To the Graduate Council:

I am submitting herewith a thesis written by Reuben D. Morris, III, entitled "A Study of Pole-zero Placement Options for Frequency Sampling FIR Filter Realizations." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Electrical Engineering.

Bruce W. Bomar

Bruce W. Bomar, Major Professor

We have read this thesis and

recommend its acceptance:

Roy D Joseph

L Montgomery Smith

Accepted for the Council:

C W Minkel

Associate Vice Chancellor and

Dean of The Graduate School

A Study of Pole-Zero Placement Options for Frequency Sampling FIR Filter Realizations

A Thesis Presented for the

Master of Science Degree

The University of Tennessee, Knoxville

Reuben D. Morris, III

December 1995

DEDICATION

This thesis is dedicated to my parents

Reuben Daniel Morris, Jr.

and

Alfrieda Blackwell Morris

who taught me the most important lessons in life.

ACKNOWLEDGEMENTS

I would like to thank my advisor and major professor Dr. Bruce Bomar for his guidance in producing this thesis. I would also like to thank Dr. Roy Joseph and Dr. L. Montgomery Smith for serving on my thesis committee.

ABSTRACT

This thesis investigates finite impulse response (FIR) digital filters implemented using a frequency-sampling structure with a different radius for the poles than for the zeros. Even though it is well known that both the poles and zeros of the filter should ideally fall at the same locations on the unit circle, practical implementations move both the poles and zeros slightly inside to make the filter stable. Another benefit gained is that as the poles are moved farther inside the unit circle the error due to roundoff noise decreases. In a recent thesis both the poles and zeros were moved inside to the same radius. It was shown that the roundoff noise decreases as the poles are moved farther inside, but the frequency response in the passband also decreases in amplitude. It was also shown that the amplitude in the passband can be corrected by scaling, but it increases the amplitude in the stopband causing frequencies defined to have a zero gain to have a gain greater than zero.

Pole-zero separation distance and stopband zero compensation are investigated in this thesis for moving the poles inside the unit circle to a different radius than the zeros. These techniques were used in an attempt to retain the desirable effects of increased stability and reduced roundoff noise, while allowing the amplitude to be corrected in the passband by scaling without

distorting the amplitude in the stopband. It is found that neither technique is useful for the design of low roundoff noise frequency sampling structure digital filters. Pole-zero separation cannot be used because both the magnitude and phase responses become too distorted for separation distances that are large enough to reduce roundoff noise. Stopband zero compensation is not practical because of the excessive computation involved in its implementation.

Since neither pole-zero separation nor stopband zero compensation proved to be useful it is concluded that the best technique for designing low roundoff noise frequency sampling filters is to move both the poles and zeros inside the unit circle to the same radius. The floating-point roundoff noise gain is analyzed for three different realizations of both lowpass and bandpass filters where this technique is used. This roundoff noise analysis takes the extended precision of floating-point digital signal processors into account.

TABLE OF CONTENTS

FIGURES

FIGURE **PAGE**

CHAPTER I

INTRODUCTION

Frequency sampling structures are more efficient for the implementation of narrow-band finite impulse response (FIR) digital filters than direct form structures [1]. The frequency sampling structure produces a transfer function that is computationally simpler than the impulse response convolution of direct form structures. However, two problems exist with frequency sampling filters. One is that they are marginally stable when realized ideally with both the poles and zeros in the same locations on the unit circle. The other is that they have excessive roundoff noise if the poles are close to the unit circle. To avoid these problems, most practical realizations move both the poles and zeros to the same radius inside the unit circle to make the filter stable and reduce the amount of roundoff noise.

Unfortunately, as the poles and zeros are move inside the unit circle, both the magnitude and phase response become distorted. The passband response becomes attenuated; the stopband response loses attenuation; and the phase response becomes nonlinear. Although the passband response can easily be corrected by scaling, this increases the problem in the stopband [1].

The major issue addressed by this thesis is whether the magnitude and phase response problems associated with moving both the poles and zeros to the same radius inside the unit circle can be solved by allowing the poles and zeros to lie on different radii. In Chapter II the background equations for this thesis are developed. These equations were used in Borland® C/C++ 3.0 language programs written to accomplish the work presented in this thesis. The programs were executed on a personal computer with an Intel® 80386 processor and a math co-processor. In these programs all calculations where performed using double-precision arithmetic. In Chapter III pole-zero separation is investigated as one possible solution. Pole-zero separation involves moving the poles inside the unit circle farther than the zeros to minimize roundoff noise while preserving the stopband attenuation. In Chapter IV stopband zero compensation is investigated as another possible solution. Stopband zero compensation involves moving both the passband poles and zeros to the same radius inside the unit circle while leaving the stopband zeros on the unit circle to achieve the same goal as with pole-zero separation. Since it is concluded that poles and zeros must both be moved to the same radius inside the unit circle, Chapter V discusses the roundoff noise gain of three realizations when this is done. The three realizations are the canonical form structure of [3], the minimum noise structure of [4], and a Type III structure of [3]. The case of implementation on a floating-point digital signal processor is considered. Finally, Chapter VI presents conclusions from this research.

CHAPTER II

BACKGROUND

This section derives the equations for a linear phase FIR frequency sampling filter that has a different radius for the poles than for the zeros and develops a set of state-space equations that can be used to implement the filter.

Derivation of the Frequency Sampling Filter Equation

If given the complex-valued frequency samples, *H(k)*, of the odd *M* point frequency response of a filter, the real-valued impulse response, *h(n)*, can be calculated by taking the Inverse Discrete Fourier Transform (IDFT) of the frequency samples as shown below.

$$h(n) = \frac{1}{M}\sum_{k=0}^{M-1} H(k)\cdot e^{j\frac{2\pi kn}{M}} \cdot u(n)\,; \qquad n = 0, 1, 2, \ldots, M\text{-}1. \quad \textbf{(1)}$$

Although the unit-step sequence, *u(n)*, is not needed in **(1)** to calculate the IDFT, it aids in what follows and does not change *h(n)*.

Taking the z-transform, *H(z)*, of *h(n)* yields

$$H(z) = \sum_{n=0}^{M-1} h(n)\cdot z^{-n}. \qquad \textbf{(2)}$$

If the right side of equation **(1)** is substituted into equation **(2)** for *h(n)*, an equation for *H(z)* in terms of the complex-valued frequency samples *H(k)* is produced as shown below.

$$H(z)=\sum_{n=0}^{M-1}\left[\frac{1}{M}\sum_{k=0}^{M-1}H(k)\cdot e^{j\frac{2\pi kn}{M}}\right]\cdot u(n)\cdot z^{-n}; \qquad n=0,1,2,\ldots,M\text{-}1. \quad \textbf{(3)}$$

The form of the equation for the frequency sampling structure can be obtained from equation **(3)**. The mathematical manipulation required is as follows:

$$H(z)=\frac{1}{M}\sum_{k=0}^{M-1}H(k)\left[\sum_{n=0}^{M-1}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n)\right];$$

$$H(z)=\frac{1}{M}\sum_{k=0}^{M-1}H(k)\left[\sum_{n=0}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n)-\sum_{n=M}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n-M)\right];$$

changing the variables ($L = n\text{-}M$) in the terms after the third summation gives

$$H(z)=\frac{1}{M}\sum_{k=0}^{M-1}H(k)\left[\sum_{n=0}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n)-\sum_{L=0}^{\infty}e^{j\frac{2\pi k}{M}(L+M)}\cdot z^{-(L+M)}\cdot u(L)\right];$$

$$H(z)=\frac{1}{M}\sum_{k=0}^{M-1}H(k)\left[\sum_{n=0}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n)-z^{-M}\sum_{L=0}^{\infty}e^{j\frac{2\pi k}{M}L}\cdot z^{-L}\cdot u(L)\right];$$

changing the variables ($L = n$) in the terms after the third summation gives

$$H(z)=\frac{1}{M}\sum_{k=0}^{M-1}H(k)\left[\sum_{n=0}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n)-z^{-M}\sum_{n=0}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n)\right];$$

combining like terms and simplifying yields

$$H(z)=\frac{1-z^{-M}}{M}\sum_{k=0}^{M-1}H(k)\sum_{n=0}^{\infty}e^{j\frac{2\pi k}{M}n}\cdot z^{-n}\cdot u(n);$$

$$H(z)=\frac{1-z^{-M}}{M}\sum_{k=0}^{M-1}\frac{H(k)}{1-e^{j\frac{2\pi k}{M}}z^{-1}};$$

$$H(z)=\frac{z^{M}-1}{z^{M}\cdot M}\sum_{k=0}^{M-1}\frac{H(k)\cdot z}{z-e^{j\frac{2\pi k}{M}}}. \quad \textbf{(4)}$$

Equation **(4)** is the z-transform representation of a frequency sampling filter that has both poles and zeros in the same location on the unit circle. If the filter is implemented with poles on the unit circle, it is marginally stable, and it can become unstable due to coefficient quantitization, and it can have excessive roundoff noise [1]. Therefore, in most implementations both the poles and zeros are moved inside the unit circle by the same amount to avoid instability and large roundoff noise, but it is possible to move both inside by different amounts. The radius of the poles is determined by the denominator of the term after the summation, $z-e^{j\frac{2\pi k}{M}}$. The poles occur when $|z|=\left|e^{j\frac{2\pi k}{M}}\right|=1$. The poles can be moved off the unit circle by a radius (r_p) by changing the denominator to $z-r_p\cdot e^{j\frac{2\pi k}{M}}$. Similarly, the radius of the zeros is determined by the numerator of the term that precedes the summation, $z^{M}-1$. The zeros occur when $z^{M}=1$.

