

Distortion in Elementary Transistor Circuits

Willy Sansen, *Fellow, IEEE*

Abstract—In this paper the distortion components are defined for elementary transistor stages such as a single-transistor amplifier and a differential pair using bipolar transistors or MOST's. Moreover, the influence of feedback is examined. Numerical examples are given for sake of illustration.

Index Terms—Amplifiers, distortion, feedback, intercept point.

I. INTRODUCTION

DISTORTION analysis has gained renewed interest because it is responsible for the generation of spurious frequency bands in telecommunication circuits. Therefore, it is reviewed starting with the most elementary circuit blocks [2], [4]–[6].

Distortion actually refers to the distortion of a voltage or current waveform as it is displayed versus time, i.e., as seen on an oscilloscope. Any difference between the shape of the output waveform versus time and the input waveform, except for a scaling factor, is called distortion. For example, the flattening of a sinusoidal waveform is distortion. The injection of a spike on a sinusoidal waveform is called distortion as well. Several kinds of distortion occur. They are defined first.

A. Linear and Nonlinear Distortion

Linear distortion is caused by the application of a linear circuit, with a nonconstant amplitude or phase characteristic. As an example, the application of a high-pass filter (of first order) to a square waveform causes distortion, as shown in Fig. 1. In a similar way, the application of a low-pass filter reduces the high-frequency content in the output waveform, as shown in Fig. 2.

Nonlinear distortion is caused by a nonlinear transfer characteristic. For example, the application of a sinusoidal waveform to the exponential characteristic of a bipolar transistor causes a sharpening of one top and flattening of the other one (see Fig. 3). This corresponds to the generation of a number of harmonic frequencies of the input sinusoidal waveform. These are the nonlinear distortion components.

B. Weak and Hard Distortion

When the nonlinear transfer characteristic has a gradual change of slope (as shown in Fig. 3), then the quasi-sinusoidal waveform at the output is still continuous. This is not the case when the transfer characteristic has a sharp edge, as shown in Fig. 4 for a class B amplifier. Part of the sinusoidal waveform

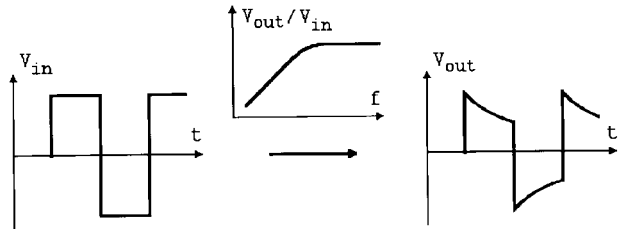


Fig. 1. Application of a high-pass filter causes linear distortion because of the reduction of the low frequencies.

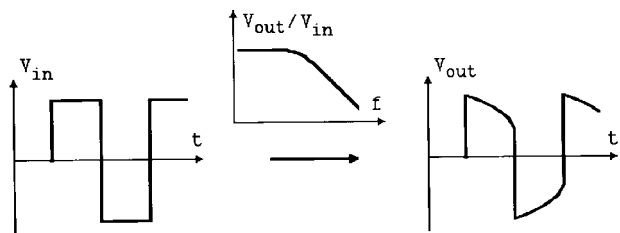


Fig. 2. Application of a low-pass filter causes linear distortion because of the reduction of the high frequencies.

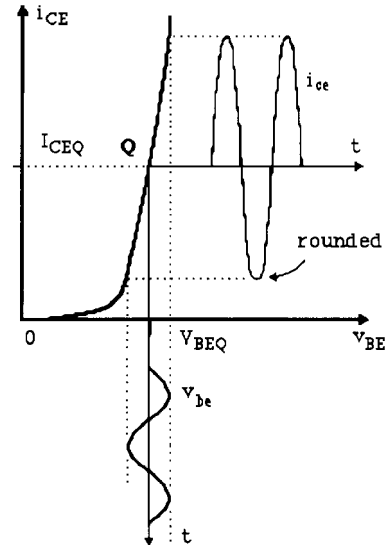


Fig. 3. Generation of nonlinear distortion caused by the nonlinear i_C - v_{BE} characteristic.

is then simply cut off, leaving two sharp corners. These corners generate a large number of high-frequency harmonics. They are sources of hard distortion.

In the case of weak distortion, the harmonics gradually disappear when the signal amplitude becomes smaller. They are never zero, however. They can easily be calculated from a Taylor series expansion around the quiescent or operating

Manuscript received July 31, 1997; revised June 15, 1998. This paper was recommended by Guest Editor A. Rodriguez-Vazquez.

The author is with ESAT-MICAS, K.U. Leuven, Leuven, Belgium.

Publisher Item Identifier S 1057-7130(99)01766-8.

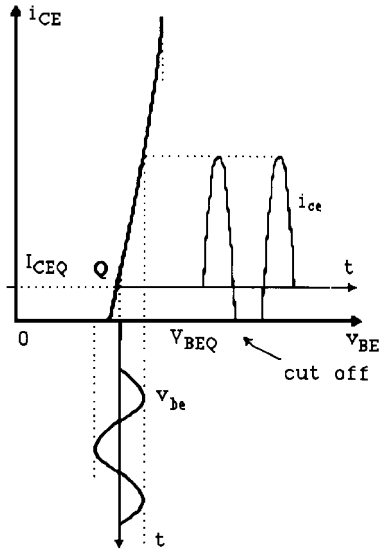


Fig. 4. Generation of hard nonlinear distortion.

point, as will be carried out in next paragraph.

Hard distortion harmonics, on the other hand, suddenly disappear when the amplitude of the sinusoidal waveform falls below the threshold, i.e., the edge of the transfer characteristic. Also they are much more difficult to calculate. Since they can be avoided altogether by limiting the output signal amplitudes to sufficiently low levels, they will not be discussed any further.

In this paper, the nonlinear distortion will be calculated for elementary bipolar and MOST amplifier and buffer stages. Also the influence of negative feedback is derived. First, however, the several definitions have to be reviewed to describe the weak nonlinear distortion components.

II. WEAK-DISTORTION COMPONENTS

Let us consider an amplifier with a weak nonlinearity as in Fig. 3. Both input and output signals vary with time. They are denoted by $u(t)$ and $y(t)$ or, in shorthand, u and y . At low frequencies, the output y of this amplifier can be expressed in terms of its input u by a power series

$$y = a_0 + a_1 u + a_2 u^2 + a_3 u^3 + \dots \quad (1)$$

Coefficient a_0 represents the dc component of output signal y . Coefficient a_1 represents the linear gain of the amplifier, whereas coefficients a_2, a_3, \dots , represent its distortion.

