Are Op Amps Really Linear?

By Barrie Gilbert,
ADI Fellow; Manager, NW Labs, Beaverton, OR

Everyone knows that op amps are the most linear building blocks in the analog repertoire. If you want nonlinear behavior, you had better look to multipliers or other arcania. But we recognize that every real amplifier has a *bit* of nonlinearity, like it or lump it, so maybe we ought to take a look at just how good is that standard op amp, the voltage-mode amplifier identified as an 'OPA' in previous columns, when used at frequencies above a few Hertz.

It wasn't so many years ago, in the heyday of the uA741, that a guy called Otala discovered this early 1-MHz OPA performed a lot worse than might expected at audio frequencies, so he wrote a paper in the Journal of the Audio Engineering Society in which he coined a brand new term -- transient intermodulation distortion, or 'TIM' -- to describe in audio lingo what most users of op amps had already discovered, namely, the little matter of a *limited slew-rate*. It was a phenomenon peculiar to IC op amps. If you had grown up with vacuum tubes, you had plenty of reason to raise your eyebrows, since, unlike bipolar transistors (with their extremely high gm and a correspondingly small range over which linearity prevails in the familiar differential-pair gain cell), tube amplifiers rarely, if ever, got themselves into this pathological state of affairs.

Predictably, Otala's 'discovery' only added welcome fuel to the fire of those who were quite convinced that these new-fangled transistor amplifiers were genetically incapable of pleasing the audiophile ear, a notion which, like all myths, persists unabated, in spite of the fact that the root cause of TIM is now completely understood, and can easily be avoided in transistor amplifiers expressly designed for the most demanding audio applications. Nevertheless, Mister Otala was on to something, and although there are many splendid op amps out there, not all deliver quite what the textbooks promise. Accordingly, for a few minutes, we will examine a major source of distortion in op amps. Writing this in Kailua-Kona, Hawaii, with the thunder of the surf just feet away from this lanai, as the palms and hibiscus breathe the balmy post-dawn air, I have ample time on my hands.

The dominant mantra intoned over the application of op amps goes something like this: Never mind what's inside that cute little triangle; the function of the overall linear block -- invariably just a simple amplifier or filter stage -- is determined (almost entirely) by the passive components that are added to provide feedback. The triangle thing is merely the power-house that, come rain or shine, makes everything work out right in the end, by a kind of op-amp magic. The parenthetical inclusion is a concession to those who know a bit better, and it's the focus of this piece.

Reduced to essentials, most voltage-mode op amps, OPAs, are based on a topology like that shown in Fig. 1. To develop the theory, our device is here connected as a simple amplifier with a closed-loop gain of G, determined by the ratio (R1+R2)/R2, which can alternatively be expressed in terms of the feedback fraction b = 1/G. Because the dominant source of nonlinearity is in the *input* cell, the distortion will be lowest in the voltage-follower mode.

In the interests of clarity and analytical simplicity, we will assume here that the output is a *perfect sinusoid*, having an amplitude E and angular frequency w, that is, E*sin*wt, and work backwards from this output to deduce what VIN *would need to be* to generate that output. This may seem an odd approach, and it's certainly not essential to do it this way. However, many analyses involving the exponential behavior of transistors lead to transcendental solutions when pursued in the forward direction, and that's true in this case. A reverse-direction analysis quickly generates the key insights, and points towards the required modifications to effect a solution, at the price of a slight but not serious loss of rigor.

A *bipolar* implementation is shown, since many monolithic OPAs use this technology. A basic differential-pair Q1, Q2, also cast in pnp form, and having a near-perfect current-source IT in its 'tail', senses the difference between the applied input VIN and some fraction of the output, while being insensitive to common-mode levels. In the case of a voltage follower, of course, the distinction between the signal and the common-mode voltage is somewhat fuzzy; the *real* value of the high common-mode rejection ratio (CMRR) afforded by OPAs is much more apparent when the amplifier is connected in a high-gain mode, and a small input signal is accompanied by an interfering common-mode signal.