The zeros can be moved off the unit circle by a radius (r_z) by changing the numerator to $z^M - r_z^M$. Incorporating these changes in equation **(4)** yields

$$H(z) = \frac{z^M - r_z^M}{z^M \cdot M} \sum_{k=0}^{M-1} \frac{H(k) \cdot z}{z - r_p \cdot e^{j\frac{2\pi k}{M}}} \tag{5}$$

which is the z-transform representation of a frequency sampling filter that has poles and zeros at different radii.

Given the z-transform representation of a frequency sampling filter, as shown in equation **(5)**, it can be put in the form $H(z) = C(z) \cdot G(z)$ which is the cascade of two filters [1]. The term $C(z) = 1 - z^{-M} r_z^M$ is the z-transform representation of a comb filter that is easily implemented. The other term can be represented as follows:

$$G(z) = \frac{1}{M} \sum_{k=0}^{M-1} \frac{H(k)}{1 - r_p \cdot e^{j\frac{2\pi k}{M}} \cdot z^{-1}};$$

$$G(z) = \frac{H(0)/M}{1 - r_p \cdot z^{-1}} + \sum_{k=1}^{\frac{M-1}{2}} \frac{H(k)/M}{1 - r_p \cdot e^{j\frac{2\pi k}{M}} \cdot z^{-1}} + \sum_{\frac{M+1}{2}}^{M-1} \frac{H(k)/M}{1 - r_p \cdot e^{j\frac{2\pi k}{M}} \cdot z^{-1}}. \tag{6}$$

Now, since the filter has a linear phase shift, $h(n) = h(M-n-1)$, and the complex-valued term, *H(k)*, can be represented as $H_r(k) \cdot e^{-j\frac{M-1}{2}\left(\frac{2\pi k}{M}\right)}$ where $H_r(k)$ is real [2]. Also, since *h(n)* is real, $H^*(k) = H(M-k)$. Using this information with equations **(6)** yields

$$G(z) = \frac{H(0)/M}{1 - r_p \cdot z^{-1}} + \sum_{k=1}^{\frac{M-1}{2}} \frac{H_r(k) \cdot e^{-j\frac{M-1}{2}\left(\frac{2\pi k}{M}\right)} \Big/ M}{1 - r_p \cdot e^{j\frac{2\pi k}{M}} \cdot z^{-1}} + \sum_{\frac{M+1}{2}}^{M-1} \frac{H_r(k) \cdot e^{-j\frac{M-1}{2}\left(\frac{2\pi k}{M}\right)} \Big/ M}{1 - r_p \cdot e^{j\frac{2\pi k}{M}} \cdot z^{-1}};$$

(7)

Since the two summation terms are complex conjugates of each other in equation **(7)**, they can be combined to form a second order term with real coefficients as follows:

$$G(z) = \frac{H(0)/M}{1 - r_p \cdot z^{-1}} + \frac{1}{M} \sum_{k=1}^{\frac{M-1}{2}} H_r(k) \cdot \left[\frac{e^{-j\frac{M-1}{2}\left(\frac{2\pi k}{M}\right)}}{1 - r_p \cdot e^{j\frac{2\pi k}{M}} \cdot z^{-1}} + \frac{e^{j\frac{M-1}{2}\left(\frac{2\pi k}{M}\right)}}{1 - r_p \cdot e^{-j\frac{2\pi k}{M}} \cdot z^{-1}} \right];$$

$$G(z) = \frac{H(0)/M}{1 - r_p \cdot z^{-1}} + \frac{1}{M} \sum_{k=1}^{\frac{M-1}{2}} H_r(k) \cdot \left[\frac{2\cos\left(\frac{M-1}{M}\pi k\right) - 2r_p \cdot \cos\left(\frac{M+1}{M}\pi k\right) z^{-1}}{1 - 2r_p \cdot \cos\left(\frac{2\pi k}{M}\right) z^{-1} + r_p^2 \cdot z^{-2}} \right].$$

(8)

To facilitate putting the equation in a form for realization via a state-space structure, the constant terms are extracted from both the first and second order terms by division. Then, using the trigonometric identity for *cos(a+b)* on the second order term yields

$$G(z) = \frac{1}{M} \sum_{k=1}^{\frac{M-1}{2}} H_r(k) \left[\frac{2r_p \cdot \cos\left(\frac{M-3}{M}\pi k\right) z^{-1} - 2r_p^2 \cdot \cos\left(\frac{M-1}{M}\pi k\right) z^{-2}}{1 - 2r_p \cdot \cos\left(\frac{2\pi k}{M}\right) z^{-1} + r_p^2 \cdot z^{-2}} + 2\cos\left(\frac{M-1}{M}\pi k\right) \right]$$

$$+ \frac{H_r(0)}{M} \left(1 + \frac{r_p \cdot z^{-1}}{1 - r_p \cdot z^{-1}} \right)$$

(9)

The filter, $G(z)$, can be implemented as the sum of a first order section, $G_1(z)$, and $\frac{M-1}{2}$ second order sections, $G_2(z)$, as shown.

$$G(z) = G_1(z) + G_2(z)$$

$$G_1(z) = \frac{H_r(0)}{M}\left(1 + \frac{r_p \cdot z^{-1}}{1 - r_p \cdot z^{-1}}\right) \qquad \textbf{(10)}$$

$$G_2(z) = \frac{1}{M}\sum_{k=1}^{\frac{M-1}{2}} H_r(k)\left[\frac{2r_p \cdot \cos\left(\frac{M-3}{M}\pi k\right)z^{-1} - 2r_p^2 \cdot \cos\left(\frac{M-1}{M}\pi k\right)z^{-2}}{1 - 2r_p \cdot \cos\left(\frac{2\pi k}{M}\right)z^{-1} + r_p^2 \cdot z^{-2}}\right]$$
$$+ \frac{1}{M}\sum_{k=1}^{\frac{M-1}{2}} H_r(k) \cdot 2\cos\left(\frac{M-1}{M}\pi k\right)$$

(11)

The first and second order sections in equations **(10)** and **(11)** can be implemented using state-space filter structures.

Derivation of the Frequency Sampling Filter State-Space Equations

A first order transfer function of the form, $d + \frac{\alpha \cdot z^{-1}}{1 + \beta \cdot z^{-1}}$ can be implemented by the state-space equations

$$y(n) = c_f \cdot x(n) + d \cdot v(n)$$
$$x(n+1) = a_f \cdot x(n) + b_f \cdot v(n)$$

(12)

$$\alpha = b_f \cdot c_f$$
$$\beta = -a_f$$

where *y(n)* is the output, *x(n)* is the state variable, and *v(n)* is the input [1]. Applying this to $G_1(z)$ in equation **(10)** yields

$$
\begin{aligned}
y(n) &= \frac{H_r(0)\cdot r_p}{M}\cdot x(n) + \frac{H_r(0)}{M}\cdot v(n) \\
x(n+1) &= r_p \cdot x(n) + v(n) \\
\\
a_f &= r_p \\
b_f &= 1 \\
c_f &= \frac{H_r(0)\cdot r_p}{M}
\end{aligned}
\qquad \textbf{(13)}
$$

which are the state-space equations for the first order filter section.

A second order transfer function of the form, $d + \dfrac{\alpha_1 \cdot z^{-1} + \alpha_2 \cdot z^{-2}}{1 + \beta_1 \cdot z^{-1} + \beta_2 \cdot z^{-2}}$ can be implemented by the state-space equations

$$
\begin{aligned}
y(n) &= \mathbf{C}\mathbf{x}(n) + \mathbf{D}v(n) \\
\mathbf{x}(n+1) &= \mathbf{A}\mathbf{x}(n) + \mathbf{B}v(n) \\
\mathbf{x}(n) &\equiv \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} \\
\mathbf{A} &\equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \mathbf{B} \equiv \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad \mathbf{C} \equiv \begin{bmatrix} c_1 & c_2 \end{bmatrix} \quad \mathbf{D} \equiv d
\end{aligned}
\qquad \textbf{(14)}
$$

where *y(n)* is the output; ***x****(n)* is the state-variable matrix; and *v(n)* is the input [2]. The transfer function coefficients and the matrices ***A***, ***B***, and ***C*** are related by the following equations [3].

$$
\begin{aligned}
\alpha_1 &= c_1 \cdot b_1 + c_2 \cdot b_2 \\
\alpha_2 &= c_1 \cdot b_2 \cdot a_{12} + c_2 \cdot b_1 \cdot a_{21} - c_1 \cdot b_1 \cdot a_{22} - c_2 \cdot b_2 \cdot a_{11} \\
\beta_1 &= -(a_{11} + a_{22}) \\
\beta_2 &= a_{11} \cdot a_{22} - a_{12} \cdot a_{21}
\end{aligned}
\qquad \textbf{(15)}
$$

From equation **(11)**, the transfer function coefficients are

$$
\begin{aligned}
\alpha_1 &= \frac{2 \cdot H_r(k) \cdot r_p \cdot \cos\left[\left(\frac{M-3}{M}\right) \cdot \pi k\right]}{M} \\
\alpha_2 &= -\frac{2 \cdot H_r(k) \cdot r_p^2 \cdot \cos\left[\left(\frac{M-1}{M}\right) \cdot \pi k\right]}{M} \\
\beta_1 &= -2 \cdot r_p \cdot \cos\left(\frac{2\pi k}{M}\right) \\
\beta_2 &= r_p^2
\end{aligned}
\qquad \textbf{(16)}
$$

