APPLICATION NOTE 15

DISTORTION AND INTERMODULATION

Barney Oliver

In a completely linear system each input produces a proportional response independently of all other inputs. That is if $f_1(t)$ produces the response $g_1(t)$ and if $f_2(t)$ produces the response $g_2(t)$, then the input $a\,f_1(t) + b\,f_2(t)$ will produce the response $a g_1(t) + b\,g_2(t)$, when a and b are real constants. In fact, this statement may be said to define a linear system. It can be shown that any linear system so defined which is also stable and invarient with time will have the property that the spectrum of the output, $G(\omega)$, is the spectrum of the input, $F(\omega)$, multiplied by a transmission function $K(\omega) \neq \infty$. As a result:

1. All frequencies in the output will be proportional to those in the input- $K(\omega)$ being the weighting factor, independent of level.

2. No frequencies not in the input will appear in the output.

These two facts form the basis of all measurements of system non-linearity- that is all measurements of non-linearity measure the amount by which a given system fails to obey one or the other of these properties. Methods which make use of property 1 include

1. Direct measurements of transfer characteristic.

2. Measurements of saturation characteristic.

Methods which utilize property 2 include

3. Harmonic distortion measurements

4. Intermodulation measurements

5. Shift of operating point (by rectification of signal)

6. Measurement of slope of transfer characteristic by small ac superposed on dc.

All of these methods are useful - though not necessarily equally so in a given situation.

Method 1 is especially suited to devices which pass dc - for it is then simple to carry out and often gives exactly the information desired. For example transfer characteristic linearity is a convenient test for a dc oscilloscope - or a dc meter. Harmonic distortion or intermodulation methods would be hard to apply with the spot or needle deflection as the output and what would they tell?

Method 2 is useful with many ac devices, e.g. in RF and IF amplifier measurements and is indeed sometimes the only practical method with the tools at hand (TWT saturation). The method requires at most a signal source, an input attenuator, an output attenuator, and a detector (which need not be linear). So long as the reading stays constant as we transfer db from input to output the device between is linear. AGC systems and the like must be disabled, of course, to make the device invarient.

Methods 3 and 4 are the most commonly used in audio work where it is in fact the spurious frequencies produced by non-linearity which disturb the spectrum analyzers we wear in our heads.

Method 5 is sometimes used as an index of distortion in single ended amplifiers, or of overmodulation in transmitters. It has the advantages of extreme simplicity, often requiring no additional equipment, and providing a constant operating check. It is in general responsive only to even order distortion.

Method 6 is really a special case of 4 (as in method 5, for that matter) where one signal is a dc or a wave with constant plateaux, and the modulation of the ac by this "dc" is measured. It is used in television, for example, to measure the transfer characteristics of video equipment under actual operating conditions.

In general, non-linearity arises from a transfer characteristic which is not a straight line. Thus, if Method 1 could always be used it should be possible to correlate the results of any other method to those of Method 1. Now things correlated to the same thing are not necessarily correlated with each other so it does not follow that the results of the other methods should show correlation. However such correlation does often exist and may be quite strong. In the following sections we will discuss the correlation which exist between harmonic distortion and intermodulation measurements.

In harmonic distortion and intermodulation tests we are primarily concerned with systems that should be linear, but aren't quite - quasi-linear systems, if you like, as distinguished from highly non-linear systems such as harmonic generators. Assuming the non-linearity to be gentle, then, it is appropriate to express the transfer characteristic which produces the non-linearity as a power series expansion about the operating point. Thus let

$e_{out} = k_1 e_{in} + k_2 e_{in}^2 + k_3 e_{in}^3 +$ higher order terms and let's assume the input consists of two sinusoids

$$e_{in} = e_1 \sin \omega_1 t + e_2 \sin \omega_2 t.$$

Then the output of the non-linear stage will be (hold your hats!)

$$e_{out} = \frac{k_2}{2}\left(e_1^2 + e_2^2\right) \qquad \text{(dc)}$$

$$+ \left(k_1 e_1 + \frac{3k_3}{2} e_1 e_2^2 + \frac{3k_3}{4} e_1^3\right) \sin \omega_1 t$$
$$+ \left(k_1 e_2 + \frac{3k_3}{2} e_2 e_1^2 + \frac{3k_3}{4} e_2^3\right) \sin \omega_2 t \qquad \text{(fundamentals)}$$

$$- \frac{k_2}{2} e_1^2 \cos 2\omega_1 t - \frac{k_2}{2} e_2^2 \cos 2\omega_2 t \qquad \text{(2nd harmonics)}$$

$$- \frac{k_3}{4} e_1^3 \sin 3\omega_1 t - \frac{k_3}{4} e_2^3 \sin 3\omega_2 t \qquad \text{(3rd harmonics)}$$

$$+ k_2 e_1 e_2 \left[\cos(\omega_1 - \omega_2)t - \cos(\omega_1 + \omega_2)t\right]$$
$$+ \frac{3k_3}{4} e_1^2 e_2 \left[\sin(2\omega_1 - \omega_2)t - \sin(2\omega_1 + \omega_2)t\right]$$
$$+ \frac{3k_3}{4} e_1 e_2^2 \left[\sin(2\omega_2 - \omega_1)t - \sin(2\omega_2 + \omega_1)t\right]$$

