Sensitivity Analysis of Digital Filter Structures

by

Victor Earl DeBrunner

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

APPROVED:

A. A. (Louis) Beex, Chairman

K-B. Yu

D. K. Lindner

May 1986

Blacksburg, Virginia

Sensitivity Analysis of Digital Filter Structures

by

Victor Earl DeBrunner

A. A. (Louis) Beex, Chairman

(ABSTRACT)

A coefficient sensitivity measure for state space recursive, finite wordlength, digital filters is developed and its relationship to the filter output quantization noise power is derived. The sensitivity measure is simply the sum of the L_2 norm of all first order partials of the system function with respect to the system parameters; alternatively, the measure may be viewed as the output variance of the error system created by the inherent parameter quantization. Since the measure uses only the first order partials, it is a lower bound approximation to the output quantization noise power. During analysis, numerically unstable conditions may occur because ideal filter characteristics imply system poles which are almost on the unit circle in the z-plane; therefore, it is proposed to scale the radii of the pole and zero magnitudes. Thus, the scaled system has the same frequency information as the original system, but performs better numerically. The direct II form sensitivity, which is shown to be inversely proportional to the product of the system pole and zero distances, can be reduced by the judicious placement of added pole/zero cancellation pairs which increase the order of the system but do not change the system function.

Acknowledgements

I would like to thank Dr. A. A. (Louis) Beex for all of his input and help in getting this work done; he provided the needed motivation in a manner which helped me get the most out of this project. I would also like to thank Dr. D. K. Lindner for his constructive criticisms and Dr. K-B. Yu for his time and effort. Finally, I would like to thank my wife Linda for her support and her review of my grammar and word usage.

Table of Contents

List of Illustrations

1.0 Introduction

Recently, much effort has been concentrated on the development of filters with low, or minimum, output quantization noise power. L. B. Jackson [11], M. Kawamata and T. Higuchi [15], V. Tavsanoglu and L. Thiele [23] and D. V. B. Rao [22] have all noted the relationship between coefficient sensitivity and output quantization noise power; they are directly related and the minimization of one implies the minimum of the other. Mullis and Roberts [19] and Hwang [10] developed, by different methods, the theoretical aspects of minimum noise filters as well as the practical computation of this optimal form. Further, recognizing that this optimal form has, in general, a full state space description, Mullis and Roberts developed a block-optimal form which is near optimal and has approximately 4n coefficients instead of the approximately $n(n+2)$ coefficients of the optimal form. The block-optimal form has sub-filters which have been cascaded or placed in parallel.

Later, L. B. Jackson, A. G. Lindgren and Y. Kim [12] developed a set of design equations for optimal second-order filter sections. Easily computed, this section-optimal filter type is identical to the block-optimal form for parallel sub-filters, while for cascaded sub-filters the section-optimal form is less optimal than the block-optimal form. However, its performance is still quite good in most cases. Continuing this process, B. W. Bomar and J. C. Hung [2] and B. W. Bomar [3,4] have developed near optimal second-order structures with constraints placed on the coefficients to further

reduce the total number of coefficients in the system description. These constraints force some of the coefficients to be structural ones and zeros, while others are forced to be exact powers of two (thus making multiplications equivalent to shifts of the binary point).

From an alternative viewpoint, several researchers have devised design methods for filters using structures with known low sensitivity properties. Among these are the wave digital filters of Fettweis [8] and their special case, the wave lattice digital filters [9]. Constantinides [5,6] noted the applicability of using all-pass functions to implement these filter types. Then, realizing the low pass-band sensitivity of these forms, P. P. Vaidyanathan, S. K. Mitra and Y. Neuvo [24] developed a synthesis approach suitable for the design of low-pass digital filters which have low output quantization noise power in the filter pass-band; however, the stop-band sensitivity may be extremely poor. This form requires about the same number of parameters as the direct II state space form but has lower output quantization noise power, although the noise power is not optimal.

Clearly, much interest exists in the implementation of reduced sensitivity filters with the constraints that this lower sensitivity filter not have too many added coefficients and that it is reasonably easy to compute. This thesis looks at the concept of increasing the order of the system for the purpose of reducing the sensitivity of the filter without changing the input/output relationship of the filter (i.e. adding pole/zero cancellation pairs). This freedom leads to a design methodology for producing low sensitivity filters for low-pass, high-pass, band-pass and band-stop functions consisting of the following steps:

1. Determine the required filter specifications.
2. Find the proper low-pass prototype filter specifications with the constraint that the band-width of the low-pass prototype be the band-width of the desired filter.
3. Find the location of the pole/zero cancellation pair(s) for minimum sensitivity.
4. Frequency translate the low-pass prototype filter to the desired filter function.

Our design methodology is developed along the lines of the following organization. In Chapter 2, the background and interpretation of the sensitivity measure is developed and then the method of its calculation is presented. Chapter 3 contains the description of several state space forms which are analyzed later. The relationship of pole and zero sensitivities to the sensitivity measure of direct II filter implementations is shown in Chapter 4, as well as how to use this knowledge to advantage with respect to low sensitivity designs for low-pass and high-pass filters. In Chapter 5, the effects of frequency transformations of low-pass filters to band-pass and band-stop filters are studied with an eye on sensitivity. We close with summarizing our conclusions and recommendations for further study.

2.0 The Sensitivity Calculation

Since this work deals with roundoff noise of digital filters described in state space form, some necessary background in the mathematics of state space is provided first. Next the sensitivity measure is described, and then the strong relationship between the sensitivity measure and the filter output quantization noise power is shown. Two methods for calculating estimates of the output quantization noise power are derived. Finally the numerical algorithms required in computing both the filter sensitivity measure and the filter output quantization noise power are described; the associated numerical problems are discussed and some solutions to reduce their effect are proposed.

2.1 The State Space Filter Description

As is well known, a digital filter with impulse response h_n and rational transfer function H(z) can be described in the state space form:

$$x_{k+1} = Ax_k + Bu_k \tag{2.1.1}$$

$$y_k = Cx_k + du_k \tag{2.1.2}$$

where x is the state vector, u is the input and y is the output. Note that A is an (nxn) matrix, B is an (nx1) vector, C is a (1xn) vector and d is a scalar. Taking the z transform of (2.1.1) yields

$$zX(z) - zx(0) = AX(z) + BU(z) \tag{2.1.3}$$

Rearranging terms gives

$$X(z) = (zI - A)^{-1}zx(0) + (zI - A)^{-1}BU(z) \tag{2.1.4}$$

For causal systems initially at rest, x(0) is 0 and (2.1.4) becomes

$$X(z) = (zI - A)^{-1}BU(z) \tag{2.1.5}$$

And substituting equation (2.1.5) into the z transform of (2.1.2) yields

$$Y(z) = [C(zI - A)^{-1}B + d]U(z) \tag{2.1.6}$$

and thus,

$$H(z) = \frac{Y(z)}{U(z)} = C(zI - A)^{-1}B + d \tag{2.1.7}$$

It is well known that the state space representation {A, B, C, d} is not unique. This property can be seen by defining a new state vector, $\tilde{x}_k = T^{-1}x_k$, where T is any nonsingular (nxn) matrix. Substitution into equations (2.1.1) and (2.1.2) yields

$$T^{-1}(T\tilde{x}_{k+1} = AT\tilde{x}_k + Bu_k) \tag{2.1.8}$$

$$y_k = CT\tilde{x}_k + du_k \tag{2.1.9}$$

which reduces to

$$\tilde{x}_{k+1} = T^{-1}AT\tilde{x}_k + T^{-1}Bu_k \tag{2.1.10}$$

$$y_k = C\tilde{T}x_k + du_k \tag{2.1.11}$$

with the algebraically equivalent state space description $\{T^{-1}AT, T^{-1}B, CT, \mathrm{d}\}$.

2.2 *The Sensitivity Measure*

Two different interpretations, one deterministic and the other probabilistic, exist for determining the sensitivity measure. In the deterministic view, the classic linearization procedure is used to approximate the non-linear quantization effects. In the probabilistic view, the non-linear quantization effects are modeled by injected noise sources. Both of these interpretations have merit and since they both generate the same final sensitivity measure, they lend credence to each other.

First, we examine the deterministic case. The filter H(z) is a function of the parameter set $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_l]$, where both l and γ depend on the particular implementation used. The set γ is the quantization of the set γ_∞, which is the set of ideal coefficients. If we expand the filter using a Taylor series around the ideal filter, the actual filter H(z) which is implemented can be represented, as in Figure 1 on page 7, by the parallel combination of the ideal transfer function $H_\infty(z)$ described by γ_∞ and the error or stray transfer function $H_{\text{stray}}(z)$.

Considering only the first-order terms by truncating the higher order terms of $H_{\text{stray}}(z)$ (i.e. linearizing around the ideal transfer function) gives

$$\begin{aligned} H(z) &\cong H_\infty(z) + \frac{\partial H^t(z)}{\partial \gamma}\Big|_{\gamma_\infty} \delta\gamma \\ &\cong H_\infty(z) + \frac{\partial H^t(z;\gamma_\infty)}{\partial \gamma}\,\delta\gamma \end{aligned} \tag{2.2.1}$$

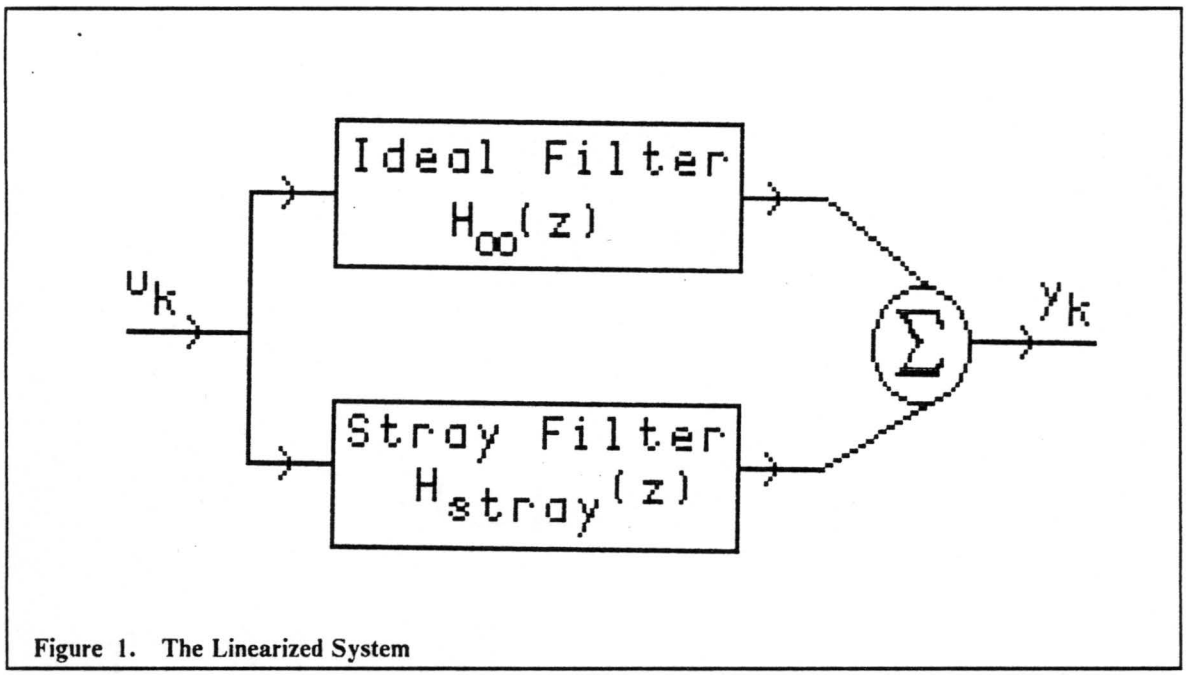

Figure 1. The Linearized System

where

$$\frac{\partial H^t(z)}{\partial \gamma} = \left[\frac{\partial H(z)}{\partial \gamma_1}, \frac{\partial H(z)}{\partial \gamma_2}, \ldots, \frac{\partial H(z)}{\partial \gamma_l} \right] \tag{2.2.2}$$

The L_2 norm, as described first by V. Tavsanoglu and L. Thiele [23] and later by Rao [22], yields a sensitivity measure; the square of the L_2 norm of the error (stray filter) is given by[1]

$$\frac{1}{2\pi j} \int \left| \frac{\partial H^t(z; \gamma_\infty)}{\partial \gamma} \delta\gamma \right|^2 \frac{dz}{z} \leq ||\delta\gamma||_2^2 \sum_{i=1}^{l} \frac{1}{2\pi j} \int \left| \frac{\partial H(z; \gamma_\infty)}{\partial \gamma_i} \right|^2 \frac{dz}{z} \tag{2.2.3}$$

From the probabilistic viewpoint, the exact nature of the quantization effects is uncertain, which leads to the statistical model of Figure 2 on page 8. Note that the quantized branch is modeled as the ideal branch with a quantization noise term added. This added quantization noise is such

---

[1] Note that in this, and all other integrations in this work, $\int$ denotes contour integration along the unit circle of the z-plane in the counterclockwise direction.

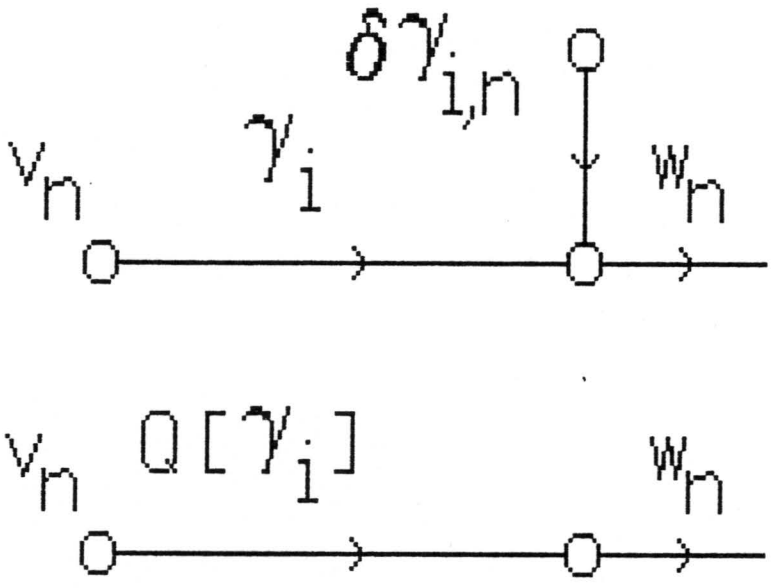

Figure 2. The Probabilistic Model

that for the same input signal both branch models have the same output signal. The quantization noise terms are modeled using the following standard assumptions [21]:

1. The sequence $\{\delta\gamma_{i,n}\}$ is a white noise process.
2. The error sequences are uncorrelated with the other error sequences.
3. The error sequences are uncorrelated with the input v_n.
4. The probability density function of the error process is uniform over the range of quantization error.

These assumptions lead to a linear probabilistic model for coefficient quantization. Heuristically, the model is valid when the input signal is sufficiently complex and the quantization steps are sufficiently small so that the amplitude of the input signal is likely to traverse many quantization levels

from sample to sample. This model is supported empirically [21], where speech signals quantized to as low as eight bits exhibited the properties of the above assumptions.

The use of the above probabilistic model leads to the following state space descriptions for the effect of quantizing single parameter branches:

$$H(z) = (C + \delta c_i e_i^t)(zI - A)^{-1}B \tag{2.2.4a}$$

$$H(z) = C(zI - A)^{-1}(B + \delta b_j e_j) \tag{2.2.4b}$$

$$H(z) = C(zI - (A + \delta a_{ij} e_i e_j^t))^{-1}B \tag{2.2.4c}$$

where e_i is the unit length vector with a 1 in the i^{th} position and 0's elsewhere. Note that assumptions 2 and 3 above allow the separation of the errors as described in equations (2.2.4). Clearly, the coefficient quantization errors in the C vector (equation (2.2.4a)) are propagated through the system as

$$\delta c_i e_i^t (zI - A)^{-1}B = \delta c_i \frac{\partial H(z)}{\partial c_i} \tag{2.2.5}$$

(see equation (2.2.12)) while the coefficient quantization errors in the B vector (equation (2.2.4b)) are propagated through the system as

$$C(zI - A)^{-1}\delta b_j e_j = \delta b_j \frac{\partial H(z)}{\partial b_j} \tag{2.2.6}$$

(see equation (2.2.11)). To separate the coefficient quantization errors in the A matrix, we use the Sherman-Morrison formula [25]

$$[(zI - A) - e_i e_j^t \delta a_{ij}]^{-1} = (zI - A)^{-1} + \frac{(zI - A)^{-1} e_i e_j^t (zI - A)^{-1} \delta a_{ij}}{1 - e_j^t (zI - A)^{-1} e_i \delta a_{ij}}$$

Thus the output error is given by

$$\frac{C(zI - A)^{-1}e_i e_j^t (zI - A)^{-1}B\delta a_{ij}}{1 - e_j^t (zI - A)^{-1}e_i \delta a_{ij}}$$

By the assumptions on the quantization, the denominator is very close to one; thus the error term is approximately given by

$$C(zI - A)^{-1}e_i e_j^t (zI - A)^{-1}B\delta a_{ij} = \delta a_{ij} \frac{\partial H(z)}{\partial a_{ij}} \tag{2.2.7}$$

(see equation (2.2.13)). Thus, we finally can describe the system as in Figure 3 on page 11. Taking the mean square value of the error (output noise) terms gives

$$E\left[\frac{1}{2\pi j}\int \left|\frac{\partial H^t(z;\gamma_\infty)}{\partial \gamma}\delta\gamma\right|^2 \frac{dz}{z}\right] = \sigma_o^2 \sum_{i=1}^{l} \frac{1}{2\pi j}\int \left|\frac{\partial H(z;\gamma_\infty)}{\partial \gamma_i}\right|^2 \frac{dz}{z} \tag{2.2.8}$$

where σ_o^2 is the noise variance of a single quantizer of the system. Since the quantization assumed is rounding, $E[\delta\gamma_i] = 0$ and the variance is given by

$$\sigma_o^2 = \frac{2^{-2b}}{12} \tag{2.2.9}$$

where b is the coefficient wordlength in bits. This probabilistic criterion has been used by several researchers to quantify the transfer function degradation caused by finite wordlength effects. Noting the similarity between equations (2.2.3) and (2.2.8), Rao [22] defined the L_2 norm sensitivity measure S_2 as

$$\begin{aligned} S_2 &\equiv \sum_i \frac{1}{2\pi j}\int \frac{\partial H(z)}{\partial \gamma_i}\frac{\partial H(z^{-1})}{\partial \gamma_i}\frac{dz}{z} \\ &= \sum_i \frac{1}{2\pi j}\int \left|\frac{\partial H(z)}{\partial \gamma_i}\right|^2 \frac{dz}{z} \end{aligned} \tag{2.2.10}$$

where the γ_i are the non-structural coefficients (i.e. the coefficients which are $\not\equiv 0$ or $\not\equiv \pm 1$) of the {A, B, C} state space description.

