1.0

1.1

1.25   1.4   1.6

45
50

28   25

32   2.2

36

40   2.0

1.8

MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963

☰SL COORDINATED SCIENCE LABORATORY

ROUNDOFF NOISE IN FIXED-POINT RECURSIVE DIGITAL FILTERS

UNIVERSITY OF ILLINOIS - URBANA, ILLINOIS

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ROUNDOFF NOISE IN FIXED-POINT RECURSIVE DIGITAL FILTERS		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER R-746, UILU-ENG-76-2234
7. AUTHOR(s) DALE JOHN MAYER		8. CONTRACT OR GRANT NUMBER(s) DAAB-07-72-C-0259
9. PERFORMING ORGANIZATION NAME AND ADDRESS Coordinated Science Laboratory University of Illinois at Urbana-Champaign Urbana, Illinois 61801		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 74 p.
11. CONTROLLING OFFICE NAME AND ADDRESS Joint Services Electronics Program		12. REPORT DATE September 1976
		13. NUMBER OF PAGES 69
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

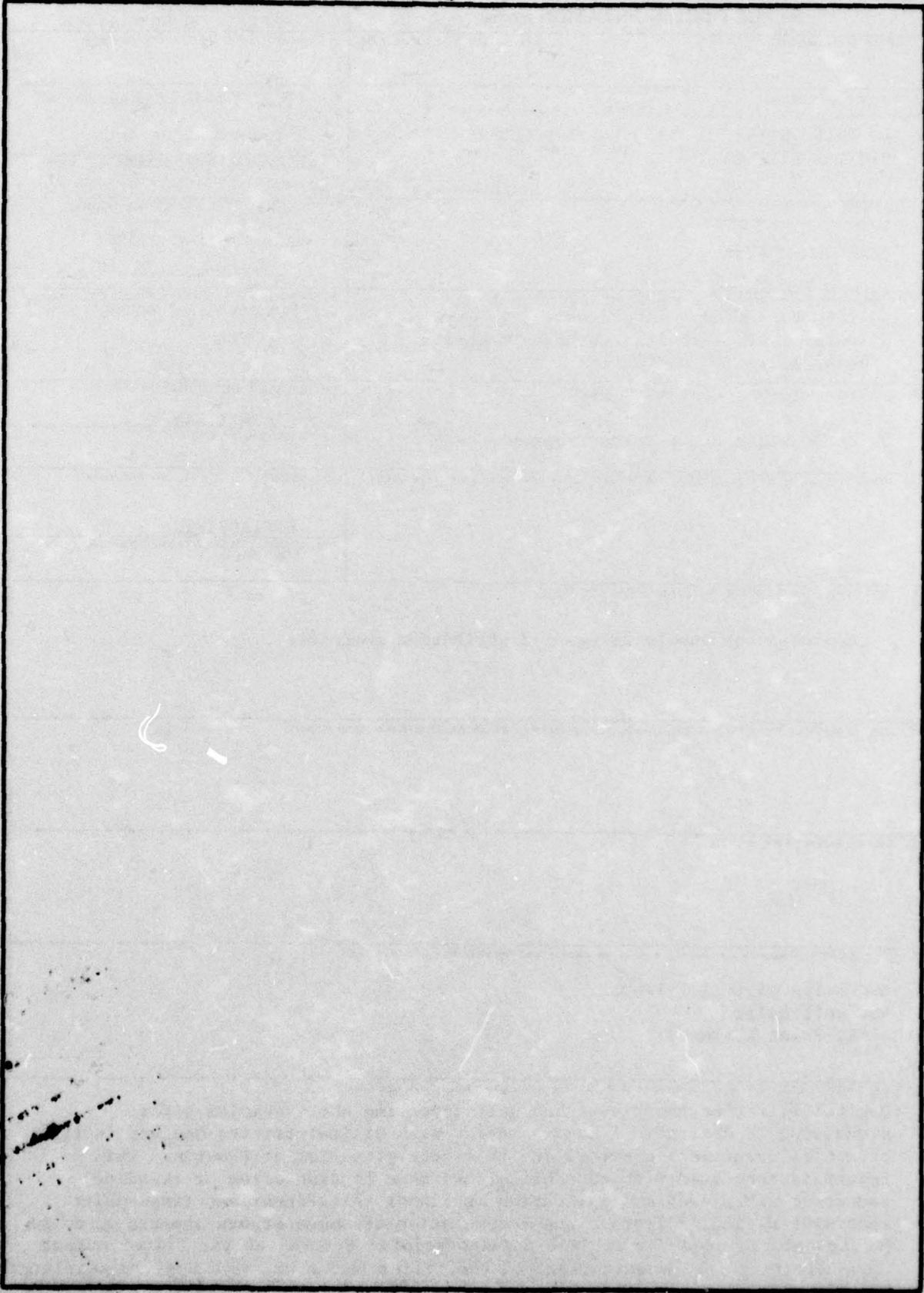Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Recursive Digital Filters
Roundoff Noise
Fixed-Point Arithmetic

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Digital filtering techniques are very appealing where complex signal processing is desired. A major problem with digital filters however is the effect of quantization errors due to finite precision arithmetic. This report is concerned with roundoff errors (due to truncation or rounding) occurring after each multiplication in linear shift-invariant fixed-point recursive digital filters. The accumulation of these errors appears as noise (error uncorrelated) or limit cycles (correlated errors) at the filter output thus limiting the dynamic range of the filter operating with a given wordlength.

ROUNDOFF NOISE IN FIXED-POINT
RECURSIVE DIGITAL FILTERS

by

Dale John Mayer

ROUNDOFF NOISE IN FIXED-POINT RECURSIVE DIGITAL FILTERS

BY

DALE JOHN MAYER

B.S., Southern Methodist University, 1975

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1976

Thesis Advisor: Professor Timothy N. Trick

Urbana, Illinois

ACKNOWLEDGMENT

I would like to thank my advisor, Dr. Timothy N. Trick, for his guidance during the past year as this research progressed.

The assistance of Mrs. Rose Harris in typing part of the thesis is appreciated.

I would also like to thank the University of Illinois for its financial support.

1. INTRODUCTION

Digital filtering techniques are very appealing where complex signal processing is desired. A major problem with digital filters however is the effect of quantization errors due to finite precision arithmetic. This thesis is concerned with roundoff errors(due to truncation or rounding) occuring after each multiplication in linear shift-invariant fixed-point recursive digital filters. The accumulation of these errors appears as noise(error uncorrelated) or limit cycles(correlated errors) at the filter output thus limiting the dynamic range of the filter operating with a given wordlength. This paper will be concerned with uncorrelated errors.

Any transfer function expressible as a rational fraction in z can be realized by many filter configurations. The effect of roundoff errors however is dependent upon the particular structure used. Thus structures are sought which decrease the effect of roundoff errors. Such structures often lead to increased hardware costs and this will also be considered.

Long[1] presents the design of a lowpass Chebyshev filter using a multiple feedback structure which is significantly less sensitive to roundoff errors than the popular cascade structure. He was unable to get similar results with a bandpass design using the multiple feedback structure however. In chapter 3 of this thesis, Long's lowpass filter is transformed to bandpass and to narrowband lowpass in an effort to obtain other multiple feedback filters with low roundoff noise output. Also an equivalent multiple feedback structure using less hardware is

presented. Chapter 2 discusses the modeling of roundoff errors as uncorrelated random noise for analysis purposes, discusses the proper scaling of filter structures, and gives a direct relationship between certain network sensitivities and roundoff noise. Chapter 4 analyzes the second order coupled form structure which can give better roundoff noise performance than the popular direct form realization of a complex pole pair. In chapter 5, the numerical analysis technique used to obtain the noise output of a filter structure is discussed. Each of four transfer functions is realized by four different structures and the roundoff noise is determined for each.

2. ROUNDOFF NOISE, SCALING, and SENSITIVITY

A. Modeling Roundoff Noise

In order to determine the roundoff noise properties of a digital filter structure, the roundoff errors will be modeled as random noise sources injected at the outputs of infinite precision multipliers with the following assumptions:

1. The error sequence from a noise source is a sample function of a stationary random process.

2. Samples from the same noise source are uncorrelated with each other and are uncorrelated with samples from other noise sources and the input sequence.

3. The probability distribution of the error process is uniform over the quantization range.

This method of analysis has been shown to give good results when the quantization level is not low and when the spectral content of the filter input is fairly high[2]. Since the filter is linear shift invariant, the total output noise variance can be determined by superposition.

Each noise source results from the quantization of a multiplier output word and its noise variance is given by (2.1) for the uniform probability density function where d is the quantization increment. Rounding will be assumed so that the mean is zero.

$$\sigma_o^2 = d^2/12 \tag{2.1}$$

The variance of the noise at the filter output due to noise source i is given by

$$\sigma_i^2 = \sigma_o^2 \sum_{k=-\infty}^{\infty} |h_{ni}(k)|^2 \tag{2.2}$$

where $h_{ni}(k)$ is the impulse response from the point where the noise is injected to the filter output. The output noise variance can be represented as a ratio of σ_i^2 to σ_o^2 to eliminate the dependence on the quantization increment d. (2.3) and (2.4) are equivalent to (2.2) and are generally of more use in noise calculations.

$$\frac{\sigma_i^2}{\sigma_o^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H_{ni}(e^{j\omega})|^2 \, d\omega \tag{2.3}$$

$$\frac{\sigma_i^2}{\sigma_o^2} = \frac{1}{2\pi j} \oint H_{ni}(z) H_{ni}(z^{-1}) z^{-1} dz \tag{2.4}$$

$H_{ni}(z)$ is the transfer function in the z-transform domain from the noise source to the output and $H_{ni}(e^{j\omega})$ is the corresponding Fourier transform transfer function. The total noise variance at the filter output is found using superposition and is given by

$$\frac{\sigma_n^2}{\sigma_o^2} = \sum_{i=1}^{M} \frac{\sigma_i^2}{\sigma_o^2} \tag{2.5}$$

where M is the number of noise sources.

B. Scaling

Before the noise variance of a filter can be calculated, overflow constraints must be imposed and met at each summation point. This is accomplished by scaling appropriate multipliers in the filter. The theoretically calculated variance of an improperly scaled filter can be radically different from that of the same filter properly scaled. For example, consider the following lowpass filters.
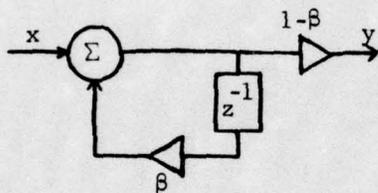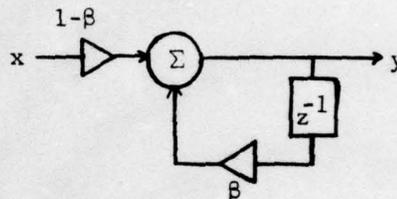

Figure 2.1a. Filter a   Figure 2.1b. Filter b

The transfer functions of both filters are given by (2.6) yet their noise variances due to the feedback multiplier β are as given in (2.7a) and (2.7b) corresponding to Figures 2.1a and 2.1b respectively. Filter b has been scaled in a manner to be described below.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1-\beta}{1-\beta z^{-1}} \qquad 0 < \beta < 1 \tag{2.6}$$

$$\frac{\sigma_a^2}{\sigma_o^2} = \frac{(1-\beta)^2}{1-\beta^2} \tag{2.7a}$$

$$\frac{\sigma_b^2}{\sigma_o^2} = \frac{1}{1-\beta^2} \tag{2.7b}$$

The noise output of filter a is less than that of filter b, however filter a is improperly scaled allowing overflows to occur at the

summation node and therefore is useless.

Various methods for scaling filters exist and each yields slightly different scaling multipliers[2,3]. Scaling in this thesis is accomplished by satisfying the constraint

$$\max_{-\pi < \omega < \pi} \left| G_{i1}(e^{j\omega}) \right| = 1 \qquad (2.8)$$

at each node i where a summation occurs. $G_{i1}(e^{j\omega})$ is the transfer function from the input to the summation node i. This is the L_∞-norm discussed by Jackson[3]. Filters scaled with norms other than that of (2.8) will have slightly different noise properties, however general trends in the noise performance of a structure should be about the same. Two's complement arithmetic is assumed so that overflows may occur internally at a summation node with more than two inputs while the net sum still meets the requirement of (2.8).

C. Use of Extra Scaling Multipliers to Reduce Noise

It was demonstrated above that scaling affects the output noise variance of a filter. It is possible to improve the noise performance of some scaled filter structures by selectively introducing extra multipliers in the circuit. This will increase the cost of the filter, but if power-of-two shifts are used to accomplish the scaling as described by Jackson[2] the increase in cost will not be as much as it would be if full precision multipliers were used. Consider the internal summation node i of a filter as illustrated in Figure 2.2a. Such a node would occur in the common cascade filter structure for instance. The
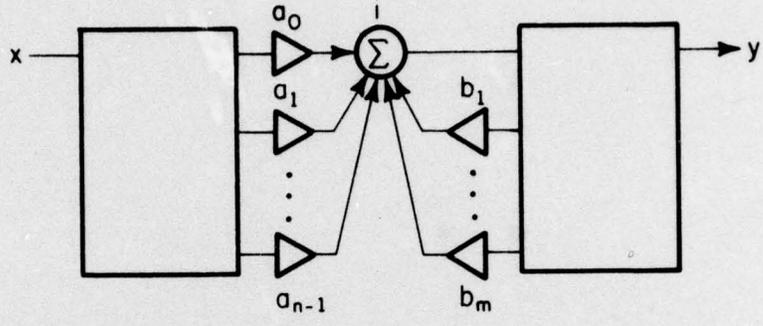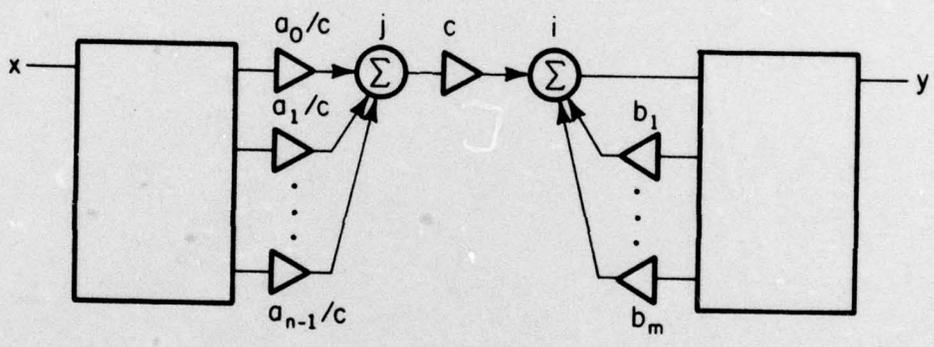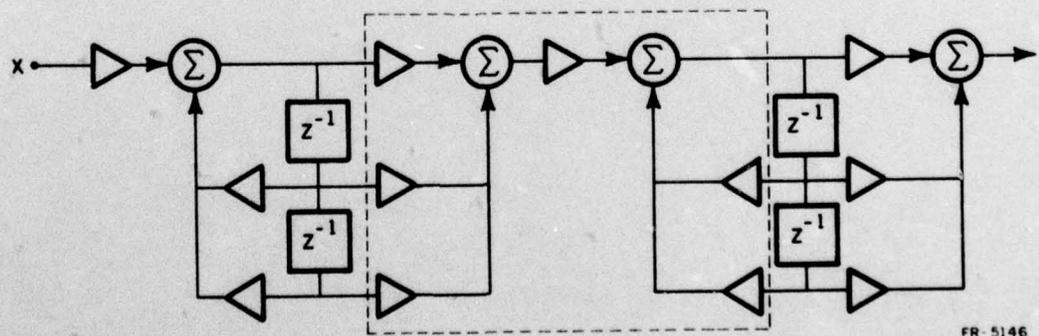
Figure 2.2a

Figure 2.2b

FR-5147

Figure 2.2. Introduction of Multiplier c
to Reduce Noise

FR-5146

Figure 2.3. Noise Reduction for Cascaded
Second Order Sections

a's include proper scaling for node i. The noise output due to errors introduced at this node is given by

$$\frac{\sigma_n^2}{\sigma_o^2} = (m+n)\,\frac{\sigma_i^2}{\sigma_o^2}\,. \tag{2.9}$$

In Figure 2.2b, an extra scaling multiplier c is introduced so that there are two summation nodes j and i. The value c is chosen so that when the a's are divided by c, the new overflow constraint at node j is met exactly. By multiplying by c at the output of node j, the overflow constraint at node i is also met exactly as before. The noise variance from the m+n+1 multipliers in Figure 2.2b is given by

$$\frac{\sigma_n^2}{\sigma_o^2} = (m+1+c^2 n)\,\frac{\sigma_i^2}{\sigma_o^2}\,. \tag{2.10}$$

From (2.9) and (2.10) it can be seen that the introduction of the multiplier c will reduce the output noise if

$$c < \sqrt{\frac{n-1}{n}}\,. \tag{2.11}$$

The condition of 2.11 is often met. The effect c has on reducing the noise becomes greater as c becomes smaller. For high Q filters, c could easily take on values of 0.1 or less. In such a case the noise introduced by the n multipliers entering node j is insignificant compared to that of the m+1 multipliers entering node i. The noise from the m+n+1 multipliers for this case is approximately

$$\frac{\sigma_n^2}{\sigma_o^2} = (m+1) \frac{\sigma_i^2}{\sigma_o^2} \tag{2.12}$$

Figure 2.3 shows a cascade of two second order sections with the multipliers of interest outlined by the dotted box. Here, n=3 and m=2. If the second section has high Q, then the improvement in noise generated by the outlined multipliers over an equivalent filter without the scaling multiplier c is -2.2dB as calculated below.

$$\text{Improvement} = 10 \log \frac{m+1}{m+n} = -2.2\text{dB} \tag{2.13}$$

The overall improvement in total output noise of a filter obtained by this technique will be determined by the rest of the filter and may or may not result in a significant improvement in the total output noise.

D. A Relationship Between Roundoff Noise
and Certain Network Sensitivities

This section shows how the roundoff noise may be expressed exactly in terms of certain network sensitivities. Fettweis[4] has developed a similar relationship however it does not express the total roundoff noise exactly.

