- 10 -
DISTORTION

TRANSDUCER NONLINEARITY

Loudspeaker distortion is a broad subject, especially if one considers linear distortion – frequency response – as part of the subject matter. In this chapter, we are interested only in nonlinear distortion. A discussion of nonlinear distortion is by its very nature fundamentally different than the subjects that we have dealt with thus far. We have looked at tools and techniques for synthesizing and analyzing various designs of transducers and transducer systems. When dealing with nonlinear systems, one seldom "designs" the distortion characteristics, although this is not without some attraction. We are mostly interested in eliminating distortion. The goal of zero distortion is attractive, but not really practical, for it ignores the realities of cost trade-offs required by the marketplace. Lowering distortion below the point at which it is objectionable is not a good cost-benefit trade-off.

In order to perform the task of making an optimum trade-off between distortion and perception, one must have capabilities in two areas: the analysis of nonlinear systems and the subjective impression of nonlinear distortion. When these two subjects are understood, then we can begin to optimize the design. In this chapter, we will deal with both of these aspects of the problem. Nonlinear systems theory is well established and we can discuss it with certainty. However, the subjective perception of distortion is not well understood, mostly because there is little data on which to support or disprove a hypothesis.

10.1 Nonlinear Systems Theory

There are a multitude of mechanisms that can create distortion products in a transducer. If we are going to design out these mechanisms, then we must know what these mechanisms are and how they effect the end result – the distorted signal. While there are many mechanisms, fortunately there is one feature that is common among the majority of them. We can define a function that relates the instantaneous output level of some quantity versus the instantaneous input level of this same quantity. When this relationship is not a straight line, then the system is said to be nonlinear. Several examples of this function are shown in Fig. 10-1. These curves denotes what is known in the literature as a memoryless nonlinear transfer function, memoryless because it has no frequency dependence. It is sometimes called a static nonlinearity. The importance of this distinction will become apparent later.

The input-output relationship can be between any two variables of the same type, displacement, velocity, current, voltage, any variable of the system. The only requirement is that this "block" must be placed in the domain in which it is defined and it must relate identical quantities. These are not the most general restrictions, but they make our analysis simpler without compromising its validity.

Note in Fig. 10-1 that the clipping transfer characteristic is completely linear, as long as the input remains below $|.6|$. However, if we allow the input to go to .7 or 1.0, then the distortion becomes highly dependent on the input level. This makes clear an important point that we must always consider. What values are we going to allow as inputs?

If we scale the output values to be unity when they reach some predefined level, x_{peak} (we would have preferred the term x_{max}, but its historical usage, which is inconsistent with our usage here, prohibits us from doing that), then we can see that the output scale would go from -1 to 1. What we use as an input scale is not arbitrary and it should be arranged so that the output level never exceeds ±1 for any valid input level. If this is not done, then there can be an ambiguity (and a failure of the applicability of the theory) in the series expansions that we will use. If the system has a scalable gain, then we can always scale the gain to accomplish this task. If, on the other hand, the system gain is set by the systems characteristics then we must be careful in selecting the allowed input range so as to get a valid mapping curve.

We should realize by now (remember orthogonal sets of functions) that the curves shown in Fig. 10-1 can be expanded in many different ways. For example we could expand the curves into Legendre Polynomials and study theses expansions in that domain. As we shall see, there are very good reasons to do that. We could also expand them as Chebycheff Polynomials, or Laguerre Polynomials (as Weiner does[1]) etc. For our purposes right now, a simple polynomial expansion is attractive because of its simplicity. We will let

$$T(x) = \sum_n a_n x^n \tag{10.1.1}$$

The solid curve in Fig. 10-1 has the equation

$$T(x) = .8x + .1x^2 - .2x^3 \tag{10.1.2}$$

Here, the gain of the system is given as .8, but since the output does not come close to either ±1, we should adjust the gain to be 1.1 to better normalize the curves or we should readjust the allowed input scaling. The results of any nonlinear analysis depend on the choice of x_{peak} and the gain values. The choice of too large a value for x_{peak} and/or too small a gain will result in larger higher order coefficients. It should be apparent that the value of $x_{peak} = 1.0$ in the transfer characteristic of Eq. (10.1.2) can never be reached with the input levels as defined. It is always desirable to define x_{peak} in such a way that one of the limits is reached

[1] see Schetzen, *The Volterra and Weiner Theories of Nonlinear Systems*

when the limits of the input is reached, i.e. the maximum output should always be normalized to the maximum input. When this is done we will get a more consistent and useful specification of the system.

The first term in equation Eq.(10.1.1) is called the *offset* term. None of the curves in the figure have an offset. The second term is the *gain*. The third term is know as the second order or *quadratic* nonlinearity. The fourth term is know as the third order or *cubic* nonlinearity. Higher order terms are simply defined by their powers, i.e. fifth order, x^5 and so on. There is no limit to the number of orders that can be required to represent a given transfer characteristic. For example, the two curves with sharp slope discontinuities would require very high orders to fit them over the range of -1 to 1. This is an extremely important point, as we shall see.

As an example of the effect of a nonlinearity in a system, consider a nonlinear transfer function with only gain and a *quadratic* nonlinearity. We know that for a sinusoidal excitation, the output will contain harmonics of the input. Given an input $x(t)$, this can be shown as follows

$$x(t) = A\cos(\omega t) = A\left(\frac{e^{-i\omega t} + e^{i\omega t}}{2i}\right) \tag{10.1.3}$$

then the output $y(t)$ will be

$$\begin{aligned} y(t) &= T(x(t)) = a_{.1}x(t) + a_2 x(t)^2 \\ &= a_1 A\left(\frac{e^{-i\omega t} + e^{i\omega t}}{2i}\right) + a_2 A^2\left(\frac{e^{-i\omega t} + e^{i\omega t}}{2i}\right)^2 \\ &= a_1 A\left(\frac{e^{-i\omega t} + e^{i\omega t}}{2i}\right) + a_2 A^2\,\frac{1}{2}\left(\frac{e^{-i2\omega t} + e^{i2\omega t}}{2i} + 1\right) \end{aligned} \tag{10.1.4}$$

The output contains the original input scaled by the gain a_1 and a second harmonic, at 2ω scaled by $a_2A^2/2$. If the output is normalized to the input then A can be taken as one. There is also an offset term in the output that results from a second order nonlinearity. The use of complex exponentials is desirable due to the simplicity of taking powers, but we need to remember that we must always use two complex exponentials (both signs) or we get an incorrect result. (Why?)

We can now see an interesting relationship by considering the well know expansion of powers of x into Legendre Polynomials[2]. These results are

$$x^n = \frac{n!}{1\cdot 3\cdot\ldots(2n-1)}\left[P_n(x) + \frac{2n-3}{2}P_{n-2}(x) + \frac{(2n-7)(2n-1)}{2\cdot 4}P_{n-4}(x) + \cdots\right] \tag{10.1.5}$$

which means that

[2] See Skudryk, *The Foundations of Acoustics*

$$\begin{aligned} x^0 &= P_0(x) \\ x^1 &= P_1(x) \\ x^2 &= \tfrac{2}{3} P_2(x) + \tfrac{1}{3} P_0(x) \\ x^3 &= \tfrac{2}{5} P_3(x) + \tfrac{3}{5} P_1(x) \end{aligned} \tag{10.1.6}$$

It may not be obvious how we would use this information, so let's reconsider our previous example.