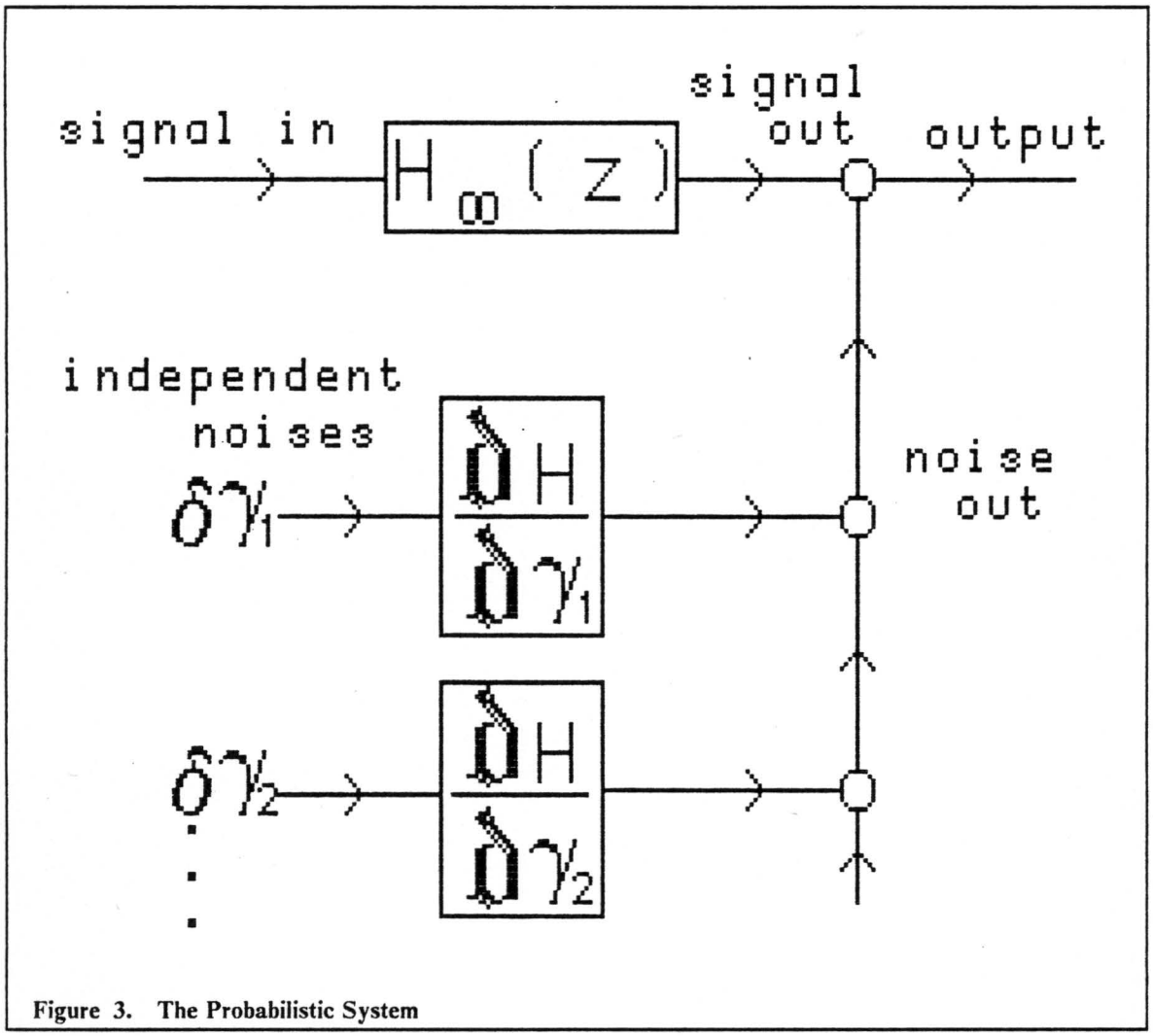

Figure 3. The Probabilistic System

The necessary partial derivatives are now determined. Partial derivatives, with respect to b_i and c_i, of H(z) in equation (2.1.7) directly lead to the first order sensitivity functions:

$$\frac{\partial H(z)}{\partial b_i} = C(zI - A)^{-1}e_i \tag{2.2.11}$$

$$\frac{\partial H(z)}{\partial c_i} = e_i^t(zI - A)^{-1}B \tag{2.2.12}$$

Somewhat more difficult to determine are the sensitivity functions for the coefficients of the A matrix. Using the mathematical identity

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1}$$

gives

$$\frac{\partial H(z)}{\partial a_{ij}} = -C(zI - A)^{-1}\frac{\partial (zI - A)}{\partial a_{ij}}(zI - A)^{-1}B$$

or,

$$\frac{\partial H(z)}{\partial a_{ij}} = C(zI - A)^{-1}e_i e_j^t (zI - A)^{-1}B$$

which is simply,

$$\frac{\partial H(z)}{\partial a_{ij}} = \frac{\partial H(z)}{\partial b_i}\frac{\partial H(z)}{\partial c_j} \tag{2.2.13}$$

Notice that these sensitivity functions are rational functions with the same poles as the original transfer function, H(z); thus, the unit circle is in the region of convergence of the integrand.

The only difference between the sensitivity measure of Tavsanoglu and Thiele [23] and that of Rao [22] is that the former uses the product of the norms when determining the sensitivity of the coefficients of the A matrix (see equation (2.2.13)) while the latter calculates the norm of the product. Both use finite sum approximations of infinite series to evaluate these sensitivities. As described later in this chapter, an exact, closed form solution to the necessary integrations is available, thus the use of norm of the product (i.e. the use of the measure given by Rao) since this is an exact evaluation, whereas the product of the norms by the Schwarz inequality yields only an upper bound. This sensitivity measure is an indication of the frequency response deviation caused by small changes in the system coefficients, i.e. by coefficient quantization.

For further justification of using the S_2 measure as an indication of output quantization noise power, L. B. Jackson [11] has derived roundoff noise bounds from these coefficient sensitivities. Of special interest is the lower bound

$$\sigma_o^2 S_2 \leq \sigma_e^2 \tag{2.2.14}$$

where σ_e^2 is the filter output quantization noise variance. That S_2 is a lower bound for σ_e^2 is also evident from Figure 1 on page 7, remembering that S_2 is the output power of a truncated form of the stray transfer function. Calculating the output variance of $H_{\text{stray}}(z)$ (with all the terms present) gives an infinite sum of auto-covariance terms because all the cross terms go to zero under the assumption that the quantization noise sources are statistically independent from each other and the input signal source. Remember that S_2 is only one of these auto-covariances, although it will be the largest because of the order. The lower bound of equation (2.2.14) was shown empirically by Jackson to be a rather tight bound; thus S_2 is closely related to the output noise power (data is presented later in Chapter 4 to confirm the boundedness). Since one number, i.e. the coefficient sensitivity measure S_2, describes the filter quantization noise power, the problem of identifying low roundoff noise filters is made conceptually easy.

2.3 *The Roundoff Noise Power Calculation*

For completeness, it is necessary to compute an estimate for the roundoff noise power. This calculation is tedious, but the results do corroborate the S_2 measure (as seen in equation (2.2.14)). Hence the calculation of the roundoff noise power is important for verification purposes. Note that we are calculating the true output noise variance of the error transfer function, $H_{\text{stray}}(z)$, which was described in the previous section.

2.3.1 The Error State Space Description

Under finite wordlength conditions, the elements of A, B and C, as well as the scalar d, are constrained and the corresponding state space representation becomes

$$\hat{x}_{k+1} = \hat{A}\hat{x}_k + \hat{B}u_k \tag{2.3.1.1}$$

$$\hat{y}_k = \hat{C}\hat{x}_k + \hat{d}u_k \tag{2.3.1.2}$$

where $\hat{\cdot}$ denotes a quantized entity. Thus the error, $e_k = y_k - \hat{y}_k$, is the difference between the ideal (infinite wordlength) output y_k and the quantized (finite wordlength) output $\hat{y}_k$.

The error state space filter can be constructed as follows:

$$e_k = y_k - \hat{y}_k = Cx_k - \hat{C}\hat{x}_k + (d - \hat{d})u_k$$

or,

$$e_k = [C, -\hat{C}]\begin{bmatrix} x_k \\ \hat{x}_k \end{bmatrix} + (d - \hat{d})u_k \tag{2.3.1.3}$$

with a corresponding system equation:

$$\begin{bmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix}\begin{bmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{bmatrix} + \begin{bmatrix} B \\ \hat{B} \end{bmatrix}u_k \tag{2.3.1.4}$$

Consequently, the error transfer function $H_e(z)$ is given by

$$H_e(z) = [C, -\hat{C}]\{zI - \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix}\}^{-1}\begin{bmatrix} B \\ \hat{B} \end{bmatrix} + (d - \hat{d}) \tag{2.3.1.5}$$

Now the output error variance is

$$\sigma_e^2 = \frac{1}{2\pi j}\int H_e(z)H_e(z^{-1})\frac{dz}{z} \qquad (2.3.1.6)$$

Clearly, the quantizations which occur when forming $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{d}$ in equation (2.3.1.5) depend on the form of the state space realization (i.e. on the form of the A, B, and C matrices). Further, the quantizations determine the exact form of $H_e(z)$. Thus, the state space realization affects σ_e^2 in equation (2.3.1.6) and we expect to be able to classify filter realizations which minimize the output noise power.

2.3.2 Block Diagram View of Output Noise

Alternatively, we may view the output error as the difference in output of H(z) and $\hat{H}(z)$ when driven by the same input (see Figure 4 on page 16). The output noise, $E\{e_k^2\} \equiv \sigma_e^2$, can be readily found as follows:

$$\begin{aligned} E\{e_k^2\} &= E\{(y_k - \hat{y}_k)^2\} \\ &= E\{y_k^2\} + E\{\hat{y}_k^2\} - 2E\{\hat{y}_k y_k\} \end{aligned}$$

or,

$$\sigma_e^2 = \sigma_y{}^2 + \sigma_{\hat{y}}{}^2 - 2\sigma_{y\hat{y}} \qquad (2.3.2.1)$$

This computation is easily performed as follows:

$$\sigma_e^2 = \frac{1}{2\pi j}\int [H(z)H(z^{-1}) + \hat{H}(z)\hat{H}(z^{-1}) - 2H(z)\hat{H}(z^{-1})]\frac{dz}{z} \qquad (2.3.2.2)$$

Thus, σ_e^2 can alternatively be computed using the cross-correlation terms $\sigma_{y\hat{y}}$ together with the two auto-correlation terms $\sigma_y{}^2$ and $\sigma_{\hat{y}}{}^2$

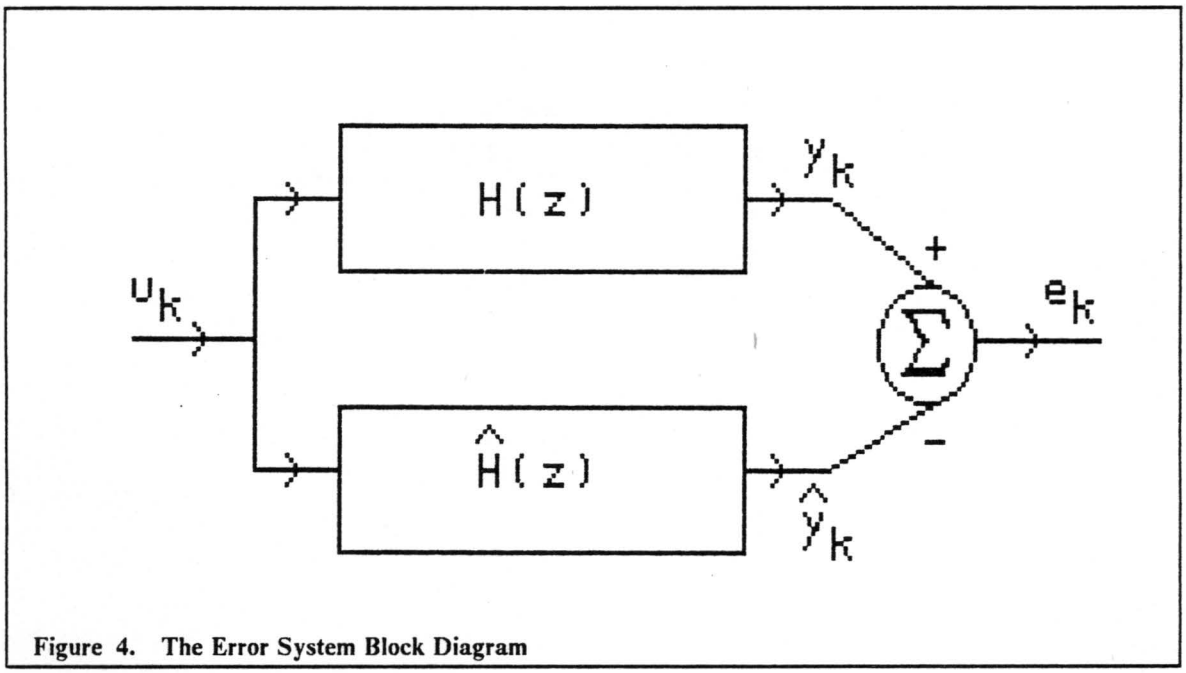

Figure 4. The Error System Block Diagram

2.4 Computing the Sensitivity Measure and Roundoff Noise

The calculations of both S_2 and σ_e^2 require the following two basic steps:

1. Determining the necessary transfer functions.

2. Calculating the variance and cross-covariance terms.

2.4.1 The Subroutine TRAN

The routine for calculating the necessary transfer functions is based on code developed by James Melsa [17] as the routine STVARFDBK. The routines were structured, updated to FORTRAN77 and adapted for the sole use of finding the transfer function of a given state space input--the A matrix, B vector and C vector (Melsa's routine calculates optimal feedback for state variable controllers). The routine uses a similarity transformation to place the state space input into the direct II form, i.e.

$$T^{-1}AT = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}$$

$$T^{-1}B = [0, 0, \ldots, 0, 1]^t$$

and

$$CT = [0, \ldots, 0, c_m, c_{m-1}, \ldots, c_1] \quad ; \quad m \leq n$$

In this form, the transfer function coefficients are directly equal to the difference equation coefficients, giving the transfer function

$$H(z) = \frac{c_1 z^{-1} + c_2 z^{-2} + \cdots + c_m z^{-m}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}} \quad ; \quad m \leq n.$$

The transformation matrix T is generated in two steps. The first is to determine the characteristic polynomial of the A matrix, i.e. the transfer function denominator coefficients. The second is a recursive formulation of the form:

$$t^n = B$$

$$t^{n-i} = At^{n-i+1} + a_i B \quad ; \quad i = 1,2, \ldots ,n-1$$

Then the transformation matrix is

$$T = [t^1|t^2| \ldots |t^n].$$

The transformation matrix can always be calculated as long as the A matrix is nonsingular. This method is amenable to implemention on the computer and yields numerically acceptable results for low system orders according to Melsa[17].

