

## CANADIAN THESES ON MICROFICHE

## THÈSES CANADIENNES SUR MICROFICHE



National Library of Canada  
Collections Development Branch

Canadian Theses on  
Microfiche Service

Ottawa, Canada  
K1A 0N4

Bibliothèque nationale du Canada  
Direction du développement des collections

Service des thèses canadiennes  
sur microfiche

### NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

### AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

**THIS DISSERTATION  
HAS BEEN MICROFILMED  
EXACTLY AS RECEIVED**

**LA THÈSE A ÉTÉ  
MICROFILMÉE TELLE QUE  
NOUS L'AVONS REÇUE**

NEW IMPROVED STRUCTURES

FOR

RECURSIVE DIGITAL FILTERS

Paulo Sergio Ramirez Diniz

A Thesis

in

The Department

of

Electrical Engineering

Presented in Partial Fulfilment of the Requirements  
for the degree of Doctor of Philosophy at  
Concordia University  
Montréal, Québec, Canada

October 1984

© Paulo Sergio Ramirez Diniz, 1984

ABSTRACT

New Improved Structures for Recursive Digital Filters

Paulo Sergio Ramirez Diniz, Ph.D.  
Concordia University, 1984

A digital-filter synthesis procedure is developed by applying the concept of wave characterization to an analog configuration which realizes a continuous-time biquadratic transfer function by means of voltage-conversion type generalized-immittance converters (VGIC's). Then through the use of transposition a digital structure (TVGIC) is obtained which realizes simultaneously a lowpass, a bandpass and a highpass transfer function. This structure also realizes a transfer function with zeros on the unit circle using the minimum number of multipliers. In addition, a universal digital biquad is derived, which realizes simultaneously all the standard second-order transfer functions.

A special case of the TVGIC structure is shown to be amenable to the application of error spectrum shaping (ESS). ESS brings about a dramatic reduction in the output roundoff noise.

A systematic procedure is proposed which can be used for the generation of low-sensitivity digital filter structures which are amenable to ESS. The procedure is illustrated by the generation of several new structures.

The VGIC and TVGIC structures are shown to be free of zero-input and overflow limit cycles. Then a theorem is proved which establishes

sufficient conditions that will ensure freedom from constant-input limit cycles in a general digital-filter structure in which zero-input limit cycles can be eliminated. The application of this theorem to the TVGIC structure, to a sub-class of the VGIC structures, and to the universal digital biquad leads to efficient elimination of constant-input limit cycles.

Conditions for efficient elimination of constant-input limit cycles in second-order state-space structures are derived. These lead to new and more economical state-space structures. A design procedure is then described based on these structures which yields near-optimal noise performance.

A detailed noise analysis is undertaken whereby the proposed structures are compared with the direct-canonic and section-optimal structures. The results show that reduction can be achieved in the output noise by using some of the new structures.

Finally, several sensitivity aspects pertaining to the VGIC and TVGIC structures, the low-sensitivity structures and the new state-space structures are considered in detail. Techniques for the reduction of the sensitivity in the VGIC and TVGIC structures are presented. Then an optimal subset of the low-sensitivity structures is identified.

A sensitivity comparison of the proposed structures with the direct canonic and section-optimal structures shows that some of the proposed structures lead to significant improvements in the sensitivity.

TO MY PARENTS  
MY WIFE  
MY BROTHER  
AND PAULA

## ACKNOWLEDGEMENTS

I am greatly indebted to Dr. A. Antoniou for his guidance and for his careful correction of the manuscript.

I also wish to thank Profs. Luiz P. Calôba and Francira Sanches of UFRJ, and Profs. V. Ramachandran and R.V. Patel of Concordia University for their very valuable help.

My wife Mariza and my daughter Paula deserve special thanks for their understanding and love.

The encouragement of my mother, Hirlene, and my brother, Fernando have also made this work possible.

I am also grateful to Mrs. Madeleine Klein for typing this thesis and to Mr. Anthony Antoniou for drawing the figures. Both have done an excellent job.

My friends at Concordia have assisted me greatly, especially, P.C. Balla, G.V. Mendonça, L. Datta, J.C.M. Bermudez, P. Misra, Madeleine, Monica, Lina, Anita and Lynne.

The financial support of Capes (Brazilian Post-Graduate Education Federal Agency) and UFRJ (Federal University of Rio de Janeiro) is greatly acknowledged.

TABLE OF CONTENTS

	<u>Page</u>
List of Figures	xi
List of Tables	xv
List of Abbreviations and Symbols	xviii
<u>CHAPTER 1</u>	
1. Introduction	1
1.1 General	1
1.2 Quantization Effects in Digital Filters	2
1.2.1 Product Quantization	3
1.2.2 Overflow Limit Cycles	6
1.2.3 Coefficient Quantization	9
1.2.4 Input Quantization	12
1.3 Digital Filter Structures	12
1.3.1 Wave Structures	12
1.3.2 Cascade and Parallel Direct Canonic Structures	12
1.3.3 State-Space Section-Optimal Structures	13
1.3.4 Error-Spectrum Shaping Technique	18
1.4 Scope of the Thesis	21
<u>CHAPTER 2</u>	
2. VGIC-Based Structures	25
2.1 Introduction	25
2.2 Generation of the Active Analog-Filter Configuration	26
2.3 Digital VGIC Structure	28

	<u>Page</u>
2.4 Universal Digital Biquads with Simultaneous Outputs	37
2.5 TVGIC Structure with Error Spectrum Shaping	40
2.6 Comparison of Computational Complexity	42
2.7 Conclusions	42

### CHAPTER 3

3. Low-Sensitivity Structures Which are Amenable to Error-Spectrum Shaping	46
3.1 Introduction	46
3.2 Synthesis Procedure	47
3.3 Zero Placement	55
3.4 Application of ESS	58
3.5 Conclusions	58

### CHAPTER 4

4. Elimination of Limit Cycles	64
4.1 Introduction	64
4.2 Stability of VGIC Structures under Infinite- Precision Arithmetic	64
4.3 Stability of VGIC Structures under Finite- Precision Arithmetic	65
4.4 Elimination of Constant-Input Limit-Cycles	70



	<u>Page</u>
4.5 Elimination of Constant-Input Limit Cycles in VGIC and TVGIC Structures	75
4.6 Conclusions	78
 <u>CHAPTER 5</u>	
5. New Improved State-Space Structures	80
5.1 Introduction	80
5.2 Elimination of Zero-Input Limit Cycles	81
5.3 Elimination of Constant-Input Limit Cycles	82
5.4 Design Considerations	83
5.5 Design of High-Order Filters	89
5.6 Comparison of Computational Complexity	92
5.7 Conclusions	92
 <u>CHAPTER 6</u>	
6. Roundoff Noise Analysis	95
6.1 Introduction	95
6.2 VGIC Structures	95
6.3 Low-Sensitivity Structures of Chapter 3	111
6.4 New State-Space Structure	118
6.5 Section-Ordering	123
6.6 Conclusions	129

CHAPTER 7

7. Sensitivity Analysis	130
7.1 Introduction	130
7.2 Sensitivity Measures	130
7.3 VGIC Structures	132
7.4 Low-Sensitivity Structures of Chapter 3	144
7.5 State-Space Structures	155
7.6 Conclusions	156

CHAPTER 8

8. Conclusions	163
8.1 Introduction	163
8.2 Results of the Thesis	163
8.3 General Comparisons	166
8.4 Suggestions for Future Research	167

REFERENCES	171
------------	-----

APPENDIX A: Digital-Filter Designs	178
------------------------------------	-----

# LIST OF FIGURES

- Fig. 1.1 Noise Analysis
  - (a) Noise-Model for a Multiplier
  - (b) Noise Transfer Function
- Fig. 1.2 Signal Scaling
- Fig. 1.3 Direct Canonic Structure
- Fig. 1.4 (a) Cascade Realization
  - (b) Parallel Realization
- Fig. 1.5 Second-Order State-Space Structure
- Fig. 1.6 Application of ESS
  - (a) First-Order ESS
  - (b) Second-Order ESS
- Fig. 2.1 VGIC
- Fig. 2.2 VGIC-Based Active Analog-Filter Configuration
- Fig. 2.3 (a) Symbol for the Digital VGIC
  - (b) Digital VGIC with  $r(s) = s$
- Fig. 2.4 General VGIC Biquadratic Structure
- Fig. 2.5 General VGIC Biquadratic Structure
- Fig. 2.6. Standard Second-Order Digital VGIC Sections
  - (a) Lowpass
  - (b) Bandpass
  - (c) Notch
  - (d) Highpass
  - (e) Allpass
- Fig. 2.7 Allpass Section Using the VGIC Bandpass Structures of Fig. 2.5
  - (a) Block Diagram
  - (b) Structure
- Fig. 2.8 TVGIC General Structure
- Fig. 2.9 Universal Multiple-Output Biquad
- Fig. 2.10 TVGIC Structure with ESS

- Fig. 3.1 General Structure Suitable for the Application of ESS
- Fig. 3.2 General Second-Order Structure of Low Complexity
- Fig. 3.3 Structure for Case I
- Fig. 3.4 Structure for Case II
- Fig. 3.5 Zero Placement
- Fig. 3.6 Alternative Zero Placement
- Fig. 3.7 Application of ESS
- Fig. 4.1  $m_2$  versus  $m_1$  Plane
- Fig. 4.2 (a) VGIC Recursive Structure  
(b) Corresponding Block Diagram
- Fig. 4.3 Elimination of Overflow Limit Cycles
- Fig. 4.4  $m$ -th Order Digital-Filter Structure
- Fig. 4.5 Modified  $m$ -th Order Digital-Filter Structure
- Fig. 4.6 Elimination of Constant-Input Limit Cycles in the VGIC Structures of Figs. 2.6(b), 2.7(b) and 2.9
- Fig. 4.7 Elimination of Constant-Input Limit Cycles in the TVGIC Structure
- Fig. 5.1 Structure for Case I  
( $b_1=1-a_{11}$ ,  $b_2=-a_{21}$ )
- Fig. 5.2 Structure for Case II  
( $b_1=-a_{12}$ ,  $b_2=1-a_{22}$ )
- Fig. 5.3 Structure for Case III  
( $b_1=1-a_{11}-a_{12}$ ,  $b_2=1-a_{22}-a_{21}$ )

- Fig. 6.1 Transfer Functions of Interest
- Fig. 6.2 RPSD versus Frequency (Elliptic Lowpass)
- Fig. 6.3 RPSD versus Frequency (Elliptic Lowpass)
- Fig. 6.4 RPSD versus Frequency (Elliptic Bandpass)
- Fig. 6.5 RPSD versus Frequency (Elliptic Bandpass)
- Fig. 6.6 RPSD versus Frequency (Butterworth Bandstop)
- Fig. 6.7 RPSD versus Frequency (Butterworth Bandstop)
- Fig. 6.8 RPSD versus Frequency (Chebyshev Highpass)
- Fig. 6.9 RPSD versus Frequency (Chebyshev Highpass)
- Fig. 6.10 RPSD versus Frequency (Elliptic Lowpass)
- Fig. 6.11 RPSD versus Frequency (Elliptic Bandpass)
- Fig. 6.12 RPSD versus Frequency (Butterworth Bandstop)
- Fig. 6.13 RPSD versus Frequency (Chebyshev Highpass)
- Fig. 6.14 RPSD versus Frequency (Elliptic Lowpass)
- Fig. 6.15 RPSD versus Frequency (Elliptic Bandpass)
- Fig. 6.16 RPSD versus Frequency (Butterworth Bandstop)
- Fig. 6.17 RPSD versus Frequency (Chebyshev Highpass)
- Fig. 7.1 Improved TVGIC Structures
- (a) For Pole Angle  $0 < \omega_0 < \pi/3$
  - (b) For Pole Angle  $2\pi/3 < \omega_0 < \pi$
  - (c) For Pole Angle  $\pi/3 < \omega_0 < 2\pi/3$
- Fig. 7.2 Maximum Sensitivity  $\hat{S}$  versus  $\omega_0$  with  $\alpha_2=0.985$
- (a) VGIC
  - (b) Improved VGIC
  - (c) Direct Canonic
  - (d) Section-Optimal
- Fig. 7.3 Amplitude Response of the Lowpass Filter (Coefficient Wordlength = 11 bits)

- Fig. 7.4      Amplitude Response of the Bandpass  
Filters (Coefficient Wordlength = 9 bits)
- Fig. 7.5      Amplitude Response of the Bandstop  
Filters (Coefficient Wordlength = 7 bits)
- Fig. 7.6      Amplitude Response of the Highpass  
Filters (Coefficient Wordlength = 8 bits)
- Fig. 7.7      Maximum Sensitivity  $\hat{S}$  versus  $\alpha_1$  with  $\alpha_2 = 0.985$
- Fig. 7.8      Amplitude Response of the Lowpass  
Filters (Coefficient Wordlength = 8 bits)
- Fig. 7.9      Amplitude Response of the Bandpass  
Filters (Coefficient Wordlength = 8 bits)
- Fig. 7.10     Amplitude Response of the Bandstop  
Filters (Coefficient Wordlength = 5 bits)
- Fig. 7.11     Amplitude Response of the Highpass  
Filters (Coefficient Wordlength = 5 bits)
- Fig. 7.12      $\hat{S}$  versus  $\omega_0$  for the State-Space Structure  
( $\alpha_2=0.985$ )
- Fig. 7.13     Amplitude Response of the Lowpass  
Filters (Coefficient Wordlength = 8 bits)
- Fig. 7.14     Amplitude Response of the Bandpass  
Filters (Coefficient Wordlength = 9 bits)
- Fig. 7.15     Amplitude Response of the Bandstop  
Filters (Coefficient Wordlength = 7 bits)
- Fig. 7.16     Amplitude Response of the Highpass  
Filters (Coefficient Wordlength = 7 bits)

LIST OF TABLES

Table 2.1	Arithmetic Operations
Table 3.1	Specific Structures Based on the General Structures of Fig. 3.3
Table 3.2	Specific Structures Based on the General Structure of Fig. 3.4
Table 3.3	Equations for the Design of Biquadratic Transfer Functions
Table 3.4	Alternative Equations for the Design of Biquadratic Transfer Functions
Table 5.1	Arithmetic Operations
Table 6.1	Transfer Functions for VGIC and TVGIC Structures
Table 6.2	Filter Specifications
Table 6.3	Average Noise in db
Table 6.4	Transfer Functions in the Low-Sensitivity Structures
Table 6.5	Filter Specifications
Table 6.6	Average Noise in db
Table 6.7	Average Noise in db
Table 7.1	Numerator Polynomials of Sensitivity and Design Equations
Table 7.2	Numerator and Denominator Polynomials of Sensitivities
Table 7.3	Optimum Structures
Table 8.1	General Comparisons
Table A.1	Transfer Function Coefficients of Filters Described in Table 6.2

Table A.2	Multiplier Coefficients for Lowpass Design (a) VGIC Structure of Fig. 2.6 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10 (d) TVGIC Structure with First-Order ESS of Fig. 2.10 (e) Direct Canonic Structure (f) Section-Optimal Structure
Table A.3	Multiplier Coefficients for Bandpass Design (a) VGIC Structure of Fig. 2.6 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10 (d) TVGIC Structure with First-Order ESS of Fig. 2.10 (e) Direct Canonic Structure (f) Section-Optimal Structure
Table A.4	Multiplier Coefficients for Bandstop Design (a) VGIC Structure of Fig. 2.6 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10 (d) TVGIC Structure with First-Order ESS of Fig. 2.10 (e) Direct Canonic Structure (f) Section-Optimal Structure
Table A.5	Multiplier Coefficients for Highpass Design (a) VGIC Structure of Fig. 2.6 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10 (d) TVGIC Structure with First-Order ESS of Fig. 2.10 (e) Direct Canonic Structure (f) Section-Optimal Structure
Table A.6	Transfer Function Coefficients of Filters Described in Table 6
Table A.7	Multiplier Coefficients for Lowpass Design (Structure II-1)
Table A.8	Multiplier Coefficients for Bandpass Design Described in Table 6.5 (a) Structure I-5 (b) Section-Optimal Structure
Table A.9	Multiplier Coefficients for Bandstop Design Described in Table 6.5 (a) Structure I-7 (b) Section-Optimal Structure



- Table A.10 Multiplier Coefficients for Highpass  
Design (Structure II-6)
- Table A.11 Multiplier Coefficients for Lowpass  
Design (New State-Space Structure)
- Table A.12 Multiplier Coefficients for Bandpass  
Design (New State-Space Structure)
- Table A.13 Multiplier Coefficients for Bandstop  
Design (New State-Space Structure)
- Table A.14 Multiplier Coefficients for Highpass  
Design (New State-Space Structure)

LIST OF ABBREVIATIONS

AP	Allpass
BP	Bandpass
BS	Bandstop
ESS	Error Spectrum Shaping
HP	Highpass
LP	Lowpass
N	Notch
PSD	Power Spectrum Density
RPSD	Relative Power Spectrum Density
$r(s)$	Conversion Function of the VGIC
$s$	Complex Frequency Variable in the Analog Domain
$T$	Sampling Period
TVGIC	Transposed Voltage Conversion Generalized-Immittance Converter
VGIC	Voltage Conversion Generalized-Immittance Converter
$\omega_s$	Sampling Frequency, Rad/s
$z$	Complex Frequency Variable in the Digital Domain

## CHAPTER 1 INTRODUCTION

### 1.1 General

The rapid development of digital integrated circuit technology in the last two decades has made digital signal processing not only a tool for the simulation of analog systems but also a technique for the implementation of very complex systems. It has found applications in many areas such as picture processing, speech analysis and synthesis, biomedical engineering, sonar, radar, seismology and many others.

The main advantages of digital systems relative to analog systems are high reliability, ease of modifying the filter's characteristics and low cost. These advantages led to the digital implementation of many signal processing systems, which were usually implemented with analog circuits in the past. Also some new applications became viable after the development of digital integrated-circuit technology.

The digital filter is in general the most important component in most digital signal processing systems. In this thesis, we are primarily concerned with linear, shift-invariant digital filters which are realized using finite-precision arithmetic.

In practice, a digital filter is implemented using software on a general-purpose digital computer or by using special-purpose hardware. In both implementations quantization errors are inherent due to finite-precision arithmetic. These errors can be categorized as follows:

1. Roundoff errors committed when the internal data like outputs of multipliers are quantized before or after summations.
2. Errors in the amplitude response due to the use of finite wordlength for the representation of multiplier constants.
3. Errors due to the quantization of the input signal into a set of discrete levels.

The errors described above are dependent upon the type of arithmetic used in the implementation. If the digital filter is implemented on a general-purpose computer, floating-point arithmetic is in general available and so this type of arithmetic is the natural choice. However, if the filter is implemented by means of special-purpose hardware, fixed-point arithmetic is in general the best choice since it is simple to perform and less costly in terms of hardware. In this thesis only fixed-point arithmetic will be considered.

## 1.2 Quantization Effects in Digital Filters

The choice of a digital filter structure for a given application is based on evaluating the performance of known structures and choosing the most suitable one. The effects of quantization are important factors to be considered when assessing the performance of digital filter structures.

### 1.2.1 Product Quantization

#### I - Roundoff Noise

A finite-wordlength multiplier can be modelled in terms of an ideal multiplier followed by a single noise source  $e(n)$  as shown in Fig. 1.1(a).

If product quantization is carried out by rounding and the signal levels throughout the filter are much larger than the quantization step  $q$ , it can be shown [1] - [2] that the power spectral density (PSD) of  $e_i(n)$  is

$$S_{e_i}(z) = \frac{q^2}{12} \quad (1.1)$$

that is,  $e_i(n)$  is a white-noise process. Also  $e_i(n)$  and  $e_j(n+k)$  are in practice statistically independent for any value of  $n$  or  $k$  ( $i \neq j$ ). As a consequence the PSD of  $e_i(n) + e_j(n)$  is equal to the sum of their respective PSDs, i.e.

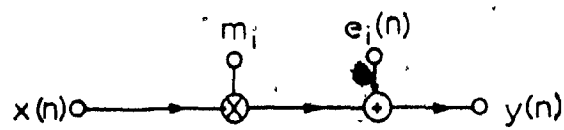
$$S_{e_i+e_j}(z) = S_{e_i}(z) + S_{e_j}(z) \quad (1.2)$$

Eqn. 1.2 indicates that superposition can be employed in the evaluation of the output PSD in a digital filter in which several noise sources  $e_i(n)$  are embedded.

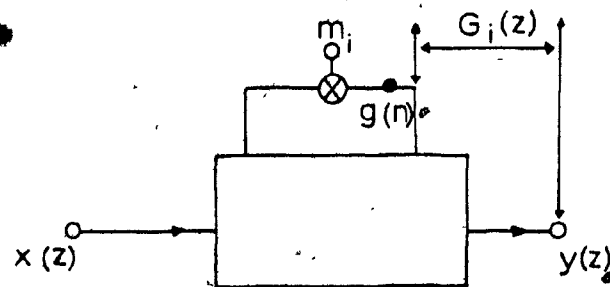
In a fixed-point digital-filter implementation, the PSD of the output noise is given by [1] - [2] ,

$$S_y(z) = \frac{2^{-2b}}{12} \sum_{i=1}^k G_i(z) G_i(z^{-1}) \quad (1.3)$$

where  $G_i(z)$  are the transfer functions from each multiplier output  $g(n)$  to the filter output as shown in Fig. 1.1(b). The wordlength, including sign, is  $b+1$  bits and  $k$  is the number of multipliers of the filter.



(a)



(b)

Fig. 1.1: Noise Analysis  
 (a) Noise Model for a Multiplier  
 (b) Noise Transfer Function

The relative power spectral density (RPSD) of the output noise in decibels (dbs) is given by

$$\text{RPSD} = 10 \log_{10} \frac{S_y(e^{j\omega T})}{S_{e_i}(e^{j\omega T})} \quad (1.4)$$

The RPSD is a useful measure for noise performance of digital filters, because it eliminates the dependence of the output noise on the wordlength. Hence the RPSD is a measure of the extent to which the output noise depends upon the internal structure of the filter.

Another useful performance measure for the assessment of roundoff-noise in digital filters is the noise gain or relative variance of the noise given by

$$\frac{\sigma^2}{\sigma_0^2} = \frac{2}{\omega_s} \int_0^{\omega_s/2} \sum_{i=1}^k |G_i(e^{j\omega T})|^2 d\omega \quad (1.5)$$

## II. Granularity Limit Cycles

On many occasions, signal levels in a digital filter can become constant or very low, at least for short periods of time. Under such circumstances, the noise signals become highly correlated from sample to sample and from source to source. This correlation can cause autonomous oscillations called granularity limit cycles [1] - [3]. In effect, the filter can become unstable under certain circumstances.

Limit-cycles oscillations can occur in recursive digital filters implemented with rounding, magnitude truncation and other types of quantization [4] - [5].

In many applications, the presence of limit cycles can be a serious problem. Thus, it is desirable to eliminate limit cycles or to keep their amplitude bounds low.

For wave structures and some other second-order structures, the stability can be demonstrated in the linear case (infinite precision arithmetic) as well as in the nonlinear case (finite precision arithmetic) by means of the second method of Liapunov [6]. In the nonlinear case magnitude truncation is applied to quantize suitable signals inside the structure such that a defined positive definite pseudo-energy function is proved to be a Liapunov function [7] - [9].

The concept of pseudo-energy function was first applied for the elimination of zero-input limit cycles in [7] - [9]. More recently some papers dealing with the control of constant-input limit-cycles have been published [4], [10] - [12]. In [4], [9] and [10], controlled rounding arithmetic was introduced and applied successfully to certain types of wave structures and in [12] the concept of pseudo-energy has been extended to the elimination of constant-input limit cycles in a particular second-order structure.

### 1.2.2 Overflow Limit Cycles

Overflow limit cycles can occur when the magnitudes of internal signals exceed the available register range [13]. In order to prevent the increase of the signal wordlength in recursive digital filters, nonlinear signal operations, referred to as overflow nonlinearities, must be applied. Since overflow nonlinearities influence the most



significant bits, they cause severe signal distortion. An overflow may start self-sustained, high-amplitude oscillations called overflow limit cycles.

A structure is considered to be free of overflow limit cycles or has a stable forced response if the error which is introduced in the filter after an overflow decreases with time in such a way that the output of the nonlinear filter converges to the output of the ideal linear filter [13].

Overflow can occur in any structure in the presence of an input signal, and input scaling is generally required to reduce the probability of overflow to an acceptable level. In order to maximize the dynamic range in fixed-point digital-filter implementations, signal scaling must be applied so as to ensure that the probability of overflow is the same at each node.

The scaling technique proposed by Jackson [1], which is applicable to one's-and two's-complement implementations, requires only the multiplier inputs to be scaled. In this technique a scaling multiplier is used at the input of a filter section, as depicted in Fig. 1.2. The constant  $\lambda_i$  can be chosen on the basis of the  $L_p$  norm of  $F_i(z)$  [1] - [2], depending on the properties of the input signal. The  $L_p$  norm of  $F_i(z)$  is defined as

$$\|F_i\|_p = \left( \frac{1}{\omega_s} \int_0^{\omega_s} |F_i(e^{j\omega T})|^p d\omega \right)^{1/p} \quad (1.6)$$

where  $F_i(z)$  is the transfer function from the filter input to the input of multiplier  $P_i$ . The scaling ensures that the amplitudes of multiplier inputs are bounded by  $M$  if  $|x(n)| < M$ . Therefore, in order to ensure that all multiplier inputs are bounded by  $M$ , we must assign

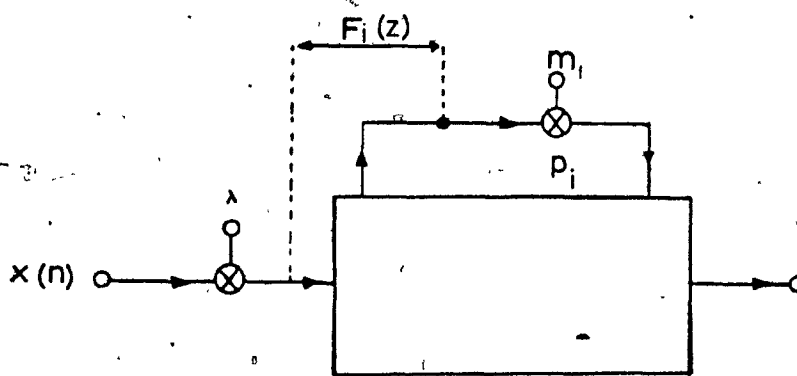


Fig. 1.2: Signal Scaling.

$$\lambda = \frac{1}{\max(\|F_1\|_p, \dots, \|F_i\|_p, \dots, \|F_m\|_p)} \quad (1.7)$$

where  $m$  is the number of multipliers in the section.

The order of the norm  $p$  is usually chosen to be  $\infty$  or 2. The  $L_\infty$  norm is used for input signals which have some dominating frequency component, whereas the  $L_2$  norm is most commonly used for filters with a random input signal.

It is common in practice to use powers of 2 for the scaling coefficients, provided they satisfy the overflow constraints. In this way, the scaling multipliers can actually be implemented by simple shift operations.

In the case of cascade or parallel realizations of digital filters, optimum scaling is accomplished by applying one scaling multiplier per section. In some cascade designs, the scaling multiplier of each section can be eliminated by incorporating it in the output multipliers of the previous section [1], [14]. This, in general, leads to an improvement in the roundoff noise performance.

### 1.2.3 Coefficient Quantization

During the approximation step the coefficients of a digital filter are calculated with high accuracy. If these coefficients are quantized, the frequency response of the realized digital filter will deviate from the ideal response. In fact, the quantized filter may even fail to meet the desired specifications.


It is well known that the sensitivity of the filter response to errors in the coefficients is highly dependent on the type of

structure used [15] - [16]. This fact motivated many researchers to develop low-sensitivity structures [16] - [21].

In [17] and [18], low-sensitivity structures are obtained by transforming equally-terminated, LC-ladder filters into corresponding digital filters. The low sensitivity of the digital filters is due to the inherently low sensitivity of the equally-terminated LC filters.

In many instances digital filters are realized using a cascade or parallel connection of second-order sections. If the sections are realized using the direct canonic form of Fig. 1.3, the sensitivity problems become serious when the poles are close to the unit circle  $|z| = 1$ . In order to solve this problem, Agarwal and Burrus [19] proposed a delay replacement scheme whereby  $z^{-1}$  is replaced by  $1/(z-1)$ , which yields low-sensitivity structures. On the other hand, Nishimura, Hirano and Pal [20] introduced a delay replacement scheme whereby  $z^{-1}$  is replaced by  $z/(z-1)$ , which was also shown to be useful.

Agarwal and Burrus [19] also proposed modifications to the recursive part of the direct canonic form of Fig. 1.3 such that the absolute values of  $m_1$  and  $m_2$  become as small as possible for poles close to  $z = 1$ . In this way, the sensitivities of the transfer function with respect to  $m_1$  and  $m_2$  become low. However, when the poles are not close to  $z = 1$  the sensitivities of the structures proposed in [19] and [20] are no longer low.



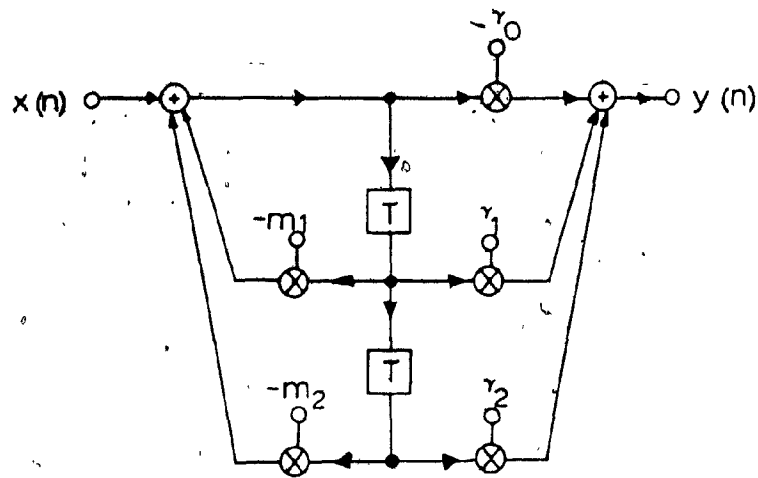


Fig. 1.3: Direct Canonic Structure

#### 1.2.4 Input Quantization

Input quantization is similar to product quantization and can be represented by including a noise source at the input of the structure. The output PSD due to input noise can be evaluated by using an equation like Eqn. 1.3.

### 1.3 Digital Filter Structures

Many types of digital filter structures have been proposed in the literature [1] - [3]. The choice of the structure for a specific application must be based on the cost and performance under finite-precision arithmetic. In this section some important types of structures are briefly described.

#### 1.3.1 Wave Structures

Wave structures are obtained by applying the concept of wave characterization [17] - [18] to an analog prototype filter. Capacitances and inductors are transformed into delays characterized by  $z^{-1}$  and  $-z^{-1}$ , respectively. These components are interconnected through parallel and series adaptors [2], [17] - [18], [22]. The synthesis of wave digital filters is described in detail in Chapter 12 of [2].

#### 1.3.2 Cascade and Parallel Direct Canonic Structures

The realization of high-order digital filters by using a connection of direct canonic second-order sections in cascade or in

parallel, as depicted in Fig. 1.4(a) and (b), is quite popular. These realizations are attractive because they are simple and economical with respect to the number of delays, multipliers and adders.