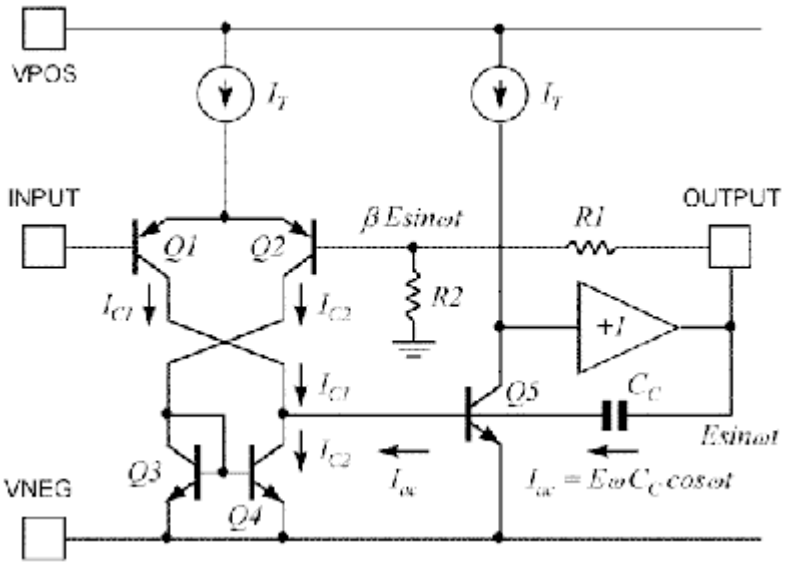

Fig. 1 Topology of a Typical OPA

Formally, this input cell is a *nonlinear transconductance*, whose output currents IC1 and IC2 are applied to the npn current-mirror Q3, Q4, which generates the difference IC1 - IC2. This current is then *integrated* by the 'HF compensation' capacitor, CC, in the main voltage-gain stage provided by the common-emitter stage Q5, which is forced to operate at a constant collector current of IT.

The resulting output voltage is buffered by what is here shown as an ideal voltagecontrolled voltage-source (VCVS) but which in most cases will be the familiar Class-AB complementary emitter-follower that provides the high current-gain to drive the load, RL.

In a real OPA, some of the open-loop distortion will arise in this VCVS stage, but in order to keep our sights focused on the main distortion mechanism, we can ignore that here. Notice that CC is connected to the *final* output node, not to the collector of Q5, as is often the case. This minor modification means that backand-forth flow of the HF displacement current in CC is not supported by Q5 but, rather, by the output stage. Consequently, there is no variation in IC5 (it's held steady at IT) and thus *the VBE of Q5 is likewise constant with output voltage*.

All this groundwork may seem very tedious, but it is with a view to getting at the *root cause* of distortion. Numerous such detailed considerations are crucial to the design of an op amp capable of ultra-linear HF performance, and each needs to be eliminated independently. We're only considering the first of many, here.

Now we're ready to start the sums. Unfortunately, there is no painless way to avoid mathematics, but what we *can* do is use simple, even rudimentary, models for the transistors, in the spirit of Foundation Design. This approach to the quest for insight was mentioned in the "Spicing Up The Op Amp", and it's time to be a little more specific about this notion. It works very well for the bipolar junction transistor (BJT), which will probably remain a major technology -- certainly in the high-performance arena -- for at least the next decade, and beyond, in spite of the impressive advances in analog CMOS design, only made more difficult by the almost total emphasis on digital applications in the on-going development of sub-micron technologies. (Some of the reasons for holding to this view were stated in "Why Bipolar?".)

The Level-0 model for the BJT is simply a voltage-controlled current-source (VCCS) having an exact exponential relationship between its collector current, IC, and its base-emitter voltage, VBE; this is the *heart* of the BJT:

IC = IS *exp* (VBE/VT) (1)

where IS is the saturation current (and, only slightly whimsically, can be regarded as the BJT's *soul*, since it mediates so much of the device's personality) having a value of some 3.6E-18 amps for a VBE of 800 mV at IC = 100 mA and temperature of 27˚C. It's amazing to me, after having stared at this equation for the better part of my life, how very profound it is, traceable to fundamental aspects of carrier statistics in semiconductor materials. It is the well-spring of BJT magic. It takes but a moment to find that the transconductance dIC/dVBE of a single device is:

gm = IC/VT (2)

where VT is, of course, the thermal voltage kT/q, 26 mV at 30˚C.