2.4.2 The Subroutine XCOV

The complex integration necessary in computing the variance and cross-covariance terms is performed using an algorithm for the calculation of ARMA cross-covariances presented by A. A. Beex [1]. Note that an auto-covariance is generated if $G(z) = H(z)$ in Figure 5 on page 19. The algorithm requires four steps:

1. Imbed the polynomials

$$\tilde{A}(z) = A(z)C(z)$$
$$\tilde{C}(z) = A(z)C(z)$$
$$\tilde{B}(z) = B(z)C(z)$$
$$\tilde{D}(z) = A(z)D(z)$$

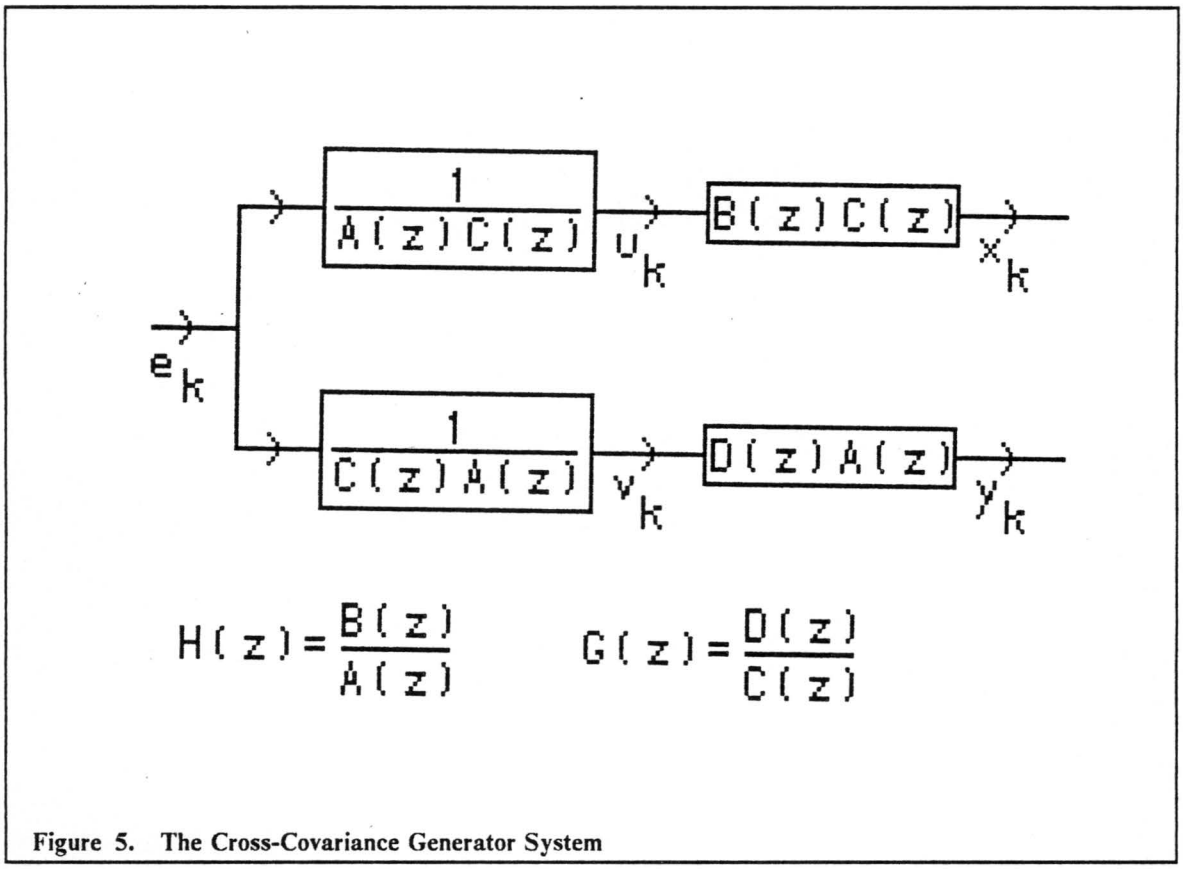

Figure 5. The Cross-Covariance Generator System

2. With $\bar{A}$, use scalar Levinson recursion to generate the auto-covariance of the AR part, checking the magnitude of the reflection coefficients to determine system stability and thus covariance generation sensibility.

3. From $\bar{D}$ and $\bar{B}$ determine $\bar{F}$ where

$$\bar{f} = \bar{d}_{-n} * \bar{b}_n$$

where the $\bar{d}$ are the coefficients of $\bar{D}$ and the $\bar{b}$ and the $\bar{f}$ are similarly defined.

4. Convolve the AR auto-covariance with $\bar{f}$ to get the ARMA cross-covariance sequence R(n).

R(0) is the variance term for auto-covariances and thus is equal to the value of the required contour integral. This algorithm is simple to implement on the computer and yields adequate results except where noted in the following section.

2.4.3 Practical Computation Notes

During the course of implementing and using the above sensitivity measures, two numerical problems in the cross-covariance generator were noted:

1. Poles close to the unit circle may migrate to unstable positions outside the unit circle as a result of creating the higher order (imbedded) polynomials in step 1 of the algorithm.

2. Precision error in the last convolution of the algorithm (step 4) may actually produce a resulting negative auto-covariance, especially when large numbers are alternately added and subtracted!

The first problem may be eliminated by progressively increasing precision since the backward Levinson recursion of step 2 of the algorithm may generate large errors from small errors caused by the polynomial multiplication of step 1. This error was studied in detail by Cybenko [7]. The second problem is more difficult to anticipate and so, as discussed earlier, two methods to determine σ_e^2 were developed. The second method described by equation (2.3.2.2) appears to be numerically superior to that of equation (2.3.1.6), and so is generally preferred in calculating σ_e^2 . This superiority was determined empirically from the example systems of Chapter 4.

It should be noted that these problems show up specifically when designing and analyzing filters approximating ideal characteristics in which poles are located almost on the unit circle. In an effort to alleviate this characteristic, the idea of scaling down the radii of the poles and zeros of the filter was considered. The effect is to replace z^{-1} with rz^{-1}, where r is a scaling constant which is ≤ 1 .

This scaling leaves the original filter frequency information intact, but the scaled filter has much better numerical properties; the poles have some flexibility of movement without throwing the system into instability. The S_2 measure decreases monotonically as the scaling factor decreases (shown in Chapter 4) and preserves the relationship of the sensitivity measure magnitudes between the various implementations of the system. Analysis of the scaled system is then equivalent to analysis of the original system, in a relative sense.

3.0 Description of State Space Filter Implementations

A description of several state space filter implementations is necessary because within the class of state space descriptions it is possible to (almost) continuously vary the filter structure in order to minimize filter sensitivity without changing the system transfer function. Here, several forms of state space digital filters are considered:

1. Direct II
2. Parallel
3. Cascade
4. Optimal
5. Block- and Section-Optimal
6. Dual Generalized Hessenberg Representation

The three primary reasons for interest in these forms are commonality, ease of use and low coefficient sensitivity designs.

3.1 The Direct II, Parallel and Cascade Forms

The direct II, cascade and parallel forms are the most well known forms of recursive state space digital filter implementations because of their direct relationships to the system transfer function, either one to one, factored multiplicatively or factored additively.

3.1.1 The Direct II Form

The direct II form is the easiest canonic state space form to implement. Given the transfer function,

$$H(z) = d + \frac{b_1 z^{-1} + \cdots + b_m z^{-m}}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n^{-n}} \tag{3.1.1.1}$$

the direct II state space coefficients are the a's and the b's, i.e.

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} u_k \tag{3.1.1.2}$$

$$y_k = [0, \ldots, 0, b_m, b_{m-1}, \ldots, b_1]x_k + du_k \qquad (3.1.1.3)$$

3.1.2 The Cascade and Parallel Forms

The cascade form uses a product form of H(z), i.e.

$$H(z) = A \prod_{i=1}^{\left[\frac{n+1}{2}\right]} \frac{\gamma_{0i} + \gamma_{1i}z^{-1} + \gamma_{2i}z^{-2}}{1 - \alpha_{1i}z^{-1} - \alpha_{2i}z^{-2}} \qquad (3.1.2.1)$$

where $\left[\frac{n+1}{2}\right]$ means the largest integer which is $\leq \frac{n+1}{2}$. The parallel form uses a partial fraction expansion of H(z), i.e.

$$H(z) = \sum_{i=1}^{\left[\frac{n+1}{2}\right]} \frac{\gamma_{0i} + \gamma_{1i}z^{-1}}{1 - \alpha_{1i}z^{-1} - \alpha_{2i}z^{-2}} \qquad (3.1.2.2)$$

In each case, the second-order sections are usually implemented in the direct II form, with the corresponding state space description of the system being the combination of these second-order sections. While the parallel form second-order sections are completely decoupled from each other, the cascade form second-order sections are not usually decoupled at all, thereby making the total state space system formulation more complex. Both of these forms require knowledge of pole locations and can be tedious and wasteful to compute when the digital filter is, as is often the case, not given in factored form.

3.2 *The Optimal Form*

Recently the optimal form has received much attention because it minimizes the quantization noise and thus the coefficient sensitivity, producing filters with more robust quantization noise power characteristics. Mullis and Roberts [19] have shown in their paper that the output quantization noise is proportional to the products of the diagonal elements of the controllability grammian matrix, K, and the observability grammian matrix, W, that is,

$$\sigma_e^2 \propto \sum_{i=1}^{n} k_{ii} w_{ii} \tag{3.2.1}$$

The formulation of the transformation matrix required to convert the filter to its optimal form is then a matter of finding a transformation matrix T which will minimize this sum. One formulation for the construction of the transformation matrix T required to convert the direct II state space form to the optimal state space structure is presented by S. Y. Hwang [10]. In the transformation, Hwang forces the controllability grammian diagonals, the k_{ii}, to be equal to 1. Thus, the problem has been reduced to that of minimizing

$$\sum_{i=1}^{n} w_{ii} \tag{3.2.2}$$

which is equivalent to minimizing the trace of W, tr(W). Alternatively, given the non-optimized system grammian W_0, find the transformation matrix T which minimizes

$$\mathrm{tr}(T^t W_0 T) \tag{3.2.3}$$

noting that K_0 and W_0 are the solutions of the following Lyapunov equations:

$$K_0 = AK_0A^t + BB^t \tag{3.2.4}$$

and

$$W_0 = A^t W_0 A + C^t C \tag{3.2.5}$$

K and W satisfy the equivalent Lyapunov equations of the transformed (optimal) system. Interestingly, these Lyapunov equations need not be solved directly. The controllablility grammian, K, is just the covariance matrix given by [19]

$$r_k = \frac{1}{2\pi j}\int |\frac{\partial H(z)}{\partial c_i}|^2 z^{k-1} dz \quad ; \quad k = 0, \ldots, n-1 \tag{3.2.6}$$

where i is any valid subscript (since K is a covariance matrix). The observability grammian, W, can be evaluated as [19]

$$w_{ij} = \frac{1}{2\pi j}\int \frac{\partial H(z)}{\partial b_i} \frac{\partial H(z^{-1})}{\partial b_j} \frac{dz}{z} \quad ; \quad i \le n, \quad j \le i \tag{3.2.7}$$

Note that since W is symmetric, we need only calculate the limits of equation (3.2.7) to identify the complete matrix. Since the diagonal elements of W are variances and the off-diagonal elements are cross-covariances and K is a covariance matrix, both K and W can be solved for via a closed form solution by using the cross-covariance generator described in Chapter 2. This closed form solution is an important result, as the Lyapunov equations need not be directly solved (which is usually done by a finite sum approximation to an infinite sum). Thus, both K and W can be efficiently computed.

The construction requires the solution of three basic matrix equations [10]:

1. Solve for the orthogonal matrix R_0, from

$$R_0(\Lambda^*)^{-2}R_0^t = \begin{bmatrix} 1 & x & x & \dots & x \\ x & 1 & x & \dots & x \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ x & x & \dots & 1 & x \\ x & x & \dots & x & 1 \end{bmatrix} = K \tag{3.2.8}$$

and noting that

$$\Lambda^* = \begin{bmatrix} \lambda_1^* & 0 & 0 & \dots & 0 \\ 0 & \lambda_2^* & 0 & \dots & 0 \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ 0 & 0 & \dots & \lambda_{n-1}^* & 0 \\ 0 & 0 & \dots & 0 & \lambda_n^* \end{bmatrix} \tag{3.2.9}$$

where

$$\lambda_i^* = \left[\frac{\sum_{m=i}^{n} \theta_m}{n\theta_i} \right]^{\frac{1}{2}} \tag{3.2.10}$$

and the θ_i^2 are the eigenvalues of K_0W_0 . Hwang [10] provides an algorithm for determining R_0, which is difficult to compute.

2. Solve for T_0 where

$$T_0T_0^t = K_0 \tag{3.2.11}$$

3. Finally, solve for the orthogonal matrix R_1 where

$$R_1^t T_0^t W_0 T_0 R_1 = \begin{bmatrix} \theta_1^2 & 0 & 0 & \dots & 0 \\ 0 & \theta_2^2 & 0 & \dots & 0 \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ 0 & 0 & 0 & \dots & \theta_2^2 \end{bmatrix} \tag{3.2.12}$$

The transformation matrix, T, is then given by

$$T = T_0 R_1 \Lambda^* R_0^t \tag{3.2.13}$$

and the optimal state space structure obtained is $\{T^{-1}AT, T^{-1}B, CT\}$. Note that the grammians are transformed as $\{T^{-1}K_0T^{-t}, T^tW_0T\}$. The design trade-off is a resulting state space structure which has been completely filled with non-trivial coefficients, requiring the maximum number of multiplies and adds possible for the particular system order. This increased complexity may prohibit the use of the optimal form in some cases, due to time and hardware constraints.

3.3 The Block- and Section-Optimal Forms

The generally increased complexity of the optimal form had already been recognized by Mullis and Roberts [19] and therefore they proposed a block-optimal form which is near optimal in the total system sense. This form is proposed to optimize the sub-sections of a parallel or cascade implementation of the original system. This form has approximately the same low roundoff noise power as the optimal form, but the number of non-trivial coefficients has been reduced from order n^2 to 4n.

L. B. Jackson, A. G. Lindgren and Y. Kim [12] have developed conditions on a second-order state space filter which are necessary and sufficient for an optimal second-order state space filter. From these conditions, a set of design equations are developed and a design method is proposed. This method involves the determination of the parallel or cascade form of the original system and then the application of the design equations to the second-order sub-sections. Using this technique on a parallel form yields a system equivalent to the block-optimal form of Mullis and Roberts [19] since the sections are decoupled; because the sections are constructed separately, a system resulting from a cascaded filter is not block-optimal, hence the term section-optimal. However, examples are provided by Jackson et al. which show that the difference in quantization noise power performance between the section-optimal form and the block-optimal form for the cascaded filter is not significant.

Using arguments similar to those given by Hwang [10], the second-order design equations developed are determined:

1. Given the second-order transfer function

$$H(z) = d + \frac{\gamma_1 z^{-1} + \gamma_2 z^{-2}}{1 + \beta_1 z^{-1} + \beta_2 z^{-2}}$$

which is usually implemented in the direct II form

$$x_{k+1} = \begin{bmatrix} 0 & 1 \\ -\beta_2 & -\beta_1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k$$

$$y_k = [\gamma_2, \gamma_1] x_k + d u_k$$

2. The optimal form can then be constructed by first calculating $\{\hat{A}, \hat{B}, \hat{C}, d\}$ where

$$\begin{aligned} \hat{a}_{11} &= \hat{a}_{22} \\ &= \frac{-\beta_1}{2} \end{aligned} \tag{3.3.1}$$

$$\hat{b}_1 = \frac{1 + \gamma_2}{2} \tag{3.3.2}$$

$$\hat{b}_2 = \frac{\gamma_1}{2} \tag{3.3.3}$$

$$\hat{c}_1 = \frac{\gamma_1}{1 + \gamma_2} \tag{3.3.4}$$

$$\hat{c}_2 = 1 \tag{3.3.5}$$

$$\hat{a}_{12} = \frac{1 + \gamma_2}{\gamma_1^2}\left[(\gamma_2 - \frac{\beta_1\gamma_1}{2}) + (\gamma_2^2 - \gamma_1\gamma_2\beta_1 + \gamma_1^2\beta_2)^{\frac{1}{2}}\right] \tag{3.3.6}$$

and

$$\hat{a}_{21} = \frac{1}{1 + \gamma_2}\left[(\gamma_2 - \frac{\beta_1\gamma_1}{2}) - (\gamma_2^2 - \gamma_1\gamma_2\beta_1 + \gamma_1^2\beta_2)^{\frac{1}{2}}\right] \tag{3.3.7}$$

3. The optimal form is then the scaled network $\{T^{-1}\hat{A}T, T^{-1}\hat{B}, \hat{C}T, d\}$, where the scaling transformation matrix T is

$$T = \begin{bmatrix} t_{11} & 0 \\ 0 & t_{22} \end{bmatrix} \tag{3.3.8}$$

whose diagonal elements are

$$t_{ii} = \left[\frac{1}{2\pi j}\int \frac{\partial H(z)}{\partial c_i}\frac{\partial H(z^{-1})}{\partial c_i}\frac{dz}{z}\right]^{\frac{1}{2}} \quad ; \quad i = 1, 2 \tag{3.3.9}$$

which is the output variance of the stable system with transfer function $\frac{\partial H(z)}{\partial c_i}$ driven by unit variance white noise.

3.4 *The Dual Generalized Hessenberg Representation*

The Dual Generalized Hessenberg Representation or Dual GHR is a form of interest in control applications, Lindner [16]. The state space representation for a single-input single-output system has the following scalar form:

$$x_{k+1} = \begin{bmatrix} a_1 & \pm\gamma_2 & 0 & 0 & \cdots & 0 \\ \gamma_2 & a_2 & \pm\gamma_3 & 0 & \cdots & 0 \\ & & \cdot & & & \\ & & \cdot & & & \\ & & \cdot & & & \\ 0 & \cdots & 0 & \gamma_{n-1} & a_{n-1} & \pm\gamma_n \\ 0 & \cdots & 0 & 0 & \gamma_n & a_n \end{bmatrix} x_k + \begin{bmatrix} \gamma_1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix} u_k \qquad (3.4.1)$$

$$y_k = [\pm\gamma_1,\ 0,\ \cdots,\ 0]x_k + du_k \qquad (3.4.2)$$

Further, the similarity of this canonic form to that of the continued fraction expansion proposed by S. K. Mitra and R. J. Sherwood [18] and outlined by Oppenheim and Schafer [21] was discerned. This form may then be computed in the following manner:

1. Given the system transfer function

$$H(z) = \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 + a_1 z^{-1} + \cdots + a_n z^{-n}} \quad ; \quad m \le n$$

Multiply the numerator and denominator of H(z) by z^n so that the transfer function can be expressed as

$$H(z) = \frac{b_0 z^n + b_1 z^{n-1} + \cdots + b_m z^{n-m}}{z^n + a_1 z^{n-1} + \cdots + a_n}$$

2. Divide the denominator into the numerator to obtain

$$H(z) = A_0 + G_0(z)$$

where $A_0 = b_0$ and

$$G_0(z) = \frac{c_1 z^{n-1} + \cdots + c_m}{z^n + a_1 z^{n-1} + \cdots + a_n}$$

3. If $c_1 \neq 0$ and the numerator is divided into the denominator, $G_0(z)$ can be expressed as

$$G_0(z) = \frac{1}{A_1 + B_1 z + G_1(z)}$$

where $G_1(z)$ has the form

$$G_1(z) = \frac{d_2 z^{n-2} + \cdots + d_n}{c_1 z^{n-1} + \cdots + c_m}$$

4. Repeating the above process of dividing the numerator into the denominator (assuming $d_2 \neq 0$), $G_1(z)$ can be expressed as

$$G_1(z) = \frac{1}{A_2 + B_2 z + G_2(z)}$$

This process is repeated until the numerator is 1.