Fettweis has shown[5] that the sensitivity of the input-output transfer function H_{n1} with respect to a multiplier α directed from node k to node i is given by

$$\frac{\partial H_{n1}}{\partial \alpha} = H_{k1} \, H_{ni} \tag{2.14}$$

H_{ni} is the transfer function from node i to the output and is the function required for the computation of the gain of the noise power from node i. Solving (2.14) for H_{ni} and substituting this in (2.3) yields

$$\frac{\sigma_i^2}{\sigma_o^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1}{H_{k1}} \, \frac{\partial H_{n1}}{\partial \alpha} \right|^2 d\omega \tag{2.15}$$

This expression gives the roundoff noise variance in terms of the sensitivity of H_{n1} with respect to α divided by the gain H_{k1} from the input to the multiplier source node k. This gain H_{k1} is dependent on the filter structure. However if k=n, then (2.14) can be written as

$$\frac{\sigma_i^2}{\sigma_o^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \frac{1}{H_{n1}} \, \frac{\partial H_{n1}}{\partial \alpha} \right|^2 d\omega \tag{2.16}$$

H_{n1} is the input-output transfer function and is not related to the filter structure. Thus the noise variance due to a noise source at node i is expressed exactly in terms of the sensitivity of a feedback multiplier α from the output to node i. This multiplier is often present in a structure although it need not be for (2.16) to hold.

3. THE MULTIPLE FEEDBACK STRUCTURE

A. Motivation

The multiple feedback(MFB) or leapfrog digital filter structure is shown in Figure 3.1. The transfer functions $T_i(z)$ are of the form

$$T_i(z) = \frac{\alpha_{0i} + \alpha_{1i}z^{-1} + \alpha_{2i}z^{-2}}{1 - \beta_{1i}z^{-1} - \beta_{2i}z^{-2}} \tag{3.1}$$

and are realized by first or second order sections. Long[1] used the technique of numerical coefficient matching in the z-domain by solving simultaneous nonlinear equations to design several filters of this type. He determined the noise variances of these filters and compared them to those of filters with the same transfer functions but realized with different structures. In particular, comparisons were made to the cascade of direct form second order sections. He designed three fifth order lowpass Chebyshev filters using first order sections for the $T_i(z)$'s as in Figure 3.2. The α_{02} and α_{04} multipliers are missing in order to make the filter physically realizable. The transfer functions of these filters had 0.5 dB ripple and cutoff frequencies of $\frac{\pi}{2}$, $\frac{\pi}{4}$, and $\frac{\pi}{8}$ and are designated LP2, LP4, and LP8 respectively. For the wideband LP2 transfer function, the MFB design was about 4 dB worse than the cascade LP2 design. The LP4 MFB and cascade realizations showed no appreciable difference but the LP8 MFB realization proved to be 10 dP better than the LP8 cascade structure indicating that for narrowband lowpass filters the MFB structure can significantly improve the signal to noise ratio of a filter. Long also designed 6th order bandpass
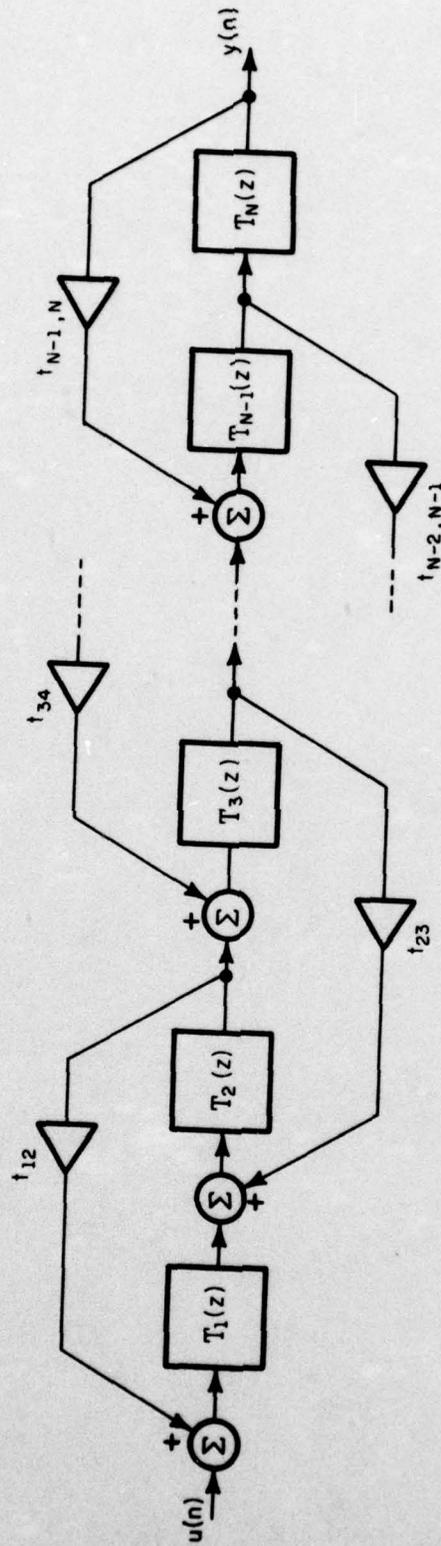
Figure 3.1. General multiple feedback configuration.

12

FR-3737

13
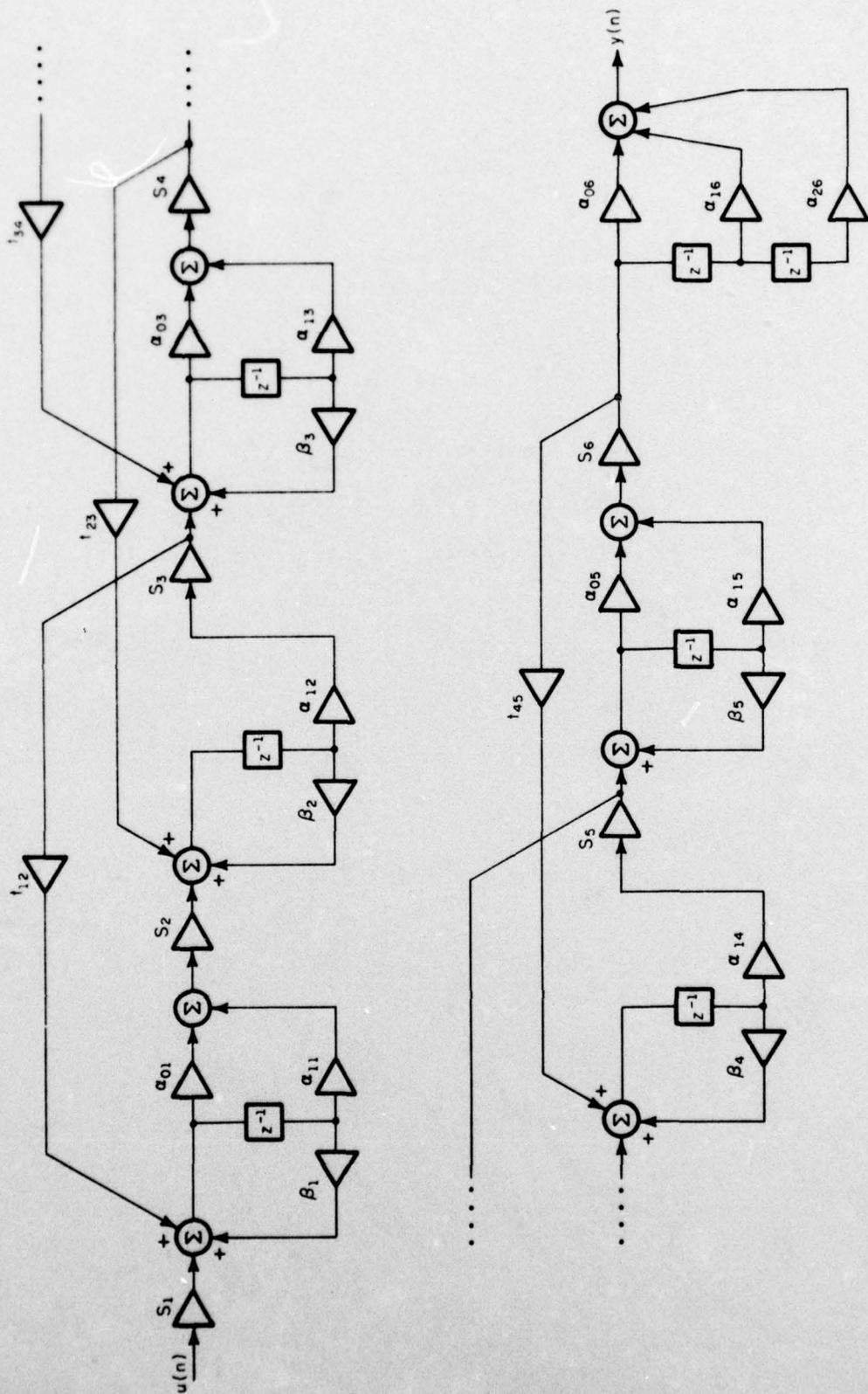
Figure 3.2. Multiple feedback realization of fifth order Chebyshev lowpass filter.

FR-3761

Chebyshev MFB filters using second order sections for the $T_i(z)$'s as shown in Figure 3.3. These filters performed poorly compared to the cascade structure. He attributes this result to a poor selection of the free parameters in the filter design stage. Little is known about the proper selection of the free parameters to obtain low noise filters.

The approach taken in this thesis is to take a proven low noise MFB design and perform frequency transformations on the structure in an effort to create low noise bandpass and narrowband lowpass filters. Long's LP8 MFB filter was used for this.

B. Modified MFB Structure

For the MFB structure in Figure 3.2, the transfer function for $T_i(z)$ where i is even has a zero at z=0 in order to make the filter realizable. The zeroes missing at z=-1 are added at the end of the filter. This causes the filter in Figure 3.2 to have two unnecessary zeros at z=0 which add phase shifts to the filter output. These extra zeroes can be removed by reconfiguring the filter as shown in Figure 3.4. This new configuration also requires less hardware to implement. To show that the filter of Figure 3.4 realizes the same magnitude function as the Chebyshev filter of Figure 3.2, consider the general case depicted in Figures 3.5a and 3.5b where

$$D_i = 1 - \beta_i z^{-1} \tag{3.2a}$$

$$N_i = \alpha_{0i} + \alpha_{1i} z^{-1} \tag{3.2b}$$

The transfer functions for Figures 3.5a and 3.5b are given by (3.3a) and
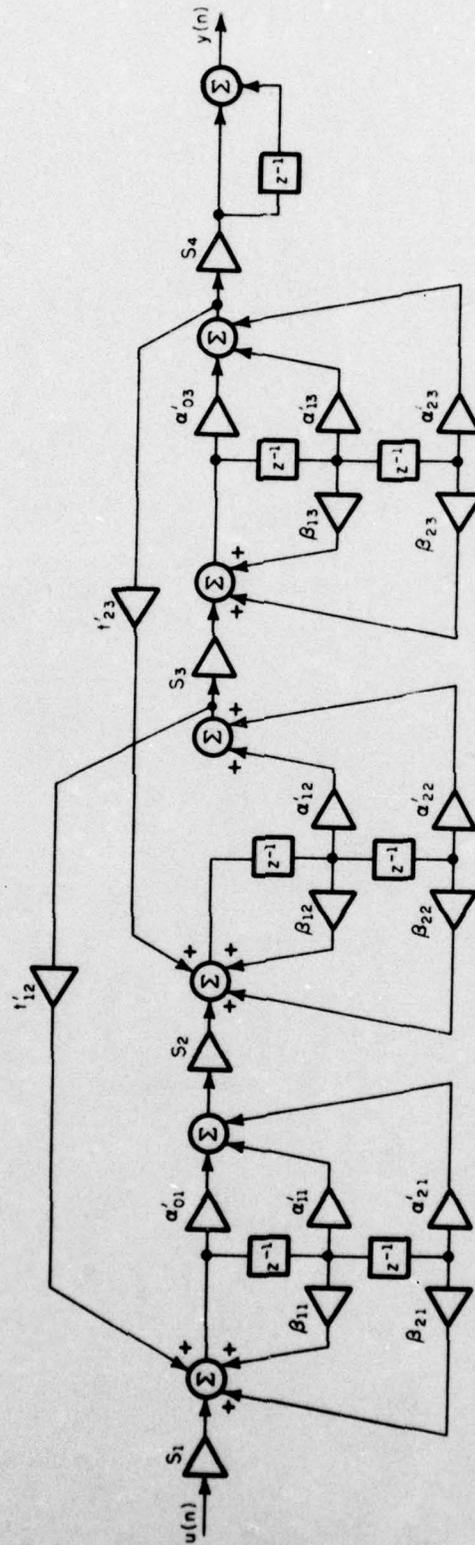
Figure 3.3 Multiple feedback realization of sixth order Chebyshev bandpass filter.

FR-3762

Figure 3.4 The modified multiple feedback structure.

Figure 3.5a Long's Realization

Figure 3.5b Modified Realization
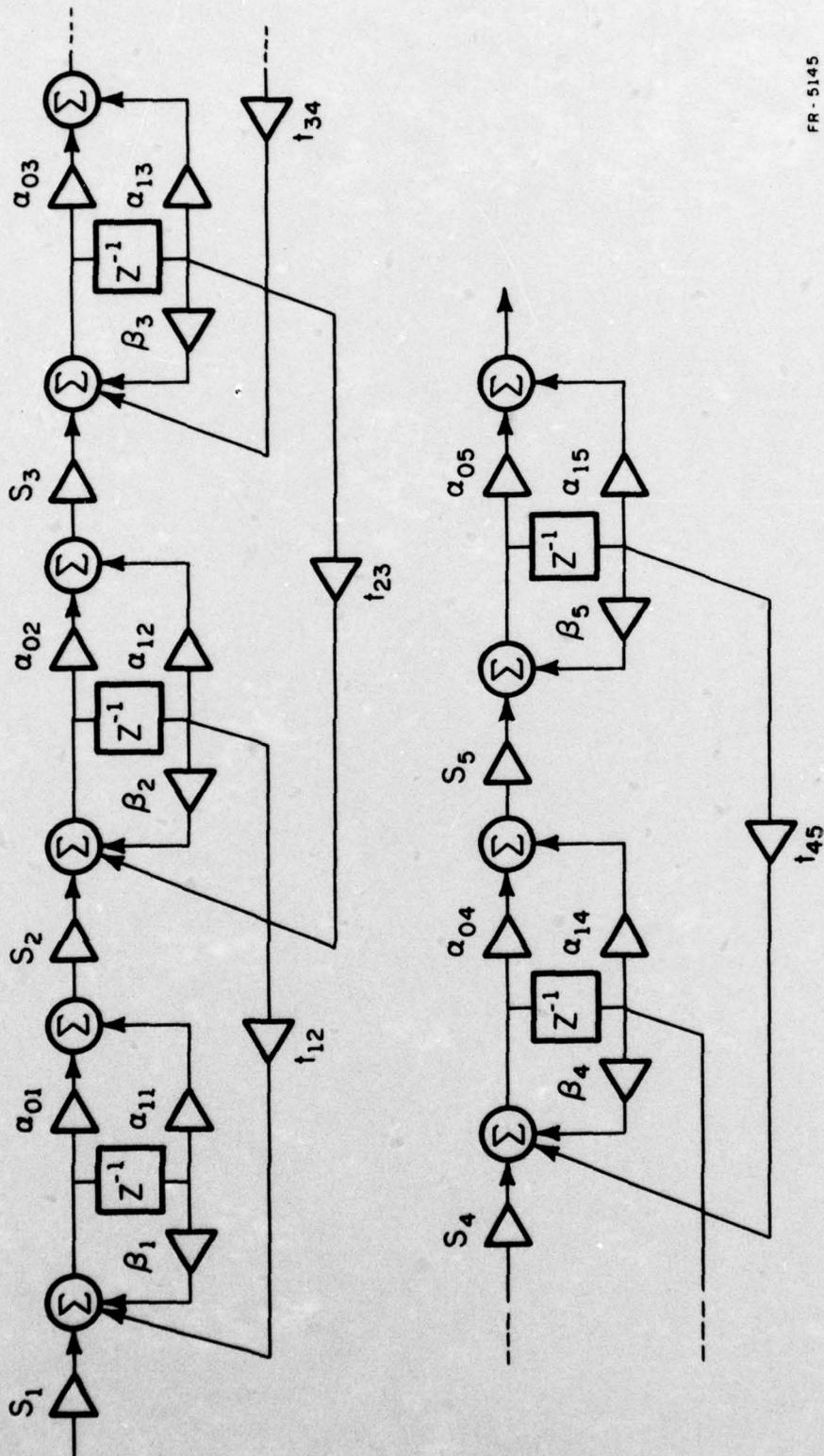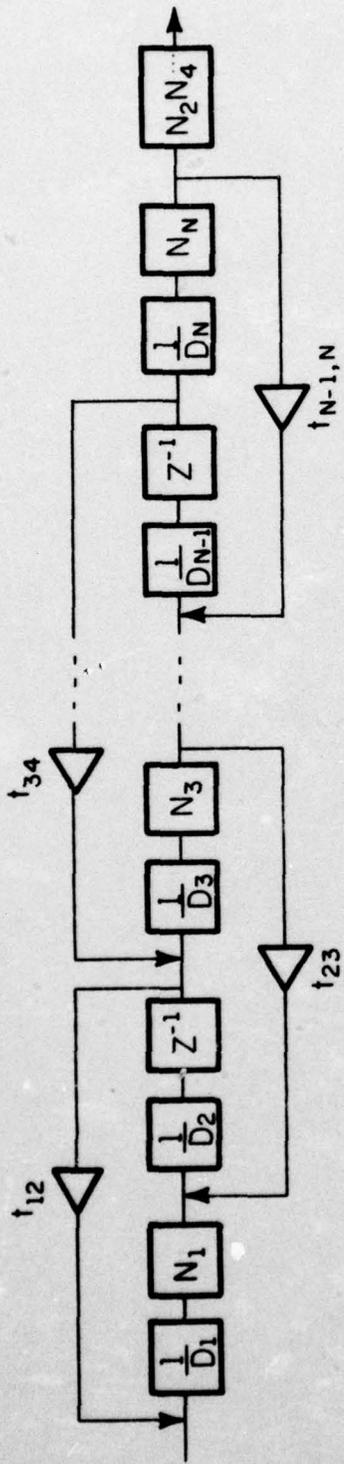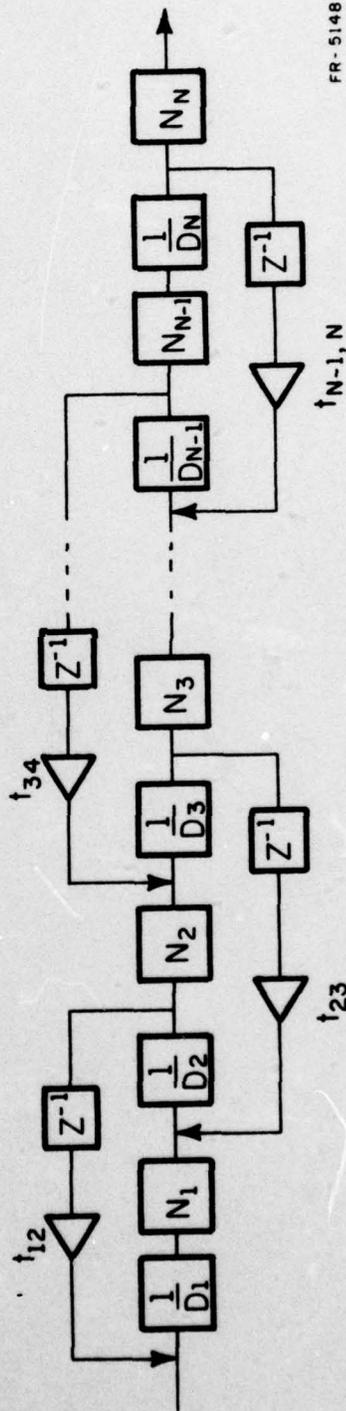
FR-5148

