive and negative supply rails respectively, which is awkward but not impossible; I have built amplifiers of precisely this topology[8].

The transistor and its power supply on each side of Fig. 1a are in series around a loop with

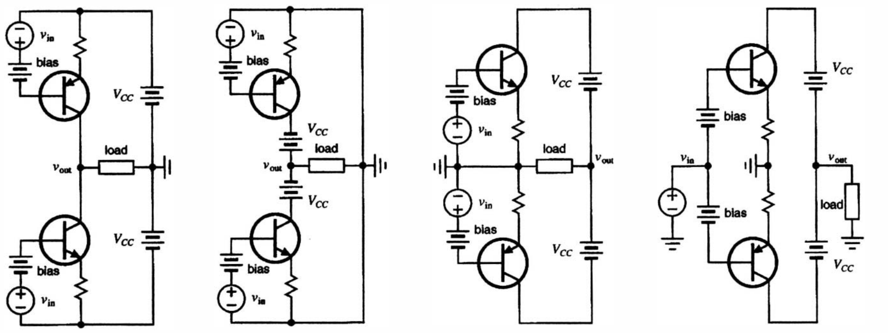

Fig. 1. Evolution of a common-emitter output stage. Figure 1a) is Self's Fig. 9a), a conventional common-emitter stage with the signal input and biasing for each side referenced to the supply rail and with the output collector currents combined in the load. In Fig. 1b) the order of each transistor and its power supply is reversed; these are in series, so operation of the circuit is unchanged. Figure 1c) is a purely cosmetic re-drawing, and Fig. 1d) is a further re-drawing with the input signal generators combined.

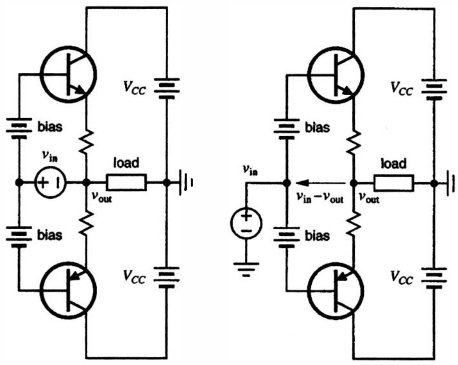

Fig. 2. Evolution of a common-collector stage from common-emitter. Figure 2a) is a true common-emitter stage, derived from Fig. 1d) by simply moving the ground point; the input signal voltage is required to float on top of the output. Figure 2b) is a conventional common-collector stage; the signal voltage between base and emitter in Fig. 2b) is vin – vout, showing that a common-collector stage can be regarded as a common-emitter stage with 100% local voltage feedback.

the load. As Self points out, the order of series components can be altered without changing the operation in any way. Accordingly the supplies are moved as in Fig. **1b)**. And Figure **1c)** is a drastic but purely cosmetic re-drawing – no change whatever to the circuit.

Finally, Fig. **1d)** is a further re-drawing in which the the two input signal generators are combined – again no change to the circuit. Figure 1d) is identical to Fig. 1a).

Figure 1d) is a true common-emitter output stage. The emitters are grounded (neglecting the ballast resistors), the full input signal (plus bias voltage, of course) appears between each base and emitter, and the full output signal (plus quiescent voltage) appears between each collector and emitter.

I believe that this arrangement of a common-emitter output stage was original when published in 1968[9], although it has appeared several times since, in reference 6 for example, without ever really catching on.

Benefits of common emitter

The arrangement of a common-emitter output stage in Fig. 1d) has an enormous practical advantage over any common-collector output stage: *the input signal amplitude is just a few volts peak-to-peak.* Therefore, the only transistors in the complete amplifier which ever need to withstand high voltages are the output transistors. Everything else – including the drivers – can operate from low supplies of, say, ±15V.

A high-voltage transistor must be lightly doped in order to achieve a wide collector depletion layer and reduce the electric field for a given collector voltage. Inevitably this reduces f_T and increases the saturation voltage.