Coefficients a_1, a_2 , and a_3 can be obtained from the analytic expression of the function $y(u)$ as given by

$$a_n = \frac{1}{n!} \left. \frac{d^n y}{du^n} \right|_{u=0}. \quad (2)$$

Application of a cosine waveform of frequency ω and amplitude U at the input of that amplifier yields output components at all multiples of ω . It is obtained by trigonometric manipulation. Under low-distortion conditions, only second- and third-order distortion components are considered. By use of the expressions

$$u = U \cos(\omega t) \quad (3)$$

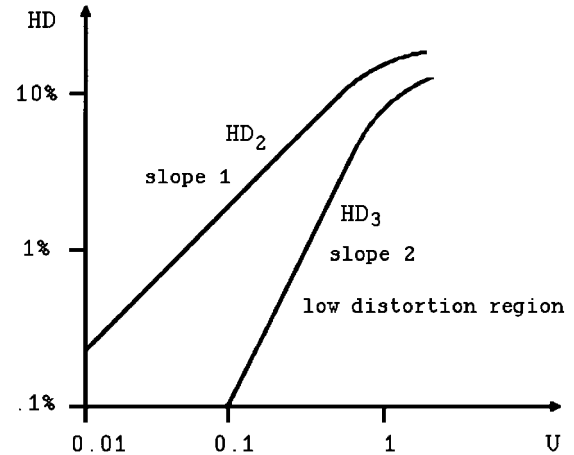


Fig. 5. Distortion components versus normalized input voltage.

and

$$\begin{aligned} \cos^2 x &= 1/2(1 + \cos 2x) \\ \cos^3 x &= 1/4(3 \cos x + \cos 3x) \end{aligned}$$

(1) thus becomes

$$\begin{aligned} y = \left(a_0 + \frac{a_2}{2} U^2 \right) + \left(a_1 + \frac{3}{4} a_3 U^2 \right) U \cos(\omega t) + \frac{a_2}{2} U^2 \\ \cdot \cos(2\omega t) + \frac{a_3}{4} U^3 \cos(3\omega t) + \dots \quad (4) \end{aligned}$$

Odd-order distortion, and especially a_3 , thus modifies the signal component at the fundamental frequency. Term $a_3 U^2$ can be neglected, however, with respect to a_1 , provided the signal amplitude U is sufficiently small.

Harmonic distortion is then defined as follows. The n th harmonic distortion (HD_n) is defined as the ratio of the component of frequency $n\omega$ to the one at the fundamental ω . Application to (4) yields

$$HD_2 = \frac{1}{2} \frac{a_2}{a_1} U \quad (5a)$$

and

$$HD_3 = \frac{1}{4} \frac{a_3}{a_1} U^2. \quad (5b)$$

It is important to note that HD_2 is proportional to U and HD_3 to U^2 . Increasing the input signal level by 1 dB thus increases the HD_2 by 1 dB and the HD_3 by 2 dB. These relationships hold true for all values of U , which are not too large. This is the region where the so-called low-distortion conditions are valid. For even larger values of U , the values of HD_2 and HD_3 flatten off with increasing U as shown in Fig. 5.

In this paper, the analyzes are limited to the region of low distortion, i.e., where U is sufficiently small, i.e., where HD_2 is proportional to U and HD_3 to U^2 .

Also the total harmonic distortion THD is given by

$$THD = \sqrt{HD_{2^2} + HD_{3^2} + \dots} \quad (5c)$$

It is not very useful as it does not give a clear dependence on the input signal level.

Application of the sum of two cosine waveforms of frequencies ω_1 and ω_2 and both of amplitude U at the input gives rise to output signal components at all combinations of ω_1, ω_2 and their multiples. Under low-distortion conditions, the number of terms can be reduced to the ones caused by coefficients a_2 and a_3 only. They are mapped versus frequency in Fig. 6(a) for $\omega_1 = 2\pi$ (10 MHz) and $\omega_2 = 2\pi$ (11 MHz). A real frequency spectrum for frequencies 10.695 and 10.705 MHz is shown in Fig. 6(b).

Second-order *intermodulation distortion* (IM_2) is then defined by the ratio of the component at frequency $\omega_1 \pm \omega_2$ to the one at ω_1 or ω_2 . Under low-distortion conditions

$$IM_2 = \frac{a_2}{a_1} U. \quad (6a)$$

Third-order intermodulation distortion (IM_3) can be detected at the frequencies $2\omega_1 \pm \omega_2$ and $2\omega_2 \pm \omega_1$ [see Fig. 6(a)].

It is given by the ratio of the component at frequency $2\omega_2 - \omega_1$ (or one of the other three frequencies), which is $\frac{3}{4}a_3U^3$, to the fundamental, which is a_1U , as given by

$$IM_3 = \frac{3}{4} \frac{a_3}{a_1} U^2. \quad (6b)$$

Comparison of the four equations above shows that

$$IM_2 = 2HD_2 \quad (7a)$$

$$IM_3 = 3HD_3. \quad (7b)$$

Under low-distortion conditions, there is thus a one-to-one correspondence between harmonic and intermodulation distortion. It is thus sufficient to specify only one of them.

Note that two of the four equal IM_3 components, i.e., the ones at the frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, occur closely to the two fundamentals. This is one reason why they are more important than the HD_3 components. In music signals for instance, it is quite conceivable that two peaks which are close together in frequency, generate intermodulation products in the same frequency range. At high frequencies, these products may already be reduced by the amplitude-frequency characteristic. In Fig. 6(b), the IM_3 peaks are clearly visible at frequencies 10.685 and 10.715 MHz. The IM_3 is thus about -40 dB. The other two IM_3 components around 30 MHz are already heavily attenuated (not in the picture).

A second reason why the measurement of IM_3 is preferred above the one of HD_3 is that the value of IM_3 is three times larger than the one of HD_3 and hence easier to measure.

For these reasons, the value of IM_3 is always preferred.

Another important characteristic and often used point is the IM_3 *intercept*, or IP_3 . It is the value of the input signal where the extrapolated curves of the components of IM_3 and the fundamental coincide. This is shown in Fig. 7.

The output components at the fundamental frequencies and at the IM_3 frequencies are plotted versus the input voltage V_{in} . They are given by, respectively, a_1V_{in} and $\frac{3}{4}a_3V_{in}^3$ (note that a_1 is again dimensionless but that a_3 has V^{-2} as dimension). IM_3 is the ratio of both components. The point where both components coincide is IP_3 . It is thus also the point where