5. As long as the numerator of $G_k(z)$ has degree $n - k - 1$ and the denominator of $G_k(z)$ has degree $n - k$, H(z) can be expressed as

$$H(z) = A_0 + \cfrac{1}{A_1 + B_1 z + \cfrac{1}{A_2 + B_2 z + \cfrac{\ddots}{\cfrac{1}{A_n + B_n z}}}} \qquad (3.4.3)$$

The second-order section signal flow graphs are shown in Figure 6 on page 34. The Dual GHR canonic form can then be determined by a simple scaling transformation matrix T whose diagonal elements are determined in a manner which produces the form of equations (3.4.1) and (3.4.2). It should be noted that the Dual GHR canonic form can be computed whether the numerators and denominators of the $G_k(z)$ are of the proper order or not, because the more general state space form is only block tri-diagonal. Lindner [16] provides an algorithm which will transform the original state space description directly into the Dual GHR canonic state space form without leaving state space. Thus, no polynomial divisions are required as is the case with the continued fraction expansion algorithm given above. This transformation yields the following canonic form:

$$x_{k+1} = \begin{bmatrix} F_1 & H_2 & 0 & 0 & \dots & 0 \\ G_2 & F_2 & H_3 & 0 & \dots & 0 \\ & & \ddots & & & \\ & & & & & \\ 0 & \dots & 0 & G_{n-1} & F_{n-1} & H_n \\ 0 & \dots & 0 & 0 & G_n & F_n \end{bmatrix} x_k + \begin{bmatrix} G_1 \\ 0 \\ \vdots \\ \\ 0 \\ 0 \end{bmatrix} u_k \qquad (3.4.4)$$

$$y_k = [H_1, \ 0, \ \dots \ , 0]x_k + du_k \qquad (3.4.5)$$

where

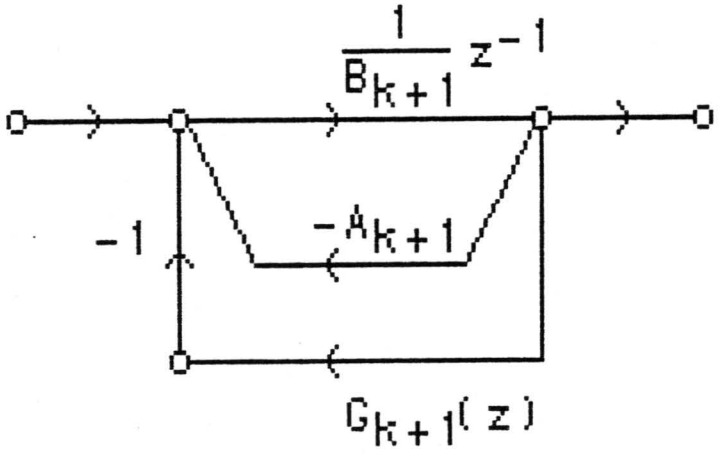

Figure 6. The Block Diagram for the Dual GHR

$$H_k = \begin{bmatrix} \\ \pm \gamma_k \end{bmatrix} \tag{3.4.6}$$

and

$$G_k = \begin{bmatrix} \\ \gamma_k \end{bmatrix} \tag{3.4.7}$$

and

$$F_k = \begin{bmatrix} a_{k1} & 1 & 0 \ldots 0 \\ a_{k2} & 0 & 1 \ldots 0 \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ a_{kp-1} & 0 \ldots 0 & 1 \\ a_{kp} & 0 \ldots 0 & 0 \end{bmatrix} \tag{3.4.8}$$

Note that the F_k are in pseudo-companion form and represent the characteristic equations of the dividend of steps 3 and 4 of the above procedure. This form can always be calculated, as previously noted, and is a canonic form.

4.0 Low-Pass Digital Filters

A commonly used filter design technique is to determine the desired filter characteristics, then translate these characteristics to their corresponding low-pass filter equivalent, and finally design this normalized low-pass filter. The low-pass filter is then frequency transformed back to the desired type, either low-pass, band-pass, band-stop or high-pass filter. Because of this practice, it is logical to first look at low-pass filters and then determine the characteristics related to their sensitivity measure which can be used to advantage.

4.1 Basic Low-Pass Filter Description

Figure 7 on page 37 shows the typical pole placements of a digital low-pass filter. The fact that the poles are clustered near z = 1 inside the unit circle, and have magnitudes which are close to one is of great interest to us. The sensitivity measure, S_2, of a direct II form filter is claimed to be inversely proportional to the system pole distances. This relationship is shown as follows (see Oppenheim and Schafer [21]). Given the ideal system transfer function

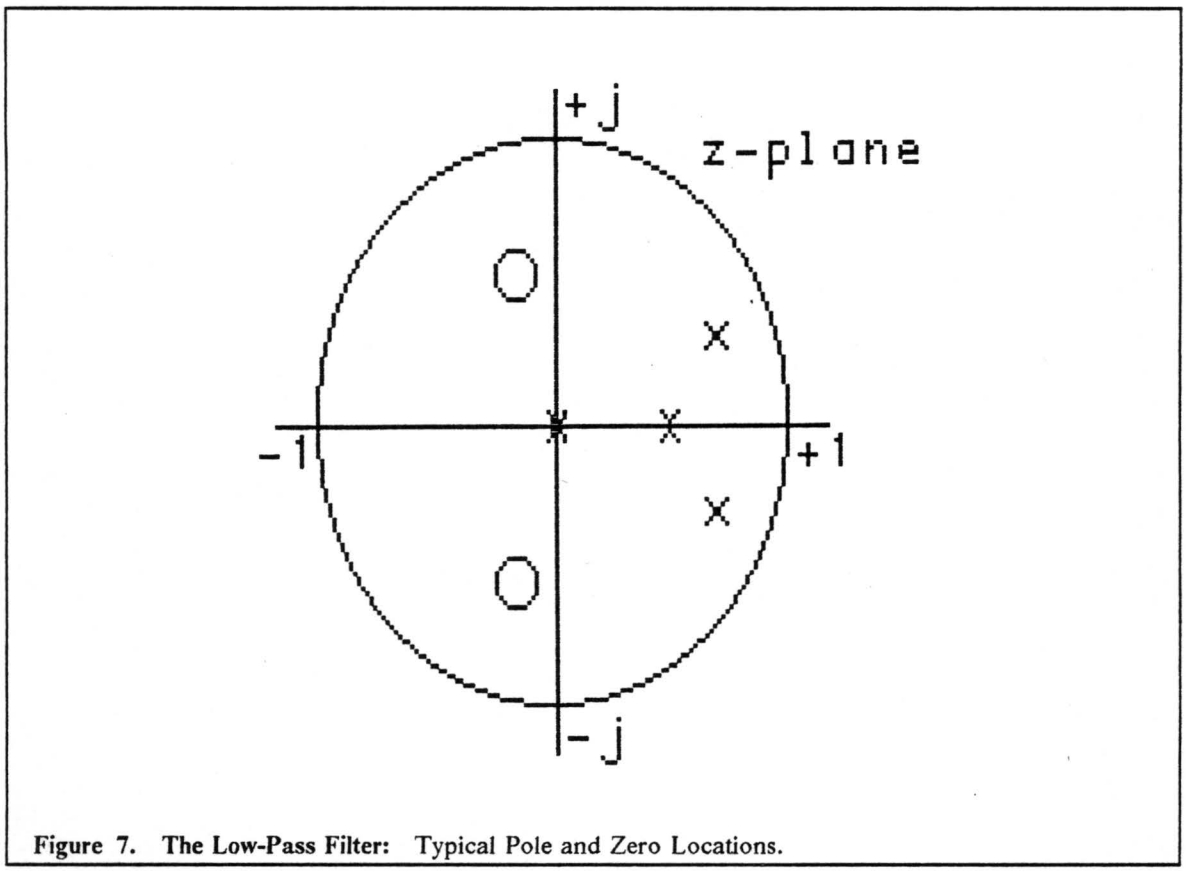

Figure 7. The Low-Pass Filter: Typical Pole and Zero Locations.

$$H(z) = d + \frac{b_1 z^{-1} + b_2 z^{-2} + \cdots + b_m z^{-m}}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} \ldots - \alpha_n^{-n}} \quad ; \quad m \leq n$$

$$= \frac{N(z)}{D(z)} \tag{4.1.1}$$

Express the denominator, D(z), as

$$D(z) = 1 - \sum_{j=1}^{n} \alpha_j z^{-j} = \prod_{j=1}^{n} (1 - p_j z^{-1}) \tag{4.1.2}$$

where the p_j are the simple poles of H(z). From calculus,

$$\frac{\partial H(z)}{\partial p_i}\Big|_{z=p_i} \frac{\partial p_i}{\partial \alpha_j} = \frac{\partial H(z)}{\partial \alpha_j}\Big|_{z=p_i} \tag{4.1.3}$$

which can be rewritten as

$$\frac{\partial p_i}{\partial \alpha_j} = \frac{\frac{\partial H(z)}{\partial \alpha_j}\Big|_{z=p_i}}{\frac{\partial H(z)}{\partial p_i}\Big|_{z=p_i}} \tag{4.1.4}$$

$$= \frac{\frac{N(z)}{D(z)^2}\frac{\partial D(z)}{\partial \alpha_j}\Big|_{z=p_i}}{\frac{N(z)}{D(z)^2}\frac{\partial D(z)}{\partial p_i}\Big|_{z=p_i}}$$

$$= \frac{\frac{\partial D(z)}{\partial \alpha_j}\Big|_{z=p_i}}{\frac{\partial D(z)}{\partial p_i}\Big|_{z=p_i}}$$

Taking the required derivatives using equation (4.1.2), the pole sensitivity can be rewritten as

$$\frac{\partial p_i}{\partial \alpha_j} = \frac{p_i^{n-j}}{\prod_{\substack{l=1 \\ l \neq i}}^{n} (p_i - p_l)} \tag{4.1.5}$$

Similarly, the numerator N(z) can be written

$$N(z) = \sum_{j=0}^{m} b_j z^{-j} = \prod_{j=1}^{m} (1 - z_j z^{-1}) \tag{4.1.6}$$

As for the denominator, the z_j are the simple zeros of H(z). Parallel to equation (4.1.3), the zero sensitivity can be determined from

$$\frac{\partial H(z)}{\partial z_i}\Big|_{z=z_i} \frac{\partial z_i}{\partial b_j} = \frac{\partial H(z)}{\partial b_j}\Big|_{z=z_i} \tag{4.1.7}$$

which can be rewritten as

$$\frac{\partial z_i}{\partial b_j} = \frac{\frac{\partial H(z)}{\partial b_j}\Big|_{z=z_i}}{\frac{\partial H(z)}{\partial z_i}\Big|_{z=z_i}} \tag{4.1.8}$$

$$= \frac{\frac{1}{D(z)}\frac{\partial N(z)}{\partial b_j}\Big|_{z=z_i}}{\frac{1}{D(z)}\frac{\partial N(z)}{\partial z_i}\Big|_{z=z_i}}$$

$$= \frac{\frac{\partial N(z)}{\partial b_j}\Big|_{z=z_i}}{\frac{\partial N(z)}{\partial z_i}\Big|_{z=z_i}}$$

which from equation (4.1.6) reduces to

$$\frac{\partial z_i}{\partial b_j} = \frac{z_i^{m-j}}{\prod_{\substack{l=1 \\ l \neq i}}^{m} (z_i - z_l)} \tag{4.1.9}$$

It is important for us to interpret this latter result in terms of the sensitivity measure S_2. From the definition of the S_2 sensitivity measure, an alternate way of writing S_2 for the direct II state space form is

$$S_2 = \frac{1}{2\pi j}\int \sum_{j=1}^{n} \left|\frac{\partial H(z)}{\partial \alpha_j}\right|^2 + \sum_{i=1}^{m} \left|\frac{\partial H(z)}{\partial b_i}\right|^2 \frac{dz}{z} \tag{4.1.10}$$

Equations (4.1.4) and (4.1.8), along with equation (4.1.10), show that the pole and zero sensitivities are proportional to the S_2 sensitivity measure. Note that

$$\frac{\partial H(z)}{\partial \alpha_k} = -\frac{N(z)}{D^2(z)} z^{-k} \tag{4.1.11}$$

which can be rewritten using equation (4.1.2) as

$$\frac{\partial H(z)}{\partial \alpha_k} = -\frac{N(z)}{D(z)} \frac{z^{n-k}}{\prod_{j=1}^{n} (z - p_j)} \tag{4.1.12}$$

and similarly,

$$\begin{aligned} \frac{\partial H(z)}{\partial b_k} &= \frac{1}{D(z)} z^{-k} \\ &= \frac{N(z)}{D(z)} \frac{z^{n-k}}{\prod_{j=1}^{m} (z - z_j)} \end{aligned} \tag{4.1.13}$$

Both equations (4.1.12) and (4.1.13) are similar to the pole and zero sensitivities of equations (4.1.5) and (4.1.9), with H(z) as a weighting function. Further, the S_2 sensitivity measure is evaluated as the complex contour integral on the unit circle of the z-plane while the pole and zero sensitivities of equations (4.1.5) and (4.1.9) are only point evaluations at the pole and zero locations of the system (i.e. the sensitivity measure is essentially an integration over all z on the unit circle). In practice, this difference is of limited consequence because the partial derivatives of the transfer function appear to approximate delta functions, thus making the integration itself close to a point evaluation. Since the pole and zero sensitivities are inversely proportional to the system pole and zero distances, the S_2 measure is also approximately inversely proportional to the system pole and zero distances. Also, since the sensitivity measure is weighted by the system transfer function, only that output quantization noise power in regions of practical importance (i.e. the noise power in frequencies passed by the filter) are considered.

Using equation (4.1.5), J. F. Kaiser [13,14] showed that small errors in the coefficients can create large pole displacements from the ideal design. Coefficient quantization errors belong to the category of small errors, and therefore one would expect large S_2 sensitivities (and thus large quantization noise power) in narrow bandwidth low-pass filters, where the poles are tightly clustered.

4.2 *Reducing Direct II Form Low-Pass Filter Sensitivities*

Until recently, the principal method of reducing large roundoff noise power has been the classical analog filter design approach of breaking large order filters into cascaded or parallel second-order sections. In this approach, the complex conjugate poles are isolated from each other, and so the error in each pole is independent from its distance to all the other poles in the higher order system, thus reducing the overall system output quantization noise.

However, forms have been developed which minimize the roundoff noise; these were mentioned in the preceding chapter. The cost of this form is increased complexity; the transformation to the optimal form causes the {A,B,C} state space description to be filled with non-trivial coefficients. Also, as noted previously, Mullis and Roberts [19] presented their block-optimal form and Jackson et al. [12] their section-optimal forms which have near-optimal output quantization noise power and reduced complexity.

To illustrate the range of sensitivities for different implementations of the same system function, the third order low-pass filter used by Hwang [10] was examined. The system has transfer function

$$H(z) = \frac{.079306721z^{-1} + .023016947z^{-2} + .0231752363z^{-3}}{1 - 1.974861148z^{-1} + 1.556161235z^{-2} - .4537681314z^{-3}} \tag{4.2.1}$$

The forms and their sensitivities are as follows:

1. The direct II system is

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ .4537681314 & -1.556161235 & 1.974861148 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u_k \tag{4.2.2}$$

$$y_k = [.0231752363 \;\; .023016947 \;\; .079306721] x_k \tag{4.2.3}$$

The sensitivity measure is $S_2 = 93.714442$.

2. The cascade form of the system is

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 \\ -.689750194 & 1.316988002 & 0 \\ -.397527345 & 1.607214942 & .657873146 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u_k \tag{4.2.4}$$

$$y_k = [0 \;\; 0 \;\; .079306721] x_k \tag{4.2.5}$$

The sensitivity measure is $S_2 = 43.511076$.

3. The parallel form of the system is

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 \\ -.689750194 & 1.316988002 & 0 \\ 0 & 0 & .657873146 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u_k \tag{4.2.6}$$

$$y_k = [.262118112 \;\; -.204296974 \;\; .283603691] x_k \tag{4.2.7}$$

The sensitivity measure is $S_2 = 15.698915$.

4. The optimal form of the system is

$$x_{k+1} = \begin{bmatrix} .6672421816 & .0588820057 & .1297010701 \\ .0951152564 & .6488117266 & .5866572779 \\ .089399279 & -.4660588199 & .6588073673 \end{bmatrix} x_k + \begin{bmatrix} .6221731984 \\ -.1549534962 \\ .6111579978 \end{bmatrix} u_k \tag{4.2.8}$$

$$y_k = [.2917887397 \;\; .2806760077 \;\; -.09612048753] x_k \tag{4.2.9}$$

The sensitivity measure is $S_2 = 8.816327$.

5. The block-optimal form of the system is

$$x_{k+1} = \begin{bmatrix} .658494001 & .684463705 & 0 \\ .3742139062 & .658494001 & 0 \\ 0 & 0 & .657873146 \end{bmatrix} x_k + \begin{bmatrix} .312887592 \\ -.652953035 \\ .753128756 \end{bmatrix} u_k \qquad (4.2.10)$$

$$y_k = [\,-.326470236 \;\; .156440787 \;\; .376567338]x_k \qquad (4.2.11)$$

The sensitivity measure is $S_2 = 7.338480$. Note that this sensitivity is lower than the optimal form sensitivity of Hwang [10] in 4 above, probably because of numerical inaccuracies incurred by Hwang when calculating K_0 and W_0. However, the representation is nearly optimal.

6. The section-optimal form of the system is

$$x_{k+1} = \begin{bmatrix} .658494001 & .506281166 & 0 \\ -.5059162 & .658494001 & 0 \\ 1.787303811 & .851076483 & .657873146 \end{bmatrix} x_k + \begin{bmatrix} .338621722 \\ .711122804 \\ .753128756 \end{bmatrix} u_k \qquad (4.2.12)$$

$$y_k = [0 \;\; 0 \;\; .105303004]x_k \qquad (4.2.13)$$

The sensitivity measure is $S_2 = 24.787467$.