### 1.3.3 State-Space Section-Optimal Structures

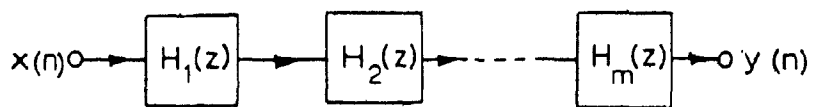
A synthesis method for obtaining low-roundoff-noise state-space realizations of recursive digital filters has been developed independently in [23] - [24] and [25]. For an  $n$ th-order filter, this synthesis results in a structure which requires  $(n+1)^2$  multipliers. The number of multipliers is, therefore, large and hence these structures are avoided in practice. In order to eliminate this problem, some authors proposed that the state-space approach be applied only to the individual second-order sections of the cascade or parallel designs [21], [23], [26]. The simplest design of optimal second-order state-space structure was provided by Jackson, Lindgren, and Kim [21].

The structure of Fig. 1.5 is characterized by

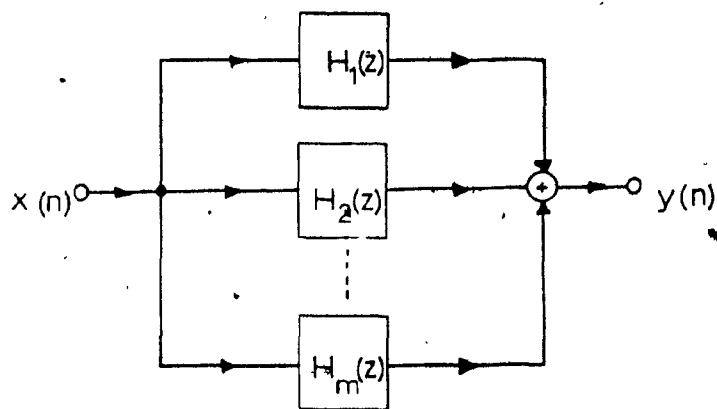
$$\left. \begin{aligned} \underline{x}(n+1) &= \underline{A} \underline{x}(n) + \underline{B} u(n) \\ y(n) &= \underline{C} \underline{x}(n) + \underline{D} u(n) \end{aligned} \right\} \quad (1.8)$$

where  $\underline{x}(n)$  is a column vector,  $y(n)$  is a scalar and

$$\underline{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \underline{B} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \underline{C} = [c_1 \ c_2], \quad \underline{D} = [d]$$



(a)



(b)

Fig. 1.4: (a) Cascade Realization  
(b) Parallel Realization



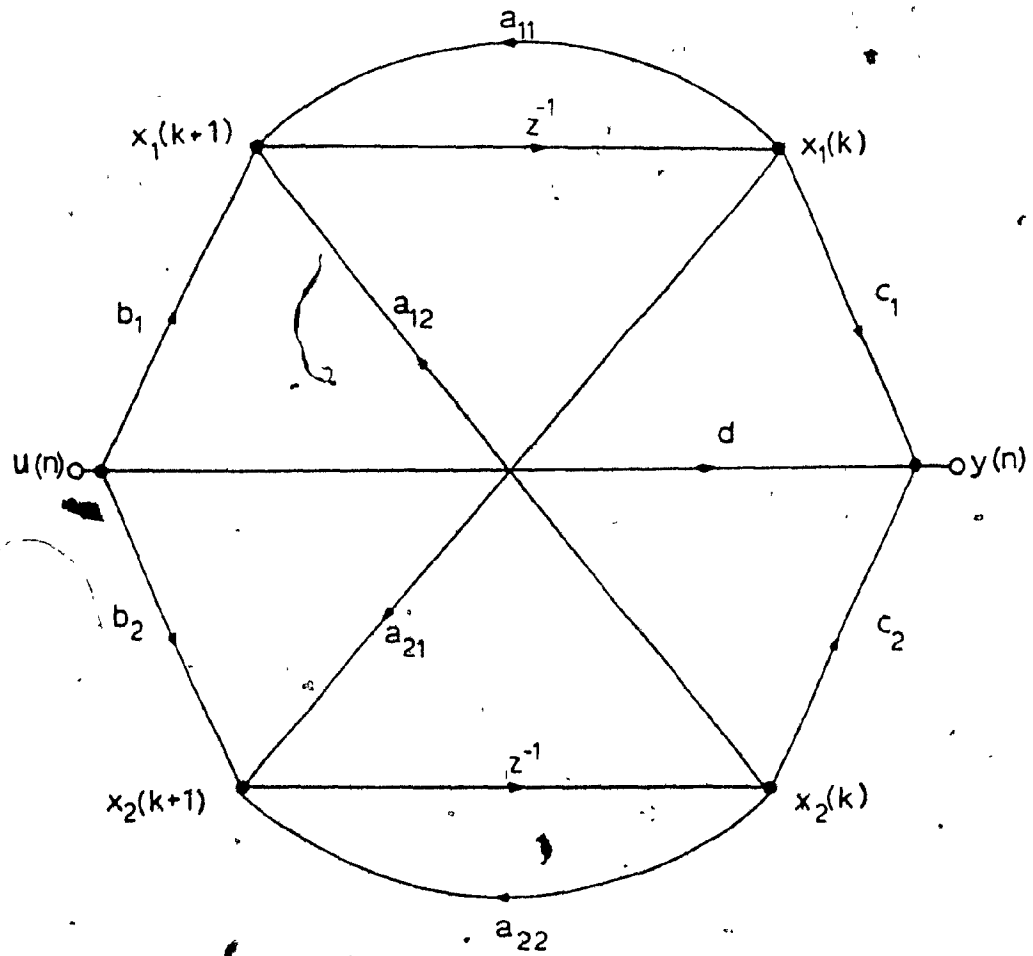


Fig. 1.5: Second-Order State-Space Structure

It can readily be used to realize an arbitrary second-order transfer function of the form

$$H(z) = d + \frac{\beta_1 z + \beta_2}{z^2 + \alpha_1 z + \alpha_2} \quad (1.9)$$

Minimum roundoff noise can be achieved by using the following procedure [21]:

1) Define an unscaled system represented by  $\tilde{A}'$ ,  $\tilde{B}'$ ,  $\tilde{C}'$ ,  $\tilde{D}'$  such that

$$\begin{aligned} a'_{11} &= a'_{22} = -\frac{\alpha_1}{2} \\ b'_1 &= \frac{1}{2}(1+\beta_2); \quad b'_2 = \frac{1}{2}\beta_1 \end{aligned} \quad (1.10)$$

$$c'_1 = \frac{\beta_1}{1+\beta_2}; \quad c'_2 = 1$$

$$a'_{12} = \beta_1^{-2}(1+\beta_2) \left[ (\beta_2 - \frac{1}{2}\alpha_1\beta_1) \pm \sqrt{\beta_2^2 - \beta_1\beta_2\alpha_1 + \beta_1^2\alpha_2} \right]$$

$$a'_{21} = (1+\beta_2)^{-1} \left[ (\beta_2 - \frac{1}{2}\alpha_1\beta_1) \mp \sqrt{\beta_2^2 - \beta_1\beta_2\alpha_1 + \beta_1^2\alpha_2} \right]$$

$$d' = d$$

2) Obtain an optimal  $L_p$ -scaled system, represented by  $\tilde{A}$ ,  $\tilde{B}$ ,  $\tilde{C}$ ,  $\tilde{D}$ , as

$$\left. \begin{aligned} \tilde{A} &= \tilde{I}^{-1} \tilde{A}' \tilde{I} & \tilde{B} &= \tilde{I}^{-1} \tilde{B}' \\ \tilde{C} &= \tilde{C}' \tilde{I} & d &= d' \end{aligned} \right\} \quad (1.11)$$

where

$$\tilde{I} = \begin{bmatrix} \|F'_1(z)\|_p & 0 \\ 0 & \|F'_2(z)\|_p \end{bmatrix} \quad (1.12)$$

$F'_i(z)$  are the transfer functions from input node  $x(n)$  to the state variable node  $x_i(n+1)$  in the system  $(\underline{A}', \underline{B}', \underline{C}', \underline{D}')$  and are given by

$$\left. \begin{aligned} F'_1(z) &= \frac{b'_1 z + b'_2 a'_{12} - b'_1 a'_{22}}{z^2 - (a'_{11} + a'_{22})z + a'_{11} a'_{22} - a'_{12} a'_{21}} \\ F'_2(z) &= \frac{b'_2 z + b'_1 a'_{21} - b'_2 a'_{11}}{z^2 - (a'_{11} + a'_{22})z + a'_{11} a'_{22} - a'_{12} a'_{21}} \end{aligned} \right\} \quad (1.13)$$

The transfer functions from the state variable nodes  $x_1(n+1)$  and  $x_2(n+1)$  to the filter output are given by

$$\left. \begin{aligned} G_1(z) &= \frac{c_1 z + a_{21} c_2 - a_{22} c_1}{z^2 - (a_{22} + a_{22})z + a_{11} a_{22} - a_{12} a_{21}} \\ G_2(z) &= \frac{c_2 z + a_{12} c_1 - a_{11} c_2}{z^2 - (a_{11} + a_{22})z + a_{11} a_{22} - a_{12} a_{21}} \end{aligned} \right\} \quad (1.14)$$

These are needed for noise analysis.

The resulting second-order structure is optimal for  $L_2$  scaling. However, it may be scaled in terms of  $L_\infty$  scaling which yields a near-optimal structure. The optimal second-order state-space section is often referred as section-optimal structure.

Another important class of state-space sections are the normal sections which satisfy the relation [26] - [27]

$$\underline{A} \underline{A}^T = \underline{A}^T \underline{A}.$$

In section-optimal and normal sections, overflow and zero-input limit cycles can be eliminated by applying magnitude truncation to the state variables [28] - [29].

#### 1.3.4 Error Spectrum Shaping Technique

The error spectrum shaping technique (ESS) comprises of the generation of an error signal and the application of local feedback for the purpose of forcing zeros in the PSD of the output noise [30] - [35]. The technique is implemented by replacing each and every adder whose inputs include at least one nontrivial product\* by the recursive sub-structure shown in Fig. 1.6(a) or (b). The complexity of hardware increases if the number of adders of the type described is increased. This fact makes the cascade or parallel connection of direct canonic sections particularly useful for ESS application, because it requires only one ESS sub-structure per second-order section.

The choice of the coefficients  $b_0$  and  $b_1$  must be made such that the output PSD is minimized. The optimum values of  $b_0$  and  $b_1$  for a given pole position, where  $b_0$  is restricted to 0,  $\pm 1$  and  $b_1$  to 0,  $\pm 1$  and  $\pm 2$  are given in reference [35].

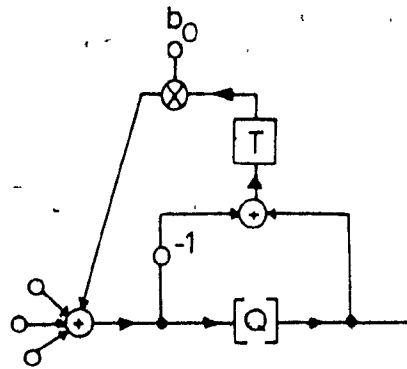
The values of  $b_{1j}$  and  $b_{0j}$  that minimize the output noise in a cascade design are given by solving the following optimization problem:

$$\min_{b_{0j}, b_{1j}} \left\| (z^{-2}b_{0j} + z^{-1}b_{1j} + 1) \prod_{i=j}^m H_i(z) \right\|_2^2 \quad (1.15)$$

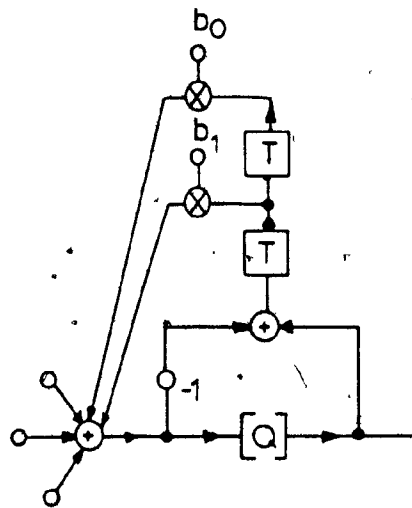
The optimum value of coefficient  $b_{0j}$  for first-order ESS is given by

---

\*A signal multiplied by a noninteger constant



(a)



(b)

Fig. 1.6: Application of ESS

- (a) First-Order ESS
- (b) Second-Order ESS

$$b_{0j} = \frac{-p_1}{p_3}$$

(1.16)

whereas the optimum values of  $b_{0j}$  and  $b_{1j}$  for second-order ESS are given by [34]

$$\left. \begin{aligned} b_{1j} &= \frac{p_1 p_2 - p_1 p_3}{p_3^2 - p_1^2} \\ b_{0j} &= \frac{p_1^2 - p_2 p_3}{p_3^2 - p_1^2} \end{aligned} \right\} \quad (1.17)$$

where

$$p_1 = \int_{-\pi}^{\pi} \left| \prod_{L=j}^m H_L(e^{j\omega T}) \right|^2 \cos(\omega) d\omega,$$

$$p_2 = \int_{-\pi}^{\pi} \left| \prod_{L=j}^m H_L(e^{j\omega T}) \right|^2 \cos(2\omega) d\omega,$$

$$p_3 = \int_{-\pi}^{\pi} \left| \prod_{L=j}^m H_L(e^{j\omega T}) \right|^2 d\omega$$

and

$$T = 1.$$

The output RPSD is given by

$$\text{RPSD} = 3 + \sum_{i=1}^m \left| (1/\lambda_i) (1 + b_{1i} z^{-1} + b_{0i} z^{-2}) \prod_{L=i}^m H_L(z) \right|^2 \quad (1.18)$$

where  $z = e^{j\omega T}$  and  $\lambda_i$  is the scaling constant for section  $i$ .

Limit cycles are likely to occur in structures using ESS.

However, their amplitudes are expected to be low [36] - [37], and for some choices of coefficients  $b_0$  and  $b_1$ , zero-input granularity limit cycles can be eliminated [36].

#### 1.4 Scope of the Thesis

A number of different approaches to the realization of digital filters have been proposed in the literature. However, no general agreement has emerged so far as to which realization is the most advantageous. Two approaches that seem to have attractive features in the realization of high-order digital filters are the cascade and parallel realizations, where second-order sections are connected in cascade or in parallel. The main advantages of these realizations are their higher modularity, which simplifies VLSI implementation, the simplicity of roundoff-noise and sensitivity analysis, and the ease with which limit cycles can be studied. In addition, when fixed-point arithmetic is used, these realizations can be less noisy and/or sensitive when compared with other forms of realization, such as ladder-based wave structures [16] - [18].

The motivation of this work is to generate second-order digital-filter structures, which present several desired characteristics together, such as reduced roundoff noise and sensitivity, elimination of overflow and granularity limit-cycles, reduction of the number of multipliers, and simultaneous realization of different types of transfer functions by one and the same structure.

Although all of the above desired features are unlikely to occur in the same second-order structure, we have been able to develop some structures that achieve most of these features.

In Chapter 2, we begin by developing a digital-filter synthesis based on the analog voltage-conversion type generalized immittance

converter (VGIC) by using the wave characterization. By means of this synthesis, several VGIC structures are obtained which are canonical with respect to the number of multipliers and delays. Through the application of Tellegen's theorem, a transposed VGIC structure (TVGIC) is obtained which realizes simultaneously a lowpass, a bandpass, and a highpass transfer function. It can also realize a transfer function with zeros on the unit circle  $|z| = 1$  using the minimum number of multipliers. A special case of the TVGIC structure is shown to be amenable to the application of ESS technique. Subsequently, a universal digital biquad is developed which realizes all the standard transfer functions simultaneously.

Chapter 3 describes a systematic and exhaustive procedure for the generation of second-order, low-sensitivity, digital-filter structures which are amenable to ESS. The procedure is used to generate two sets of structures which include many new as well as a few known structures. Collectively, these structures can realize any stable second-order transfer function. The emphasis is placed on generating second-order structures which can be used in cascade or in parallel for the realization of high-order transfer functions.

Chapter 4 begins by showing that zero-input and overflow limit cycles can be eliminated in the VGIC and TVGIC structures. A new theorem is then proved which establishes sufficient conditions that will ensure freedom from constant-input limit cycles in a general structure in which zero-input limit cycles can be eliminated. By using this theorem, it is shown that constant-input limit cycles can be eliminated in a subset of the VGIC structures, in the TVGIC structure, and also the universal digital biquad proposed.



In Chapter 5, the theorem proposed in Chapter 4 for the elimination of constant-input limit cycles is applied to a general state-space second-order structure, and the conditions for efficient elimination of constant-input limit cycles are established. These conditions lead to three different state-space second-order structures. In addition, a design procedure is described which yields, near-optimal new state-space structures.

Chapter 6 studies the effect of product quantization in the VGIC and TVGIC structures and also the TVGIC structure with ESS incorporated, and several comparisons are made with the direct canonic and section-optimal structures. Also the low-sensitivity structures of Chapter 3 are compared with the section-optimal structure. Subsequently, the new state-space and section-optimal structures are compared. The various comparisons are made on the basis of output-noise spectra in several sixth-order filter designs which include lowpass, highpass, bandpass and bandstop filters.

In Chapter 7, a procedure for improving the sensitivity performance of the VGIC structures is described. The sensitivity properties of the low-sensitivity structures of Chapter 3 are then discussed. For every transfer function, a choice of at least two structures is available but, through a sensitivity analysis, the optimum one can always be identified. This chapter also deals with sensitivity properties of the new state-space and section-optimal structures. The effect of coefficient quantization is treated at length by computing the actual amplitude responses under fixed-point finite arithmetic in several sixth-order filter designs.

Extensive comparisons are then undertaken. The VGIC structures of Chapter 2 and the low-sensitivity structures of Chapter 3 are compared with the direct canonic and section-optimal structures. In addition, the new state-space structure of Chapter 5 is compared with the section-optimal structure.

The thesis concludes with overall comparisons of the structures proposed relative to the conventional structures. Specifically, a table is constructed which gives the advantages and disadvantages of the various structures.

## CHAPTER 2

### VGIC-BASED STRUCTURES

#### 2.1 Introduction

Given an analog-filter configuration, a corresponding digital-filter structure can be obtained by utilizing the method proposed by Fettweis [17], [18]. In this chapter, the method of Fettweis is used to generate a general biquadratic digital-filter structure from an active analog-filter configuration which is based on the voltage-conversion generalized-immittance converter (VGIC) [38]. Then several specific structures are deduced which collectively can realize most of the standard transfer functions encountered in the design of cascade or parallel digital filters.

Through the application of Tellegen's theorem [2], a transposed version of the general VGIC (TVGIC) structure is derived which offers several advantages. First, it realizes simultaneously a lowpass, a bandpass, and a highpass transfer function; second, it realizes a transfer function with zeros on the unit circle using the minimum number of multipliers; and third, it is amenable to the application of error spectrum shaping (ESS), which is a technique for the reduction of roundoff noise (see Sec. 1.3.4).

The chapter concludes with a comparison of the computational complexity encountered in VGIC and TVGIC structures relative to that in the direct canonic and the section-optimal structures [21]. The latter is considered as a very-low-noise structure and is in direct competition with TVGIC structures incorporating ESS.

## 2.2 Generation of the Active Analog-Filter Configuration

The analog voltage-conversion type generalized-immittance converter (VGIC) [38] is a two-port network as depicted in Fig. 2.1.

It is characterized by the voltage-current equations

$$\left. \begin{aligned} V_1(s) &= r(s) - V_2(s) \\ I_1(s) &= -I_2(s) \end{aligned} \right\} \quad (2.1)$$

where  $r(s)$  is said to be the conversion function and the pairs  $(V_1, I_1)$  and  $(V_2, I_2)$  are the voltage-current pairs of the VGIC at ports 1 and 2, respectively.

The VGIC has not been used extensively in the design of RC-active circuits, because it is difficult to implement with conventional active devices. However, no such difficulty is encountered in using the VGIC in the design of digital filters.

The analog configuration shown in Fig. 2.2 realizes a continuous-time biquadratic current transfer function of the form

$$\begin{aligned} \frac{\tilde{I}_1(s)}{I_1(s)} &= \frac{K_2 R_2 s^2 + K_1 R_1 s + K_0 R_0}{R_2 s^2 + R_1 s + R_0} \\ &= \frac{c_2 s^2 + c_1 s + c_0}{R_2 s^2 + R_1 s + R_0} \end{aligned} \quad (2.2)$$

where  $r(s)$  has been assumed to be equal to  $s$ . Since  $\tilde{I}_0 = \tilde{I}_1 = \tilde{I}_2$ , according to Eqn. 2.1 and Fig. 2.1, the same transfer function is obtained when current  $\tilde{I}_0$  or  $\tilde{I}_2$  is taken as the output.

The motivation to start with the configuration of Fig. 2.2 has been based on the fact that several equivalent outputs are available, namely,  $\tilde{I}_0$ ,  $\tilde{I}_1$  and  $\tilde{I}_2$ , and additional equivalent outputs can be

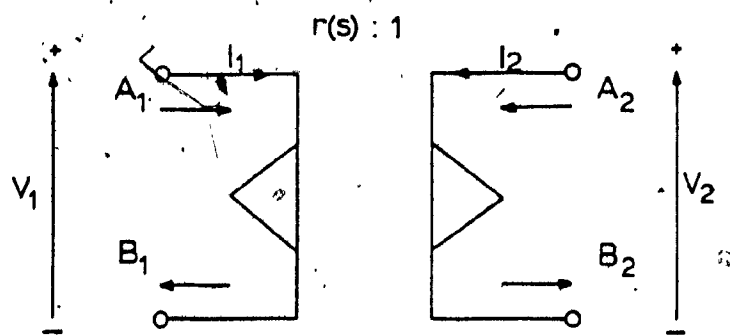


Fig. 2.1: VGIC

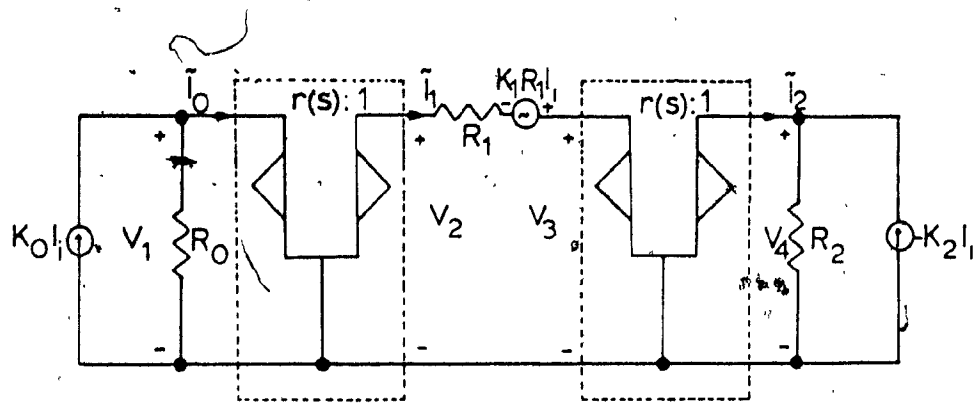


Fig. 2.2: VGIC-Based Active Analog-Filter Configuration

located after the prototype is converted into a digital structure. This property provides a design flexibility which will allow us to generate some interesting structures in Secs. 2.3 - 2.4.

### 2.3 Digital VGIC Structure

It is well known that any analog n-port network can be characterized by using the concepts of incident and reflected wave quantities. Through the application of the wave characterization, and the use of the bilinear transformation, digital realizations can be obtained for passive and active elements as described by Fettweis [17] - [18]. By this means, analog n-port networks can be converted into corresponding digital networks. In order to obtain a digital-filter structure from the network of Fig. 2.2, we need only develop an economical digital realization for the VGIC.

The VGIC of Fig. 2.1 can be described in terms of the wave characterization as

$$\left. \begin{aligned} A_1 &= V_1 + \frac{I_1}{G_1}, & A_2 &= V_2 + \frac{I_2}{G_2} \\ B_1 &= V_1 - \frac{I_1}{G_1}, & B_2 &= V_2 - \frac{I_2}{G_2} \\ I_1 &= -I_2, & V_1 &= r(s)V_2 \end{aligned} \right\} \quad (2.3)$$

where  $A_i$  and  $B_i$  ( $i=1,2$ ) are the incident and reflected wave quantities, respectively, and  $G_i$  is the port conductance assigned to port  $i$ .

After some manipulation, we obtain the values of  $B_1$  and  $B_2$  in terms of  $A_1$ ,  $A_2$ ,  $G_1$ ,  $G_2$  and  $r(s)$  as

$$\left. \begin{aligned} B_1 &= \frac{r(s)G_1 - G_2}{r(s)G_1 + G_2} A_1 + \frac{2r(s)G_2}{r(s)G_1 + G_2} A_2 \\ B_2 &= \frac{2G_1}{r(s)G_1 + G_2} A_1 + \frac{G_2 - G_1 r(s)}{r(s)G_1 + G_2} A_2 \end{aligned} \right\} \quad (2.4)$$

Now by applying the bilinear transformation

$$s \rightarrow \frac{z-1}{z+1}$$

and by letting  $G_1 = G_2$  and  $r(s) = s$ , we obtain

$$\left. \begin{aligned} B_1 &= -z^{-1} A_1 + (1-z^{-1})A_2 \\ B_2 &= (1+z^{-1})A_1 + A_2 z^{-1} \end{aligned} \right\} \quad (2.5)$$

In this way an economical digital realization of the VGIC can be derived from Eqn. 2.5 as shown in Fig. 2.3.

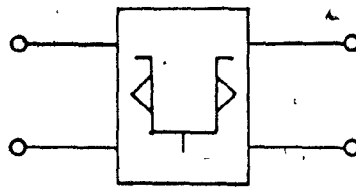
Now by using known digital realizations for current and voltage sources and series interconnections [22], a digital structure for the analog network of Fig. 2.2 can be obtained as shown in Fig. 2.4. Outputs  $y_0(n)$ ,  $y_1(n)$  and  $y_2(n)$  correspond to the cases where currents  $\tilde{I}_0$ ,  $\tilde{I}_1$  and  $\tilde{I}_2$  are taken as outputs, respectively.

The actual transfer function realized turns out to be a transimpedance and is given by

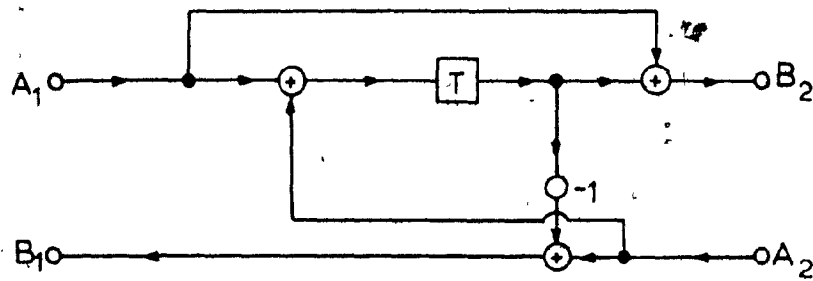
$$\left. \frac{Y_j(z)}{X_i(z)} = \frac{2 R_j \tilde{I}_j(s)}{I_i(s)} \right|_{s = \frac{z-1}{z+1}} \quad (2.6)$$

where  $j = 0, 1, 2$ .





(a)



(b)

Fig. 2.3: (a) Symbol for the Digital VGIC  
(b) Digital VGIC with  $r(s)=s$

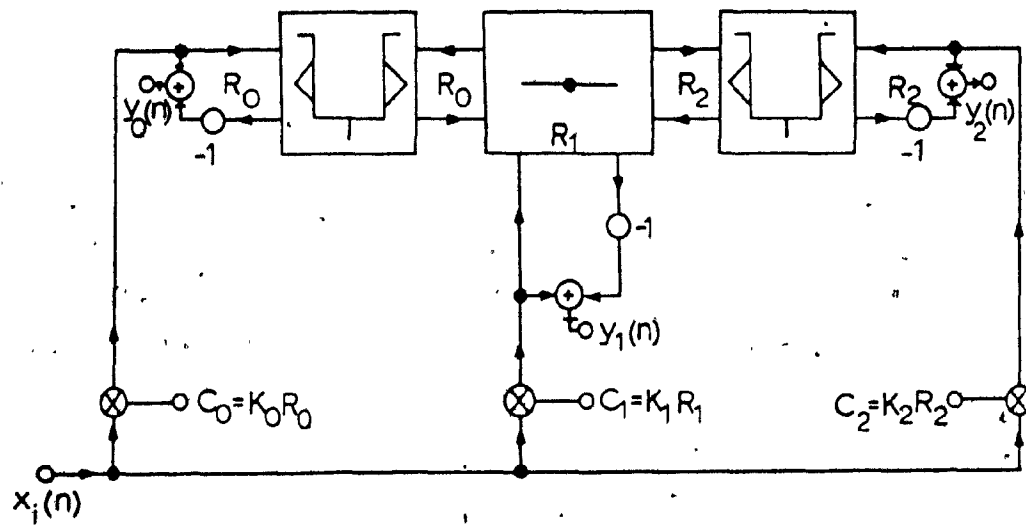


Fig. 2.4: General VGIC Biquadratic Structure

The structure of Fig. 2.4 can assume several forms, depending on the type of series adaptor used. If the adaptor reported in [22] is used the digital-filter structure of Fig. 2.5 is obtained, where  $y_0(n)$  to  $y_7(n)$  represent the various equivalent outputs. The equivalence of outputs  $y_3(n)$  to  $y_7(n)$  with respect to  $y_0(n)$ ,  $y_1(n)$  and  $y_2(n)$  can be demonstrated by inspection.

The availability of several equivalent outputs in Fig. 2.5 leads to a reduction in the number of adders. In addition, through the use of transposition, a digital structure can be obtained with several equivalent inputs. By choosing the input appropriately, structures can be obtained in which either limit cycles can be eliminated or the level of quantization noise can significantly be decreased through the use of error spectrum shaping (see Chapter 3).

If signal  $y_3(n)$  is taken as output in Fig. 2.5, the number of adders is decreased and, furthermore, the scaling of the filter can be simplified. The transfer function in this case becomes

$$\frac{Y_3(z)}{X_i(z)} = \frac{(c_0+c_1+c_2)z^2 + 2(c_0-c_2)z + c_0-c_1+c_2}{z^2+(m_1-m_2)z + m_1+m_2-1} \quad (2.7)$$

Fig. 2.6 shows the different types of digital sections that can be obtained from the general structure Fig. 2.5, in the case where signal  $y_3(n)$  is taken as output. The new structures are canonical with respect to the number of multipliers and unit-delays, except for the allpass structure which is not canonical with respect to the number of multipliers. The number of adders is always comparable to that in the corresponding direct canonic second-order structure.