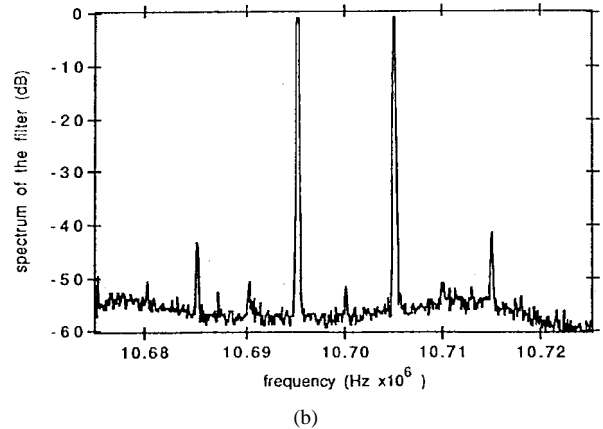
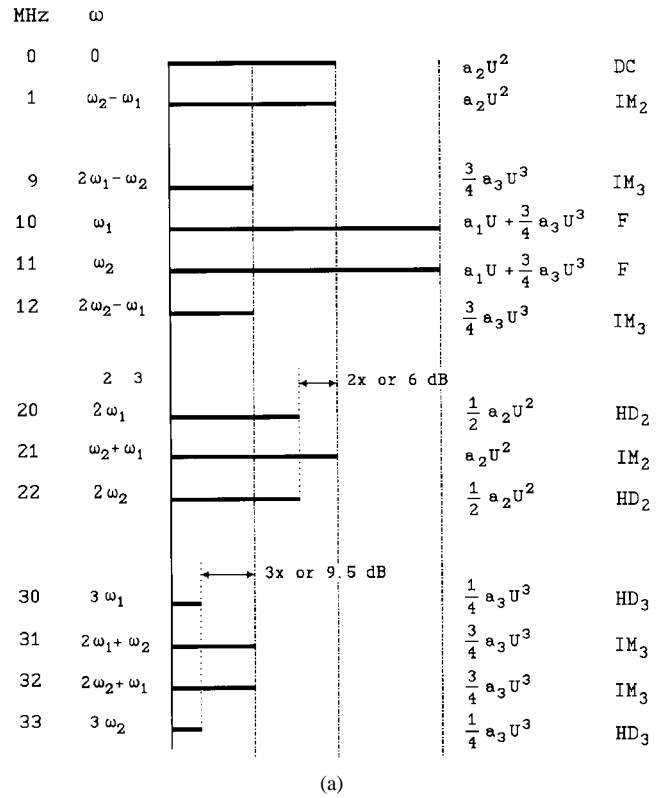


Fig. 6. (a) Second- and third-order harmonic and intermodulation components. (b) Intermodulation distortion of a 10.7 MHz filter [3].

$IM_3 = 1$. This point is easy to calculate from (6b) and is given by

$$IP_3 = \sqrt{\frac{4}{3} \frac{a_1}{a_3}} \quad (8)$$

or

$$IP_3 = V_{IN} \frac{1}{\sqrt{IM_3}}$$

or

$$IP_{3dB} = V_{INdB} - \frac{1}{2}IM_{3dB}.$$

Obviously, the smaller a_3 , the larger the value of IP_3 .

Another related measure for the distortion is the *Intermodulation free dynamic range* ($IMFDR_3$). The dynamic range is

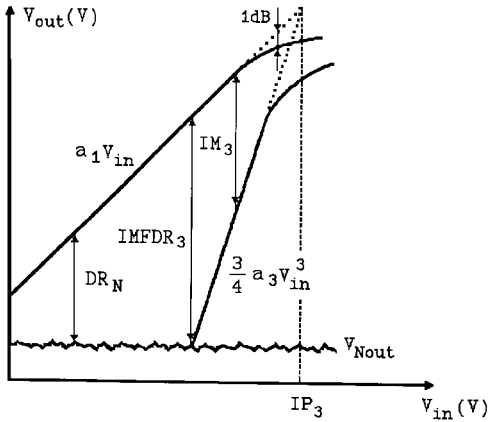


Fig. 7. Fundamental and IM_3 components versus input voltage.

the ratio of the maximum output signal $a_1 V_{in}$ to the output noise V_{Nout} , as shown in Fig. 7. It is thus given by

$$DR = \frac{a_1 V_{IN}}{V_{Nout}} = \frac{V_{IN}}{V_{Nin}} \quad (9)$$

or

$$DR_{dB} = V_{IN,dB} - V_{Nin,dB}$$

in which $V_{Nin} = V_{Nout}/a_1$ is the input noise (in V_{RMS}).

The IMFDR is the largest possible DR without IM_3 distortion. It is thus obtained at the input voltage where the output noise equals the IM_3 component (see Fig. 7), or where

$$V_{Nout} = \frac{3}{4} a_3 V_{in}^3$$

which yields

$$V_{in,IMF} = \sqrt[3]{\frac{4}{3} \frac{1}{a_3} V_{Nout}} \quad (10a)$$

Substitution of this value in (9) finally gives

$$IMFDR_3 = \frac{a_1 V_{in,IMF}}{V_{Nout}} = \sqrt[3]{\frac{4}{3} \frac{a_1^3}{a_3} \frac{1}{V_{Nout}^2}} = \sqrt[3]{\frac{4}{3} \frac{a_1}{a_3} \frac{1}{V_{Nin}^2}}$$

or

$$IMFDR_3 = \left(\frac{IP_3}{V_{Nin}} \right)^{2/3} \quad (10b)$$

and

$$IMFDR_{3,dB} = \frac{2}{3}(IP_{3,dB} - V_{Nin,dB}).$$

An alternative, albeit less accurate, way to characterize distortion is the -1 dB compression point (see Fig. 7). It is the value of V_{in} where the fundamental component is compressed by 1 dB, and is denoted by $V_{in,1dBc}$. This value can be approximately calculated from (4). Indeed, the compression is caused by the second term (in a_3) of the coefficient of $\cos(\omega t)$. A reduction of 1 dB is a reduction to 0.122. The resultant value of $V_{in,1dBc}$ is thus about given by

$$V_{in,1dBc} = \sqrt[3]{0.122 \frac{4}{3} \left| \frac{a_1}{a_3} \right|} \quad (11)$$

or

$$V_{in,1dBc} = \sqrt{0.122} IP_3$$

and

$$V_{in,1dBc,dB} = IP_{3,dB} - 9.64 \text{ dB}.$$

The difference between both is thus almost 10 dB. The measurement of the -1 dB compression point is thus an easy way to obtain the value of IP_3 .

There are several other ways to describe the distortion caused by coefficients such as cross-modulation distortion, triple beat, etc. There is nevertheless always a constant relationship of the type (7) between them. Therefore, only one more distortion is shortly discussed. It is the cross-modulation distortion.

For the determination of *cross-modulation distortion*, again, two carrier frequencies ω_1 and ω_2 are required. The first one, however, is modulated by a modulating signal at low frequency ω_{lo} . The modulation index, which is ratio of the amplitude of the modulating signal to the one of the carrier, is denoted by m_c . A nonlinear transfer characteristic causes the modulation to be transferred from the first carrier to the other one. As a result, the second carrier is modulated as well. This causes mixing of the channels in cable TV, etc., and is thus to be avoided.

The modulation index of the other channel is a measure of the distortion, and is called the cross-modulation distortion. It is given by

$$CM_3 = \frac{3}{4} m_c \frac{a_3}{a_1} U^2 \quad (12a)$$

or

$$CM_3 = m_c IM_3. \quad (12b)$$

Note that CM_3 is only generated by the third-order terms of the power series, which describe the nonlinearity. Since it is closely related to IM_3 , it will not be discussed any further.

III. DISTORTION IN A BIPOLAR TRANSISTOR AMPLIFIER

In a bipolar transistor, the collector current i_C is controlled by the base-emitter voltage v_{BE} as given by

$$i_C = I_{CS} \exp\left(\frac{v_{BE}}{V_t}\right) \quad (13)$$

in which I_{CS} is the collector saturation current (see [1, Ch. 1]) and $V_t = kT/q = 26$ mV at 29°C (or 302 K).