Figure 3.5 A Modification of the Multiple Feedback Structure

(3.3b) respectively where N is the order of the filter.

$$H_a(z) = \frac{z^{-k} \prod\limits_{i=1}^{N} \frac{N_i}{D_i}}{\Delta_a} \qquad (3.3a)$$

where $\qquad k = N/2 \qquad$ N even

$\qquad\qquad\quad = (N-1)/2 \qquad$ N odd

and

$$H_b(z) = \frac{\prod\limits_{i=1}^{N} \frac{N_i}{D_i}}{\Delta_b} \quad . \qquad (3.3b)$$

The denominators of (3.3a) and (3.3b) are given by

$\Delta = 1-(\text{sum of all individual loop gains})+(\text{sum of gain products}$
$\qquad$ of all possible combinations of two non-touching loops)
$\qquad$ -(sum of the gain products of all possible combinations
$\qquad$ of 3 non-touching loops)+... $\qquad (3.4)$

The gain products in (3.4) for Figure 3.5a consist of terms of the form

$$t_{i,i+1} z^{-1} \frac{N_i}{D_i D_{i+1}} \qquad \text{for} \quad i \text{ odd}$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.5a)$$

$$t_{i,i+1} z^{-1} \frac{N_{i+1}}{D_i D_{i+1}} \qquad \text{for} \quad i \text{ even.}$$

The gain products in (3.4) for Figure 3.5b consist of terms of the form

$$t_{i,i+1} z^{-1} \frac{N_i}{D_i D_{i+1}} \qquad \text{for all } i. \qquad (3.5b)$$

Thus the denominators of $H_a(z)$ and $H_b(z)$ will be the same when $N_i=N_{i+1}$ for all even i. This is the case for the Chebyshev filter of figure 3.2 since all zeroes lie at z=-1. The LP8 filter was designed using the new structure of Figure 3.4. The noise variances were computed(as in Chapter 5) for both old and new structures and differed by only 0.1 dB.

C. Frequency Transformations

In order to transform from a lowpass filter of cutoff frequency θ_c to a bandpass filter centered at ω_c with bandwidth B the following substitution is made for z^{-1} in the lowpass prototype[6]

$$z^{-1} \rightarrow \frac{z^{-2} - az^{-1} + b}{bz^{-2} - az^{-1} + 1} \tag{3.6}$$

where

$$a = \frac{2\alpha k}{k+1} \qquad b = \frac{k-1}{k+1}$$

and

$$\alpha = \frac{\cos \omega_c}{\cos(B/2)}$$

$$k = \cot(B/2)\tan(\theta_c/2).$$

This substitution transforms the general first order section shown in Figure 3.6a to the second order section shown in Figure 3.6b. To transform the MFB filters in Figures 3.2 and 3.4 to bandpass filters, the substitution depicted in Figures 3.6a and 3.6b is made for each first order section occuring in the prototype MFB filters. For the transformed filter to have no delay-free loops, it is required that $B = \theta_c$ ($k=1, a=\alpha, b=0$). Performing this transformation on the filter of Figure 3.2 is straightforward and results in the filter shown in Figure 3.7 where use has been made of the fact that all zeroes are at $z=-1$ in Long's LP8 filter. It should be noted that four more full precision multipliers, two more delays, and three more summations are necessary here than would be necessary for a 10th order MFB bandpass filter designed like the 6th order filter in Figure 3.3.
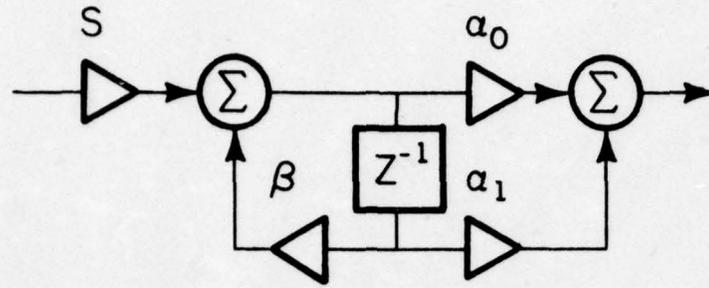
Figure 3.6a General First Order Lowpass Section
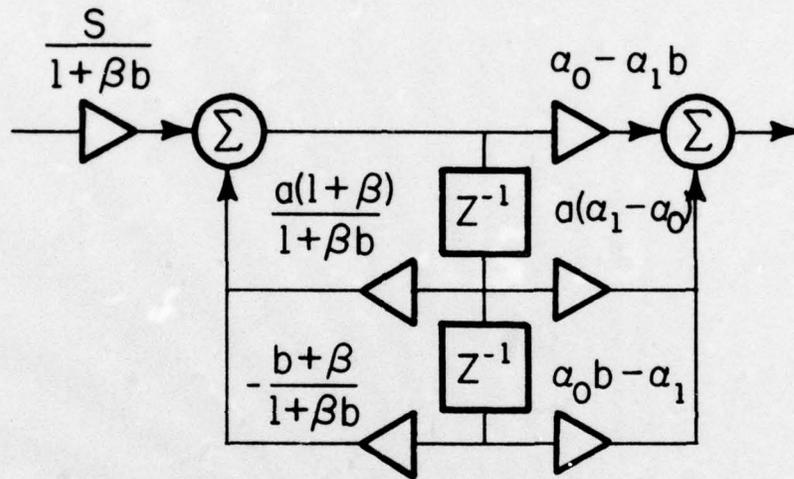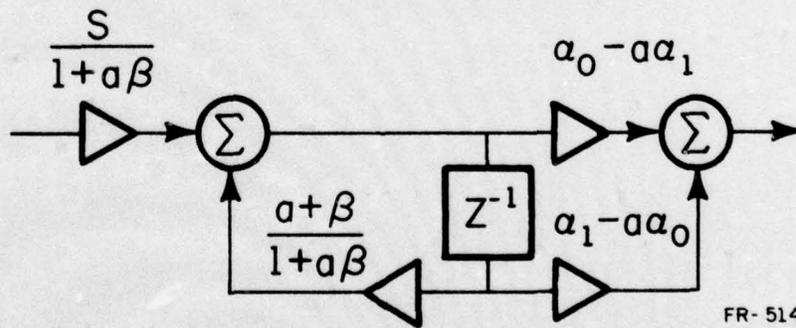

Figure 3.6b Bandpass Transformation

FR- 5149

Figure 3.6c Lowpass Transformation

Figure 3.6 The Transformation of a First Order Lowpass
Section to Bandpass and Lowpass
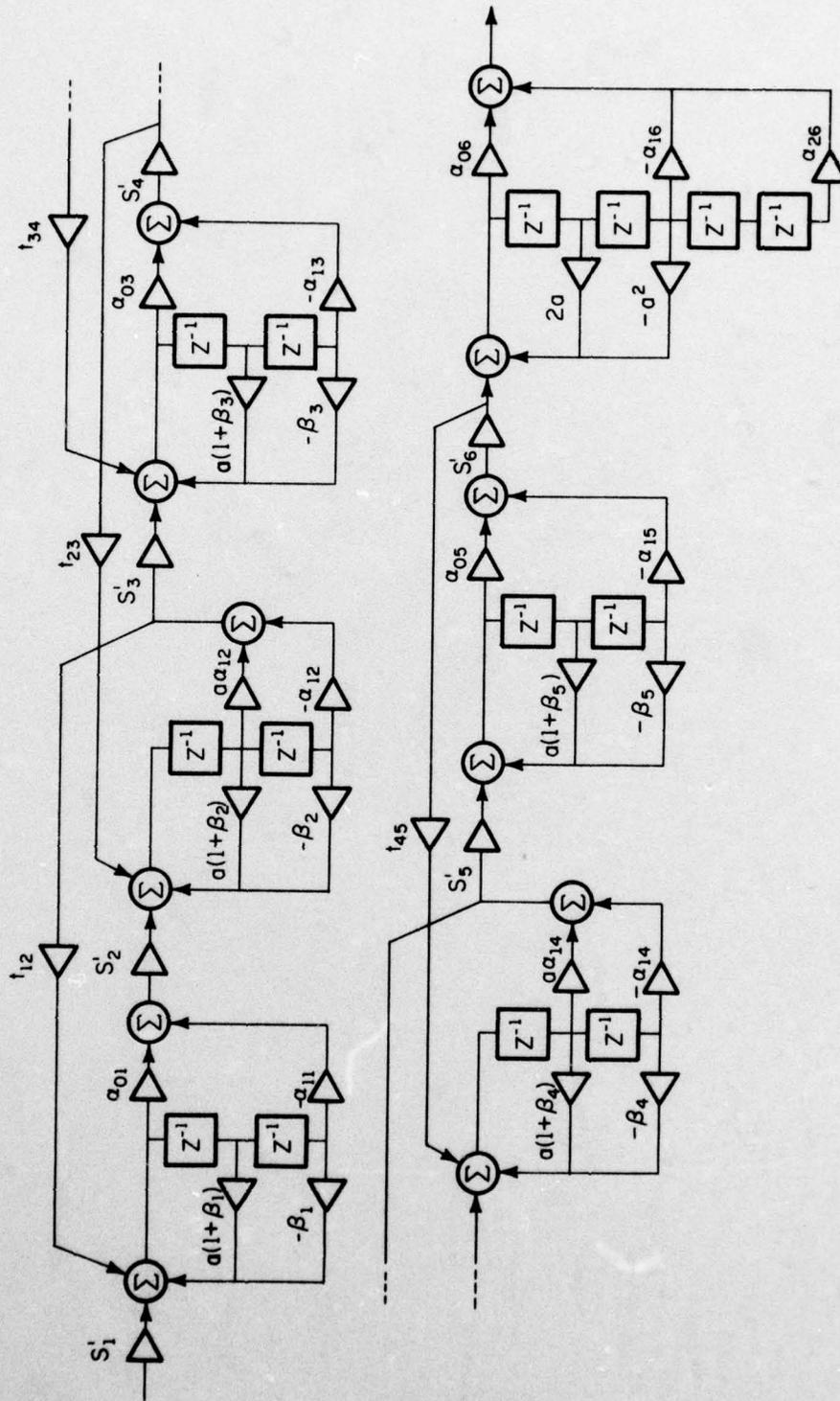
Figure 3.7 Bandpass Transformation of Long's Fifth Order Lowpass MFB Filter

Performing this transformation on the modified MFB filter in Figure 3.4 is slightly more complicated. Each feedback path is considered to be a multiplier α_1 in Figure 3.6a and results in two feedback paths. The transformed filter is shown in Figure 3.8 and has the same magnitude characteristic as the filter of Figure 3.7. This filter requires the same number of full precision multipliers as the other transformed bandpass filter(Figure 3.7) however it requires fewer delays and summations. In an analysis of the noise variance of a bandpass filter design, the noise variance of the filter in Figure 3.8 was only 0.7 dB worse than that of the filter in Figure 3.7. Since the noise properties of both the old and modified MFB structures appear to be similar, only the modified structure and its transforms will be considered in the noise analysis of Chapter 5.

In order to transform from a lowpass filter of cutoff frequency θ_c to another lowpass filter of cutoff frequency ω_c, the following substitution is made for z^{-1} in the lowpass prototype[6]

$$z^{-1} \to \frac{z^{-1} - a}{1 - az^{-1}} \tag{3.8}$$

where

$$a = \frac{\sin\left(\frac{\theta_c - \omega_c}{2}\right)}{\sin\left(\frac{\theta_c + \omega_c}{2}\right)}$$

This substitution transforms the general first order section of Figure 3.6a to that of 3.6c. Using this transformation on the modified MFB structure(Figure 3.4) yields the filter shown in Figure 3.9. It should be noted that this structure is physically unrealizable as it contains
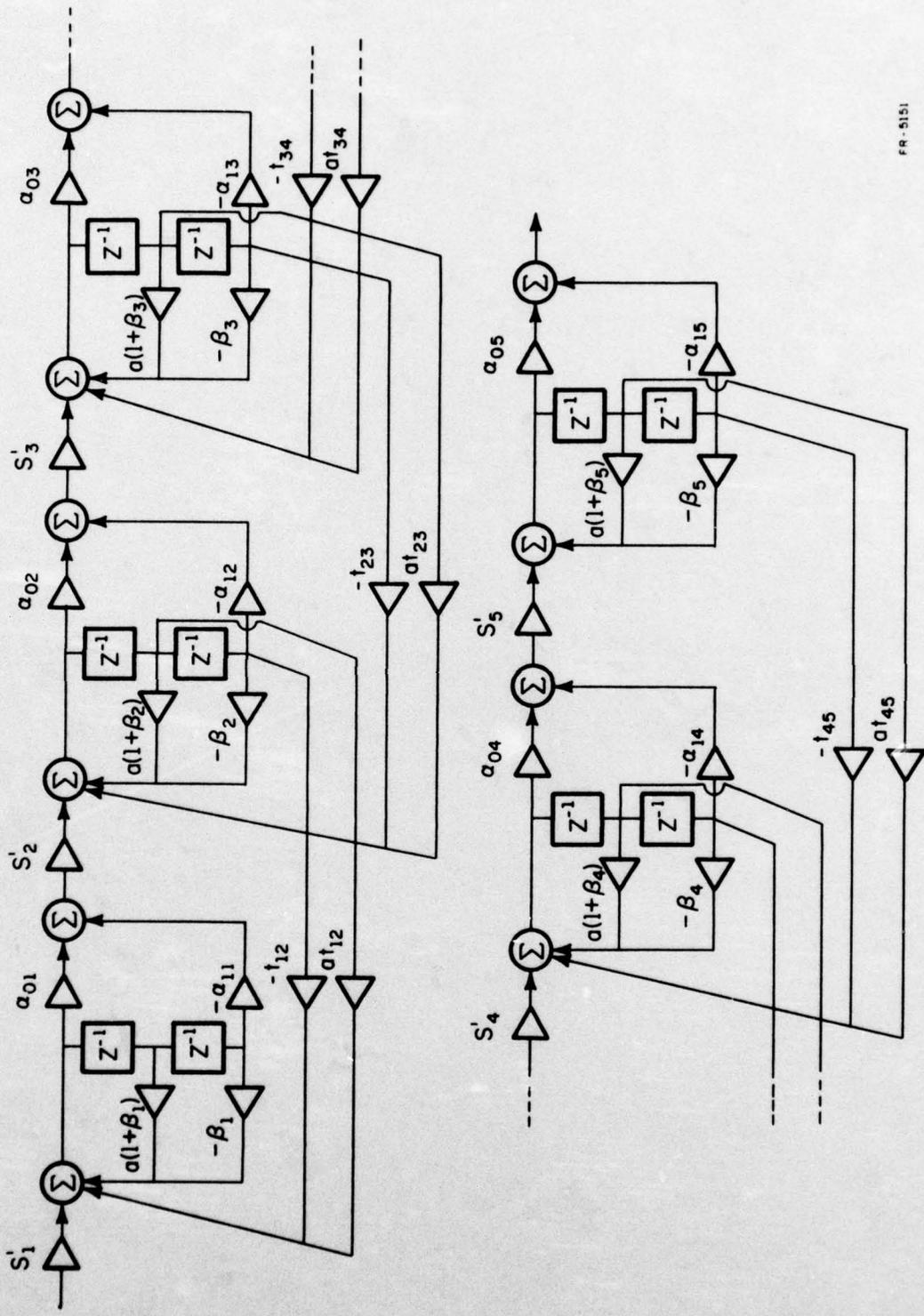
Figure 3.8 Bandpass Transformation of the Modified MFB Structure

23

FR-5151
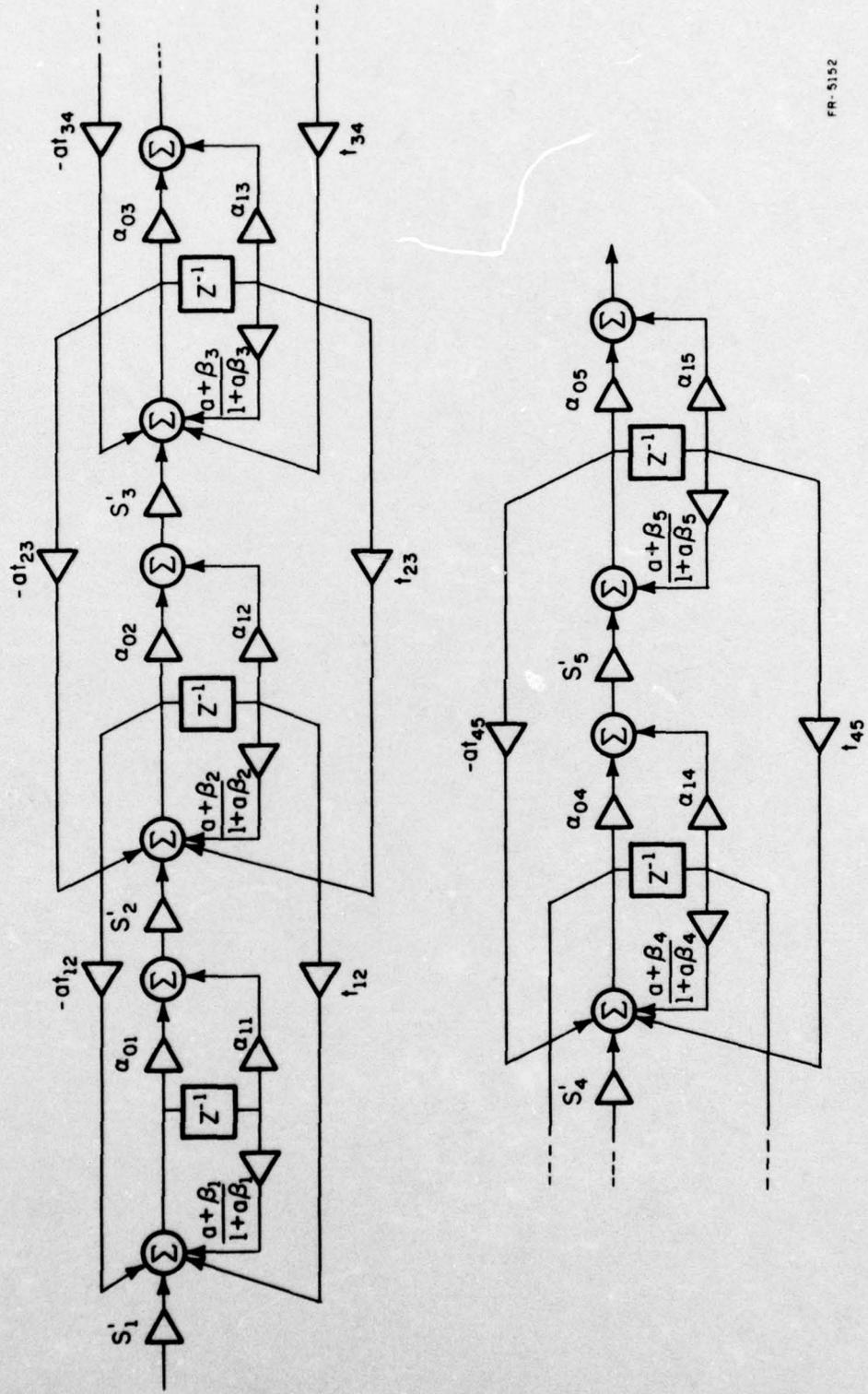
Figure 3.9 Lowpass Transformation of the Modified MFB Structure

24

FR-5152

delay free loops. It is included for comparison purposes only in

Chapter 5.