Compared with any common-collector amplifier, much better transistors can be used in all the low-voltage low-level and driver sections of an amplifier based on Fig. 1d). Non-dominant poles can therefore be moved further out, making it easier to stabilise the feedback loop. Higher quiescent and peak currents can be used in order to achieve high slewing rate, without running into either power-dissipation or secondary-breakdown limits.

The only disadvantage of Fig. 1d) is that the main V_{CC} supplies float. Separate supplies are therefore needed for each channel of a stereo amplifier. But then, many highly-regarded amplifiers use separate power supplies anyway.

Figure 2 explains the relationship between common-emitter and common-collector output stages. Figure **2a)** is is identical to Fig. 1d except that the ground point has been moved: the V_{CC} supplies are grounded but now the input signal source must float on top of the output. This is a thoroughly impracticable arrangement, but circuit operation is not changed. The amplifier is still strictly common-emitter: the full input signal voltage appears between base and emitter – neglecting ballast resistors – and the full output signal voltage appears between collector and emitter.

Figure **2b)** shows the conversion from common-emitter to common-collector: the neutral end of the signal source is simply grounded. Now the signal voltage between base and emitter, neglecting ballast resistors, becomes v_{in}–v_{out} rather than v_{in}, which demonstrates that a common-collector stage is nothing more than a common-emitter stage with 100% local voltage feedback. All the output voltage is subtracted from the input voltage to give the drive voltage for the transistors.

Perhaps this gives physical insight as to why the output resistance of a common-emitter amplifier with overall feedback is the same as for a common-collector amplifier. The intrinsic output resistance of a common-emitter stage is high, but this is reduced in **Fig. 3a)** by the overall feedback.

By comparison, the intrinsic output resistance of a common-collector stage is low; this low resistance is attributable to the local feedback, and in Fig. **3b)** it is further reduced by the overall feedback. However, the voltage gain of a common-emitter stage is large whereas the gain of a common-emitter stage is near unity. Therefore the overall loop gain around the common-emitter amplifier is larger than around the common-collector amplifier.

It turns out that the extra overall feedback around the common-emitter stage compensates exactly for its higher intrinsic output resistance. It also turns out – but is much more difficult to prove – that the stability of the feedback loop is the same, higher loop gain not withstanding[7].

Output resistance – a new method

You might be interested in a simple new, general and precise method for finding the input and output resistances of a feedback amplifier[10]. In the same paper is an approximation which appears more reliable than any of the "multiply or divide the resistance-without-feedback by loop gain" types of formula:

- Write down the loop gain. The expression doesn't need to be exact, merely an approximation of the same order of accuracy as the required output resistance.

- Equate this loop gain to unity, and solve for the load resistance. In other words, the output resistance of a feedback amplifier is equal to the load resistance that would reduce the loop gain to unity.

The method is easy, because it requires only the loop gain and not the output resistance without feedback. It works for all feedback amplifiers, not just common-emitter-output or not just common-collector-output, and not just voltage-feedback or current-feedback; you don't need sometimes to multiply by loop gain and sometimes divide.

There is a corresponding method for finding input resistance.

Slewing rate

Self's discussion of slewing rate[3] is correct, but it falls into the category of analysing a bad

*Fig. 1a) on p. 632 of August 1993 issue, or my Fig. 1 on p. 15 of January 1995.

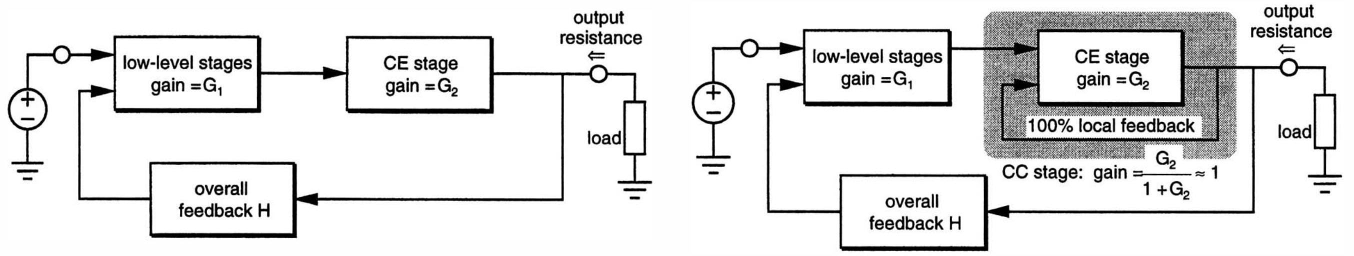

Fig. 3. Output resistance of common-emitter and common-collector amplifiers. The total feedback around the output stage is the same, and therefore the output resistances are equal.