The transistor is biased at a specific dc value of v_{BE} , i.e., V_{BE} in quiescent point Q of the characteristic (see Fig. 3). A small variation of this voltage causes a variation in collector current. These variations or ac components of the collector current and the base-emitter voltage can be expressed as given by

$$i_C = I_C + i_c \quad (14)$$

and

$$v_{BE} = V_{BE} + v_{be}.$$

Expression (13) thus results in

$$I_C + i_c = I_{CS} \exp\left(\frac{V_{BE} + v_{be}}{V_t}\right). \quad (15)$$

After division of both terms by the value of the quiescent current I_C , we obtain

$$1 + i = \exp\left(\frac{v_{be}}{V_t}\right) \quad (16)$$

with $i = i_c/I_C$, which is called the relative current swing. It is the current variation in the transistor, normalized to the quiescent or dc current. It is a measure of the fraction of the dc current in the transistor, which is used to generate ac output signal. It will be used throughout this section to compare distortion performance.

For small peak base-emitter voltages $v_{be} < V_t$, the exponential of (16) can be expanded in a Taylor series. Indeed, for $x < 1$, we know that

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots \quad (17)$$

and application of this expansion to (16) yields

$$i_p = \frac{I_{cp}}{I_C} = \frac{V_{bep}}{V_t} + \frac{1}{2} \left(\frac{V_{bep}}{V_t}\right)^2 + \frac{1}{6} \left(\frac{V_{bep}}{V_t}\right)^3 + \dots \quad (18)$$

in which i_p is the peak value of the relative current swing, and V_{bep} is the peak value of the ac base-emitter input signal.

For small input signals, only the first term in (18) has to be retained, which leads to

$$I_{cp} = g_m \cdot V_{bep} \quad (19)$$

which is well expected. Moreover, in first order, the peak value of the relative current swing is derived from the peak input voltage as given by

$$i_p = \frac{V_{bep}}{V_t}. \quad (20)$$

Finally, identification of (18) with (1) shows that for $i_p = y$ and $V_{bep} = u$, the coefficients are $a_0 = 0$, $a_1 = 1$, $a_2 = 1/2$, and $a_3 = 1/6$. Use of the (5)–(10) and substitution of V_{BEp} by i_p as given by (20) yields

$$\text{IM}_2 = 2\text{HD}_2 = \frac{1}{2} \frac{V_{bep}}{V_t} = \frac{i_p}{2} \quad (21a)$$

and

$$\text{IM}_3 = 3\text{HD}_3 = \frac{1}{8} \left(\frac{V_{bep}}{V_t}\right)^2 = \frac{i_p^2}{8}. \quad (21b)$$

For example, a peak ac current of 100 μA in a bipolar transistor, carrying 1 mA, causes a peak relative current swing of $i_p = 0.1$, and $\text{IM}_2 = 5\%$ (or $\text{HD}_2 = 2.5\%$), and also $\text{IM}_3 = 0.125\%$ (or $\text{HD}_3 = 0.04\%$). For this ac current, a peak input voltage is required of only 2.6 mV or to 1.84 mV_{RMS}. A larger peak current swing of 0.5 leads to $\text{IM}_3 = 3.1\%$, for which an input signal amplitude of 9.2 mV_{RMS} is sufficient. For $R_L = 5.2 \text{ k}\Omega$, the voltage gain then equals -200 .

Finally, the value of the

$$\text{IP}_3 = \sqrt{8} = 2.8 \quad (22)$$

on the scale of the current swing

$$\text{IP}_3 = \sqrt{8}V_t = 2.8V_t$$

or on the input voltage scale.

This corresponds with an input voltage of 73 mV.

This is quite small. A bipolar transistor with 1 mA has a $g_m = 38.4 \text{ mS}$. With a base resistor of $r_B = 100 \Omega$, its equivalent input noise is the noise of $r_B + 1/2g_m = 152 \Omega$ (see [1]). This corresponds with 1.56 nV_{RMS}/√Hz. For a bandwidth of 200 kHz, the noise level $V_{Nim} = 0.70 \mu\text{V}_{RMS}$. As a result $\text{IMFDR}_3 = 2215$ or 67 dB.

It is important to note that distortion components can always be described by means of the input voltage drive and by the current swing. The latter way has a number of advantages. The current swing already includes the effects of the transconductance and of the feedback such that the expressions become simpler and very much comparable. They will be used throughout this paper.

From these numbers, it is clear that only small input signal amplitudes can be applied to a bipolar transistor. Also, a current swing of 0.5 already corresponds with a high distortion region, as shown in Fig. 5. For small values of V_{bep} and i_p , relations (21) and (22) hold. On a double logarithmic scale, straight lines result with slopes of 1 and 2, respectively. Doubling V_{bep} and i_p thus quadruples the third-order distortion. At higher values of V_{bep} and i_p , however, the values of the distortion are quite high but do not increase (see Fig. 5) any further. These values have been calculated by means of transient analyzes in SPICE, followed by Fourier analyses.

IV. DISTORTION IN A MOSFET AMPLIFIER

For a MOST, the analysis is very similar as for a bipolar transistor. Only the transfer characteristic $I_{DS}-V_{GS}$ is quadratic and not exponential. Less distortion is thus expected.

The drain current i_{DS} and gate-source voltage v_{GS} of a MOST are in first-order related by

$$i_{DS} = \frac{\beta}{2} (v_{GS} - V_T)^2 \quad (23)$$

in which β is the transconductance factor, which includes the size W/L , and V_T is the threshold voltage.

The transistor is biased at a specific dc value of v_{GS} , i.e., V_{GS} in a quiescent or operating point Q . A small variation of this voltage causes a small variation in drain current. They are related by

$$i_{DS} = I_{DS} + i_{ds} \quad (24)$$

and

$$v_{GS} = V_{GS} + v_{gs}.$$

Expression (23) thus becomes

$$I_{DS} + i_{ds} = \frac{\beta}{2} (V_{GS} - V_T + v_{gs})^2. \quad (25)$$

Subtraction of I_{DS} from both sides, and division by the value of the quiescent current I_{DS} , yields

$$i = \frac{i_{ds}}{I_{DS}} = \left(1 + \frac{v_{gs}}{V_{GS} - V_T}\right)^2 - 1 \quad (26)$$

or

$$i_p = \frac{I_{dsp}}{I_{DS}} = \frac{2V_{gsp}}{V_{GS} - V_T} + \frac{1}{4} \left(\frac{2V_{gsp}}{V_{GS} - V_T}\right)^2 \quad (27)$$

in which i_p and V_{gsp} are the peak values of the relative current swing and the gate-source input voltage, respectively.

For small signals, only the first term in (27) has to be retained, which yields

$$I_{dsp} = \frac{2I_{DS}}{V_{GS} - V_T} V_{gsp} \quad (28)$$

or

$$= g_m V_{gsp}$$

as expected.

Also, the peak relative current swing is related to the input drive by

$$i_p = \frac{2V_{gsp}}{V_{GS} - V_T}. \quad (29)$$

Finally, identification of (27) with (1) shows that for $i_p = y$ and $2V_{gsp}/(V_{GS} - V_T) = u$, the coefficients are $a_0 = 0$, $a_1 = 1$, $a_2 = 1/4$, and $a_3 = 0$.