Different realizations can be obtained by imposing different constraints on the elements of ***A***, ***B***, and ***C*** [3]. The three realizations used in this thesis are the canonical form structure of [3], the minimum noise structure of [4] , and a Type III structure of [3]. Choosing $a_{11} = b_1 = 0$, and $a_{12} = b_2 = 1$, gives the canonical form structure

$$
\begin{aligned}
a_{11} &= 0 \\
a_{12} &= 1 \\
a_{21} &= -\beta_2 \\
a_{22} &= -\beta_1 \\
b_1 &= 0 \\
b_2 &= 1 \\
c_1 &= \alpha_2 \\
c_2 &= \alpha_1
\end{aligned}
\qquad \textbf{(17)}
$$

which has 4 multiplies per section. Choosing $a_{11} = a_{22}$, and $b_1 = b_2 = 1$, gives the minimum noise structure

$$
\begin{aligned}
a_{11} &= \frac{-\beta_1}{2} \\
a_{12} &= \frac{\alpha_2}{\alpha_1} - \frac{\beta_1}{2} + \sqrt{\left(\frac{\alpha_2}{\alpha_1} - \frac{\beta_1}{2}\right)^2 - \left[\left(\frac{\beta_1}{2}\right)^2 - \beta_2\right]} \\
a_{21} &= \frac{\alpha_2}{\alpha_1} - \frac{\beta_1}{2} - \sqrt{\left(\frac{\alpha_2}{\alpha_1} - \frac{\beta_1}{2}\right)^2 - \left[\left(\frac{\beta_1}{2}\right)^2 - \beta_2\right]} \\
a_{22} &= \frac{-\beta_1}{2} \\
b_1 &= 1 \\
b_2 &= 1 \\
c_1 &= \frac{\alpha_1}{2} \\
c_2 &= \frac{\alpha_1}{2}
\end{aligned}
\tag{18}
$$

which has 6 multiplies per section. Choosing $a_{11} = a_{22}$, and $b_1 = 0$, gives the Type III structure

$$
\begin{aligned}
a_{11} &= \frac{-\beta_1}{2} \\
a_{12} &= 1 \\
a_{21} &= \left(\frac{\beta_1}{2}\right)^2 - \beta_2 \\
a_{22} &= \frac{-\beta_1}{2} \\
b_1 &= 0 \\
b_2 &= 1 \\
c_1 &= \alpha_2 - \frac{\alpha_1\beta_1}{2} \\
c_2 &= \alpha_1
\end{aligned}
\tag{19}
$$

which has 5 multiplies per section [3]. By comparing the number of multiplies per section, it is clear that the structure of a digital filter realization affects the computational load of its implementation. The structure of a digital filter realization also affects the roundoff noise gain of the filter. The effect that structures, (17), (18), and (19) have on the roundoff noise gains of filters used in this thesis is examined in Chapter V.

CHAPTER III

POLE-ZERO SEPARATION

This section examines the effects of pole-zero separation distance on the frequency response of two example FIR frequency sampling filters. One is a lowpass filter, and the other is a bandpass filter. Both were implemented using a filter length of 15. Then, the magnitude response, phase response, and group delay were graphed from 0 to π to characterize the frequency response of each filter as a function of pole and zero radius. The equations for the magnitude response, phase response, and group delay are as follows:

magnitude response $= |H(\omega)| = \sqrt{\mathrm{Re}(H(\omega))^2 + \mathrm{Im}(H(\omega))^2}$;

phase response $= \Theta(\omega) = \arctan\left(\dfrac{\mathrm{Im}(H(\omega))}{\mathrm{Re}(H(\omega))}\right)$ and

group delay $= \tau_g(\omega) = \dfrac{-d\Theta(\omega)}{d\omega}$.

The Lowpass Filter

The lowpass filter has a length of 15 and is constructed using 5 non-zero, positive, integer, frequency samples as follows:

$$H_r\left(\frac{2\pi k}{15}\right)=\begin{cases}1, & k=0,\pm 1,\pm 2,\pm 3;\\ 0.4, & k=\pm 4;\\ 0, & k=\pm 5,\pm 6,\pm 7.\end{cases}$$

The effect of pole-zero separation distance on the frequency response was determined by graphing the response for $r_p = r_z - 2^{-N}$, with $8 \le N \le \infty$. Although a number of radii were examined, the ones graphed are sufficient to show the effects. The radius of the zeros was set slightly inside the unit circle at 0.9999 for two reasons. One reason is to ensure stability when $r_p = r_z$. The other reason is to prevent the magnitude response from being zero at $\omega = \frac{2\pi k}{15}$ for all pole-zero separation distances due to the zero on the unit circle that is not canceled by a pole.

It was found that the effect of pole-zero separation distance is negligible on the magnitude response until $N < 24$ as shown in Figure 1. For $N < 24$, the zeros dominate the equation causing the response to be distorted in the passband at $\omega = \frac{2\pi k}{15}$. For $N \le 8$, the distortion reduces the magnitude to less than 0.05. As shown in Figure 2, pole-zero separation distance causes the phase response to become distorted, and the filter loses its linear phase characteristics for $N < 16$.

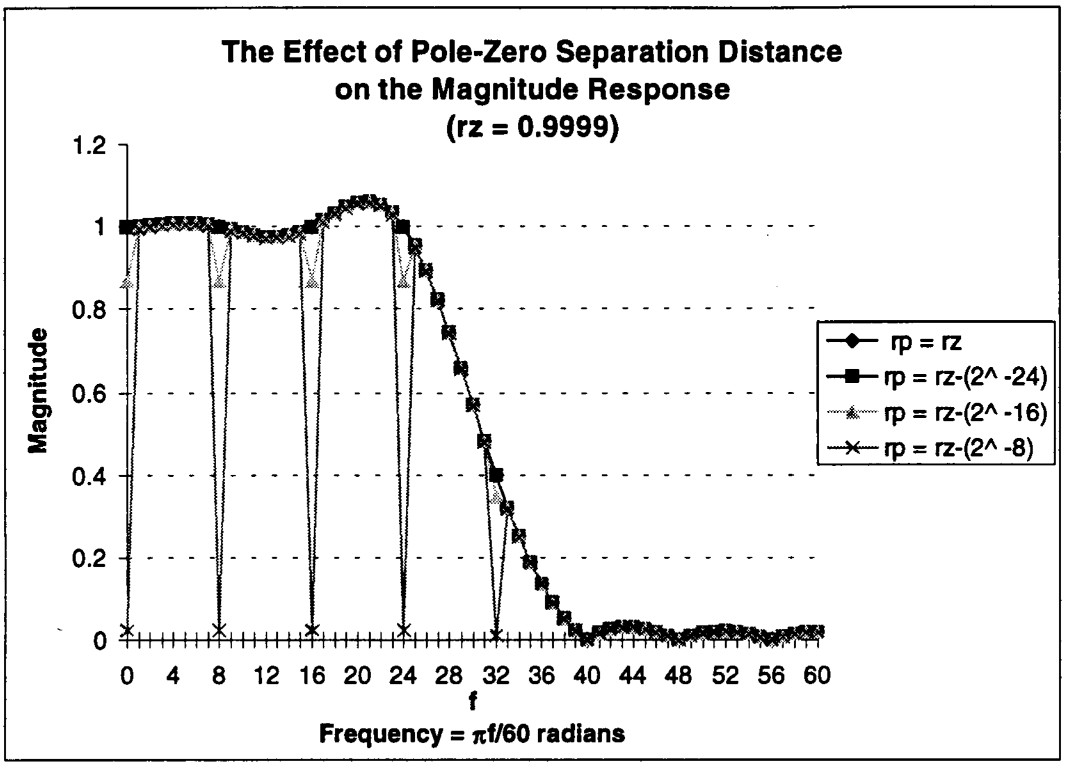

Figure 1: The effect of pole-zero separation distance on the magnitude response of the lowpass filter.

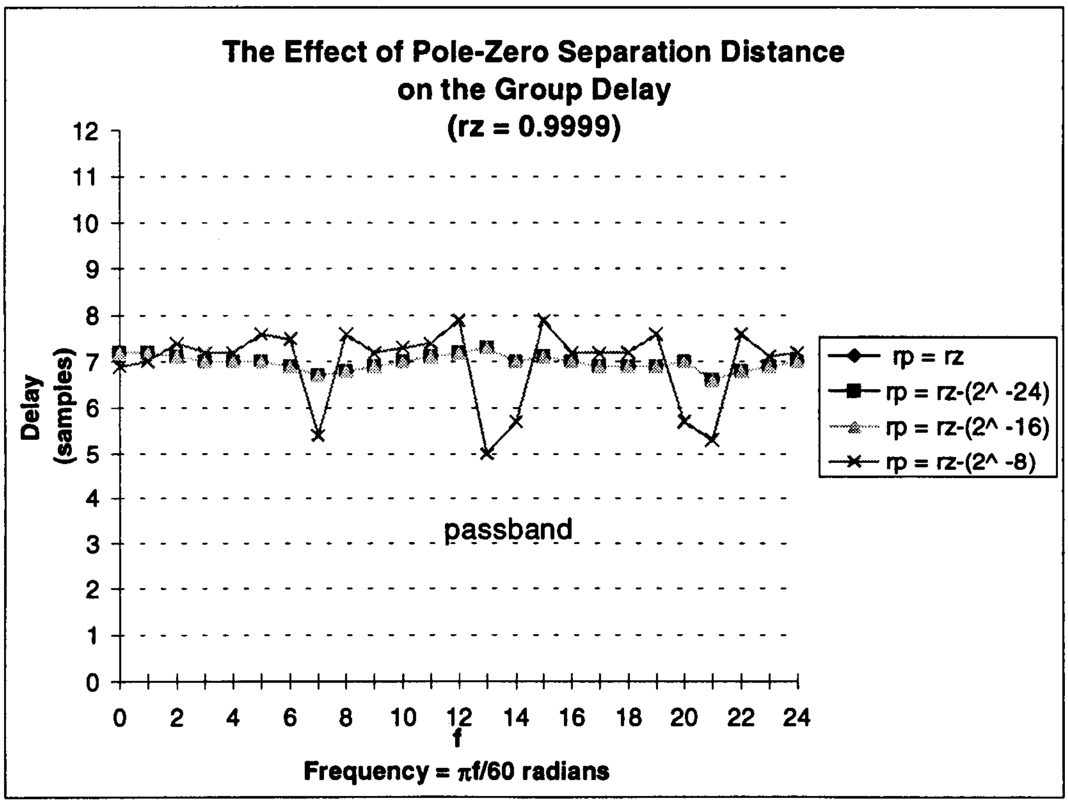

Figure 2: The effect of pole-zero separation distance on the group delay of the lowpass filter.