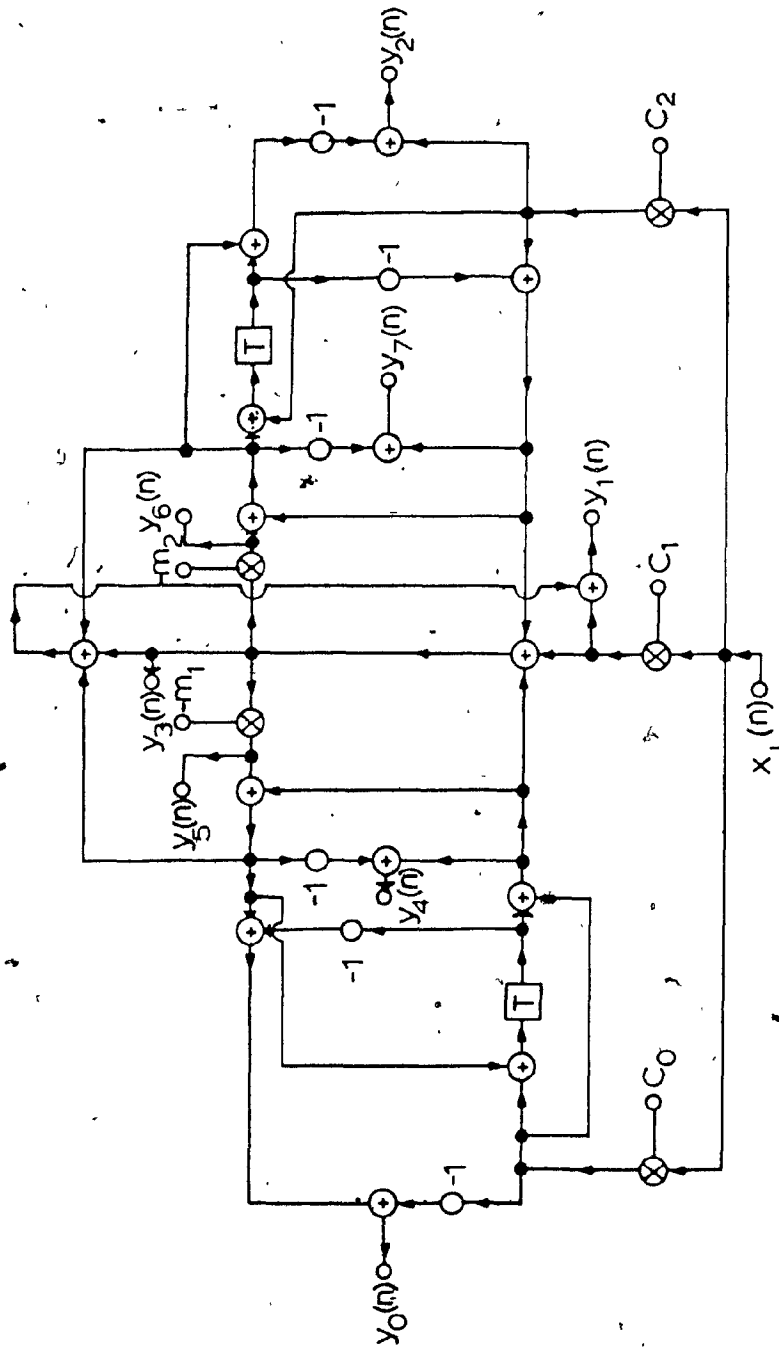
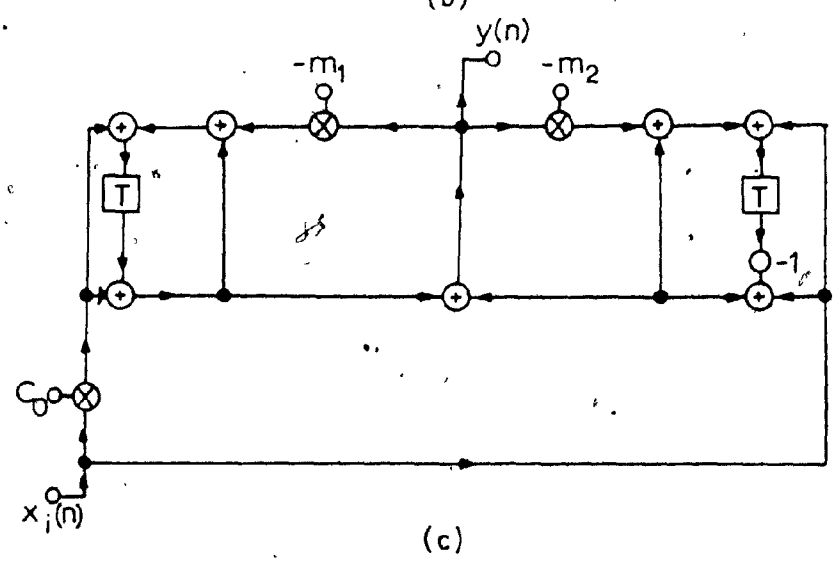
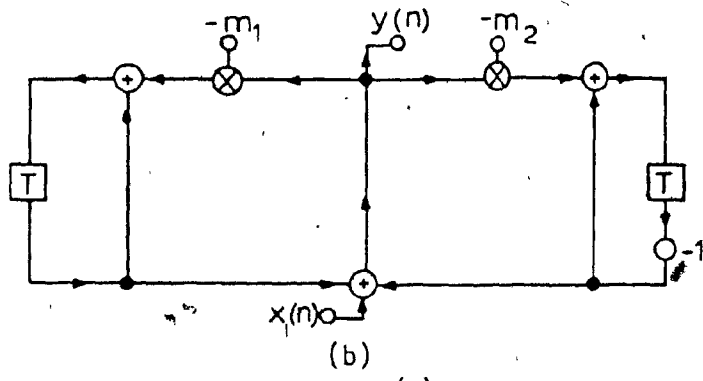
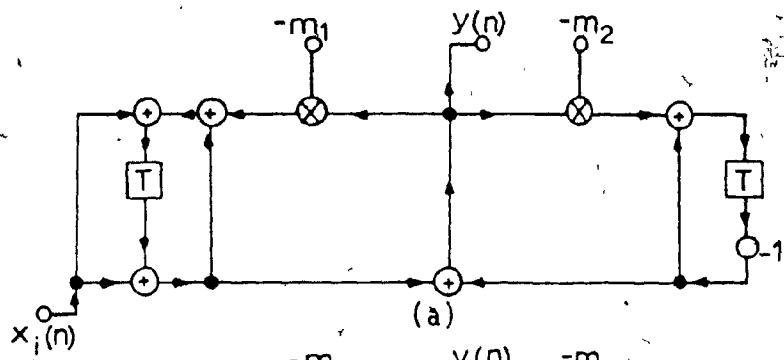


Fig. 2.5: General VGIC Biquadratic Structure



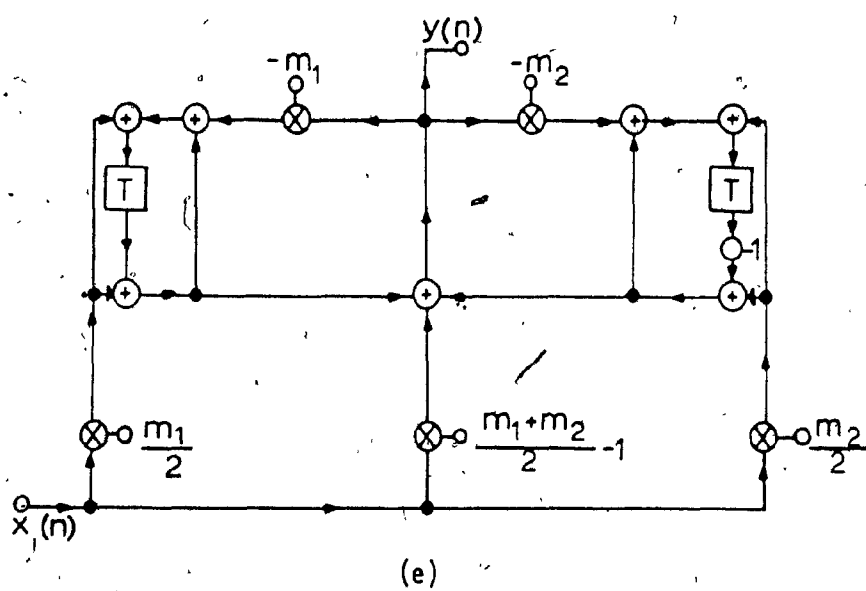
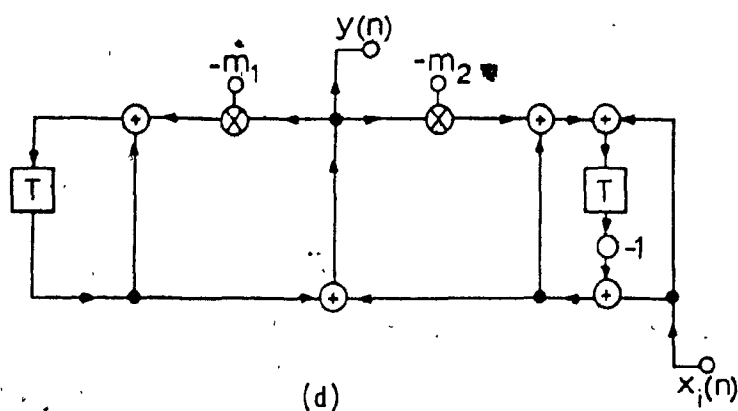


Fig. 2.6: Standard Second-Order VGIC Sections

- (a) Lowpass
- (b) Bandpass
- (c) Notch (zeros on the unit circle)
- (d) Highpass
- (e) Allpass

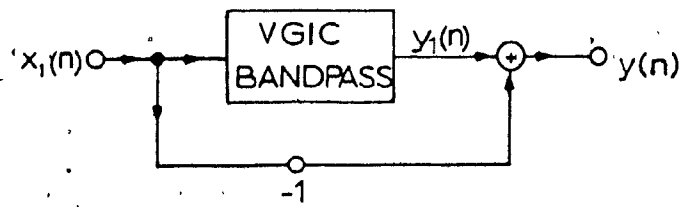
An alternative, two-multiplier allpass structure can be generated as illustrated in Fig. 2.7(a) and (b) and is characterized by

$$\begin{aligned} Y_{AP}(z) &= H_{BP}(z) X(z) - X(z) \\ &= \left[ (2-m_1-m_2) \frac{z^2-1}{z^2+(m_1-m_2)z+m_1+m_2-1} - 1 \right] X(z) \\ &= \frac{(1-m_1-m_2)z^2-(m_1-m_2)z-1}{z^2+(m_1-m_2)z+m_1+m_2-1} X(z) . \end{aligned} \quad (2.8)$$

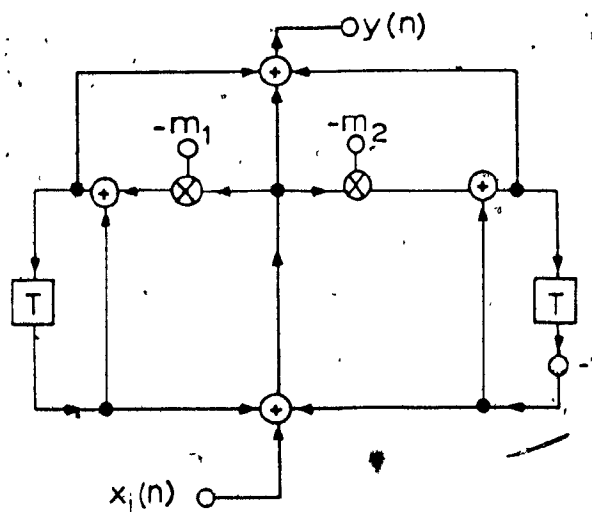
The VGIC bandpass structure in Fig. 2.7(a) can be generated utilizing  $y_1(n)$  as output in the general structure of Fig. 2.5.

#### 2.4 Universal Digital Biquads with Simultaneous Outputs

An interesting property of the general VGIC structure of Fig. 2.5 is that three different types of transfer functions can be obtained by setting any two of the input multiplier constants,  $c_0$ ,  $c_1$  and  $c_2$  to zero. That is, a lowpass transfer function with zeros at  $z=-1$  is obtained if  $c_1=c_2=0$ , a bandpass transfer function with zeros at  $z = \pm 1$  is obtained if  $c_0=c_2=0$ , and a highpass transfer function is obtained if  $c_0=c_1=0$ . This is a very useful feature because if the transpose of the VGIC structure is formed [1], a second-order section can be generated, as illustrated in Fig. 2.8, which realizes simultaneously a lowpass, a bandpass, and a highpass transfer function. This structure also realizes a transfer function with zeros anywhere on the unit circle using the minimum number of multipliers. Such a transfer function is useful for the design of elliptic filters. In addition, an allpass transfer function can be obtained, which unfortunately is not canonical with respect to the number of multipliers. The various simultaneous outputs of this structure are



(a)



(b)

Fig. 2.7: Allpass Section Using the VGIC Bandpass Structure of Fig. 2.5

- (a) Block Diagram
- (b) Structure



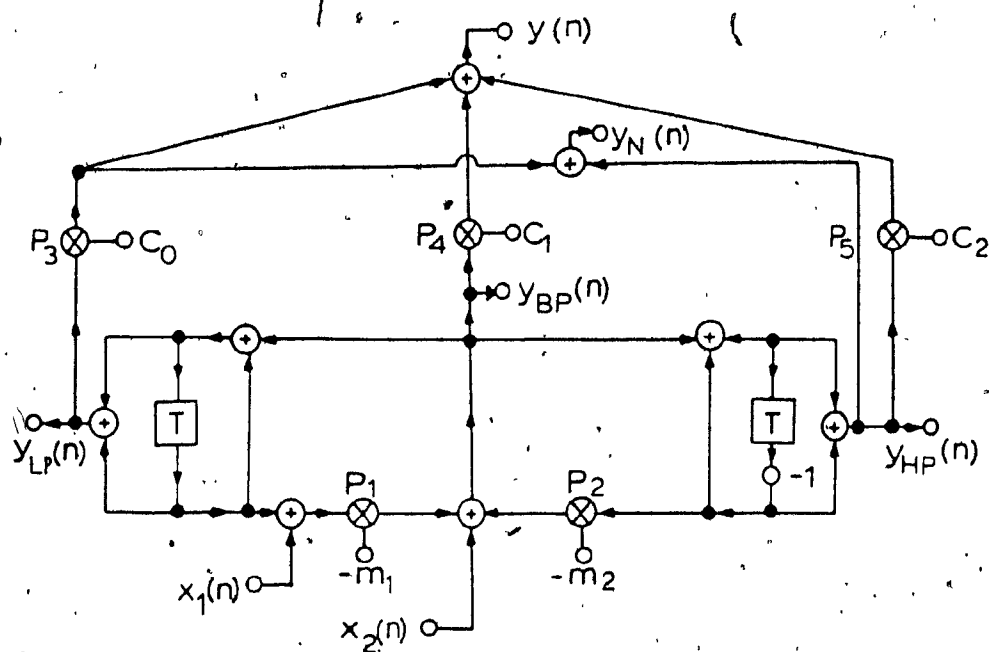


Fig. 2.8: TVGIC General Structure

illustrated in Fig. 2.8, where  $y_{LP}(n)$ ,  $y_{BP}(n)$  and  $y_{HP}(n)$ , indicate the lowpass, bandpass, and highpass outputs, respectively. Output  $y_N(n)$  corresponds to the transfer function with zeros on the unit circle, while  $y(n)$  corresponds to a general biquadratic transfer function. Note that only two inputs are indicated in Fig. 2.8, namely,  $x_1(n)$  and  $x_2(n)$ , although eight are possible. This choice appears to be optimum at this point since limit-cycles can be eliminated by using  $x_1(n)$  as input (see Chapter 4). Alternatively, the output noise can be reduced by using  $x_2(n)$  as input (see Sec. 2.5).

An alternative universal digital biquad which realizes simultaneously a lowpass, a highpass, a bandpass, an allpass, and a general biquadratic transfer function can be derived by combining the VGIC bandpass and allpass structures of Figs. 2.6(b) and 2.7(b) with the lowpass and highpass structures of Verkröost [11] as depicted in Fig. 2.9.

## 2.5 TVGIC Structure With Error Spectrum Shaping

By examining further the available inputs of the TVGIC structure of Fig. 2.8 it is noted that  $x_2(n)$  is incident to a node where the outputs of multipliers  $P_1$  and  $P_2$  are also incident. This property can be explored further by noticing that if a filter is realized as a cascade connection of TVGIC sections, the outputs of multipliers  $P_3$ ,  $P_4$  and  $P_5$  of one section are incident to the same node as the outputs of multipliers  $P_1$  and  $P_2$  of the following section. This is a very important property because it leads to a very efficient application of

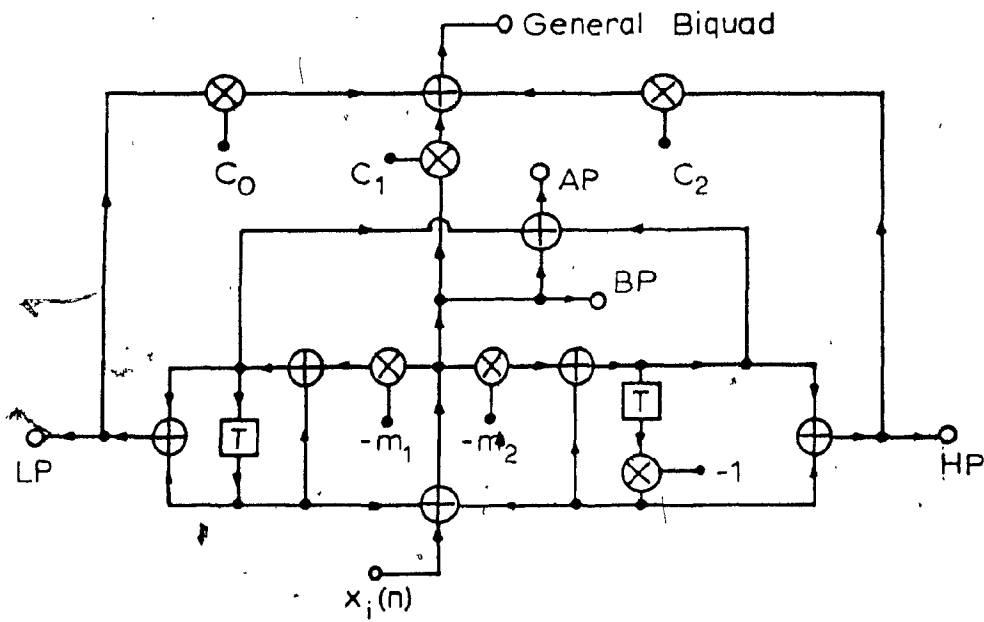


Fig. 2.9: Universal Multiple-Output Biquad

error spectrum shaping (ESS) [4] (see Chapter 3). The TVGIC structure with a second-order ESS can be obtained, as illustrated in Fig. 2.10. Its application leads to a considerable reduction in the roundoff noise, as will be demonstrated in Chapter 6.

## 2.6 Comparison of Computational Complexity

The number of arithmetic operations in a digital structure is an important factor in evaluating its efficiency. Table 2.1 shows the number of arithmetic operations in the various VGIC-based, in the direct canonic and in the section-optimal structures. As can be noted, the difference of the VGIC and TVGIC structures relative to the direct canonic structure is marginal. The TVGIC with ESS is always more economical than the section-optimal structure, because the latter structure utilizes an excessive number of multipliers.

## 2.7 Conclusions

In this chapter new second-order digital filter structures have been obtained, by applying the concept of wave characterization to an active analog-filter configuration comprising resistors and VGIC's.

The TVGIC universal digital biquad realizes simultaneously a lowpass, a bandpass and a highpass transfer function. It can also realize a transfer function with zeros on the unit circle using the minimum number of multipliers. Furthermore, this structure was shown to be amenable to the application of error spectrum shaping.

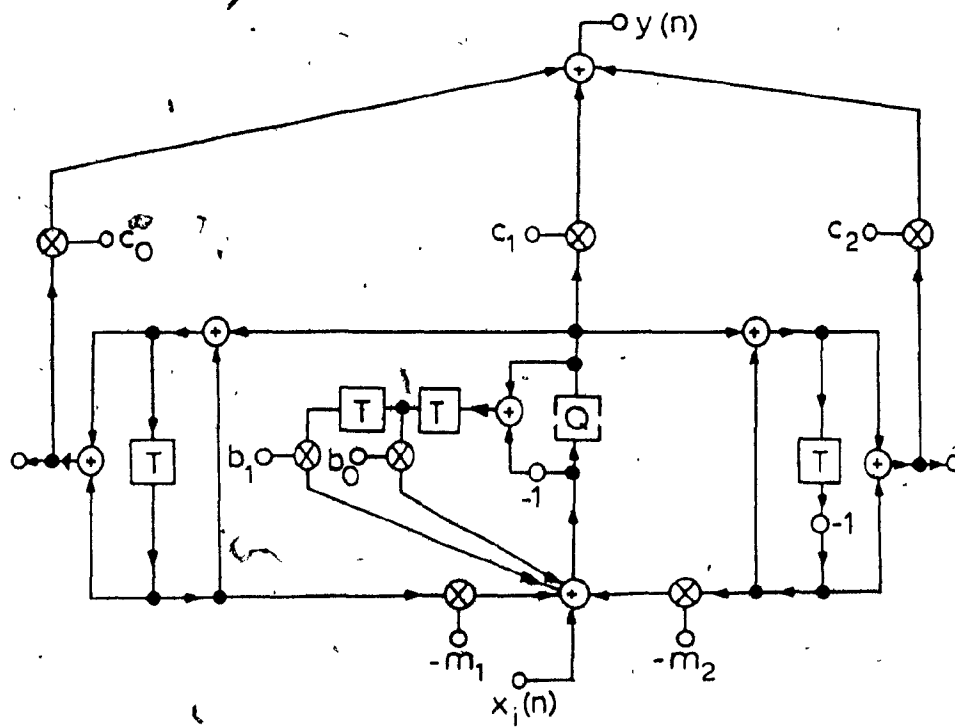


Fig. 2.10: TVGIC Structure with ESS

TABLE 2.1  
Arithmetic Operations

		Arithmetic Operations														
		Lowpass			Bandpass			Highpass			T.F. with zeros on the unit circle			General Biquad		
		+	X	T	+	X	T	+	X	T	+	X	T	+	X	T
VGIC		5	2	2	4	2	2	5	2	2	7	3	2	8	5	2
TVGIC		5	2	2	4	2	2	5	2	2	7	3	2	8	5	2
	1st-order ESS	7	3	3	6	3	3	7	3	3	9	4	3	10	6	3
	2nd-order ESS	8	4	4	7	4	4	8	4	4	10	5	4	11	7	4
Direct Canonic		5	2	2	3	2	2	5	2	2	4	3	2	4	5	2
Section-Optimal		6	9	2	6	9	2	6	9	2	6	9	2	6	9	2

+ = Additions  
X = Multiplications  
T = Unit delays

The VGIC universal digital biquad obtained in Sec. 2.4 by using the bandpass and allpass sections of Figs. 2.6(b) and 2.7(b) realizes simultaneously a lowpass, a highpass, a bandpass and an allpass transfer function, in addition to a general biquadratic transfer function.

The flexibility provided by the TVGIC and VGIC universal digital biquads in terms of the availability of multiple outputs renders these structures attractive for the fabrication of a universal VLSI chip which can be used in a multiplicity of digital-filter applications.

The VGIC and TVGIC structures were compared with the direct canonic and the section-optimal structures from the point of view of computational complexity. It was found that the VGIC and TVGIC structures are comparable to the direct canonic structure except for a small difference in the number of additions. The TVGIC structure with ESS, on the other hand, was shown to be more economical than the low-noise section-optimal structure.

## CHAPTER 3

### LOW-SENSITIVITY STRUCTURES WHICH ARE AMENABLE TO ERROR-SPECTRUM SHAPING

#### 3.1 Introduction

Roundoff noise can be reduced in recursive digital filters by increasing the wordlength [2], by choosing the structure appropriately [16] - [17], or by applying error-spectrum shaping (ESS) [30] - [35].

As was mentioned in Sec. 1.3.4, ESS is a quantization technique which involves the generation of an error signal and the application of local feedback for the purpose of forcing zeros in the power spectral density of the output noise. The technique can be implemented by incorporating a quantizer and a corresponding sub-structure between the output and input of each and every adder whose inputs include at least one nontrivial product, but its application entails an increase in the complexity of hardware.

The application of ESS to the direct form tends to bring about a dramatic reduction in the output roundoff noise [30] - [35]. Unfortunately, however, the sensitivity to coefficient quantization is not affected to any significant extent by ESS and, as is well known, it can be large, in particular if the poles of the transfer function are close to the unit circle of  $z$  plane [19] - [20].

In this chapter, a systematic procedure is described which can be used for the generation of low-sensitivity digital-filter structures which are amenable to ESS. The procedure is then used to



generate a set of structures which includes several new structures as well as some known structures like those of Agarwal and Burrus [19], and Nishimura, Hirano, and Pal [20]. The emphasis is placed on generating economical second-order structures which can be used in cascade or in parallel for the realization of high-order transfer functions.

### 3.2.5 Synthesis Procedure

The application of ESS in a second-order structure is economically attractive only if product quantization can be achieved by using no more than one quantizer. Consequently, a second-order structure is amenable to ESS only if all nontrivial signal-coefficient products in the structure are inputs to only one adder.

A general second-order structure can be constructed as depicted in Fig. 3.1, and if coefficients  $m_i$  are assumed to be noninteger constants and substructure N is assumed to be free of noninteger multipliers, a general second-order structure is obtained which is amenable to ESS. By applying a systematic and exhaustive search to the general structure of Fig. 3.1, all possible low-sensitivity structures can be generated and appropriate design formulas can be deduced.

In this section, the above approach is used for the realization of allpole transfer functions. The realization of biquadratic transfer functions is accomplished through zero-placement techniques and is considered in Section 3.3.

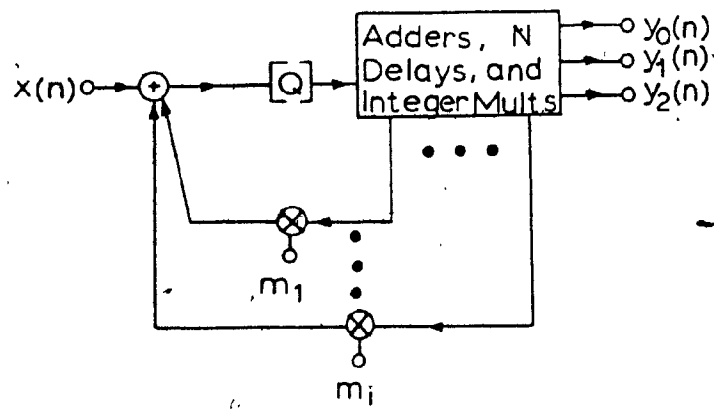


Fig. 3.1: General Structure Suitable for Application of ESS

The simplest structure that can realize any second-order allpole transfer function must have three nodes [39] where each node may represent a distribution or a summation point. Thus, the structure shown in Fig. 3.2 is the simplest special case of the general structure of Fig. 3.1. Branches A, B, C, D and E represent unit delays or integer multipliers whose multiplier constants are restricted to 0,  $\pm 1$ ,  $\pm 2$ . The structure of Fig. 3.2 realizes the following characteristic polynomial

$$\begin{aligned} D(z) &= (1 - BD - AC - m_1 A + ABE + m_2 AB + ABCD + m_1 ABD) z^2 \\ &= z^2 + \alpha_1 z + \alpha_2 \end{aligned} \quad (3.1)$$

In order to avoid delay-free loops [2] and maintain the number of delays to the minimum of two, the following constraints are needed

$$A = z^{-1} \text{ and } B \text{ or } D = z^{-1}. \quad (3.2)$$

Therefore, two cases are possible, as follows:

Case I:  $A = B = z^{-1}$

Case II:  $A = D = z^{-1}$ .

The corresponding structures are illustrated in Figs. 3.3 and 3.4.

#### Case I:

For Case I, the characteristic polynomial of Eqn. 3.1 becomes

$$D(z) = z^2 - z(C + D + m_1) + CD + m_1 D + m_2 + E \quad (3.3)$$

In order to obtain low-sensitivity multipliers C, D and E must be chosen as follows [19]

$$C + D = \text{Int.} [-\alpha_1] \quad (3.4)$$

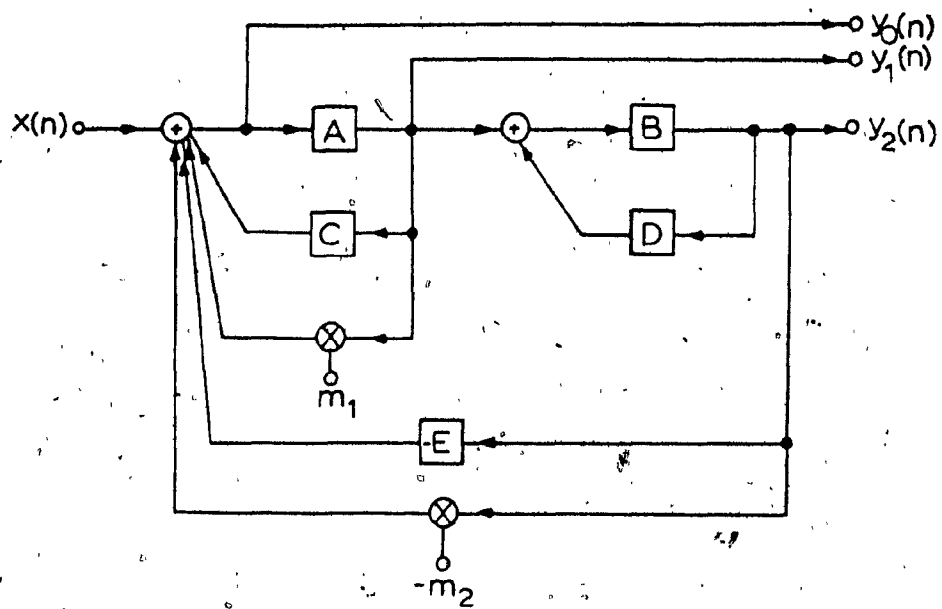


Fig. 3.2: General Second-Order Structure of Low-Complexity

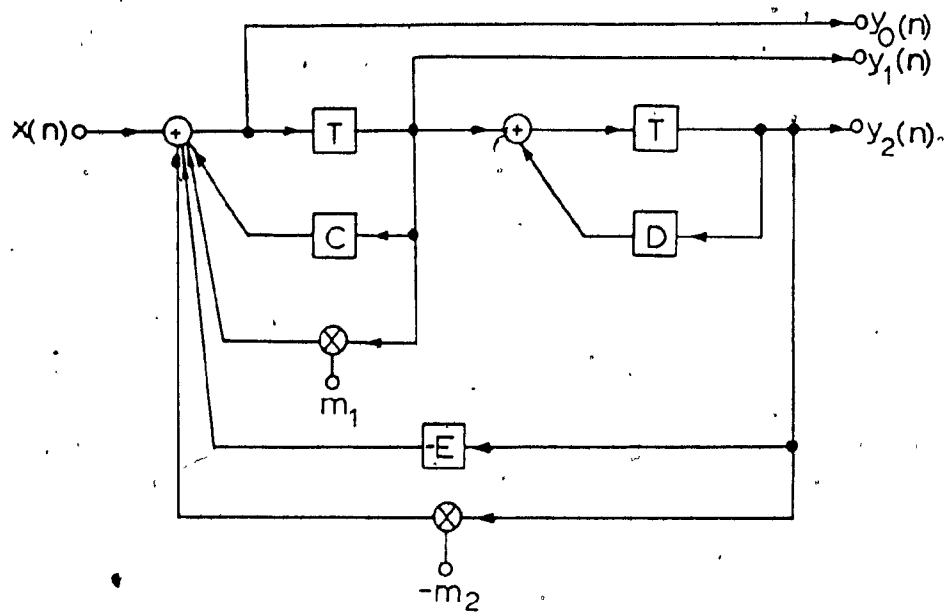


Fig. 3.3: Structure for Case I

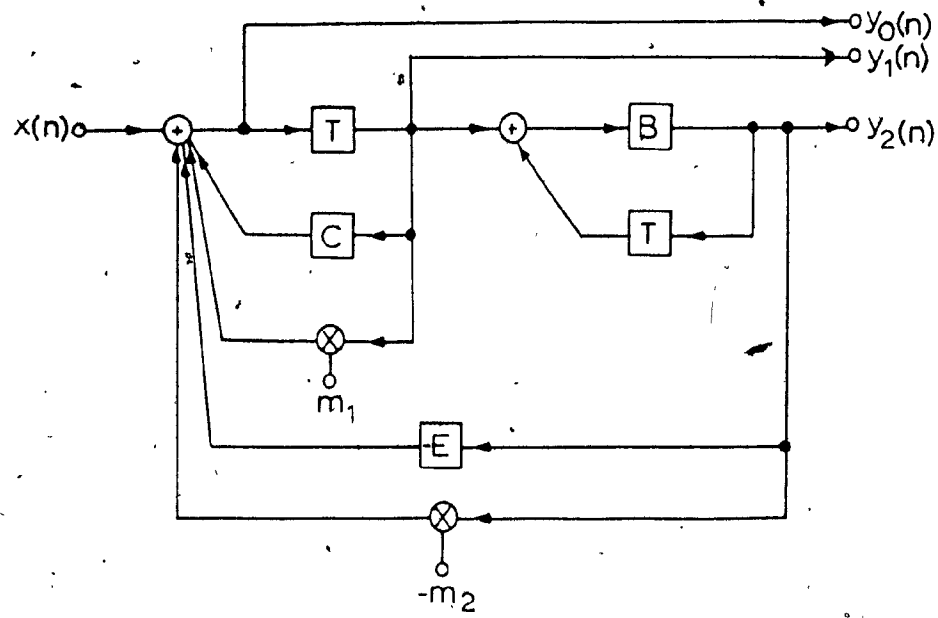


Fig. 3.4: Structure for Case I II

and

$$E = \text{Int. } [\alpha_2 + \alpha_1 D + D^2] \quad (3.5)$$

where  $\text{Int. } [x]$  means the closest integer to  $x$ . Eqns. 3.4 and 3.5 force the values of  $m_1$  and  $m_2$  to be low in order to ensure low sensitivity of the transfer function with respect to changes in  $m_1$  and  $m_2$ .