Use of definitions (5)–(7) thus yields

$$IM_2 = 2HD_2 = \frac{1}{2} \frac{V_{gsp}}{V_{GS} - V_T} = \frac{i_p}{4} \quad (30)$$

and

$$IM_3 = 0. \quad (31)$$

Note that no third-order distortion occurs. Indeed, the transfer characteristic [expression (23)] is only quadratic, and hence no third-order terms can be generated. Hence, IM_3 is zero and IP_3 infinite.

Comparison of (30) with (21) shows that a MOST only generates half as much (second-order) distortion as a bipolar transistor. The main advantage of a MOST, however, is that the input voltage is scaled to $(V_{GS} - V_T)$, which can be made quite large, whereas for a bipolar transistor, the input voltage is fixed and scaled to $V_t = 26$ mV.

For example, if again a peak relative current swing is taken of 0.1 (for $I_{DS} = 1$ mA and 100 μ A peak ac current), then $IM_2 = 2.5\%$. Even more important, however, is that a peak input voltage is allowed of $V_{gsp} = 50$ mV (35 mV_{RMS}) for $V_{GS} - V_T = 1$ V, or of 10 mV (7 mV_{RMS}) only, for $V_{GS} - V_T = 0.2$ V.

The smaller the aspect ratio W/L is made, the larger the value $V_{GS} - V_T$ and the larger the peak input voltage can be allowed for the same distortion. The input voltage is indeed related to the distortion (or the relative current swing i_p) as given by

$$V_{gsp} = \frac{i_p}{2} (V_{GS} - V_T) = \frac{i_p}{2} \sqrt{\frac{I_{DS}}{\beta/2}}. \quad (32)$$

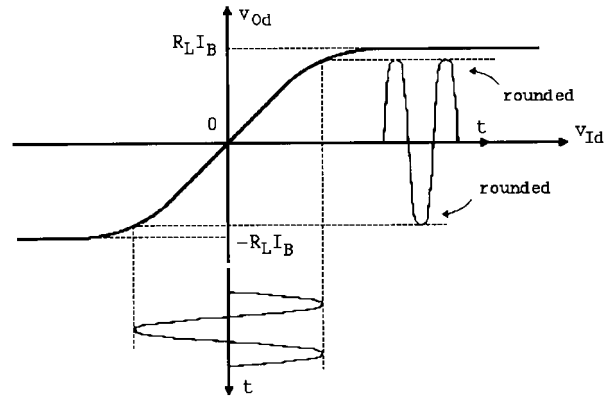


Fig. 8. Generation of nonlinear distortion (compression), caused by a symmetrical differential stage.

For a given amount of distortion (i_p) and dc current (I_{DS}), the maximum value of V_{gsp} is inversely proportional to the square root of β and hence, W/L .

Finally note that no third-order distortion is generated as long as the first-order model of a MOST is guaranteed. As soon as the complete expression is taken of MOST, including the terms with $3/2$ exponents, then third-order distortion does occur, but nevertheless in very limited amounts.

V. DISTORTION IN A BIPOLAR TRANSISTOR DIFFERENTIAL AMPLIFIER

Phase inversion of the input signal changes the sign of the fundamental and third-order components but not of the second-order component. This is exploited in a balanced or differential circuit, to which two input signals of equal amplitude but opposite phase are applied. The difference of the output signals does not contain even-order distortion at least if no unbalance is caused by mismatch. This is the case for a differential amplifier as discussed next.

As shown by the transfer characteristic (see Fig. 8), the operating point occurs now at zero output and input voltage. The transfer characteristic is indeed perfectly symmetrical with respect to the crosspoint of the axis. Application of a sinusoidal waveform in v_{Id} causes a flattening of both tops of the quasi sinusoidal waveform in v_{Od} . Compression thus occurs.

The transfer characteristic has been derived in [1, Ch. 4]. The differential output current i_{Od} is twice the ac current in each transistor. The relative current swing i is thus given by

$$i = \frac{i_{Od}}{I_B} = \tanh \frac{v_{Id}}{2V_t} \quad (33)$$

in which i_{Od} is the ac current circulating through both transistors T_1 and T_2 and v_{Id} is the differential input voltage. If two load resistors were added, then the output voltage would be $v_{OUT} = -R_L i_{Od}$.

For small input voltages ($v_{Id} < 2V_t$), the tanh function can be expanded in a power series. Indeed, for $x < 1$, we know that

$$\tanh(x) = x - \frac{x^3}{3} + \dots \quad (34)$$

Application to (33) yields

$$i_p = \frac{I_{Odp}}{I_B} = \frac{1}{2} \left[\frac{V_{Idp}}{V_t} - \frac{1}{12} \left(\frac{V_{Idp}}{V_t} \right)^3 + \dots \right] \quad (35)$$

in which i_p and V_{Idp} represent peak values of the relative current swing and the input voltage, respectively.

Truncation of this power series after its first term is sufficient an approximation for small signals. It leads to the well-known result that

$$I_{Odp} = g_{m1} \cdot V_{Idp} \quad (36)$$

in which g_{m1} is the transconductance of both transistors T_1 and T_2 , both carrying current $I_B/2$. In a first-order approximation, a simple relation is also obtained between the input voltage and the relative current swing, as given by

$$i_p = \frac{V_{Idp}}{2V_t}. \quad (37)$$

Finally, identification of (35) with (1) shows that for $i_p = y$ and $V_{Idp}/V_t = u$, the coefficients are $a_0 = 0$, $a_1 = 1$, $a_2 = 0$, and $a_3 = -1/12$.

Use of (5)–(10) and of relation (37) yields

$$IM_2 = 2HD_2 = 0 \quad (38a)$$

as expected and

$$IM_3 = 3HD_3 = \frac{1}{16} \left(\frac{V_{Idp}}{V_t} \right)^2 = \frac{i_p^2}{16}. \quad (38b)$$

Also,

$$IP_3 = 4V_t. \quad (39)$$

Coefficient a_3 is negative, hence, the distortion causes compression of the waveform.

For example, a total dc current $I_B = 1$ mA is used again. Now, however, each bipolar transistor only carries a dc current of 0.5 mA. The peak ac current in each transistor is also reduced to 50 μ A. For $R_L = 10.4$ k Ω , the voltage gain also equals -200 .

The peak relative current swing is again $i_p = 0.1$. As a result, $IM_3 = 0.062\%$. For this, a peak input voltage is obtained of 5.2 mV or 3.7 mV_{RMS}. The distortion is thus 2 times lower than in the case of a single transistor carrying a dc current I_B and providing the same gain. This factor of 2 is also found by comparison of (38) with (21).

This conclusion is especially true because no second-order distortion is present. In practice, mismatch will generate some second-order distortion as well. It is usually much smaller than the third-order distortion.

For a peak relative current swing of 0.5, $IM_3 = 1.56\%$ for which a signal amplitude of 8.4 mV_{RMS} is required. Again, a factor of 2 difference is found. It can be concluded that a differential stage can take $\sqrt{2}$ or 1.4 times more input voltage to generate the same third-order distortion as a single transistor amplifier with the same total dc current.