Let $x(t)$=A cos(ωt) as in our example above (a_1x+a_2x^2) which would yield

$$\begin{aligned} & a_1 A P_1(\cos(\omega t)) + a_2 A^2 \left[\tfrac{2}{3} P_2(\cos(\omega t)) + \tfrac{1}{3} P_0(\cos(\omega t))\right] \\ &= a_1 A \cos(\omega t) + a_2 A^2 \left[\tfrac{2}{3}\tfrac{1}{4}(3\cos(2\omega t) + 1) + \tfrac{1}{3}\right] \\ &= a_1 A \cos(\omega t) + a_2 A^2 \ \tfrac{1}{2}\left(\cos(2\omega t) + 1\right) \end{aligned} \tag{10.1.7}$$

the identical result. This means that the Legendre Polynomials offer us a convenient and concise alternative to determining the harmonic content of any order of nonlinearity. We could of expanded the nonlinear transfer functions directly in terms of the Legendre Polynomials, but this might not have been as clear.

10.2 Loudspeaker Component Nonlinearity

In considering a loudspeaker as a nonlinear system, we must consider how each element in the system contributes to the nonlinear portion of the problem. The fundamental elements are listed in the table below:

component	displacement variation	temperature variation	importance
mass	none	none	none
compliance	high	low	medium
mechanical resistance	low	low	none
Bl	high	medium	high
R_e	None	high	medium
inductance	medium	low	low

Table 10.1:

It may seem unusual to see thermal variations listed as a nonlinear characteristic, but in fact they are. What makes them fundamentally different from what we usually think of as nonlinearity is the fact that the thermal variations are very slow. They don't happen at acoustic frequencies and hence they don't generate nonlinear by-products that lie in band as sound. They do have a strong effect on system performance, however, and must not be ignored. We will discuss the thermal aspects or our investigation at the end of this chapter.

The importance column is based on our judgment as to the propensity of the component to generate objectionable types of audible distortion. The only high on the list for displacement variations is *Bl* and it is certainly well understood that *Bl* variation with displacement is a major, if not the major source of loudspeaker nonlinearity. Next on the list would be the thermal variations of R_e and the *Bl* product – principally the magnet flux. Finally, would be the compliance variation with displacement (although the stiffness variation with displacement is more useful).

The *Bl* product nonlinearity is complicated as we shall see. On the other hand, the stiffness is relatively easy to handle. The inductance is a little more complicated than the stiffness, but certainly easier than the *Bl* product. All of these nonlinear components have memory – i.e. they are frequency dependent.

When we describe the *Bl* product and suspension stiffness nonlinearity as being the principal ones we have excluded some important issues regarding the nonlinearity of the medium. The medium of air is considered to be very linear – at least when compared to other mediums like water. It is still true that air can be a significant source of nonlinearity in some types of transducers, particularly compression drivers. The pressures just adjacent to the diaphragm in these devices can become so high as to generate nonlinear distortion which is comparable to that in the driver itself. The issue here is that there is no solution to this problem other than reducing the actual amplitudes of the pressure changes. (We will not consider electronic means of canceling this distortion. They are not within our scope here.) The primary consideration for the pressure magnitudes is the compression ratio, which for a typical 10:1 ratio increases the SPL at the diaphragm by 20dB – a significant amount. Reducing this compression ratio substantially reduces the SPL, but does so at the sacrifice of efficiency. One must make the choice in these devices between low distortion or high efficiency. We should point out, however, that nonlinearity of the medium is extremely low order, dominantly quadratic, and as such would not, by our hypothesis (to follow), be found to be highly objectionable. The trade-off between the perception of objectionable distortion in a compression driver versus efficiency has not been studied to our knowledge.

It has been shown[3,4] that most nonlinear systems can be considered to be a parallel combination of subsystems, each branch having a particular nonlinear transfer function order. In general, each leg has its own frequency dependence. This is shown schematically in the figure below. There can be an infinite number of branches, but we have only shown three.

[3] see Schetzin, *The Volterra and Wiener Theories of Nonlinear Systems*

[4] see Bendat, *Nonlinear Systems Techniques and Applications*

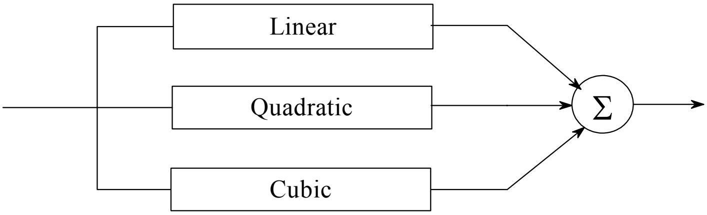


Figure 10-1 - Schematic of a nonlinear system

It is actually amazing that a nonlinear system can be represented in this way, for this is a very simple model. Unfortunately, a loudspeaker does not fit this simplified approach or at least not exactly.

Considering the nonlinear form for differential equation for a voltage driven moving coil transducer with negligible load impedance as shown in Eq.(1.9.22) on page 16. We can write this differential equation for the diaphragm motion as

$$M_m \frac{d^2x}{dt^2} + \left(R_m + z_m + \frac{Bl(x)^2}{R_e} \right) \frac{dx}{dt} + k(x)x = \frac{e(t)\, Bl(x)}{R_e} \qquad (10.2.8)$$

M_m = *mechanical mass of system*

R_m = *mechanical resistance of system*

z_m = *external mechanical load on system*

x = *diaphragm displacement*

$e(t)$ = *voltage input*

R_e = *coil resistance*

$Bl(x)$ = *displacement dependent force factor*

$k(x)$ = *displacement dependent diaphragm stiffness*

There are two complications to fitting this equation into the block diagram form of Fig.10-1. The first is that the forcing function itself is nonlinear and second is that the electromagnetic resistance terms are proportional to the square of a nonlinear function. These problems result from the fact that we have a transformer (or a gyrator depending on the model chosen) with a nonlinear coupling factor, a situation not usually found in the classical study of nonlinear systems. Most nonlinear theory considers systems whose nonlinear coefficients only appear on the left hand side of Eq.(10.2.8).

The situation described above requires a model like that in Fig.10-2. Models like this can be extremely hard to analyze analytically. It is not too hard to analyze with numerical techniques that have coefficients fit to measured data, but it is very difficult to derive simple equations for the nonlinear terms. The question arises as to what simplifications would we have to make in order to allow a transducer to be modeled as shown in Fig.10-1. This form is highly desirable since the analysis of systems like these are a simple extension of the well established linear systems theory (see Sec.1.13 on page 21). Simplification of a nonlinear transducer to the simpler model would allow us analyze this system in a manner which is analogous to that of a multiple input single output linear system[5].