3. THE COUPLED FORM

A. Structure

The coupled form realization of a complex pole pair as proposed by Gold and Radar[7] is illustrated in Figure 4.1.
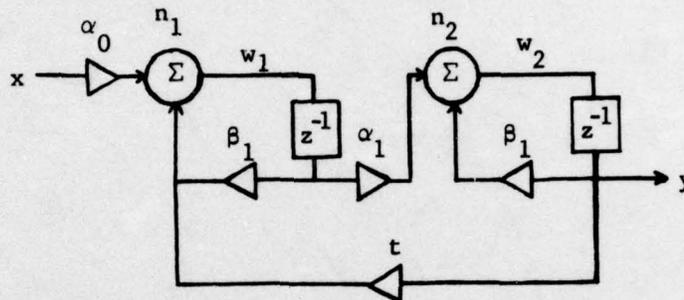
Figure 4.1. The Coupled Form

The noise properties of this structure will be compared to those of the direct form implementation shown in Figure 4.2.
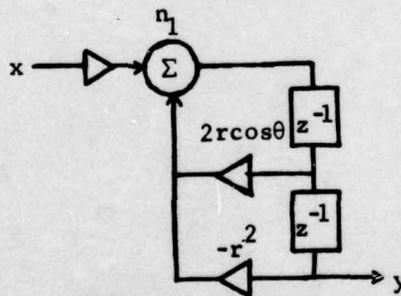
Figure 4.2. The Direct Form

The transfer function for Figure 4.1 is

$$H(z) = \frac{\alpha_0 \alpha_1 z^{-2}}{1 - (\beta_1 + \beta_2)z^{-1} + (\beta_1 \beta_2 - \alpha_1 t)z^{-2}} \qquad (4.1)$$

and that of Figure 4.2 is

$$H(z) = \frac{\alpha z^{-2}}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}} \qquad (4.2)$$

Thus these two filters realize the same pole pair if

$$\beta_1 + \beta_2 = 2r\cos\theta \qquad (4.3a)$$

$$\beta_1\beta_2 - a_1 t = r^2 \qquad (4.3b)$$

α_0 and α_1 are set by the scaling required at nodes n_1 and n_2 in Figure 4.1. There is one free parameter available in designing this filter. It is desired to select that parameter to reduce the output noise.

B. Sensitivity Analysis

The sensitivity function S_y^x is defined by

$$S_y^x = \frac{y}{x} \frac{\partial x}{\partial y} \qquad (4.4)$$

The sensitivity functions of (4.1) are found to be

$$S_{\beta_1}^H = \frac{\beta_1(1 - \beta_2 e^{-j\omega})e^{-j\omega}}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \qquad (4.5a)$$

$$S_{\beta_2}^H = \frac{\beta_2(1 - \beta_1 e^{-j\omega})e^{-j\omega}}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \qquad (4.5b)$$

$$S_t^H = \frac{(\beta_1\beta_2 - r^2)e^{-2j\omega}}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \qquad (4.5c)$$

$$S_{\alpha_1}^H = 1 + S_t^H \tag{4.5d}$$

$$S_{\alpha_0}^H = 1 \tag{4.5e}$$

where z has been replaced by $e^{j\omega}$ to point out the dependence on frequency. (4.5c) can be minimized with respect to β_1 and β_2 to obtain

$$\beta_1 = \beta_2 = r\cos\theta. \tag{4.6}$$

Also, since (4.5a) and (4.5b) together are symmetric in β_1 and β_2, this seems a good choice for them. This selection is given by Oppenheim and Schafer[6].

C. Noise Analysis

To perform a closed form noise analysis, the transfer functions from nodes n_1 and n_2 to the output are required. These are given by

$$\frac{y}{n_1} = \frac{\alpha_1 z^{-2}}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}} \tag{4.7a}$$

$$\frac{y}{n_2} = \frac{(1 - \beta_1 z^{-1})z^{-1}}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}}. \tag{4.7b}$$

The transfer function from node n_1 to the output of the direct form is

$$\left(\frac{y}{n_1}\right)_{direct} = \frac{z^{-2}}{1 - 2r\cos\theta z^{-1} + r^2 z^{-2}}. \tag{4.7c}$$

It can be seen that the magnitudes of both (4.7a) and (4.7b) can be less

than the magnitude of (4.7c). Substituting (4.7) into (2.4) and using Jury's tables of that integral[8], the output noise variances for a single noise source at each node are found to be

$$\frac{\sigma^2_{n_1}}{\sigma^2_0} = \frac{\alpha^2_1(1+r^2)}{(1-r^2)(1+r^4-2r^2\cos2\theta)} \qquad (4.8a)$$

$$\frac{\sigma^2_{n_2}}{\sigma^2_0} = \frac{(1+\beta^2_1)(1+r^2)-4\beta_1 r\cos\theta}{(1-r^2)(1+r^4-2r^2\cos2\theta)} \; . \qquad (4.8b)$$

The output noise variance for a noise source at node n_1 of the direct form version is given by

$$\left(\frac{\sigma^2_{n_1}}{\sigma^2_0}\right)_{direct} = N_2 = \frac{1+r^2}{(1-r^2)(1+r^4-2r^2\cos2\theta)} \; . \qquad (4.9)$$

Therefore (4.8a) and (4.8b) can be rewritten as

$$\frac{\sigma^2_{n_1}}{\sigma^2_0} = \alpha^2_1 N_2 \qquad (4.10a)$$

$$\frac{\sigma^2_{n_2}}{\sigma^2_0} = (1+\beta^2_1 - \frac{4r\cos\theta}{1+r^2}\beta_1)N_2. \qquad (4.10b)$$

Thus if $\alpha_1 < 1$, then $\frac{\sigma^2_{n_1}}{\sigma^2_0} < N_2$ and if $\beta_1 < \frac{4r\cos\theta}{1+r^2}$, then $\frac{\sigma^2_{n_2}}{\sigma^2_0} < N_2$. If these conditions are met, then the noise gain from each node of the coupled form is less than that of the noise gain N_2 of the direct form. The total noise output for each form will depend on the number of noise sources at each node in the filters. For the coupled form, there are

three noise sources at node n_1 due to α_0, β_1, and t and there are two noise sources at node n_2 due to α_1 and β_2. For the direct form there are three noise sources at node n_1. The total noise outputs of the direct and coupled forms are given by (4.11a) and (4.11b) respectively.

$$\left(\frac{\sigma_n^2}{\sigma_0^2}\right)_{coupled} = 3\alpha_1^2 N_2 + 2(1+\beta_1^2 - \frac{4r\cos\theta}{1+r^2}\beta_1)N_2 \qquad (4.11a)$$

$$\left(\frac{\sigma_n^2}{\sigma_0^2}\right)_{direct} = 3N_2 \qquad (4.11b)$$

Therefore the coupled form noise is less than the direct form noise if

$$\alpha_1^2 + \frac{2}{3}(1+\beta_1^2 - \frac{4r\cos\theta}{1+r^2}\beta_1) < 1. \qquad (4.12)$$

D. Closed Form Noise Calculations

The objective here is to select the parameters β_1, β_2, and t in Figure 4.1 so as to minimize the noise variance given by (4.11a). Since N_2 is a function only of the pole locations (4.9), it is sufficient to minimize the following expression

$$NUM = 3\alpha_1^2 + 2(1+\beta_1^2 - \frac{4r\cos\theta}{1+r^2}\beta_1). \qquad (4.13)$$

To do this, an expression for α_1 must be obtained. The overflow constraints which must be met at the summation nodes n_i and n_2 are given by

$$\max_{-\pi < \omega < \pi} \left| H_{w_1 x}(e^{j\omega}) \right| = \max_{-\pi < \omega < \pi} \left| \frac{\alpha_0 (1 - \beta_2 e^{-j\omega})}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \right| = 1 \qquad (4.14a)$$

$$\max_{-\pi < \omega < \pi} \left| H_{w_2 x}(e^{j\omega}) \right| = \max_{-\pi < \omega < \pi} \left| \frac{\alpha_0 \alpha_1 e^{-j\omega}}{1 - 2r\cos\theta e^{-j\omega} + r^2 e^{-2j\omega}} \right| = 1 \qquad (4.14b)$$

A closed form solution for α_0 and α_1 from (4.14) is not readily obtainable. It can be seen that α_0 and α_1 will be functions of β_2 and not of t. If a closed form solution for α_1 were available, it could be used in (4.13) with β_2 replaced by $2r\cos\theta - \beta_1$ and a closed form expression for the minimum noise could be found by taking the derivative of (4.13) with respect to β_1 and substituting the resulting β_1 into (4.11a). Since these expressions were not available, the following algorithm was used to find the minimum variance on a digital computer.

> Step 1: Select a value for β_1 ($\beta_2 = 2r\cos\theta - \beta_1$)
> Step 2: Solve (4.14a) iteratively for α_0
> Step 3: Solve (4.14b) iteratively for α_1
> Step 4: Solve (4.13) for NUM
> Step 5: Repeat steps 1-4 until NUM is minimized

An alternative to selecting β_1 to minimize the variance of the coupled form is to set β_1 to unity. This eliminates the β_1 multiplier and reduces the number of noise sources at node n_1 from 3 to 2. This will be referred to as coupled form A and the result of the previous design algorithm(variance minimized) will be referred to as coupled form B. The variance for coupled form A is given by

$$\frac{\sigma_n^2}{\sigma_0^2} = 2 \frac{\sigma_{n_1}^2}{\sigma_0^2} + 2 \frac{\sigma_{n_2}^2}{\sigma_0^2} \qquad (4.15)$$

By selecting $\beta_1 = 1$, the parameters β_2 and $\alpha_1 t$ are determined by (4.3). α_0 and α_1 are found as in steps 2 and 3 of the previous algorithm. Then the variance can be calculated from (4.15).

These calculations were made for various pole locations and the results are plotted in Figures 4.3, 4.4, and 4.5. Each figure is a plot of the noise variance in dB verses the angle θ of the pole location for a fixed pole radius r. Also included is the variance of the direct form realization of the filter. The plots go to $\theta = 90^\circ$ and they are symmetric from 90° to 180° ($\beta_1 = -1$ in form A for $\theta > 90^\circ$). From the graphs it can be seen that for large pole radius r and small pole angle θ, both coupled forms A and B offer significant improvement over the direct form. At $\theta = 0$, coupled form B is 36.4 dB better than the direct form for r=.99 and it is 27.2 and 16.5 dB better for r=.95 and r=.90 respectively. For small angles, coupled form B is seen to be slightly better than coupled form A. However for most angles less than 50°, form A which has fewer noise sources is seen to be slightly better. These results are tabulated in Table 4.1. The structures are listed in descending order according to their performance. θ_0 is shown in Figures 4.3-4.5.

Angle Range

Rank	$0 - \theta_0$	$\theta_0 - 50^\circ$	$50^\circ - 90^\circ$
1	Form B	Form A	Direct
2	Form A	Form B	Form B
3	Direct	Direct	Form A

Table 4.1. Noise Performance of Second Order Structures
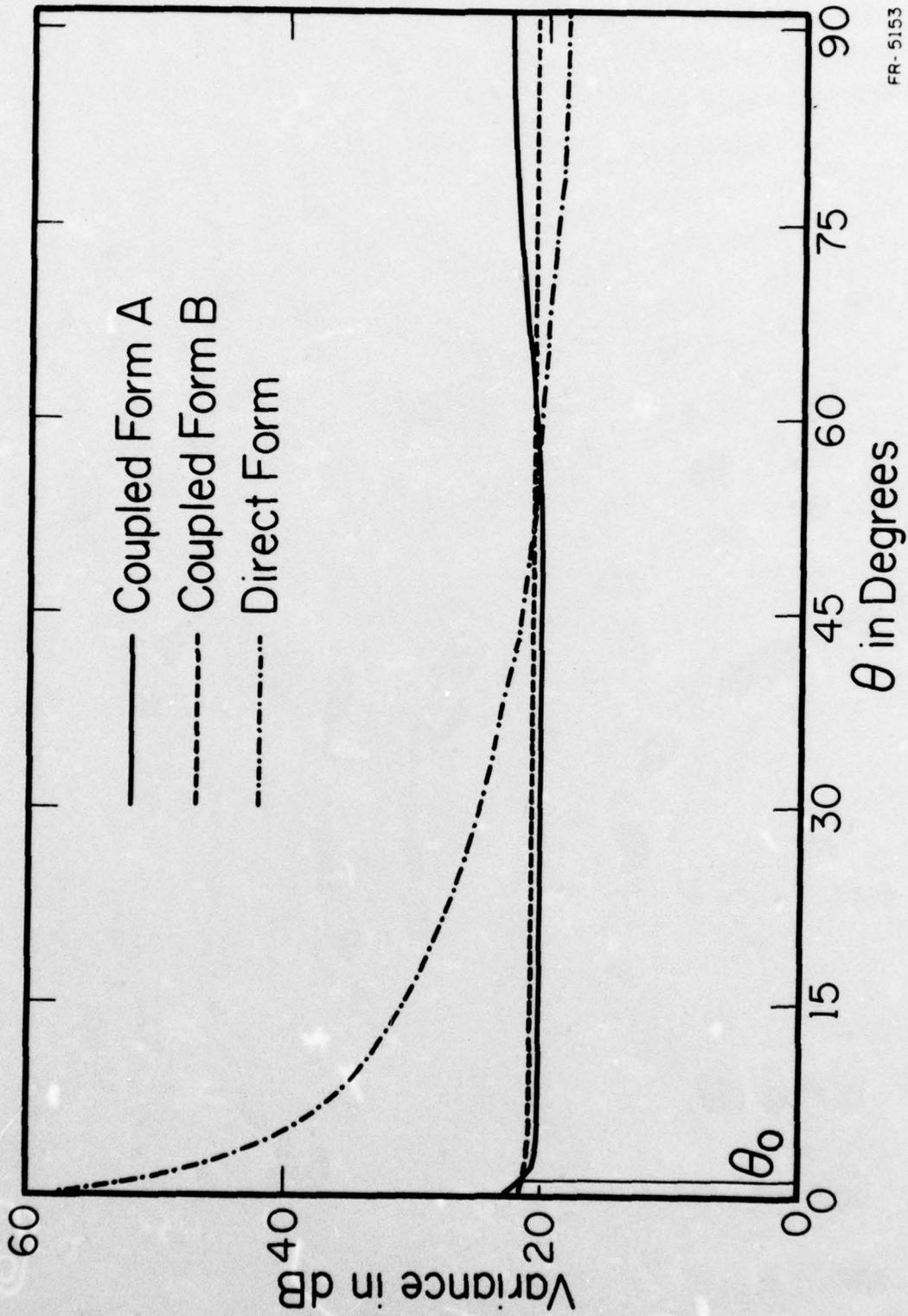
Figure 4.3 Noise Variance versus θ for Second Order Sections (radius = .99)
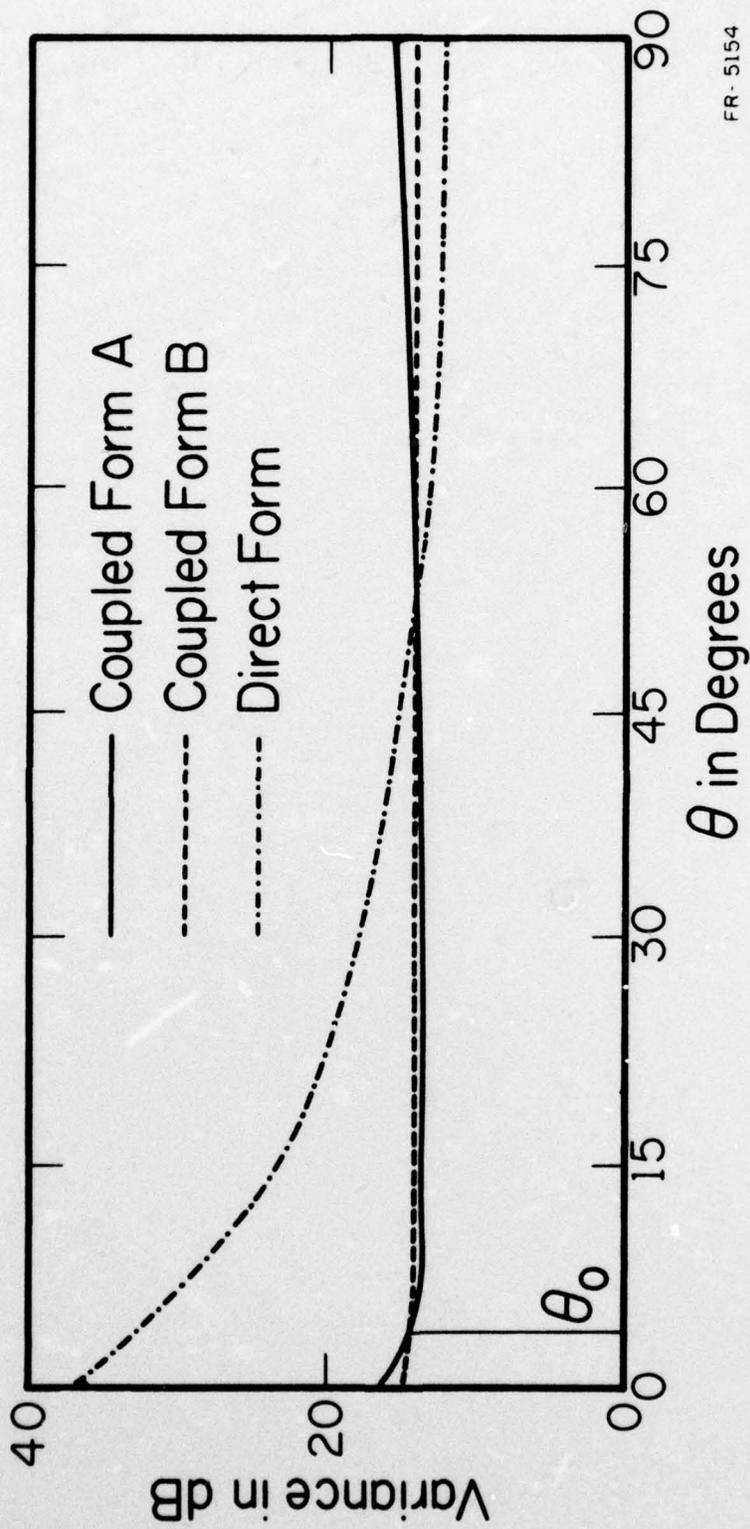
FR-5153