Benefits of IGBTs in power driving

I would like to place it on record that a complementary common-emitter or common-source amplifier with floating power supplies as in Fig. 1d provides an elegant solution to driving the gates of large insulated-gate bipolar transistors, or igbts, in high-power pulse-width modulated drives and inverters. In this application the gate must be driven between about +15V and –10V in a few nanoseconds; transient current in the gate capacitance during switching amounts to several amperes.

Figure A shows the arrangement. Parasitic gate-lead inductance has little influence on switching speed, because it is in series with the driver mosfet drains which behave like current sources.

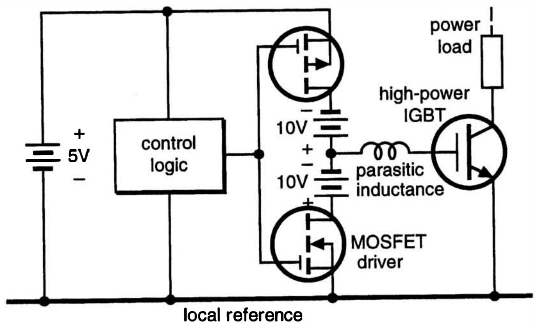

Fig. A. Complementary common-source enhancement-mode mosfet amplifier based on Fig. 1d, for driving an igbt gate between +15V and –10V in a few nanoseconds.

circuit rather than recommending a good one. Slewing rate is set in amplifiers of the basic common-collector-output topology* by the circuit's ability to charge and discharge the second-stage compensating capacitor, **Fig. 4**.

The charging current flows at both sides of this capacitor, and slewing rate is restricted by whichever side first reaches the available current limit. On the left-hand side of the compensating capacitor the available charging current is near enough to the output from the first stage and, if this stage is a long-tailed pair with current mirror, the positive and negative limiting values are symmetrical and equal to the tail current. On the right-hand side the situation is more complicated: the current available to the capacitor is the left-overs from the algebraic sum of the second-stage collector current, its current-source load current, and the input base currents of the third-stage transistors.

Self noticed that the 'current source' in his amplifier (Tr_6 in Fig. 1 on p. 761 of September 1994 issue) did not supply constant current when its collector voltage was changing rapidly. He observed a 'spike' of current.

Said differently, Self observed that, although his current source might have had a high output resistance, it also has a high shunt capacitance; recall that a capacitor draws a spike of current when the voltage across it changes rapidly. This spike is in the direction which subtracts from the peak current available to the right-hand side of the compensating capacitor.

Current source analyses

Figure 5a) is an n-p-n current source, the "flip" of Self's p-n-p circuit. Note the collector-base capacitance C_{CB} of the transistor. In the vacuum-tube era a circuit of this topology was known as a 'reactance-tube modulator'. Its function was to provide a voltage-variable capacitor to modulate the frequency of an *LC* oscillator. The capacitance looking into the anode of Fig. **5b** is,

$$C_{\text{modulator}}=g_m R_G C_{AG}$$

The similarity of Figs 6a) and 6b) is apparent, and the capacitance looking into the current source is,

$$C_{source}=\left(\frac{R_B}{R_E}\right)C_{CB}$$

Figure **5c** is Self's *p-n-p* current source with actual component values marked; the capacitance looking into the collector is about 100pF – equal to his compensating capacitor. No wonder the slew rate was affected.

If the circuit topology is fixed, obvious reductions in current-source capacitance will accrue from reducing R_B in the C_{source} equation (to zero, ideally) and from increasing R_E.