It turns out that there is really only one major assumption that we need to make in order to make the desired simplifications, the system must be approximately linear; the nonlinearity can be thought of as first and second order perturbations of the linear response. In other words, we are simply saying that the nonlinear terms are small relative to the linear ones. This is not too severe a limitations for some uses, but it is for others. Our intention in this chapter is to give the reader sufficient background in nonlinear theory so that the techniques that we will introduce in a later chapter will have a clear foundation in theory. Our primary consideration then is to obtain the ability to analyze the major components nonlinearities.

From a perceptual standpoint the main contributor to transducer distortion will be nonlinearity in the motor. Take for instance the Bl product of a moving coil loudspeaker. Even the simplest form of nonlinearity, second-order, will generate fourth order systems response terms in the electromagnetic damping (the dominant one), along with second-order nonlinearities in the driving function. This means that this simplest of all Bl nonlinearity will cause distortion products up to the sixth order. Third-order nonlinearity of this component will cause distortion products up to the ninth order. Clearly, motor nonlinearity for a loudspeaker must be small if one is to consider the system to be even quasi-linear. This becomes the most limiting restriction on our model. Significant orders higher than the third would create a transducer which would have an almost chaotic output for large inputs. We have probably all heard this type of distortion.

Proceeding on with the traditional multiple leg nonlinear system analysis we have to allow each leg to have two transfer functions as shown below. The nonlinearity block is now a memory-less nonlinearity of a single order $2...n$.

As an example of how we might apply this model, consider a transducer example. The block diagram shown in Fig.10-2 has a voltage as an input and a pressure as an output. The nonlinearity that we are considering for the Bl and

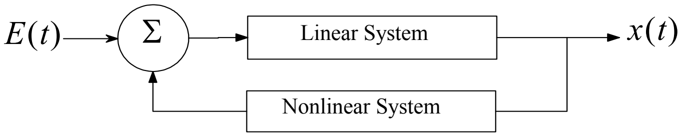


Figure 10-2 - System block diagram for a loudspeaker

[5] Bendat and Piersol, *Engineering Applications of Correlation*

stiffness terms are both functions of the displacement of the diaphragm. Therefore, the first block in each leg for each nonlinearity would be the transfer function from the voltage to the diaphragm displacement, a high-pass function (see Fig. 1-9 on page 17). The first transfer functions for each leg in our example should therefore become identical (we will find this to be mostly true and we have assumed this in the figure above). If there are nonlinear characteristics which depend on the diaphragm velocity, like Doppler distortion, or viscous flow type nonlinearity (most viscous related resistance goes as the velocity squared), then this pre-transfer function would map from the voltage to the velocity, a bandpass function. There are only two types of nonlinear transfer curves that are encountered in audio systems; those nonlinear in a variable and those nonlinear in the slope of a variable. There could be acceleration nonlinearity, but these are not found to be significant in transducers. Acceleration and velocity nonlinearity will not be discussed owing to the fact that they are so much smaller than the ones that we will be discussing. The analysis shown here, however, is directly applicable.

The next block is the memoryless nonlinear transfer function for the particular leg. It is x^1, x^2, x^3, or some higher order function in x. Following this block is another transfer function which maps from the displacement for that particular leg to that components effect on the pressure response. This later transfer function contains the coefficients a_n from Eq. (10.1.1). In general there are n legs, one for each nonlinear term in the nonlinear transfer function expansion.

We now have to consider the post transfer functions. These are different for each of the different legs, and will be functions of the *Bl,* compliance, etc. nonlinear representations. Lets assume that we can represent the major nonlinear components as

$$Bl(x) = Bl_0 + b_1 x + b_2 x^2 \tag{10.2.9}$$

and

$$k(x) = k_0 + k_1 x + k_2 x^2 \tag{10.2.10}$$

Using these forms in the differential equation Eq. (10.2.8) we get

$$M_m \frac{d^2 x}{dt^2} + \left(R_m + z_m + \frac{Bl_0^2 + 2Bl_0 b_1 x + \left(b_1^2 + 2Bl_0 b_2\right) x^2}{R_e} \right) \frac{dx}{dt} + k_0 x + k_1 x^2 + k_2 x^3$$
$$= \frac{e(t)\left(Bl_0 + b_1 x + b_2 x^2\right)}{R_e} \tag{10.2.11}$$

where we have already simplified the equation by expanding the $Bl(x)^2$ term and retaining only orders up to the third. The limitations of our assumption about small nonlinearity in the terms is already evident in the above equation. The electromagnetic damping term cannot go negative which implies that

$$Bl_0^2 + 2Bl_0 b_1 x + \left(b_1^2 + 2Bl_0 b_2\right) x^2 > 0 \tag{10.2.12}$$

for all valid displacements x. Since there will always become value of x for which this is not true the value of x where the above term becomes zero makes an ideal definition of x_{peak}

$$x_{peak} = \frac{-b_1 \pm \sqrt{-2Bl_0 b_2}}{b_1^2 + 2Bl_0 b_2} Bl_0 \tag{10.2.13}$$

where we take the sign which yields the smallest value. We can immediately see that b_2 must always be negative. We cannot analyze a system beyond x_{peak} because our equations will become unstable.

Just as we have shown to be so effective in Sec. 10.1, we will assume an input that is a Legendre polynomial

$$e(\theta) = A\, P_1(\cos\theta) \tag{10.2.14}$$

and an output as a series of these polynomials

$$Y(\theta) = \sum_{n=0}^{3} a_n P_n(\cos\theta) \tag{10.2.15}$$

where $\theta = \omega t$. If we plug these two functions into the nonlinear differential equation using Eq. (10.1.6) we can expand the resulting equation into a set of coefficients of the different orders of the Legendre Polynomials. This process results a very large equation which is extremely unwieldy (we will let a computer sort out the algebra). We retained all orders of nonlinearity, but dropped terms which are higher in P_n than three. Higher orders could be calculated but with some additional complexity (we will leave that exercise ...). Finally we solve each equation in P_n for a_n.

The result of this rather elaborate algebraic manipulation is, for a_1

$$a_1(\omega) = \frac{\dfrac{Bl_0 + \frac{3}{5} b_2}{R_e} + \frac{4}{35}\left(i\omega \dfrac{b_1^1 + \frac{6}{7} b_2^2 + 2Bl_0 b_2}{R_e} - k_2 \right) a_3(\omega) + \frac{1}{5}\left(i\omega \dfrac{3b_1 b_2 + 4Bl_0 b_1}{R_e} - k_1 \right) a_2(\omega)}{-\omega^2 M_m - i\omega \left[R_m + z_m + \dfrac{Bl_0^2 + \frac{3}{5}\left(b_1^2 + \frac{2}{3} b_2^2 + 2Bl_0 b_2\right)}{R_e} \right] + k_m + \frac{3}{5} k_2} \tag{10.2.16}$$

This is a satisfying result for it quantifies several things that we know to be true.

First, there are corrections to the force factor

$$Bl_{force} = Bl_0 + \tfrac{3}{5} b_2 \tag{10.2.17}$$

which result in a compression of the output because $b_2 < 0$. We can also see that the impedance function in the denominator stiffens, if $k_2 > 0$, as it almost always is and there is also a modification of the electromagnetic damping term

$$Bl_{damping}^2 = Bl_0^2 + \tfrac{3}{5}\left(b_1^2 + \tfrac{2}{3}b_2^2 + 2Bl_0 b_1\right) \tag{10.2.18}$$

A final check is, as the nonlinear terms in Eq. (10.2.16) go to zero the correct linear system results.