VI. DISTORTION IN A MOST DIFFERENTIAL AMPLIFIER

The transfer characteristic of a differential pair with MOST is very similar to the one with bipolar transistors; it is symmetrical around the origin. No second-order distortion can thus occur. Since a single MOST amplifier does not generate third-order distortion, it will be interesting to examine what distortion performance can be obtained with a MOST differential amplifier.

The transfer characteristic has been derived in [1, Ch. 4]. The differential output current i_{Od} is again twice the ac current in each transistor. The relative current swing i is thus given by

$$i = \frac{i_{Od}}{I_B} = \frac{v_{Id}}{V_{GS} - V_T} \sqrt{1 - \frac{1}{4} \left(\frac{v_{Id}}{V_{GS} - V_T} \right)^2} \quad (40)$$

in which v_{Id} is the differential input voltage. Note that $V_{GS} - V_T$ can always be substituted by $\sqrt{I_B/\beta}$.

For small values of $v_{Id} (< V_{GS} - V_T)$, the square root can be expanded as a power series. Indeed, for $x < 1$, we know that

$$\sqrt{1-x} = 1 - \frac{x}{2} - \frac{x^2}{8} - \frac{x^3}{16} - \dots \quad (41)$$

which allows us to work out (40) into

$$i_p = \frac{I_{Odp}}{I_B} = \frac{V_{Idp}}{V_{GS} - V_T} - \frac{1}{8} \left(\frac{V_{Idp}}{V_{GS} - V_T} \right)^3 - \dots \quad (42)$$

Again, i_p , i_{Odp} , and V_{Idp} all represent peak values.

For a pure small signal analysis, the power series has to be limited to the first term only, which leads to

$$I_{Odp} = \frac{I_B}{V_{GS} - V_T} \cdot V_{Idp} \quad (43)$$

or

$$I_{Odp} = \beta(V_{GS1} - V_T) \cdot V_{Idp}$$

or

$$I_{Odp} = g_{m1} \cdot V_{Idp} \quad (44)$$

in which V_{GS1} is the gate–source voltage, and g_{m1} is the transconductance of either T_1 or T_2 , which both carry currents of $I_B/2$.

Expression (42) also provides a first-order relation between the input voltage and the relative current swing

$$i_p = \frac{V_{Idp}}{V_{GS} - V_T} = \frac{V_{Idp}}{\sqrt{I_B/\beta}}. \quad (45)$$

In order to obtain the distortion components, (42) has to be identified with (1). It shows that for $i_p = y$ and $V_{Idp}/(V_{GS1} - V_T) = u$, the coefficients are $a_0 = 0$, $a_1 = 1$, $a_2 = 0$, and $a_3 = -1/8$. No second-order distortion thus occurs, as expected indeed, since no quadratic component occurs in (42). Also, coefficient a_3 is negative, which shows that compression distortion occurs, as expected as well, from a differential stage.

Use of the definitions (5)–(10) and of relation (45) yields zero for IM_2 and

$$IM_3 = 3HD_3 = \frac{3}{32} \left(\frac{V_{Idp}}{V_{GS1} - V_T} \right)^2 = \frac{3}{32} i_p^2 \quad (46a)$$

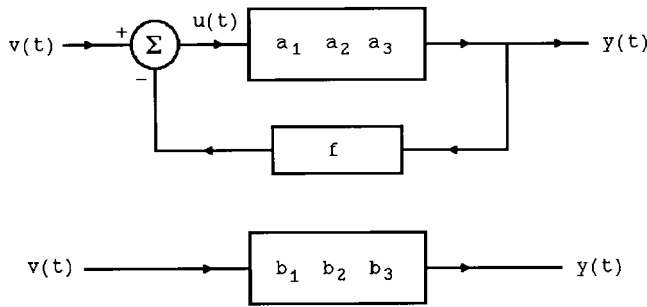


Fig. 9. Application of negative feedback f converts power series coefficients a_i into b_i .

and

$$IP_3 = 4\sqrt{\frac{2}{3}}(V_{GS1} - V_T) = 3.26(V_{GS1} - V_T). \quad (46b)$$

For example, both transistors carry a dc current of 0.5 mA and a peak ac component of 50 μ A or $I_{Odp} = 100 \mu$ A, which yields $i_p = 0.1$. As a result $IM_3 \approx 0.1\%$. This value is 1.6 times larger than the one for a bipolar differential stage with the same current swing. However, the input voltage allowed, again depends on the value of β as given by (45), which can be rewritten as

$$V_{Idp} = i_p(V_{GS} - V_T) = i_p \sqrt{\frac{I_B}{\beta}} \quad (47)$$

in which V_{GS} applies to either transistor T1 or T2. The smaller β (or W/L), the larger the input voltage allowed. For example, if $V_{GS} - V_T = 0.2$ V, then $V_{Idp} = 20$ mV or 14 mV_{RMS}. Finally, $IP_3 = 0.65$ V.

It can thus be concluded that a MOST differential stage does generate third-order distortion, because of the limiting action of its transfer characteristic. It even generates a somewhat more third-order distortion than a bipolar differential stage. The input voltage allowed is, however, much larger and can be designed to, in principle, any value, depending on the value of $V_{GS} - V_T$.

VII. THE EFFECT OF FEEDBACK ON DISTORTION

Series base and emitter resistances in the bipolar transistor linearize the exponential $I_C - V_{BE}$ relationship and thus reduce the distortion. This corresponds, however, with a reduction in gain. Also, series source resistance in the MOST reduces the distortion and the gain as well. In this section, it is examined how the application of negative feedback reduces the distortion.

A. Theory

The application of negative feedback around the nonlinear amplifier, which is characterized by coefficients a_i (see Fig. 9) gives rise to a new power series of the same form, but with coefficients b_i .

The feedback action is described by

$$u(t) = v(t) - fy(t) \quad (48)$$

in which f represents the transfer function of the unilateral feedback network. The coefficients of the new power series can be found by application of (2) on (1) and use of (48), which yields the following relations

$$b_1 = \frac{a_1}{1+T} \quad (49)$$

$$b_2 = \frac{a_2}{(1+T)^3} \quad (50)$$

$$b_3 = \frac{a_3(1+T) - 2fa_2^2}{(1+T)^5} \quad (51)$$

in which the loop gain T is given by

$$T = fa_1. \quad (52)$$

All expressions (5)–(10) are used to obtain the distortion components are still valid, provided the coefficients a_i are replaced by b_i .

The amplitude of the output signal itself is given by (49). It is reduced by a factor of $1+T$ as expected. For this reason, the input voltage V (see Fig. 9) is reduced by $1+T$ as well.

The second-order distortion is given by

$$IM_{2f} = \frac{b_2}{b_1} V = \frac{a_2}{a_1} \frac{V}{(1+T)^2} = \frac{a_2}{a_1} \frac{1}{(1+T)} \frac{V}{(1+T)}. \quad (53)$$

Also, after replacement of f by T/a_1

$$IM_{3f} = \frac{3}{4} \frac{b_3}{b_1} V^2 = \frac{3}{4} \left[\frac{a_3}{a_1} \frac{1}{(1+T)} - \left(\frac{a_2}{a_1} \right)^2 \frac{2T}{(1+T)^2} \right] \frac{V^2}{(1+T)^2}. \quad (54)$$

The first term represents third-order distortion related to a_3 , which is present as well without feedback. It is positive and thus represents expansion distortion. A sinusoidal waveform becomes more triangular.

