Errors in Determining the Center of a Resonance Line Using Sinusoidal Frequency (Phase) Modulation

Invited Paper

FRED L. WALLS

Abstract—The errors in determining the center of a resonance line, which are due to residual imperfections in practical high-precision electronic systems using sinusoidal frequency or phase modulation, are reviewed. 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 calculate analytically the frequency offsets as a function of modulation index, 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 servosystems, 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 to find the center of a resonance line. The purpose of this paper is to review the residual imperfections that 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 demodulation processes are qualitatively analyzed for a Lorentzian line in the limit of small modulation index; this permits one to calculate analytically the frequency offsets as a function of the modulation index, the transfer function of the fundamental, and various harmonics of the modulation frequency. Based on this model one can then compare the relative susceptibility 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 Fig. 1. The various subsystems and their effect on errors in determining the center of the resonance will be analyzed in later sections.

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

For the moment, let us assume that the probe output is a sine wave with a very narrow spectral width compared to the width of the resonance shown in Fig. 2, curve a. If the center of the probe is at the point A, then the output signal increases as the frequency of the voltage-controlled oscillator is increased; 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 Fig. 1) fairly accurately reproduces the derivative of curve a. 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 difference between the probe frequency and the center of the resonance.

Now let us examine this process in greater detail. More generally, assume that we have a symmetric Lorentzian line superimposed on a sloping and curved background. Then the normalized signal amplitude is

$$\text{signal amplitude} = \left[\frac{\gamma^2}{\gamma^2 + (\omega - \omega_0)^2}\right]^{1/2} + K_1(\omega - \omega_0) + K_2(\omega - \omega_0)^2 \quad (1)$$

where K_1 and K_2 are the first two coefficients of a Taylor expansion of the background about line center, $\gamma = 2\pi W$ is the half-angular linewidth, ω is the instantaneous angular frequency of the probe, and ω_0 is the true center of the resonance. A Lorentzian lineshape is chosen to sim-

Manuscript received March 18, 1986; revised November 10, 1986. This work was supported in part by the US Naval Research Laboratories.

The author is with the National Bureau of Standards Time and Frequency Division, 325 Broadway, Boulder, CO 80303.

IEEE Log Number 8716460.

0885-3010/87/1100-0592$01.00 © 1987 IEEE

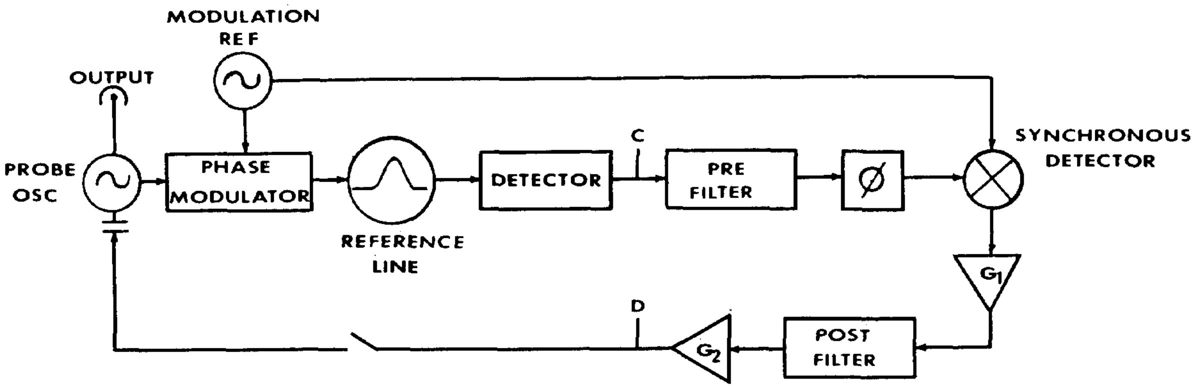

Fig. 1. Block diagram of sinusoidally modulated probe oscillator which can be locked to center of reference line.