The next term of interest is the offset term

$$a_0(\omega) = \frac{1}{3}\frac{\begin{array}{c}\left(2i\omega\dfrac{b_1 b_2 + Bl_0 b_2}{R_e} - k_1\right)a_1(\omega) + \left(i\omega\dfrac{\frac{3}{4}b_1^2 + \frac{2}{3}b_2^2 + \frac{5}{3}Bl_0 b_2}{R_e} - k_2\right)a_2(\omega) + \\ \cdots \qquad \tfrac{2}{3}i\omega\dfrac{b_1 b_2}{R_e}a_3(\omega) + \dfrac{b_1}{R_e}\end{array}}{-\omega^2 M_m - i\omega\left[R_m + z_m + \dfrac{Bl_0^2 + \frac{2}{3}\left(b_1^2 + \frac{3}{5}b_2^2 + 2Bl_0 b_1\right)}{R_e}\right] + k_0 + \tfrac{1}{3}k_2} \tag{10.2.19}$$

from which we can see that the offset is frequency dependent.

The next term will be the quadratic, P_2 term

$$a_2(\omega) = \frac{2}{3}\frac{\begin{array}{c}\left(i\omega\dfrac{b_1^2 + \frac{6}{7}b_2^2 + 2Bl_0 b_2}{R_e} - k_2\right)a_0(\omega) + \left(i\omega\dfrac{\frac{12}{7}b_1 b_2 + 2Bl_0 b_1}{R_e} - k_1\right)a_1(\omega) + \\ \cdots \qquad \tfrac{2}{3}\left(i\omega\dfrac{\frac{4}{3}b_1 b_2 + 2Bl_0 b_1}{R_e} - k_1\right)a_3(\omega) + \dfrac{b_1}{R_e}\end{array}}{-\omega^2 M_m - i\omega\left[R_m + z_m + \dfrac{Bl_0^2 + \frac{4}{10}\left(b_1^2 + \frac{2}{3}b_2^2 + 2Bl_0 b_1\right)}{R_e}\right] + k_0 + \tfrac{4}{10}k_2} \tag{10.2.20}$$

and finally we get the P_3 cubic values

$$a_3(\omega) = \frac{2}{5}\frac{\begin{array}{c}\left[i\omega\dfrac{\left(\frac{6}{7}b_2^2 + b_1^2 + 2Bl_0 b_2\right)}{R_e} - k_2\right]a_1(\omega) + \left[i\omega b_1\dfrac{\left(2b_2 + 3Bl_0\right)}{R_e} - \tfrac{4}{35}k_1\right]a_2(\omega) + \\ \cdots \qquad 2i\omega b_1 b_2 a_0(\omega) + \dfrac{b_2}{R_e}\end{array}}{-\omega^2 M_m - i\omega\left[R_m + z_m + \dfrac{Bl_0^2 + \frac{3}{8}\left(b_1^2 + \frac{2}{3}b_2^2 + 2Bl_0 b_2\right)}{R_e}\right] + k_0 + \tfrac{3}{8}k_2} \tag{10.2.21}$$

Note that the denominators in each of these equations are almost identical to the denominator for the linear terms (as we hypothesized it should be). There will not be much error in assuming that the denominators for the a_0, a_2 and a_3 coefficients are all equal to the linear one (although in our results we have not done that). These equations all depend on each other and some form of iterative solution must be performed. We first assume a_2 and a_3 are both zero in Eq. (10.2.16) and solve for $a_1(\omega)$. Most likely we would solve Eq. (10.2.21) with $a_0 = a_2 = 0$ followed by Eq. (10.2.20) with $a_0 = 0$. The process should be clear. We can reinsert new values back into previously calculated equation and continue this process

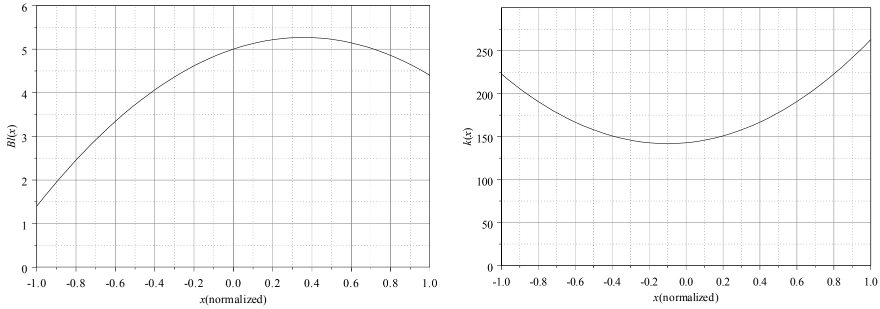


Figure 10-3 - Nonlinear parameters Bl and k versus normalized amplitude (to x_{peak})

until there are no longer major changes in the results. Iterations beyond two are seldom required.

We must not forget that what we have solved for are not the amplitudes of the harmonics (fundamental included) but the amplitudes of the Legendre Polynomials. To get the harmonic amplitudes we note that

$$\begin{aligned} x_0(\omega) &= a_0(\omega) + \tfrac{1}{4}a_2(\omega) \\ x_1(\omega) &= a_1(\omega) + \tfrac{3}{5}a_3(\omega) \\ x_2(\omega) &= \tfrac{3}{4}a_2(\omega) \\ x_3(\omega) &= \tfrac{5}{8}a_3(\omega) \end{aligned} \tag{10.2.22}$$

$x_n(\omega)$ = *the amplitudes of the harmonics*

We should make a few comments about why we used the Legendre Polynomials in these equations. Mostly, it makes the analysis easier to do since the equations only contain simple products of two P_n's which can always be reduced to a simple sum of single terms. Using complex exponentials for the calculations is possible, but long and complex.

10.3 Simulated Results of a Nonlinear Transducer

At this point, we will look at some results for a simulated example. We will use the linear values for the transducer parameters as defined in Sec.2.7 on page 39 and shown in Fig.2-7. We will let the parameters $Bl(x)$ and $k(x)$ be

$$Bl(x) = 5.0 + 1.5x - 2.1x^2 \tag{10.3.23}$$

with x normalized to x_{peak} and

$$k(x) = 150 + 20x + 100x^2 \tag{10.3.24}$$

which are both plotted in Fig.10-3.

The predicted linear displacement output for both the compressed and uncompressed calculations are shown in Fig.10-4. Note that the compressed displacement has been substantially reduced from the un-compressed output owing to the decreased motor strength. These curves are for an input signal which creates an excursion at low frequencies equal to the peak excursion x_{peak}, to which the above two expansions (Eq. (10.3.23) and Eq. (10.3.24)) have been normalized.