The choice of structure tends to depend heavily on the pole positions. Therefore, several possibilities must be examined.

If the poles are close to  $z = 1$ , then  $\alpha_1 \approx -2$  and  $\alpha_2 \approx 1$  and so

$$C + D = 2.$$

We can thus assign

$$D = 1, \quad C = 1 \quad \text{and} \quad E = 0.$$

This choice of coefficients yields a structure due to Agarwal and Burrus [19], which will be referred to as structure I-1.

Alternatively, if

$$D = 0, \quad C = 2, \quad \text{and} \quad E = 1$$

another structure due to the same authors [19] is obtained. This structure will be referred to as structure I-2. A third possibility is obtained by letting

$$D = 2, \quad C = 0, \quad \text{and} \quad E = 1.$$

This structure will be referred to as structure I-3, and is to the author's knowledge new.

By considering poles which are close to the unit circle, several other new structures can be obtained for different ranges of values of  $\alpha_1$  as shown in Table 3.1.

TABLE 3.1  
Specific Structures Based on the General  
Structure of Fig. 3.3

STRUCTURE	COEFFICIENTS			RANGE OF $\alpha_1$
	D	C	E	
I-1	1	1	0	$-2.0 < \alpha_1 < -1.5$
I-2	0	2	1	
I-3	2	0	1	$-2.0 < \alpha_1 < -1.75$
I-4	2	0	2	$-1.75 < \alpha_1 < -1.5$
I-5	0	1	1	$-1.5 < \alpha_1 < -0.5$
I-6	1	0	1	
I-7	0	0	1	$-0.5 < \alpha_1 < 0.5$
I-8	1	-1	2	
I-9	-1	1	2	
I-10	0	-1	1	$0.5 < \alpha_1 < 1.5$
I-11	-1	0	1	
I-12	-2	0	2	$1.5 < \alpha_1 < 1.75$
I-13	0	-2	1	$1.5 < \alpha_1 < 2.0$
I-14	-1	-1	0	
I-15	-2	0	1	$1.75 < \alpha_1 < 2.0$



### Case II:

For Case II the characteristic polynomial of Eqn. 3.1 becomes

$$D(z) = z^2 - z(B + C + m_1 - m_2B - BE) + BC + m_1B. \quad (3.6)$$

In order to obtain low sensitivity, the multiplier coefficients B, C and E must be chosen as

$$B = 1, \quad C = 1 \quad (3.7)$$

$$E = \text{Int.} [\alpha_1 + \alpha_2 + 1] \quad (3.8)$$

for poles with positive real part, and

$$B = -1, \quad C = -1 \quad (3.9)$$

$$E = \text{Int.} [-\alpha_1 + \alpha_2 + 1] \quad (3.10)$$

for poles with negative real part.

If the poles are close to  $z = 1$ , we can assign

$$B = 1, \quad C = 1, \quad \text{and } E = 0.$$

This choice of coefficients yields a structure due to Nishimura, Hirano and Pal [20], which will be referred to as structure II-1.

As for Case I, several new structures can be generated by considering poles which are close to the unit circle. The structures obtained are summarized in Table 3.2.

### 3.3 Zero Placement

The general second-order transfer function

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\gamma_0 z^2 + \gamma_1 z + \gamma_2}{z^2 + \alpha_1 z + \alpha_2} \quad (3.11)$$

can be realized by modifying the low-sensitivity sections generated so far as depicted in Fig. 3.5. The design equations required to

TABLE 3.2  
Specific Structures Based on the General  
Structure of Fig. 3.4

STRUCTURE	COEFFICIENTS			RANGE OF $\alpha_1$
	B	C	E	
II-1	1	1	0	$-2.0 < \alpha_1 < -1.5$
II-2	1	1	1	$-1.5 < \alpha_1 < -0.5$
II-3	1	1	2	$-0.5 < \alpha_1 < 0$
II-4	-1	-1	2	$0 < \alpha_1 < 0.5$
II-5	-1	-1	1	$0.5 < \alpha_1 < 1.5$
II-6	-1	-1	0	$1.5 < \alpha_1 < 2.0$

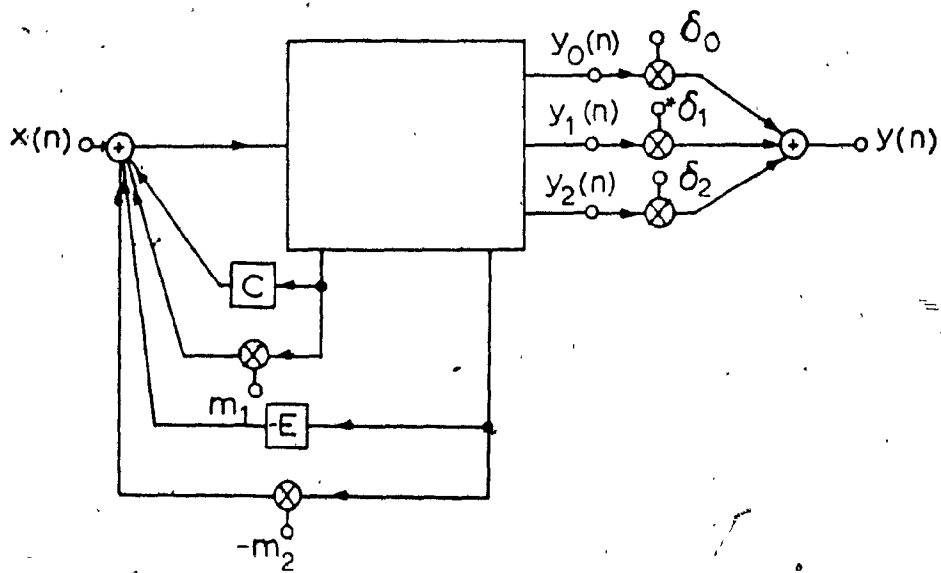


Fig. 3.5: Zero Placement

complete a design can readily be derived as shown in Table 3.3.

An alternative possibility is to take signal  $y_0(n)$  just after the input adder as shown in Fig. 3.6. In this case, the design equations for the feedforward multipliers are given in Table 3.4.

### 3.4 Application of ESS

ESS can be applied in the structures of Figs. 3.2 to 3.6 by including quantizer  $Q$  and an appropriate substructure between node  $y_0(n)$  and the input adder as illustrated in Fig. 3.7. In this way, the power spectral density of the output noise can be properly shaped. In effect, error feedback is applied, which can be adjusted to force zeros in the power spectral density of the output noise, and by choosing coefficients  $b_0$  and  $b_1$  the output noise can be reduced or minimized.

In Chapter 6, the proposed structures are compared with other known high-performance structures with respect to output roundoff noise. The sensitivity performance of these structures is considered in detail in Chapter 7.

### 3.5 Conclusions

A systematic and exhaustive procedure has been described for the generation of low-sensitivity digital-filter structures which are amenable to the application of ESS. The procedure has then been used to generate structures I-1 to I-15 and II-1 to II-6 as summarized in Tables 3.1 and 3.2.

TABLE 3.3  
Equations for the Design of  
Biquadratic Transfer Functions

Case	$\delta_0$	$\delta_1$	$\delta_2$	$m_1$	$m_2$
I	$\gamma_0$	$\gamma_1 + D\gamma_0$	$\gamma_2 + D\gamma_1 + D^2\gamma_0$	$-\alpha_1 - C - D$	$\alpha_2 + \alpha_1 D + D^2 - E$
II	$\gamma_0$	$-\frac{\gamma_2}{B}$	$\gamma_0 + \frac{\gamma_1}{B} + \frac{\gamma_2}{B^2}$	$\frac{\alpha_2}{B} - C$	$1 + \frac{\alpha_1}{B} + \frac{\alpha_2}{B^2} - E$

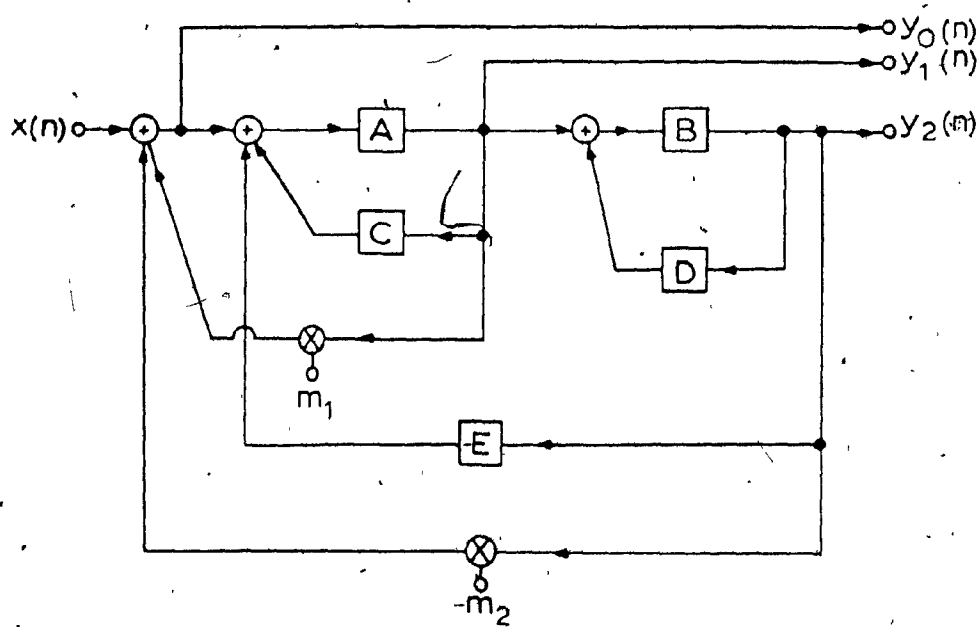


Fig. 3.6: Alternative Zero Placement

TABLE 3.4  
Alternative Equations for the Design of Biquadratic  
Transfer Function

Case	$\delta_0$	$\delta_1$	$\delta_2$
I	$\gamma_0$	$\gamma_1 + \gamma_0 C + \gamma_0 D$	$\gamma_2 + \gamma_1 D - \gamma_0 E + \gamma_0 D^2$
II	$\gamma_0$	$\gamma_0 C - \frac{\gamma_2}{B}$	$\frac{\gamma_2}{B^2} + \frac{\gamma_1}{B} - \gamma_0 E + \gamma_0$

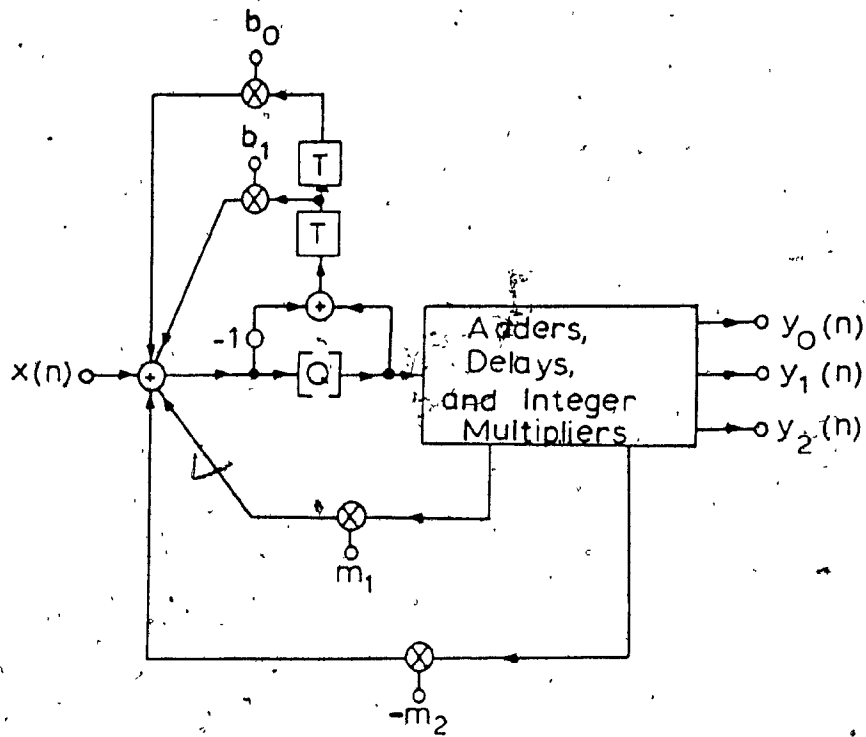


Fig. 3.7: Application of ESS



Structures I-1 and I-2 are due to Agarwal and Burrus [19] and structure II-1 is due to Nishimura, Hirano and Pál [20]. The remaining structures are to the author's knowledge new.

In addition, formulas have been derived which facilitate the realization of biquadratic transfer functions in terms of low-sensitivity structures which are amenable to ESS. The pole positions can be located anywhere in the unit-circle of the  $z$  plane.

It should be mentioned that the structures presented are also suitable for implementation based on ROM-accumulator arithmetic [40] - [41] or on stored-product arithmetic [42]. In addition, the internal scaling strategies applied in [19], [20] and [41] can also be applied to the proposed structures, as an alternative to the use of ESS.

## CHAPTER 4 ELIMINATION OF LIMIT CYCLES

### 4.1 Introduction

In this chapter, methods are explored for the elimination of zero-input and overflow limit cycles in VGIC and TVGIC structures. The approach used entails a suitable quantization scheme based on magnitude truncation. A positive-definite pseudo-energy function is defined and is then shown to remain a Liapunov function [6] - [12] upon the application of finite-precision arithmetic. In this way, it is demonstrated that zero-input and overflow limit cycles can be eliminated in all the VGIC and TVGIC structures of Chapter 2.

This chapter also deals with the elimination of constant-input limit cycles, which can often occur in certain applications. Specifically, a theorem is proved which establishes sufficient conditions that will ensure freedom from constant-input limit cycles in a general digital-filter structure in which zero-input limit cycles can be eliminated [43]. Direct application of this theorem shows that all types of known limit cycles can be eliminated in some VGIC sections as well as the TVGIC and VGIC universal multiple-output biquads of Chapter 2.

### 4.2 Stability of VGIC Structures under Infinite-Precision Arithmetic

The characteristic polynomial in the VGIC and TVGIC structures of Chapter 2 is given by

$$D(z) = z^2 + (m_1 - m_2)z + m_1 + m_2 - 1 \quad (4.1)$$

Under infinite-precision arithmetic, the range of values that the multiplier coefficients  $m_1$  and  $m_2$  can assume is constrained by the inequalities

$$m_1 > 0, \quad m_2 > 0, \quad \text{and} \quad m_1 + m_2 < 2 \quad (4.2)$$

if stability is to be assured. The permissible region of the  $m_2$  versus  $m_1$  plane is the shaded region in Fig. 4.1.

#### 4.3 Stability of VGIC Structures under Finite-Precision Arithmetic

Consider the second-order structure of Fig. 4.2(a). This structure is the recursive part of all the VGIC structures under zero-input conditions. Quantizers  $Q$  are used for the elimination of limit cycles. The same structure can be interpreted as a wave digital network comprising a series two-port adaptor terminated by a unit-delay at the left-hand port and a unit-delay in series with an inverter at the right-hand port as depicted in Fig. 4.2(b). It is the digital implementation of a closed loop containing a series capacitor and a series inductor.

The state difference equation of the structure of Fig. 4.2(a) is given by

$$\tilde{x}'(k) = \begin{bmatrix} x_1'(k+1) \\ x_2'(k+1) \end{bmatrix} = \tilde{A} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} \quad (4.3)$$

where

$$\tilde{A} = \begin{bmatrix} 1 - m_1 & m_1 \\ -m_2 & m_2 - 1 \end{bmatrix}$$

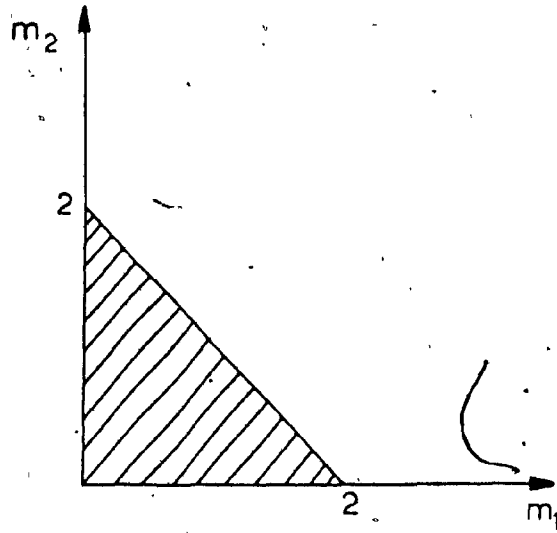


Fig. 4.1:  $m_2$  versus  $m_1$  Plane (Shaded Region  
Indicates Range of Permissible Values  
for  $m_1$  and  $m_2$ )

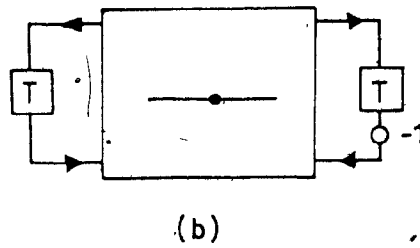
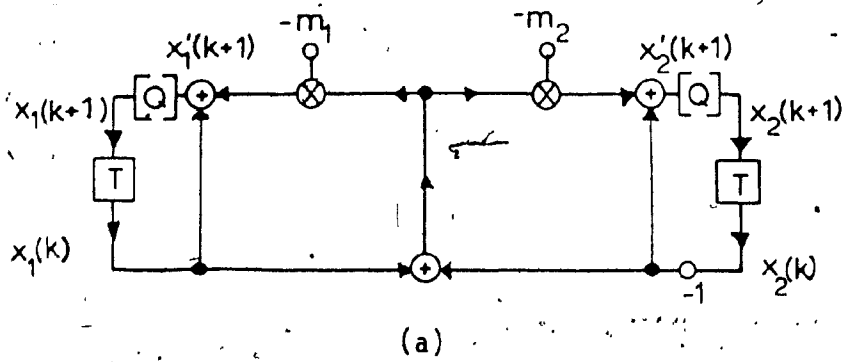


Fig. 4.2: (a) VGIC Recursive Structure  
(b) Corresponding Block Diagram

and with quantization applied to  $\tilde{x}'(k)$ , we obtain

$$\tilde{x}(k+1) = [\tilde{x}'(k+1)]_Q = [A \tilde{x}(k)]_Q \quad (4.4)$$

A positive-definite quadratic function (or pseudo-energy function) can be defined as

$$p(\tilde{x}(k)) = \tilde{x}^T(k) G \tilde{x}(k) = \frac{x_1^2(k)}{m_1} + \frac{x_2^2(k)}{m_2} \quad (4.5)$$

where

$$G = \begin{bmatrix} 1/m_1 & 0 \\ 0 & 1/m_2 \end{bmatrix}$$

and from Eqn. 4.2,  $m_1, m_2 > 0$ . From Eqn. 4.5,

$$\begin{aligned} \Delta p_0(k+1) &= p(\tilde{x}'(k+1)) - p(\tilde{x}(k)) \\ &= \tilde{x}'^T(k+1) G \tilde{x}'(k+1) - \tilde{x}^T(k) G \tilde{x}(k) \\ &= \tilde{x}^T(k) (A^T G A - G) \tilde{x}(k) \\ &= (m_1 + m_2 - 2) (x_1(k) - x_2(k))^2. \end{aligned} \quad (4.6)$$

In order to guarantee stability under infinite-precision arithmetic,  $m_1 + m_2 < 2$  and hence we conclude that

$$\Delta p_0(k+1) < 0 \quad \text{if } x_1(k) \neq x_2(k) \quad (4.7a)$$

and

$$\Delta p_0(k+1) = 0 \quad \text{if } x_1(k) = x_2(k). \quad (4.7b)$$

In the case where  $x_1(k) = x_2(k)$ , Eqn. 4.3 yields

$$x_1'(k+1) = x_1(k), \quad x_2'(k+1) = -x_2(k) \quad (4.8)$$

that is,

$$x_1'(k+1) \neq x_2'(k+1)$$

and, therefore, from Eqn. 4.7(a)

$$\Delta p_0(k+2) < 0.$$

In effect, the condition

$$\Delta p_0(k+1) = 0$$

can be satisfied in more than one cycle if and only if

$$x_1(k) = x_2(k) = 0.$$

Now if magnitude truncation is applied to quantize the state variables, then  $p(\underline{x}(k)) < p(\underline{x}'(k))$  which implies that

$$\Delta p(\underline{x}(k)) = p(\underline{x}(k+1)) - p(\underline{x}(k)) < 0.$$

We thus conclude that  $p(\underline{x}(k))$  is a Liapunov function [6].

If no quantization is applied in the structure of Fig. 4.2(a), oscillations cannot occur if the structure is assumed to satisfy the stability constraints of Eqn. 4.2. If quantization is applied as shown in Fig. 4.2(a) then limit cycles oscillation would occur when  $|x_1(k)| < |x_1'(k)|$ . Under these circumstances the Liapunov function would decrease during the following cycles by a positive amount and eventually it would become negative. However, this contradicts the fact that  $p(\underline{x}(k))$  is always greater than or equal to zero and, therefore, we conclude that

$$\underline{x}^T(k) = [0 \quad 0]$$

is the only equilibrium point possible. Consequently, the quantization scheme of Fig. 4.2(a) can be used to eliminate zero-input limit cycles in all the VGIC structures of Chapter 2.

Claasen, Mecklenbraüker and Peek [13] showed that in a digital filter in which the condition  $|x_1(k)| < |x_1'(k)|$  is sufficient to guarantee zero-input stability, then the forced response stability to overflow oscillations can also be guaranteed if the quantized signals

are restricted to the shaded regions shown in Fig. 4.3. By incorporating the nonlinearity implied by Fig. 4.3 in the quantizers used, overflow limit-cycles can also be eliminated in the VGIC structures of Chapter 2.

The results discussed so far hold for any type of finite-precision arithmetic. However, by using fixed-point two's complement arithmetic some advantages can be achieved in the implementation of  $|x_i(k)| < |x_i^1(k)|$ , such as simple rear and front chopping operations [7], [9].

The TVGIC structure of Fig. 2.8 under zero-input conditions can be described by the state-space difference equation

$$\begin{bmatrix} \bar{x}_1(k+1) \\ \bar{x}_2(k+1) \end{bmatrix} = \tilde{A} \begin{bmatrix} \bar{x}_1(k) \\ \bar{x}_2(k) \end{bmatrix} \quad (4.9)$$

where

$$\tilde{A} = \begin{bmatrix} I-m_1 & m_2 \\ -m_1 & m_2-1 \end{bmatrix}.$$

By applying the same form of quantization as for the VGIC structure of Fig. 4.2(a) and then following the procedure outlined above, one can prove that zero-input and overflow limit cycles can also be eliminated in the TVGIC structure.

#### 4.4 Elimination of Constant-input Limit Cycles

##### Theorem 4.1:

Assume that the digital-filter structure of Fig. 4.4 is free of zero-input limit cycles and that



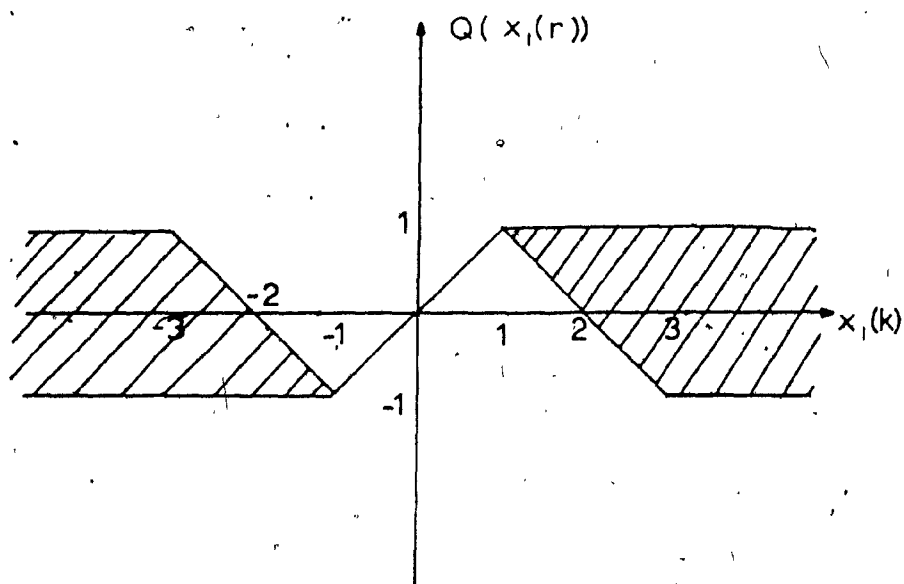


Fig. 4.3: Elimination of Overflow Limit Cycles /  
(Shaded Regions are the Permissible  
Regions for the Overflow Nonlinearities)

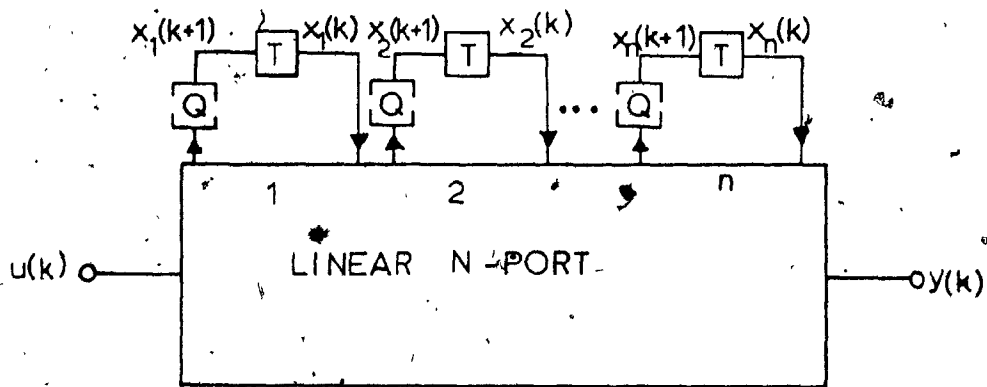


Fig. 4.4:  $n$ -th Order Digital-Filter Structure

$$\begin{aligned}\underline{\tilde{x}}(k+1) &= [\underline{\tilde{A}}\underline{\tilde{x}}(k) + \underline{\tilde{B}}u(k)]_Q \\ y(k+1) &= \underline{\tilde{C}}\underline{\tilde{x}}(k) + \underline{\tilde{D}}u(k)\end{aligned}\quad (4.10)$$

where  $[\bullet]_Q$  is the quantized value of  $[\bullet]$ .

Constant-input limit cycles can be eliminated by modifying the structure of Fig. 4.4 as depicted in Fig. 4.5, where  $\underline{\tilde{P}}$  is given by

$$\underline{\tilde{P}} = [p_1 \ p_2 \ \dots \ p_n]^T = (\underline{\tilde{I}} - \underline{\tilde{A}})^{-1} \underline{\tilde{B}} \quad (4.11)$$

and is machine representable.

#### Proof

Since the structure of Fig. 4.4 is assumed to be free from zero-input limit cycles, the equation

$$\underline{\tilde{x}}(k+1) = [\underline{\tilde{A}}\underline{\tilde{x}}(k)]_Q \quad (4.12)$$

describes a stable system such that

$$\lim_{k \rightarrow \infty} \underline{\tilde{x}}(k) = [0 \ 0 \ \dots \ 0]^T.$$

If the input is constant, i.e.  $u(k) = u_0$ , the modified structure of Fig. 4.5 is characterized by

$$\underline{\tilde{x}}(k+1) = [\underline{\tilde{A}}\underline{\tilde{x}}(k) - \underline{\tilde{P}}u_0 + \underline{\tilde{B}}u_0]_Q + \underline{\tilde{P}}u_0$$

and if Eqn. 4.11 holds, we have

$$\begin{aligned}\underline{\tilde{x}}(k+1) &= [\underline{\tilde{A}}\underline{\tilde{x}}(k) - \underline{\tilde{I}}(\underline{\tilde{I}} - \underline{\tilde{A}})^{-1}\underline{\tilde{B}}u_0 + (\underline{\tilde{I}} - \underline{\tilde{A}})(\underline{\tilde{I}} - \underline{\tilde{A}})^{-1}\underline{\tilde{B}}u_0]_Q + \underline{\tilde{P}}u_0 \\ &= [\underline{\tilde{A}}\{\underline{\tilde{x}}(k) - \underline{\tilde{P}}u_0\}]_Q + \underline{\tilde{P}}u_0.\end{aligned}$$

Hence

$$\underline{\hat{x}}(k+1) = [\underline{\tilde{A}}\underline{\hat{x}}(k)]_Q \quad (4.13)$$

where

$$\underline{\hat{x}}(k) = \underline{\tilde{x}}(k) - \underline{\tilde{P}}u_0.$$

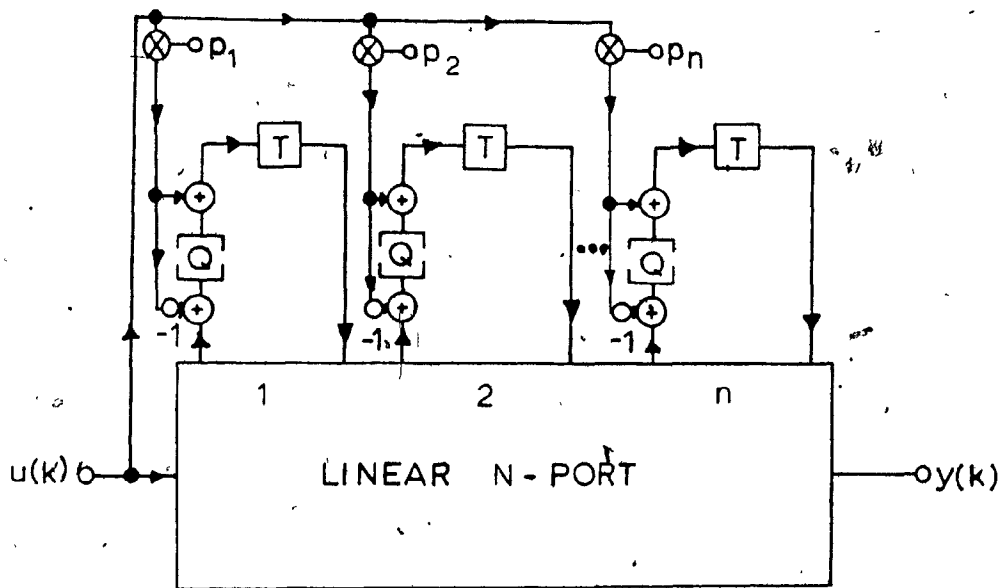


Fig. 4.5: Modified n-th Order Digital-Filter Structure

Evidently, Eqn. 4.13 is the same as Eqn. 4.12, except for the transformation in the state variables and, therefore, it represents a stable system. Stability can be guaranteed if Eqn. 4.11 is satisfied exactly and, therefore,  $\underline{P}$  is required to be machine representable (QED).