There are problems with both approaches in Self's amplifier, where the first and second stage current sources share a common voltage reference. Reducing R_B provides a kind of feedback which decreases/increases the first-stage tail current on fast positive/negative swings. Increasing R_E reduces the peak positive output voltage from the complete amplifier, hence reduces available power output.

Better by far to change the circuit topology. In my original article I showed (Fig. 5 of Ref. 1} a really solid first-stage tail-current source using a 10V zener diode. I didn't comment on Self's second-stage current source, but in fact I always use a bootstrapped resistor as described, for example, in Ref. 11 – an arrangement which Self largely dismissed.

A resistor bootstrapped as in **Fig. 6** eliminates the problem of capacitive loading, it actually increases the peak positive output voltage available from the amplifier, and it is cheap. The sum of (R_1+R_2) should be chosen to provide the desired quiescent current in the second stage,

$$I_{2(quiescent)}\approx\frac{V_{CC}}{R_1+R_2}$$

The ratio of R_1 to R_2 is not critical, except that R_1 should be as large as possible consistent with $R_2>>R_L$.

Time constant R_2C_B should be chosen having regard to the lower 3dB cut-off frequency of the amplifier: $\omega_{low}=R_2C_B>>1$.

Shifted compensating capacitor

If you want to change the circuit, then shifting the compensating capacitor as shown grey in Fig. 6 has many advantages; I have described it[11] as "...the greatest bargain of all time...". It is very effective in reducing cross-over distortion – the major residual distortion in Self's 'blameless' amplifier. It also helps with slew symmetry, because the loading effect of the capacitor is transferred from the second-stage collector, where the available current is milliamperes, to the output, where the current is amperes.

Until now there has been no reaction from readers to this recommendation in my article, but I have in the past been told that shifting the compensating capacitor provokes high-frequency oscillation. This is not my experience: I have built dozens of amplifiers, and I have published an analysis[12] which has never been challenged.

All this prompts the question "Is the supposed oscillation for real?" Is it perhaps that believers in the oscillation are merely reporting what someone else has told them? I would be interested to hear from anyone with first-hand experience of the problem. However, before making contact with me, please read what I said on pp. 19-20 of *Electronics World* for January 1995:

- I can believe in local parasitic oscillation of the first member of the output-stage Darlington – the 'drivers' – as distinct from oscillation of the main feedback loop. Driver transistors such as *BD139/140* with f_T around 100MHz usually oscillate when biased to 5-10mA at the end of 10-20cm leads. Check the frequency of the oscillation: is it near the unity-loop-gain frequency, or is it significantly higher? For a common-collector-output amplifier with the low-level stages shown in Fig. 4,

$$\omega_{\text{unity loop gain}} = \left(\frac{1}{R_{E1}C}\right)\times\left(\frac{R_{F1}}{R_{F1}+R_{F2}}\right)$$

- My amplifiers always feature impeccable layout and bypassing with separate quiet and noisy ground tracks, and short leads to the drivers. All of this discourages parasitics.
- For the same reason I routinely provide 'stopper' capacitors of 30–50pF between collector and base of the drivers, using the shortest possible leads – no more than 1cm.
- My amplifiers always incorporate a correctly-designed load-stabilising network.
- My amplifiers always include judicious emitter degeneration in the second stage as shown in Fig. 6.
- In common-collector amplifiers – as distinct from my preferred common-emitter – I use a bootstrapped resistor as the second-stage current source.

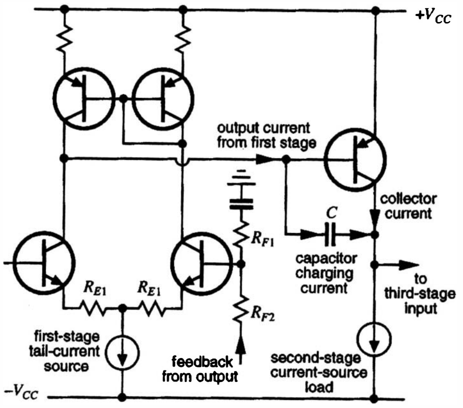
+V_CC
output current from first stage
collector current
C
R_F1
capacitor charging current
to third-stage input
R_E1
R_E1
R_F2
first-stage tail-current source
feedback from output
second-stage current-source load
-V_CC

Fig. 4. Low-level stages of a common-collector-output amplifier. Note that the polarity is flipped relative to Self's articles, but the same as in Ref. 1.