The predicted higher order displacement outputs are all shown in Fig. 10-5. It is an easy matter to convert these displacements into sound pressures. Note that these curves have a small dip at about resonance. This implies that resonance is not a good place to evaluate distortion, as is so often done. The real problem is the extremely large displacements below resonance, which would modulate all frequencies above resonance, if there are any signals present in this region. For this reason, it is extremely important to control excursion below resonance in a loudspeaker. The transducer must not be allowed to produce sound at these frequencies because it will likely produce more distortion than actual sound. With proper design of the excursion capability a driver can be operated in this region, but this takes careful consideration of the nonlinear components.

In concluding this section, we would like to point out that we have only done a cursory job of analyzing the nonlinearity of a transducer. The subject in its complete form is massive, albeit quite interesting. The interested reader should first read Schetzen[6] for a complete understanding of the theory or Bendat[7] for a more practical view of the subject. The main point that we want to make here is that the orders, whether in terms of Legendre Polynomials or harmonics, have a fre-

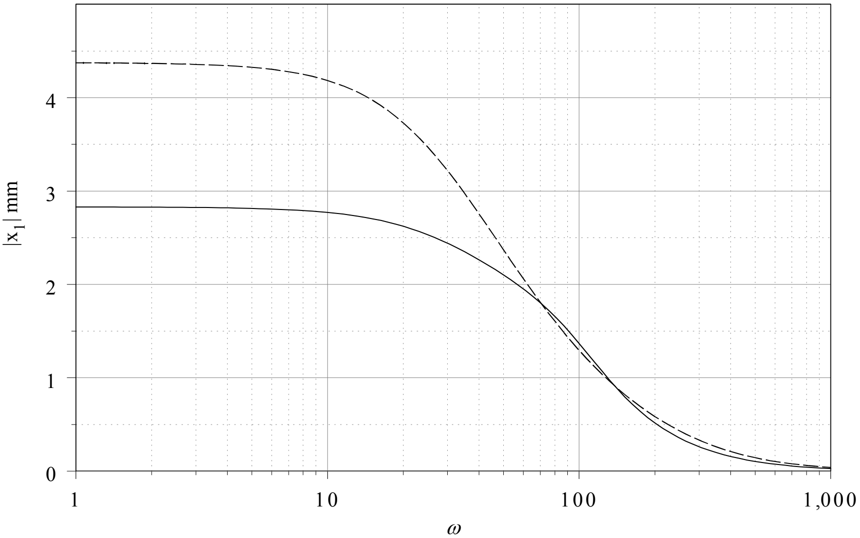


Figure 10-4 - Linear displacement output and the compressed linear displacement

[6] Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*

[7] Bendat, *Nonlinear Systems Techniques and Applications*

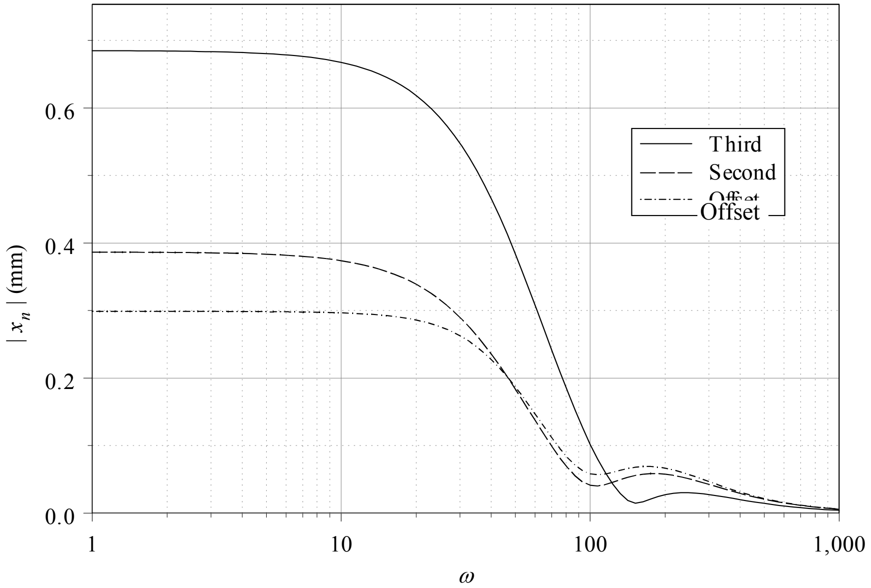


Figure 10-5 - Higher order displacement transfer functions

quency response in an analogous manner to the linear response and that the higher order terms frequency response are direct functions of the nonlinear parameters. From this, we can conclude that a knowledge of the frequency response of the orders should allow us to calculate the terms in Eq.(10.2.9) and Eq.(10.2.10). It would be a real advantage if we could use standard techniques from linear systems theory on our multi-legged model in order to analyze the nonlinear components of a transducer. We will develop this technique in Chap.12.

10.4 Background On Distortion Perception

Now that we have seen how we can analyze a loudspeaker to determine its nonlinearity in terms of the nonlinear transfer functions expanded into polynomial orders, we would like to have some way to relate this concrete mathematical theory to the less quantitative subjective aspects of the perception of distortion. We will propose a hypothesis for such a relationship, but we will do so without proof, since there is no data to either validate or invalidate it. Our hypothesis will be based on what we do know about nonlinear systems and human auditory perception.

Historically, the audio community has viewed distortion in the context of a systems nonlinear response to a sinusoid or sometimes, two or more sinusoids, basically a signal based metric. A metric is a value which is given to a system to indicate its relative scaling within some predefined context. For instance, temperature is a metric when the context is human perception. We can describe the perception of temperature in words like hot, warm, cool or cold. Since temperature also has an exact scientific scaling, it is a simple matter to map from the subjective metric to the physical one, although we must always remember that the subjective terms are relative and precise mapping is not possible. Whenever human perception is involved, metrics can only ever be statistically relevant.

The current metrics of distortion are, Total Harmonic Distortion (THD); Inter-Modulation Distortion (IM), multi-tone intermodulation, etc., all expressed as a percentage – the ratio of the distortion by-products to the total system output. In an absolute sense this view of distortion is satisfactory. If our goal is to eliminate all distortion then clearly any measure of its level is adequate. This goal, as we have stated, is naive. It is neither reasonable nor desirable to set as our goal the complete elimination of all distortion. In this context, the signal-based metrics fall short of the mark, for they fail to correlate with, or even consider, subjective impression.

By their very nature all transducers have limitations. We have seen some very real limitations on frequency response and directivity in previous chapters, but up until now we have not considered any limitations on output from these devices. As the transducers output increases, the range of motion of the mechanical system generating (or receiving) the sound must also increase. To have unlimited output, the system would have to be capable of unlimited motion – a physical impossibility. Therefore the physical constraints that exist in any transducer will inherently limit its output. The manner in which this limitation occurs is important for, as we have seen, there are an infinite number of ways that this can occur. It is this immense variety of nonlinear mechanisms that prevents a single metric from being very meaningful.

With a reliable metric we could base psychoacoustic studies on it and the same mapping could be done for transducer distortion as we described for temperature. But to be useful a metric must be consistent – the same number must mean the same thing in every context and there must be a close correlation between the metric and the subjective response. This is where the signal-based distortion metrics fail. It can be shown that .01% THD in an amplifier can be perceived as unacceptable while 1% THD in a loudspeaker can be perceived as inaudible. This simple fact invalidates THD as a viable metric for discussion of the perception of distortion. Furthermore, 1% THD is not at all the same as 1% IM. Some of the signal-based metrics may be "better" than others, but in our opinion they all fall short of what we are seeking.