It should be mentioned that if quantization in Fig. 4.5 is carried out by means of magnitude truncation, an implementation of the controlled-rounding arithmetic proposed by Butterweck [4] is achieved.

An apparent limitation of the above stabilization technique is imposed by the requirement that  $\underline{P}$  be machine representable. Nevertheless, in many second-order structures as well as higher-order wave structures this requirement is easily satisfied.

#### 4.5 Elimination of Constant-Input Limit Cycles in VGIC and TVGIC Structures

In this section Theorem 4.1 is applied to a subclass of the VGIC structures and to the TVGIC structure of Fig. 2.8.

The VGIC bandpass structure of Fig. 2.6(b) and the improved allpass structure of Fig. 2.7(b) with a constant input  $x_1(k) = u_0$  can be described by

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \underline{A} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} -m_1 \\ -m_2 \end{bmatrix} u_0 \quad (4.14)$$

where

$$\tilde{A} = \begin{bmatrix} 1-m_1 & m_1 \\ -m_2 & m_2-1 \end{bmatrix}.$$

From Eqns. 4.11 and 4.14

$$\tilde{P} = \begin{bmatrix} m_1 & -m_1 \\ m_2 & 2-m_2 \end{bmatrix}^{-1} \begin{bmatrix} -m_1 \\ -m_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad (4.15)$$

and since zero-input limit cycles can be eliminated, and  $\tilde{P}$  is machine representable, constant-input limit cycles can also be eliminated in these structures according to Theorem 4.1. Fig. 4.6 illustrates the application of Theorem 4.1 in these structures.

It should be mentioned that any second-order section which has the structure of Fig. 4.6 as its recursive part is free of limit cycles. Two such examples are the general second-order sections described by Ver Kroost and Butterweck [10], and by Ver Kroost [11].

The universal TVGIC digital biquad of Fig. 2.8 with a constant input  $x_1(n) = u_0$ , can be described by

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \tilde{A} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} -m_1 \\ -m_1 \end{bmatrix} u_0 \quad (4.16)$$

where

$$\tilde{A} = \begin{bmatrix} 1-m_1 & m_2 \\ -m_1 & m_2-1 \end{bmatrix}.$$

From Eqns. 4.2 and 4.16

$$\tilde{P} = \begin{bmatrix} m_1 & -m_2 \\ m_1 & 2-m_2 \end{bmatrix}^{-1} \begin{bmatrix} -m_1 \\ -m_1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad (4.17)$$

and since zero-input limit cycles can be eliminated and  $\tilde{P}$  is machine representable, constant-input limit cycles can also be eliminated in

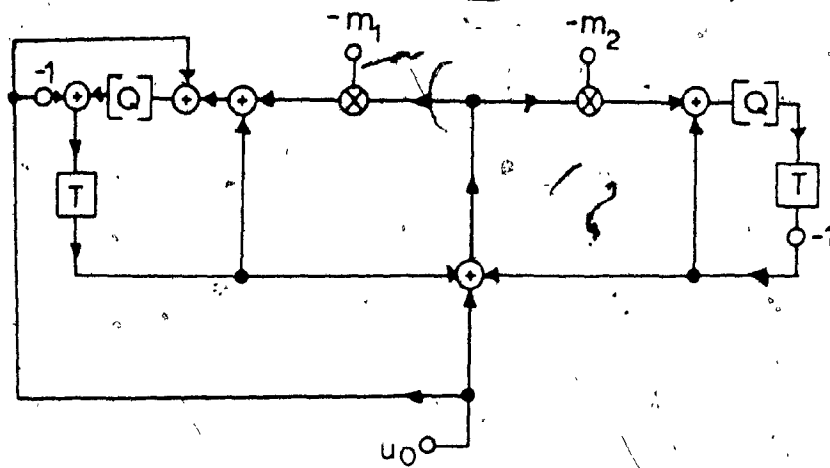


Fig. 4.6: Elimination of Constant-Input Limit Cycles in the VGIC Structures of Figs. 2.6(b), 2.7(b) and 2.9

this structure according to Theorem 4.1. Fig. 4.7 illustrates the application of Theorem 4.1.

#### 4.6 Conclusions

It has been shown that both zero-input and overflow limit cycles can readily be eliminated in the VGIC and TVGIC structures generated in Chapter 2, by utilizing an appropriate quantization scheme.

A theorem has been proved which establishes sufficient conditions that will ensure freedom from constant-input limit cycles in a general digital filter structure in which zero-input limit cycles can be eliminated.

Upon application of the aforementioned theorem, it was found that constant-input limit cycles can also be eliminated in a subclass of the VGIC second-order sections as well as the VGIC and TVGIC universal digital biquads of Chapter 2, that is, all known types of limit cycles can be eliminated in these structures.



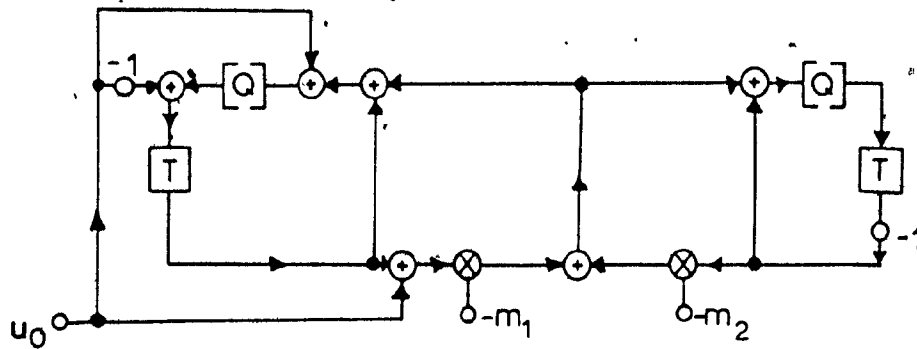


Fig. 4.7: Elimination of Constant-Input Limit Cycles in the TVGIC Structure

## CHAPTER 5

### NEW IMPROVED STATE-SPACE STRUCTURES

#### 5.1 .Introduction

This chapter presents an alternative procedure for the realization of second-order digital-filter structures. The procedure is based on the state-space characterization and leads to structures with several important advantages such as reduced number of multipliers, and elimination of overflow oscillations and granularity limit-cycles both under zero-input as well as under constant-input conditions.

In Sec. 5.2, the conditions that lead to minimum roundoff noise and to the elimination of zero-input limit cycles and overflow oscillations in a particular class of state-space structures are reviewed. Then in Sec. 5.3, Theorem 4.1, which establishes sufficient conditions for the elimination of constant-input limit cycles, is applied to a general state-space second-order structure. The conditions for the elimination of constant-input limit cycles are thus established. These conditions lead to three new state-space structures. The design aspects of the three new structures are considered in detail in Sec. 5.4. It is shown that under certain circumstances two of the new structures lead to optimal designs with respect to roundoff noise. The application of one of the new structures for the design of high-order parallel and cascade filters is considered in Sec. 5.5.

The chapter concludes with a comparison of the computational complexity of one of the new structures relative to that in the

section-optimal structure described in [21]. This comparison demonstrates the new structure to be more economical with respect to the number of multipliers required.

The roundoff noise and sensitivity properties of the new structures are discussed in Chapter 6 and 7, respectively.

## 5.2 Elimination of Zero-Input Limit Cycles

Given a second-order transfer function,

$$H(z) = \frac{\gamma_0 z^2 + \gamma_1 z + \gamma_2}{z^2 + \alpha_1 z + \alpha_2} = d + \frac{\beta_1 z + \beta_2}{z^2 + \alpha_1 z + \alpha_2} = d + H'(z) \quad (5.1)$$

a digital-filter structure can be obtained which is characterized by the state equations

$$\left. \begin{aligned} \underline{x}(k+1) &= \underline{A} \underline{x}(k) + \underline{B} u(k) \\ y(k) &= \underline{C} \underline{x}(k) + \underline{D} u(k) \end{aligned} \right\} \quad (5.2)$$

where  $\underline{A}$ ,  $\underline{B}$ ,  $\underline{C}$  and  $\underline{D}$  are given by

$$\underline{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \underline{B} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

$$\underline{C} = [c_1, c_2], \quad \underline{D} = [d].$$

The sufficient conditions that ensure minimum roundoff noise in a second-order state-space structure have been established in [21] and are given by

$$a_{11} = a_{22} \quad (5.3a)$$

$$\frac{b_1}{b_2} = \frac{c_2}{c_1} \quad (5.3b)$$

The state-space structure in which these conditions are satisfied is referred to as the section-optimal structure. Its transition matrix  $\underline{A}$  is of the form [21], [26]

$$\underline{A} = \begin{bmatrix} a & -\frac{e}{\alpha} \\ e\alpha & a \end{bmatrix} \quad (5.4)$$

where  $a$ ,  $e$ , and  $\alpha$  are constants.

It is well known [8] that zero-input limit cycles can be eliminated in a recursive digital filter if there exists a positive-definite diagonal matrix  $\underline{G}$  such that  $\underline{G} - \underline{A}^T \underline{G} \underline{A}$  is positive definite. This condition is satisfied if [28]

$$\left. \begin{array}{l} a_{12} a_{21} > 0 \text{ or} \\ \text{if } a_{12} a_{21} < 0 \text{ and } |a_{11} - a_{22}| + \det(\underline{A}) < 1. \end{array} \right\} \quad (5.5)$$

Matrix  $\underline{A}$  given by Eqn. 5.4 satisfies these conditions and, in effect, zero-input limit cycles can be eliminated in the section-optimal structure.

Zero-input limit cycles can be eliminated by quantizing the state variables such that [8]

$$|x_i(k)|_Q < |x_i(k)| \quad \forall k \quad (5.6)$$

where  $[\bullet]_Q$  is the quantized value of  $[\bullet]$ . Overflow oscillations can also be eliminated as in wave digital filters, by using the approach described in [13].

### 5.3 Elimination of Constant-Input Limit Cycles

In a structure in which zero-input limit cycles can be eliminated, according to Theorem 4.1 constant-input limit cycles can also be eliminated if the vector  $\underline{p}$  of Eqn. 4.11 is machine representable. For a second-order structure the following forms of  $\underline{p}$  are possible:

Case I:  $\underline{p} = [+1 \ 0]^T$

Case II:  $\underline{p} = [0 \ +1]^T$

Case III:  $\underline{p} = [\pm 1 \pm 1]^T$ .

In order to force  $\underline{p}$  to assume any one of these forms, vector  $\underline{b}$  should be chosen as

Case I:  $b_1 = \pm (1 - a_{11})$  and  $b_2 = \mp a_{21}$

Case II:  $b_1 = \mp a_{12}$  and  $b_2 = \pm (1 - a_{22})$

Case III:  $b_1 = \pm (1 - a_{11}) \mp a_{12}$  and  $b_2 = \mp a_{21} \pm (1 - a_{22})$ .

The structure for case I is shown in Fig. 5.1. As can be seen, coefficients  $b_1$  and  $b_2$  can be formed without the need of multipliers and thus the structure requires fewer multipliers relative to the section-optimal structure. The structure for case II is depicted in Fig. 5.2. As for case I, no multipliers are needed to form coefficients  $b_1$  and  $b_2$ . The structure for case III is depicted in Fig. 5.3. Although no multipliers are needed to form coefficients  $b_1$  and  $b_2$ , five extra additions are required relative to the number of additions in the structures for cases I and II. This structure will not be considered further because of its higher computational complexity.

#### 5.4 Design Considerations

If  $\underline{A}$  is of the form given in Eqn. 5.4, the transfer functions  $F_i(z)$  from the input node  $x(k)$  to the state-variable nodes  $x_i(k+1)$  are given by

$$\left. \begin{aligned} F_1(z) &= \frac{(1-a)z + (e^2 - a + a^2)}{z^2 - 2az + a^2 + e^2} \\ F_2(z) &= \alpha \frac{-ez + e}{z^2 - 2az + a^2 + e^2} = \alpha F_2'(z) \end{aligned} \right\} \quad (5.7)$$

for case I and by

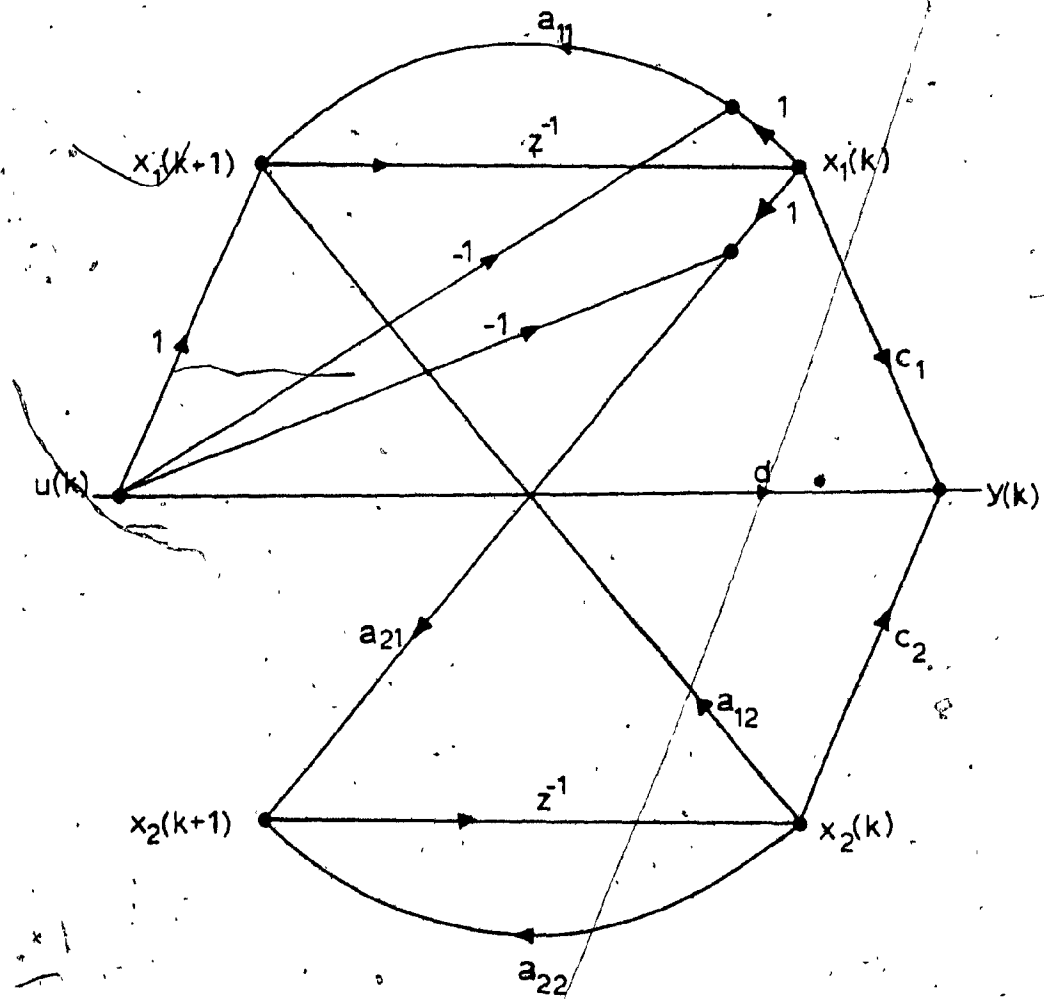


Fig. 5.1: Structure for Case I ( $b_1 \neq 1 - a_{11}$ ,  $b_2 = -a_{21}$ )

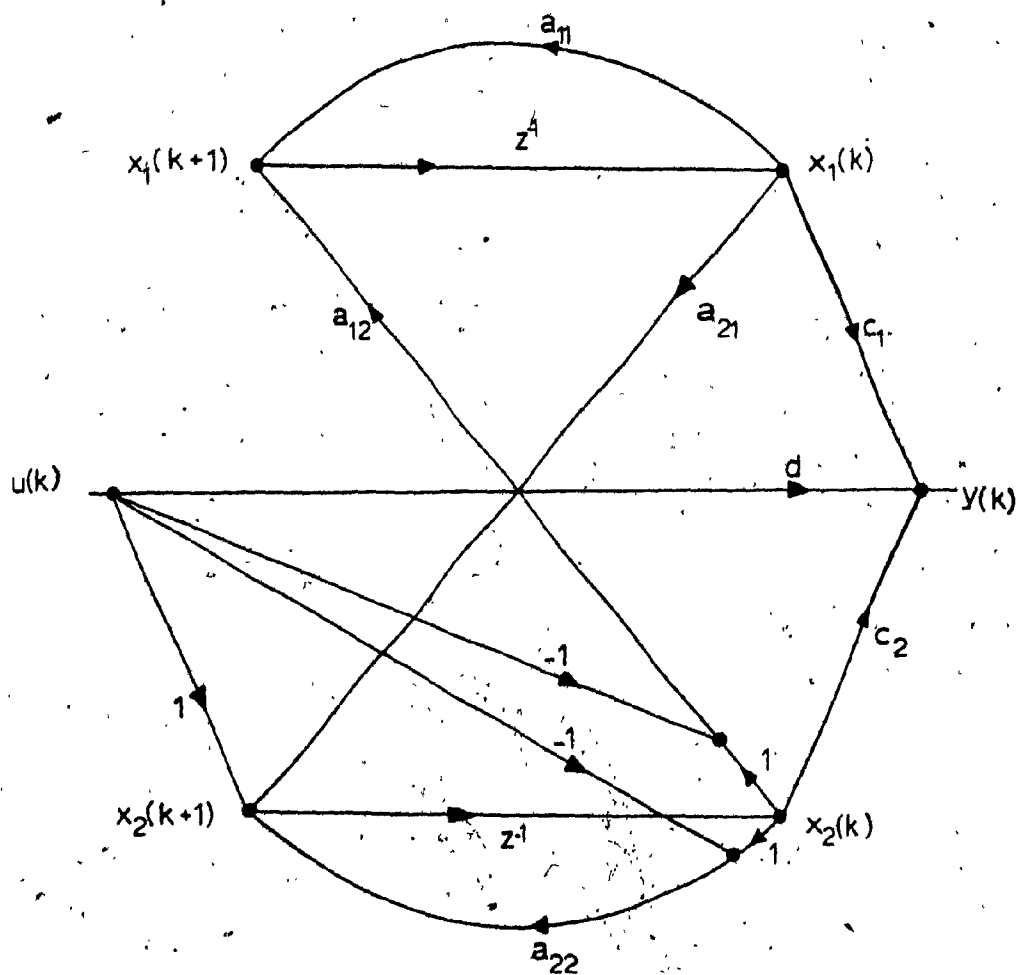


Fig. 5.2: Structure for Case II ( $b_1 = -a_{12}$ ,  $b_2 = 1 - a_{22}$ )

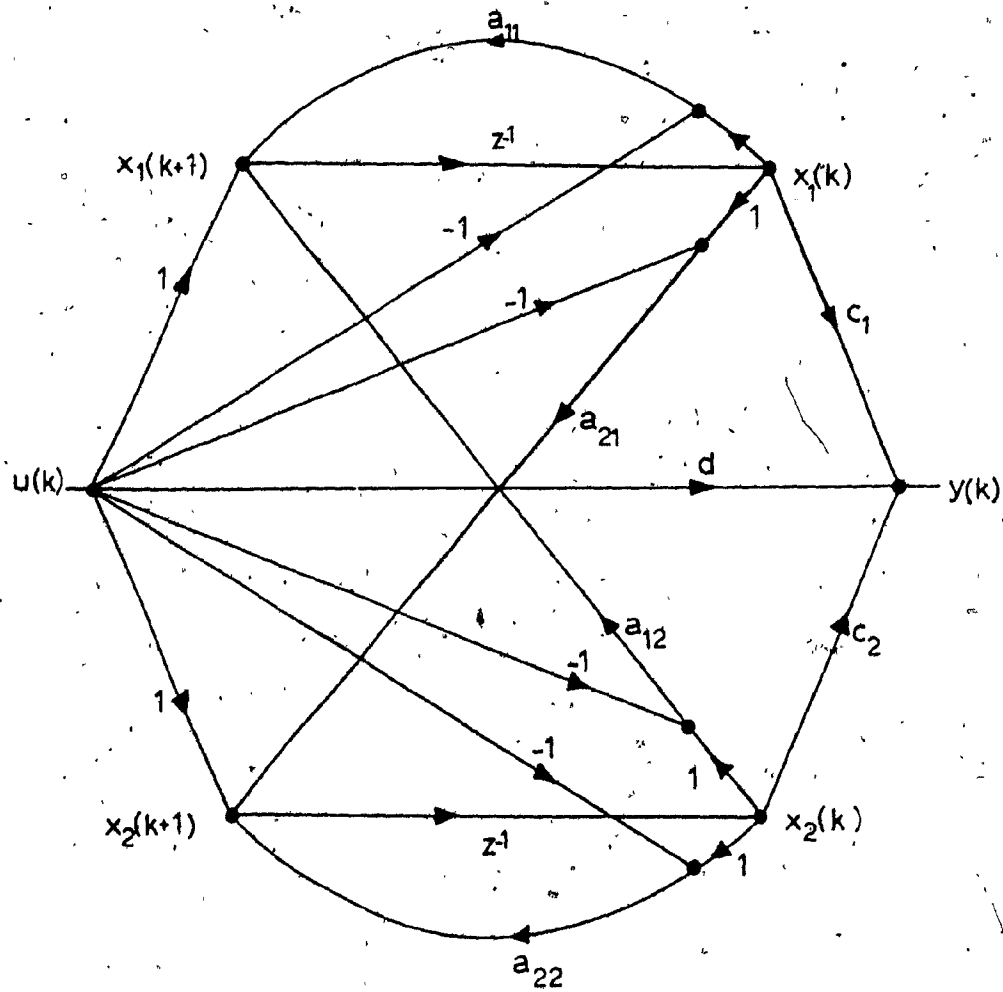


Fig. 5.3: Structure for Case III ( $b_1=1-a_{11}-a_{12}$ ,  
 $b_2=1-a_{22}-a_{21}$ )



$$\left. \begin{aligned} F_1(z) &= \frac{1}{\alpha} \frac{ez - e}{z^2 - 2az + a^2 + e^2} = \alpha F_1'(z) \\ F_2(z) &= \frac{(1-a)z + (e^2 - a + a^2)}{z^2 - 2az + a^2 + e^2} \end{aligned} \right\} \quad (5.8)$$

for case II. Parameter  $\alpha$  is chosen such that the Lp norms of transfer functions  $F_1(z)$  and  $F_2(z)$  are equal. This choice of  $\alpha$  leads to a distributed dynamic range of the state variables  $x_i(k)$ , and also to a simple scaling procedure. For cases I and II,  $\alpha$  is given by

$$\alpha = \frac{\|F_1(z)\|_p}{\|F_2'(z)\|_p} \quad (5.9)$$

and

$$\alpha = \frac{\|F_1'(z)\|_p}{\|F_2(z)\|_p} \quad (5.10)$$

respectively. By using Eqns. 5.7 to 5.10, it can easily be shown that  $\alpha$  for case I is the inverse of  $\alpha$  for case II and, as a consequence, the two structures turn out to be equivalent. From now on only the structure for case I is discussed.

The elements of vector  $\underline{c}$ , namely  $c_1$  and  $c_2$  are given by

$$c_1 = \frac{\beta_1 + \beta_2}{1 + \alpha_1 + \alpha_2} \quad (5.11)$$

$$c_2 = \frac{\left(\frac{-\alpha_1}{2} - \alpha_2\right)\beta_1 + \left(1 + \frac{\alpha_1}{2}\right)\beta_2}{(1 + \alpha_1 + \alpha_2)e^\alpha} \quad (5.12)$$

Although structures obtained by the method presented here are not optimum with respect to roundoff noise [21], one of the conditions

for optimality, i.e.  $a_{11} = a_{22}$ , is satisfied. In addition, the distributed dynamic range that is also a property of the scaled section-optimal structure is preserved. The optimal normal design, i.e.  $\alpha=1$ , proposed in [27] also has distributed dynamic range for  $L_2$  scaling.

An important feature of the structure of Fig. 5.1 can, at this point, be identified through the following theorem.

Theorem 5.1

The new state space structure is optimal with respect to roundoff noise if the zeros of the transfer function  $H(z)$  are located at  $z=1$ .

Proof:

The first condition for optimality, namely Eqn. 5.3(a) is always satisfied since  $A$  is of the form given in Eqn. 5.4 in the new state-space structure.

The second condition, namely Eqn. 5.3(b), can be shown to hold as follows. From the values assigned to  $b_1$  and  $b_2$  for Case I, and Eqns. 5.11 and 5.12, we have

$$\frac{b_1}{b_2} = \frac{1 + \alpha_1/2}{-e^\alpha} \quad (5.13)$$

$$\frac{c_2}{c_1} = \frac{(-\frac{\alpha_1}{2} - \alpha_2)\beta_1 + (1 + \frac{\alpha_1}{2})\beta_2}{(\beta_1 + \beta_2)e^\alpha} \quad (5.14)$$

If the zeros of  $H(z)$  are located at  $z=1$ , then from Eqn. 5.1 it can easily be shown that

$$\beta_1 = -\gamma_0(2+\alpha_1) \quad (5.15)$$

and

$$\beta_2 = \gamma_0(1-\alpha_2) \quad (5.16)$$

Now from Eqns. 5.13 to 5.16

$$\frac{b_1}{b_2} = \frac{c_2}{c_1}$$

and, therefore, Eqn. 5.3 holds (CQD).

This theorem indicates that the design of Butterworth and Chebyshev highpass filters by means of parallel sections of the type shown in Fig. 5.1 results in a structure which is optimal with respect to roundoff noise.

The fact that an optimal structure has been obtained using fewer multipliers relative to the section-optimal structure, leads us to believe that economical solutions exist when the zeros are located at  $z=-1$  or  $z=+1$ . However, if a solution exists it may not be possible to eliminate constant-input limit cycles.

## 5.5 Design of High-Order Filters

The design of high-order parallel or cascade filters using the structure of Fig. 5.1 can be accomplished as follows:

### Parallel Design:

- 1) Express the transfer function of the filter as

$$T(z) = \sum_{i=1}^m H_i^1(z)$$

where each  $H_i^1(z)$  is of the form given by Eqn. 5.1. Then for each

$H_i^1(z)$  compute

$$a_i = -\alpha_1/2 \quad (5.17)$$

and

$$e = \sqrt{\alpha_2^2 - \alpha_1^2/4}. \quad (5.18)$$

If the  $L_2$  norm is used for scaling,  $\alpha$  is given by

$$\alpha = \sqrt{\frac{(1+\alpha_1/2)^2[(1+\alpha_2)(1+u^2)-2\alpha_1 u]}{2e^2(1+\alpha_1+\alpha_2)}} \quad (5.19)$$

where

$$u = \frac{\alpha_2 + \alpha_1/2}{1 + \alpha_1/2}.$$

- 2) Compute the scaling parameter  $\lambda$ , and  $a_{12}$  as well as  $a_{21}$  as follows

$$\lambda = \frac{1}{\|F_2(z)\|_2} = \sqrt{\frac{(1-\alpha_1+\alpha_2)(1-\alpha_2)}{2e^2\alpha^2}}, \quad (5.20)$$

$$a_{12} = -e/\alpha, \quad (5.21)$$

$$a_{21} = e\alpha. \quad (5.22)$$

- 3) Compute  $c_1$  and  $c_2$  by using Eqs. 5.11 and 5.12, respectively. In order to restore the signal level at the output of the filter let

$$\left. \begin{aligned} c'_1 &= \frac{c_1}{\lambda} \\ c'_2 &= \frac{c_2}{\lambda} \end{aligned} \right\} \quad (5.23)$$

where  $c'_i$  is the value of  $c_i$  after scaling.

- 4) The design is completed by connecting the scaled sections in parallel.

If the  $L_\infty$  norm is used for scaling,  $\alpha$  is given by

$$\alpha = \frac{\|F_1(z)\|_\infty}{\|F'_2(z)\|_\infty} = \frac{1+\alpha_1/2}{e} \sqrt{\frac{(\cos \omega_0 + f)^2 + \sin^2 \omega_0}{(\cos \omega_0 - 1)^2 + \sin^2 \omega_0}} \quad (5.24)$$

where  $\omega_0$  is the angle of the pole and

$$f = \frac{e^2}{1+\alpha_1/2} + \frac{\alpha_1}{2}.$$

The scaling parameter in this case is given by

$$\lambda = \frac{1}{\|F_2(z)\|_\infty} \approx \frac{1-r}{e\alpha} \sqrt{\frac{1+r^2-2r \cos 2\omega_0}{2(1-\cos \omega_0)}}. \quad (5.25)$$

### Cascade Design:

The transfer function is expressed as

$$T(z) = \prod_{i=1}^m H_i(z) \quad (5.26)$$

where each  $H_i(z)$  is of the form given by Eqn. 5.1. Then compute  $a_i$  and  $e_i$  by using Eqns. 5.17 and 5.18. Next compute  $\alpha$  and  $\lambda_i$  for each section as follows

$$\alpha = \frac{\left\| \left( \prod_{j=1}^{i-1} H_j(z) \right) F_{1i}(z) \right\|_p}{\left\| \left( \prod_{j=1}^{i-1} H_j(z) \right) F_{2i}(z) \right\|_p} \quad (5.27)$$

and

$$\lambda_i = \frac{1}{\left\| \left( \prod_{j=1}^{i-1} H_j(z) \right) F_{1i}(z) \right\|_p}. \quad (5.28)$$

The design is completed computing  $a_{12i}$ ,  $a_{21i}$ ,  $c_{1i}$  and  $c_{2i}$  using Eqns. 5.21, 5.22, 5.11 and 5.12, respectively.

The scaling multiplier  $\lambda_1$  is placed at the input of the filter while the remaining  $\lambda_i$ 's are incorporated in the output multipliers  $c_{1i}$ ,  $c_{2i}$  and  $d_i$  as follows

$$c'_{1i} = c_{1i} \frac{\lambda_{i+1}}{\lambda_i} \quad (5.29)$$

$$c'_{2i} = c_{2i} \frac{\lambda_{i+1}}{\lambda_i} \quad (5.30)$$

$$d'_i = d_i \frac{\lambda_{i+1}}{\lambda_i} \quad (5.31)$$

## 5.6 Comparison of Computational Complexity

Table 5.1 shows the number of arithmetic operations in the section-optimal structure and the structure of Fig. 5.1 for  $n$ th-order filters realized in parallel or cascade forms. As can be seen for an even-order filter, the new structure reduces the number of multipliers by  $n/2$  in a parallel design or by  $n-1$  in a cascade design, relative to the number of multipliers in the section-optimal structure. The number of adders, however, is increased by  $n/2$ .