If any reader has taken care of all these matters and still experienced oscillation when the lag-compensating capacitor is shifted, I would be pleased to make contact. But I do urge you to try shifting the capacitor. It effects a remarkable reduction in cross-over distortion.

Load-stabilising Zobel networks

Load-stabilising networks are used to ensure that the amplifier proper is presented with something like its nominal load resistance at high frequencies. This is the case even if the external loudspeaker load is highly reactive. Secondly, they are used to prevent rf interference, picked up by the loudspeaker leads acting as antennae, from finding its way back into the first stage via the feedback network.

Since submitting the final manuscript of Ref. 1, I have realised that there are in fact two families of load-stabilising network – not just the two networks which Thiele proposed[13]. **Figure 7** shows the general models. For both circuits the inductance and capacitance should satisfy,

$$\frac{L}{R_O} = R_O C = \frac{1}{\omega_x}$$

where R_0 is the nominal loudspeaker load resistance, probably 8Ω, and ω_X is the network cut-off frequency. In addition, for **Figs. 7a** and **7b** respectively,

$$R_2 = \frac{R_O^2}{R_1 - R_O} \qquad \text{(a)}$$

$$R_2 = \frac{R_O^2}{R_1 + R_O} \qquad \text{(b)}$$

Note the sign change in the denominator.

In Fig. 7a), if R_1 is chosen as infinity, i.e. the capacitor branch is open-circuited, then from equation (a) above R_2 needs to equal 0 – short-circuit the inductor in other words. As a result, the whole network disappears. This corresponds to the limiting case of an amplifier without a load-stabilising network.

On the other hand if R_1 is chosen as its minimum allowed value of R_0, then R_2 is infinity and Fig. 7a) reduces to Thiele's original (Fig. 9a of Ref. 1). Between these extremes is a continuum of allowed resistance values. Is any especially desirable?

Thiele's original with R_2=∞ gives the greatest isolation between amplifier and load, and the greatest attenuation of rf interference. But the circuit does ring badly if the external load is made pure capacitance.

Whether a pure capacitive load is realistic of anything practical is a moot point, and in any case the ringing is a simply a resonance between the inductor and this capacitance – not an indication of approaching instability in the amplifier.

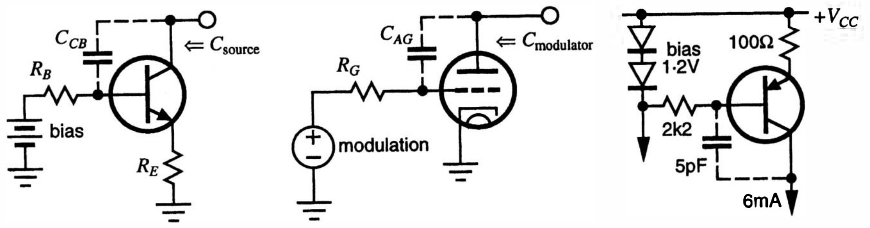

Fig. 5. (a) n-p-n current source; (b) reactance-tube modulator, a voltage-variable capacitor from the vacuum-tube era; (c) Self's p-n-p current source.

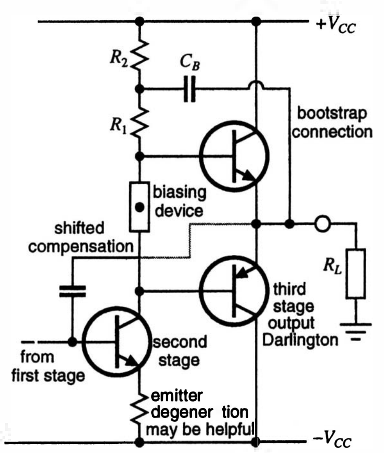

Fig. 6. Bootstrapped resistor as second-stage current source. Shifting the compensating capacitor (shown grey) reduces crossover distortion and can improve slew symmetry. Second-stage emitter degeneration helps with stability.