Figure 4.4 Noise Variance versus θ for Second Order Sections (radius = .95)

FR- 5154

34
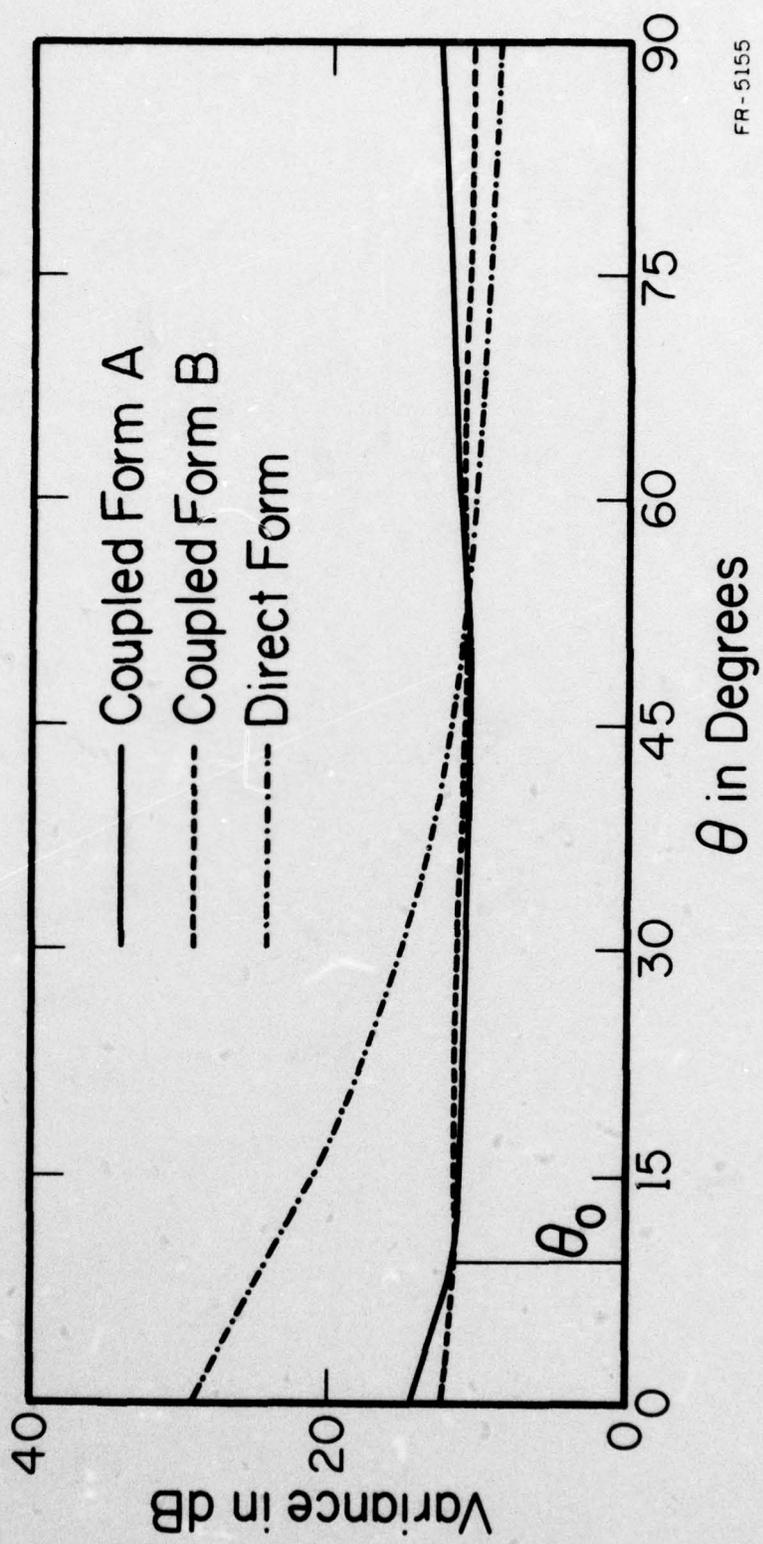
Figure 4.5 Noise Variance versus θ for Second Order Sections (radius = .90)

FR-5155

35

A plot of β_2 verses β_1 for r=.95 is shown in Figure 4.6. It can be seen that at θ =90°, β_1=0 and β_2=0 and coupled form B reduces to the direct form. The discrepancy in the noise variances of Figure 4.4 at this point is due to the fact that the noise calculation for the minimum variance in form B was made with 5 noise sources while that of the direct form was made with 3. This difference in noise sources also explains why coupled form B did not perform as well as the direct form for larger angles. It can also be seen in Figure 4.6 that $\beta_1 \approx \beta_2$ for coupled form B which is the choice indicated by the sensitivity analysis of Section B. Experimentally, it was found that if the noise gains at nodes n_1 and n_2 are weighted equally as in (4.15), then the minimum variance occured at exactly $\beta_1 = \beta_2 = r\cos\theta$.

The result that coupled form A performs better than coupled form B for angles between θ_0 and 50° and is not much worse for angles less than θ_0 makes this form appealing since it requires only two full precision multipliers as compared to three required for coupled form B. Thus the low noise performance of the coupled form can be obtained with the same number of full precision multipliers that are required in the direct form.
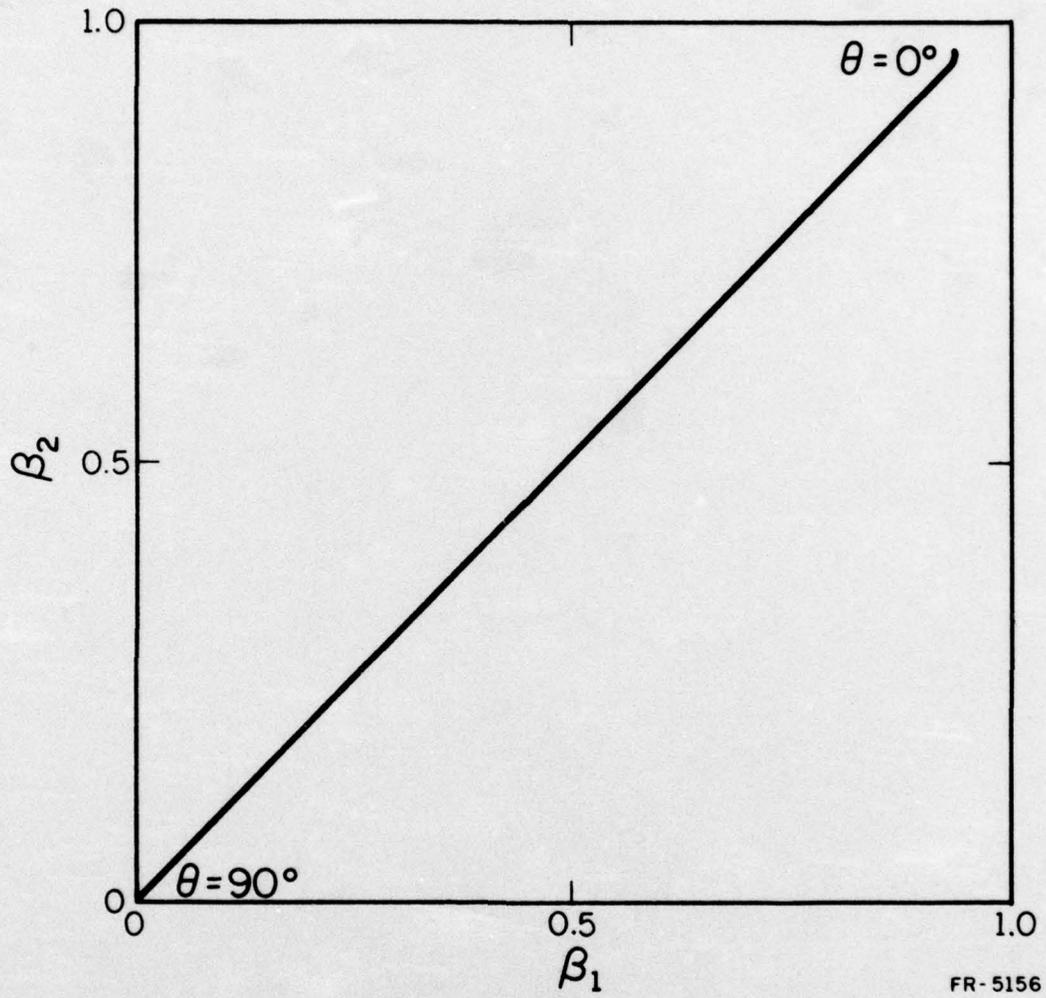
Figure 4.6 β_2 versus β_1 for Coupled Form B (r = .95)

5. NUMERICAL NOISE ANALYSIS

A. Analysis Technique

The analysis method used for computing roundoff noise is based on the signal flow graph representation of a digital filter. The actual noise analysis program used(NOISE) was written in FORTRAN by Long[1] based on the method of Jackson[2]. This method is briefly described below.

Using the signal flow graph representation, any digital filter structure with N nodes can be put in the matrix form

$$Y(z) = z^{-1}H_d Y(z) + H_c Y(z) + U(z) \qquad (5.1)$$

where

$Y(z)$ = Nx1 vector of node output values

$U(z)$ = Nx1 vector of node input values

H_c = NxN matrix of coefficients for branches with no delay

H_d = NxN matrix of coefficients for branches with one delay

The $(m,n)^{th}$ element of H_c(or H_d) is the coefficient of the multiplier in the branch with no delay(or with one delay) directed form node n to node m. (5.1) can be rewritten as

$$(I - z^{-1}H_d - H_c)Y(z) = U(z) \qquad (5.2)$$

The complex gain from node i to node k at any frequency $\omega_0 = \omega/T$ where T is the sampling rate can be found by solving the linear simultaneous

equations (5.2) with the following substitutions

$$U(z) = [0......010......0]^t \qquad \text{(5.3a)}$$
$$\qquad\qquad\qquad \underset{\uparrow}{ith} \text{ element}$$

and

$$z = e^{j\omega_0} \qquad \text{(5.3b)}$$

The gain from node i to node k is then given by the kth element of the solution vector $Y(e^{j\omega_0})$. A modified version of the IBM scientific subroutine SIMQ was used to solve the set of complex linear equations (5.2). It was found that for filters with very narrow bandwidths, double precision arithmetic was required in computing (5.2).

To compute the noise power output of a filter based on the model of Section 2.A, (2.3) and (2.5) are used. Substituting (2.3) into (2.5) and summing over the network nodes rather than the noise sources yields

$$\frac{\sigma_n^2}{\sigma_0^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{N}(\omega)d\omega \qquad \text{(5.4)}$$

where

$$\hat{N}(\omega) = \sum_{i=1}^{N} k_i \left| H_{ni}(e^{j\omega}) \right|^2 \qquad \text{(5.5)}$$

and N = number of nodes

k_i = number of error sources at node i

$H_{ni}(e^{j\omega})$ = transfer response from node i to the output

$\hat{N}(\omega)$ is the power spectral density of the noise. All $H_{ni}(e^{j\omega})$ are obtained at once by substituting

$$U(z) = [0.......01]^t \tag{5.6}$$

in (5.2). By the interreciprocity property, $H_{ni}(e^{j\omega})$ is given by the ith element of $Y(e^{j\omega})$. The integral of (5.4) is approximated by the trapezoidal rule as

$$\frac{\sigma_n^2}{\sigma_0^2} \approx \frac{1}{2N_p} \left[\hat{N}(0) + \hat{N}(\pi)\right] + \frac{1}{N_p} \sum_{i=1}^{N_p-1} \hat{N}(\omega_i) \tag{5.7}$$

where

$$\omega_i = i \cdot \frac{\pi}{N_p} \quad , \quad i = 0,1,\ldots,N_p.$$

A value of $N_p = 50$ was used for all examples listed in Section 5.B. This value for N_p gave results accurate to at least six decimal places when NOISE was run on second order structures for which the closed form variance could easily be computed by hand.

A measure of the noise power over a band of frequencies can be made by changing the limits of the integral of (5.4) to obtain

$$\frac{\sigma_{PB}^2}{\sigma_0^2} = \frac{1}{B} \int_{\omega_1}^{\omega_2} \hat{N}(\omega) d\omega \tag{5.8}$$

where

$$B = \omega_2 - \omega_1$$

ω_2 = upper frequency (normalized w.r.t. T)

ω_1 = lower frequency (normalized w.r.t. T).

The integral of (5.8) has been normalized with respect to the bandwidth B and is approximated by (5.9) for computer evaluation.

$$\frac{\sigma_{PB}^2}{\sigma_0^2} \approx \frac{1}{2N_p} \left[\hat{N}(\omega_1) + \hat{N}(\omega_2)\right] + \frac{1}{N_p} \sum_{i=1}^{N_p-1} \hat{N}(\omega_i) \tag{5.9}$$

where

$$\omega_i = i \frac{B}{N_p} \quad , \quad i = 0, 1, \ldots, N_p.$$

B. Examples For Comparison of Filter Structures

The filter structures which will be compared in this chapter are the standard cascade structure, the modified multiple feedback structure, and the coupled form second order sections in cascade form. The coupled forms of Chapter 4 were modified by placing a zero after each first order section. This and the fact that the total noise variance of a section in a cascade structure depends on the other sections means that the use of coupled form $B(\beta_1 = \beta_2)$ no longer guarantees minimized noise power. It is expected that the performance will be comparable however.

Two lowpass Chebyshev filters and two bandpass Chebyshev filters will be used for the comparisons. All have 0.5 dB passband ripple. Also, all have fairly narrow bandwidths since Long's results and those of Chapter 4 indicate that the MFB and coupled form structures will be of greatest interest for the narrow bandwidth case. The magnitudes of the transfer functions are pictured in Figures 5.1-5.4. The two lowpass filters have 0.5 dB cuttoff frequencies of $\frac{\pi}{8}$ and $\frac{\pi}{80}$ and will be designated LP8 and LP80 respectively. The two bandpass filters have 0.5 dB bandwidths of $\frac{\pi}{8}$. Their center frequencies are $.437\pi$ and $\frac{\pi}{8}$ and these filters will be designated BP1 and BP2 respectively.

The filters and their coefficients are given in Figures 5.5-5.11 and Tables 5.1-5.8. The coefficients for the MFB structures come from Long's MFB filter and transformations of it. The ordering of the cascaded second order sections in Figures 5.5, 5.8, 5.9, and 5.11 is based on Jackson[2]. All filters satisfy the scaling requirement of (2.8). This was accomplished on the computer using the techniques of Section 5.A to determine the maximum gain at each summation node. Also, the technique of Section 2.C was used to reduce the noise of each filter where possible. The BP2 filter was not constructed with coupled form B since its noise properties are similar to those of coupled form A.

