More on Nonlinear Distortion Correction*

J. ROSS MACDONALD†, FELLOW, IRE

Summary—Further consideration is given to basic amplitude limitations which may apply to the complementary distortion method of nonlinear distortion correction. It is found, in disagreement with others, that points at which the differential gain is zero or infinite do not limit the amplitude over which complete correction is possible but that relative maxima, minima, gain zeros, and infinite-gain points in the characteristic do set limitations when the usual simply connected tandem configuration is employed. When the characteristic to be corrected is multiple valued or passes through points of zero or infinite gain within a given amplitude range, a multiply connected correction circuit must be used for perfect correction of distortion over the amplitude range in question.

THERE has recently been a certain amount of controversy concerning amplitude limitations in the complementary distortion method of nonlinear distortion reduction. Two such limitations, which will be further discussed herein, were pointed out in the original paper[1] and Waldhauer[2] later suggested a specific configuration for complete distortion correction which is stated to be limited to the amplitude range over which $|de_1/de_0|>0$, where $e_0=A\cos\omega t$ is an input signal and $e_1\equiv f(e_0)$ is the output signal obtained when e_0 is applied to the input of a predistortion network which is to correct the distortion of a given black box whose input is e_1 and whose output, as shown in Fig. 1(a), is $e_2\equiv g(e_1)$. Perfect distortion correction only occurs when $e_2=Ke_0=g\{f(e_0)\}$, where K is the over-all amplification factor. As pointed out by Pritchard,[3] perfect correction is only achieved in Waldhauer's configuration provided the two amplifiers he uses are assumed to have zero input and infinite output impedance respectively. These conditions, which cannot be met in practice over a nonzero amplitude range, can still be well approximated over a limited range. Within this range, the important advantage of Waldhauer's approach is that the same elements and circuit configuration appearing in the black box whose characteristic is to be corrected appear also in the pre- or postdistortion correcting network.

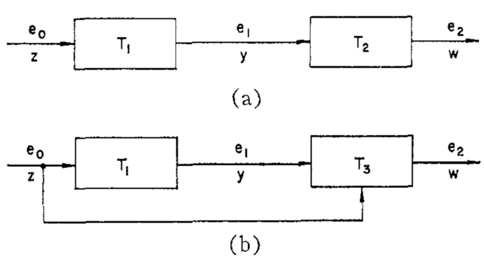

Fig. 1—(a) Usual configuration for correcting nonlinear distortion by complementary distortion. z, y, and w are normalized signal variables. (b) A multiply connected configuration for correcting nonlinear distortion generated in the left-hand circuit.

In the rest of this paper, we shall be concerned with amplitude restrictions for complete distortion correction. In the limit of complete correction, the distinction between pre- and postdistortion vanishes.[2,3] Therefore, we shall consider two nonlinear black boxes connected in tandem as shown in Fig. 1(a) and shall make no distinction between which represents the correcting circuit and which the circuit to be corrected. No significant generality will be lost if we take $K=1$, making the final output equal to the input when complete correction is achieved. In the simplest case, the transfer functions $e_1/e_0\equiv T_1=f(e_0)/e_0$ and $e_2/e_1\equiv T_2=g(e_1)/e_1$ may be considered as real, single-valued operators which operate on a single input to give a single output. Then, the condition

$$T_1T_2 = I \tag{1}$$

where I is the identity operator, leads to complete correction. In this case, the boxes may clearly be interchanged and $T_2T_1=I$ as well. Thus, the operators commute, a result which may also readily be established formally.

In the latter part of the Appendix of the author's paper,[1] a method of complete distortion correction was described which depends on the condition $T_1T_2=I$. This method was later generalized by Waldhauer[2] and has been recently mentioned again by Holbrook and Todosiev.[4] Further discussion of any amplitude restriction applying to this method is needed since the conclusions in the author's paper,[1] those of Waldhauer, and those of Holbrook and Todosiev are inconsistent with each other in some cases. Note that this method is more general than, and is distinct from, Waldhauer's approximate configuration for obtaining response inversion.

