ERRORS IN SERVO SYSTEMS USING SINUSOIDAL FREQUENCY (PHASE) MODULATION*

Fred L. Walls
Time and Frequency Division
National Bureau of Standards
Boulder, Colorado 80303

Abstract

This paper reviews the errors in determining the center of a resonance line which are due to residual imperfections in practical electronic systems using sinusoidial frequency or phase modulation. In particular the effects of residual amplitude modulation, baseline distortion, and harmonic distortion in the modulation process and the demodulator are qualitatively analyzed for a Lorentzian line in the limit of small modulation index. This permits one to easily calculate analytically the frequency offsets as a function of modulation index and the transfer function of the fundamental and various harmonics of the modulation frequency. Using this model one can easily formulate accurate tests for experimentally measuring the frequency errors in practical servo systems, even if the original assumptions about small modulation index and a pure Lorentzian line are not exactly fulfilled.

Introduction

Many systems use sinusoidal frequency or phase modulation of a probe frequency in order to find the center of a resonance line. The purpose of this paper is to review the residual imperfections which occur in practical systems and the subsequent errors in determining line center. In particular the effects of residual amplitude modulation, baseline distortion, and harmonic distortion in the modulation and the demodulation process are qualitatively analyzed for a Lorentzian line in the limit of small modulation index, this permits one to easily calculate analytically the frequency offsets as a function of modulation index and the transfer function of the fundamental and various harmonics of the modulation frequency. Based on this model one can then compare the relative susceptability of various servo configurations to residual electronic imperfections. Additionally one can easily formulate accurate tests for experimentally measuring the frequency errors in practical servo systems, even if the original assumptions about small modulation index and a pure Lorentzian line are not exactly fulfilled.

Model of a Resonance Line and Error Signal

One of the most common, methods for determining the center of a resonance line with high precision is to sinusoidally modulate the frequency (or phase) of the probe and detect the phase of the resulting amplitude modulated signal at the fundamental of the modulating signal. The general scheme is shown in Figure 1.

The various subsystems and their effect on errors in determining the center of the resonance will be analyzed in later sections.

*Contribution of the National Bureau of Standards, not subject to copyright.

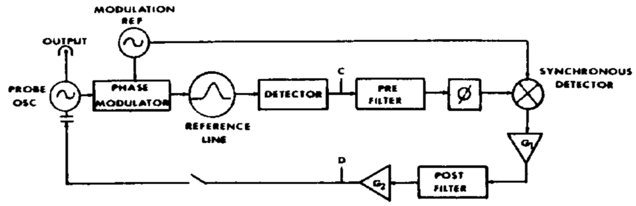

Fig. 1. Block diagram of a sinusoidally modulated probe oscillator which can be locked to the center of a reference line.

Curve a of Figure 2 shows a typical resonance line that would be observed at point c of Figure 1 as a function of slowly sweeping the frequency of the probe (without modulation) across the resonance. Curve b of Figure 2 is the derivative of curve a.

For the moment, let's assume that the probe output is a sinewave with a spectral width very narrow compared to the width of the resonance shown in Figure 2, curve a. If the center of the probe is at the point A, then the output signal increases as the frequency of the VCO is increased, and at point B, the signal decreases as the probe frequency increases. If the frequency of the probe is swept back and forth (FM), then the signal has both a dc and an ac component. If the deviation of the FM is small compared to the half linewidth W, then the demodulated and filtered output of the synchronous detector (measured at Point D of Figure 1) fairly accurately reproduces the derivative of curve a. Note that in curve b, the point of zero signal, which also has the steepest slope, nominally occurs at the center of the resonance line. This curve is referred to as a frequency discriminator curve. The signal at point D can be used to steer the probe frequency because near line center we now have a dc signal proportional to the frequency error between the probe frequency and the center of the resonance.

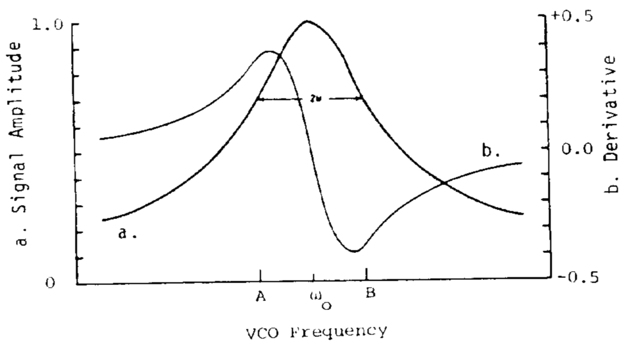

Fig. 2. Reference line (a) and its first derivative (b).