Each multiplier in a filter with coefficient other than one is counted as a noise source including those scaling multipliers which may not be full precision multipliers. The multipliers whose coefficients are one may be replaced by a direct path between nodes in the filter.
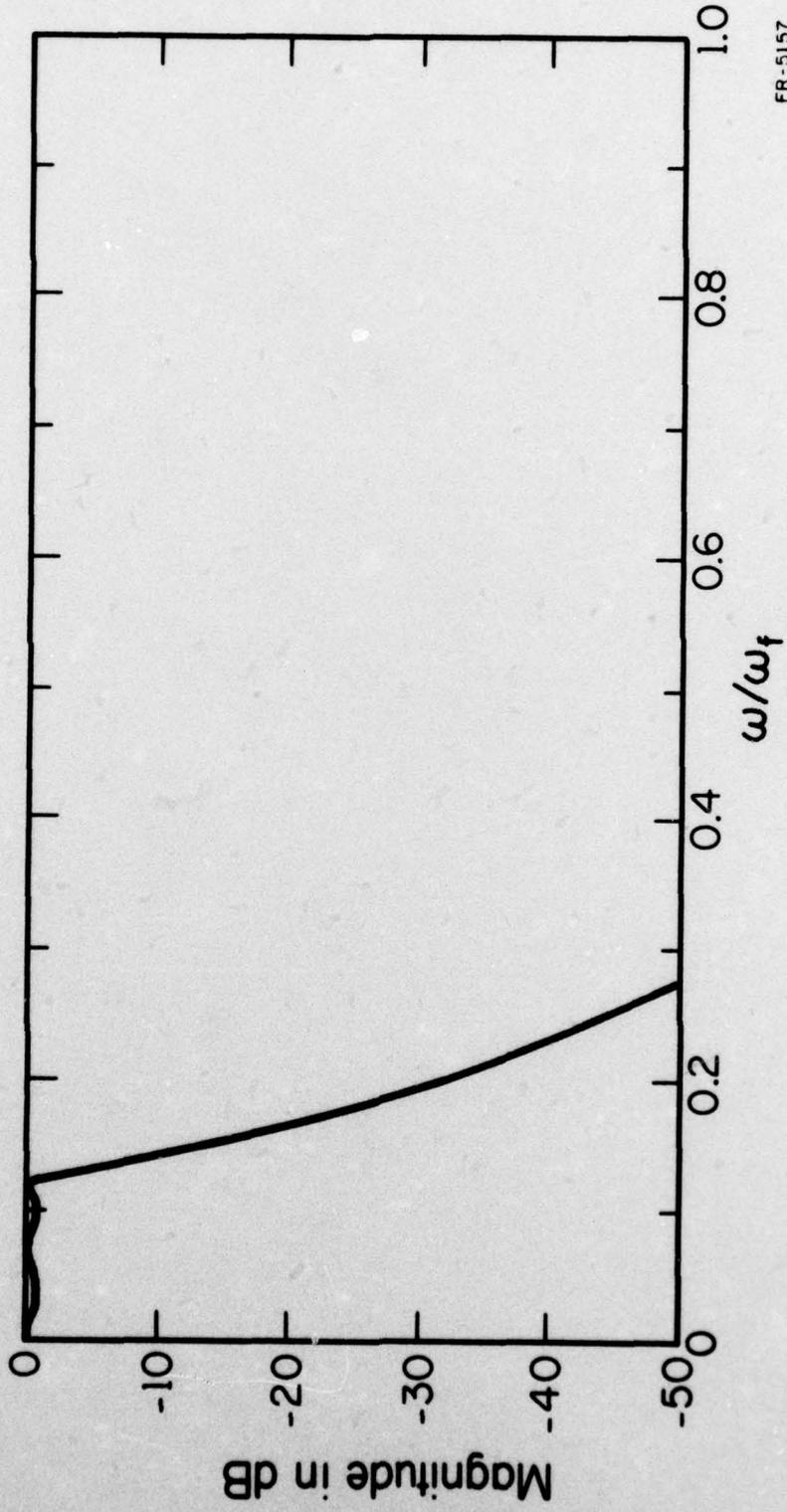
FR-5157

Figure 5.1 LP8 Magnitude Characteristic
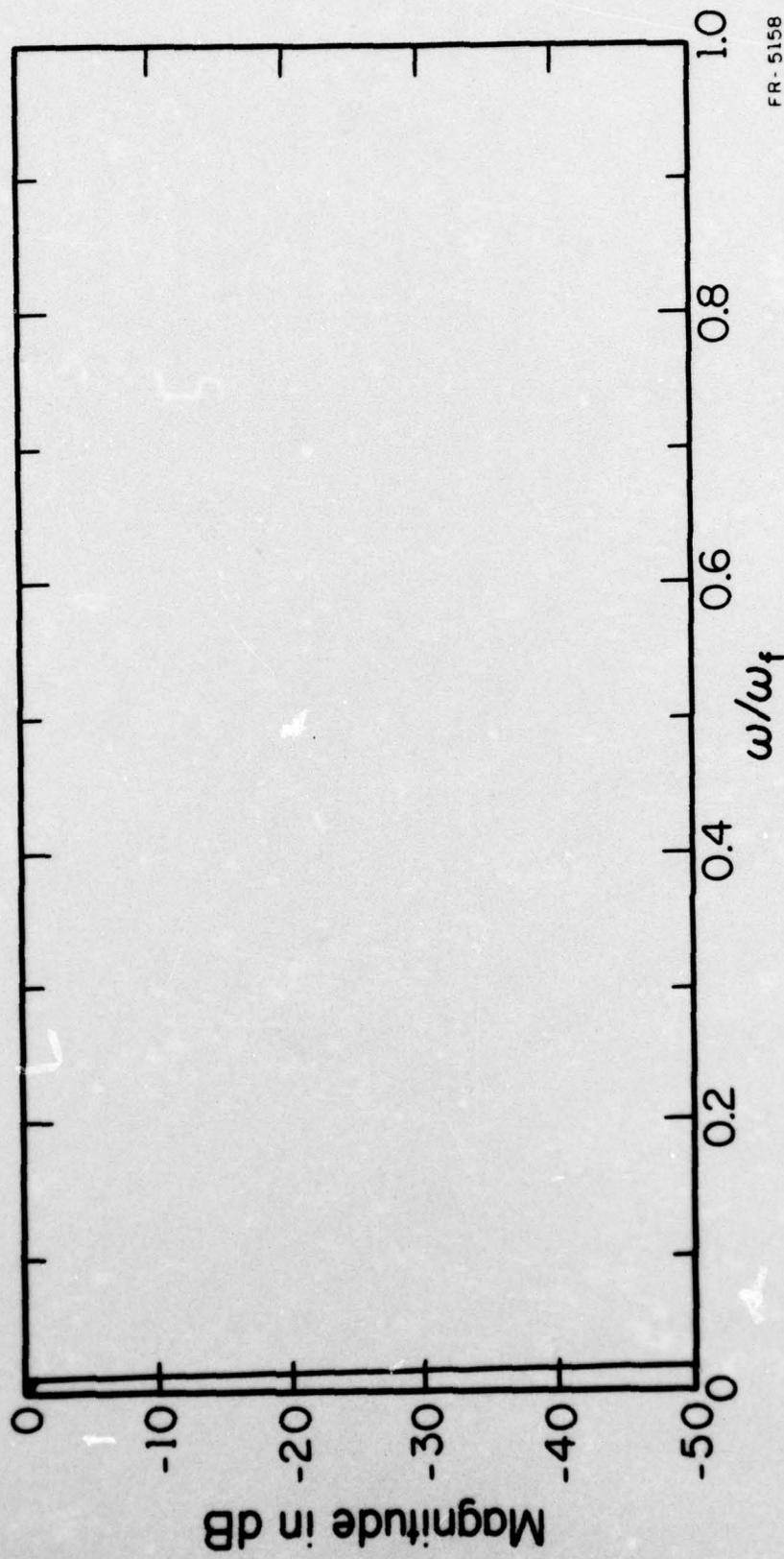
Figure 5.2 LP80 Magnitude Characteristic

FR - 5158

44

Figure 5.3 BPl Magnitude Characteristic

FR - 5159

45
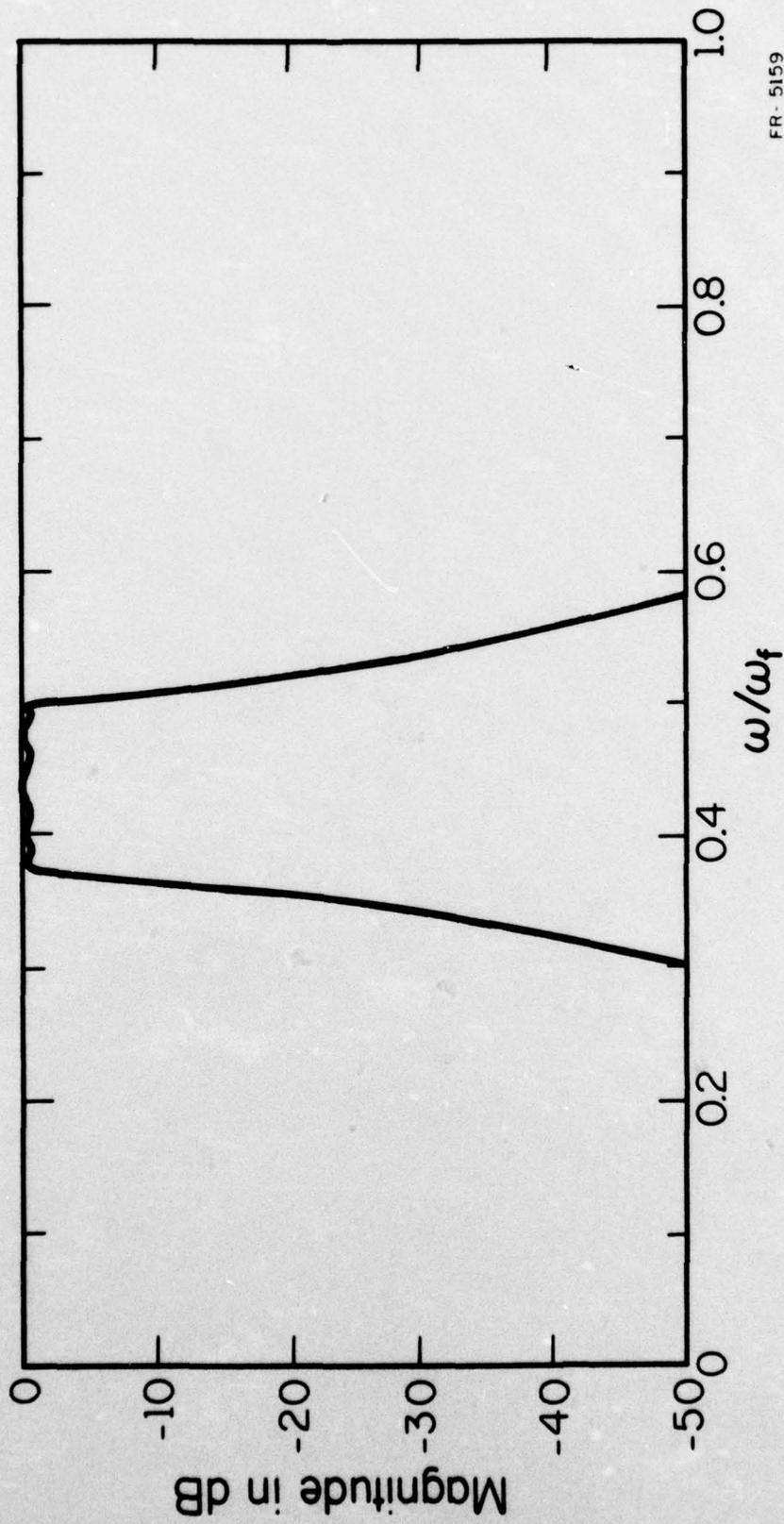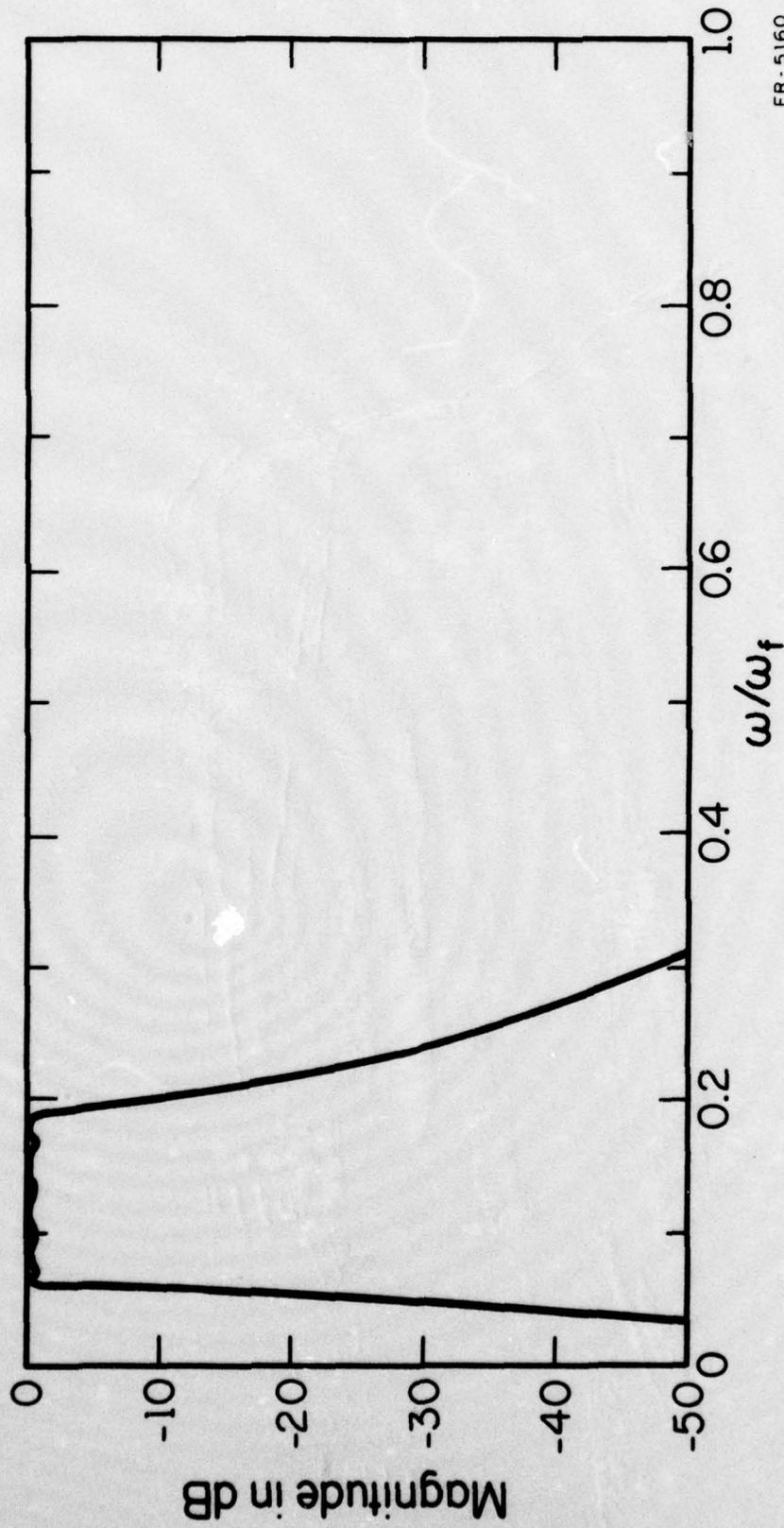
Figure 5.4 Magnitude Characteristic for BP2

FR - 5160

46
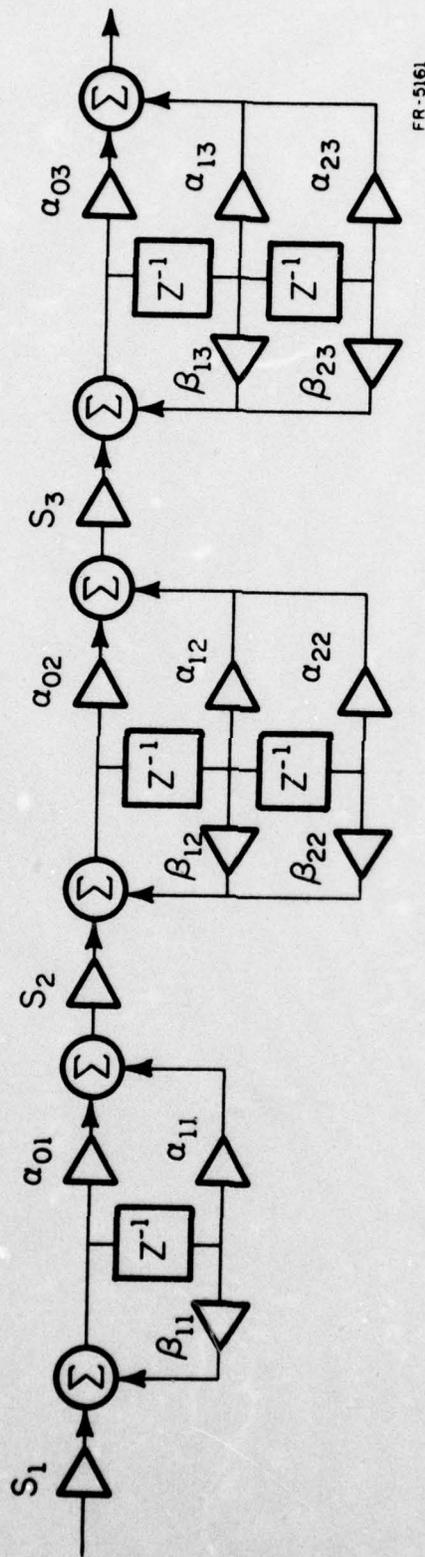
Figure 5.5 Cascade Realization of Fifth Order Filters LP8 and LP80

FR-5161

LP8 Filter

Section i	1	2	3
β_{1i}	.86555015	1.72815082	1.72669227
β_{2i}	-	-.79460336	-.91793529
α_{0i}	.5	.25	.25881416
α_{1i}	.5	.5	.51762833
α_{2i}	-	.25	.25881416
S_i	.13444980	.066452714	.14586934

LP80 Filter

Section i	1	2	3
β_{1i}	.98587045	1.97651463	1.98965721
β_{2i}	-	-.97724146	-.99124730
α_{0i}	.5	.25	.25
α_{1i}	.5	.25	.5
α_{2i}	-	.25	.25
S_i	.014129550	.000726930	.001590090

Table 5.1. Cascade Realization of Fifth-Order Chebyshev
Lowpass Filters (Figure 5.5)