The Bandpass Filter

The bandpass filter has a length of 15 and is constructed using 5 non-zero, positive, integer, frequency samples as follows:

$$H_r\left(\frac{2\pi k}{15}\right)=\begin{cases}1, & k=\pm 2,\pm 3,\pm 4;\\ 0.4, & k=\pm 1,\pm 5;\\ 0, & k=\pm 0,\pm 6,\pm 7.\end{cases}$$

The effect of pole-zero separation distance on the frequency response was determined by graphing the response in the same manner as in the lowpass case, and the results were similar as shown in Figures 3 and 4.

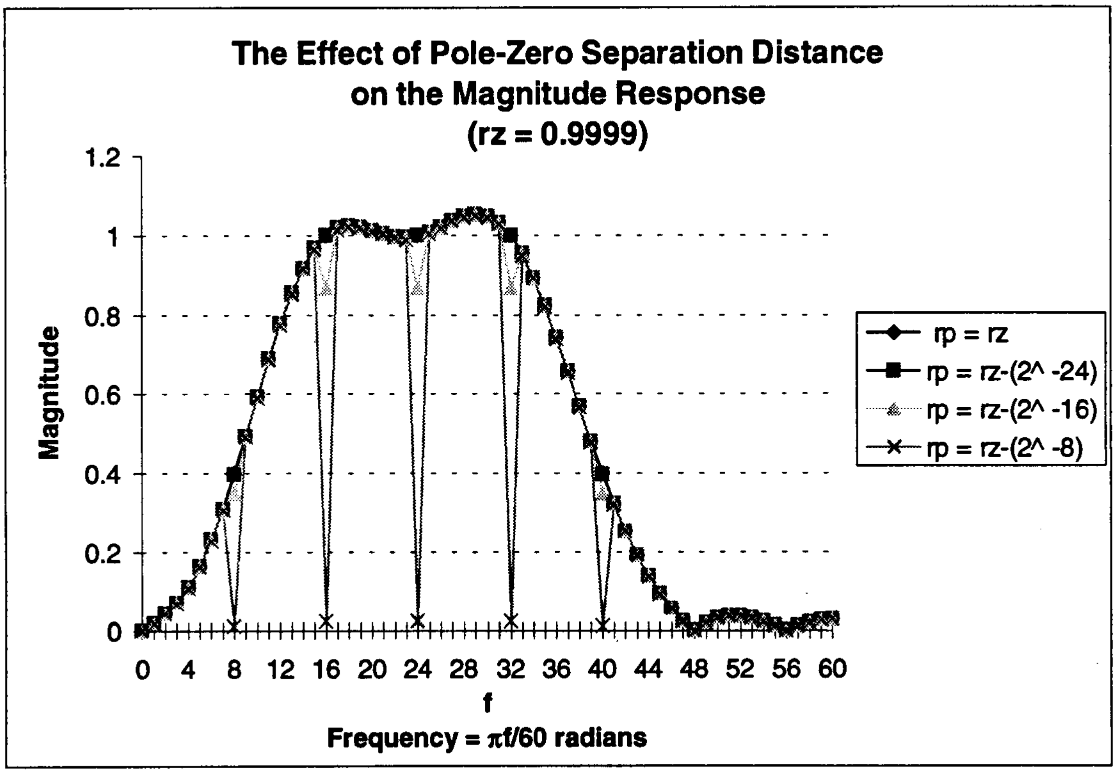

Figure 3: The effect of pole-zero separation distance on the magnitude response for the bandpass filter.

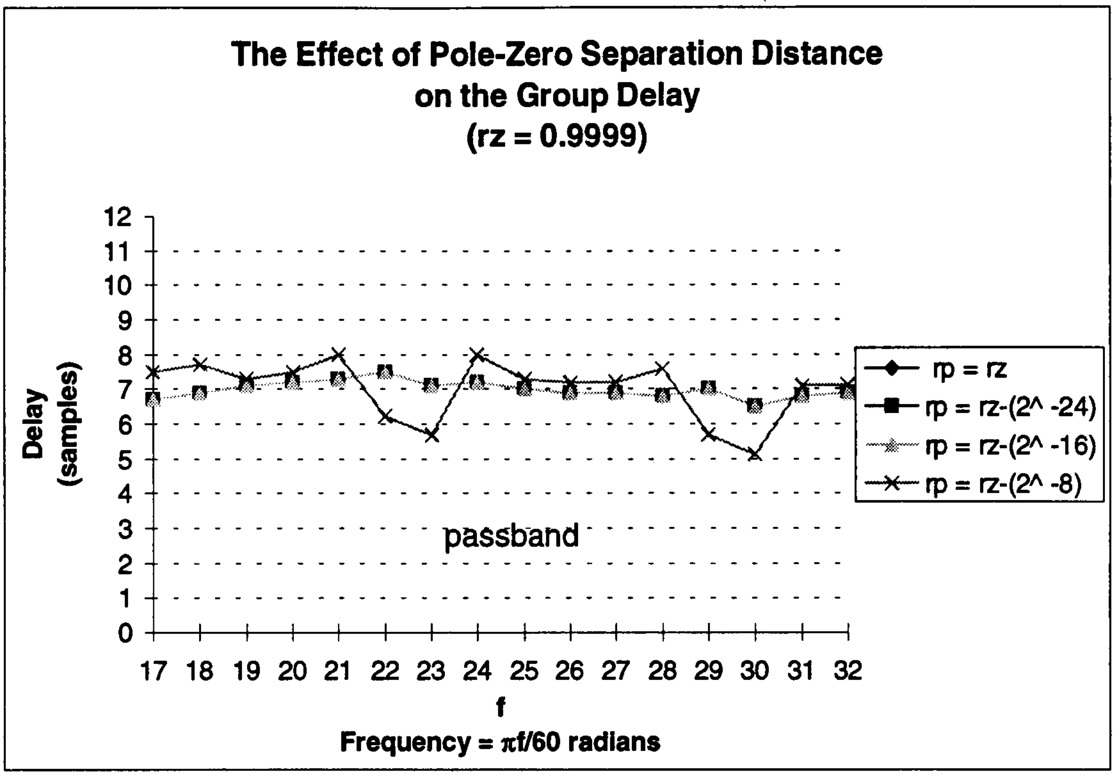

Figure 4: The effect of pole-zero separation distance on the group delay of the bandpass filter.

Summary

Pole-zero separation distance affects the frequency response of both the lowpass and bandpass frequency sampling filters identically. Both the magnitude and phase response become distorted as the separation distance increases. The distortion is negligible for separation distances that are less than or equal to 2^{-24} and may be acceptable for greater distances in some applications. The distortion causes the passband frequencies where $\omega = \frac{2\pi k}{15}$ to be greatly reduced in magnitude. These frequencies are effectively "notched-out" for significant pole-zero separation distances. This result indicates pole-zero separation cannot be used in the design of low round-off noise frequency sampling filters to solve the problem of the stopband losing attenuation when both the poles and zeros are moved inside the unit circle to the same radius.

CHAPTER IV

STOPBAND ZERO COMPENSATION

Since pole-zero separation does not offer much improvement, another possible modification is to treat the passband differently from the stopband. The stopband zero compensation technique leaves the stopband zeros on the unit circle to provide zero gain, but moves the passband poles and zeros inside the unit circle to the same radius. For this technique, the transition band poles and zeros are treated the same as those in the passband. The form of the equation to achieve this is as follows:

$$H(z)=\prod_{k_{passband}}\frac{z-r_p.e^{\frac{j2\pi k}{M}}}{z}\prod_{k_{stopband}}\frac{z-e^{\frac{j2\pi k}{M}}}{z}\sum_{k=0}^{M-1}\frac{H(k).z}{z-r_p.e^{\frac{j2\pi k}{M}}}. \quad (20)$$

It is clear from equation **(20)** that the mathematical computation for this technique is large because the zeros do not form a comb filter but are present individually in the equation. Although the computational complexity of this technique may prevent it from being practical when compared to a direct form realization, this section discusses the effects of stopband zero compensation on the frequency response of the two FIR filters considered in the last chapter.

The Lowpass Filter

The effect of stopband zero compensation on the frequency response was determined by graphing the response. As shown in Figure 5, it was found that stopband zero compensation improves the attenuation of frequencies in the stopband by ensuring that the frequencies which are defined to be zero actually are zero without distorting the passband. As shown in Figure 6, this improved stopband attenuation allows the magnitude response to be scaled to correct the attenuation in the passband caused by moving the poles and zeros inside the unit circle without causing the stopband response to become unacceptable.