## 5.7 Conclusions

In this chapter, Theorem 4.1 has been used to develop three new state-space structures in which zero-input and constant-input limit cycles as well as overflow oscillations can be efficiently eliminated.

The design aspects of the new structures have been examined in detail. It was found that one of the three structures is somewhat uneconomical whereas the other two turn out to be equivalent, if optimum signal scaling is applied. These two structures, like the section-optimal structure are optimal with respect to roundoff noise if the zeros of the transfer function are located at  $z = 1$ .

TABLE 5.1  
Arithmetic Operations

		Parallel		Cascade	
		n-even	n-odd	n-even	n-odd
New State-Space	Multipliers	$\frac{7n}{2} + 1$	$\frac{7n}{2} + \frac{1}{2}$	$\frac{7n}{2} + 1$	$\frac{7(n-1)}{2} + 4$
	Adders	$\frac{7n}{2} + 1$	$\frac{7}{2}n - \frac{3}{2}$	$\frac{7n}{2}$	$\frac{7(n-1)}{2} + 2$
Section-Optimal	Multipliers	$4n + 1$	$4n$	$\frac{9n}{2}$	$\frac{9(n-1)}{2} + 3$
	Adders	$3n + 1$	$3n - 1$	$3n$	$3n-1$

A procedure has then been developed for the design of high-order parallel and cascade filters by means of the state-space structure of Fig. 5.1. This approach yields designs which are more economical relative to designs based on the section-optimal structure. For an  $n$ th-order design, the number of multipliers is reduced by  $n/2$  for a parallel design or by  $n-1$  for a cascade design, although the number of adders is increased by  $n/2$ . This represents a significant saving in the amount of computation and the cost of hardware.



## CHAPTER 6

### ROUND-OFF NOISE ANALYSIS

#### 6.1 Introduction

In order to assess the quality of the structures proposed in Chapters 2, 3, and 5 and also to compare their performance with that of other known structures, several sixth-order filters are designed. Cascade designs are obtained with the proposed structures and also the direct canonic and section-optimal structures. Signal scaling is applied and the ordering of sections is chosen to minimize the roundoff noise in each design.

This chapter deals with a roundoff noise analysis which involves the computation of output-noise spectra for the various designs. Two's complement, fixed-point arithmetic is assumed in all examples.

Extensive noise comparisons are then undertaken. Sec. 6.2 compares the VGIC and TVGIC structures of Chapter 2 with the direct canonic and section-optimal structures; Sec. 6.3 compares the low-sensitivity structures of Chapter 3 with the section-optimal structure; and Sec. 6.4 compares one of the state-space structures of Chapter 5 with the section-optimal structure.

A sensitivity analysis for these designs is reported in Chapter 7.

#### 6.2 VGIC Structures

The transfer function in a cascade realization is given by

$$H(z) = H_0 \prod_{j=1}^m H_j(z) \quad (6.1)$$

where

$$H_j(z) = \frac{z^{2+\gamma_{1j}} z^{\gamma_{2j}}}{z^{2+\alpha_{1j}} z^{\alpha_{2j}}}$$

The partial transfer functions of a second-order section, denoted by  $F_{pi}(z)$  and  $G_{pi}(z)$ , are defined in Fig. 6.1. These are needed for scaling and roundoff noise analysis. The partial transfer functions for the VGIC and TVGIC structures are given in Table 6.1.

For a realization comprising  $m$  cascaded VGIC sections, the relative power spectrum density (RPSD) of the output noise can be shown to be

$$\text{RPSD} = \sum_{i=1}^m \left\{ \sum_{\ell=1}^n |G_{p\ell i}(e^{j\omega T})|^2 \right\} \lambda_{m+1} \prod_{j=i+1}^m \lambda_j |H_j(e^{j\omega T})|^2 \quad (6.2)$$

where  $n$  is the number of multipliers per second-order section,  $\lambda_j$  is the scaling constant of section  $j$  and

$$\lambda_{m+1} = \frac{H_0}{\prod_{i=1}^m \lambda_i}$$

The scaling constants  $\lambda_j$  are calculated by using the formula

$$\lambda_j = \frac{1}{\max_{\ell=1, n} \left\{ \left| F_{p\ell j}(e^{j\omega T}) \prod_{i=1}^{j-1} \lambda_i H_i(e^{j\omega T}) \right| \right\}} \quad (6.3)$$

Similarly, for a corresponding cascade TVGIC design

$$\text{RPSD} = \sum_{i=1}^m \left\{ \sum_{\ell=1}^n |G_{p\ell i}(e^{j\omega T})|^2 \right\} \frac{H_0}{\lambda_i} \prod_{j=i+1}^m |H_j(e^{j\omega T})|^2 \quad (6.4)$$

where

$$\prod_{j=m+1}^m H_j(e^{j\omega T}) = 1,$$

$$\lambda_j = \frac{1}{\max_{\ell=1, n} \left\{ \left| F_{p\ell j}(e^{j\omega T}) \prod_{i=1}^{j-1} H_i(e^{j\omega T}) \right| \right\}} \quad (6.5)$$

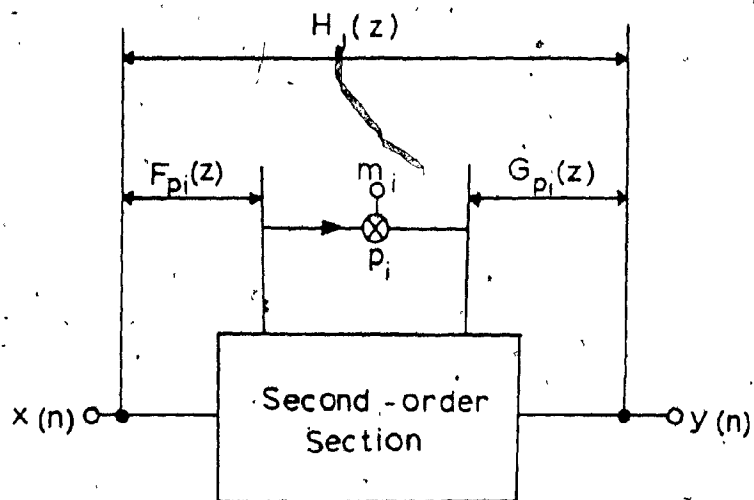


Fig. 6.1: Transfer Functions of Interest

TABLE 6.1  
Transfer Functions for  
VGIC and TVGIC Structures

VGIC Structure of Fig. 2.6	$H(z) = \frac{(c_0 + c_1 + c_2)z^2 + 2(c_0 - c_1)z + c_0 - c_1 + c_2}{D(z)}$ $F_{m_1}(z) = F_{m_2}(z) = H(z)$ $G_{m_1}(z) = \frac{z+1}{D(z)}$ $G_{m_2}(z) = \frac{z-1}{D(z)}$ $G_{c_0}(z) = \frac{(z+1)^2}{D(z)}$ $G_{c_1}(z) = \frac{z^2-1}{D(z)}$ $G_{c_2}(z) = \frac{(z-1)^2}{D(z)}$ $F_{c_0}(z) = F_{c_1}(z) = F_{c_2}(z) = 1$
TVGIC Structure * of Fig. 2.8 and 2.10	$H(z) = \frac{K[(c_0 + c_1 + c_2)z^2 + 2(c_0 - c_1)z + c_0 - c_1 + c_2]}{D(z)}$ $G_{m_1}(z) = G_{m_2}(z) = H(z)/K$ $F_{m_1}(z) = \frac{K(z+1)}{D(z)}$ $F_{m_2}(z) = \frac{K(z-1)}{D(z)}$ $F_{c_0}(z) = \frac{K(z+1)^2}{D(z)}$ $F_{c_1}(z) = \frac{K(z^2-1)}{D(z)}$ $F_{c_2}(z) = \frac{K(z-1)^2}{D(z)}$ $G_{c_0}(z) = G_{c_1}(z) = G_{c_2}(z) = 1$
$D(z) = z^2 + (m_1 - m_2)z + m_1 + m_2 - 1$	

\*  $K=1$  for the TVGIC structure that is amenable to ESS  
 $K=-m_1$  for the limit-cycle-free TVGIC structure

and

$$\lambda_{m+1} = H_0.$$

It should be mentioned that in the TVGIC design, the output multiplier coefficients  $c_{0j}$ ,  $c_{1j}$  and  $c_{2j}$  of section  $j$  are replaced by  $c_{0j}\lambda_{j+1}/\lambda_j$ ,  $c_{1j}\lambda_{j+1}/\lambda_j$  and  $c_{2j}\lambda_{j+1}/\lambda_j$ , respectively, in order to avoid the use of scaling multipliers [1].

For the sake of comparison, the cascade approach was used to design several sixth-order filters as follows:

- 1) An elliptic lowpass filter
- 2) An elliptic bandpass filter
- 3) A Butterworth bandstop filter
- 4) A Chebyshev highpass filter.

Designs were obtained with the VGIC and TVGIC structures, with the direct canonic structure [1], and with the section-optimal structure described in [21].

The specifications assumed are summarized in Table 6.2 where

- |                              |                                |
|------------------------------|--------------------------------|
| $A_p$ :                      | maximum passband loss          |
| $A_a$ :                      | minimum stopband loss          |
| $\omega_{p1}, \omega_{p2}$ : | lower and upper passband edges |
| $\omega_{a1}, \omega_{a2}$ : | lower and upper stopband edges |
| $\omega_s$ :                 | sampling frequency.            |

The transfer function coefficients of the lowpass, bandpass, bandstop and highpass filters are given in Table A.1.

Table A.2 gives the multiplier constants for lowpass designs based on the

TABLE 6.2  
Filter Specifications

	Elliptic Lowpass	Elliptic Bandpass	Butterworth Bandstop	Chebyshev Highpass
$A_p$ , db	1.0	0.5	2.0	0.6
$A_a$ , db	72.9	65.6	32.4	48.5
$\omega_{p1}$ , rad/s	250.0	980.0	500.0	4000.0
$\omega_{p2}$ , rad/s	---	1020.0	1000.0	---
$\omega_{a1}$ , rad/s	400.0	850.0	650.0	3300.0
$\omega_{a2}$ , rad/s	---	1150.0	850.0	---
$\omega_s$ , rad/s	10000.0			

- 1) VGIC structure of Fig. 2.6
- 2) TVGIC structure of Fig. 2.8 with  $x_1(n)$  as input
- 3) TVGIC structure of Fig. 2.10 with first- and second-order ESS
- 4) Direct canonic structure [1]
- 5) Section-optimal structure [21].

Signal scaling was applied in all designs, and the scaling constants  $\lambda_j$  were chosen on the basis of the  $L_\infty$  norm. All possible section sequences were scaled and the optimum section sequence from the point of view of signal/noise ratio was chosen for each design. The implementation was assumed to be in terms of fixed-point arithmetic, and quantization of products was assumed to be performed after the multiplications, except for the structures with ESS where product quantization was assumed to be performed after additions.

The RPSD of the output roundoff-noise for the various designs was then computed. In this analysis roundoff errors were assumed to be uncorrelated from sample to sample and from one source to another. These assumptions were shown to be valid by several experimental results presented in [1].

For a better exposition of the results, the RPSD plots are divided in two categories:

- (i) Plots for medium-noise designs like the VGIC, limit-cycle-free TVGIC, and direct-canonic designs, as illustrated in Fig. 6.2.
- (ii) Plots for low-noise designs like the section-optimal design and the TVGIC design with ESS incorporated, as illustrated in Fig. 6.3.

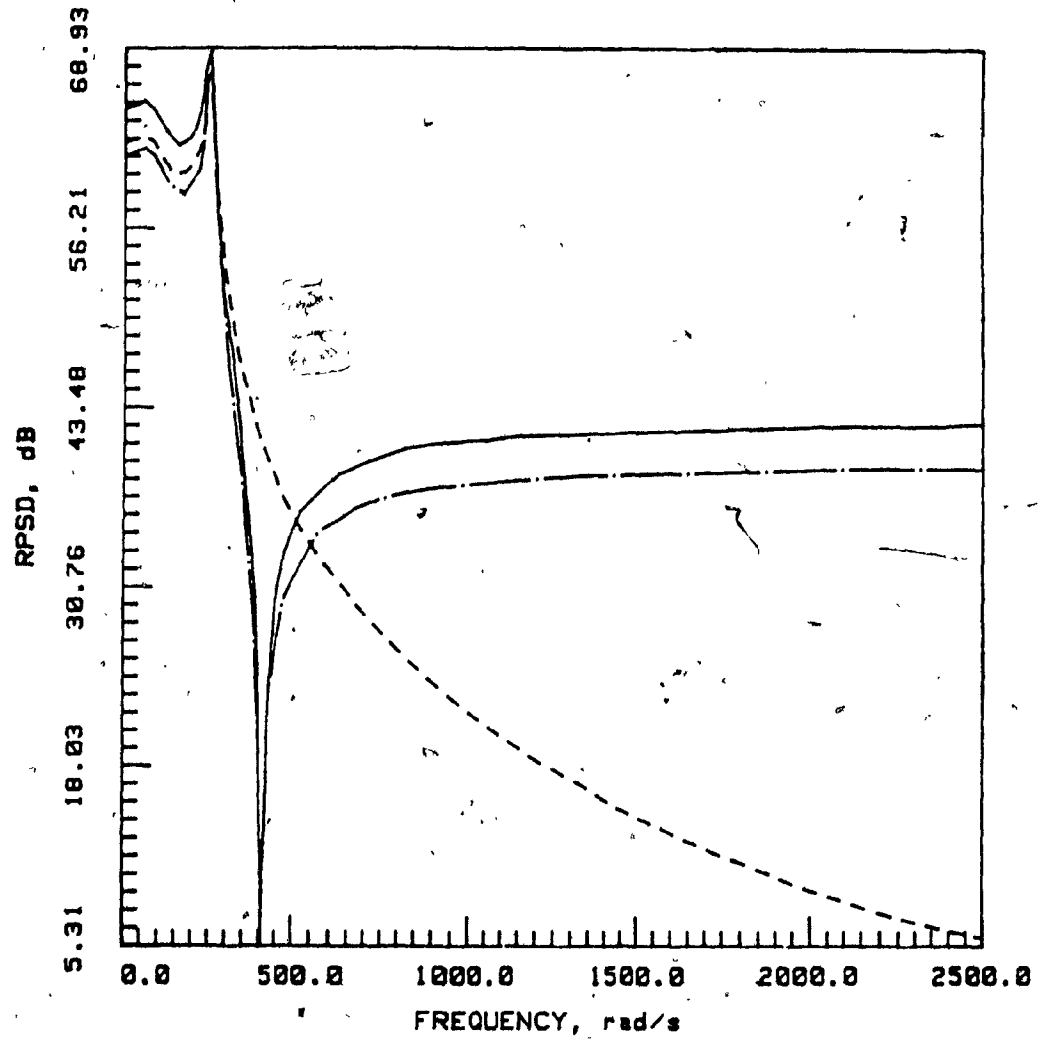


Fig. 6.2: RPSD versus Frequency (Elliptic Lowpass)

- VGIC
- .- Direct Canonic
- Limit-Cycle-Free TVGIC



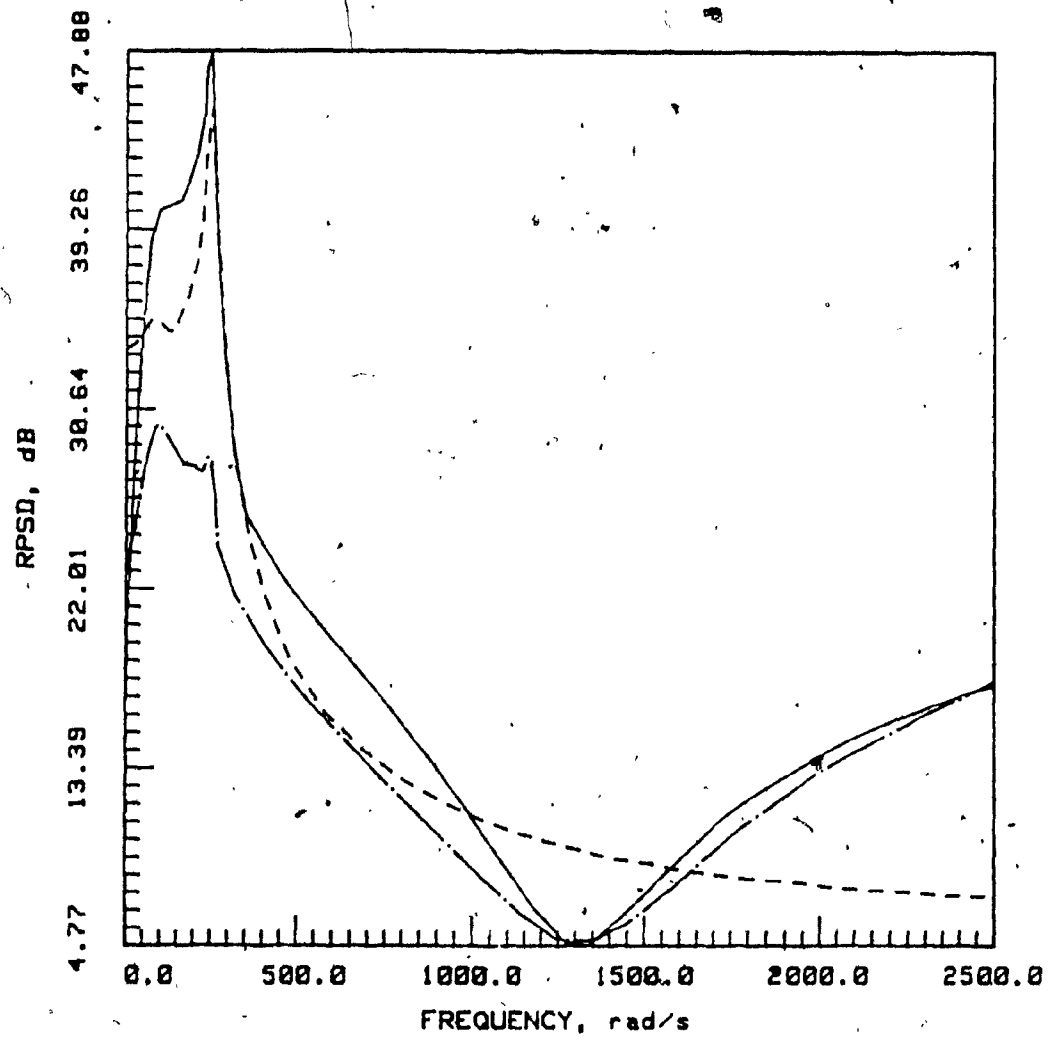


Fig. 6.3: RPSD versus Frequency (Elliptic Lowpass)

- Section-Optimal
- . - . - . TVGIC with 2<sup>nd</sup>-Order ESS
- TVGIC with 1<sup>st</sup>-Order ESS

The roundoff-noise plots of Fig. 6.2 show that the VGIC and the direct canonic designs are very close in terms of inband noise, while the limit-cycle-free TVGIC design is somewhat noisier. Fig. 6.3 shows that the TVGIC design with first-order ESS is comparable with the section-optimal design while the TVGIC design with second-order ESS is definitely the best.

Table A.3 gives the multiplier constants for the bandpass designs for the same structures used in the lowpass design described above.

Figs. 6.4 and 6.5 show the RPSD plots for the medium- and low-noise designs, respectively. The direct canonic design is shown to be the best among the medium-noise designs followed by the VGIC design. In the low-noise class, the TVGIC design with second-order ESS has a much lower output noise than the section-optimal design and TVGIC design with first-order ESS.

Table A.4 gives the multiplier constants for the bandstop designs for the same structures compared above. The RPSD plots of the various designs are depicted in Figs. 6.6 and 6.7. The VGIC design is clearly better than the direct-canonic design, while the limit-cycle-free TVGIC design is the worst. The section-optimal design in this case outperforms the TVGIC design with ESS because ESS is not very effective for filters with wide passband(s).

Table A.5 gives the multiplier constants for the highpass design for the same structures discussed so far. The RPSD plots of these filters are depicted in Figs. 6.8 and 6.9. Fig. 6.8 shows that the VGIC design is the best followed by the direct canonic design. Fig. 6.9 shows that the TVGIC design with second-order ESS is the best

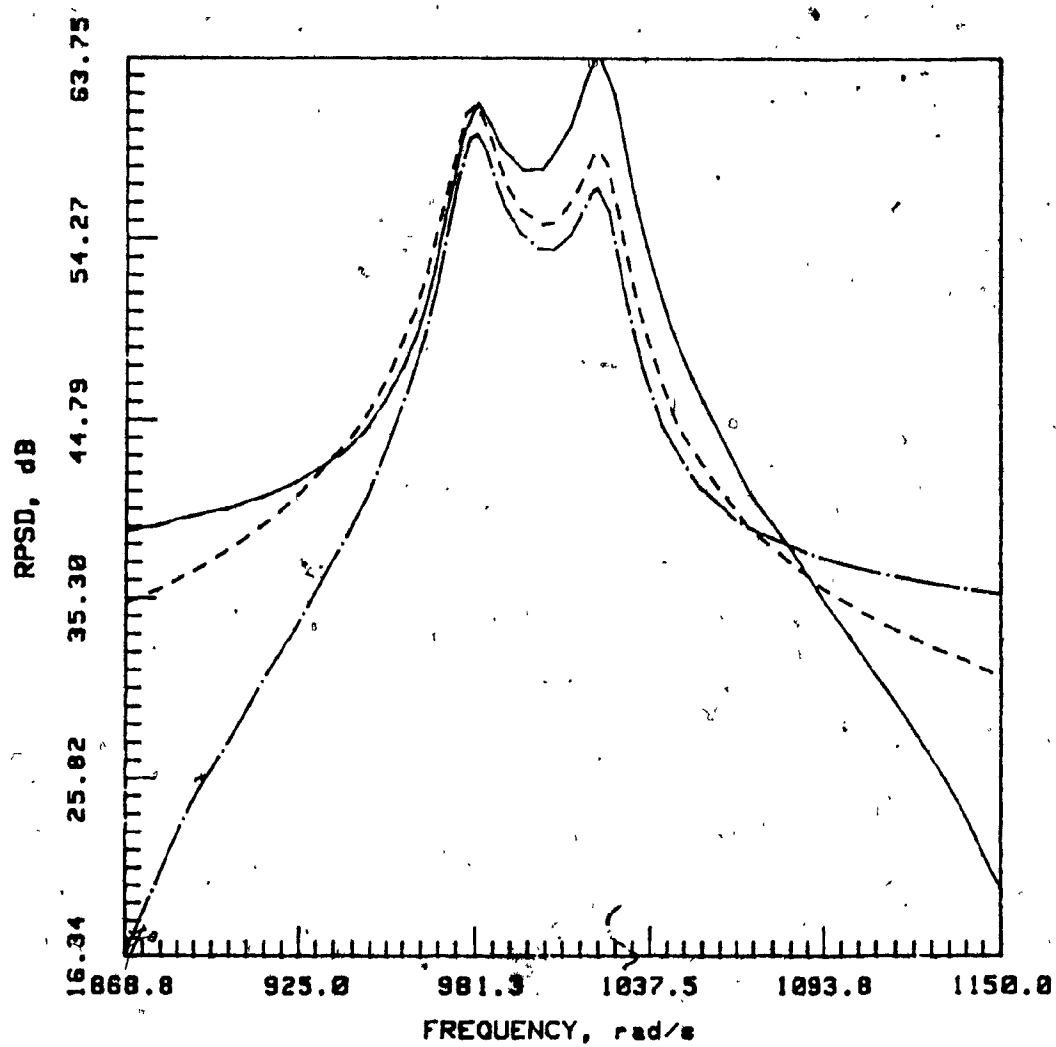


Fig. 6.4: RPSD versus Frequency (Elliptic Bandpass)

- VGIC
- . - . - Direct Canonic
- Limit-Cycle-Free TVGIC

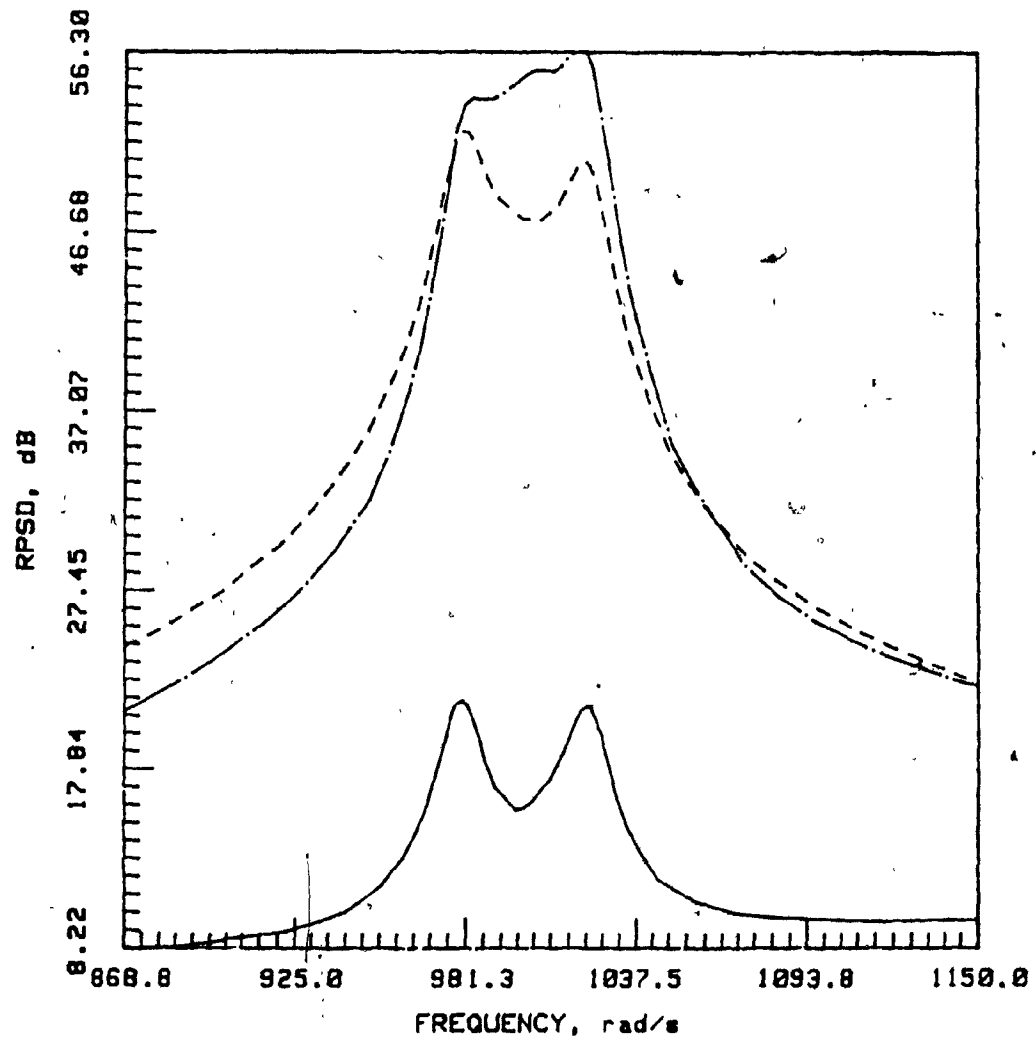


Fig. 6.5: RPSD versus Frequency (Elliptic Bandpass)

- Section-Optimal
- TVGIC with 2<sup>nd</sup>-Order ESS
- · - · - TVGIC with 1<sup>st</sup>-Order ESS

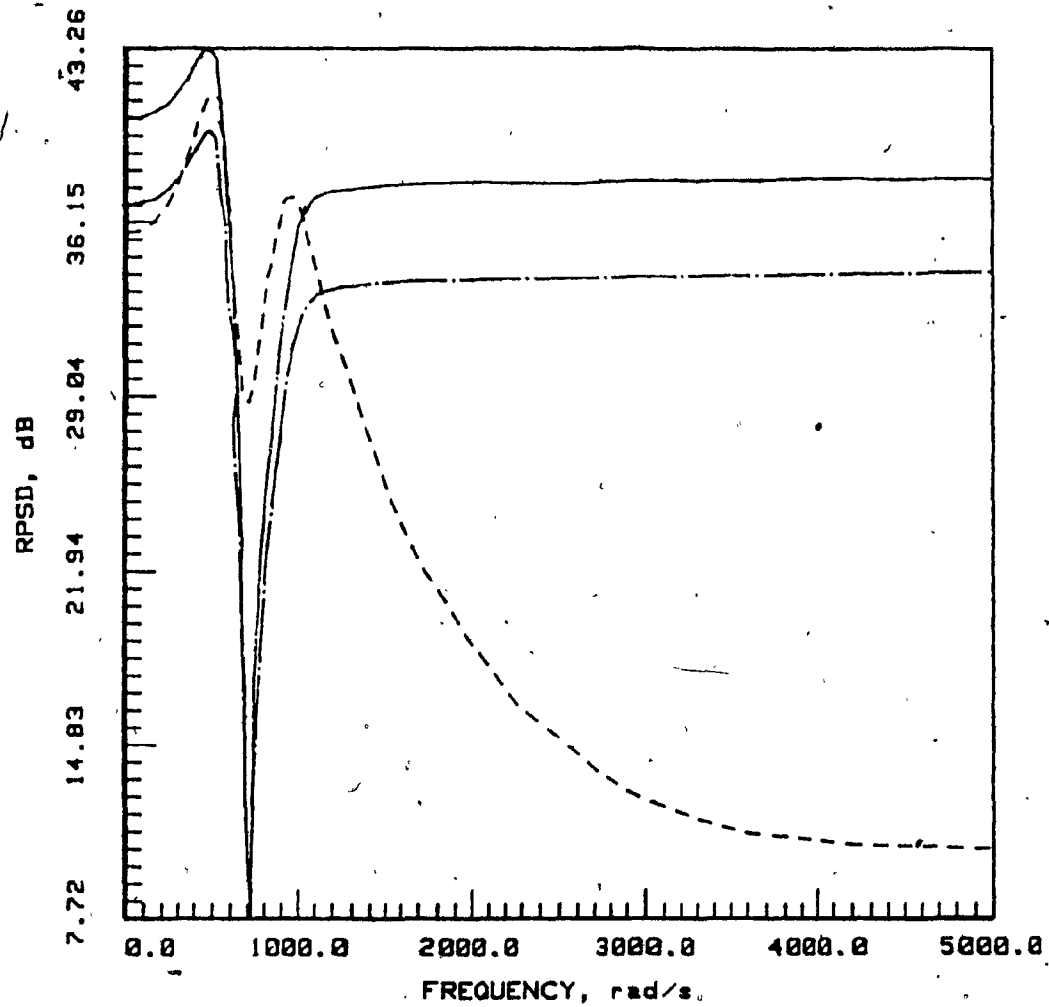


Fig. 6.6: RPSD versus Frequency (Butterworth Bandstop)