How then does one establish a metric for the quantification of distortion that is consistent, reliable and (hopefully) correlates with subjective impression? Based on what is known about the human hearing system and what we have learned about the nonlinear systems analysis, we will propose such a metric. In keeping with our promise – that this would be a theoretical text – we will not offer up any experimental results or supporting or refuting data. *This will be left as an exercise for the reader!*

10.5 The Psychoacoustics of Distortion Perception

In Chap. 13, we will attempt to supply sufficient information on the human hearing system to support our usage of the concepts here. The reader may wish to consult this chapter if the psychoacoustic terminology or the concepts being used are unfamiliar.

One reason that the perception of nonlinear distortion is so complex is that the hearing mechanism itself is not linear and taken as a "system" it is quite complex. It should thus be expected that it will be a difficult task to ascertain what levels and types of nonlinearity the ear can perceive and even more difficult will be the scaling of the subjective impression of these nonlinear functions.

The attribute of hearing that overwhelmingly dominates the perception of distortion is that of masking. Masking is also the principal effect used in the creation of all modern techniques of perceptual coders (MP3, AAC, etc.). When masking effects allows us to reduce the data by 90% or more, in a way that is subjectively benign, then one has to suspect that masking would have a profound effect on the perception of nonlinear distortion. Masking has no analog in linear systems theory, and it is not very intuitive since it does not occur in common systems other than the ear.

From our knowledge of masking we postulate the following two fundamental characteristics.

- Masking is predominately upward toward higher frequencies although masking does occur in both directions.
- The masking effect widens – masking occurs further away from the masker – at a substantial rate with excitation level.

Given these characteristics we will now hypothesize the following three *Perception Principles.*

- Distortion by-products that are created upward in frequency are likely to be less perceptible (masked to a greater extent) than those that fall lower in frequency (postulate 1).
- Distortion by-products that lie closer to the excitation are likely to be less perceptible (masked) than those that lie farther away (masking is a localized effect – it only occurs in the vicinity of the masker).
- Distortion by-products of any kind are likely to be more perceptible at lower levels than at higher levels (postulate 2).

The following discussion relies on these "principles," given without proof, as its foundation. If one accepts these principles as valid, then what we say in the following sections should have substantial validity.

We have already seen the following facts.

- Odd and even orders do not interact, odd orders generate only odd harmonics, even orders generate only even harmonics.
- An *n*th order nonlinearity generates *n*th order harmonics and every other harmonic below it.

- Harmonics of pure tones are generated only above the input signal (this is true only for nonlinear transfer functions which can be represented by Eq. (10.1.1), which, fortunately for us, is true for everything that we will talk about.)
- For multi-tones an nth order nonlinearity causes sidebands at ±*n* times the modulation frequency and every other value of *n* below it as well as harmonics (as above).

These may all seem obvious since we have already provided the mathematical foundation for these points. The interesting part comes next.

Consider Fig. 10-6 where typical distortion products are shown for tones in four situations: a low order nonlinearity and a high order nonlinearity, at a low signal level and a high signal level. Approximate masking curves of the principle masker tone are also shown. We can see that the higher order distortion products are not masked as well as the lower order ones and that the masking effect is greater at the higher signal level. The low order distortion at a high signal level is completely masked in this figure. The high order distortion is never masked, but it would be more audible at low levels.

If we take these facts and join them up with our Perception Principles then we can make the following statements, which are, perhaps, not exact, but are, none the less, more valid than not.

- The masking effect of the human ear causes higher order nonlinearities to be more audible than lower order ones.
- Distortion that rises with level can be completely masked if the order of the nonlinearity is low.

Again these may seem intuitively obvious.

These statements give rise to our hypothesis for a new approach to quantifying nonlinearity (distortion):

- Nonlinearity within the specified operating output range should be of low order – the importance of the order being weighted by $(n - 1)^2$ where n is the order of the nonlinearity ($n > 1$).
- No order should increase with decreasing input level.

As qualitative measures these objectives are reasonable, but only with extensive subjective testing will we be able to put quantitative values to the metric proposed here – an interesting study that has yet to be done.

Consider now our first example of the failure of THD to differentiate between loudspeaker distortion and amplifier distortion. If the amplifier has crossover distortion then this type of nonlinearity violates both of our principles – it is both very high order and it increases (as a proportion of the linear terms) with decreasing signal level. One would expect, based on our hypothesis, that this type of distortion would be highly objectionable and it is. Now consider a loudspeaker. Unless it has some severe design or manufacturing problems, it will have lower orders of nonlinearity and the distortion will only rise with level. Based on our

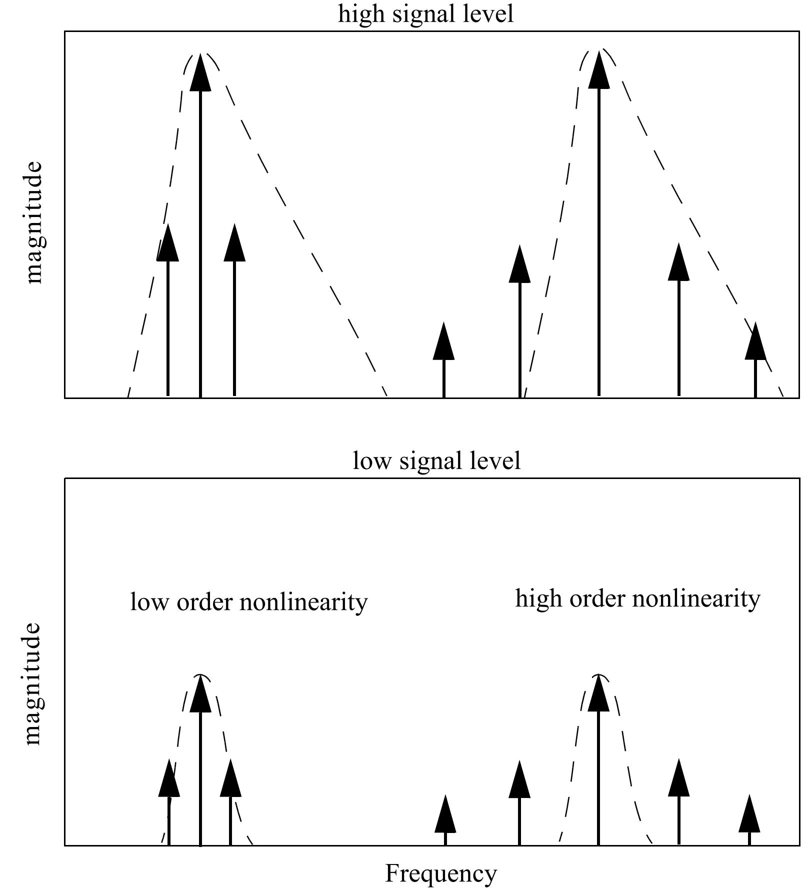


Figure 10-6 - Schematic representation of the effect of masking on the perception of distortion

principles, we should expect this type of distortion to be benign, almost inaudible, and this is in fact what we find to be true. Generally speaking, electronics and mechanics have different nonlinear characteristics. It is not at all uncommon to see very high orders of nonlinearity in electronics, but it is rare to see higher orders in mechanical systems. Our new view of distortion explains a lot of the THD based metric paradoxes.