In order to discuss amplitude limitations, it will be convenient to consider various classes of transfer functions for the left black box and to ask over what input-signal amplitude range the nonlinearity introduced by

* Received by the PGA, March 16, 1961; revised manuscript received, April 3, 1961.

† Texas Instruments Incorporated, Dallas, Texas.

1 J. R. Macdonald, "Nonlinear distortion reduction by complementary distortion," IRE TRANS. ON AUDIO, vol. AU-7, pp. 128–133; September–October, 1959.

2 F. D. Waldhauer, "Comments on 'nonlinear distortion reduction by complementary distortion,' " IRE TRANS. ON AUDIO (Correspondence), vol. AU-8, p. 103; May–June, 1960.

3 J. R. Macdonald, "Reply to comments on 'nonlinear distortion reduction by complementary distortion,' " IRE TRANS. ON AUDIO (Correspondence), vol. AU--8, pp. 104–105; May–June, 1960.

4 G. W. Holbrook and E. P. Todosiev, "Amplitude limitations in nonlinear distortion correction," IRE TRANS. ON AUDIO (Correspondence), vol. AU-8, p. 235; November–December, 1960.

the left box of Fig. 1(a) can be completely corrected by the right box. Let $e_1 = a_1e_0 + a_2e_0^2 + a_3e_0^3$, a sufficiently general expression for illustrative purposes. For simplicity, we shall restrict attention to the upper right-hand e_1–e_0 quadrant; thus e_1 and e_0 will both be positive. The behavior in the other quadrants can be easily specified in practical cases. The above expression for e_1 is, in general, asymmetric. If a nonlinear push-pull characteristic is to be represented, e_1 must be made an odd or antisymmetric function of e_0, while a symmetric or even dependence of e_1 on e_0 would give a kind of rectifier characteristic. For a nonlinear amplifier, the over-all response characteristic will fall entirely in the first and third quadrants in cases of practical importance. If the curve is antisymmetric in e_0, the following treatment for a given curve in quadrant 1 can be applied without a significant change to quadrant 3. If the curve is asymmetric, however, the amplitude limitations (if any) which follow from the first and third quadrant responses may be different. Since the input signal is assumed sinusoidal, and hence changes sign, the largest input amplitude which can still be used with complete distortion correction will be determined by the smaller of the two amplitude limitations provided the input ac signal is zero-biased.

It is desirable to express the e_1 vs e_0 characteristic in terms of normalized variables. First, take $a_1 = 1$, consistent with the choice $K = 1$ for the over-all gain. Then, let $a_2 = \epsilon a$; $\epsilon = \operatorname{sgn} a_2$, the sign of a_2. Define $z \equiv ae_0$, $y \equiv ae_1$, $w \equiv ae_2$, and let $a_3/a^2 \equiv \delta$. Then the cubic characteristic becomes

$$y = z + \epsilon z^2 + \delta z^3. \tag{2}$$

A number of cases of this characteristic are plotted in Fig. 2 and will be discussed in terms of amplitude limitations.

In Fig. 2(a), δ has been set to zero and ϵ taken as $+1$. The resulting square-law-distortion characteristic will be completely linearized if the normalized output, $w(y)$ of the second black box in Fig. 1(a) is identically equal to z, the normalized input to the system. When this is the case, the right box has an inverse or complementary characteristic to the left one and the solution of the quadratic yields the single-valued response

$$w = z = \tfrac{1}{2}[\sqrt{1 + 4y} - 1]. \tag{3}$$

Substitution of $y = z + z^2$ yields an identity as it should. Here there is no ideal, or mathematical, input amplitude limitation when the right side of (3) is synthesized exactly. In practice, the inversion of $y = z + z^2$ cannot be carried out exactly, and Holbrook and Todosiev's[4] reference to *merely* obtaining the inverse is an undue simplification. Waldhauer's[2] specific configuration for distortion correction mentioned earlier represents a useful method of achieving approximate inversion, but it will only be approximate in any real circuit. Thus, in