- VGIC
- . - . - Direct Canonic
- \_\_\_\_\_ Limit-Cycle-Free TVGIC

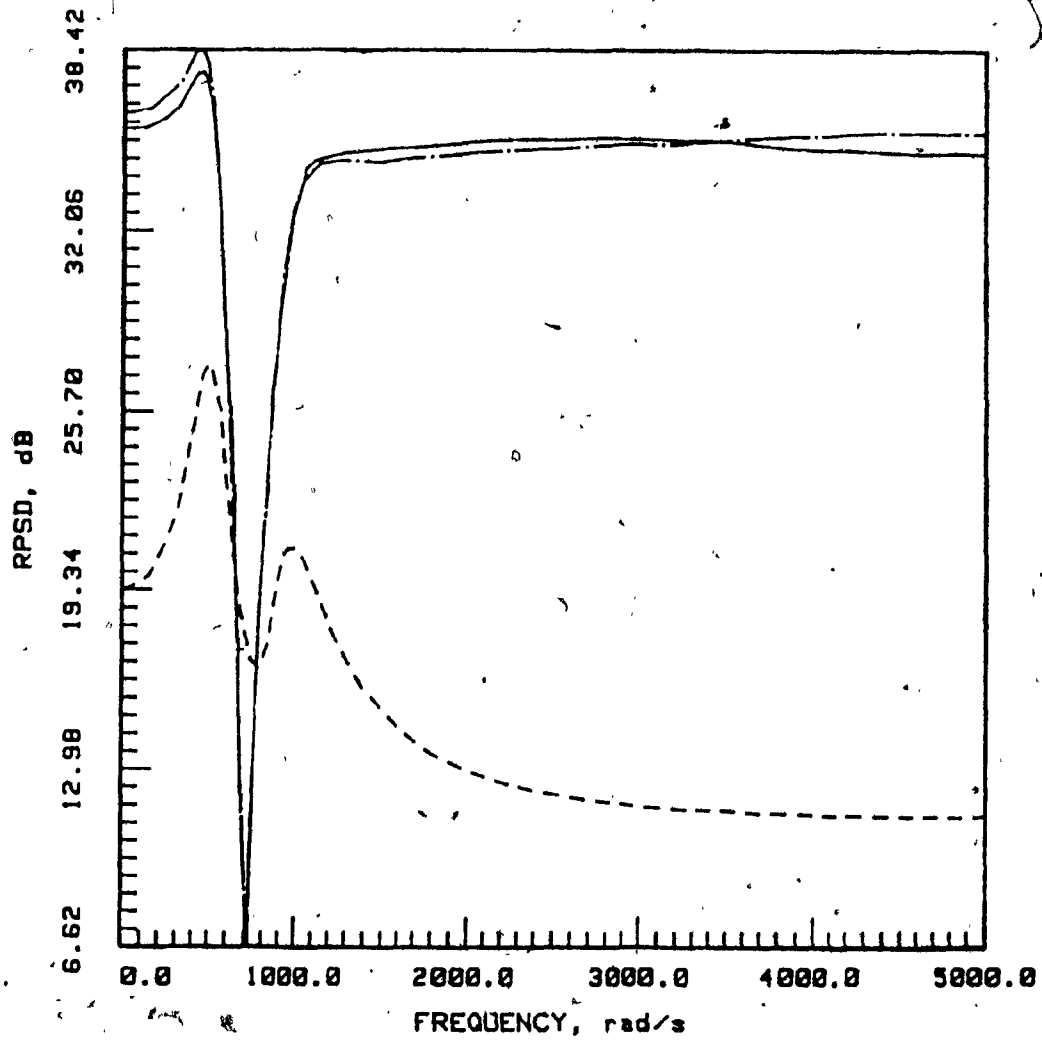


Fig. 6.7: RPSD versus Frequency (Butterworth Bandstop)

- Section-Optimal /
- TVGIC with 2<sup>nd</sup>-Order ESS
- · - · - TVGIC with 1<sup>st</sup>-Order ESS

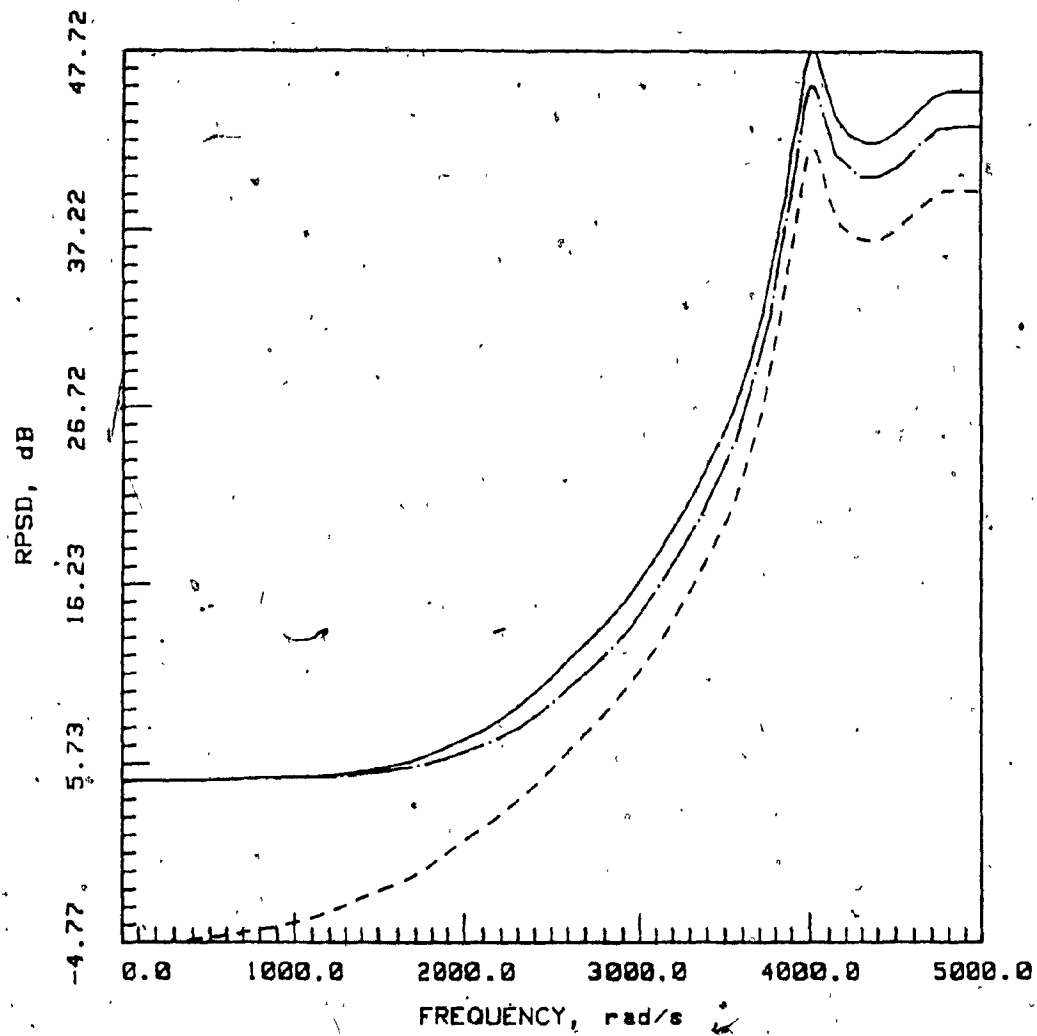


Fig. 6.8: RPSD versus Frequency (Chebyshev Highpass)

- VGIC
- Direct Canonic
- Limit-Cycle-Free TVGIC

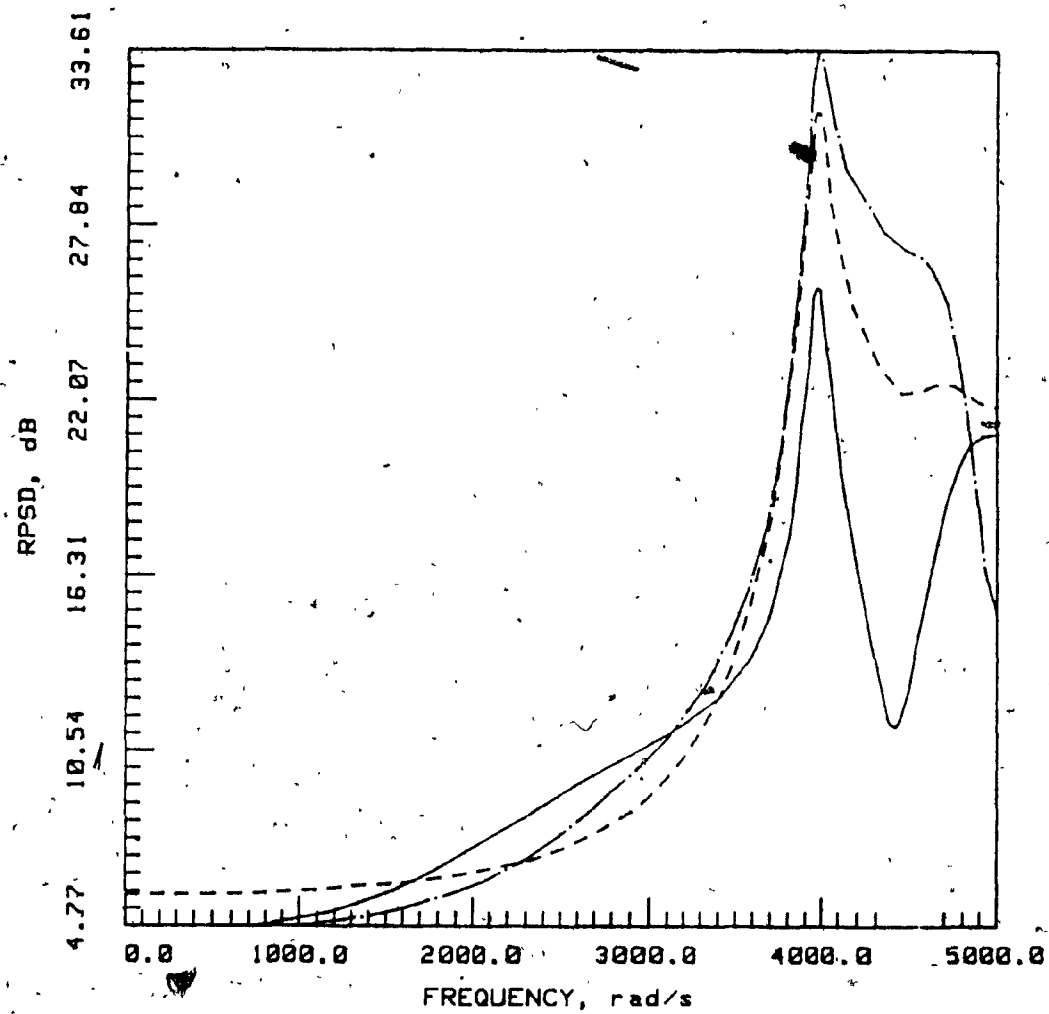


Fig. 6.9: RPSD versus Frequency (Chebyshev Highpass)

- Section-Optimal
- TVGIC with 2<sup>nd</sup>-Order ESS
- . - . - . TVGIC with 1<sup>st</sup>-Order ESS



followed by the section-optimal design.

Finally, the average noise for each of the designs considered so far is given in Table 6.3. Average values are tabulated for each design, one with the scaling based on the  $L_\infty$  norm and the other with the scaling based on the  $L_2$  norm. This table confirms the results based on the RPSD plots described above. It should be mentioned the choice of scaling norm, either  $L_\infty$  or  $L_2$ , does not affect the choice of the optimum design.

### 6.3 Low-Sensitivity Structures of Chapter 3

In this section, the low-sensitivity structures of Chapter 3 are compared with the section-optimal structure.

Transfer functions  $F_{pi}(z)$  and  $G_{pj}(z)$  (see Fig. 6.1) for the various low-sensitivity structures are given in Table 6.4.

For a realization comprising  $m$  cascade low-sensitivity sections with ESS incorporated, the output noise RPSD can be shown to be

$$\text{RPSD} = 3 + \sum_{i=1}^m \left| \frac{H_0}{\lambda_i} (1 + b_{1i}e^{-j\omega T} + b_{0i}e^{-2j\omega T}) \prod_{\ell=i}^m H_\ell(e^{j\omega T}) \right|^2 \quad (6.6)$$

where  $\lambda_i$  is the scaling constant of section  $i$ ,  $b_{1i}$  and  $b_{0i}$  are the parameters of the ESS substructure of section  $i$ . The scaling constants are given by

$$\lambda_i = \frac{1}{\max_{\ell=1,3} \left\{ \left\| F_{\delta_{\ell i}}(e^{j\omega T}) \prod_{j=1}^{i-1} H_j(e^{j\omega T}) \right\|_p \right\}} \quad (6.7)$$

As in the TVGIC structure, the output multipliers  $\delta_{0i}$ ,  $\delta_{1i}$  and  $\delta_{2i}$  of section  $i$  have been multiplied by  $\lambda_{i+1}/\lambda_i$  in order to avoid the use of scaling multipliers constant. The scaling constant  $\lambda_{m+1}$  of the last section, is given by  $H_0$ .

TABLE 6.3  
Average Noise in dbs

Design	Lowpass Elliptic		Bandpass Elliptic		Bandstop Butterworth		Highpass Chebyshev	
	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$
VGIC	50.00	35.04	40.29	19.27	30.92	26.19	32.12	23.16
Limit-cycle-free TVGIC	52.37	37.08	43.05	21.25	38.09	28.09	37.92	29.02
TVGIC with 1st-order ESS	29.41	13.08	35.43	13.95	34.89	24.89	22.04	12.39
TVGIC with 2nd-order ESS	21.44	9.41	9.94	4.87	34.77	24.74	14.50	7.16
Direct Canonic	49.06	33.44	38.43	16.85	34.27	24.09	35.84	26.93
Section-Optimal	25.54	10.74	31.43	11.24	17.12	11.46	19.47	11.01

TABLE 6.4

Transfer Functions in the Low-Sensitivity Structures\*

CASE I	$F_{m_1}(z) = F_{\delta_1}(z) = \frac{(z-D)}{D(z)}$ $F_{m_2}(z) = F_{\delta_2}(z) = \frac{1}{D(z)}$ $F_{\delta_0}(z) = z F_{\delta_1}(z)$ $G_{m_1}(z) = G_{m_2}(z) = H(z)$ $G_{\delta_0}(z) = G_{\delta_1}(z) = G_{\delta_2}(z) = 1$ $D(z) \text{ is given by Eqn. 3.3}$
CASE II	$F_{m_1}(z) = F_{\delta_1}(z) = \frac{(z-B)}{D(z)}$ $F_{m_2}(z) = F_{\delta_2}(z) = \frac{z}{D(z)}$ $F_{\delta_0} = z F_{\delta_1}(z)$ $G_{m_1}(z) = G_{m_2}(z) = H(z)$ $G_{\delta_0}(z) = G_{\delta_1}(z) = G_{\delta_2}(z) = 1$ $D(z) \text{ is given by Eqn. 3.6}$

\*These formulas are valid for the design equations of Table 3.3

Several sixth-order filters, namely, a lowpass, a bandpass, a bandstop, and a highpass filter, were designed using the new structures and the section-optimal structure. Specifications and transfer functions for the lowpass and highpass filters are the same as in Tables 6.2 and A.1, respectively. The specifications and transfer functions for the bandpass and bandstop filters are summarized in Table 6.5 and A.6, respectively.

As was demonstrated in Chapter 3, two or more low-sensitivity structures are possible for each second-order transfer function and according the analysis in Sec. 7.4, it is possible to select the least sensitive structure. On the basis of this analysis, the least sensitive structures for the lowpass, bandpass, bandstop, and highpass filters were found to be structures II-1, I-5, I-7, and II-6, respectively.

The multiplier constants for the lowpass design, which was based on structure II-1, are given in Table A.7. The RPSD plots for this design with first- and second-order ESS and for the section-optimal design are depicted in Fig. 6.10. The superior performance of structure II-1 is clearly observed.

Table A.8 gives the multiplier constants for the bandpass designs which were based on structure I-5 and on the section-optimal structure, respectively. In Fig. 6.11 the output noise RPSD plots for these designs are depicted. As can be seen the design based on structure I-5 is superior relative to the design based on the section-optimal structure.

TABLE 6.5  
Filter Specifications

	Elliptic Bandpass	Butterworth Bandstop
$A_p$ , db	0.5	2.0
$A_a$ , db	66.2	33.9
$\omega_{p1}$ , rad/s	1180.0	2300.0
$\omega_{p2}$ , rad/s	1220.0	2700.0
$\omega_{a1}$ , rad/s	1050.0	2450.0
$\omega_{a2}$ , rad/s	1350.0	2550.0
$\omega_s$ , rad/s	10000.0	

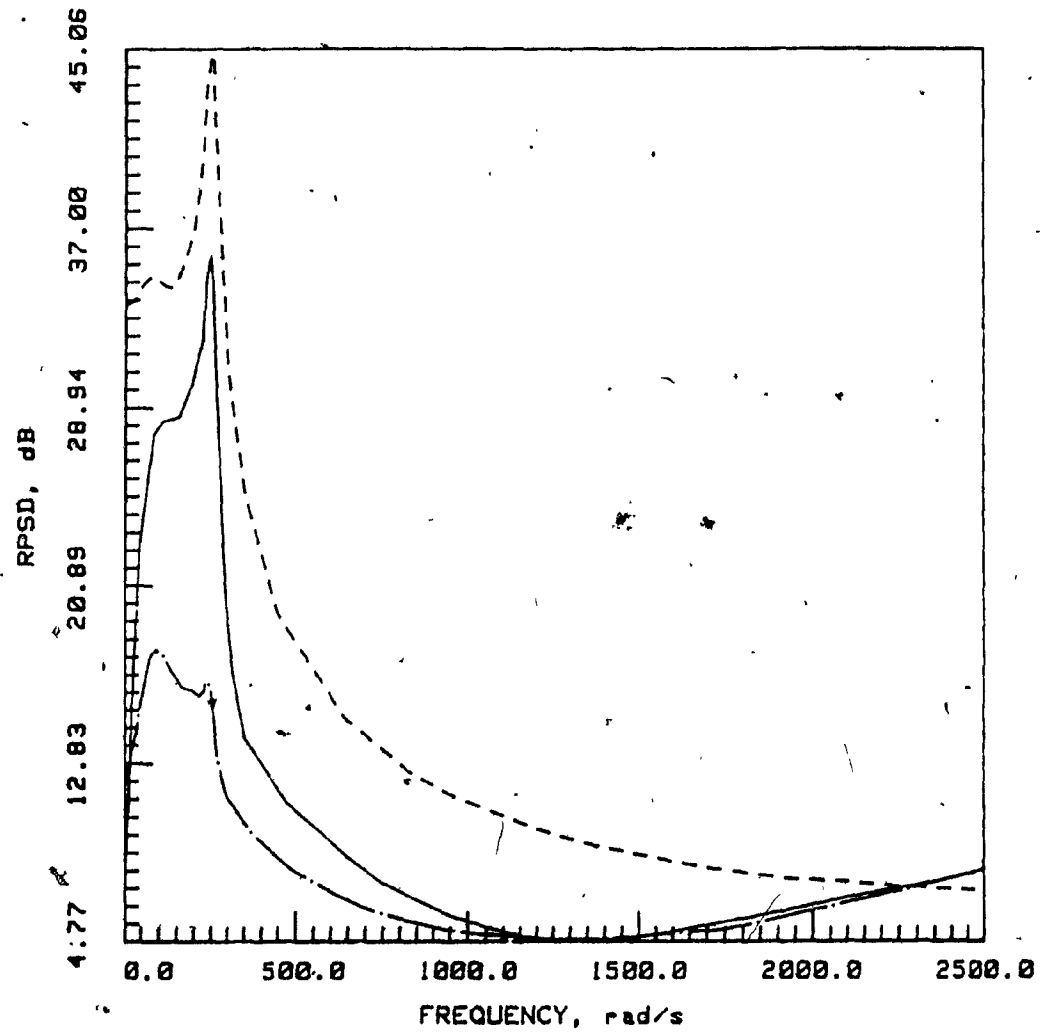


Fig. 6.10: RPSD versus Frequency (Elliptic Lowpass)

- Section-Optimal
- . - . - . II-1 with 2<sup>nd</sup>-Order ESS
- \_\_\_\_\_ II-1 with 1<sup>st</sup>-Order ESS

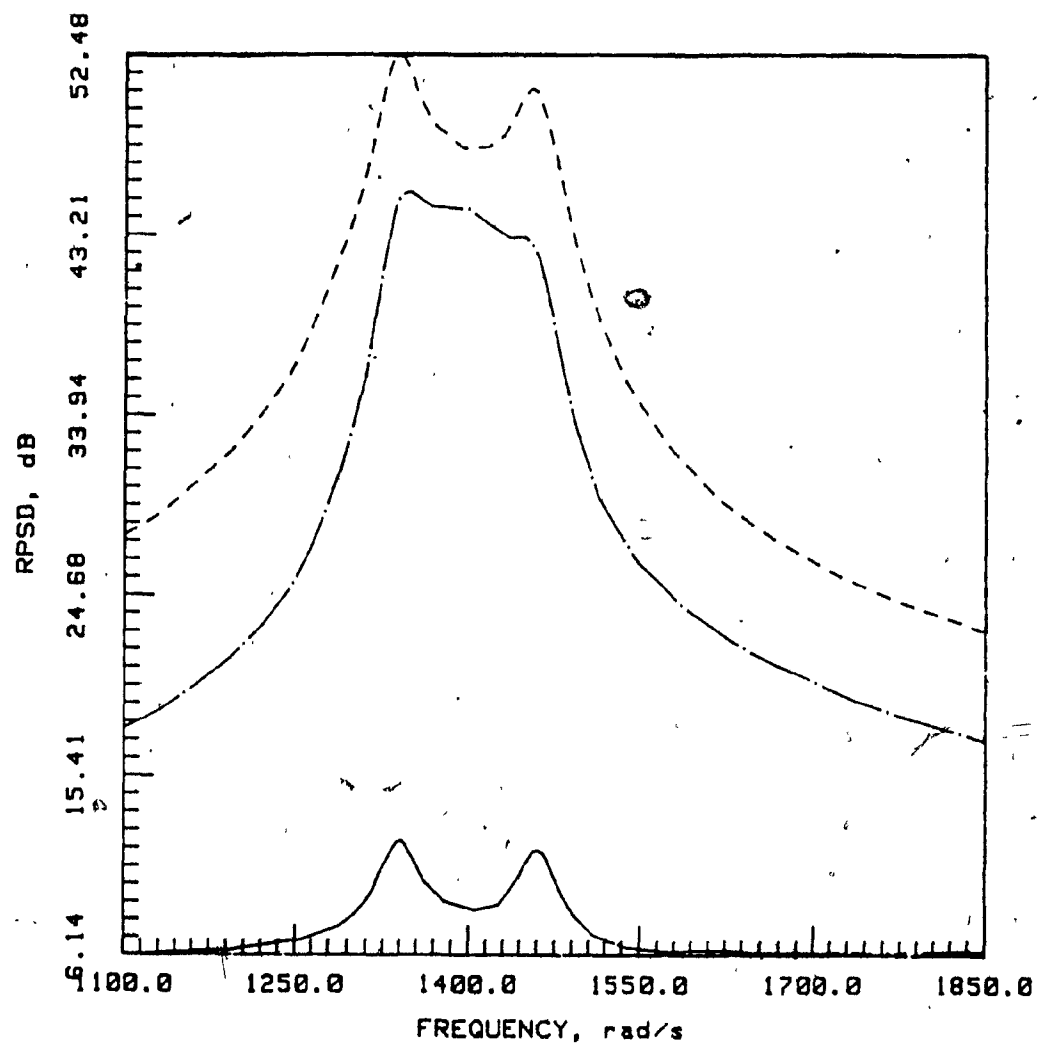


Fig. 6.11: RPSD versus Frequency (Elliptic Bandpass)

- Section-Optimal
- I-5 with 2<sup>nd</sup>-Order ESS
- . - . - . I-5 with 1<sup>st</sup>-Order ESS

The multiplier constants of the bandstop designs, which were based on structure I-7 and the section-optimal structure, are given in Table A.9. The RPSD plots for these designs are shown in Fig. 6.12 for this design. The section-optimal structure gives better results than structure I-7.

Finally, the multiplier constants of the highpass design, which was based on structure II-6, are given in Table A.10. The RPSD plots for the design based on structure II-6 and that based on the section-optimal structure are depicted in Fig. 6.13. The design based on structure II-6 is clearly superior.

In terms of roundoff noise, all the structures of Chapter 3 are similar because the noise transfer functions are the same in all cases and the dominant transfer function needed in the scaling process is given by  $1/D(z)$ .

The average noise for each example presented in this section is given in Table 6.6 for signal scaling based on  $L_\infty$  and  $L_2$  norms. The new structures are superior in the lowpass, bandpass and highpass filters, while the section-optimal structure is better in the bandstop case.

#### 6.4 New State-Space Structure

The transfer functions from the state variable nodes of the new state-space structures presented in Chapter 5, to the output are given by

$$G_1(z) = c_1' \frac{z + \alpha_1/2 + \xi}{z^2 + \alpha_1 z + \alpha_2} \quad (6.8)$$



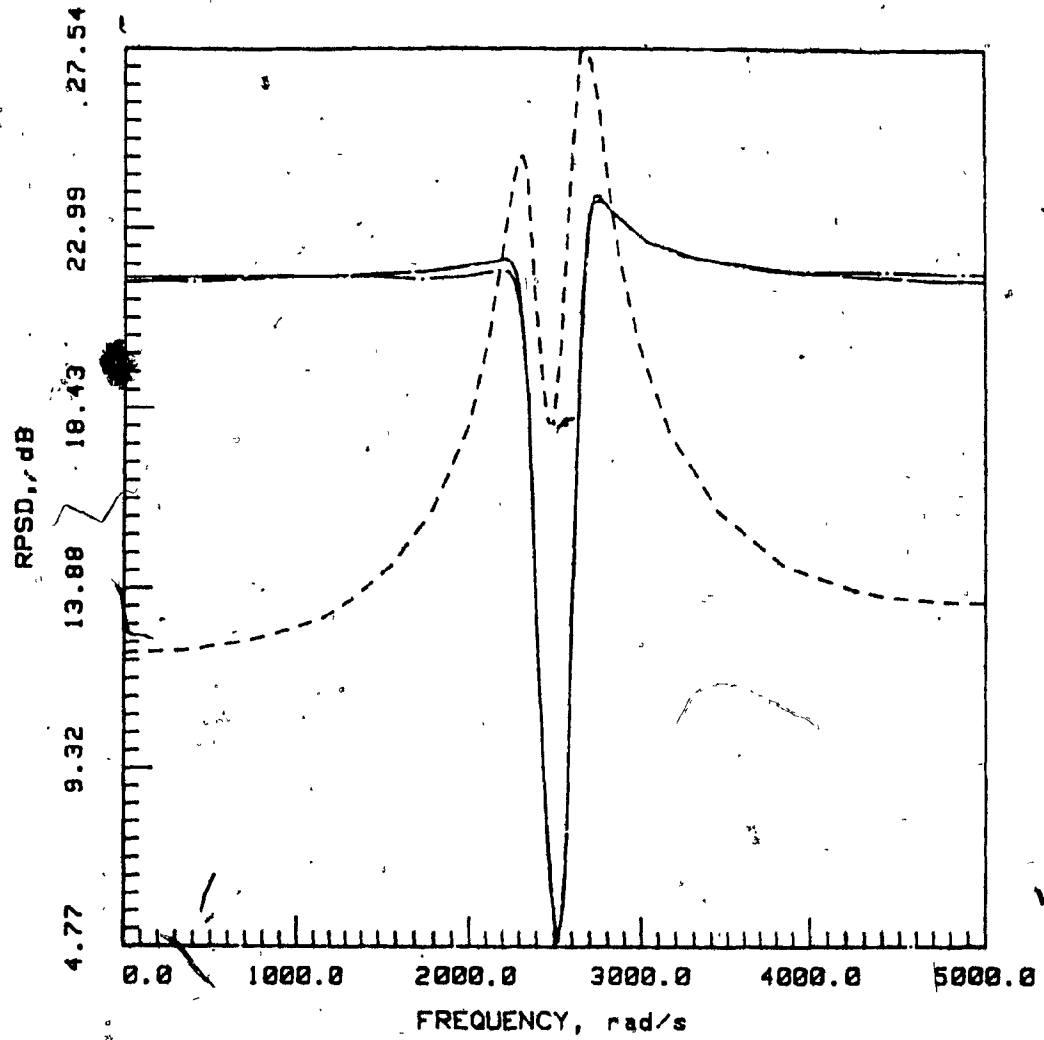


Fig. 6.12: RPSD versus Frequency (Butterworth Bandstop)

- Section-Optimal
- I-7 with 2<sup>nd</sup>-Order ESS
- . - . - . I-7 with 1<sup>st</sup>-Order ESS

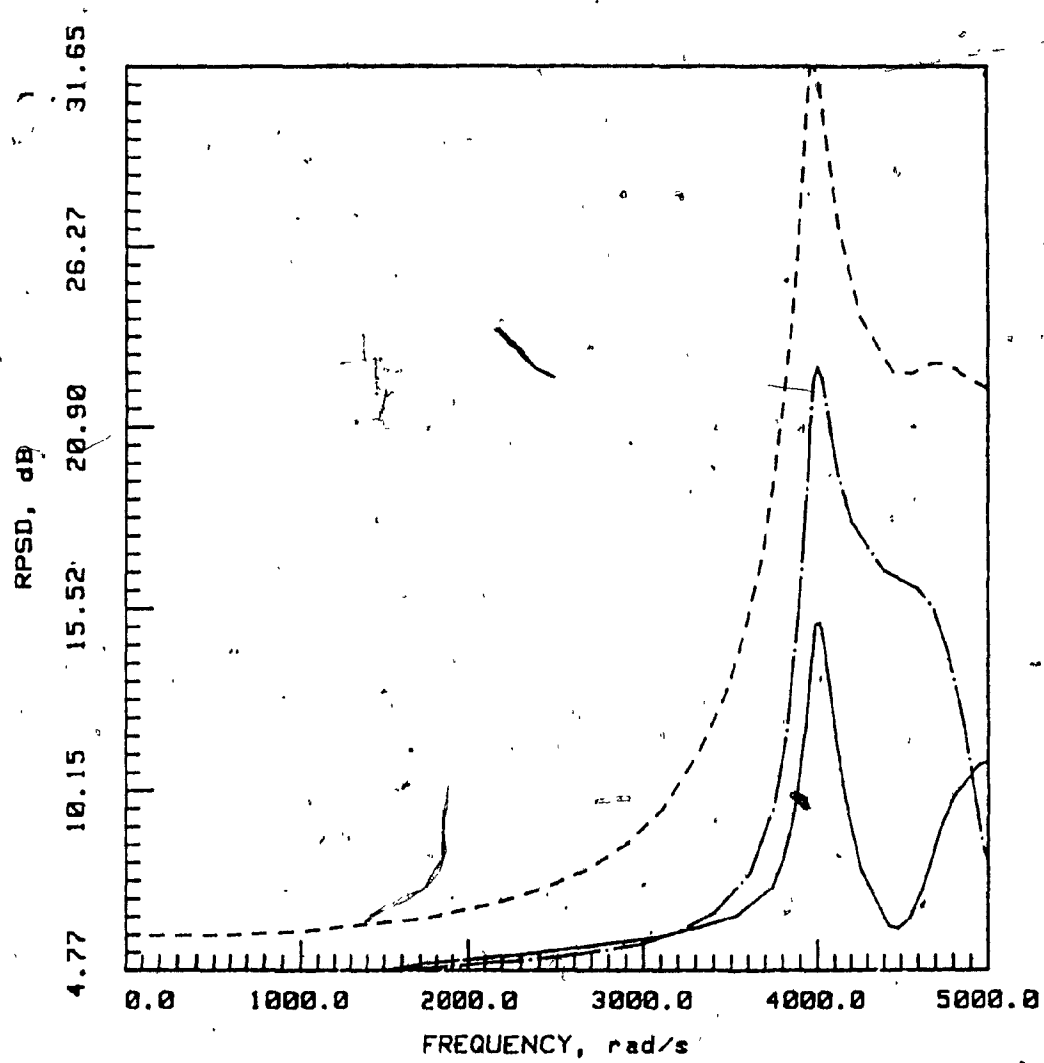


Fig. 6.13: RPSD versus Frequency (Chebyshev Highpass)

- Section-Optimal
- II-6 with 2<sup>nd</sup>-Order ESS
- · - · - II-6 with 1<sup>st</sup>-Order ESS

TABLE 6.6  
Average Noise in db

	Lowpass Elliptic		Bandpass Elliptic		Bandstop Butterworth		Highpass Chebyshev	
Scaling	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$
Low Sensitivity Structure with 1st-order ESS	17.63	6.12	25.09	7.09	21.59	11.96	12.52	6.03
Low Sensitivity Structure with 2nd-order ESS	10.65	5.26	6.31	4.79	21.57	11.95	6.61	4.99
Section- Optimal Structure	25.54	10.74	31.77	11.52	18.01	11.32	19.47	11.01

and

$$G_2(z) = c_2' \frac{z + \alpha_1/2 - e^{2/\xi}}{z^2 + \alpha_1 z + \alpha_2} \quad (6.9)$$

where

$$\xi = \frac{-(\alpha_1/2 + \alpha_2)\beta_1 + (1 + \alpha_1/2)\beta_2}{\beta_1 + \beta_2}$$

For a design comprising  $m$  cascade state-space structures, the output noise RPSD can be shown to be

$$\text{RPSD} = \sum_{i=1}^m \frac{1}{\lambda_{m+1}} \{2|G_{1i}(e^{j\omega T})|^2 + 2|G_{2i}(e^{j\omega T})|^2 + 3\} \left| \prod_{j=i+1}^m H_j(e^{j\omega T}) \right|^2 \quad (6.10)$$

where,  $H_j(z)$  is given in Eqn. 5.1,  $\lambda_i$  is given by Eqn. 5.29 and  $\lambda_{m+1}$  is given by

$$\lambda_{m+1} = 1.$$

The coefficient multipliers  $d_i$  for  $i = 1, 2, \dots, m-1$ , should be determined as

$$d_i = \frac{1}{\left\| \prod_{j=1}^i H_j(z) \right\|_p} \quad (6.11)$$

in order to satisfy the required overflow constraints at the output of each section and

$$d_m = \frac{H_0}{\prod_{i=1}^{m-1} d_i} \quad (6.12)$$

The sixth-order filters described in Tables 6.2 and A.1 were designed using the state-space structure proposed in Chapter 5. The multiplier constants for the lowpass, bandpass, bandstop and highpass designs are given in Tables A.11 to A.14, respectively.