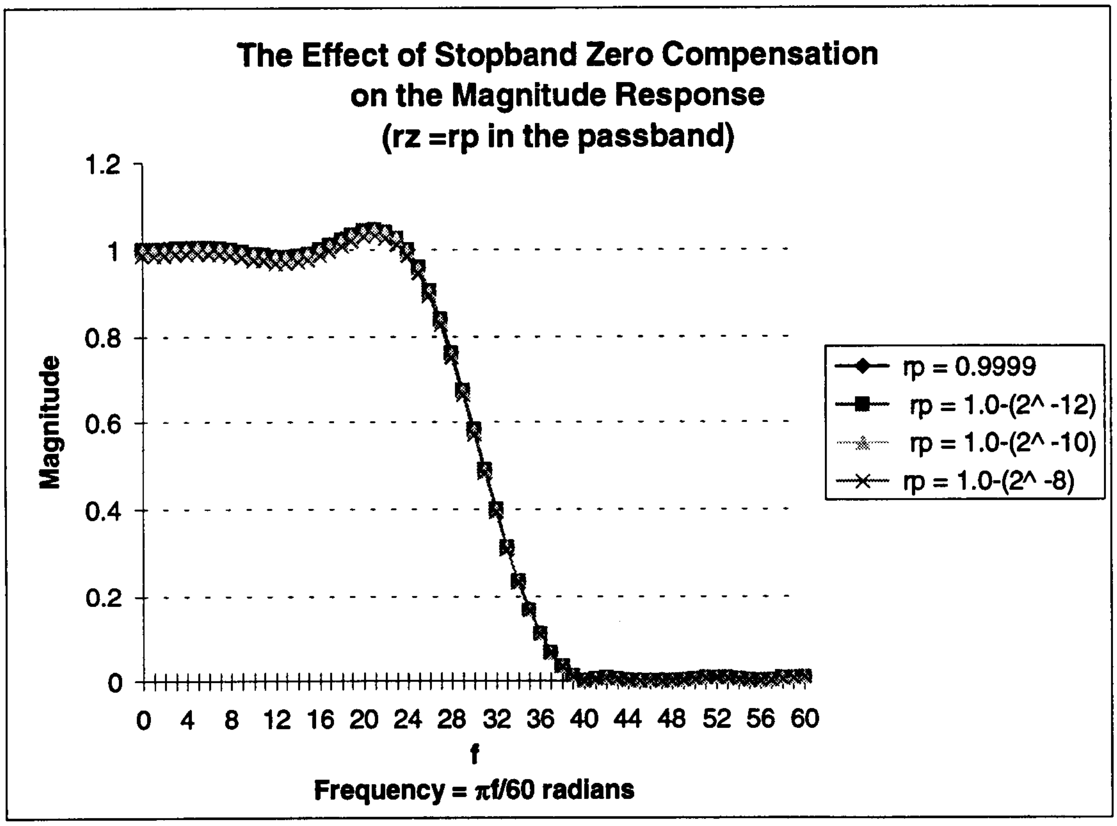

Figure 5: The effect of stopband zero compensation on the unscaled magnitude response of the lowpass filter.

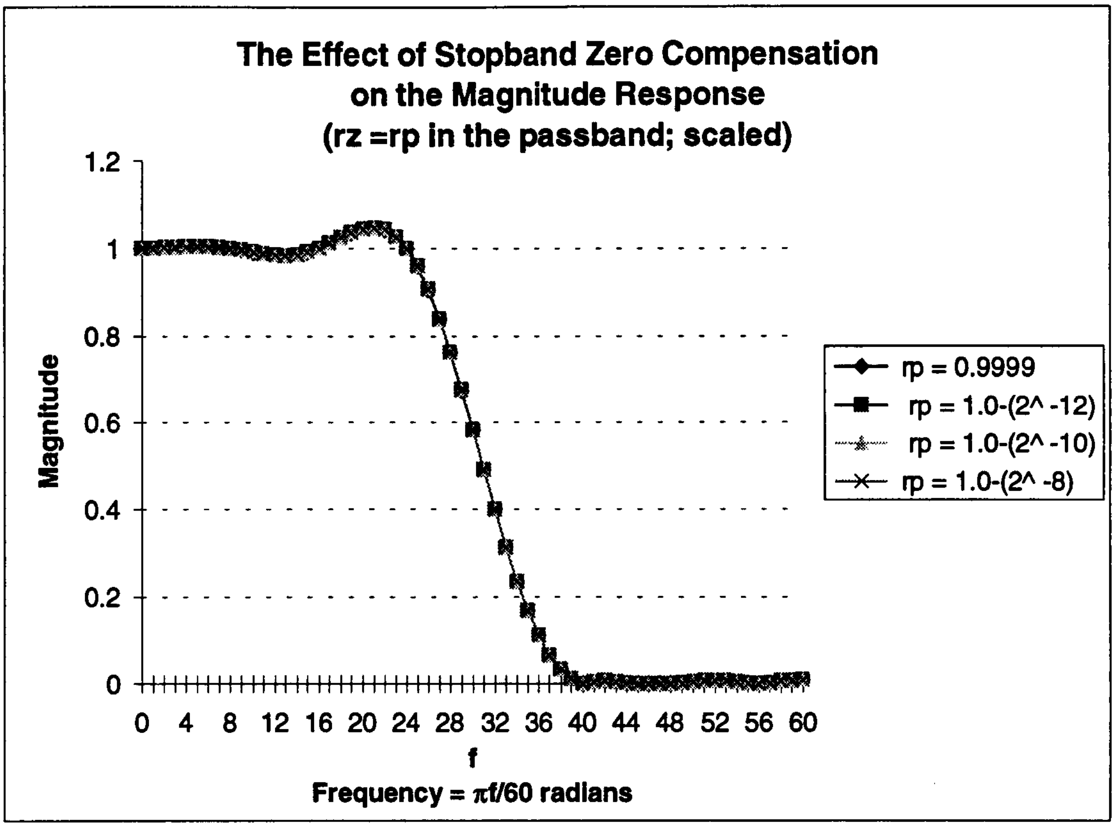

Figure 6: The effect of stopband zero compensation on the scaled magnitude response of the lowpass filter.

Although the magnitude response is improved because the stopband zeros on the unit circle cause the zero gain frequencies to remain zero, the phase response, as shown in Figure 7, becomes nonlinear when the passband poles and zeros are moved inside the unit circle. This effect occurs because the stopband zeros are no longer on the same radius as the passband zeros. The nonlinearity increases as the passband poles and zeros are moved farther inside the unit circle. Although the phase response becomes distorted, the nonlinearity is negligible for radii greater than $(1.0\text{-}2^{-10})$.

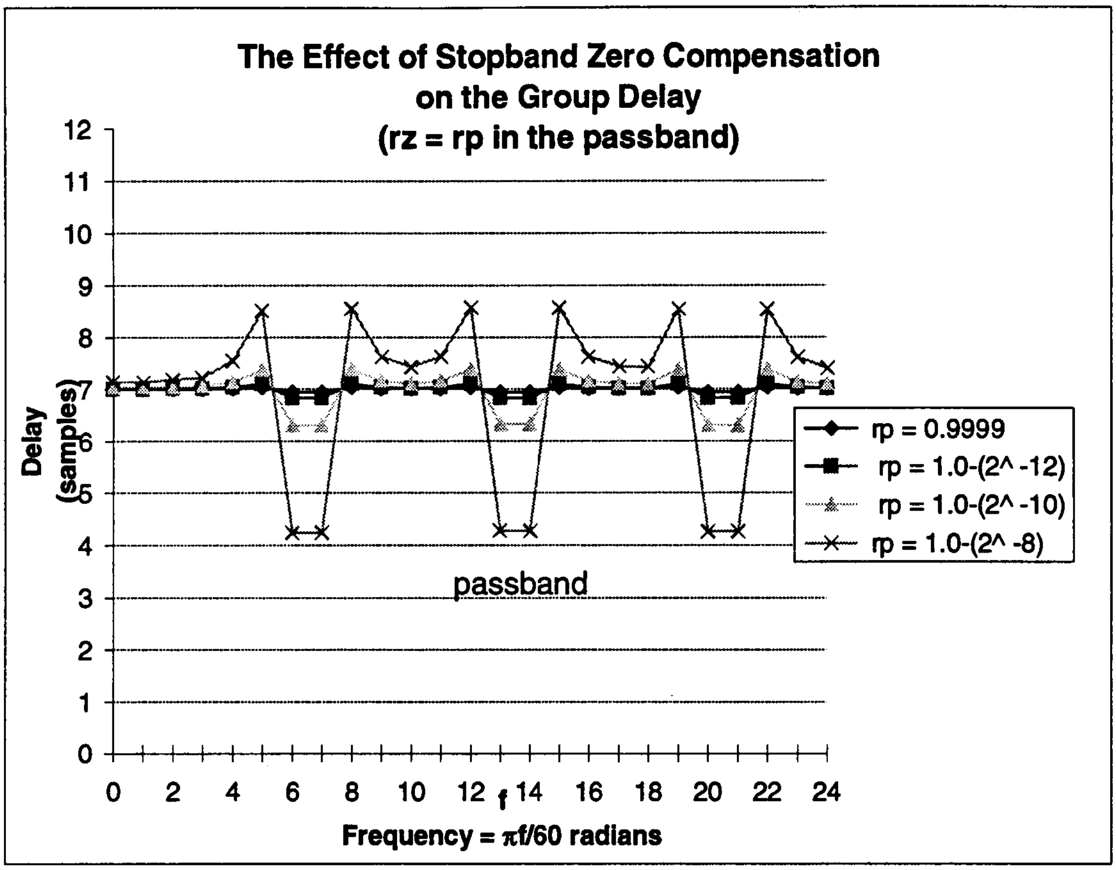

Figure 7: The effect of stopband zero compensation on the group delay for the lowpass filter.

The Bandpass Filter

The effect of stopband zero compensation on the frequency response of the bandpass filter was similar to the lowpass case. The magnitude response is improved, but the phase response becomes nonlinear. This nonlinearity is negligible for radii greater than ($1.0\text{-}2^{-10}$). The results are shown in Figures 8, 9, and 10.