This remains as true for a modern complex SiGe heterojunction transistor as it was for the primitive junction-alloy devices that came along quite shortly after Bardeen, Brattain and Shockley went whooping up and down the halls of Bell Labs jubilantly shouting *Eureka!* It's fair to say that (1) and (2) are the most remarkable of all equations in modern electronics.

The Level-0 model also conveniently omits such pesky set-backs as the finite base current and Early voltage of a BJT, its ohmic resistances and parasitic capacitances, base transit time and other effects. Accordingly, we set BF = BR = VAF = VAR = 1E6, and most other parameters to zero. Crazy? Not really: This is just what first-order textbook analyses do, without drawing attention to the fact. It is nonetheless surprising just how much of the reality of an IC's behavior emerges from the application of this simple *translinear* model to circuit analysis. For example, the minimum permissible supply voltage will usually be correctly modeled; the shot noise will be right on the money; most of the temperature behavior; and, for the present purposes, an important distortion mechanism in OPAs will be quite accurately predicted.

The usual starting point in analyzing something like the circuit of Fig. 1 is to figure out the effective gm of the input stage, and note that its output current is integrated in CC, whose impedance is $j/2\pi fCC$, thus providing a voltage gain of magnitude $gm/2\pi fCC$ with a constant phase-shift of -90˚. Then the discussion turns to such practical matters as the finite dc gain (much less important than often believed), which will be dominated by Early voltage effects, the additional phase shift at frequencies above the unity-gain frequency, $gm/2\pi CC$, and so on. All this attention lavished on the small-signal view, however, blithely overlooks the darker side of the OPA's character.

While equation (2) is delightfully clear, we need to keep in mind that it is a *derivative*, the limit value of the ratio $\Delta IC/\Delta VBE$ as ΔVBE tends to zero. It doesn't take much ΔVBE to change the gm: it will double for a mere 18 mV (at T = 27˚C); just 18 mV less and the gm will halve. This is the root of all kinds of evil in supposedly linear circuits built using BJTs, with their sometimes valuable but at other times unwelcome exponential behavior. Nothing like this had ever been seen in vacuum-tubes, and, for all the perversities lurking in CMOS transistors, this is certainly not one of them. Indeed, vacuum-tube worshipers should be right at home with the mushy V-I curves of CMOS devices.

The lower, more linear, gm of field-effect transistors (FETs) is very useful in lowering the open-loop (and thus, the closed-loop) distortion of an op amp. Of course, a low input bias current is also valuable in a general-purpose op amp. Partly for historical reasons (but also because of the lower input-offset voltage than possible in CMOS), *junction* FETs are often employed in commercial standalone op amps. Incidentally, the first reported monolithic JFET op amp was designed by my good friend George Wilson, back in our Tektronix days; for good measure, he threw in a new type of BJT current mirror, now widely known as the Wilson mirror.

But, here we are, talking about *cures* before we've discussed the *sickness*. Let's go back to our starting point, the exact sinusoidal output voltage E*sin*wt in Fig.

1. That circuit shows that the ac current in the capacitor is simply:

Iac = EwCC *cos*wt (3)

and this must also be the difference current, IC2 - IC1, generated by the input pair. The first thing to notice, then, is that the input stage has to do some work so as to make the output happen: it's not just an 'error detector' which, by virtue of the integrator inside the loop, merely keeps returning to a fully-balanced state, as is sometimes suggested in op-amp textbooks. That is true (roughly) at dc and even at very low frequencies (a few Hertz), but obviously cannot ever be true at frequencies which are a substantial fraction (even as small as one-thousandth) of the unity-gain frequency. Note, too, that this situation is in no way eased by including more open-loop gain, for example, by using a Darlington-connected mirror or a Darlington stage in place of Q5. These often help to increase the dc gain, and lower the input offset voltage by improving the balance of the circuit, but *they do nothing to increase the ac gain*.