49
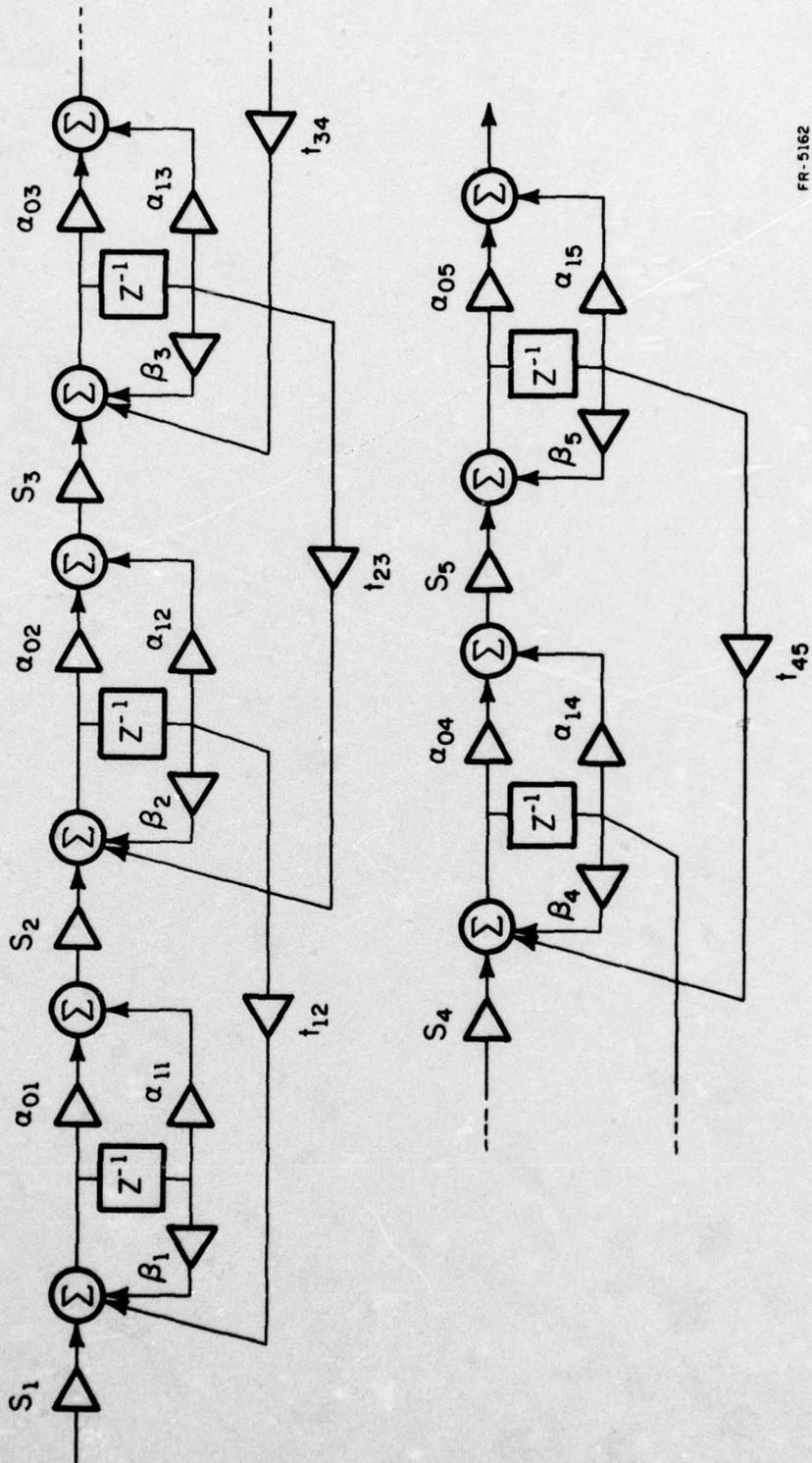
Figure 5.6 Multiple Feedback Realization of Fifth Order Filter LP8
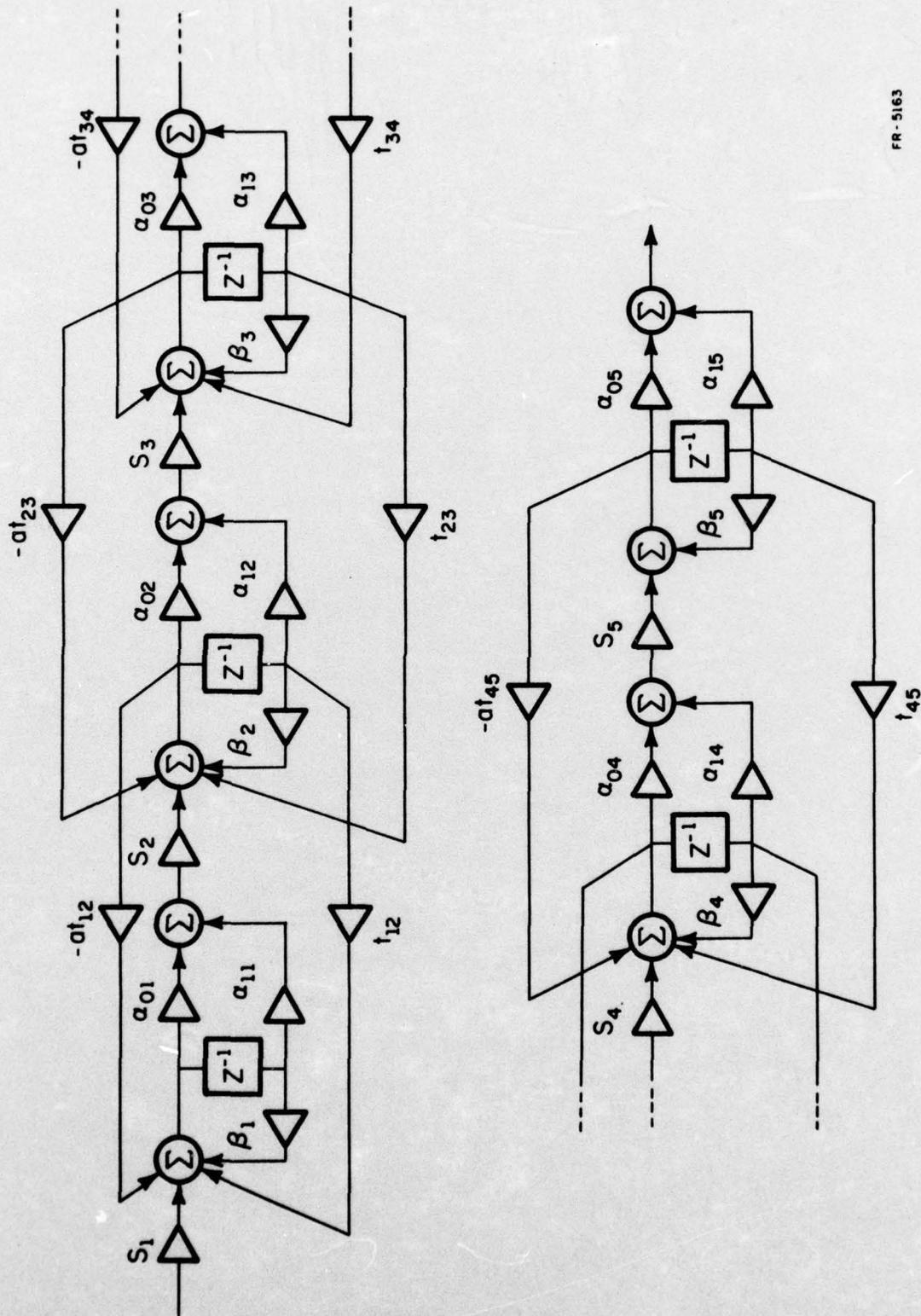
FR-5162

Figure 5.7 Multiple Feedback Realization of Fifth Order Filter LP80
(not physically realizable)

LP8 Filter (Figure 5.6)

Section i	1	2	3	4	5
β_i	.73984200	1	1	1	.73984366
α_{0i}	.50580116	.50142226	.5	.51741791	.52696216
α_{1i}	.50580116	.50142226	.5	.51741791	.52696216
S_i	.27256869	.21088026	.20847000	.25423218	.40029000
$t_{i,i+1}$	-.27027892	-.10586551	-.18621970	-.26794146	-

LP80 Filter (Figure 5.7)

Section i	1	2	3	4	5
β_i	.97090509	1	1	1	.97090530
α_{0i}	.50005464	.50003204	.5	.50008994	.5
α_{1i}	.50005464	.50003204	.5	.50008994	.5
S_i	.030617308	.021013212	.020581892	.025968065	.047181503
$t_{i,i+1}$	-.16848062	-.058317808	-.098877889	-.14450219	-
$-at_{i,i+1}$	.13820345	.047837684	.081108833	.11853412	-

Table 5.2. Multiple Feedback Realization of Fifth-Order Chebyshev Lowpass Filters
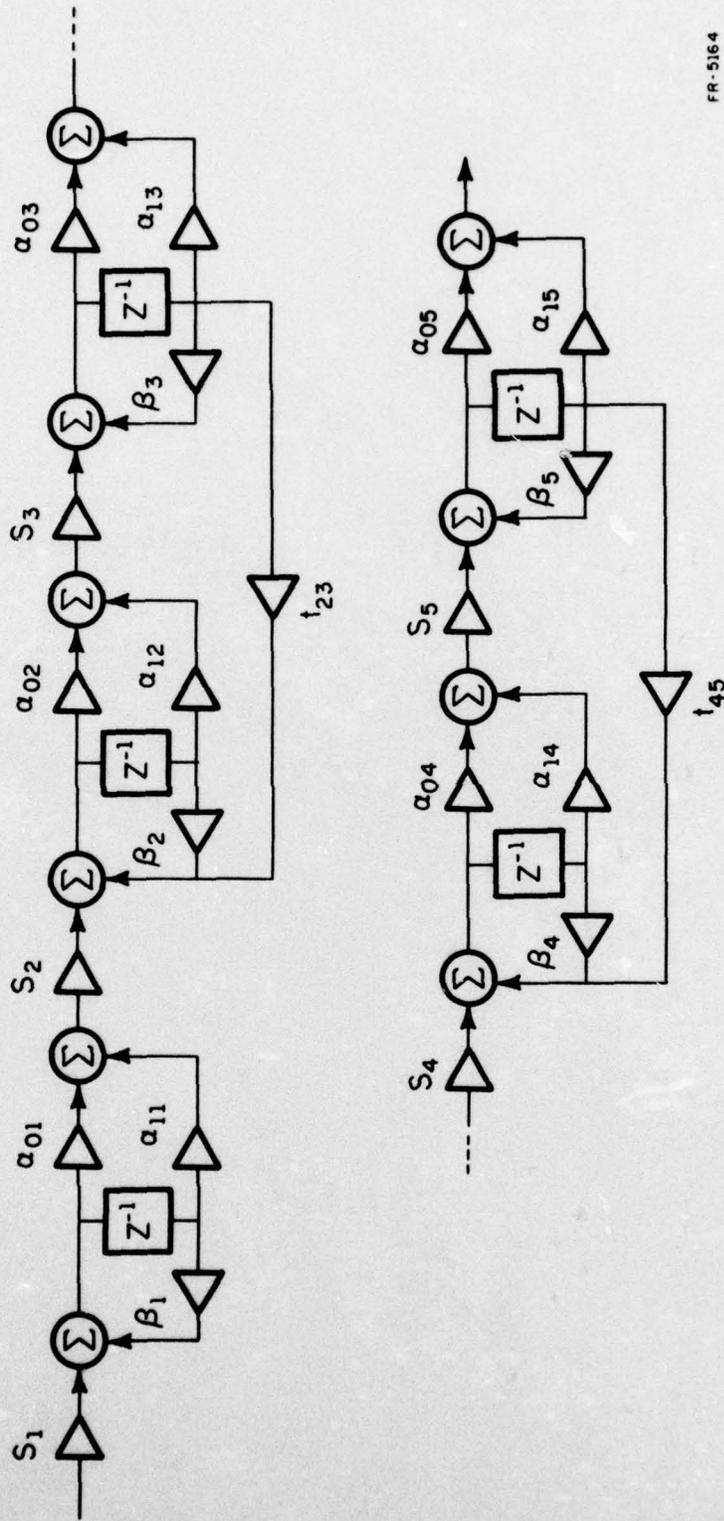
Figure 5.8 Coupled Form Realization of Fifth Order Filters LP8 and LP80

FR-5164

LP8 Filter

Section i	1	2	3	4	5
β_i	.86555015	1	.76137695	1	.84242903
α_{0i}	.5	.5	.5	.50925680	.50887813
α_{1i}	.5	.5	.5	.50925680	.50887813
s_i	.13444992	.27848330	.23862360	.38390200	.37947373
$t_{i,i+1}$	–	–.27848290	–	–.39071902	–

LP80 Filter

Section i	1	2	3	4	5
β_i	.98587045	1	.97687799	1	.99045223
α_{0i}	.5	.5	.5	.5009135	.50008438
α_{1i}	.5	.5	.5	.5009135	.50008438
s_i	.01412955	.031438434	.023122020	.04092788	.03883737
$t_{i,i+1}$	–	–.031438400	–	–.040934790	–

Table 5.3. Coupled Form A Realization of Fifth-Order Chebyshev Lowpass Filters (Figure 5.8)

LP8 <u>Filter</u>

Section i	1	2	3	4	5
$\hat{\beta}_i$	.86555008	.87690008	.87690008	.91959842	.91959842
α_{0i}	.5	.50310306	.5	.50929864	.50887813
α_{1i}	.5	.50310306	.5	.50929864	.50887813
S_i	.13444992	.38532380	.17139564	.39371173	.36998840
$t_{i,i+1}$	-	-.29745568	-	-.38354994	-

LP80 <u>Filter</u>

Section i	1	2	3	4	5
β_i	.98587045	.98840539	.98840539	.99522041	.99522041
α_{0i}	.5	.50003097	.5	.50009156	.50009166
α_{1i}	.5	.50003097	.5	.50009156	.50009166
S_i	.014129550	.041564330	.017488070	.041767851	.038055630
$t_{i,i+1}$	-	-.033877570	-	-.041175320	-

Table 5.4. Coupled Form B Realization of Fifth-Order Chebyshev Lowpass Filters (Figure 5.8)
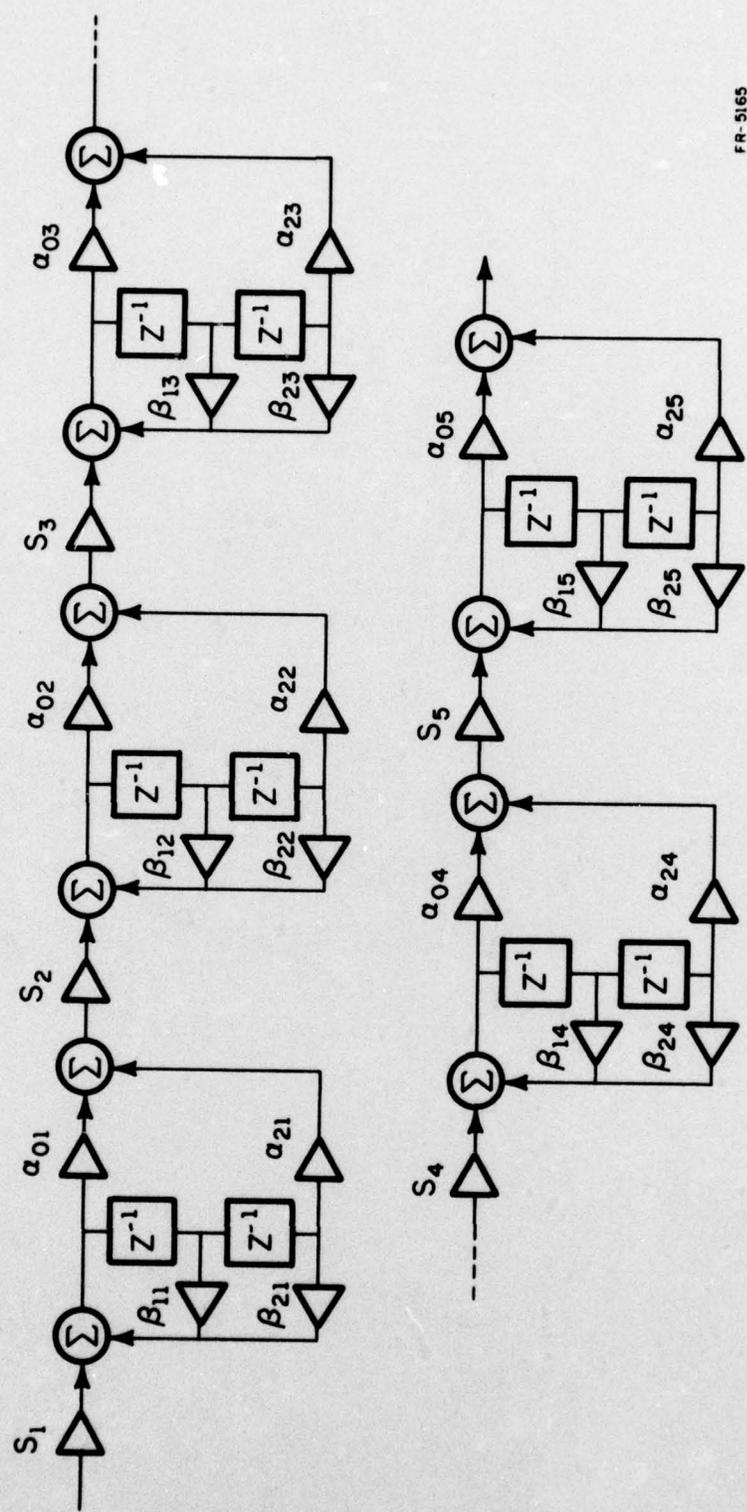
Figure 5.9 Cascade Realization of Tenth Order Filters BP1 and BP2

FR-5165

55

BP1 Filter

Section i	1	2	3	4	5
β_{1i}	.37311002	.14314536	.60248480	-.00246690	.75585144
β_{2i}	-.86555015	-.88881844	-.89399960	-.95643961	-.95974207
α_{0i}	.51036701	.50257077	.51057782	.50572018	.5
α_{2i}	-.51036701	-.50257077	-.51057782	-.50572018	-.5
s_i	.13171900	.20517000	.31547000	.20439000	.73774000

BP2 Filter

Section i	1	2	3	4	5
β_{1i}	1.64022622	1.60767978	1.75730982	1.87161502	1.94068386
β_{2i}	-.85804902	-.93840000	-.86555015	-.92605818	-.97819190
α_{0i}	1.07547972	.90399127	.90956961	.84101291	2.50940195
α_{2i}	-1.07547972	-.90399127	-.90956961	-.84101291	-2.50940195
s_i	.065995086	.053728000	.21435000	.28722000	.10346000