Sum and difference products.

$$+ \text{ higher order terms.}$$

The purpose of all this, dear reader, is not to frighten you but to illustrate certain general rules of harmonic and intermodulation product production. These are

1. Only frequencies of the form $a\,\omega_1 + b\,\omega_2$ are generated, where a and b are integers.

2. If the highest power term in the power series expansion is of exponent n, then no frequencies for which $|a| + |b| > n$ will be generated.

3. If all k's up to k_n are not zero, all frequencies for which $|a| + |b| \leq n$ will be present.

4. The parity of $|a| + |b|$ is that of the power term which produces that component. (For even powers $|a| + |b|$ is even, for odd powers $|a| + |b|$ is odd.)

5. All terms produced by the m^{th} power term vary as $e_1^c e_2^d$ where $c + d = m$.

6. Corresponding sum and difference frequencies are generated at equal amplitude.

These properties follow from the laws of expansion of powers of sines and cosines and of their products, and from the fact that each power term generates only trignometric product terms of that total degree.

Property 6 illustrates the fact that we have in a non-linear transfer characteristic a device which produces pure AM, no FM. The corresponding sum and difference products are the "sidebands" produced by the modulation of one frequency by another. For example the $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$ terms result from the modulation of ω_1 by ω_2 or ω_2 by ω_1, or to say it symmetrically: The inter-modulation of ω_1 and ω_2. The dc term $\frac{k_2}{2} e_1^2$ and the term $-\frac{k_2}{2} e_1^2 \cos 2\omega_1 t$ represent the modulation of ω_1 by ω_1, that is the self-modulation of ω_1. With self modulation the "sum" and "difference" frequencies are separated by multiples of the carrier and thus lie in harmonic relation to each other.

Where the distortion is small so thatterms of higher degree than cubic can be neglected, then a simple rule gives the way in which the amplitude of the distortion coefficients vary with level. A frequency of the form $a\omega_1 + b\omega_2$ where $|a| + |b| \geq 2$ varies as $e_1^{|a|} e_2^{|b|}$. Thus the frequency $2\omega_1 - \omega_2$ has an amplitude proportional to $e_1^2 e_2$; the frequency $3\omega_2$ has an amplitude proportional to e_2^3 etc.

With two oscillators, a mixing means and a wave analyzer one can measure the individual distortion products of a device and get a complete picture of the distortion. Such a process is time consuming and short-cut methods are commonly employed. These will now be considered.

Harmonic Distortion - Fundamental Rejection Method

In this method a single sine wave input is used and the output voltage is read first directly and then through a filter which rejects the fundamental. If an RMS meter is used, the ratio of these readings will be

$$R = \frac{\sqrt{\sum(\text{harmonics})^2}}{\sqrt{(\text{fundamental})^2 + \sum(\text{harmonics})^2}}$$

The total harmonic distortion, definded as the total RMS of all harmonics divided by the fundamental is thus

$$(HD) = \frac{R}{\sqrt{1-R^2}} \approx R, \text{ for } R \ll 1.$$

When R< .1 the error is less than 1/2% of (HD).

More serious is the error which results from using an average reading meter. Measurements indicate that readings of R made with an average meter may be as much as 20 to 30% low. Usually the error will be less than this, but even this amount is not serious in most distortion measurements.

If the residual wave after fundamental rejection is applied to the vertical plates of a 'scope deflected horizontally by the fundamental (input), a stationary pattern is obtained. This pattern often gives a lot more direct information about the cause of the distortion than mere measurement of the amount of distortion by any method. Grid current being drawn produces sharp peaks in the wave (one if single ended, two if push-pull); parasitic oscillations show up as r-f bursts on the wave, or if too high frequency to be passed, as sharp breaks. If the harmonics are passed with equal transmission amplitude and delay and if the horizontal deflection phase is correct a single trace (rather than a loop) pattern will result. This trace is a presentation of the actual departure of the transfer characteristic from a straight line, and quickly reveals defects such as improper bias etc. This visual presentation is one of the chief virtues of the fundamental suppression method.