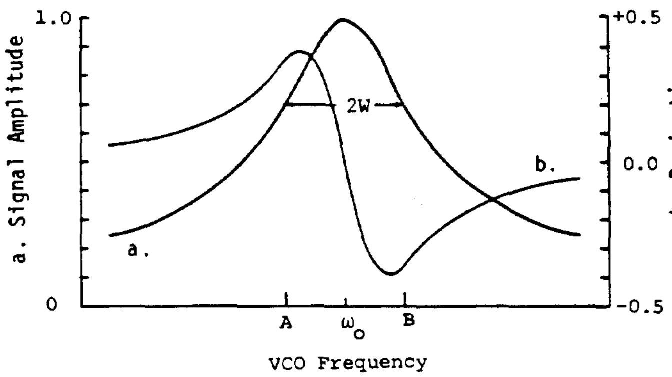

Fig. 2. Reference line a and its first derivative b.

plify the calculations. Other kinds of symmetric resonance lineshapes yield the same leading terms; however, the coefficients differ somewhat. Likewise, the first two terms of the Taylor expansion of the nonflat background yield the leading odd and even terms due to the background.

Real frequency or phase modulators have small nonlinearities and therefore generate small components of modulation at multiples of the modulation frequency. Also the modulation reference signal generally has some higher harmonic components as well. Therefore, let us assume that the modulated probe signal is of the form

$$\omega = \omega_1 + B\cos\Omega t - M_{2s}\sin 2\Omega t + M_{2c}\cos 2\Omega t + M_{3s}\sin 3\Omega rt + M_{3c}\cos 3\Omega t \tag{2}$$

where Ω is the modulation frequency and ω_1 is the average frequency of the probe signal.

The effects of distortion in the reference and the modulation process are contained in coefficients M_{2s}, M_{2c}, M_{3c}, and M_{3s}. This model assumes that residual modulation at harmonics or subharmonics of Ω and especially at $\Omega/2$, Ω, 2Ω, and 3Ω due to spurious signals on the probe control line or other sources is small compared to that imposed by the modulator. This places a heavy burden on the postfilter (see Fig. 1) and on the decoupling between circuits, 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, $A_m\cos\Omega t$, needs to be added to (1). Substituting for ω in (1) and adding the A_m term yields (3). It has been assumed that the amplitude modulation is the phase with the frequency modulation, which yields the maximum offset from this term:

signal amplitude =

$$\left\{1 + \frac{1}{\gamma^2}[\omega_1 - \omega_0 + B\cos\Omega t - M_{2s}\sin 2\Omega t + M_{2c}\cos 2\Omega t + M_{3s}\sin 3\Omega t + M_{3c}\cos 3\Omega t]^2\right\}^{-1/2} + K_1[(\omega_1 - \omega_0) + B\cos\Omega t - M_{2s}\sin 2\Omega t \cdots] + K_2[(\omega_1 - \omega_0) + B\cos\Omega t - M_{2s}\sin 2\Omega t \cdots]^2 + A_m\cos\Omega t. \tag{3}$$

If the coefficients M_{2s}, M_{2c}, M_{3s}, and M_{3c} are very small compared to γ, $A_m < 1$, and if the modulation amplitude B is smaller than γ, then near line center the denominator of the first term can be expanded using the approximation

$$(1+\delta)^{-1/2} = 1 - \frac{\delta}{2} + \frac{3}{8}\delta^2 - \frac{5}{16}\delta^3 \cdots \quad \text{for } \delta < 1. \tag{4}$$

We then have

signal amplitude = dc terms

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

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

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

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

$$+\frac{B}{2\gamma^2}M_{3s}(\sin 2\Omega t + \sin 4\Omega t)(1+\delta'') \tag{5e}$$

$$-\frac{B}{2\gamma^2} M_{3c}(\cos 2\Omega t + \cos 4\Omega t)(1 + \delta'') \quad (5f)$$

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

$$+K_2 \left(2\Delta\omega B \cos \Omega t + \frac{B^2}{2} \cos 2\Omega t\right) \quad (5h)$$

$$+ \; A_m \cos \Omega t \quad (5i)$$

where

$$\delta' = -\frac{3}{4}\frac{B^2}{\gamma^2} \cos^2 \Omega t + 5/8 \frac{B^4}{\gamma^4} \cos^4 \Omega t \cdots$$

$$\delta'' = -\frac{3}{2}\frac{B^2}{\gamma^2} \cos^2 \Omega t + 5/8 \frac{B^4}{\gamma^4} \cos^4 \Omega t$$

and

$$\Delta\omega = \omega_1 - \omega_0.$$

Dc terms, and terms involving the product of two or more small coefficients, e.g., $M_{2s}M_{2c}$, have been dropped in (3) and (5).