In a linear analysis, we'd take the view that the current in CC is caused by the difference between the applied input VIN and the output E*sin*wt acting on the gm of the input pair. Now, from (2), and noting that the collector currents in Q1 and Q2 are equally IT/2 in a small-signal situation, we find that the net gm to the output of the current mirror is:

gm = IT/2VT (4)

Only a fraction, b, of the output is present at the base of Q2, and there is a *sign reversal* of the gm for this pnp stage, so we have:

(VIN - b E*sin*wt) IT/2VT = IC2 - IC1 = Iac
= EwCC*cos*wt, (5)

from which we can calculate that the *input required to generate this output* is:

VIN = b E*sin*wt + EwCC*cos*wt / (IT/2VT) (6)

That is, there has to be an additional *quadrature* component of *relative* magnitude:

2G VTwCC*cos*wt / IT, (7)

present at the input. Since the unity-gain frequency for this OPA, w1 = gm/CC, can alternatively be written as IT/2VTCC it follows that the relative magnitude of this quadrature component can be written as G (w/w1) *cos*wt. Thus, it will be equal in amplitude to the sine term *at the input* when the signal frequency is equal to one-tenth the crossover frequency. In general, the phase angle will be:

j = *arctan* G (w/w1) (8)

For the voltage-follower case, G=1, the sum of these two input terms is root-2, or about 1.4 times, that of the output, so the gain is down to 0.71 at the corner frequency, and the phase angle *measured from input to output* is -45˚. All this is standard op amp analysis. Unfortunately, it is considerably in error, because of the original assumption of *a linear gm cell* at the input.

If you have the patience to follow through with the rest of this analysis, you'll discover that the gain magnitude is different; that there is significant *odd-harmonic distortion* even at frequencies *well below* the unity-gain crossover; and that the phase angle is not only a function of frequency, but also of amplitude -- that is, *the op amp generates a peculiar kind of amplitude-to-phase modulation,* something which one should not expect of a linear system. But, then, an op amp is by no means as pristine in this respect as the textbooks suggest, and that gm cell has very strong open-loop distortion for even quite small inputs.

To anticipate our analysis, it is found that the third-harmonic distortion of a simple bipolar differential pair reaches 1% for a sinewave input amplitude of about 18 mV at 27˚C. Now, recall that the *open-loop ac gain* is only w1/w, and consider what the *nonlinear* differential input (the error voltage DV in Fig. 2) must be to provide a 10-V sinewave output at, say, one-hundredth of w1, we will find it to be a whopping 100 mV, much higher that the 1% voltage, and some 10% of the 1-V input at a gain of 10. Our question is, what will be *actual* differential error voltage be in practice, and the actual distortion? Finding the answers requires a bit more mathematics, with the device nonlinearities included.

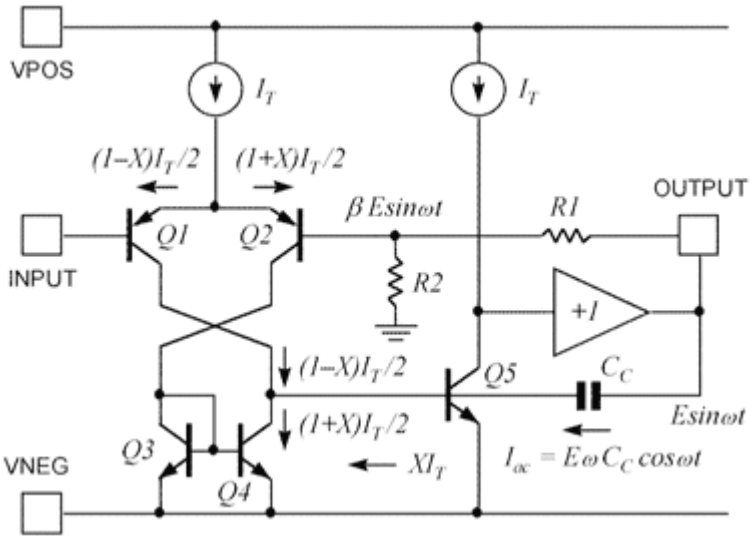

Fig. 2 The Typical OPA with Further Annotations

The approach used here is to introduce the *modulation factor*, X, (Fig. 2) as an alternative way of defining IC1 and IC2. If we now apply the basic junction equation (1) it is readily found that the input voltage can be expressed as:

DV = VT *log* (1+X)/(1-X) = 2VT *arctanh* (X) (9)

Using the first two terms in a polynomial expansion of the inverse hyperbolic tangent function, we have:

DV = 2VT (X+X3/3) (10)

It will be apparent from inspection of the figure that:

Iac = XIT (11)

so,

X = EwCC*cos*wt/IT (12)

Inserting this value into (10):

DV = 2VT {EwCC*cos*wt/IT + (1/3)(EwCC/IT)3*cos*3wt } (13)

and, with the substitution w1 = IT/2VTCC from the linear analysis, above, and a little bit more rearranging, we can write:

DV = E { (w/w1)*cos*wt + (1/3)(E/2VT)2(w/w1)3*cos*3wt } (14)

There's just one more substitution, then we're almost home. It is to expand the *cos*3 factor into its fundamental and third-harmonic components:

*cos*3wt (3/4)*cos*wt + (1/4)*cos*3wt

This gets us:

DV = E(w/w1) [{1+(3/48) (w/w1)2(E/VT)2} *cos*wt
+ (1/48) (w/w1)2(E/VT)2 *cos*3wt] (15)

This expression for the difference voltage at the input of the OPA is just a little complicated, and we have yet to sum the sinusoidal component of the output delivered to the base of Q2 via the gain-setting feedback network. It illustrates how even a single source of distortion can quickly generate rather messy expressions. On the other hand, it is not too hard to reduce all this to something quite tractable and useful.

We can start with the phase angle at the fundamental frequency. Without the third harmonic term, the full input may be written:

VIN = bE*sin*wt + E(w/w1){1+(1/16) (w/w1)2(E/VT)2} *cos*wt (16)

so the phase angle of VIN relative to the output is:

j = *arctan* G (w/w1) {1+(1/16) (w/w1)2(E/VT)2} (17)

(Of course, the phase *from input to output* has a negative sign). Now, if it were not for that term involving E2 and w2, this would be the same as for the linear

case: the phase for G = 10 would be 5.71˚ at w = w1/100. However, if we include this extra term, it is clear that the phase angle is higher, by an amount that increases with the square of frequency and the square of the amplitude E. Assume again that the signal frequency is 1/100th the unity-gain frequency, that E = 2 V, and G = 10, all of which represent quite moderate conditions for an op amp. Then, the actual phase angle is 6.01˚ by E = 5, it has increased to 9.14˚. This is not what we would expect of a *linear* amplifier, whose phase should be quite independent of amplitude. In certain applications, this excess phase and its variation with signal level will be very troublesome.

The third-harmonic distortion is the ratio of the second term in (15) to the *vector sum* of the sine and cosine fundamentals in (17). Since we have chosen to work 'backwards', from input to output, this is the distortion referred to VIN. For moderate levels, though, the percentage is very similar whichever way round the calculation is made. Further, to simplify the calculation, we can assume that the magnitude of the cosine term (in 18) is fairly small; with this assumption, the distortion calculation will be pessimistic. Taking the ratio, we find:

HD3 = G(1/48)(E/VT)2(w/w1)3, (19)

that is, the third-harmonic distortion increases with the square of the output voltage and the cube of the signal frequency. Once again assuming G = 10, w = w1/100, we find that HD3 evaluates to 0.125% (-58 dBc) for E = 2 V, 0.5% (-46 dBc) for E = 4 V, and 0.78% (-42 dBc) for E = 5V. Of course, these are all higher than would be calculated if the usual 'open-loop gain' figure, AO, of several hundred thousand, were used to estimate distortion.

The results found by simulation, using a 'forward-causal' path, are slightly better for low values of E (partly because we neglected the increase in the vector sum of the input due to the fundamental cosine term), being -55.6 dBc for E = 2 V, somewhat worse than predicted at E = 4 V, where the simulation shows -42 dBc, and considerably higher, by a factor of 2.3, at E = 5 V (-34.6 dBc). This is because of the failure of the polynomial expansion of *arctanh*, and the imminent onset of full *slew-rate limiting*, which occurs at an output amplitude of 5.17 V for this 1-MHz OPA.