Now let's examine this process in a little more detail. More generally, assume that we have a symmetric Lorentzian line superimposed on a sloping and curved background. Then:

$$\text{Signal Amp} = \left(\frac{\gamma^2}{\gamma^2+(\omega-\omega_o)^2}\right)^{1/2} + K_1(\omega-\omega_o) + K_2(\omega-\omega_o)^2 \qquad (1)$$

where K_1 and K_2 are the first two coefficients of a Taylor expansion of the background about line center, $\gamma = \pi W$ is the half angular linewidth, ω is the instantaneous angular frequency of the probe, and ω_o is the true center of the resonance.

Real frequency or phase modulators have small non-linearities and therefore generate small components of modulation at multiples of the modulation frequency. Also the modulation reference generally has some higher harmonic components as well. Therefore let's assume that the modulated signal is of the form

$$\omega = \omega_1 + B\cos\Omega t - K_3\sin 2\Omega t + K_5\cos 2\Omega t + K_6\sin 3\Omega t \qquad (2)$$

where Ω is the modulation frequency.

The effects of distortion in the reference and the modulation process are contained in coefficients K_3, K_5 and K_6. This model assumes that the residual modulation at $\Omega/2$, Ω, 2Ω, etc., due to spurious signals on the probe control line is small compared to that imposed by the modulator. This places a heavy burden on the postfilter (see Figure 1) especially in servos using a square wave reference for the demodulation. The modulation process can and usually does cause some amplitude modulation, therefore another term, $K_4\cos\Omega t$, needs to be added to equation 1. Substituting for ω in equation 1 and adding the K_4 term yields equation 3. It has been assumed that the amplitude modulation is in phase with the frequency modulation, which yields the maximum offset.

Signal Amp =

$$\left[1 + \frac{1}{\gamma^2}[(\omega_1-\omega_o) + B\cos\Omega t - K_3\sin 2\Omega t + K_5\cos 2\Omega rt + K_6\sin 3\Omega t]\right]^{-1/2}$$

$$+K_1[\omega_1 - \omega_2) + B\cos\Omega t]$$
$$+ K_2[\omega_1 - \omega_o) + B\cos\Omega t]^2$$
$$+ K_4\cos\Omega t \qquad (3)$$

Near line center, with the coefficients K_3, K_4, K_5, K_6 very small compared to γ, and the modulation amplitude B only slightly smaller than γ, the denominator can be expanded using the approximation

$$\frac{1}{1+\delta} = 1 - \frac{\delta}{2} + \frac{3}{8}\delta^2 - \frac{15}{48}\delta^3 \ldots \text{for } \delta<1 \qquad (4)$$

Signal Amplitude =

$$- \frac{B^2}{4\gamma^2}\cos 2\Omega t\ (1+\delta') \qquad (5a)$$

$$- \frac{B}{\gamma^2}\ \Delta\omega\cos\Omega t\ \ (1+\delta'') \qquad (5b)$$

$$+ \frac{B}{2\gamma^2}K_3\ (\sin\Omega t+\sin 3\Omega t)\ (1+\delta'') \qquad (5c)$$

$$- \frac{B}{\gamma^2}K_5\ (\cos\Omega t+\cos 3\Omega t)\ (1+\delta'') \qquad (5d)$$

$$- \frac{B}{\gamma^2}K_6\ (\sin 2\Omega t + \sin 4\Omega t)\ (1+\delta'') \qquad (5e)$$

$$+ K_1 \quad B\cos\Omega t \qquad (5f)$$

$$+ K_2\ (2\Delta\omega B\cos\Omega t + \frac{B^2}{2}\cos 2\Omega t) \qquad (5g)$$

$$+ K_4\cos\Omega t \qquad (5h)$$

where $\delta' = -\frac{3}{4}\frac{B}{\gamma^2}\cos^2\Omega t + .267\frac{B^6}{\gamma^6}\cos^6\Omega t \ldots$

$$\delta'' = +\ \frac{3}{2}\ \frac{B^2}{\delta^2}\cos^2\Omega t - 2.491\ \frac{B^6}{\delta^6}\cos^6\Omega t\ .$$

$$\Delta\omega \equiv \omega_1-\omega_o$$

Dc terms, and terms involving the product of two or more small coefficients, eg. K_3K_5, have been dropped.