δ' and δ'' are even-power series of $(B/\gamma)^2 \cos^2 \Omega t$ and could have been given in terms of Bessel functions. They contain mixtures of $\cos 2\Omega t$, $\cos 4\Omega t$, $\cos 6\Omega t$, etc., and have a nonzero value averaged over multiple periods of the modulation frequency Ω [2].

Term (5a) contains only even harmonics of Ω (mostly second) due to sweeping over the line profile. Expanding this term yields

$$(5a) =$$

$$1 - \frac{B^2}{4\gamma^2}\left(1 - 3/16 \frac{B^2}{\gamma^2} + 0.39 \frac{B^4}{\gamma^4} \cdots\right)$$

$$-1/4 \frac{B^2}{\gamma^2}\left(1 - 3/4 \frac{B^2}{\gamma^2} + 0.58 \frac{B^4}{\gamma^4} \cdots\right) \cos 2\Omega t$$

$$+3/64 \frac{B^2}{\gamma^2}\left(1 - 5/4 \frac{B^2}{\gamma^2} \cdots\right) \cos 4\Omega t. \quad (6)$$

Term (5b) contains the desired error signal proportional to the frequency error $\Delta\omega$. It contains odd harmonics at Ωt, $3\Omega t$, $5\Omega t$, etc., coming from the expansion of $(1 + \delta'') \cos \Omega t$. The functional dependence of this term on $(B/\gamma)^2$ is the same as for the unwanted error terms (5c)–(5f):

$$(5b) = \frac{-\Delta\omega B}{\gamma^2}\left[\left(1 - \frac{9}{8}\frac{B^2}{\gamma^2} + 1.6 \frac{B^4}{\gamma^4} \cdots\right) \cos \Omega t + \frac{3}{8}\frac{B^2}{\gamma^2}\left(1 + 0.78 \frac{B^2}{\gamma^2} \cdots\right) \cos 3\Omega t\right]. \quad (7)$$

Fundamental Sine-Wave Demodulation

The most common types of demodulators used to recover the error signal displayed in (5) are the sine-wave demodulator and the square-wave demodulator. The primary distinction between the two is in the type of reference signal. 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 sine-wave demodulator operating at the fundamental of the modulation. The detector of Fig. 1 is assumed to be linear. This is very important as nonlinearities can cause intermodulation between the various terms of (5), yielding large errors. This type of error will not be explicitly analyzed here as the important intermodulation terms would have a harmonic content similar to terms (9d)–(9h) discussed later. 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 reducing the potential errors originating from terms (5c), (5d), (5e), and (5f), this permits the demodulator to be operated at the highest possible level to minimize the relative effects of dc offsets in the demodulator output. The prefilter also reduces the effect of intermodulation in the following stages.

Before filtering, the signals at $2\Omega t$, $4\Omega t$, etc. generally far exceed the noise near line center ($\Delta\omega \sim 0$) and therefore would limit the useful dynamic range of the demodulator if not attenuated in the prefilter. Assume that the prefilter transmission at $2\Omega t$ is T_2, at $3\Omega t$ is T_3, etc. The reference signal is further assumed to be the same as that used in the modulator, phaseshifted 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

$$\text{ref} = \cos \Omega t \quad (8a)$$

$$-\phi \sin \Omega t \quad (8b)$$

$$-D_{2s} \sin 2\Omega t \quad (8c)$$

$$+D_{2c} \cos 2\Omega t \quad (8d)$$

$$+D_{3s} \sin 3\Omega t \quad (8e)$$

$$+D_{3c} \cos 3\Omega t. \quad (8f)$$

Mathematically, the effect of the demodulator is to multiply the signal of (5) by the reference signal given in (8) [2].

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 average demodulator output is zero [2]. If the modulator output has been averaged over at least six full periods of the modulation frequency Ω, the various trigometric functions can be replaced by their average values, yielding

$$\Delta\omega(\overline{1 + \delta''}) \quad (9a)$$

$$= K_1\gamma^2 \quad (9b)$$

$$+2K_2\Delta\omega\gamma^2 \quad (9c)$$

$$+1/2M_{2s}(-\phi + T_3D_{3s})(\overline{1+\delta''}) \quad (9d)$$

$$-1/2M_{2c}(1 + T_3D_{3c})(\overline{1+\delta''}) \quad (9e)$$

$$+1/2M_{3s}T_2D_{2s}(\overline{1+\delta''}) \quad (9f)$$

$$-1/2M_{3c}T_2D_{2c}(\overline{1+\delta''}) \quad (9g)$$

$$-1/4BT_2D_{2c}(\overline{1+\delta'}) \quad (9h)$$

$$+\frac{A_m\gamma^2}{B} \quad (9i)$$

$$+\frac{K_{dc}\gamma^2}{B} \quad (9j)$$

where $(\overline{1+\delta''})$ refers to the value of $(1+\delta'')$ averaged over six or more periods of Ω.