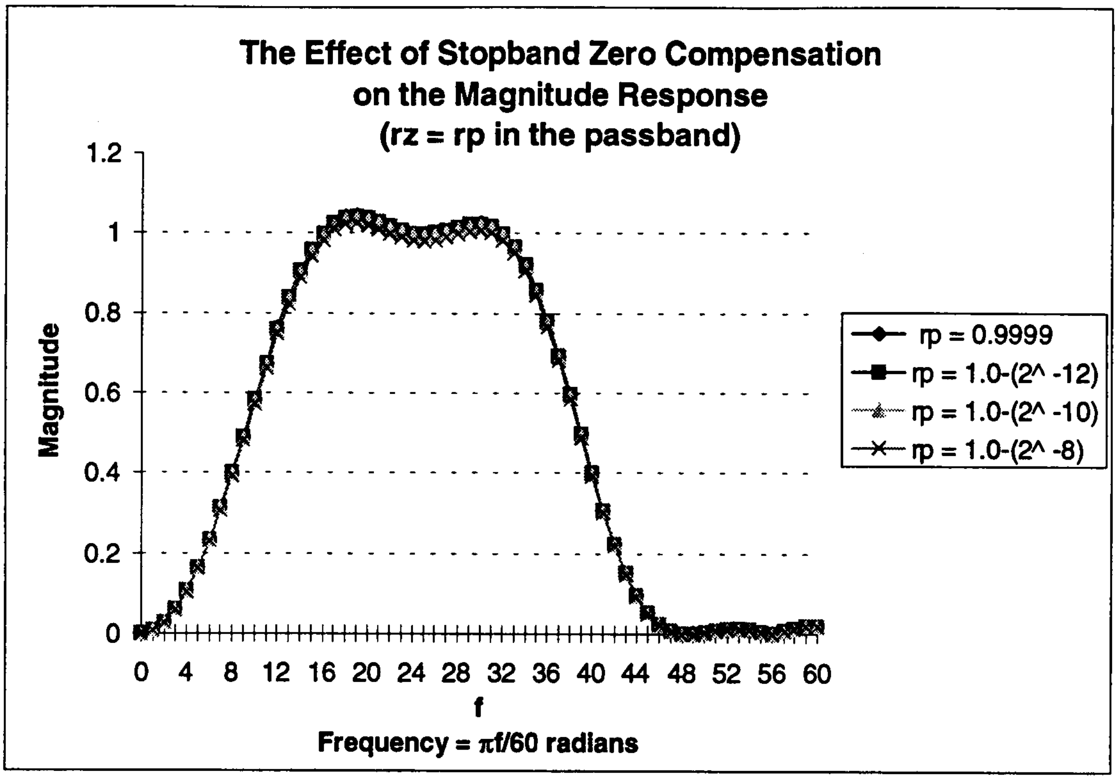

Figure 8: The effect of stopband zero compensation on the unscaled magnitude response of the bandpass filter.

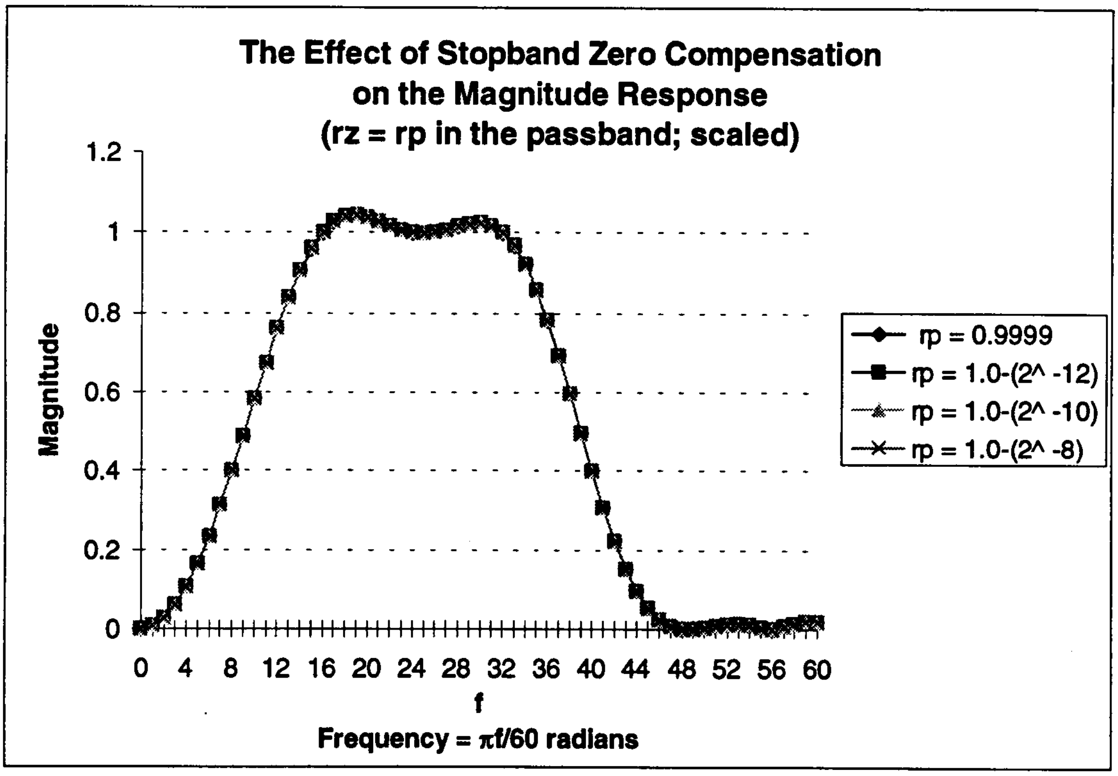

Figure 9: The effect of stopband zero compensation on the scaled magnitude response of the bandpass filter.

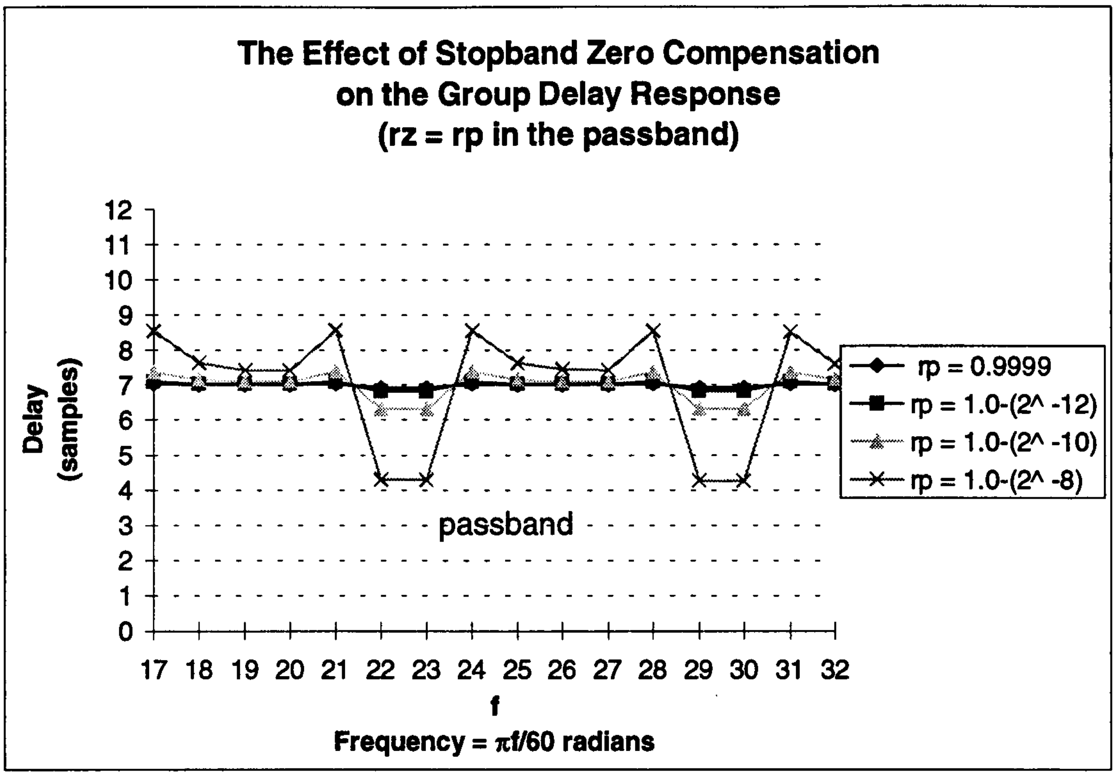

Figure 10: The effect of stopband zero compensation on the group delay of the bandpass filter.

Summary

Stopband zero compensation affects the frequency response of both the lowpass and bandpass filters similarly. The stopband attenuation is improved because the zeros in that area are on the unit circle and cause the zero frequencies to remain zero even when scaled. The phase response of the filters becomes nonlinear as the passband zeros are moved inside the unit circle because they lie on a radius different from the stopband zeros, but this effect is negligible for radii greater than ($1.0\text{-}2^{-10}$).

Although the frequency response of this technique is good, it may not be practical because the computation required to implement the filter is excessive.

This excessive computation is the result of the zeros no longer forming a comb filter. This technique might be useful if the numerical calculation involved with the zeros was practical.

CHAPTER V

ROUNDOFF NOISE GAIN

The best technique for implementing low roundoff noise frequency sampling filters moves all the poles and zeros inside the unit circle to the same radius until the roundoff noise gain is as low as possible while retaining acceptable passband, stopband, and phase responses after scaling. This section determines the floating-point roundoff noise gain curves as a function of pole-zero radius for both the lowpass and bandpass filters used during this research. Floating-point realization is assumed rather than fixed-point realization due to the difficulties of realizing filters with poles very near the unit circle using fixed-point arithmetic. This work differs from that in [1] since the extended precision capability of floating-point digital signal processors is taken into account to give a much more accurate prediction of noise gain.

Background Equations

Since the floating point roundoff noise gain of a general N^{th}-order floating-point state-space digital filter is identical to the fixed-point roundoff noise gain when the extended precision of a floating-point digital signal processor is used [4], the equations for fixed-point roundoff noise gain will be used.