7. The Dual GHR form of the system is

$$x_{k+1} = \begin{bmatrix} 2.265088088 & -1.386119935 & 0 \\ 1.386119926 & -.181894411 & -.522032404 \\ 0 & .522032404 & -.108332552 \end{bmatrix} x_k + \begin{bmatrix} .28161449 \\ 0 \\ 0 \end{bmatrix} u_k \qquad (4.2.14)$$

$$y_k = [.28161449 \;\; 0 \;\; 0]x_k \qquad (4.2.15)$$

The sensitivity measure is $S_2 = 155.135468$.

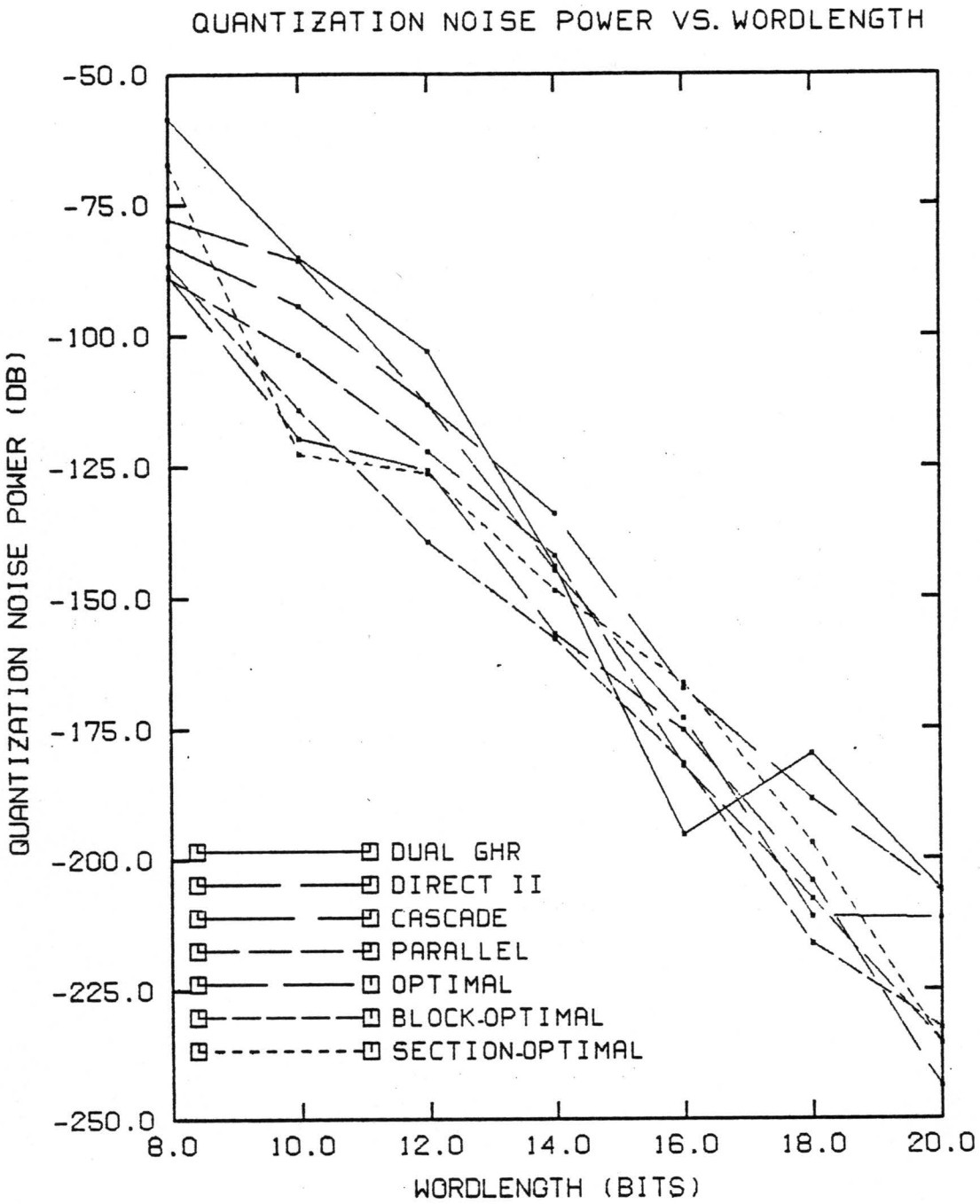

Figure 8. Output Quantization Noise Power

The quantization noise estimates for the above implementations at various wordlengths are shown in Figure 8 on page 44 and the close relationship (refer to equation (2.2.10)) between S_2 and σ_e^2 is shown in Figure 9 on page 45 for the direct II and the optimal forms. Clearly, for this filter, all

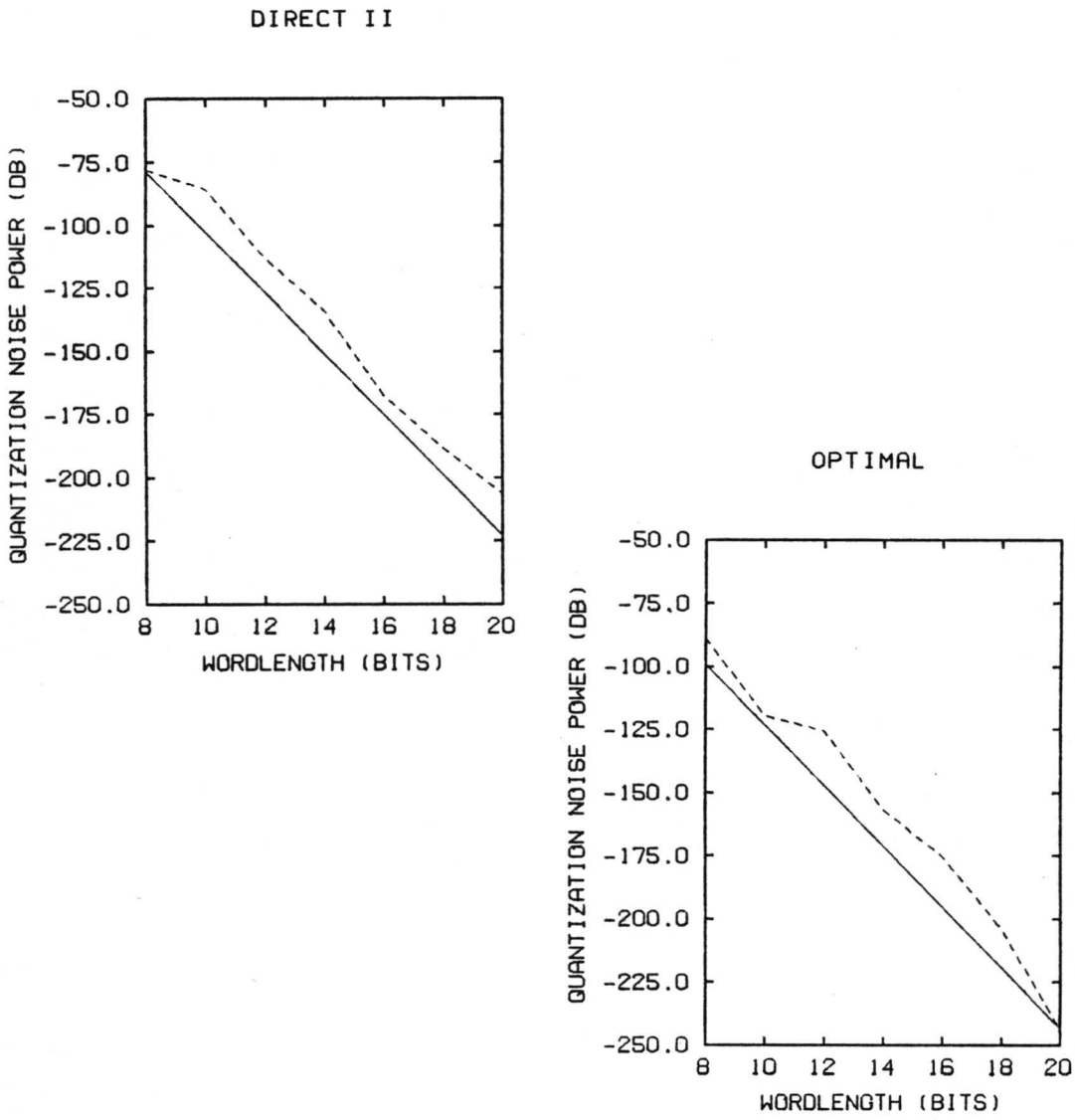

Figure 9. The Sensitivity as a Lower Bound of the Quantization Noise Power

forms are relatively insensitive to coefficient quantization; however the direct II and the Dual GHR forms are an order of magnitude more sensitive than the optimal forms. The result that the Dual GHR implementation is more sensitive than the direct II form is surprising; this high sensitivity may mean that the Dual GHR form could be used to advantage in system identification.

Since the direct II state space form is trivial to compute, the idea of reducing its sensitivity without altering the form is a seemingly attractive idea. Equations (4.1.5) and (4.1.9) suggest a procedure for reducing the direct II form sensitivity by adding poles and zeros; because the transfer function must remain unchanged, the added pole must have a corresponding zero while any added zero should also have its identically related pole. In the case of low-pass filters, all the poles are at low frequencies and so are grouped around $z = 1$ in the z-plane. Clearly, if a pole/zero cancellation pair is added at a high frequency (near $z = -1$ in the z-plane), the sensitivity must be reduced because the added pole distances are greater than one. Also, the newly added sensitivity term (equations (4.1.12) and (4.1.13)) is weighted by H(z) which at high frequencies is close to zero.

To experiment, a real pole/zero pair is added to the direct II form filter of equations (4.2.2) and (4.2.3). A graph of the sensitivity measure as a function of the location of a real pole/zero cancellation pair is given in Figure 10 on page 47, and it reveals a minimum sensitivity comparable to the sensitivity of the cascade realization of equations (4.2.4) and (4.2.5). Note that the plot shows the fourth-order system with the pole/zero cancellation pair at $z = 0$ having a sensitivity of 125, not the 93.7 which one would expect since the filter has the same coefficients as the original third-order system. This increase occurs because the sensitivity was not actually calculated at this point, but merely interpolated. This interpolation does not take into account the fact that the fourth-order direct II realization with the pole/zero cancellation pair added at $z = 0$ has only six non-trivial coefficients in state space, not the eight which the fourth-order filters with all other added pole/zero cancellation pairs require. The added pole/zero pair for minimum sensitivity has increased the system order by one, thus adding two non-trivial coefficients above the number required for the original order direct II model.

As seen above, the sensitivity of the filter has been reduced, but has the transfer function been changed in the process? Ideally, a pole/zero cancellation will leave the system transfer function unchanged, but without infinite precision wordlengths, the impulse response will change, however imperceptibly. Figure 11 on page 48 shows the magnitude characteristic of the original third-order filter and the change which occurs when the filter has the minimum sensitivity real pole/zero added.

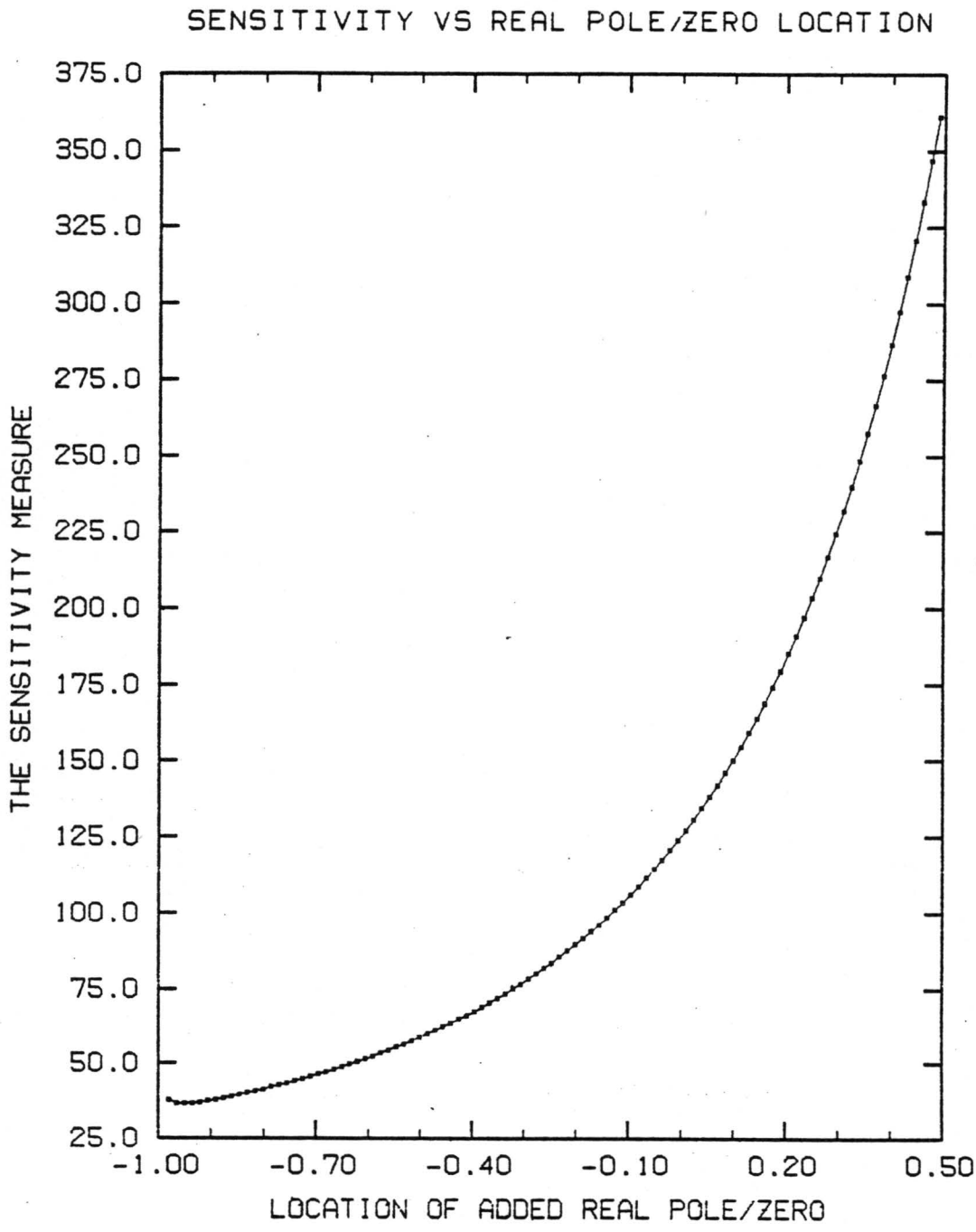

Figure 10. Sensitivity of Fourth-Order Implementations

Figure 12 on page 49 shows the phase characteristic of the original third-order filter and the change which occurs when the filter has the same added pole/zero. Clearly, the system function has not been changed significantly.

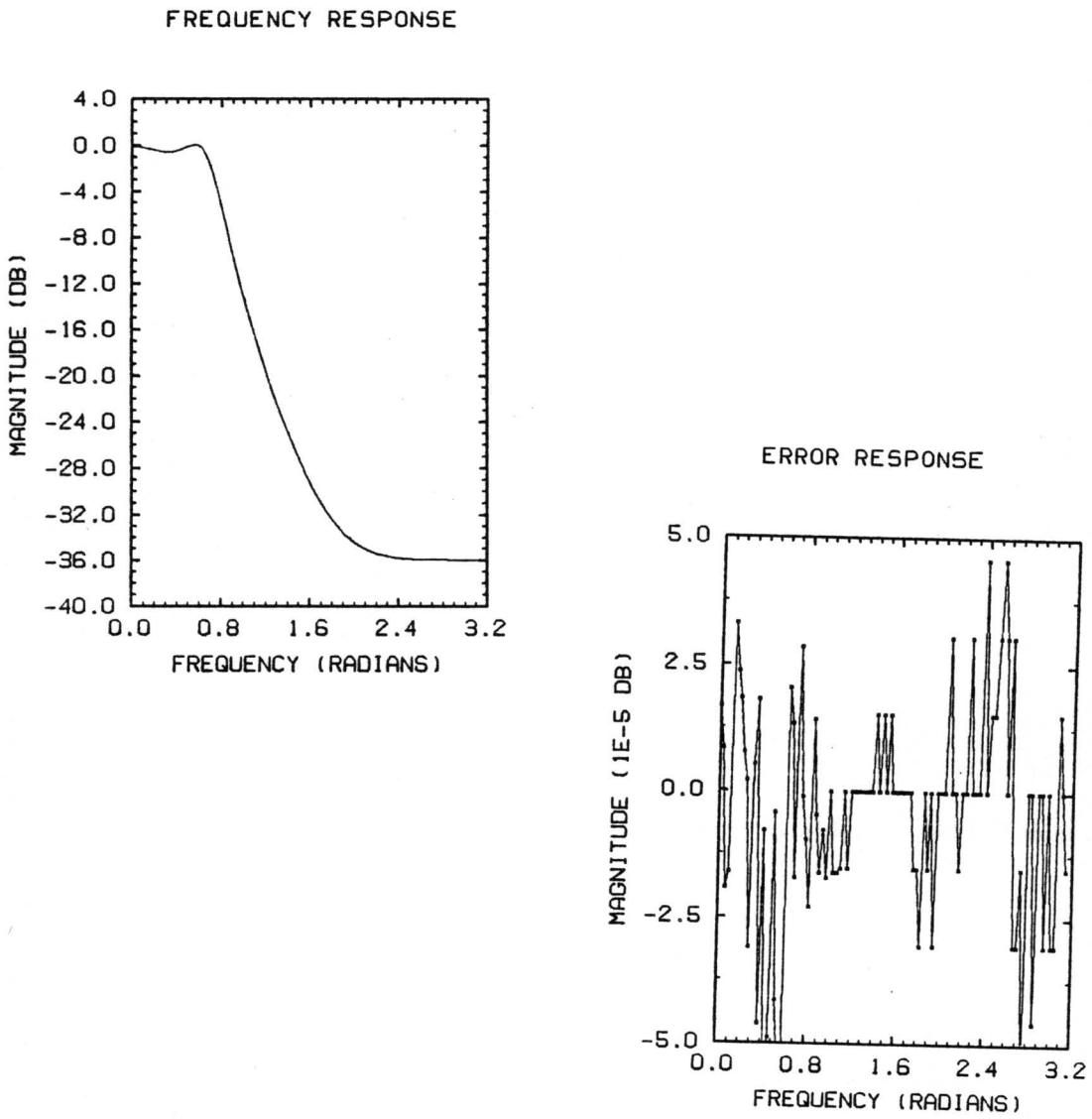

Figure 11. The Magnitude Response of the Fourth-Order System: The Original-Order Filter and the Error of the Fourth-Order, Reduced Sensitivity System.