It must be again noted that this is for operation at w1/100, and a gain of ten; op amps are often used at frequencies as high as w1/5, where the HD3, according to (19), should be 203 or 8,000 times higher! Clearly, something is badly wrong if the analysis predicts distortion larger than 100 %. The reason for the failure in our analysis is not hard to find; the basic assumption of a purely sinusoidal output is inappropriate under such conditions, since for high frequencies and amplitudes at the input, the nonlinearities become extreme, and the rate-ofchange at the output is then determined by the slew-rate, determined simply by the maximum current available to charge CC, which is ˚IT when the gm stage is overdriven. Thus, the slew-rate is ˚IT/CC. Since w1 = IT/2VTCC, we can express the slew rate as 2w1VT. It is a miserable 0.325 V/ms for a 1-MHz OPA, and there's only one way to increase it, using this particular gm stage, which is to raise w1.

Modern op amps -- even ones of this unity-gain frequency -- are, of course, far better than this. One reason is the use of a modified input (gm) stage having the capacity to cope with a *large open-loop error signal.* This is achieved either by using some type of emitter degeneration -- in the case of BJT input stages -- or the use of optimized multi-*tanh* cells, which can exhibit ultra-low distortion for a DV of up to ˚200 mV. Various sorts of Class-AB cells, which idle at a low bias current but are able to generate large peak currents, essentially proportional to DV, are sometimes used. Complementary BJT processes are valuable in such input stages, as well as in low-distortion output stages. The use of JFETs or MOSFETs, which inherently have a 'weak' gm and a large signal capacity, is common in improving the open-loop linearity.

There is an interesting trend here: operational amplifiers started out (back in analog computing days) with a totally different set of objectives to those surrounding amplifiers for ac amplification with low distortion. In the 'op amp paradigm,' the idea was to make the open-loop gain *so very high* that the function is determined solely by the external components. However, as we've seen, that is far from the case for a typical OPA operating at moderate gains and frequencies. In the second design style, typified by audio power amplifiers, the objective is to achieve almost *perfect linearity without feedback*, and then use a small amount of feedback to squeeze out the last gram of linearity. Designers of modern high-performance op amps now understand the critical importance of pursuing the latter paradigm in all high-frequency applications. This equates almost completely to achieving extremely high slew rates; value of over 5,000 V/ms are nowadays available in monolithic OPAs.

In another approach to improving HF linearity, the gm input stage is replaced by a *current-conveyor*, which is another type of Class-AB cell. (In fact, high-slew OPAs often used what is effectively two such cells in a differential arrangement.) A current conveyor is the starting point of a *current-feedback amplifier*, also called a transimpedance amplifier (TZA). This type of amplifier can, in principle, be designed to be *totally free* of slew-rate limitations, one of its main attractions. Another is that, unlike an OPA, in which the so-called 'virtual ground' or 'summing node' exhibits a low impedance only by the action of global feedback, becoming very high at frequencies close to w1, the TZA provides a low summing-node impedance (ohms) even with the feedback completely removed, from dc to many hundreds of megahertz.

This useful and interesting special-purpose amplifier topology will be the subject of my next column. Later, another fundamental building block, which I call the Active Feedback Amplifier, or AFA, will be discussed. In my opinion, this structure, which, unlike the OPA, has high open-loop linearity and excellent closed-loop linearity due to its distortion-canceling topology, and a very high degree of versatility arising from its dual fully-differential inputs, has the potential to eclipse the OPA in all applications involving the manipulation of purely voltagemode
signals. Being a super-set, it can do anything that a conventional op amp can do, plus a whole lot more. Watch this space!

Barrie Gilbert (IEEE Member 1962, Fellow, 1984), b. 1937, in Bournemouth, England, pursued an early interest in solid- state devices at Mullard Ltd, working on first-generation planar ICs. Emigrating to the US in 1964, he joined Tektronix, in Beaverton, OR, where he developed the first electronic knob-readout system, and other advances in instrumentation. Between 1970-1972 he was Group Leader at Plessey Research Laboratories. He later joined Analog Devices Inc. and was appointed ADI Fellow in 1979. He manages the development of highperformance analog ICs at the NW Labs in Beaverton.

For work on merged logic he received the IEEE "Outstanding Achievement Award" (1970) and the IEEE Solid-State Circuits Council "Outstanding Development Award" (1986). He was Oregon Researcher of the Year in 1990, and received the Solid-State Circuits Award (1992) for "Contributions to Nonlinear Signal Processing". He has written extensively about analog design and has five times received ISSCC Outstanding Paper Award. He has been issued over 40 patents and holds an Honorary Doctorate from Oregon State University.