Note that δ' and δ'' are even power series of $(B/\gamma)^2\cos^2\Omega t$ and could have been given in terms of a Bessel function. They contain mixtures of $\cos 2\Omega t$, $\cos 4\Omega t$, $\cos 6\Omega t$ etc. and have a finite value averaged over a period of the modulation frequency Ω.

Term 5a contains a dc contribution plus even harmonics of Ω (mostly 2nd) due to sweeping over the line profile. Expanding term 5a yields

$$5a = 1 - \frac{B^2}{2\gamma^2}\ (1 - .375\frac{B^2}{\gamma^2} + .16\frac{B^6}{\gamma^6}\ldots)$$

$$-1/4\ \frac{B^2}{\gamma^2}\cos 2\Omega t\ (1 - 3/4\frac{B^2}{\gamma^2} + .267\frac{B^2}{\gamma^6}\ldots.)$$

$$-1/4\ \frac{B^2}{\gamma^2}\cos 4\Omega t\ (.129\ \ \frac{B^2}{\gamma^6}\ldots) \qquad (6)$$

(6)

Term 5b contains the desired error signal proportional to the frequency error $\Delta\omega$. Its harmonic content is odd with contributions at Ωt, $3\Omega t$, $5\Omega t$, etc. coming from the expansion of δ''. Note that dependence of this term on $(B/\gamma)^2$ is the same as for the unwanted error terms 5c through 5e.

$$5b = \frac{-\Delta\omega B}{\gamma^2}[\cos \Omega t \left(1 + \frac{9}{8}\frac{B^2}{\gamma^2}\ldots\right) + \frac{3}{8}\frac{B^2}{\gamma^2}\cos 3\Omega t(1+\ldots)] \quad (7)$$

Fundamental Sinewave Demodulation

The most common types of demodulators used to recover the error signal displayed in Equation 5 are the sinewave demodulator and the squarewave demodulator. The primary distinction between the two is the type of reference. The reference can be at the frequency of modulation or at a higher harmonic - typically the third.

The first type to be considered is the fundamental sinewave demodulator. The detector of Figure 1 is assumed to be linear. This is very important as nonlinearities can cause intermodulation between the various terms of Equation 4 yielding large errors. These type of errors will not be analyzed here. The function of the prefilter is to filter noise and spurious signals from the detected signal by narrowing the bandwidth. Of particular importance is the reduction of the signals at $2\Omega t$, $3\Omega t$, $4\Omega t$, etc. In addition to the potentials errors originating from terms 5c, 5d, 5e, and 5f of equation 4, the demodulator should be operated at the highest possible level so as to minimize the relative effects of dc offsets in the demodulator output.

Before filtering, the signals at $2\Omega t$, $4\Omega t$, etc. generally far exceed the noise near line center ($\Delta\omega = 0$) and therefore limit the useful dynamic range of the demodulator if not attenuated. Assume that the prefilter attenuates the signal at $2\Omega t$ by K_{12}, and by K_{13} at $3\Omega t$, etc.

The reference signal is further assumed to be the same as that used in the modulator, phase shifted by ϕ, where ϕ is due to various delays in the electronics and can be a function of the environment-especially temperature. For $\phi \ll 1$, $\cos(\Omega t+\phi)$ can be approximated as $\cos(\Omega t) + \phi\sin(\Omega t)$ yielding

$$\text{Ref} = \underset{8a}{\cos \Omega t} + \underset{8b}{\phi \sin \Omega t} - \underset{8c}{K_7 \sin 2\Omega t} + \underset{8d}{K_8 \cos 2\Omega t} + \underset{8e}{K_9 \sin 3\Omega t} \quad (8)$$

Mathematically, the effect of the demodulator is to multiply the signal of equation 5 by the reference signal given in equation 8 [1].

The servo acts to force the output of the demodulator towards zero. The actual error depends on the servo gain. If the dc servo gain, $G_1 G_2$, is sufficiently large, one can assume that the demodulator output is zero. For simplicity the modulator output has been averaged over three full periods of the modulation frequency Ω and the desired frequency error term, 9a, set opposite to the spurious error terms in equation 9.