Further improvement in sensitivity can be realized when a complex conjugate pair of pole/zero cancellations is added. From the sensitivity surface of Figure 13 on page 50, a location in the z-plane is found which has a lower sensitivity than the minimum achieved by adding only a single, real pole/zero cancellation pair. This sensitivity is comparable to the sensitivity of the parallel description given in equations (4.2.6) and (4.2.7), while not quite twice as sensitive as the optimal

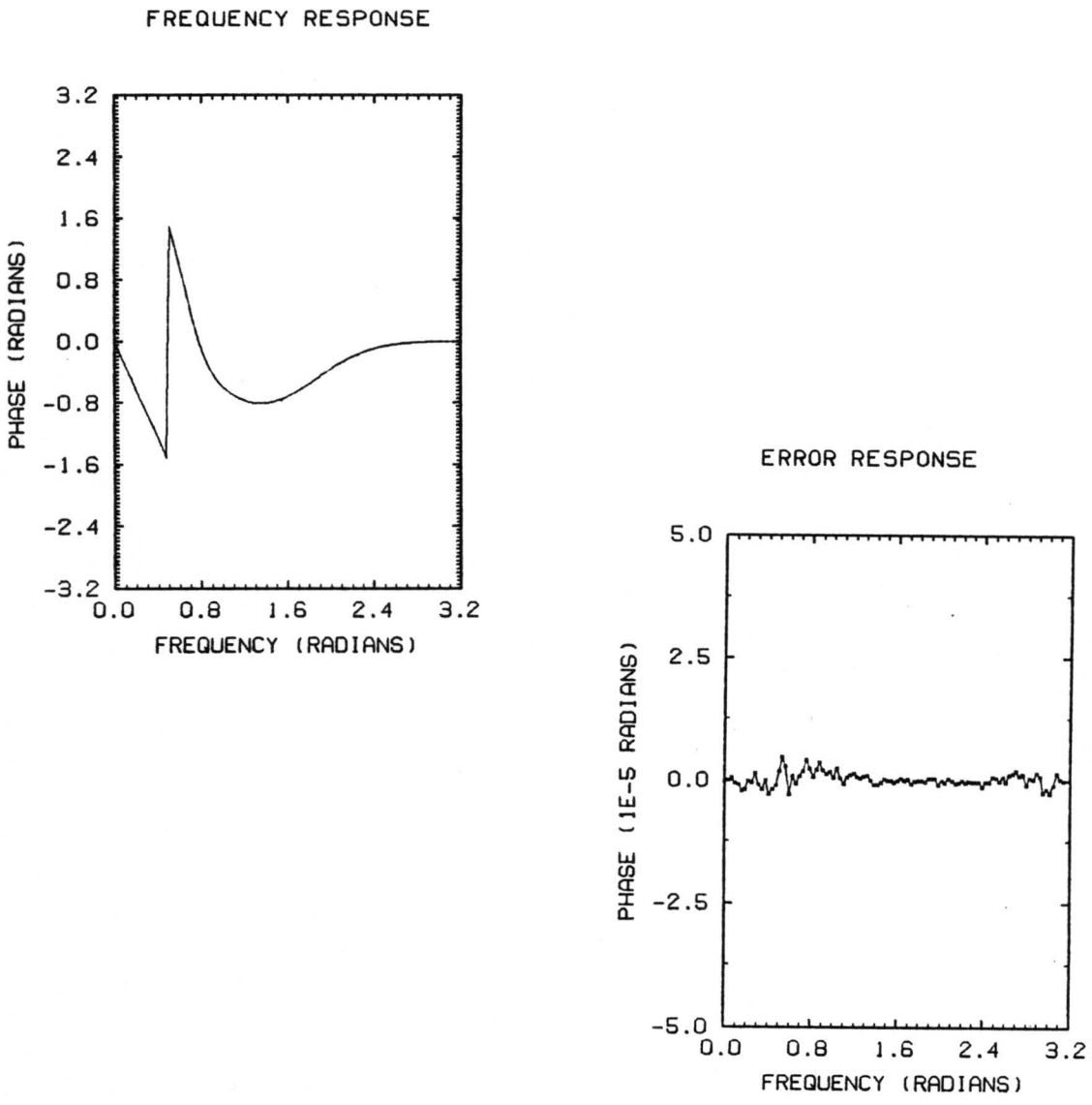

Figure 12. The Phase Response of the Fourth-Order System: The Original-Order Filter and the Error of the Fourth-Order, Reduced Sensitivity System.

form of equations (4.2.8) and (4.2.9). Again, the cost is not too great, only four non-trivial coefficients are added. For comparison of the system complexity, notice that the third-order optimal form has $n(n + 2)$ ($= 15$ for the third-order system) non-trivial coefficients and the block-optimal form has $4n$ ($= 12$), while the fifth-order direct II form has only $2(n + 2)$ ($= 10$). Clearly, as the order of the original system grows, the savings become more important. Again, the transfer func-

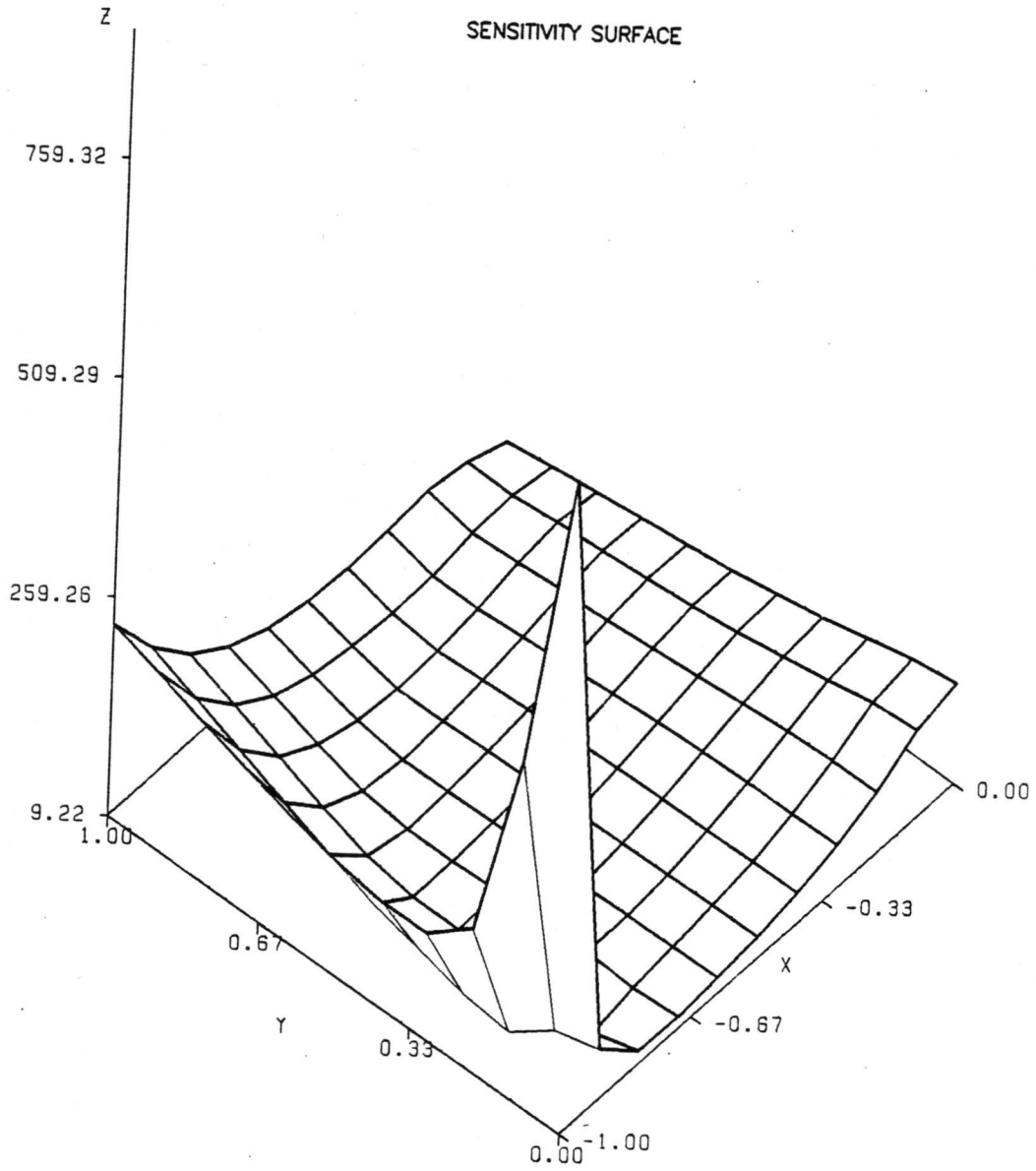

Figure 13. Sensitivity Surface of Fifth-Order Implementations

tion has not been significantly altered. Figure 14 on page 51 shows the magnitude characteristic of the original third-order filter and the change which occurs when the filter has the minimum sensitivity complex pole/zero cancellation pair added. Figure 15 on page 52 shows the phase

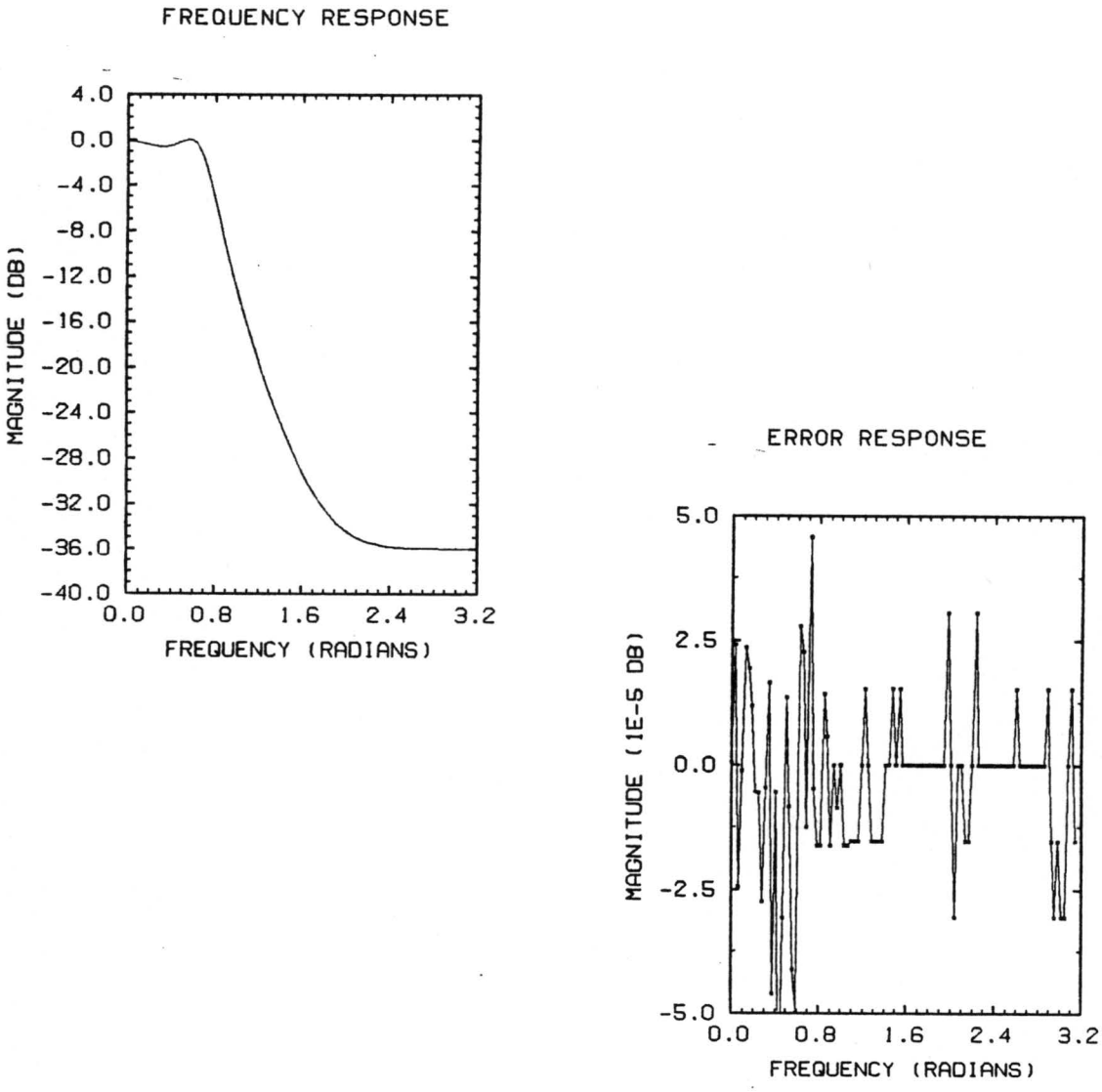

Figure 14. The Magnitude Response of the Fifth-Order System: The Original-Order Filter Response and the Error of the Fifth-Order, Reduced Sensitivity System.

characteristic of the original third-order filter and the change which occurs when the filter has the same added pole/zero cancellation pairs.

To summarize the improvements, Figure 16 on page 53 compares the output quantization noise power of the optimal and various direct II implementations of the third-order filter. Note that the higher-order reduced sensitivity direct II forms have lower output quantization noise power at every

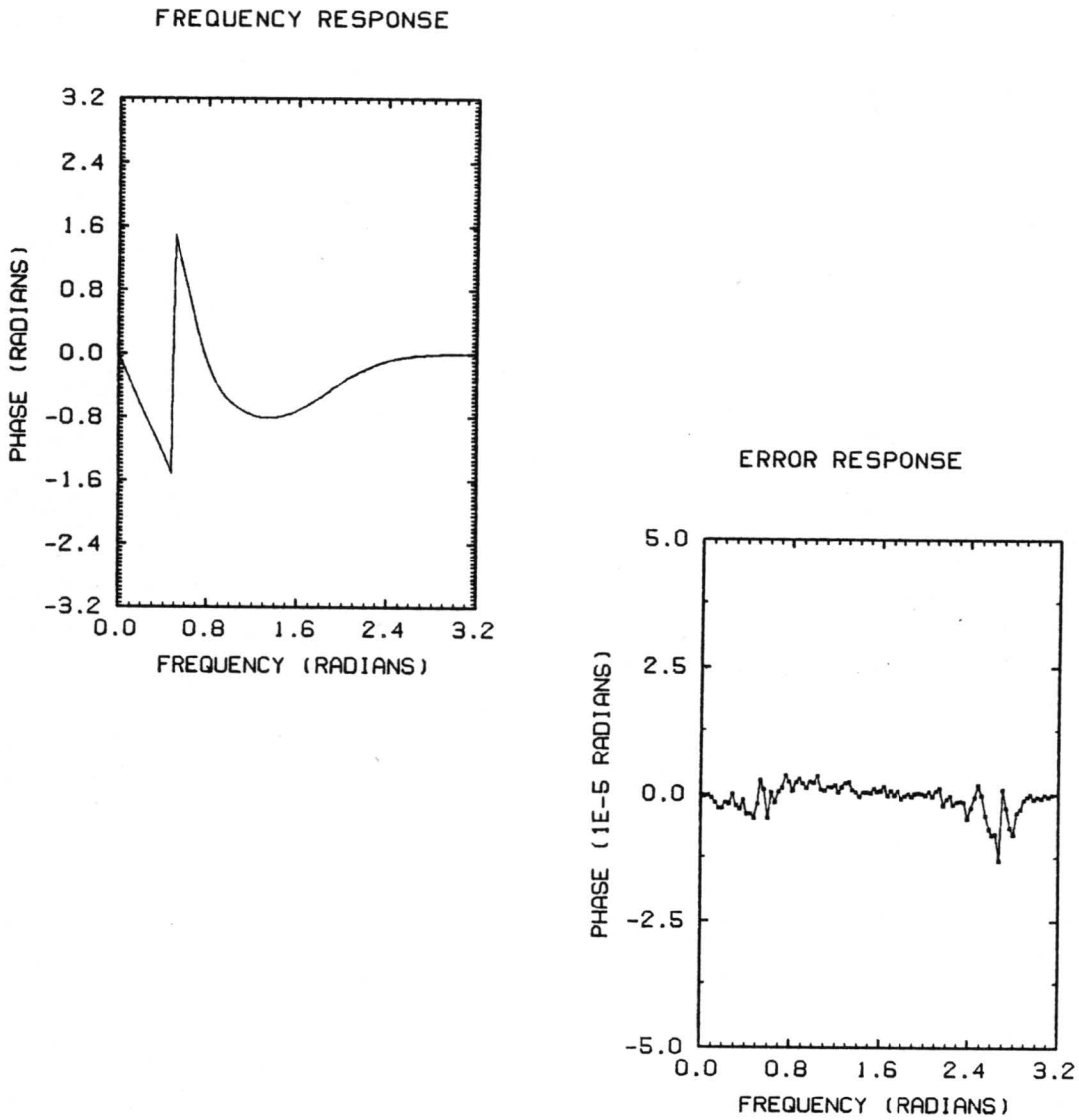

Figure 15. The Phase Response of the Fifth-Order System: The Original-Order Filter Response and the Error of the Fifth-Order, Reduced Sensitivity System.

bit wordlength than does the third-order original direct II filter. Also, both of these higher-order direct II forms actually have lower output quantization noise power than the optimal form at certain wordlengths.