The second term represents second-order interaction around the feedback loop, generating third-order distortion. It is negative and thus corresponds with compression.

The third-order distortion can cancel completely for specific values of a_i and T . This causes a null in the IM_3 characteristic, which is quite sharp and difficult to maintain over a wide range of transistor variables. Therefore, it is never a parameter to design for. Moreover, it occurs at very small values of loop gain T .

For high values of T , the second term usually dominates and compression distortion results. For small values of T or of a_2 , expansion distortion is dominant. These effects are now illustrated with several examples.

B. Emitter Resistance in Single Bipolar Transistor Amplifier

Insertion of an emitter resistance R_E in a single transistor amplifier provides local feedback. The loop gain T is given by

$$T = g_m R_E. \quad (55)$$

The second-order distortion component is then obtained from (53) and given by ($a_2/a_1 = 1/2$ for a bipolar transistor)

$$IM_{2f} = \frac{1}{2} \frac{1}{(1+T)} \frac{V_{ip}}{(1+T)V_t} = \frac{V_{ip}}{2V_t} \frac{1}{(1+T)^2} \quad (56)$$

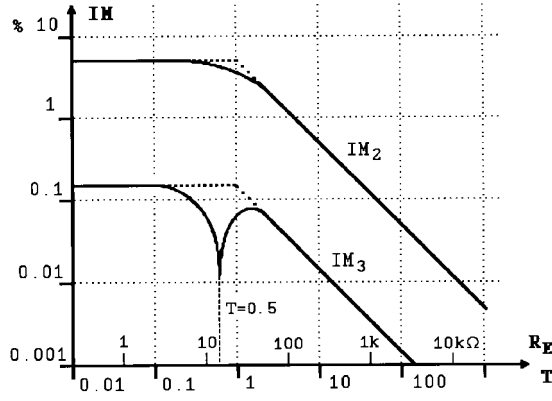


Fig. 10. Distortion components with feedback in a bipolar transistor with 1 mA collector current.

in which V_{ip} is the peak input voltage with respect to ground.

Also since

$$i_p = \frac{V_{ip}}{(1+T)V_t} \quad (57)$$

the result is

$$IM_{2f} = \frac{1}{(1+T)} \frac{i_p}{2}. \quad (58)$$

The third-order distortion is derived from (54) and given by ($a_2/a_1 = 1/2$; $a_3/a_1 = 1/6$)

$$IM_{3f} = \frac{1}{8} \frac{1-2T}{(1+T)^2} \left(\frac{V_{ip}}{(1+T)V_t} \right)^2 = \frac{1}{8} \frac{1-2T}{(1+T)^2} i_p^2. \quad (59)$$

For example, a bipolar transistor carries a dc current of 1 mA and an ac peak current of 100 μ A. The peak relative current swing is thus 0.1. Without feedback $IM_2 = 5\%$, $IM_3 = 0.125\%$, and $V_{be} = 1.84$ mV_{RMS}.

Addition of a resistance of 260 Ω causes a dc voltage across R_E of $V_E = 260$ mV which results in $T = V_E/V_t = g_m R_E = 10$. Note that the value of T is easily found by taking the dc voltage across R_E , divided by V_t .

The value of $IM_{2f} = IM_2/(1+T) = 0.45\%$. Moreover, the input voltage allowed increases to 20.2 mV_{RMS}. The value of IM_{3f} is then $0.16 \times IM_3$, or 0.02%. In order to increase IM_{3f} to the same value as without feedback, the value of i_p has to be increased by $\sqrt{6.25}$ or 2.5, yielding $i_p = 0.25$ and $V_i = 126$ mV_{RMS}.

The distortion components with feedback are plotted versus T (and R_E) in Fig. 10 for constant values of the collector current (=1 mA) and relative current swing i_p (=0.1). For low feedback ($T < 0.1$), the values are the same as on Fig. 5 for $i_p = 0.1$. For large feedback ($T > 10$), the values decrease with a slope of unity. Note that the null in IM_{3f} indeed occurs at $T = 0.5$ or $R_E = 1/2g_m = 13 \Omega$, which corresponds with a very small amount of feedback indeed.

For high values of feedback ($T > 10$ and $1+T \approx T$), expression (57) of the relative current swing can be modified into

$$i_{pT} \approx \frac{1}{g_m R_E} \frac{V_{ip}}{V_t} = \frac{V_{ip}}{I_C R_E} \quad (60)$$

which shows that the input voltage is to be compared with the voltage drop across the feedback resistance, in order to obtain i_p . For instance, for a voltage drop across R_E of 1 V ($R_E = 1$ k Ω with $I_C = 1$ mA), then $V_{ip} = 0.1$ V or 0.07 V_{RMS} gives $i_{pT} = 0.1$.

For such high values of feedback ($T > 10$), the distortion components can be simplified to

$$IM_{2fT} \approx \frac{1}{T} \frac{i_{pT}}{2} = \frac{V_t}{2} \frac{V_{ip}}{(I_C R_E)^2} \quad (61)$$

$$IM_{3fT} \approx -\frac{2}{T} \frac{i_{pT}^2}{8} = -\frac{V_t}{4} \frac{V_{ip}^2}{(I_C R_E)^3}. \quad (62)$$

Comparison with (21) shows that for IM_2 it is sufficient to divide by T , whereas IM_3 has to be divided by $T/2$. It can thus be concluded that feedback reduces distortion components indeed. All of them are reduced, however, by about the same amount.

Finally, note that emitter resistances can never fully be excluded in a bipolar transistor since the base resistance linearizes the exponential as well. The equivalent emitter resistance is then $R_{Ee} = r_B/(1+\beta)$, which is usually of the order of a few ohms.

C. Source Resistance in Single MOST Amplifier

The insertion of a source resistor provides local feedback. The value of the loop gain is again given by (55) with an emitter resistor instead of a source resistor. From (27), we find $a_2/a_1 = 1/4$ and $a_3 = 0$. As a result, (53) and (54) become

$$IM_{2f} = \frac{1}{4} \frac{2V_{ip}}{(1+T)^2(V_{GS} - V_T)} = \frac{1}{1+T} \frac{i_p}{4} \quad (63)$$

$$IM_{3f} = -\frac{3}{32} \frac{T}{(1+T)^2} \left(\frac{2V_{ip}}{(1+T)(V_{GS} - V_T)} \right)^2 = -\frac{3}{32} \frac{T}{(1+T)^2} i_p^2 \quad (64)$$

since now

$$i_p = \frac{2V_{ip}}{(1+T)(V_{GS} - V_T)}. \quad (65)$$

For the same current swing, the second-order distortion is reduced by $1+T$. Now, however, third-order distortion emerges as well. It is caused by the presence of a_2 in (54), which represents the increase in order of the second-order distortion component which is fed back to the input. It is still smaller than for a bipolar transistor.

For large feedback $T > 10$, the current swing becomes

$$i_{pT} \approx \frac{V_{ip}}{R_{S} I_{DS}} \quad (66)$$

and

$$IM_{3fT} = -\frac{3}{32} \frac{1}{T} i_p^2 \quad (67)$$

which leads to the same conclusion as for a bipolar transistor.