For a general Nth-order state-space digital filter of the form

$$H(z)=\frac{Y(z)}{U(z)}=d+\frac{\alpha_1 z^{-1}+\alpha_1 z^{-2}+\ldots+\alpha_N z^{-N}}{1+\beta_1 z^{-1}+\beta_1 z^{-2}+\ldots+\beta_N z^{-N}} \tag{21}$$

realized by the state-space structure of equation **(14)**, the output variance is given by

$$\sigma_o^2=\sigma_u^2\sum_{n=0}^{\infty}h^2(n) \tag{22}$$

and, the roundoff noise variance [4] at the output is given by

$$\sigma_y^2=\left[\sum_{n=0}^{\infty}h^2(n)+\sum_{i=1}^{N}K_{ii}W_{ii}\right]\sigma_\gamma^2\sigma_u^2 \ . \tag{23}$$

In equations **(22)** and **(23)** σ_u^2 is the variance of the input. In equation **(23)** K_{ii} represents the diagonal elements of the scaling covariance matrix [3]

$$\mathbf{K}=\sum_{n=0}^{\infty}(\mathbf{A}^n\mathbf{b})(\mathbf{A}^n\mathbf{b})^t=\mathbf{AKA}^t+\mathbf{bb}^t; \tag{24}$$

W_{ii} represents the diagonal elements of the roundoff noise covariance matrix [3]

$$\mathbf{W}=\sum_{n=0}^{\infty}(\mathbf{c}^t\mathbf{A}^n)^t(\mathbf{c}^t\mathbf{A}^n)=\mathbf{A}^t\mathbf{WA}+\mathbf{cc}^t; \tag{25}$$

and the variance of the quantitization error is $\sigma_\gamma^2=(0.167)\cdot 2^{-2B'}$ where *B'* is the number of bits to which the mantissa is quantitized [4]. Since the floating-point signal processors in use today are all single-precision devices, *B'* is typically 24 bits [4]. If the digital filters are quantitized with the same floating point precision and have the same input, the roundoff noise gain, **G**, is determined by the equation

$$G=\frac{\sigma_y^2}{\sigma_\gamma^2\sigma_u^2}=\left[\sum_{n=0}^{\infty}h^2(n)+\sum_{i=1}^{N}K_{ii}W_{ii}\right]. \tag{26}$$

As a function of **G**, the signal-to-noise ratio is

$$SNR=\frac{\sigma_o^2}{\sigma_y^2}=\frac{\sum_{n=0}^{\infty}h^2(n)}{(0.167)\left(2^{-2B'}\right)G}. \tag{27}$$

For a given digital filter realization quantitized to *B'* bits, it is clear from equation **(27)** that minimization of **G** maximizes the *SNR.*

The Lowpass and Bandpass Filters

The roundoff noise gain was calculated at different pole-zero radii for three different second order realizations of both the lowpass and bandpass filters. For the canonical form structure of **(17)**, the minimum noise structure of **(18)**, and the Type III structure of **(19)**, equation **(26)** was used on each filter section and the individual gains were added according to the equation, $Gain=G_1+G_2+\ldots+G_n$, to get the total roundoff noise gain for each digital filter realization. For all three realizations, the lowest roundoff noise obtained occurs at a radius of ($1.0\text{-}2^{-8}$) because the filters have poor stopband and phase response for smaller radii. As expected, the roundoff noise gain is the lowest when the minimum noise structure is used to realize the second order filter sections, and the lowest gain is approximately 485 for both the lowpass and bandpass filters. Using equation **(27)** with **G** replaced by the total roundoff noise gain, *Gain*, of the digital filter and *B'* equal to 24 bits, the signal-to-noise ratio (*SNR*) of each filter can be determined. With a *Gain* of 485, the SNR is 545 decibels for the lowpass filter and 546 decibels for the bandpass filter. Assuming a *SNR* ratio of at least 100 decibels is adequate for most applications, the *Gain* must be less than or equal to 56,000 for the lowpass and bandpass filters to be considered acceptable. Therefore, the pole-zero radius must be less than or equal to $(1.0\text{-}2^{-11})$ for the filters to be acceptable. The roundoff noise gain curves for each of the three realizations and the stopband response for the lowpass filter are shown in Figures 11 and 12, respectively. Similarly, Figures 13 and 14 show the roundoff noise gain curves and the stopband response for the bandpass filter.

Summary

Currently, the best technique for designing low roundoff noise frequency sampling filters is to move both the poles and zeros inside the unit circle to the same radius. The roundoff noise gain can be computed using equation **(26)**, and the *SNR* can be computed using equation **(27)**. The limiting factors in choosing a radius are the stopband response, the phase response, and the roundoff noise gain. For the three realizations studied in this thesis, the lowest obtainable

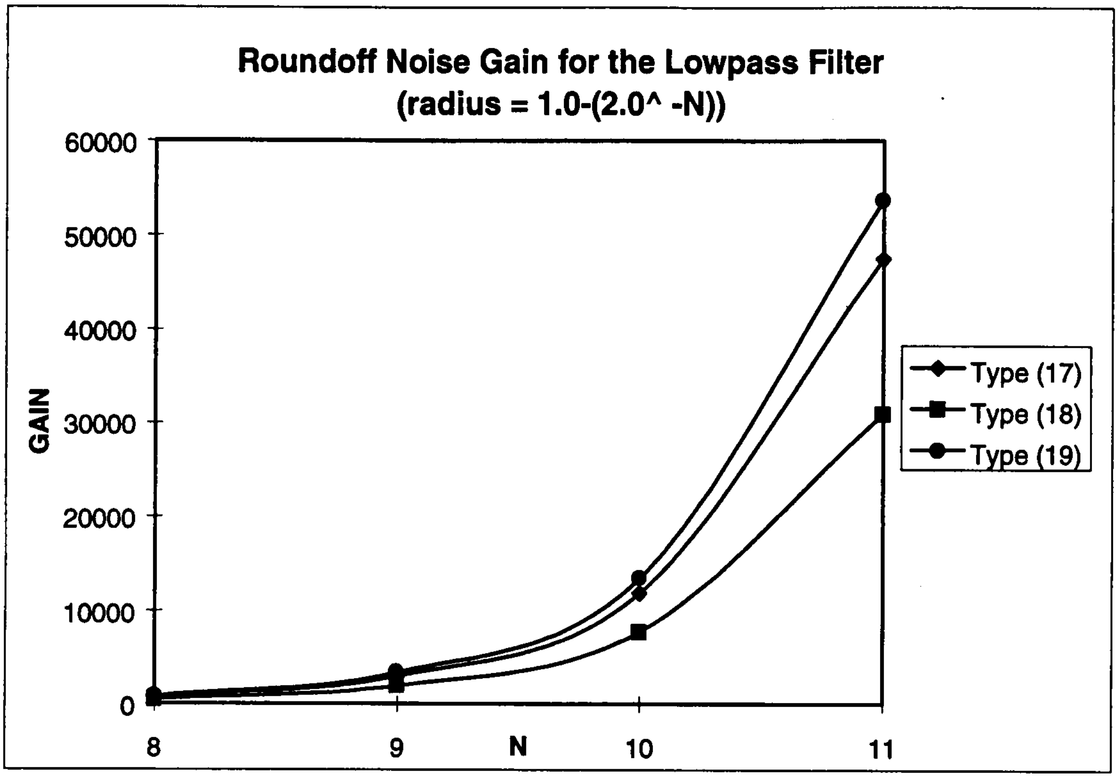

Figure 11: The roundoff noise gain curves for the lowpass filter.

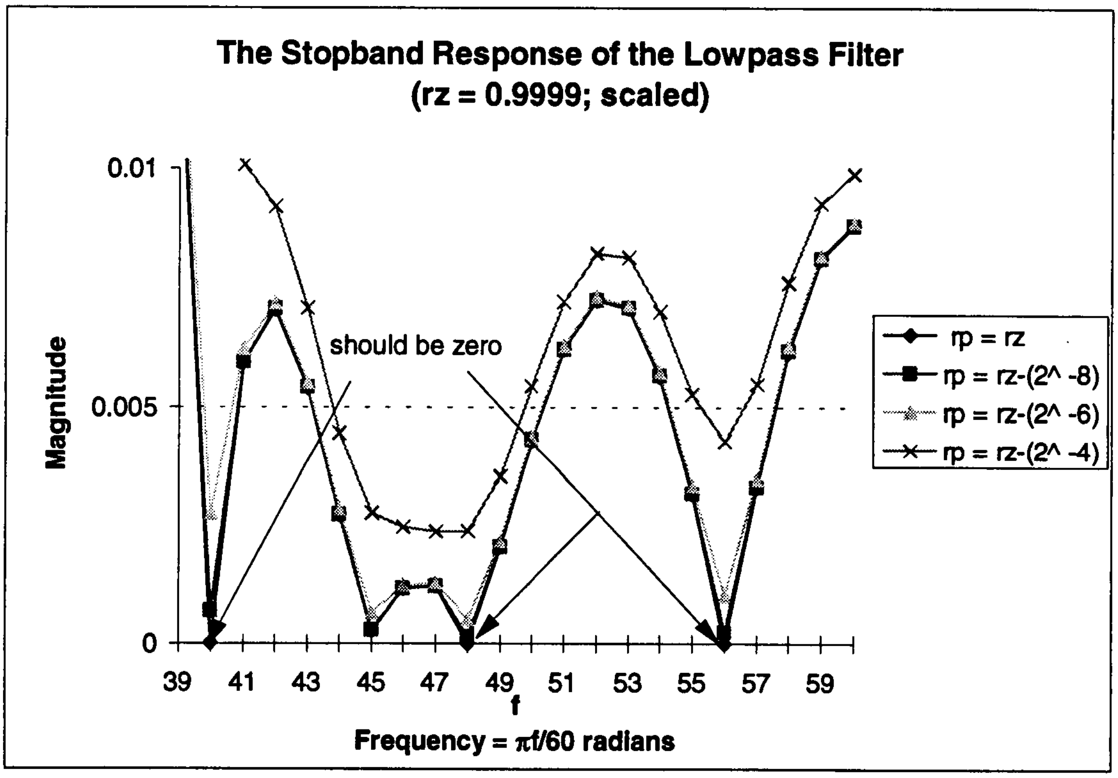

Figure 12: The scaled stopband magnitude response of the lowpass filter.