Term (9a) is the desired frequency discriminant and the other terms of (9) are spurious error terms which ideally should be zero. 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-linewidth. 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 (discussed later).

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 halfwidth γ. This effect is usually small and can be ignored.

Term (9d), selected out of the error signal by (8b) and (8e), 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 and T_3 small. Values of ϕ between 0.01 and 0.1 are generally easy to achieve and maintain. T_3 can easily be made $\leq 10^{-3}$.

Term (9e), selected out of the error signal by (8a), is due to the mixing of the in-phase component of the modulator's second-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 (3). Because of this, no method exists to surpress it other than by making M_{2c} small. The offset is just 1/2 the amplitude of the in-phase second-harmonic distortion. The added effect due to the third-harmonic distortion in the demodulator can be made small by making T_3 small.

Terms (9f) and (9g) arising from the third-harmonic distortion in the modulation process both can be made small by making the transmission at $\cos 2\Omega t$ (T_2) small as well as by using a demodulator with very little second-harmonic distortion. Term (9h) is selected out of the error signal by D_{2c} and is due to the second-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 T_2 small, one can greatly reduce the susceptibility to this effect and also second-harmonic distortion in the demodulator and permit the ac gain to be increased to the largest value consistent with the noise in the bandwidth of the demodulator.

Term (9i) is selected out of the error signal by (8a) and is due to the fractional amplitude modulation A_m at $\cos \Omega t$. Since for most systems $\gamma/B \simeq 1$, the error is approximately A_m multiplied by the half-angular bandwidth γ. Some modification of this result will occur in systems exhibiting saturation effects. This can be a major limitation in some systems.

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 K_{dc} grows exponentially with signal level. These offsets can take the form of rectification or leakage. By making T_2 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 operating level for the demodulator; this along with making $B/\gamma \sim 1$ minimize the effect of K_{dc}. Thus for systems where T_2 and T_3 are small, the most important error terms for sine-wave demodulation at the fundamental are

$$\Delta\omega(\overline{1+\delta''}) = K_1\gamma^2 - 1/2M_{2s}\phi(\overline{1+\delta''}) - 1/2M_{2c}(\overline{1+\delta''}) + A_m\frac{\gamma^2}{B} + 2K_{dc}\frac{\gamma^2}{B}. \quad (10)$$

Fundamental Square-Wave Demodulation

For many systems it is easier to implement a square-wave demodulator than it is to use a sine-wave demodulator, and K_{dc} is often much smaller for square-wave demodulation. In this instance the reference signal of (8) is replaced by

$$\text{ref} = \cos\Omega t + \frac{1}{3}\cos 3\Omega t - \phi\sin\Omega t\left(\frac{-\phi}{3} + D_{3s}\right)\sin 3\Omega t + D_{2s}\sin 2\Omega t + D_{2c}\cos 2\Omega t \quad (11)$$

where higher order harmonics such as $\cos 4\Omega t$ 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 (9) with $\phi \rightarrow 4/3\phi$ and $D_{3c} = 1/3$, for T_2 and T_3 small. The dominate terms are

$$\Delta\omega(\overline{1+\delta''}) = \gamma^2K_1 - \frac{2}{3}M_{2s}\phi(\overline{1+\delta''}) - \frac{1}{2}M_{2c}(\overline{1+\delta''}) + A_m\frac{\gamma^2}{B} + 2K_{dc}\frac{\gamma^2}{B}. \quad (12)$$

Third-Harmonic Demodulation

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

$$\text{ref} = \cos 3\Omega t - \phi \sin 3\Omega t + D_{6c} \cos 6\Omega t \cdots.$$

In this case, with $T_3 = 1$ the significant frequency errors are given by

$$\Delta\omega(\overline{1 + \delta''}) = -\frac{4}{3} M_{2s}\phi(\overline{1 + \delta''}) - \frac{4}{3} M_{2c}(\overline{1 + \delta''}) + \frac{16}{3} K_{dc} \frac{\gamma^2}{B} \tag{13}$$

where the sixth-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 second-harmonic distortion in the modulator and dc offset in the demodulator. T_2 should be kept small 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 for $B/\gamma << 1$. 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 third-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 [3]. In the optical region a number of techniques have been developed to supress nonflat background effects. (For example, see [4]-[6]. See also [7], where the effects of high-modulation indices are calculated for several servo configurations using Ramsey cavities.)