So basically our new "metric" is the actual parameters of the nonlinear components themselves, or the frequency response of the orders, weighted by their order and required to only grow with level (again relative to the linear term). It is not that uncommon to see discussions of 2nd and 3rd order nonlinearity – we did it ourselves – but it is rare to see a discussion of the higher order nonlinearity. If increasing orders are indeed more audible than lower orders then limiting our discussions to only the lower two orders is seriously flawed. The *root cause* of distortion is the underlying nonlinearity of the system or subsystem and the correct way to discuss nonlinearity is with the orders of its nonlinear transfer function. When one views the distortion problem in this way, signal based distortion metrics (IM, THD, etc.) become irrelevant. It is, and will likely remain so, unclear as to the relationship of the signal-based metrics to subjective impression. It is the authors hope that the audio community will give the outdated notion of THD, IM, signal types, etc. (signal-based concepts) as these are all just symptoms of the real problem – nonlinearity.

10.6 Thermal Nonlinearity

It may perhaps be a misnomer to call thermal effects a nonlinearity since they do not cause distortion in the usual sense of signal distortion. Thermal effects in transducers generally do not generate distortion by products, but they do distort the frequency response – i.e. they cause severe linear distortion. In an ideal system, just like nonlinearity, there would not be any dependence of the parameters on the temperature, but that, like nonlinearity, is not realistic. Since we cannot eliminate these thermal changes, we need to understand how important they are and how we might minimize their negative aspects. The subject of thermal parameter dependence is important and we felt that this is the best place to put this discussion. We will talk specifically about loudspeakers in this discussion, but the problems exist in any motor structure to some degree.

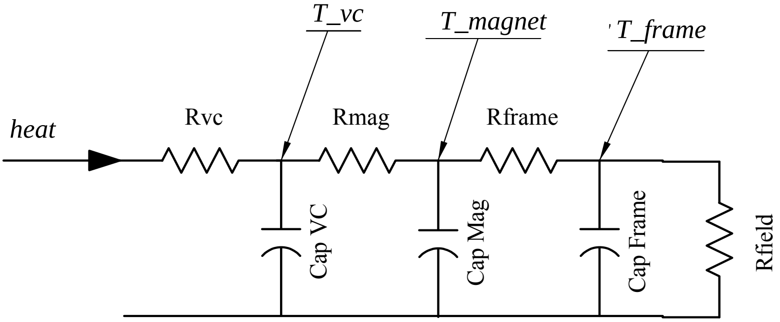
It is likely that all of the parameters of a moving coil loudspeaker are thermally dependent, but there are two that are critically so. The first is the voice coils electrical resistance R_e, which increases with increasing temperature at a predictable rate. The second is the Bl product which will also vary with temperature at a predictable rate, which almost always falls with temperature. The complexity here is that these two effects occur over substantially different time scales due to the different thermal masses involved.

The thermal problem is another differential equation that has only a single first order derivative. It can also be simplified with "lumped parameter" methods into an electrical equivalent circuit which is composed of resistors and capacitors with the voltage representing the temperature and the current the thermal flux. (There are no inductors in this model.) A simple thermal model of a loudspeaker motor structure is shown in Fig. 10-7. This is an extremely simple circuit, but it does a good job of demonstrating what we need to understand.

The current source in this model is the heat generated in the voice coil as

$$\frac{V(t)^2}{\Re(z_e)} \tag{10.6.25}$$

$\Re(z_e)$ = *the real part of the transducers complex electrical input impedance*



T_vc = voice coil temperature
T_magnet = magnet temperature
T_frame = frame and enclosure temperature
Rvc = thermal resistance from heat generation to voice coil
Rmag = thermal resistance from voice coil to motor structure
Rframe = thermal resistance from magnet to frame
Rfield = lumped thermal resistance from motor to infinity
Cap VC = thermal capacity of voice coil
Cap Mag = thermal capacity of motor

Figure 10-7 - Thermal model for a loudspeaker

This heat will take a finite, but very small amount of time to raise the voice coil temperature, denoted as *T_vc*. This thermal resistance Rvc must be finite or the voice coil's temperature would rise immediately, which cannot be true since, albeit it is small, the voice coil does have some thermal mass and it does take a finite amount of time for the heat generation to raise the temperature of this thermal mass. If this were not true and the voice coil did in fact change temperature instantaneously then there would be distortion by-products created by the thermal modulation of the voice coil resistance. In fact, only if the time constant for the voice coil heating is longer than the period of the lowest frequency of usage can we ignore the resistance modulation effects as actual signal distortion. Fortunately, most woofers have substantial voice coils and this thermal modulation distortion can usually be ignored.

The next thermal resistance, *Rmag,* is from the voice coil to what is probably the largest heat sink (thermal capacitor) in the system, the magnet structure. *Rmag* is a fairly high resistance and the one we would most like to be low. The thermal mass of the magnet causes the time constant for *T_mag* to be very long. The magnet heats slowly, but it also cools slowly. There could be a resistance to ground at *T_mag* which would represent thermal radiation off of the magnet or thermal cooling via convection, we have lumped all of these effects into *Rfield.*

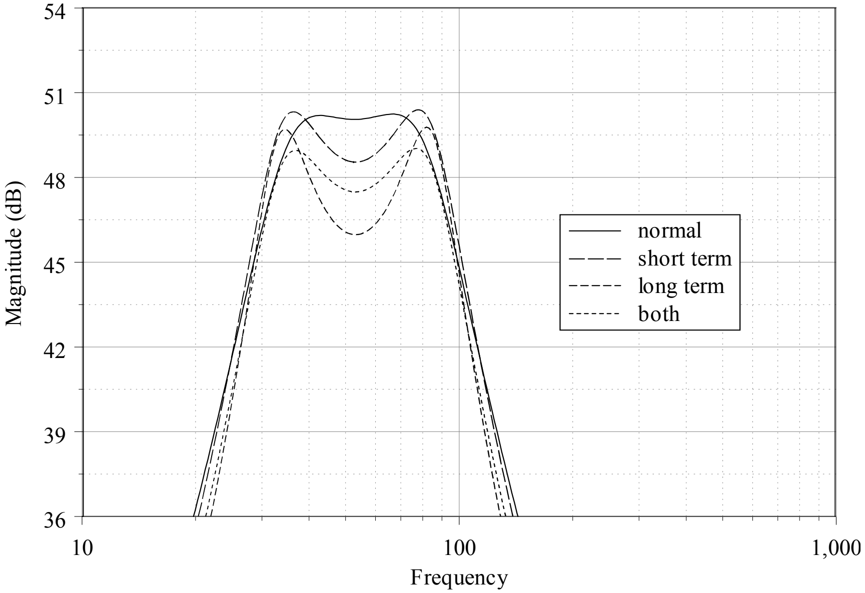


Figure 10-8 - Thermal variations for a typical bandpass system at 80°C

The voice coil temperature will follow the time variations of Eq. (10.6.25) fairly closely with only a short time averaging and lag of the input signal. The magnet, on the other hand, has a very long time constant. For our purposes, we can simply assume that the magnet temperature will continue to rise until it reaches the mean temperature of Eq. (10.6.25) over a very long time interval – perhaps hours.