If anyone is worried by the ringing, however, then damping can be increased by using some finite R_2 and the corresponding R_1. The price, of course, is reduced isolation and reduced rf attenuation. Thiele's original (or my modification of it in Ref. 1) is still my preferred choice.

Similarly, if in Fig. 7b you choose R_1=∞ and R_2=0, the whole network disappears. If you choose R_1=0 and R_2=R_0 on the other hand, the circuit reduces to Thiele's original. Again there is a continuum of allowed resistance values between these extremes.

Notice that this form of Thiele's network (Fig. 9b of Ref. 1 – the circuit I described as crazy-looking with 100nF directly across the loudspeaker) is much better in regard to ringing than the more common **Fig. 8a**. I urge you to try **Fig. 8b** – crazy-looking or not.

Distortion off the supply rails

Self's cascode-like first stage[4] is an ingenious solution to the problem of distortion on the supply rails re-entering the circuit via the lag-compensating capacitor. Congratulations. However there are at least two other solutions.

My preferred choice is the common-emitter output stage as described above. Here there is no significant signal on the low-voltage supplies to the low-level stages, hence no problem.

Alternatively, with a common-collector output stage use nested differentiating feedback loops as in Ref. 11. Here the return end of each lag-compensating capacitor is connected to a virtual ground. Again there is no problem.

First-stage c-m distortion

Related to distortion off the supply rails is distortion which enters via the finite common-mode rejection of the first stage. Signals on

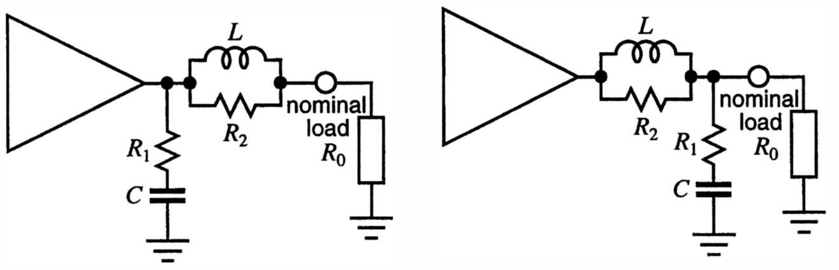

Fig. 7. Generalised load-stabilising networks. These reduce to Thiele's networks for special cases of the resistor values.

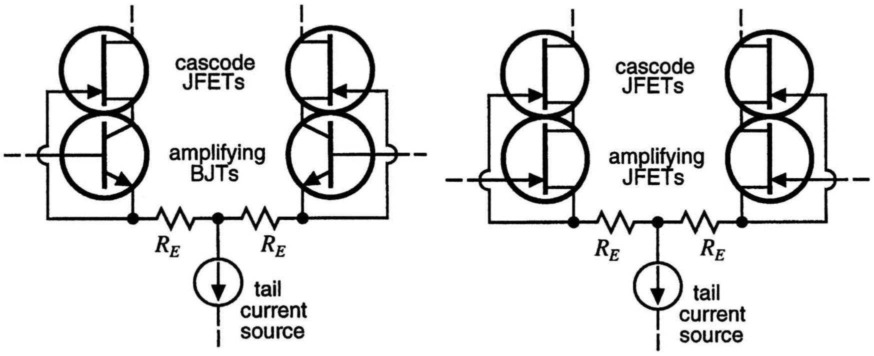

Fig. 8. Bootstrapped cascode long-tailed pairs, using a JFET as the top member. The top JFET needs appropriate values for V_P and I_{DO} (I_{DSS}) in order to provide headroom for the bottom device.

the supply rail appear more-or-less unattenuated at the collectors of the first stage unless something like Self's cascode is included, and harmonics of signal on the supply rails can introduce distortion via this mechanism.

However there is another common-mode distortion mechanism. The input and feedback signals, applied to the bases of the input long-tailed pair, can be resolved into differential and common-mode components.