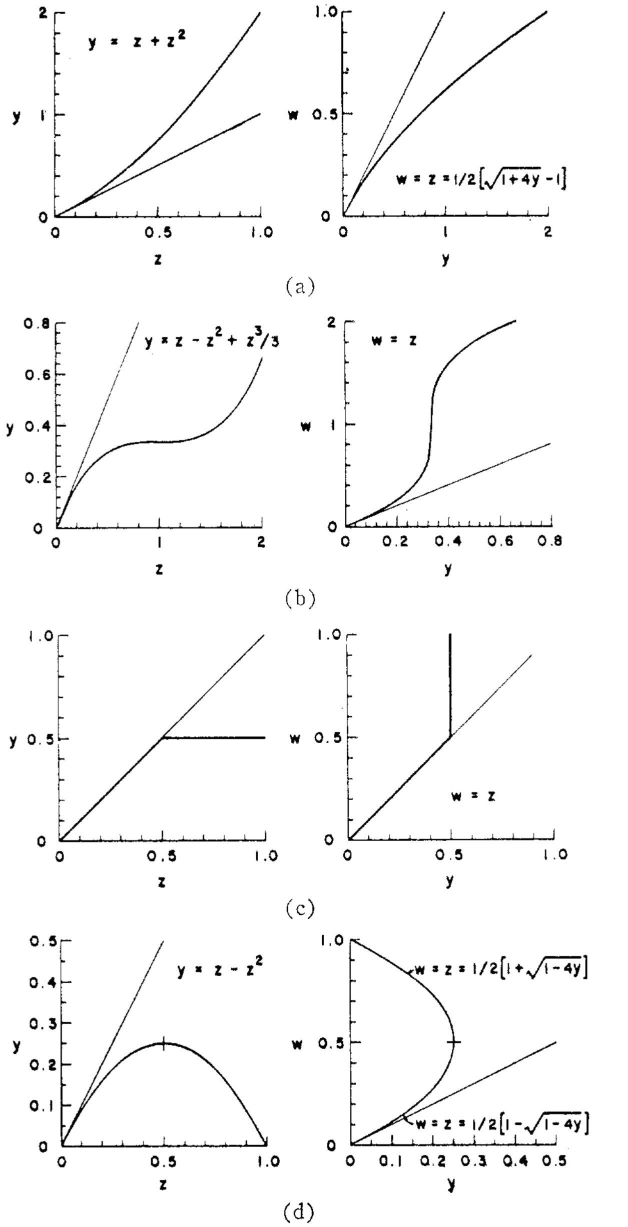

Fig. 2—Various illustrative nonlinear response functions. For each part, the right-hand curve shows the response complementary to that of the left-hand curve.

the present case, there are no mathematical amplitude limitations but there will usually be practical physical ones.[3,4]

In the treatment of the present case in the latter part of the Appendix of reference 1 it was stated that complete correction could be obtained if the characteristic (3) were realized using an analog computer, a general term for passive and active components. Since the Waldhauer method of approximate inversion requires an additional circuit having exactly the same nonlinear response as that to be corrected, it will not be appropriate in all cases.[3] Usually, given a nonlinear characteristic in terms of a e_1–e_0 response curve or its power-law representation, one must synthesize its inverse, such as that in (3), using passive and possibly active components. Such synthesis will often be difficult, but there are no intrinsic physical or mathematical prohibitions to the synthesis of a characteristic such as that in (3). Thus, while power and voltage ratings are always limited in practice, basic laws such as the second law of thermodynamics are not contravened by the synthesis.

Fig. 2(b) shows curves a degree more complicated. Here $\epsilon=-1$ and $\delta=1/3$. Both the direct and inverse characteristics are shown, and it will be noted that the form of the direct characteristic has been selected to give an inflection point in the $y-z$ curve at $z=1$ where the differential gain, dy/dz, is zero. The exact inverse characteristic correspondingly shows $dw/dy=dz/dy=\infty$ at this one point. This result does not require that the gain e_2/e_1 of the right black box be infinite at this point, as Waldhauer[2] has stated, but only that the differential gain be infinite at one point, a condition which can be achieved in a practical circuit using active elements. Here again, the inverse characteristic is still single valued, the conditions $d^2y/dz^2=dy/dz=0$ do not define an amplitude limitation, and there are only practical obstacles to the realization of complete correction over an arbitrary amplitude range.