In an effort to help the analysis procedure numerically, the concept of scaling the pole and zero radii magnitudes was introduced in Chapter 2. It was claimed without any supporting evidence or proof

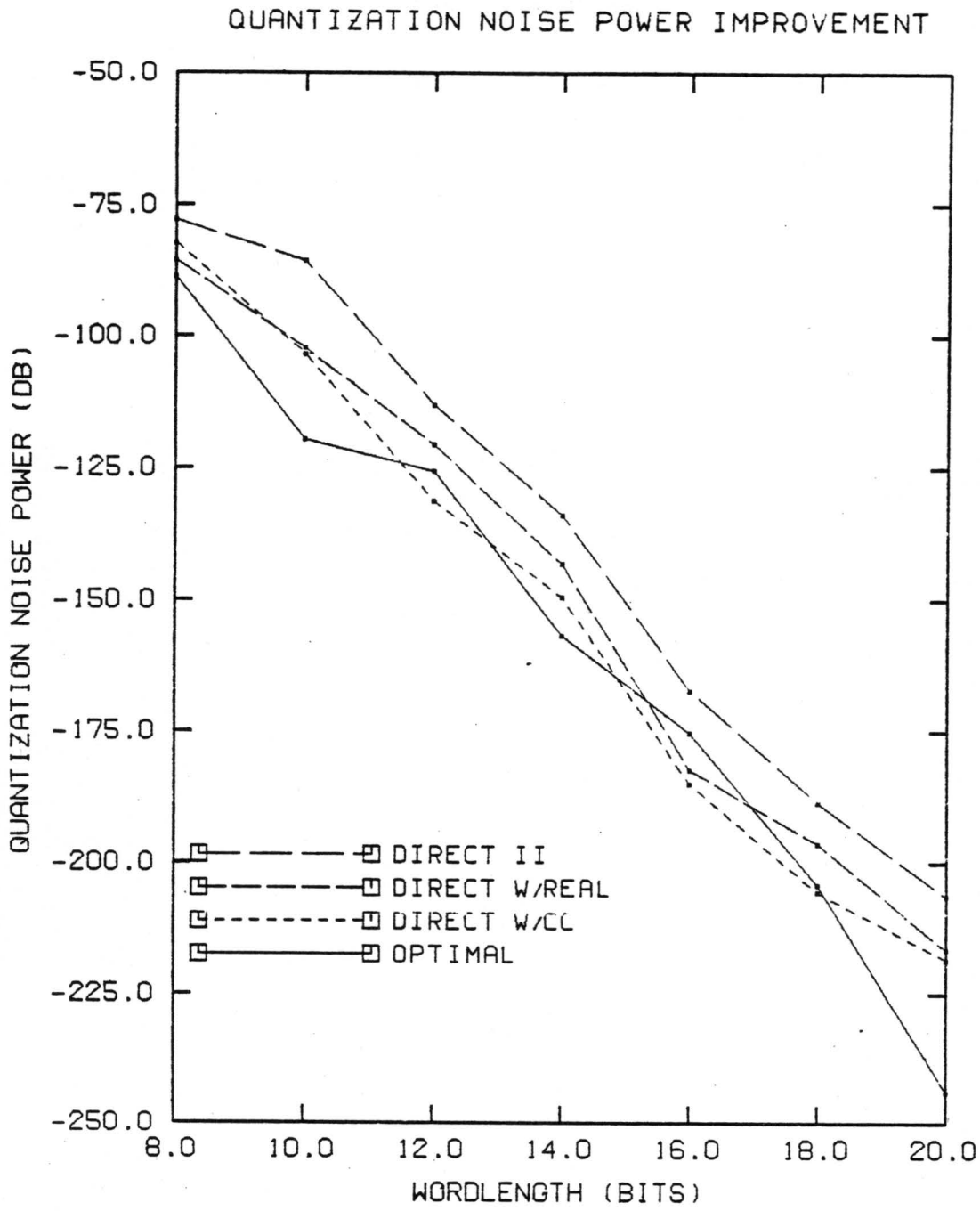

Figure 16. Comparison of the Optimal and Reduced Sensitivity Forms

that analysis could be done on the scaled system which was applicable to the unscaled system. Note that this scaling does not require finding the roots of either the numerator or the denominator, but rather only replacing the numerator and denominator coefficients, the γ_n, with $0.95^n \gamma_n$. Some

corroborating empirical evidence is now offered. Scaling the radii of the poles and zeros by 0.95 produces the following filters:

- The scaled direct II system is

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ .3890494510 & -1.404435515 & 1.876118091 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u_k \qquad (4.2.16)$$

$$y_k = [.0198698670 \;\; .020772794 \;\; .075341384] x_k \qquad (4.2.17)$$

The sensitivity measure is $S_2 = 62.828227$ as compared to the unscaled measure, 93.714442.

- The scaled parallel form of the system is

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 \\ -.622499550 & 1.251138602 & 0 \\ 0 & 0 & .624979488 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u_k \qquad (4.2.18)$$

$$y_k = [.236561596 \;\; -.194082126 \;\; .269423507] x_k \qquad (4.2.19)$$

The sensitivity measure is $S_2 = 11.790138$ as compared to the unscaled measure, 15.698915.

Notice that the relationship between the sensitivities of the filter implementations is maintained after the scaling occurs. When miminizing filter sensitivity by adding a pole/zero cancellation pair, the new, higher-order filter must first be calculated for the unscaled filter and then scaled down, otherwise it is possible to generate a pole/zero location of lowest sensitivity which when unscaled is outside the unit circle, creating an unstable system. Clearly, the scaling is not symmetric with respect to the unit circle! As for the decreasing monotonicity of the sensitivity measurement due to decreasing the scaling factor, see Figure 17 on page 55 which plots the sensitivity measure of the direct II and the parallel forms as a function of the scaling radius.

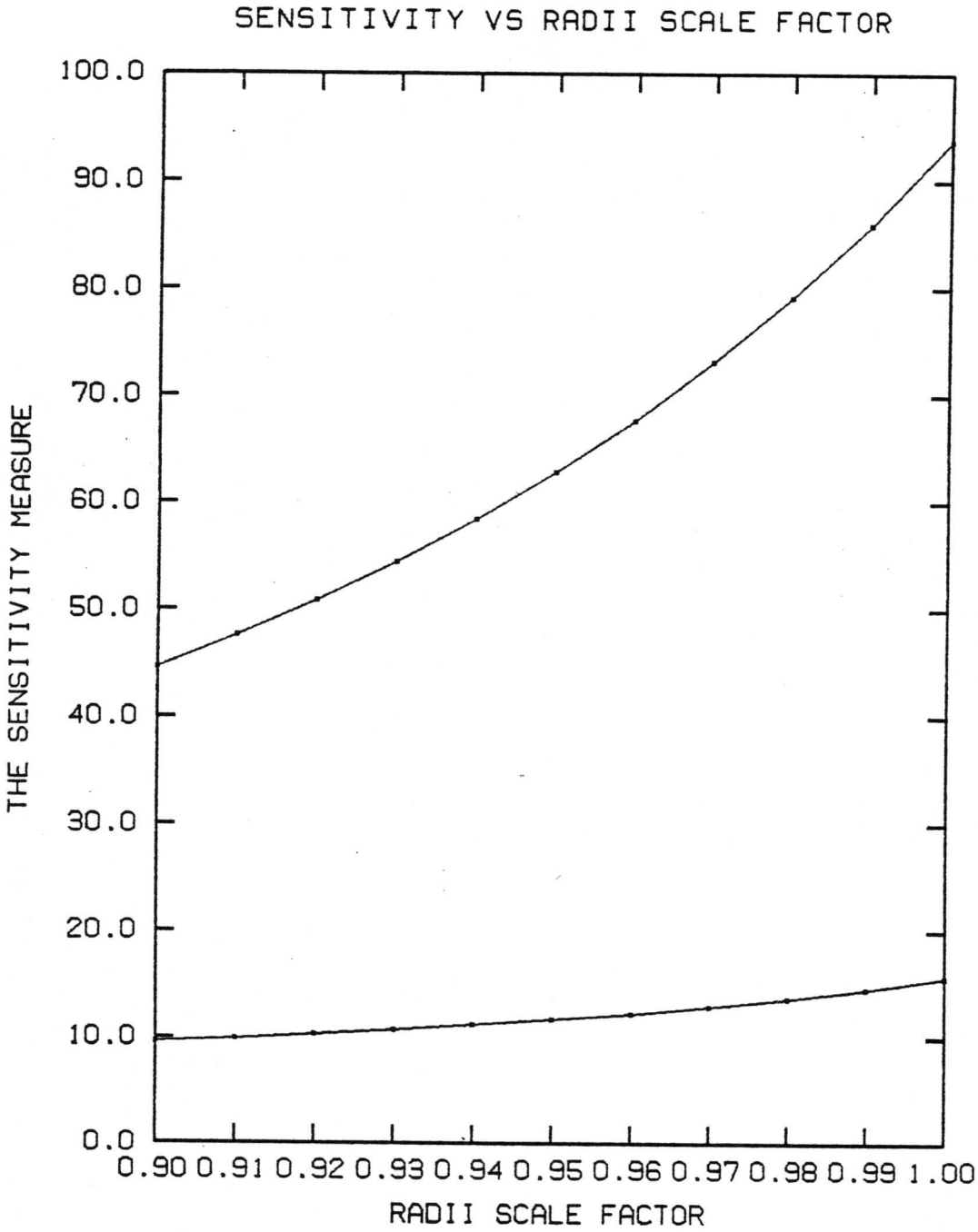

Figure 17. The Monotone Decreasing Sensitivity Measure

It is not clear why the sensitivity measure decreases as a function of the scaling radius, but all experiments demonstrate this characteristic. Several filters were examined specifically to provide a counter-example; all had the monotone decreasing sensitivity measure. Clearly, depending upon

pole and zero locations, the transfer function magnitude response can increase at local frequency ranges; however, the average system response over all frequencies is lower. How this affects the sensitivity measure is not clear, but filters designed expressly to exhibit this local increase over large frequency ranges and which have no single, dominant pole anywhere in the spectrum still exhibit the monotone decreasing characteristic.

To summarize, for the wide range of filters examined, this technique yields stable and consistent results. It is therefore a reasonable tool when used with knowledge of the system and its overall behavior to the scaling of the pole and zero radii. Caution should be exercised, however, until a non-trivial detailed mathematical description is derived for the observed behavior.

To ensure that the pole/zero cancellation will reduce the sensitivity for larger-order systems, a couple of new example filters are introduced. Larger sensitivity reductions are expected because of the pole placements (and thus the pole distances), but the direct II sensitivity will also be much higher because of the higher number of poles and corresponding pole distances which are much less than one.

The first example is a sixth-order Butterworth low-pass filter designed using the impulse invariance technique from A. V. Oppenheim and R. W. Schafer [21]. The transfer function is

$$H(z) = \frac{.00053369z^{-1} + .010300044169z^{-2} + .016007666515z^{-3}}{1 - 3.3634z^{-1} + 5.06810425z^{-2} - 4.2754936z^{-3}} \qquad (4.2.20)$$

$$\frac{+ .004129694086z^{-4} + .000117295287z^{-5} - .000006606018z^{-6}}{+ 2.106377898z^{-4} - .570561512332z^{-5} + .06606018207z^{-6}}$$

The direct II form has $S_2 = 2937.38139$. Placing a complex conjugate pole/zero pair at radius $r = 0.96$ and angle $\theta = \pm 170$ degrees reduces the sensitivity measure to $S_2 = 303.13565$.

The second example filter is a tenth-order all-pole low-pass function. The system transfer function is

$$H(z) = \frac{N(z)}{D(z)} \tag{4.2.21}$$

where

$$N(z) = .211348904z^{-1}$$

and

$$D(z) = 1 - 5.24714092z^{-1}$$
$$+ 14.6742367z^{-2} - 27.2976798z^{-3} + 37.1004172z^{-4}$$
$$- 38.082725z^{-5} + 29.9060915z^{-6} - 17.7209547z^{-7}$$
$$+ 7.66182077z^{-8} - 2.20028154z^{-9} + .339082688z^{-10}$$

The direct II form has $S_2 = 2{,}109{,}022{,}068.714$. Placing a complex conjugate pole/zero pair at radius $r = 0.99$ and angle $\theta = 180$ degrees reduces the sensitivity measure to $S_2 = 199{,}434{,}498.555$.

For narrow-bandwidth low-pass filters, the coefficient sensitivity can also be reduced using this method; however, because coefficient sensitivities of direct II, as well as cascade and parallel, implementations increase as bandwidths decrease under frequency transformations and the optimal form sensitivity is invariant to frequency transformations (Mullis and Roberts [20] and M. Kawamata and T. Higuchi [15]), the reduced sensitivity does not approach sensitivity of the optimal form. For verification, consider the example used by M. Kawamata and T. Higuchi. This example has transfer function

$$H(z) = d + \frac{N(z)}{D(z)} \tag{4.2.22}$$

where

$$d = .00000869$$

$$N(z) = .000627(.108543z^{-1} + .0067z^{-2} + .104730z^{-3} + .00193z^{-4})$$

and

$$D(z) = 1 - 3.826389z^{-1} + 5.516625z^{-2} - 3.551099z^{-3} + .86102z^{-4}$$

This system is an extremely narrow-band filter, as shown in Figure 18 on page 59. The optimal form has sensitivity measure, S_2, equal to 58.327987, as compared to the direct II form, which has a sensitivity of 18,933,029.42. Clearly, the direct II form would not normally be used when accuracy is important, as it is many orders of magnitude (3.25×10^5) more sensitive than the optimal form. Placing a double pole/zero pair at -0.98 on the real axis in the z-plane causes a reduction in coefficient sensitivity of the direct II form to 1,857,725.657534, an improvement of one order of magnitude (10.2). However, the reduced sensitivity is still not close to that of the optimal form; the sensitivity is four orders of magnitude greater.

4.3 Extension to High-Pass Digital Filters

High-pass filters are mirror images of low-pass filters, therefore one would expect to achieve the exact same sensitivity reductions as in the low-pass filter. The symmetry of the two forms is evident from the frequency transformation relationship; z^{-1} of the low-pass form is transformed into $-z^{-1}$ of the high-pass form. For canonic state space descriptions, this transformation results at most in changing the sign of some of the coefficients. This symmetric sign change does not affect the sensitivity. For clarity, the high-pass filter derived from frequency transforming the direct II low-pass filter of equations (4.2.2) and (4.2.3) is

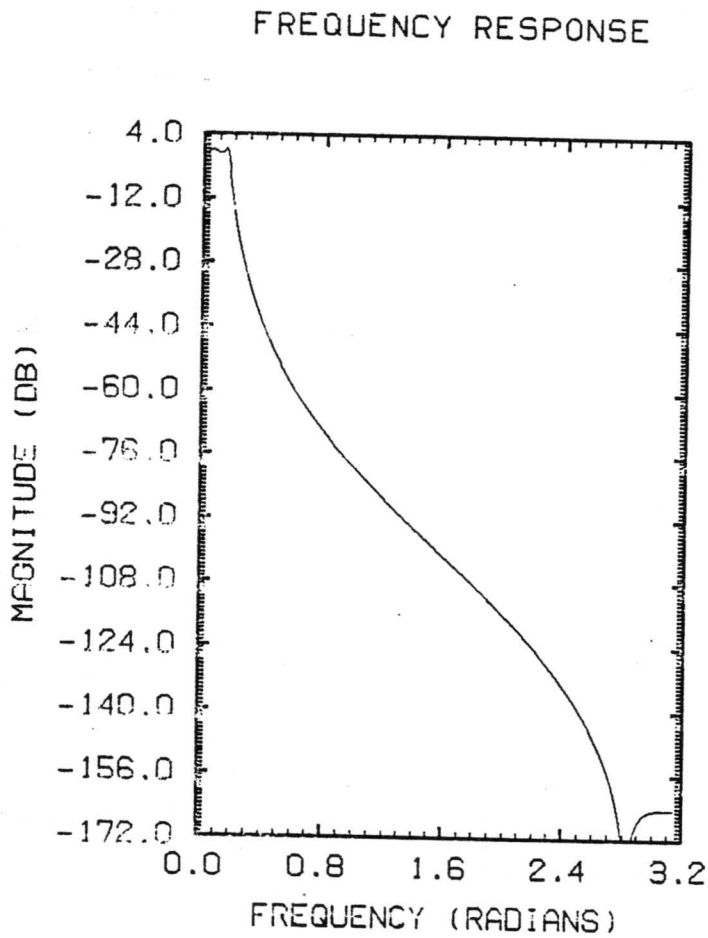

Figure 18. The Narrow Band-Width of the System of Equation (4.2.22)

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -.4537681314 & -1.556161235 & -1.974861148 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u_k \qquad (4.3.1)$$

$$y_k = [\,-.0231752363 \;\; .023016947 \;\; -.079306721]x_k \qquad (4.3.2)$$

The sensitivity measure is $S_2 = 93.714442$, which is identical to the low-pass filter sensitivity, as expected.

The above example has shown that analysis of high-pass filters is identical to analysis of low-pass systems when they are obtained by the transformation of $z^{-1} \rightarrow -z^{-1}$ applied to a low-pass prototype. The symmetrical nature of the two filtering functions allows statements about the sensitivity measurement to be identically applicable to both. This relationship is an important feature, as high-pass filters need not be separately analyzed, but prototype low-pass filters can be analyzed for proper pole/zero cancellation locations of minimum sensitivity, compared to their optimal implementation sensitivities and then transformed to the desired high-pass function only when the filter is actually implemented.

5.0 Band-Pass Digital Filters

The previous chapter shows that adding pole/zero cancellation pairs to low-pass and high-pass direct II form filters can reduce the coefficient sensitivity of the implementation. The reduction in sensitivity was explained in terms of the system pole and zero locations, and, because of the system type, these locations could be exploited to reduce the sensitivity. Logically, one asks, can the pole and zero locations of a band-pass system be exploited in the same manner and with similar results?