The principal component of current output from the first stage is proportional to the small difference between the input and feedback voltages; half this difference appears between the base of each transistor and the top of the tail. Simultaneously the average of the input and feedback voltages appears between each collector and the top of the tail as a large common-mode signal.

The differential and common-mode signals will intermodulate and produce beat frequencies if there is any dependence of first-stage differential gain on collector voltage.

In practice there is such a dependence, and the gain variation is something like linear with collector-emitter voltage. It follows that the intermodulation is proportional to the product of the differential and common-mode signal amplitudes, and its frequency is twice the input signal frequency.

In other words, the intermodulation distortion appears like second-harmonic distortion, although it truly is the result of intermodulation – perhaps auto-intermodulation would be the correct description. Adding a cascode, to hold the collector voltage constant, will not help; it is the large signal voltage at the emitter that matters.

A number of physical mechanisms are involved for bipolar-junction transistors, all associated with widening of the collector depletion layer as collector-emitter voltage increases:

- Classical text-book Early effect, by which base-emitter voltage for a specified collector current depends on collector-emitter voltage;
- Modulation of transistor base width, hence β and base current, and ultimately modulation of the signal voltage drops across any series resistance in the base circuit – the source resistance, the Thévenin equivalent resistance of the feedback network, and transistor base-spreading resistance;
- 'pinching' of the base resistance itself, hence modulation of the voltage drop associated with base current.

Field-effect transistors exhibit a corresponding dependence of differential gain on drain voltage. They too can generate second-harmonic-like distortion via auto-intermodulation. Common-mode distortion mechanisms are not confined to bjt stages nor to modulation of base current.

Simulations may not reveal common-mode distortion. Most Spice transistor models treat the Early voltage as a constant where in fact it varies as something like the square root of collector-emitter voltage. Also, simulations are likely to represent both transistors of a long-tailed pair as identical – apart, perhaps, from quiescent conditions – whereas in a real amplifier they are not identical and the unbalance is significant. Second harmonic cancels in a perfectly-balanced circuit.

Because the intermodulation is proportional to the product of the differential and common-mode signal voltages, distortion from this cause can be reduced by reducing either voltage. The differential voltage can be reduced by increasing the feedback loop gain at the signal frequency, but this means increasing the overall closed-loop cut-off frequency and hence increasing the likelihood of instability.

Increasing the overall closed-loop gain reduces the common-mode component at all frequencies and hence reduces the input voltage required for full output; this is one of the reasons why I prefer the 300mV typical of old-style vacuum-tube amplifiers, to today's more usual 0·7-1·0V.

If a hardware solution is required, I use the bootstrapped cascode arrangements shown in Fig. 8.

Power mosfets versus bipolar devices

I agree with everything Self says about the relative nonlinearity of bjts and mosfets in output stages. However he has omitted one important consideration: gain-bandwidth product. I also feel he has over-stressed the importance of crossover distortion when there is the simple fix of shifting the lag-compensating capacitor as in Fig. 6.

Fifty years ago Bode showed[14] that the amount of feedback which can be applied to an amplifier, and hence the amount by which distortion can be reduced with specified margins against instability, is proportional to active-device gain-bandwidth product *GB* exponentiated to a power that depends on the phase margin. Here gain-bandwidth product is used in Bode's precise sense, related ultimately to the transit time of carriers through the control region.

It follows that bjts are the preferred devices for the low-level stages of a feedback amplifier. Typical types, such as *BC547*s, have transit times around 500ps and hence *GB* of 300 to 500MHz, compared with 1ns and 100-200MHz for silicon j-fets such as the *2N5485.*

However, the corresponding numbers for power bjts like the *MJ802* are around 100ns and 1 to 2MHz, compared with 1ns and something over 100MHz for power mosfets such as the *IRF240.* More feedback can in theory be applied around power mosfets.

In the terminology of Refs 1 and 11, output-stage distortion for mosfets is almost entirely a consequence of nonlinearity in g_{m3}; crossover distortion for both bjts and mosfets is also a consequence of nonlinearity in g_{m3}.