Errors generated from M_{3s}, M_{3c}, D_{2s}, and D_{2c} are generally small and can be neglected, especially if T_2 and T_3 are small. Errors generated from M_{2s} can be separated out from the other terms by varying the phaseshift ϕ. For most implementations, M_{2s} varies as B^2. Modeling of the modulator can also be helpful.

The errors associated with A_m can best be obtained by measuring the fractional amplitude modulation at Ω on the probe signal. The phase chosen in this paper for the A_m term is the most likely and gives the largest error. A_m also depends on the modulation width B.

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

The errors associated with D_{2c} are unique to the fundamental demodulator systems and can be illuminated by varying T_2. For T_2 small this error can be totally neglected.

The errors originating from K_{dc} are best isolated 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 offsets is to vary the ac gain with no modulation on the probe and measure the dc error signal.

Conclusion

A simple model of a Lorentzian resonance system on a nonflat background probed by a sinusoidally modulated probe signal has been treated to expose the first-order errors in determining line center, including imperfections in the electronics. Although this approach does not produce rigorous values for the frequency errors in that it does not take into account saturation, etc., or large background distortions, it does yield the correct functional dependence of the errors on modulation index, ac gain, etc. This model permits one to compare the offsets in determining line center using various servo configurations. We have shown that in any servo system with a sine-wave demodulator reference the most serious frequency errors originate from sloping background, second-harmonic distortion in the frequency modulation, amplitude modulation on the probe signal, and dc offsets in the demodulator. Servo systems utilizing the third 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 second-harmonic distortion in the modulator and dc offsets in the demodulator. With the functional dependence outlined here it is relatively easy to design sensitive tests of these offsets, even if the original assumptions about a pure Lorentzian line and small modulation index are not exactly fulfilled.

Acknowledgment

The author is very grateful to Dr. J. Bergquist and Andrea DeMarchi for fruitful discussions.

References

[1] H. Wahlquist, "Modulation broadening of unsaturated lines," *J. Chem. Phys.*, vol. 35, pp. 1708-1710, 1981.

[2] F. M. Gardner, *Phaselock Techniques*. New York: Wiley, 1966.

[3] G. Becker, "Performance of the primary Cs-standard of the Physikalish-Technische Bundesanstalt," *Metrologia*, vol. 13, pp. 99-104, 1977.

[4] A. L. Schawlow, "Laser spectroscopy of atoms and molecules," *Science*, vol. 202, pp. 141-147, 1978.

[5] J. L. Hall, "Stabilized lasers and precision measurements," *Science*, vol. 202, pp. 147-156, 1978.

[6] V. S. Letokhov and V. P. Chebotayev, *Nonlinear Laser Spectroscopy*. Heidelberg, Germany: Springer-Verlag, 1977.

[7] A. DeMarchi, G. D. Rovera, and A. Premoli, "Effects of servo loop modulation in atomic beam frequency standards employing a Ramsey cavity," *IEEE Trans. Ultrason. Ferroelec Freq. Contr.*, this issue, pp. 582-591.

Fred L. Walls was born in Portland, OR, on October 29, 1940. He received the B.S., M.S., and Ph.D. degrees in physics from the University of Washington, Seattle, in 1958, 1964, and 1970, respectively.

From 1970 to 1973 he was a Postdoctoral Fellow at the Joint Institute for Laboratory Astrophysics in Boulder, CO, engaging in research on ion storage techniques and applications. Since 1972 he has been a Staff Member of the Time and Frequency Division of the National Bureau of Standards in Boulder. He is presently engaged in research and development of hydrogen maser devices and cesium beam standards for use as ultrastable clocks, crystal-controlled oscillators with improved short- and long-term stability, frequency synthesis from RF to infrared, and low-noise frequency stability measurement systems. He has published more than 60 scientific papers.

Dr. Walls is a member of the American Physical Society.