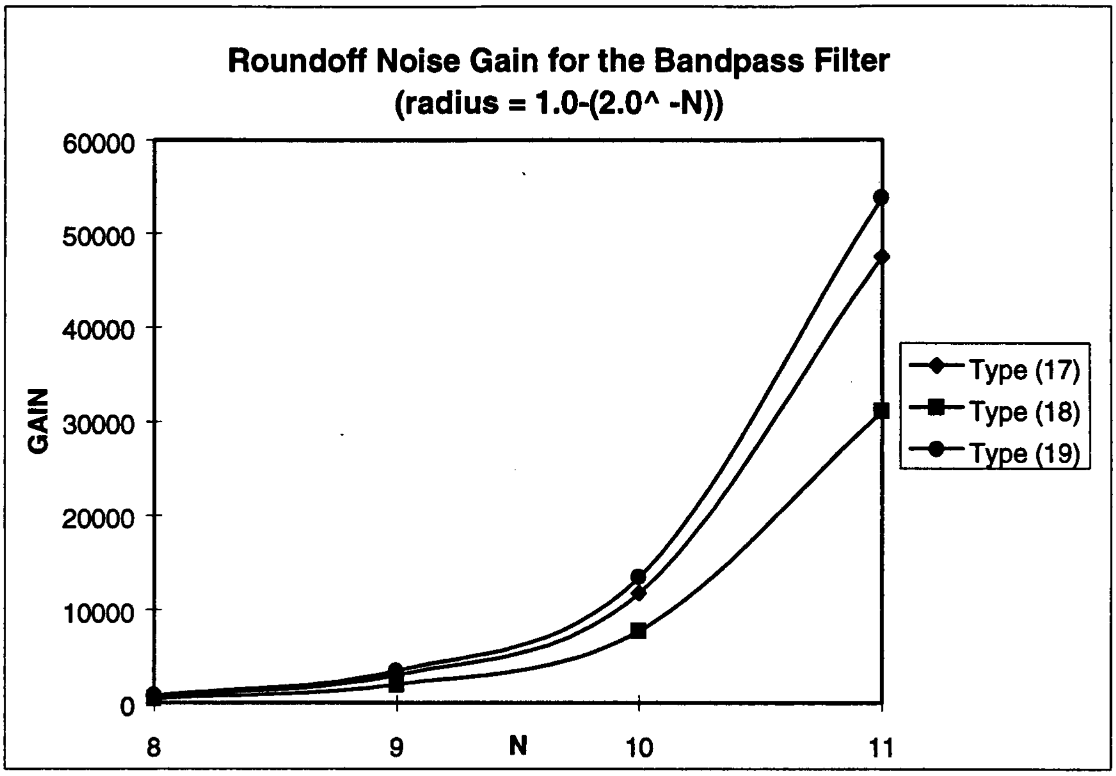

Figure 13: The roundoff noise gain curves for the bandpass filter.

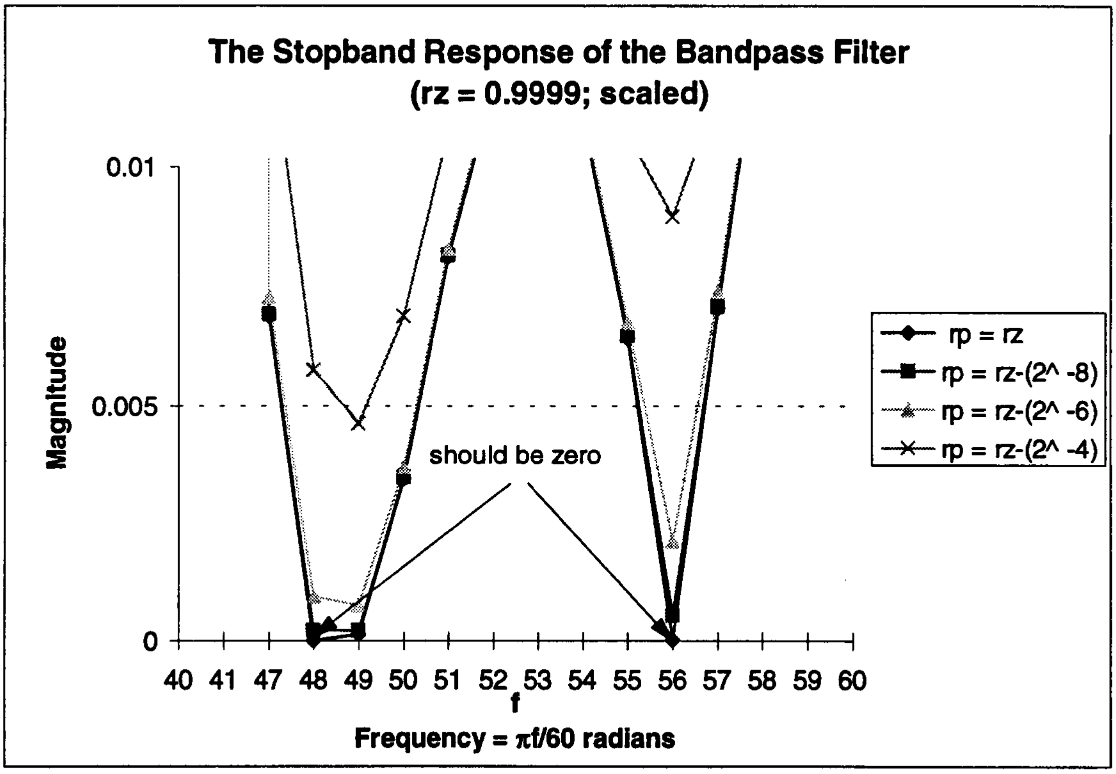

Figure 14: The scaled stopband magnitude response of the bandpass filter.

roundoff noise gain is 485 and occurs at the radius $(1.0\text{-}2^{-8})$ for both lowpass and bandpass filters with acceptable passband and stopband responses. Assuming a *SNR* of at least 100 decibels is adequate for most applications, the acceptable roundoff noise gain range for the lowpass and bandpass filters used during this research is from 485 to 56,000. Therefore, the acceptable range of pole-zero radii is from $(1.0\text{-}2^{-8})$ to $(1.0\text{-}2^{-11})$ where the lower limit is the minimum radius for an acceptable frequency response and the upper limit is the maximum radius for an acceptable signal-to-noise ratio of the filter output.

CHAPTER VI

CONCLUSIONS

Results

It was found that pole-zero separation cannot be used to design low roundoff noise frequency sampling filters because both the magnitude and phase response become too distorted for useful separation distances. The magnitude distortion causes the frequencies $\omega = \frac{2\pi k}{M}$ in the passband to be greatly reduced in magnitude. It is negligible for separation distances that are less than or equal to 2^{-24}, but these are too small to be useful in reducing roundoff noise by moving poles farther inside the unit circle while leaving the zeros close enough to maintain a good stopband response.

Although the frequency response obtained by leaving just the stopband zeros on the unit circle is good, this approach may not be practical because of the excessive computation involved in its implementation. The stopband attenuation is improved because the zeros in that area are on the unit circle and force those frequencies to be zero even when scaled. The phase response of the filters becomes nonlinear as the passband zeros are moved inside the unit circle because they lie on a radius different from the stopband zeros. Although the phase response becomes nonlinear, the effect is negligible for radii greater than $(1.0\text{-}2^{-10})$. One area for further investigation is to attempt to reduce the amount of computation required in the stopband zero compensation technique. This technique might be useful if the amount of numerical calculation involved was practical.

It is concluded that the best technique for designing low roundoff noise frequency sampling filters is to use the minimum noise structure of **(18)** to realize the second order filter sections and move both the poles and zeros inside the unit circle to the same radius. The floating-point realization roundoff noise gain can be computed using essentially the same equations as for fixed-point realizations [4]. The limiting factors in choosing a radius are the stopband response, the phase response, and the roundoff noise gain. The lowest obtainable roundoff noise with acceptable magnitude and phase distortion occurs at a radius of $(1.0\text{-}2^{-8})$ for the filters considered in this thesis. Assuming a *SNR* of at least 100 decibels is adequate for most applications, the acceptable roundoff noise gain range for the lowpass and bandpass filters used during this research is from 485 to 56,000. Therefore, the acceptable range of pole-zero radii is from $(1.0\text{-}2^{-8})$ to $(1.0\text{-}2^{-11})$ where the lower limit is the minimum radius for an acceptable frequency response and the upper limit is the maximum radius for an acceptable signal-to-noise ratio of the filter output.

REFERENCES

REFERENCES

[1] N. L. Bull, *A Study of Frequency-Sampling Filters Realized on Floating-Point Digital Signal Processors*, Masters Thesis, University of Tennessee Space Institute, May 1994.

[2] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, New York, New York: Macmillian Publishing Company, 1992.

[3] B. W. Bomar, "New Second-Order State-Space Structures for Realizing Low Roundoff Noise Digital Filters", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-33 No. 1, pp. 106-110, February 1985.

[4] B. W. Bomar, L. M. Smith, and R. D. Joseph, "Roundoff Noise Analysis of State-Space Digital Filters Implemented on Floating-Point Digital Signal Processors", Proceedings of the IEEE Int. Symposium on Circuits and Systems, Seattle, WA, pp. 2023-2026, April 1995.

VITA

Reuben Daniel Morris, III was born June 21, 1965 to parents Reuben Daniel Morris, Jr and Alfrieda Blackwell Morris. He attended Franklin County High School in Winchester, Tennessee and graduated in 1983. He enrolled at Tennessee Technological University in Cookeville, Tennessee that same year and earned a Bachelor of Science degree in Electrical Engineering in 1989. He began taking classes part-time at The University of Tennessee Space Institute in August 1989, and earned a Master of Science degree in Electrical Engineering in December 1995. He is currently employed with Sverdrup Technology at Arnold Air Force Station, Tennessee.