Table 5.5. Cascade Realization of Tenth-Order Chebyshev
Bandpass Filters (Figure 5.9)
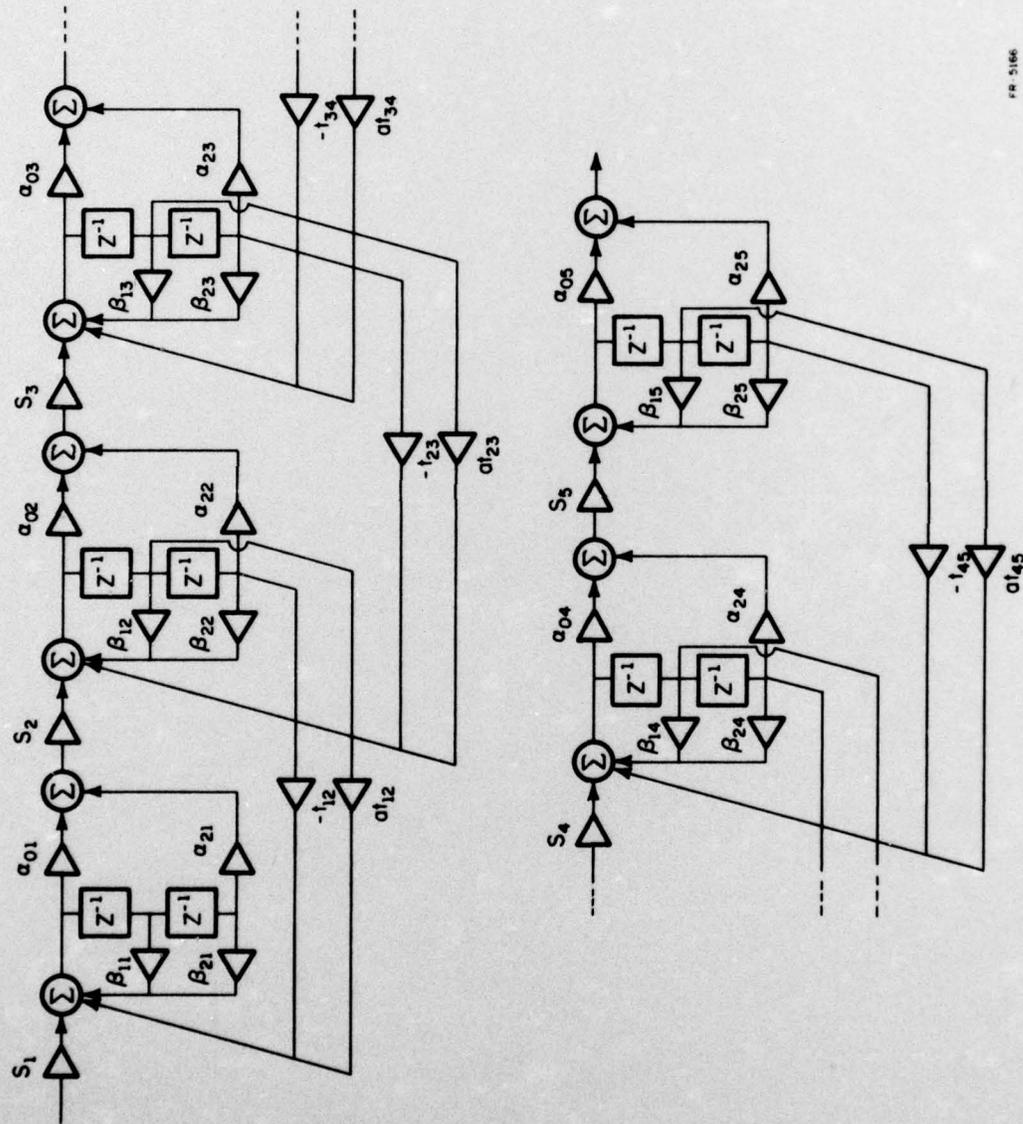
Figure 5.10 Multiple Feedback Realization of Tenth Order Filters BP1 and BP2

FR 5166

BP1 Filter

Section i	1	2	3	4	5
β_{1i}	.34796840	.4	.4	.4	.34796873
β_{2i}	-.73984200	-1	-1	-1	-.73984366
α_{0i}	.50580116	.50141044	.5	.52830427	.52679691
α_{2i}	-.50580116	-.50141044	-.5	-.52830427	-.52679691
s_i	.27256869	.21088026	.20422700	.25423219	.40019600
$-t_{i,i+1}$	.27027892	.10806750	.18621970	.26248184	-
$at_{i,i+1}$	-.054055783	-.021613500	-.037243940	-.052496368	-

BP2 Filter

Section i	1	2	3	4	5
β_{1i}	1.63889532	1.88395880	1.88395880	1.88395880	1.63889690
β_{2i}	-.73984200	-1	-1	-1	-.73984366
α_{0i}	2.49279940	2.08444795	1.54182728	2.52734089	2.50938850
α_{2i}	-2.49279940	-2.08444795	-1.54182728	-2.52734089	-2.50938850
s_i	.054918343	.051084467	.067606500	.052049375	.084058000
$-t_{i,i+1}$	.22638713	.078527574	.29496800	.26122411	-
$at_{i,i+1}$	-.21325201	-.073971351	-.27785377	-.24606773	-

Table 5.6. Multiple Feedback Realization of Tenth-Order
Chebyshev Bandpass Filters (Figure 5.10)
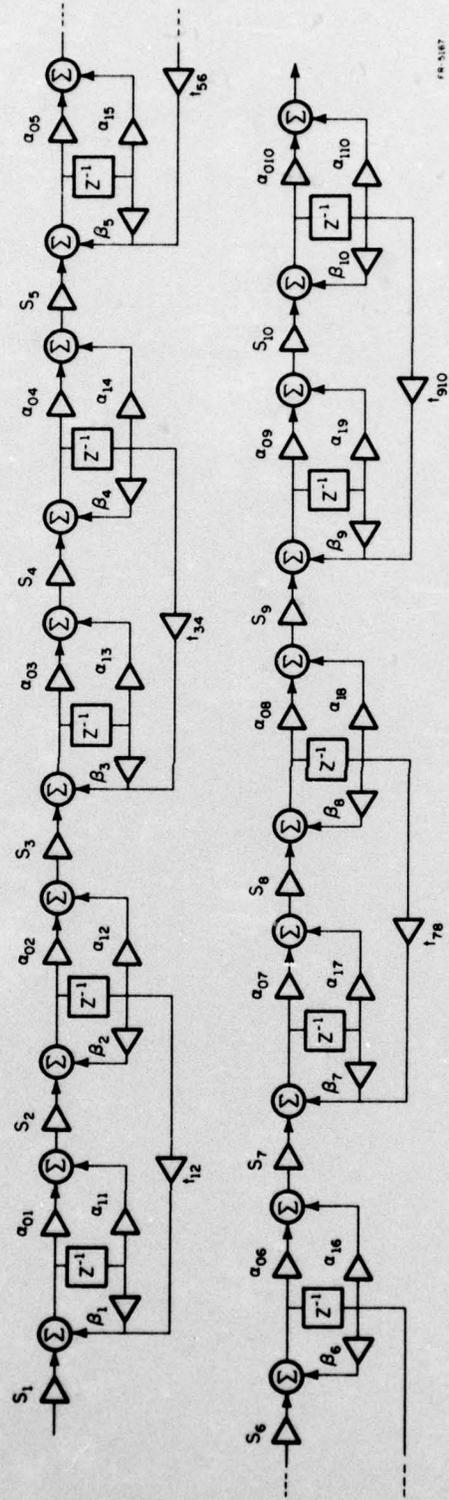
Figure 5.11 Coupled Form Realization of Tenth Order Filters BP1 and BP2

BP1 Filter

Section i	1	2	3	4	5
β_i	1	.11933007	1	.015981910	1
α_{0i}	.64514000	.79134000	.67286000	.74773000	.64509000
α_{1i}	.64514000	-.79134000	.67286000	-.74773000	.64509000
S_i	.13390000	.98267000	.20547000	.99777000	.32181000
$t_{i,i+1}$	-1.17707805	–	-1.30010299	–	-1.02131797

BP1 Filter(Continued)

Section i	6	7	8	9	10
β_i	.24824219	1	-.02301361	1	.35779676
α_{0i}	.79164000	.70407000	.71017000	.62546000	.79302870
α_{1i}	-.79164000	.70407000	-.71017000	.62546000	-.79302870
S_i	.98014000	.20438000	1.00041000	.71182000	1.0463000
$t_{i,i+1}$	–	-1.39056036	–	-.91981667	–

Table 5.7. Coupled Form A Realization of Tenth-Order Chebyshev Bandpass Filter (Figure 5.11)

BP2 Filter

Section i	1	2	3	4	5
β_i	1	.74913762	1	.77303993	1
α_{0i}	.51518560	2.0900057	.52228205	1.7309708	.52198295
α_{1i}	.51518560	-2.0900057	.52228205	-1.7309708	.52198295
S_i	.13646288	.48308654	.096567230	.55629730	.38955737
$t_{i,i+1}$	-.43760739	-	-.56913959	-	-.18840886

BP2 Filter(Continued)

Section i	6	7	8	9	10
β_i	.81142998	1	.89883662	1	.95943788
α_{0i}	1.74215060	.52175786	1.7562198	.52144475	1.76660582
α_{1i}	-1.74215060	.52175786	-1.7562198	.52144475	-1.76660582
S_i	.55030243	.479975231	.55000000	.50737774	.55474390
$t_{i,i+1}$	-	-.094859640	-	-.064832620	-

Table 5.8. Coupled Form A Realization of Tenth-Order
Chebyshev Bandpass Filter (Figure 5.11)

C. Results of the Noise Analysis

The total noise variance $\frac{\sigma_n^2}{\sigma_0^2}$ (equation (5.7)) was determined for each of the filter realizations described in Section 5.B. Also, the normalized power in the passband of each filter $\frac{\sigma_{PB}^2}{\sigma_0^2}$ was computed by (5.9) where ω_2 and ω_1 were chosen to be the upper and lower 0.5 dB cuttoff frequencies. The results are listed in Table 5.9 where the total normalized noise variance and the normalized passband noise power are expressed in decibels. The entries for the MFB LP80 filter are shaded to indicate that this filter is not physically realizable since it contains delay free loops.

Except for the nonrealizable case, the multiple feedback structure gave the best noise performance. For the fifth order LP8 transfer function, the total noise variance of the MFB structure was 10.4 dB better than the cascade form. For the tenth order BP1 and BP2 filters it was 4.4 and 11.2 dB better respectively. It was expected that for the tenth order bandpass filter BP2, the MFB structure would show even greater improvement over the cascade form than for the fifth order LP8 filter since the BP2 filter has 10 poles grouped fairly close together near z=1 while the LP8 filter has only 5. The reason that it actually showed only 0.8 dB improvement(from 10.4 to 11.2 dB) is believed to be due to the fact that the transformation from lowpass to bandpass introduced four extra feedback paths in the filter which added more noise sources to the structure. The fact that the MFB structure did not show much improvement over the cascade structure in the BP1 case was expected since in Chapter 4 it was seen that the direct form second order section has the least noise variance when its poles are at $\pm 90^\circ$.

Filter		Cascade	MFB	Coupled Form A	Coupled Form B
LP8	σ_n^2/σ_0^2	26.9	16.5	19.2	19.3
	σ_{PB}^2/σ_0^2	34.7	23.5	26.4	26.5
LP80	σ_n^2/σ_0^2	49.1	23.9	22.7	23.0
	σ_{PB}^2/σ_0^2	73.8	44.7	45.4	45.7
BP1	σ_n^2/σ_0^2	23.9	19.5	27.8	
	σ_{PB}^2/σ_0^2	40.0	26.3	34.8	
BP2	σ_n^2/σ_0^2	43.0	31.8	37.0	
	σ_{PB}^2/σ_0^2	50.3	38.6	44.6	

Table 5.9. Normalized Noise Variance $\dfrac{\sigma_n^2}{\sigma_0^2}$ and Normalized Passband Noise Power $\dfrac{\sigma_{PB}^2}{\sigma_0^2}$ in dB.

Filter	Cascade	MFB	Coupled Form A	Coupled Form B
LP8	5	6	5	7
LP80	5	10	5	7
BP1	10	15	10	
BP2	10	15	10	

Table 5.10. Number of Full Precision Multipliers Required For Each Filter.

It is interesting to note however that the normalized passband noise power of the cascaded second order sections for BP1 is 13.7 dB worse than that of the MFB filter. This indicates that while the total noise variance of the cascade form is not much worse than that of the MFB form for BP1, much more of its noise power is concentrated in the passband.

The coupled forms also performed well as expected from Chapter 4 for the lowpass filters. Coupled form A was 7.7 dB better than the cascade form for LP8 and 26.4 dB better for the very narrowband LP80 filter. Coupled form B was slightly worse than form A as expected making form A preferable to form B since it contains fewer multipliers. The MFB filter was 2.7 dB better than coupled form A for LP8 which is reasonable since it is observed that the network structure of the LP8 coupled form filter(Figure 5.8) is exactly that of the MFB filter(Figure 5.6) with every other feedback path eliminated. The extra feedback paths in the MFB filter allow more free parameters in its design. The performance of coupled form A for the BP2 filter fell halfway between that of the MFB structure and that of the cascade structure showing that for poles near z=1, this form is superior to the cascade filter. For the BP1 filter, coupled form A was 3.9 dB worse than the cascade form which was expected since the coupled form did not perform as well as the direct form for poles near the imaginary axis in Chapter 4. However the normalized passband noise power of the cascade filter was 5.2 dB worse than that of the coupled form filter, again indicating that the cascade form has much more noise power concentrated in the passband. Finally it is seen that the nonrealizable LP80 MFB filter had about the same noise variance as the coupled form although the result of the LP8 filter

design indicates that better noise performance can be obtained. This discrepancy is thought to lie in the fact that the frequency transformation used to obtain the filter introduced four extra multipliers which count as noise sources. This filter was included for comparison only and would have to be redesigned to eliminate the delay free loops for a more meaningful analysis. Such a design could result in lower output noise.

D. Hardware Considerations

Table 5.10 compares the number of full precision fixed point multipliers required for the filters of Table 5.9. Since the cost of these multipliers is a major portion of the cost of a filter it is desired to use as few as possible while at the same time reducing the noise power output of a filter. The noise output of a filter decreases by about 6 dB for each bit added to the register length of a filter[6]. Therefore for a given signal-to-noise ratio, filter structures with lower output noise can be realized with shorter register lengths. From Table 5.9, it can be seen that the lowpass LP8 MFB filter is better than the coupled form filters by less than 1 bit. However the MFB filter requires one more precision multiplier(3 multipliers are unity) than coupled form A for a fifth order filter. The choice between the two structures here is not clear cut however both offer between a one and a two bit improvement over the cascade structure for the LP8 case and a five bit improvement for the LP80 case. If the LP80 MFB filter is redesigned to eliminate the extra multipliers, it may be possible to get a six bit improvement here. For the BP1 filters, the MFB structure

offers less than a one bit improvement over the cascade form while requiring five more full precision multipliers. In this case it might be better to add one bit to the cascade structure rather than to use the MFB structure. For the BP2 filter the MFB structure is two bits better than the cascade form and the decision is not clear. It should be noted that the bandpass MFB filter design method used by Long would contain only one feedback multiplier for each stage instead of two as in Figure 3.3 however it is not known if this method will result in similar performance. Coupled form A is one bit better than the cascade form for the BP2 filter and might be used since it does not require any more full precision multipliers than the cascade form.

Finally, it was found that use of the scaling technique of Section 2.C resulted in a savings in noise power of from one to three dB for the filters constructed. Thus unless noise reduction is a much greater factor than cost, the technique would not be very useful here.

6. CONCLUSIONS

The transformed multiple feedback structures performed well with the proven LP8 MFB filter structure used as a prototype. The best performance relative to the cascade structure was achieved when the poles were grouped near z=1 as in the LP8, LP80, and BP2 filters. The MFB structure was also better than the cascade form for the BP1 filter whose poles are near z=±j although the improvement was not as great. The problem with using the frequency transformations of Section 3.C on the MFB filter is that the transformed filters are physically more complex than their prototypes. Also, if the transformation attempts to change the filter bandwidth, feedback paths are introduced which make the filter physically unrealizable. Thus the design of MFB filters cannot be accomplished in general by these frequency transformation techniques. The results of Chapter 5 indicate however that MFB filters can be designed to outperform corresponding cascade filters(direct form 2nd order sections) for any pole locations. More research should be concentrated on finding a design technique to minimize the noise of multiple feedback filters.

The coupled form second order section was shown in Chapter 4 to have less noise variance than the direct form for poles near z=±1. In Chapter 5, filters constructed with cascaded coupled form sections also performed well for poles in this region when compared to filters constructed with cascaded direct form sections. For narrowband lowpass filters(LP8 and LP80) the performance of the coupled form approached that of the MFB structure. It is interesting to note that the modified

lowpass MFB structure reduces to the cascaded coupled form structure when every other feedback path is removed. The simplicity of coupled form A combined with its noise properties make this an attractive filter structure to use when the filter poles are near $z=\pm1$.

For both the MFB and coupled form structures, the technique of selecting certain free parameters to be unity resulted in improved noise performance due to the elimination of noise sources. It was shown in Chapter 4 that the performance of a filter designed this way can actually be better than that of a filter whose noise variance has been minimized with respect to the free parameters. This method of reducing roundoff noise is also appealing because the number of costly multipliers required can be reduced. Where free parameters are available in a filter structure, the investigation of this technique for noise reduction may be profitable.

REFERENCES

1. J. L. Long, "Quantization Effects in Fixed-Point Recursive Digital Filters," Ph.D. Thesis, University of Illinois, Urbana, Illinois, Sept. 1973.

2. L. B. Jackson, "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade and Parallel Form," IEEE Trans. on Audio and Electroacoustics, Vol. AU-18, pp. 107-122, June 1970.

3. L. B. Jackson, "On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters," Bell System Tech. Jour., Vol. 49, pp. 159-184, February 1970.

4. A. Fettweis, "Roundoff Noise and Attenuation Sensitivity in Digital Filters with Fixed-Point Arithmetic", IEEE Trans. on Circuit Theory, Vol. CT-20, pp. 174-175, March 1973.

5. A. Fettweis, "On the Connection Between Multiplier Word Length Limitation and Roundoff Noise in Digital Filters," IEEE Trans. on Circuit Theory, Vol. CT-19, pp. 486-491, Sept. 1972.

6. A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Englewood Cliffs, New Jersey, Prentice-Hall, 1975.

7. B. Gold and C. M. Rader, Digital Processing of Signals, New York, McGraw-Hill, 1969.

8. E. I. Jury, Theory and Application of the z-Transform Method, New York, Wiley, 1964.