The effective dc offset [1] of the demodulator is represented by K_{dc} and is a function of the total ac gain of the system, G_{Ac}.

$$\underset{9a}{\Delta\omega(1+\delta')} = \underset{9b}{\gamma^2 L_1} \underset{9c}{+2L_2\Delta\omega\gamma^2} \underset{9d}{+1/2K_3\phi(1+\delta'')} + \underset{9e}{K_4\frac{\gamma^2}{B}} \underset{9f}{+1/2K_5(1+\delta'')} \underset{9g}{+1/2(K_{12})[K_\gamma B)(1-\delta'')} + \underset{9h}{K_2 B\gamma^2 k_g} \underset{9i}{+K_6K_7)1+\delta'')} \underset{9j}{\frac{2K_{dc}\gamma^2}{B}} \quad (9)$$

Term 9b is due to the linear component of the background slope and is selected out of the error signal by 8a. This error is just the ratio of the background slope to the slope of the derivative multiplied by the angular half bandwidth. In cases where this effect is exceptionally large and/or unmanageable, a third derivative lock can be used at the expense of signal to noise. See later discussion.

Term 9c is also selected out of the error signal by 8a and causes no frequency error by itself, however in the presence of other error terms it effectively modifies the angular half width γ. This effect is usually small and can be ignored.

Term 9d selected out of the error signal by 8b, is due to the out-of-phase component of the second harmonic distortion in the phase modulator ($\sin 2\Omega t$), the effect of this term can be reduced considerably by making ϕ small. Values of ϕ between .01 and .1 are generally easy to achieve and maintain.

Term 9e is selected out of the error signal by 8a and is due to the fractional amplitude modulation, K_4, at $\cos \Omega t$. Since for most systems $\gamma/B \approx 1$ the error is approximately K_4 multiplied by the half angular bandwidth γ. This can be a major limitation in some systems.

Term 9f, selected out of the error signal by 8a, is due to the mixing of the in-phase component of the modulator harmonic distortion ($\cos 2\Omega t$) with the fundamental of modulation by the resonance. This can be seen from the expansion of the cross products in the denominator of eq. 3. Because of this there is no method to suppress it other than by making K_5 small. Note that the offset is just 1/2 the amplitude of the in phase 2nd harmonic distortion.

Term 9g is selected out of the error signal by K_8 and is due to the 2nd harmonic generation from sweeping back and forth across the resonance. Near line center the $\cos 2\Omega t$ error signals usually dominate all other error signals. By making K_{12} small one can greatly reduce the susceptability to

2nd harmonic distortion in the demodulator and permit G_{AC} to be increased to the largest value consistent with the noise in the bandwidth of the demodulator.

Terms 9h and 9i are second order small and can be neglected in this approximation.

Term 9j is due to the dc offset in the demodulator. Usually K_{dc} is independent of level for small signal levels but at some point the errors grow exponentially with signal level. By making K_{12} very small, one can increase the signal gain to the point that the noise around frequency Ω in a bandwidth determined by the prefilter is just below the maximum level for the demodulator. This and $B/\gamma \approx 1$ minimizes the effect of K_{dc}.

Thus, for systems where K_{12} is small the most important error terms for sinewave demodulation at the fundamental are

$$\Delta\omega(1+\delta'') = \gamma^2 K_1 + 1/2 K_3 \phi(1+\delta'') + K_4 \frac{\gamma^2}{B} + 1/2\ K_5(1+\delta'') + 2Kdc \frac{\gamma^2}{B} \qquad (10)$$

Fundamental Square Wave Demodulation

For many systems it is easier to implement a squarewave demodulator than it is to use a sinewave demodulator and K_{dc} is often much smaller. In this instance the reference signal of equation 8 is replaced by:

$$\text{Ref} = \cos\Omega t + \frac{1}{3}\cos 3\Omega t + \phi\sin\Omega t + \left(\frac{\theta}{3} + K_9\right)\sin 3\Omega t) -K_7 \sin 2\Omega t + K_8 \cos 2\Omega t \qquad (11)$$

where the terms varying as sin $6\Omega t$ or cos $6\Omega t$ have been omitted.