D. Emitter Resistances in a Bipolar Differential Stage

In a differential pair, second-order distortion is absent ($a_2 = 0$). Addition of equal emitter resistances in both transistors does not degrade this symmetry. The third-order distortion is derived from (54) and is given by (for $a_3/a_1 = -1/12$)

$$\text{IM}_{3f} = -\frac{1}{16} \frac{1}{1+T} \left(\frac{V_{Idp}}{(1+T)V_t} \right)^2 = -\frac{1}{16} \frac{1}{1+T} i_p^2. \quad (68)$$

The same conclusions can thus be drawn. IM_{3f} is negative, which corresponds with compression distortion, as before. For large feedback ($T > 10$), the value of IM_{3f} decreases linearly with T and is then given by the distortion without feedback, divided by $(1+T)$.

E. Source Resistances in a MOST Differential Stage

Again, symmetry is maintained, and hence no second-order distortion occurs. From (42) we find that $a_3/a_1 = -1/8$. The third-order distortion is again derived from (54) and is given by

$$\begin{aligned} \text{IM}_{3f} &= -\frac{3}{32} \frac{1}{(1+T)} \left(\frac{V_{Idp}}{(1+T)(V_{GS} - V_T)} \right)^2 \\ &= -\frac{3}{32} \frac{1}{(1+T)} i_p^2. \end{aligned} \quad (69)$$

The same conclusion can be drawn as for a differential stage with bipolar transistors.

F. Emitter Follower

For distortion analysis, the emitter follower can be regarded as a single transistor amplifier with large feedback ($T > 10$). The output is taken at the emitter instead of at the collector; but since the relative current swing is taken as a fundamental parameter, the analysis is the same. For an emitter follower with an emitter resistance, the distortion components are thus already given by (61) and (62).

However, if a transistor is used instead of a resistance, then its output resistance r_o has to be used in the expression instead of R_E . Since $r_o = I_C/V_E$, in which V_E is the early voltage, the relative current swing i_{pT} can be derived from (60) and is given by

$$i_{pT} = \frac{1}{g_m r_o} \frac{V_{ip}}{V_t} = \frac{V_{ip}}{V_E}. \quad (70)$$

In order to obtain i_{pT} , the input voltage thus simply has to be compared with the early voltage. For instance, for $V_E = 50$ V ($I_C = 1$ mA), an input voltage of $V_{be} = 0.1$ V (or 0.07 V_{RMS}) only provides $i_{pT} = 2 \cdot 10^{-3}$.

The distortion components are then given by (61) and (62) which give ($T = g_m r_o = V_E/V_t = 2000$) $\text{IM}_{2fT} = 5 \cdot 10^{-5}\%$ and $\text{IM}_{3fT} = 5 \cdot 10^{-8}\%$. They are thus negligible, thanks to both the low values of i_{pT} and the high value of T . For an ideal follower, the current source is ideal, and its current is not modified by application of an input signal. Hence, the current swing is zero and so is the distortion (see Fig. 11).

The distortion of a source follower can also be calculated directly as a solution of a nonlinear equation.

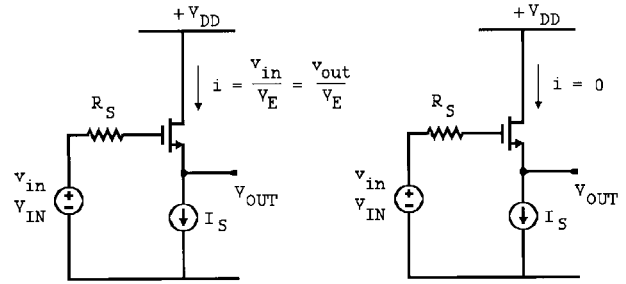


Fig. 11. The current swing in an ideal source follower is zero, and so is the distortion.

G. Source Follower

Very much the same conclusions apply to the source follower as to the emitter follower. The relative current swing is again given by

$$i_{pT} = \frac{1}{g_m r_{DS}} \frac{2V_{ip}}{V_{GS} - V_T} = \frac{V_{ip}}{V_E} \quad (71)$$

and has to be used in (63) and (67).

As an example, a source follower is taken at $I_{DS} = 1$ mA with a current source with output resistance 16 k Ω ($V_E = 16$ V). An input voltage of 4 V (or 2.8 V_{RMS}) now gives $i_{pT} = 0.25$. Now the aspect ratio is such that $V_{GS} - V_T = 1$ V, and $T = g_m r_o = 2V_E/(V_{GS} - V_T) = 32$ V. Thus, $\text{IM}_{2fT} = 0.19\%$ and $\text{IM}_{3fT} = 0.053\%$.

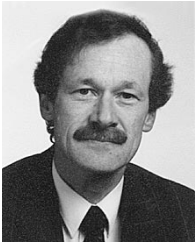
Obviously for an ideal current source, the relative current swing is zero and so is the distortion (see Fig. 11). In this consideration, the bulk is assumed to be connected to the source. If this is not the case, the parasitic JFET or the body effect has to be considered as well. In this case, the distortion is mainly caused by this effect.

To find the sources of distortion in any arbitrary circuit, the values of the relative current swing have to be found together with the feedback factor T . All distortion components are readily calculated.

In addition, the amplitude of the transfer characteristic versus frequency has to be calculated of each transistor output to the output of the total circuit. Higher harmonics are usually attenuated by the low-pass filter action of the capacitances present.

REFERENCES

- [1] K. Laker and W. Sansen, *Design of Analog Integrated Circuits and Systems*. New York: McGraw-Hill, 1994.
- [2] W. Sansen and R. Meyer, "Distortion in bipolar transistor variable-gain amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 275–282, Aug. 1973.
- [3] J. Silva-Martinez, M. Steyaert, and W. Sansen, *High-Performance CMOS Continuous-Time Filters*. Norwell, MA: Kluwer Academic, 1993.
- [4] S. Willingham and K. Martin, *Integrated Video-Frequency Continuous-Time Filters*. Norwell, MA: Kluwer, 1995.
- [5] D. Pederson and K. Mayaram, *Analog Integrated Circuits for Communications*. Norwell, MA: Kluwer, 1991.
- [6] P. Wambacq and W. Sansen, *Distortion Analysis of Analog Integrated Circuits*. Norwell, MA: Kluwer, 1998.



Willy Sansen (S'66–M'72–SM'86–F'95) received the M.Sc. degree in electrical engineering from the Katholieke Universiteit Leuven in 1967 and the Ph.D. degree in electronics from the University of California at Berkeley in 1972.

Since 1981, he has been a Full Professor at the ESAT Laboratory of the Katholieke Universiteit Leuven. He was a Visiting Professor at the Universities of Stanford (1977), Lausanne (1981), Philadelphia (1985), and Ulm (1994). He has been involved in design automation and in numerous analogue integrated circuit designs for telecom, consumer, biomedical applications and sensors. He has been supervisor of 340 papers in international journals and conference proceedings and six books, among which the textbook with K. Laker, *Design of Analog Integrated Circuits and Systems* (McGraw-Hill, 1994).