Figure 4 of Ref. 1 shows that sensitivity towards changes in g_{m3} is inversely proportional to the second-stage lag compensating capacitor. Therefore distortion is inversely proportional to this capacitor.

Comparison anomalies

In August 1995 John Linsley-Hood's published a comparison between bjts and mosfets[15]. In this comparison, the compensating capacitor C_{13} in his Fig. 1 is marked 'value depends on circuit'. In other words, the comparison of bjts, mosfets and igbts was not made on a level playing field.

The fact that measured distortion for the mosfets was about half that for the bjts is of itself meaningless; we must also be told the ratio of the compensating capacitors in the two experiments.

My guess is that the compensating capacitor was smaller in the case of the mosfets as compared with the bjts. As explained above, Bode's work shows that much more feedback can be applied to mosfets at high frequencies without approaching instability, because their transit time is shorter; mosfets require less compensation than bjts.

If the mosfet compensating capacitor was half the bjt capacitor, then Linsley-Hood's experiment shows that the open-loop distortions of mosfets and bjts are about the same. If the mosfet capacitor was smaller than half the bjt capacitor, then the open-loop bjts are better than the mosfets – as Self claims.

In the end, however, it is closed-loop distortion that matters, not open-loop, and Linsley-Hood's experiment confirms my preference for mosfets. Their open-loop distortion may be somewhat greater than for bjts, but more feedback can applied around them.

To repeat the quotation from Ref. 8: "The author's approach to designing a high-quality amplifier is to choose a simple topology based on common-emitter amplifying stages and apply negative feedback to reduce distortion. 'Clever' circuit topologies (other than push-pull operation) rarely give better than a tenfold reduction in distortion on a production basis. Feedback, however, can reduce distortion almost indefinitely." ■

References

1. Cherry, E. M., 'Ironing out distortion', *EW&WW*, vol. 101, pp. 14–20, January 1995.
2. Self, D., 'Common-emitter power amplifiers', *EW&WW*, vol. 100, pp. 548–553, July 1994.
3. Self, D., 'High-speed power', *EW&WW*, vol. 100, pp. 760–764, September 1994.
4. Self, D., 'Distortion off the rails', *EW&WW*, vol. 101, pp. 201–206, March 1995.
5. Stochino, G., 'Audio design leaps forward', *EW&WW*, vol. 100. pp. 818–824, October 1994. keith@keith-snook.info
6. Olsson, B.G., 'Better audio from non-complements', *EW&WW*, vol. 100, pp. 988–992, December 1994.
7. Cherry, E.M. and Cambrell, G.K., 'Output resistance and intermodulation distortion of feedback amplifiers,' *J. Audio Eng. Society*, vol. 30, pp. 178–191, April 1982.
8. Cherry, E.M. 'A high-quality audio power amplifier', *Proc. IREE Australia*, vol. 39, pp. 1–8, Jan/Feb. 1978.
9. Cherry, E.M. and Hooper, D.E., *Amplifying Devices and Low-Pass Amplifier Design*, Wiley, New York, 1968. See Fig. 14.26b on p. 891. keith@snook·eu
10. Cherry, E.M. 'Input impedance and output impedance of feedback amplifiers', *Proc. Inst. Elec. Engrs. – Circuits, Devices & Syst.*, vol. 143, pp. 195–201, April 1996.
11. Cherry, E.M. 'Nested differentiating feedback loops in simple audio power amplifiers', *J. Audio Eng. Society*, vol. 30, pp. 295–305, May 1982.
12. Cherry, E.M. 'Feedback, sensitivity and stability of audio power amplifiers', *J. Audio Eng. Society*, vol. 30, pp. 282–294, May 1982. Also *ibid*, vol 31, pp. 854–857, November 1983.
13. Thiele, A.N., 'Load stabilising network for audio amplifiers', *Proc. IREE Australia*, vol. 36, pp.297–300, September 1975.
14. Bode, H.W., *Network Analysis and Feedback Amplifier Design*, van Nostrand, Princeton NJ, 1945.
15. Linsley-Hood, J., 'Expert witness', *EW&WW*, vol. 101, pp. 684–685, August 1995.