It is easily shown that the frequency offset errors of the closed-loop system are functionally very similar as those derived in eq. 10 above.

$$\Delta\omega(1+\delta'') = \gamma^2 K_1 + \frac{5}{6} K_3\phi(1+\delta'') + K_4 \frac{\gamma^2}{B} + \frac{5}{6}\ K_5(1+\delta'') + 2K_{dc}\frac{\gamma^2}{B} \qquad (12)$$

Third Harmonic Demodulation

In some cases the background slope is so large and/or unstable that if is advantageous to use a 3rd harmonic reference to the demodulator. Assume it is of the form

$$\text{Ref} = \cos 3\Omega t + K_{11}\cos 6\Omega t \ldots$$

In this case the significant frequency errors are given by

$$\Delta\omega(1+\delta'') = \frac{4}{3}\ K_3\phi(1+\delta'') + \frac{4}{3} K_5(1+\delta'') + \frac{16}{3} K_{dc}\frac{\gamma^2}{B} \qquad (13)$$

where the 6th order terms have been neglected. Although the sensitivity to sloping background and amplitude modulation is virtually gone, the signal is generally also reduced by a factor of 2 or 3 which increases the relative importance of 2nd harmonic distortion in the modulator and dc offset in the demodulator. K_{12} should be kept small in order to maximize G_{ac} and thereby reduce the effect of K_{dc}.

Tests for Servo Errors

Errors generated from the K_1 coefficient have the same functional dependence on modulation width as the desired signal and are therefore difficult to separate in a fundamental demodulation system. Therefore, one generally has to measure the background slope separately and calculate the offset. One could also compare the frequency of line center for a fundamental and a 3rd harmonic demodulation system. In cases where a Ramsey structure is present, one can compare the frequency of line center when locked to pairs of successive lobes.

Errors generated from K_2, K_6, and K_7 are generally small and can be neglected.

Errors generated from K_3 can be separated out from the other terms by varying the phase shift ϕ. For most implementations, K_3 varies as B^2. Modeling of the modulator can also be helpful.

The errors associated with K_4 can best be determined by measuring the fractional amplitude modulation at Ω on the probe signal. The phase chosen for the K_4 term is the most likely and has the largest error. K_4 depends on the modulation width B.

The errors associated with K_5 are best illuminated by varying the modulation width B. A plot of frequency change vs. B^2 for $\theta \sim 0$ yields K_5 while the difference between that curve and the one obtained with $\theta \sim 0.2$ can be used to determine K_3. K_5 can also be determined from a careful characterization of the phase modulator.

The errors associated with K_8 are unique to the fundamental demodulator systems and can be illuminated by varying K_{12}. For K_{12} small this error can be totally neglected.

The errors originating from K_{dc} are best separated out by varying the ac gain. Varying the dc gain only changes the loop attack time (bandwidth) and should have no effect on these offsets [1]. Another technique for illuminating K_{dc} generated errors is to vary the ac gain with no modulation on the probe and measure the dc error signal.

Summary

A simple model of a resonance system probed by a sinusoidally modulated probe signal has been treated to expose the first order errors in determining line center due to imperfections in the electronics. Although this approach does not easily produce rigorous values for the frequency errors, in that it does not take into account saturation etc., it does yield the correct functional dependence of the errors on modulation index, ac gain, etc. This permits one to compare the offsets in determining line center using various servo configurations. As we've shown, in any servo system with a fundamental demodulator reference, the most serious frequency errors originate from sloping background, 2nd harmonic distortion in the frequency modulation, amplitude modulation on the probe signal, and dc offsets in the demodulator. Servo systems utilizing the 3rd harmonic of the modulation as a demodulator reference are generally not sensitive to baseline tilt or amplitude modulation on the probe, but have increased sensitivity to 2nd harmonic distortion in the modulator, and to dc offsets in the demodulator. With the functional dependence outlined here it's relatively easy to design sensitive tests of these offsets even if the original assumption about a pure Lorentzian line and small modulation index are not exactly fulfilled.

Acknowledgments

The author is very grateful to Dr. J. C. Bergquist and Andrea DeMarchi for fruitful discussions and to the Naval Research Laboratories for financial support.

References

[1] F. M. Gardner, Phaselock Techniques (John Wiley, New York, 1966).