It is evident that we must consider both the effect of the voice coil temperature rise as well as the magnet temperature rise, although we will do so in different ways. By simply making the flux density B and the resistance R_e functions of the temperature in our standard T-matrix models, we can easily study these two effects. The vastly different time constants makes them virtually independent and the effects will be additive. For our example, we will reexamine the system of Fig. 5-19 on page 117. Fig. 10-8 shows the response of this system in four states:

- *normal* – when the driver is first energized and a low signal is applied.
- *short term* – as above but for high level signals
- *long term* – after the magnet has heated but a low level signal is applied
- *both* – a hot magnet with a high level signal applied

Initially, the response will modulate between the *normal* and *short term* curves depending on signal level, but as time goes on it will modulate between the long term curve and the both curve. When one also considers that at high input levels, there will be nonlinear effects, such as the loss of Bl with excursion the high level system output can get very poor indeed.

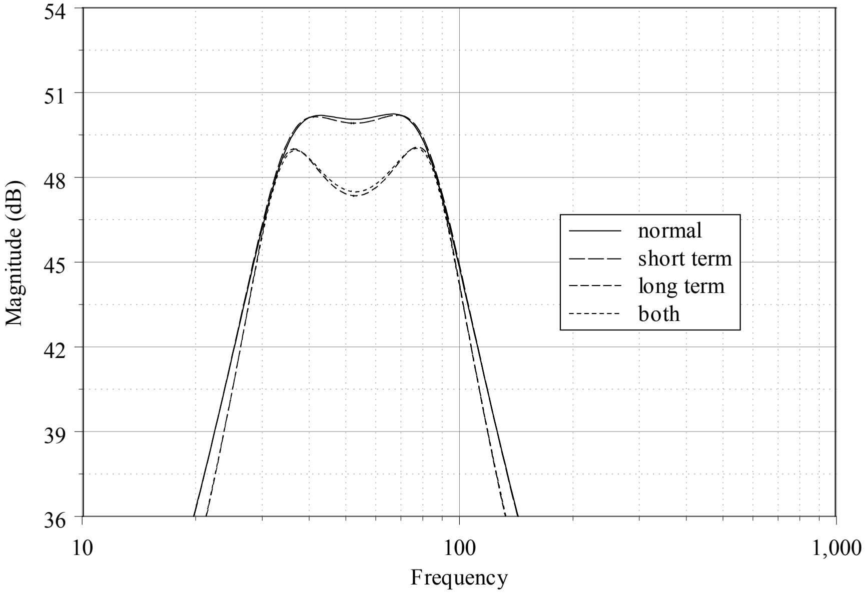


Figure 10-9 - Thermal variations with ALNICO magnet

Fig. 10-9 shows the improvement gained by the use of ALNICO as the magnet. ALNICO has almost no thermal variation in its flux. While an improvement, the critical problem that remains is the variation of the resistance with temperature. Each magnet type has its own particular thermal dependence. NdFeB has about half the variation of ceramic (the highest) and several times that of SmCo, which is comparable to but a little higher than ALNICO. We must also keep in mind that the thermal capacities of each of these magnet will differ quite substantially.

In order to control the remaining thermal modulation effects, we must look at the actual voice coil material. The common voice coil materials, copper and aluminum, have similar thermal dependence. In the case of copper however, the addition of a small percentage of nickel makes an alloy which is noted for its low change of resistance with temperature. If the voice coil is made of this material we will get the response shown in Fig. 10-10. This loudspeaker has a 6% nickel-copper alloy voice coil with an ALNICO magnet. While this result is attractive the addition of the nickel substantially raises the resistively of this copper alloy. This requires that the voice coil wire be of a larger cross section area for a given R_e and hence a heavier voice coil results, which is never an advantage. The larger wire does have the benefit of a larger surface area, which dissipates heat more readily, improving the power handling capacity of the voice coil. This could, perhaps, offset the lower output that will result from the increased coil mass. At any rate, these are all trade-offs which must be considered in the design.

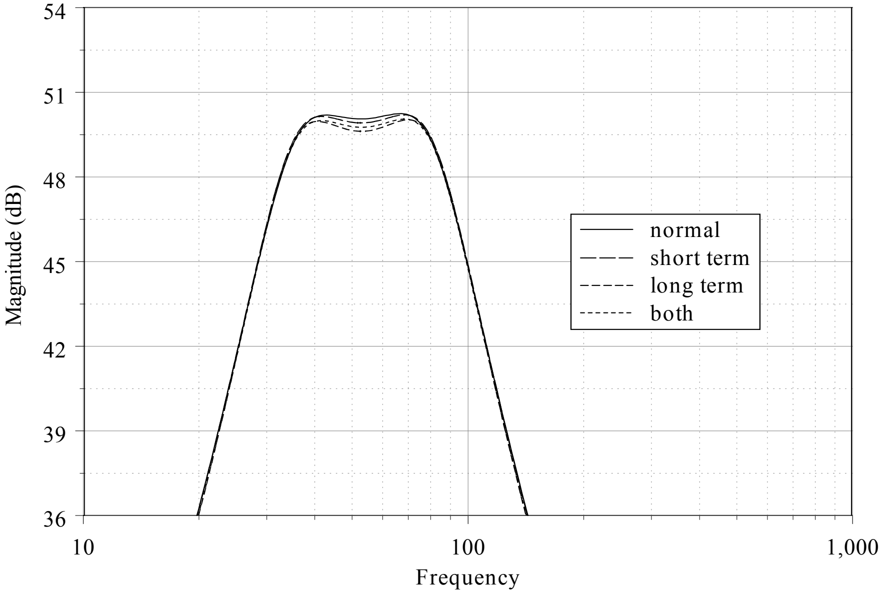


Figure 10-10 - Thermal modulation for ALNICO and 6% nickel wire

A study of the circuit in Fig.10-7 will show that a change in the thermal resistance values does little to change the effects noted here. It does help to prevent catastrophic failure due to a thermal breakdown of the voice coil wire coating and bonding, but it does not improve the thermal modulations – it only makes them happen quicker or slower. The only way to actually reduce the thermal modulation is to dissipate the heat off to ground (infinity), which could never be done quickly enough, or implement some material changes which reduce the sensitivity to temperature. Clearly any mechanism that removes heat from the motor structure is a benefit.

10.7 Summary

In this chapter, we investigated several effects which are inherent in a transducer and degrade its performance at higher levels. None of these effects can actually be eliminated, but all of them can be minimized or optimized. To design a transducer or a system without due consideration of its high level performance is almost certain to result in a less than satisfactory design. To utilize a transducer without a knowledge of its sensitivity to these effects invites trouble. These high level performance issues are often the driving forces behind trade-offs in size, weight, performance and most notably cost. Balancing these trade-offs is the art.