In Fig. 2(c), a characteristic showing complete saturation for $z\geq 1/2$ is depicted. This nonlinearity can only be corrected for $z>1/2$ with infinite direct gain, not differential gain, in the complementary box. There is here a definite mathematical and practical amplitude limitation. Note that complete correction could be achieved, however, with the circuit of Fig. 1(b). Here the transfer operator T_3 is a function of two separate inputs, one of which is the original input e_0. The complete system is multiply connected, not simply connected as it is in the configuration of Fig. 1(a), the only situation originally considered in complementary distortion correction.[1] In many cases of practical interest, the original signal e_0 is unavailable at the second black box, and sequential or tandem correction such as that shown in Fig. 1(a) is the best that is possible.

Finally, Fig. 2(d) presents a case where the complementary or inverse characteristic is multiple valued. Here $\delta=0$ and $\epsilon=-1$. As shown, the inverse characteristic is single valued up to $y=1/4$, a point where again the differential gain de_2/de_1 is infinite. Here the point $(z, y)=(1/2, 1/4)$ is a relative maximum in y rather than an inflection point, and $d^2y/dz^2\neq 0$. In the region $0\leq w=z\leq 1/2$, the inverse characteristic is single valued and given by

$$w = z = \tfrac{1}{2}[1-\sqrt{1-4y}], \tag{4}$$

a response which can be realized approximately by Waldhauer's inversion technique or by other methods. For $1/2\leq z\leq 1$, the necessary inverse characteristic for complete correction is

$$w = z = \tfrac{1}{2}[1+\sqrt{1-4y}]. \tag{5}$$

In order to achieve perfect correction over the entire range $0\leq z\leq 1$, the right black box must automatically switch its characteristic from (4) to (5) when z passes the branch point 1/2. Such a logical decision requires knowledge of the original variation of z at the input and can only be made with a circuit such as that of Fig. 1(b). In addition, it is evident that the gain e_2/e_1 is infinite when $z=1$; thus, realization of the characteristic (5) itself up to or beyond $z=1$ can only be achieved if a portion of the original input is available at the right-hand box as in Fig. 1(b). We may conclude from Fig. 2(c) and (d) that whenever a relative maximum or minimum occurs in the response function $e_1(e_0)$ or $y(z)$, as in the case of complete saturation or multiple-valued characteristics, perfect correction using the configuration of Fig. 1(a) cannot be achieved for input amplitudes greater than that which yields the first maximum or minimum. In the present case, the amplitude is thus limited to $z\leq 1/2$, equivalent to the condition $x\leq 1/2$ given in the Appendix of reference 1. This is a mathematical and physical or configurational limitation and disagrees with the conclusions of Holbrook and Todosiev.[4] It should also be mentioned that in the Appendix[1] another amplitude limitation, $x<0.207$, was given which applies to the square-law distortion case. This is a purely mathematical limitation on the inversion-of-series method[1] of determining the complementary or inverse characteristic. This method yields an infinite number of correction terms which, practically, must be realized with a finite number of correcting elements. When $x>0.207$, the power series in question does not converge and complete correction of square-law distortion by this method will be impossible.

In summary, we may conclude that, aside from practical considerations, complete correction of a given nonlinear input-output characteristic is possible using a tandem arrangement for an unlimited input amplitude range provided the initial characteristic always either monotonically increases or decreases and no zero or infinite gain points are reached. On the other hand, whenever the characteristic is multiple valued, a simply-connected tandem correcting circuit can yield complete correction only over an amplitude range extending up to the first relative maximum or minimum or zero or infinite gain point of the characteristic. In the multiple-valued case, a multiply connected correction circuit must be used to achieve complete correction. The number of logical decisions (or separate signal paths) which must be made in such a circuit is equal to the number of relative maxima, minima, gain zeros, and infinite-gain points which occur within the amplitude range over which complete correction is desired.

Acknowledgment

The author is grateful to J. P. Pritchard, Jr., for helpful discussions and criticisms.