Intermodulation - CCIF Method

In the CCIF Method of measuring intermodulation distortion, two high frequencies of equal amplitude are applied to the system and the lowest frequency difference product is extracted with a low pass filter. The distortion is defined as the ratio of the amplitude of this product to the sum of the two fundamentals is the output. From our analysis we see that this ratio will be

$$(IM)_{CCIF} = \frac{k_2 e_1 e_2 + \frac{3}{2} k_4 (e_1^3 e_2 + e_1 e_2^3) + \cdots}{k_1(e_1 + e_2) + \frac{3}{2} k_3 (e_1 e_2^2 + e_1^2 e_2) + \frac{3k_3}{4} (e_1^3 + e_2^3)}$$

Or since $e_1 = e_2 = \frac{e}{2}$ where e is the total peak drive:

$$(IM)_{CCIF} = \frac{e}{4} \frac{k_2 + \frac{3}{8} k_4 e^2 + \cdots}{k_1 + \frac{9}{16} k_3 e^3} \cong e \frac{k_2}{4k_1}$$

One glaring drawback of the CCIF method is that it measures only distortion of even order (k_2 , k_4 etc), and thus ignores the odd order distortion, usually the principal distortion in push-pull systems. Difference frequency intermodulation tests are perfectly valid tests, but if the distortion is likely to be odd-order, odd-order difference frequencies must be measured.

Intermodulation - SMPTE Method

In this method a low frequency and high frequency are used as inputs, the high frequency being on the order of 50 times the frequency of the low, and having one fourth the amplitude. The output is passed through a band pass filter to extract the high frequency with its sidebands representing the modulation products. This signal is then envelope detected and low pass filtered to obtain the modulation. The ratio of rms modulation signal to d-c obtained from this filter gives the distortion. That is the distortion is defined as the rms modulation index of the modulated wave.

Unlike the CCIF method the SMPTE method responds to both even and odd-order distortion. The response to even coming from the $\omega_2 \pm (2n-1)\,\omega_1$ sidebands and the response to odd from the $\omega_2 \pm 2n\,\omega_1$ sidebands. The bandwidth of the band-pass filter should be on the order of 20 ω_1 to pass high order products if any.

Comparison of Harmonic Distortion and Intermodulation Tests

Warren and Hewlett, in the April 1948 Proc IRE published a paper which compares the results to be expected from harmonic distortion and SMPTE intermodulation tests. Their principal results in terms of our symbols are tabulated below. e is the total peak of the fundamental, or fundamentals in the output.

Transfer Characteristic	Total HD	SMPTE IM	$\frac{\text{IM}}{\text{HD}}$
Single ended Amplifier	$\frac{e}{2}\,\frac{k_2 + k_4 e^2}{k_1 + \frac{3}{4} k_3 e^2}$	$\frac{8}{5}\,e\,\frac{k_2 + \frac{51}{50} k_4 e^2}{k_1 + \frac{99}{100} k_3 e^2}$	$\approx \frac{16}{5} = 3.2$
Push-pull Amplifier $k_{even} = 0$	$\frac{e^2}{4}\,\frac{k_3 + \frac{5}{4} k_5 e^2}{}$	$\frac{24}{25}\,e^2\,\frac{k_3 + \frac{33}{10} k_5 e^2}{}$	$\approx \frac{96}{25} = 3.84$

These results hold if the transmission of the system prior to the distorting stage is the same for all input frequencies used, and if the transmission of the system after the distortion is the same for all output frequencies produced. In other words in a flat system harmonic distortion measurements give just as much information as intermodulation tests. The fact that the IM tests give higher figures is of no consequence, the point is the ratio is constant so one test gives the result for both.

If, however, the system is not flat with frequency the results may not bear this same ratio; and one test may show up high distortion where another will not. Consider, for example an output stage with an input equalizer which compensates for a falling high frequency response in the output transformer. Harmonic tests might show no trouble at the high end because harmonics generated in the output stage (although greater because of the increased drive) are attenuated by the transformer. With two high frequency inputs, however, the difference frequencies generated would not be attenuated. If the stage is single ended the CCIF method would reveal the trouble. If push pull, components of the form $2\omega_1 - \omega_2$ or $2\omega_2 - \omega_1$ should be measured.

In general the distortion behaviour of non-flat systems can be summarized as follows:

Low End Pre-emphasis - Distortion produced mainly by low frequency components of signal as harmonics, as intermodulation of low frequency components, and as modulation of high frequency components by lows.

High End Pre-emphasis - Distortion produced mainly by high frequency components of signal - usually most noticeable in difference frequency intermodulation products.

Low End Post-emphasis - Exaggerates difference frequency intermodulation products.

High End Post-emphasis - Exaggerates harmonic distortion and sum products (at least in certain parts of frequency ranges) - suppresses difference products.

The above rules are merely qualitative. If the distortion source is localized and the frequency characteristic before and beyond are known then the actual behaviour can be calculated quantitatively and comparisons of HD and IM methods made analytically for the particular case involved.