5.1 Basic Band-Pass Filter Description

To gain an understanding of where the poles and zeros of a band-pass filter are located, note the following frequency transformation equations relating z^{-1} of a prototype low-pass filter to its bandpass equivalent.

$$z^{-1} \rightarrow -\frac{z^{-2} - \frac{2\alpha k}{k+1}z^{-1} + \frac{k-1}{k+1}}{\frac{k-1}{k+1}z^{-2} - \frac{2\alpha k}{k+1}z^{-1} + 1} \tag{5.1.1}$$

where

$$\alpha = \frac{\cos(\frac{\omega_2 + \omega_1}{2})}{\cos(\frac{\omega_2 - \omega_1}{2})} \tag{5.1.2}$$

and

$$k = \cot(\frac{\omega_2 - \omega_1}{2}) \tan\frac{\theta_p}{2} \tag{5.1.3}$$

Note that θ_p is the cut-off frequency of the low-pass prototype filter, and ω_1 and ω_2 are the desired upper and lower cut-off frequencies of the band-pass filter. Equation (5.1.1) clearly shows that the system order has been doubled. However, more importantly, the poles and zeros have migrated to the middle frequencies, i.e. they are grouped around two locations, namely $\pm j$.

The third-order system of the previous chapter, see equation (4.2.1), was frequency transformed keeping the bandwidth of the low-pass prototype and placing the center frequency at $\frac{\pi}{2}$. These conditions along with equations (5.1.2) and (5.1.3) give $k = 1$ and $\alpha = 0$. Substituting these values into equation (5.1.1) gives the transformation $z^{-1} \rightarrow -z^{-2}$. Applying this transformation yields the band-pass system function

$$H(z) = \frac{-.079306721z^{-2} + .0230169z^{-4} - .0231752363z^{-6}}{1 + 1.974861148z^{-2} + 1.556161235z^{-4} + .4537681314z^{-6}} \tag{5.1.4}$$

The direct form of the filter is then given by

$$x_{k+1} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -.4537681314 & 0 & -1.556161235 & 0 & -1.974861148 & 0 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u_k \quad (5.1.5)$$

$$y_k = [\, -.0231752363 \;\; 0 \;\; .023016947 \;\; 0 \;\; -.079306721 \;\; 0] x_k \quad (5.1.6)$$

This system has the sensitivity, $S_2 = 93.71444$. Note that the sensitivity is the same as for the low-pass prototype since the bandwidth remains unchanged by the frequency transformation and the frequency mapping is linear.

From the symmetry of the band-pass filter pole and zero locations, it is clear that finding locations for placement of pole/zero cancellation pairs which reduce the system sensitivity will be difficult to accomplish. However, if we frequency transform the higher-order, reduced sensitivity low-pass prototype filter to the desired band-pass form, the sensitivity of this function will be lower than the sensitivity of the band-pass implementation derived from the original-order low-pass prototype filter. For example, the system of equation (4.2.1) is implemented with an added real pole/zero cancellation pair at z = -0.95. Thus, the band-pass system function is given by

$$H(z) = \frac{\begin{array}{c} -.079306721z^{-2} + .098358331z^{-4} - .45041335z^{-6} \\ + .022016474z^{-8} \end{array}}{\begin{array}{c} 1 + 1.024861148z^{-2} - .3199568565z^{-4} + -1.024585042z^{-6} \\ -.4310797234z^{-8} \end{array}} \quad (5.1.7)$$

The direct II form sensitivity is 36.60593, which is identical to the low-pass prototype direct II form sensitivity. Note that the pole/zero cancellation pair is translated to $z = \pm\, 0.95^{\frac{1}{2}}$ in the band-pass filter; typical band-pass pole and zero locations are shown in Figure 19 on page 64. Empirically, this implementation gives the lowest sensitivity possible.

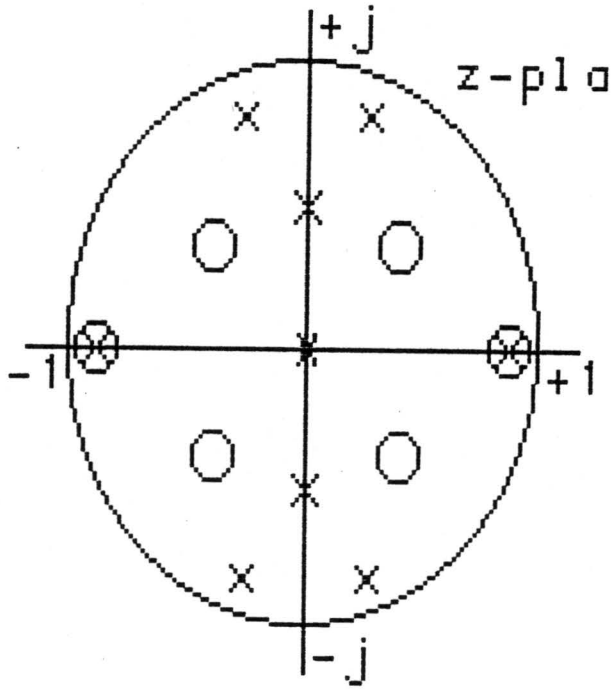

Figure 19. The Band-Pass Filter: Typical Pole and Zero Locations.

If the filter band-width is kept constant but the center frequency of the pass-band is not $\frac{\pi}{2}$ radians, an interesting situation occurs; the sensitivity increases. To investigate, the same low-pass filter as above (equation (4.2.1)) is transformed to a band-pass filter with a center frequency of 1.47 radians. The required frequency translation is $z^{-1} \rightarrow \frac{z^{-2} - .1z^{-1}}{1 - .1z^{-1}}$. The band-pass system is then

$$H(z) = \frac{\begin{array}{c} .0079306721z^{-1} - .08066268z^{-2} + .011337419z^{-3} \\ + .021988961z^{-4} + .004650876z^{-5} - .0231752363z^{-6} \end{array}}{\begin{array}{c} 1 - .497486114z^{-1} + 2.059919983z^{-2} - .711189267z^{-3} \\ + 1.620646115z^{-4} - .291746562z^{-5} + .4537681314z^{-6} \end{array}} \tag{5.1.8}$$

The direct II form sensitivity is 194.49296. Note that the sensitivity is larger than the sensitivity of the system described by equation (5.1.4), although the only difference in the two systems is the location of the center frequency. Most importantly, the band-width is preserved through the

transformation, and yet the sensitivity has increased! Why? To help clarify the details, the low sensitivity low-pass prototype used to design the system of equation (5.1.7) is frequency translated to have a center frequency of 1.47 radians; the same as the system of equation (5.1.8). The resulting system is

$$H(z) = \frac{\begin{array}{c} .0079306721z^{-1} - .080702339z^{-2} + .004206594z^{-3} \\ + .099915337z^{-4} - .006229617z^{-5} - .044088003z^{-6} \\ - .004302456z^{-7} + .022016474z^{-8} \end{array}}{\begin{array}{c} 1 - .502486114z^{-1} + 1.112407414z^{-2} - .248877057z^{-3} \\ - .332209894z^{-4} + .37578001z^{-5} - 1.084386945z^{-6} \\ + .274890393z^{-7} - .431079724z^{-8} \end{array}} \tag{5.1.9}$$

The direct II form sensitivity is 78.50374. Notice that the sensitivity is reduced by the same factor (2.5) as for the band-pass filters centered about $\frac{\pi}{2}$ radians.

The explanation of why the location of the pass-band affects the sensitivity may be viewed in one of the two following ways:

1. The frequency warping which occurs in transforming the low-pass filter prototype into the desired band-pass function is not linear unless the pass-band is centered around $\frac{\pi}{2}$ radians. Of more importance, for a band-pass filter centered anywhere else, the corresponding low-pass filter will have a smaller band-width than the prototype low-pass filter; smaller band-width has been shown previously to increase the sensitivity measure.

2. With a pass-band centered at a frequency other than $\frac{\pi}{2}$ radians, the system poles are closer to each other than they would otherwise be. As previously shown, the sensitivity measure is approximately inversely proportional to the pole distances.

5.2 *Extension of Band-Pass Filter Ideas to a Band-Stop Filter*

Just as high-pass filters are symmetrical to low-pass filters, band-stop filters are symmetrical to band-pass filters. Note the similarity of the frequency transformation equations to those of the band-pass transformation

$$z^{-1} \rightarrow -\frac{z^{-2} - \frac{2\alpha}{1+k}z^{-1} + \frac{1-k}{1+k}}{\frac{1-k}{1+k}z^{-2} - \frac{2\alpha}{1+k}z^{-1} + 1} \tag{5.2.1}$$

where again

$$\alpha = \frac{\cos(\frac{\omega_2 + \omega_1}{2})}{\cos(\frac{\omega_2 - \omega_1}{2})} \tag{5.2.2}$$

and

$$k = \cot(\frac{\omega_2 - \omega_1}{2}) \tan\frac{\theta_p}{2} \tag{5.2.3}$$

Again θ_p is the cut-off frequency of the low-pass prototype filter, and ω_1 and ω_2 are the desired upper and lower cut-off frequencies of the band-pass filter.

The order of the band-stop filter is twice that of the low-pass prototype, just as the band-pass filter order is double that of the low-pass filter prototype. Band-stop filters have two clusters of poles in the z-plane, one at low frequencies (i.e. near 1 in the z-plane) and one at high frequencies (i.e. near -1 in the z-plane). As in the band-pass filter, the locations of the pass-bands affect the sensi-

tivity; the sensitivity is either identical to the sensitivity of the low-pass prototype or larger than the low pass prototype sensitivity.

The design of low sensitivity band-pass and band-stop filters is then accomplished as follows:

1. Design the low-pass prototype with the desired band-width.

2. Find locations for pole/zero cancellation pairs which minimize the sensitivity.

3. Transform the reduced sensitivity filter to the desired form, band-pass or band-stop.

Note that because the order of the system is doubled by the frequency translation, the order of this low sensitivity form will be increased by either two or four over the minimum-order system instead of the one or two in the prototype low-pass filter.

Another possible method of sensitivity reduction in a band-pass filter would be to implement a band-pass filter as the cascade combination of a high-pass filter and a low-pass filter, and then minimize the sensitivity of each cascade section. The order of the overall band-pass filter would still be twice that of a low-pass prototype, but the structure is now one that is amenable to placing pole/zero cancellation pair(s) to reduce the coefficient sensitivity of the individual sections. A dual to the band-pass filter, the band-stop filter could possibly be implemented as the parallel combination of a high-pass and a low-pass filter. Then, the coefficient sensitivity of the individual sections could be reduced by adding pole/zero cancellation pair(s), thereby producing an overall filter of reduced coefficient sensitivity and lower output quantization noise.

6.0 Conclusions and Suggestions for Further Study

6.1 Conclusions

The linear relationship between the L_2 sensitivity measure and the output quantization noise power was shown; an efficient method for calculating both of these is given. A method for the analysis of filters which have poles that are almost on the unit circle is presented; this method scales the pole and zero radii to make the sensitivity analysis more stable numerically. The computation of the state space forms (cascade, parallel, direct II, optimal, block-optimal, section-optimal and Dual GHR) used in the thesis was presented, with special emphasis on the optimal, block-optimal, section-optimal and Dual GHR implementations.

Next, the direct relationship between the pole and zero sensitivities and the sensitivity measure was exploited to reduce the system output quantization noise power of low-pass, direct II form digital filters by the introduction of judiciously placed pole/zero cancellation pair(s). These cancellation pair(s) do not affect the system transfer function. For some filters, the present method brings the sensitivity down by a factor of 10, which for low-order low-pass sections approximates the optimal sensitivity. Further, to achieve these low sensitivities, the direct II implementation must only be

increased in order by one or two. Since a direct II implementation requires only 2n coefficients and the optimal form requires $n(n+2)$ coefficients, we actually have fewer coefficients even with the increased order. Therefore, we can increase the throughput without appreciably increasing the output quantization noise power of the filter.

The effects of frequency transformations of low-pass filters are shown to be trivial in the low-pass to high-pass transformation, but because of the frequency warping which occurs in the low-pass to band-pass or band-stop transformations the analysis of these two types is somewhat more complicated. However, the effects are easily understood and explained and they lead to the design technique below:

1. Determine the required filter specifications.
2. Find the proper low-pass prototype filter specifications with the constraint that the band-width of the low-pass prototype be the band-width of the desired filter.
3. Find the location of the pole/zero cancellation pair(s) for minimum sensitivity.
4. Frequency translate the low-pass prototype filter to the desired filter function.

This technique always yields the best sensitivity for the desired function, under the constraint of the system order.

6.2 Suggestions for Further Study

Clearly, direct minimization of the sensitivity by variation of the pole/zero cancellation parameters should be examined. We have shown that ARMA cross-covariances can be efficiently evaluated, thus gradient based methods can be implemented with little difficulty.

Further, the idea mentioned in Chapter 5 of implementing the band-pass filter as the cascade combination of low-pass and high-pass filters and the band-stop filter as a low-pass filter in parallel with a high-pass filter should also be examined. Because no frequency warping would be required and the band-widths of the separate low-pass and high-pass sections would be large, significant sensitivity reductions might be possible. Problems might be encountered in the filter stop-band regions, however further study in this direction is warranted.

Finally, the sensitivity measure has many properties which have not been explored in this work, among which are the sensitivity measure as a function of the pole and zero radii scaling factor (which appears to be a non-trivial problem), the use of other norms instead of the two norm and the examination of the partial derivative systems used in its calculation.

Bibliography

1. A. A. Beex, "Efficient Generation of ARMA Cross Covariance Sequences," *IEEE ICASSP '85 Proceedings*, pp. 327-330, March 1985.

2. B. W. Bomar and J. C. Hung, "Minimum Roundoff Noise Digital Filters With Some Power-Of-Two Coefficients," *IEEE Trans. Circuits and Systems*, v CAS-31, pp. 833-840, October 1984.

3. B. W. Bomar, "New Second-Order State-Space Structures for Realizing Low Roundoff Noise Digital Filters," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-33, n 1, pp. 106-110, February 1985.

4. B. W. Bomar, "Computationally Efficient Low Roundoff Noise Second-Order State-Space Structures," *IEEE Trans. Circuits and Systems*, v CAS-33, n 1, pp. 35-41, January 1986.

5. A. G. Constantinides and R. A. Valenzuela, "A Class of Efficient Interpolators and Decimators With Applications in Transmultiplexers," *Proc. IEEE Int. Symp. Circuits Syst.*, Rome, Italy, pp. 260-263, May 1982.

6. A. G. Constantinides and R. A. Valenzuela, "An Efficient and Modular Transmultiplexer Design," *IEEE Trans. Communications*, v COM-30, pp. 1629-1641, July 1982.

7. G. Cybenko, "The Numerical Stability of the Levinson-Durbin Algorithm for Toeplitz Systems of Equations," *SIAM J. Sci. Stat. Comput.*, v 1, n 3, pp. 303-319, September 1980.

8. A. Fettweis, "On Sensitivity and Roundoff Noise in Wave Digital Filters," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-22, pp. 383-384, October 1974.

9. A. Fettweis, "Wave Digital Lattice Filters," *Int. J. Circuit Theory Appl.*, v 2 pp. 203-211, June 1974.

10. S. Y. Hwang, "Minimum Uncorrelated Unit Noise in State-Space Digital Filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-25, n 4, pp. 273-281, August 1977.

11. L. B. Jackson, "Roundoff Noise Bounds Derived from Coefficient Sensitivities for Digital Filters," *IEEE Trans. Circuits and Systems*, v CAS-23, n 8, pp. 481-485, August 1976.

12. L. B. Jackson, A. G. Lindgren and Y. Kim, "Optimal Synthesis of Second-Order State-Space Structures for Digital Filters," *IEEE Trans. Circuits and Systems*, v CAS-26, n 3, pp. 149-153, March 1979.

13. J. F. Kaiser, "Digital Filters," *System Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, John Wiley & Sons, Inc., New York, chapter 7, 1966.

14. J. F. Kaiser, "Some Practical Considerations in the Realization of Linear Digital Filters," *Proc. 3rd Allerton Conf. Circuit System Theory*, pp. 621-633, Oct 20-22, 1965.

15. M. Kawamata and T. Higuchi, "A Unified Approach to the Optimal Synthesis of Fixed-Point State-Space Digital Filters," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-33, n 4, pp. 911-920, August 1985.

16. D. K. Lindner, "The Dual GHR," *Proc of 22nd Annual Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, pp. 745-752, 1984.

17. J. L. Melsa, *Computer Programs for Computational Assistance in the Study of Linear Control Theory*, McGraw-Hill, New York, pp. 39-55, 95-97 and 119-120, 1970.

18. S. K. Mitra and R. J. Sherwood, "Canonic Realizations of Digital Filters Using the Continued Fraction Expansion," *IEEE Trans. Audio Electroacoust.*, v AU-20, pp. 185-194, 1972.

19. C. T. Mullis and R. A. Roberts, "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters," *IEEE Trans. Circuits and Systems*, v CAS-23, n 9, pp. 551-562, September 1976.

20. C. T. Mullis and R. A. Roberts, "Roundoff Noise in Digital Filters: Frequency Transformations and Invariants," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-24, n 6, pp. 538-550, December 1976.

21. A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentiss-Hall, Inc., Englewood Cliffs, New Jersey, pp. 166-171, 186-187, 214-216, 443 and 562-570, 1975.

22. D. V. B. Rao, "Analysis of Coefficient Quantization Errors in State-Space Digital Filters," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-34, n 1, pp. 131-139, February 1986.

23. V. Tavsanoglu and L. Thiele, "Optimal Design of State-Space Digital Filters by Simultaneous Minimization of Sensitivity and Roundoff Noise," *IEEE Trans. Circuits and Systems*, v CAS-31, n 10, pp. 884-888, October 1984.

24. P. P. Vaidyanathan, S. K. Mitra and Y. Neuvo, "A New Approach to the Realization of Low-Sensitivity IIR Digital Filters," *IEEE Trans. Acoust., Speech, Signal Processing*, v ASSP-34, n 2, pp. 350-361, April 1986.

25. C. F. Van Loan and G. H. Golub, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Maryland, p. 3, 1983.

The vita has been removed from the scanned document