In Figs. 6.14 to 6.17, the RPSD plots for the new design and the section-optimal design are depicted. The new design is better in the lowpass, bandpass, and bandstop examples, while the section-optimal is better for the highpass example. The examples show that the difference in noise performance between these structures is marginal.

The average noise for each of the examples in Table 6.2 is given in Table 6.7. This table confirms the conclusions based on the RPSD plots described above.

### 6.5 Section Ordering

Although the problem of section ordering has not been studied in depth, some techniques proposed in the literature were found to be applicable in some of the proposed designs.

In designs based on the structures of Chapter 3, section ordering can be accomplished by using the technique proposed in [45] for the direct canonic design with ESS incorporated. This is possible owing to the fact that the noise transfer function is the same in both designs and the scaling constants have in general the same values. Our computer simulations on various designs have shown that noise outputs for differently-ordered sections using ESS vary less in percentage when scaling constants are calculated on the basis of the  $L_2$  norm as opposed to the  $L_\infty$  norm. When the scaling constants are chosen on the basis of  $L_\infty$  norm, the ordering of sections becomes a very important step in the design process.

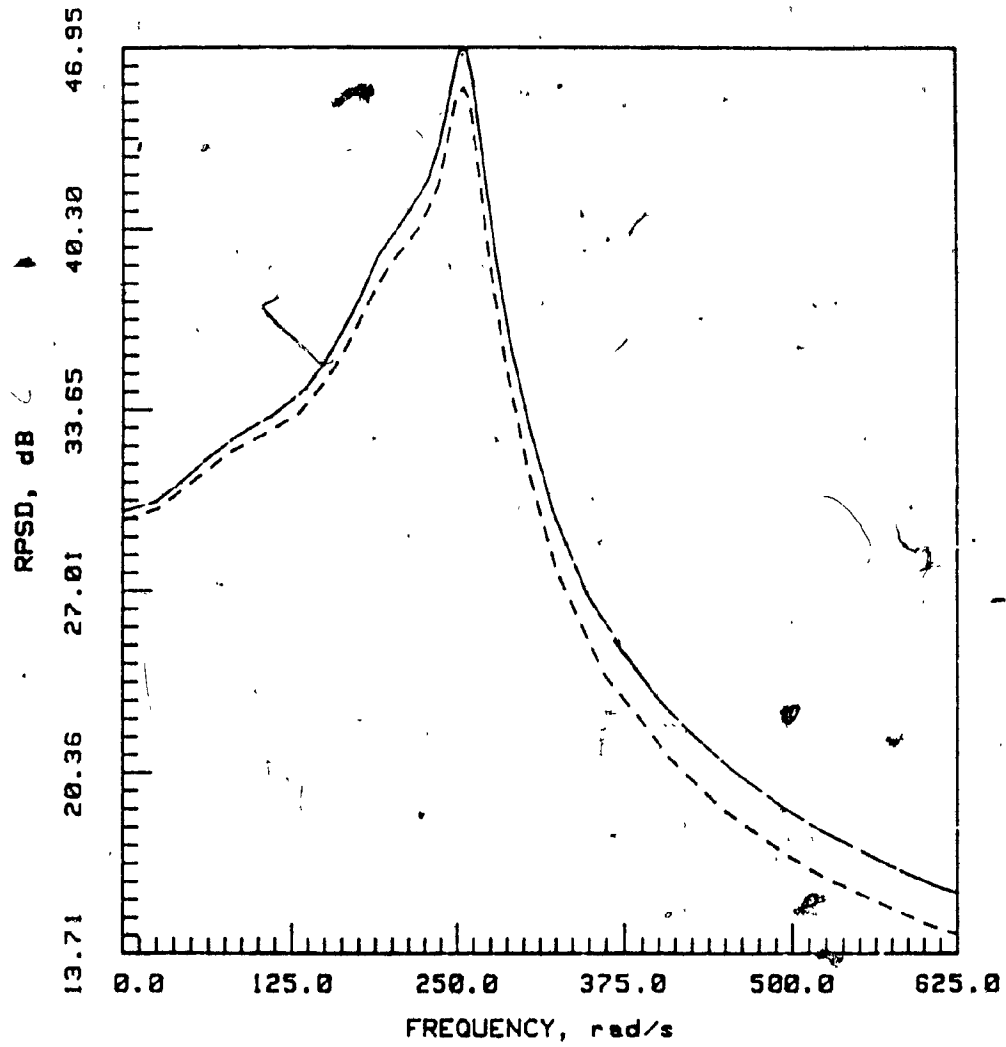


Fig. 6.14: RPSD versus Frequency (Elliptic Lowpass)

— Section-Optimal  
- - - New State-Space

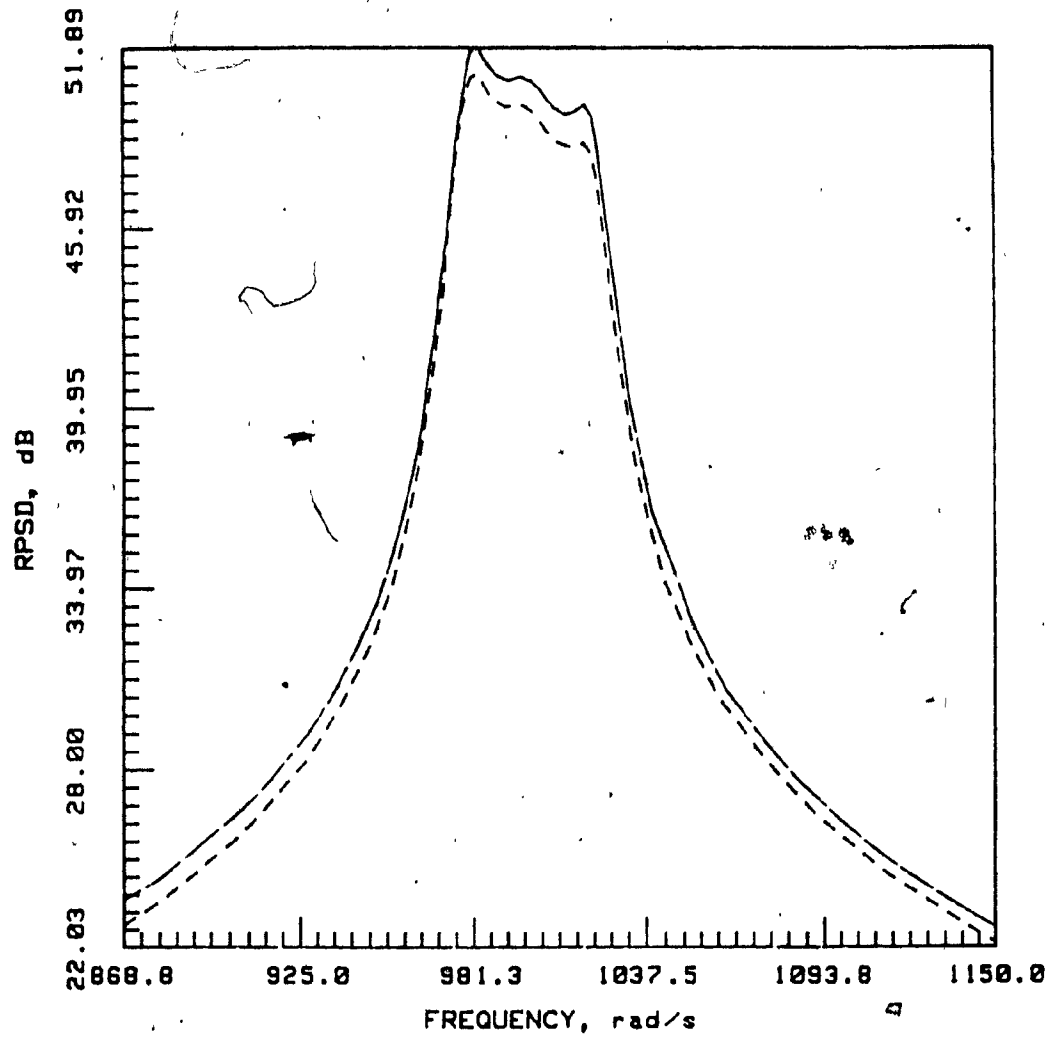


Fig. 6.15: RPSD versus Frequency (Elliptic Bandpass)

— Section-Optimal  
- - - New State-Space

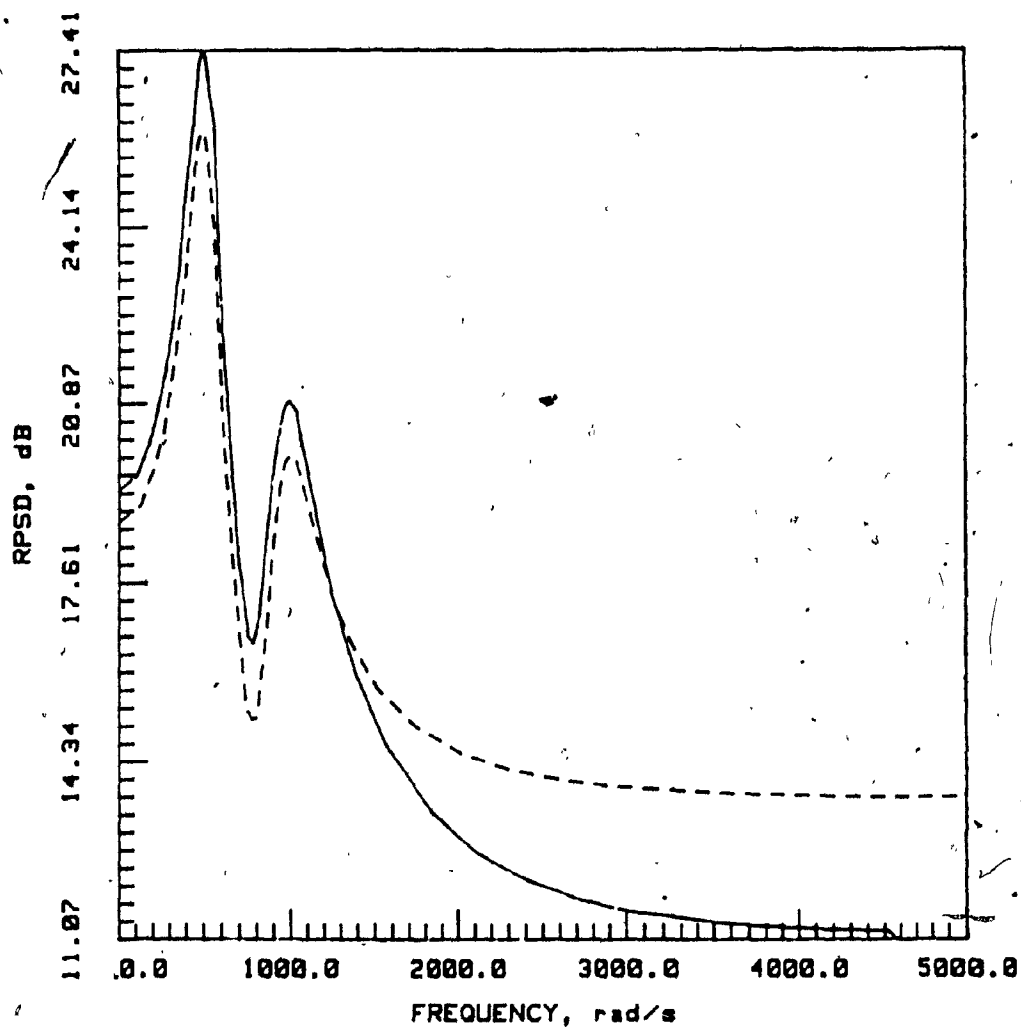


Fig. 6.16: RPSD versus Frequency (Butterworth Bandstop)

— Section-Optimal  
- - - New State-Space



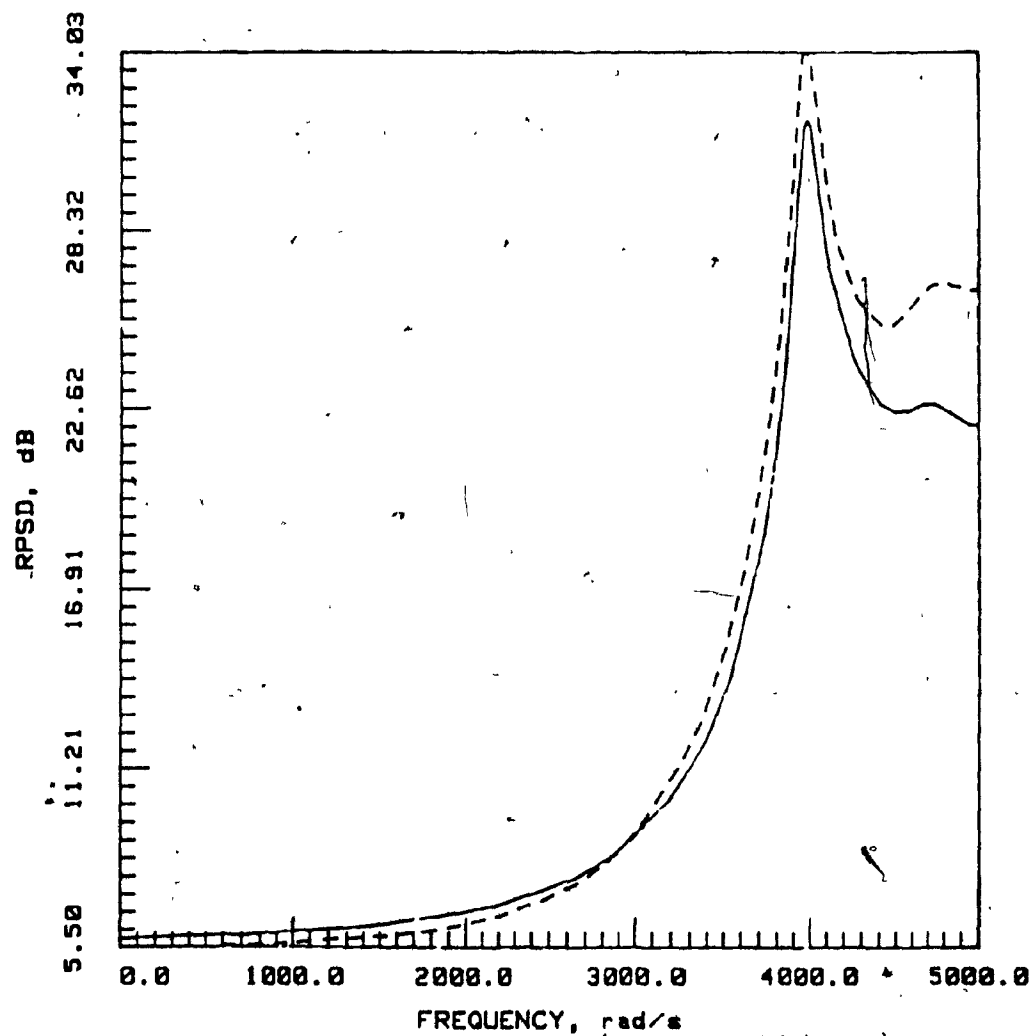


Fig. 6.17: RPSD versus Frequency (Chebyshev Highpass)

— Section-Optimal  
- - - New State-Space

TABLE 6.7  
Average Noise in dbs

	Lowpass Elliptic		Bandpass Elliptic		Bandstop Butterworth		Highpass Chebyshev	
	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$	$L_{\infty}$	$L_2$
Scaling								
New State-Space Structure	24.30	10.10	30.83	10.56	16.92	9.13	22.03	12.37
Section- Optimal Structure	25.54	10.74	31.43	11.24	17.12	11.46	19.47	11.01

In [46] an analytical procedure is described for the pole-zero pairing and section ordering for the section-optimal cascade design such that a reduction in the output noise level is achieved. This procedure has been found to be equally useful for cascade designs based on the new state-space structure of Chapter 5.

## 6.6 Conclusions

The effect of product quantization in the proposed structures has been examined by evaluating the RPSD of the output noise in several cascade designs. The various designs were then compared with corresponding designs based on the direct canonic and section-optimal structures.

The VGIC structure has been shown to give comparable results as the direct canonic structure, while the limit-cycle-free TVGIC structure gives inferior results. The TVGIC structure with second-order ESS is better than the section-optimal structure, except in the bandstop design. The TVGIC structure with first-order ESS is comparable with the section-optimal structure, except for the bandstop design where the latter is best.

The low-sensitivity structures proposed in Chapter 3 with ESS incorporated gave the best noise performance among the structures proposed, except for the bandstop design where the section-optimal structure led to comparable noise performance.

The new state-space structure proposed in Chapter 5 was shown to give comparable results as the section-optimal structure for all the designs considered.

## CHAPTER 7

### SENSITIVITY ANALYSIS

#### 7.1 Introduction

In this chapter the sensitivity aspects pertaining to the VGIC and TVGIC structures of Chapter 2, the low-sensitivity structures of Chapter 3, and the new state-space structure of Chapter 5 are studied in detail.

The techniques of Chapter 3 are used to reduce the sensitivity in VGIC and TVGIC structures.

The maximum sensitivity  $\hat{S}$  is used as a sensitivity measure, and an optimal subset of the low-sensitivity structures of Chapter 3 is identified. These structures can collectively realize any second-order transfer function.

Finally, a sensitivity comparison of the proposed structures with the direct canonic structure and section-optimal structure is undertaken.

#### 7.2 Sensitivity Measures

The sensitivity of the transfer function

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

of a digital-filter structure with respect to variations in multiplier constant  $m_i$  is defined as

$$S_{m_i}^{H(z)} = \frac{N_{m_i}(z)}{D_{m_i}(z)} = \frac{m_i}{H(z)} \frac{\partial H(z)}{\partial m_i} \quad (7.1)$$

If

$$H(e^{j\omega T}) = M(\omega)e^{j\theta(\omega)}$$

the function

$$S = \sum_{i=1}^k |S_{m_i}^H(e^{j\omega T})| \quad (7.2)$$

can be formed where  $k$  is the number of multipliers in the structures and

$$|S_{m_i}^H(e^{j\omega T})| = \{(S_{m_i}^M(\omega))^2 + |m_i|^2 (\frac{\partial \theta(\omega)}{\partial m_i})^2\}^{1/2}.$$

If the poles of the structure are close to the unit circle  $|z| = 1$ , and  $m_i$  is a noninteger denominator multiplier constant, then for  $\omega$  close to the pole frequency  $\omega_0$

$$|S_{m_i}^M(\omega)| \gg |m_i| \left| \frac{\partial \theta(\omega)}{\partial m_i} \right|$$

and hence

$$|S_{m_i}^H(e^{j\omega T})| \approx |S_{m_i}^M(\omega)| \quad (7.3)$$

In addition, if  $m_j$  is a numerator multiplier constant then

$$|S_{m_i}^H(e^{j\omega T})| \gg |S_{m_j}^H(e^{j\omega T})| \quad (7.4)$$

and, therefore, for  $\omega \approx \omega_0$  Eqns. 7.2 - 7.4 yield

$$S \approx \sum_{i=1}^{k_d} |S_{m_i}^H(e^{j\omega T})| \approx \sum_{i=1}^{k_d} |S_{m_i}^M(\omega)| \quad (7.5)$$

where  $k_d$  is the number of denominator noninteger multiplier constants. This quantity can serve as a sensitivity measure which can be used for the comparison of different high-selectivity structures.

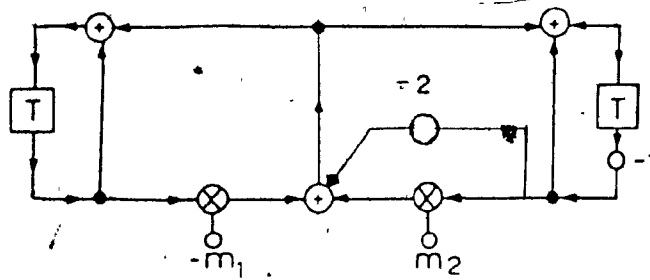
### 7.3 VGIC Structures

By using the techniques of Chapter 3, improvements can be brought about in the sensitivity performance of the VGIC and TVGIC structures of Chapter 2. Such improvements can be achieved by forcing the values of  $m_1$  and  $m_2$  to be low in order to ensure low sensitivity of the transfer function with respect to changes in  $m_1$  and  $m_2$ . For example, the sensitivity in the TVGIC structure of Fig. 2.8 can be reduced by modifying its recursive part as depicted in Fig. 7.1(a) to (c). These modifications do not affect the output roundoff noise or the limit-cycle behaviour of the structure, although the number of additions is increased by two.

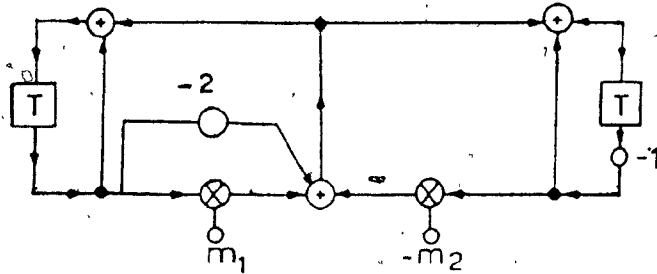
Table 7.1 gives the numerator polynomials of  $S_{m_i}^{H(z)}$  and the design equations for the original as well as the improved TVGIC structures.

The optimum structure for a given transfer function with poles close to the unit circle can easily be determined because for  $\alpha_2 \approx 1$ ,  $m_1 \approx m_2$  for all three improved TVGIC structures. The optimum structure is the one in which  $|m_1|$  is closest to zero. The structure of Fig. 7.1(a) has lower sensitivity than the structures of Fig. 7.1(b) and (c) for a pole angle  $\omega_0$  such that  $0 < \omega_0 < \pi/3$ , the structure of Fig. 7.1(b) is the best for  $2\pi/3 < \omega_0 < \pi$ , while the structure of Fig. 7.1(c) is the best for  $\pi/3 < \omega_0 < 2\pi/3$ .

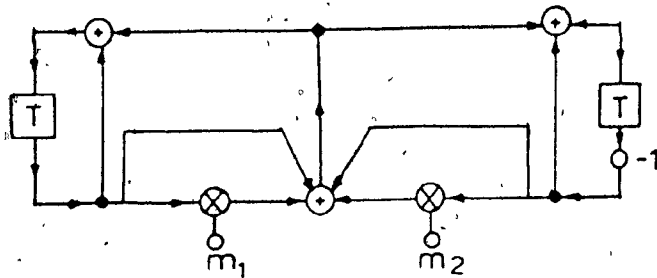
Fig. 7.2(a) to (d) shows the maximum sensitivity  $\hat{S}$  for pole angles  $\omega_0$  between 0 and  $\pi/2$  for  $\alpha_2 = 0.985$ , for the TVGIC, improved TVGIC of Fig. 7.1(a), the direct canonic, and the section-optimal structures, respectively. The direct canonic structure is the worst for pole angles  $\omega_0$  smaller than about  $\pi/3$ . The TVGIC and



(a)



(b)



(c)

Fig. 7.1: Improved TVGIC Structures

(a) For Pole Angle  $0 \leq \omega_0 \leq \pi/3$

(b) For Pole Angle  $2\pi/3 \leq \omega_0 \leq \pi$

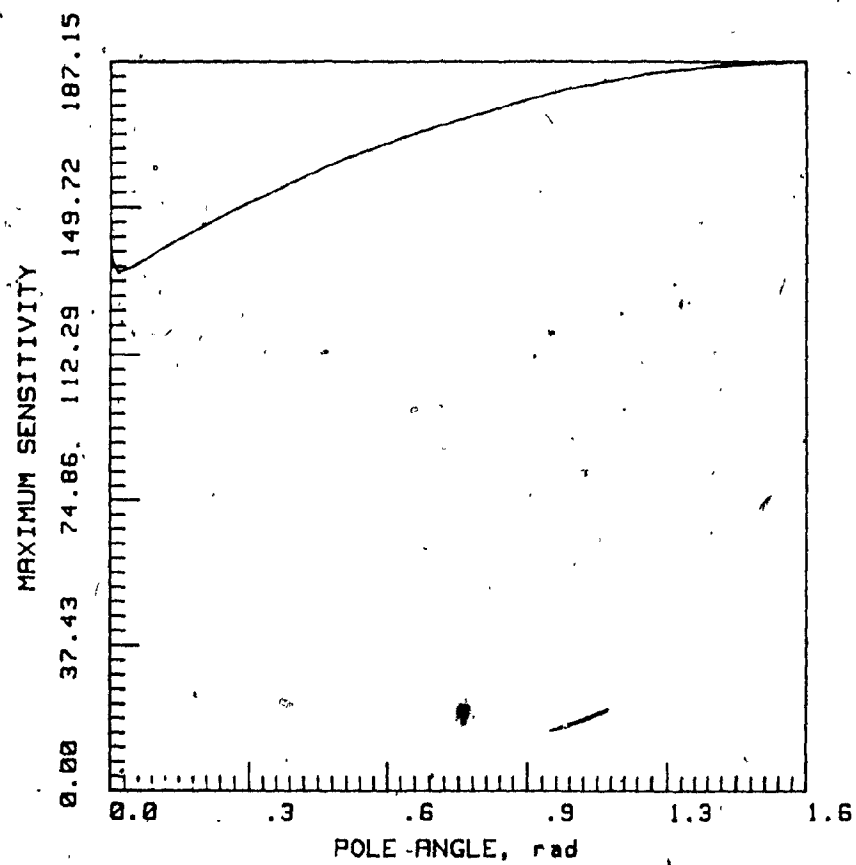
(c) For Pole Angle  $\pi/3 \leq \omega_0 \leq 2\pi/3$

TABLE 7.1

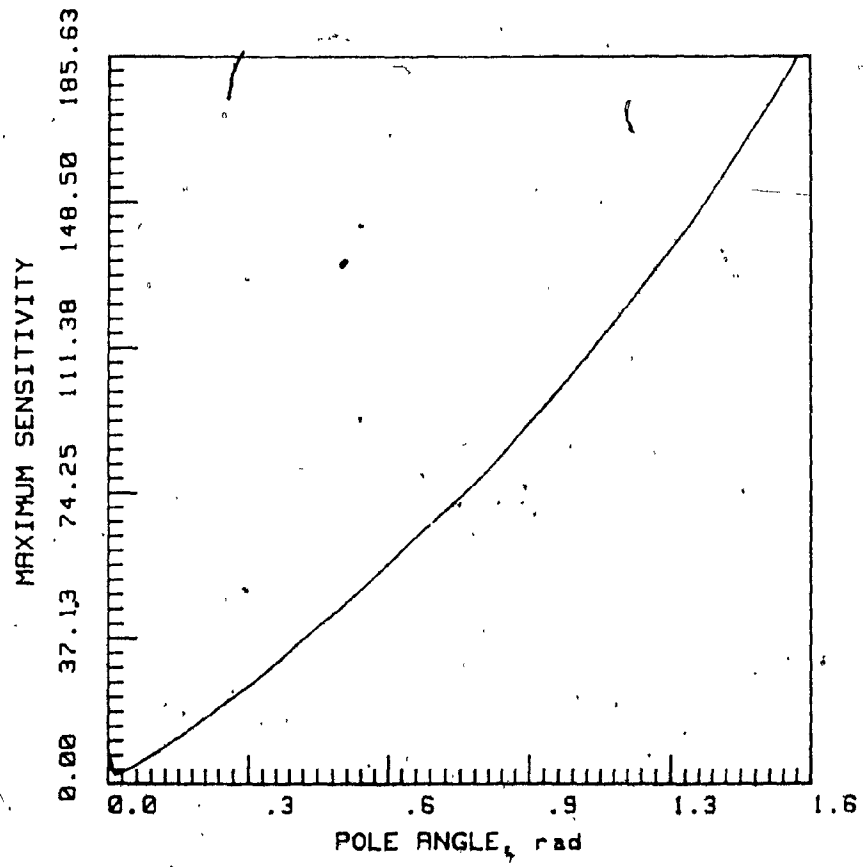
Numerator Polynomials of  
Sensitivities and Design Equations

	$N_{m_1}(z)$	$N_{m_2}(z)$	Design Equations
TVGIC	$m_1(z+1)$	$-m_2(z-1)$	$m_1 = \frac{\alpha_1 + \alpha_2 + 1}{2}$ $m_2 = \frac{-\alpha_1 + \alpha_2 + 1}{2}$
Improved TVGIC Fig. 7.1(a)	$m_1(z+1)$	$m_2(z-1)$	$m_1 = \frac{\alpha_1 + \alpha_2 + 1}{2}$ $m_2 = \frac{\alpha_1 - \alpha_2 + 3}{2}$
Improved TVGIC Fig. 7.1(b)	$-m_1(z+1)$	$m_2(z-1)$	$m_1 = \frac{1 - \alpha_2 - \alpha_1}{2}$ $m_2 = \frac{1 + \alpha_1 - \alpha_2}{2}$
Improved TVGIC Fig. 7.1(c)	$-m_1(z+1)$	$-m_2(z-1)$	$m_1 = \frac{3 - \alpha_1 - \alpha_2}{2}$ $m_2 = \frac{1 - \alpha_1 + \alpha_2}{2}$

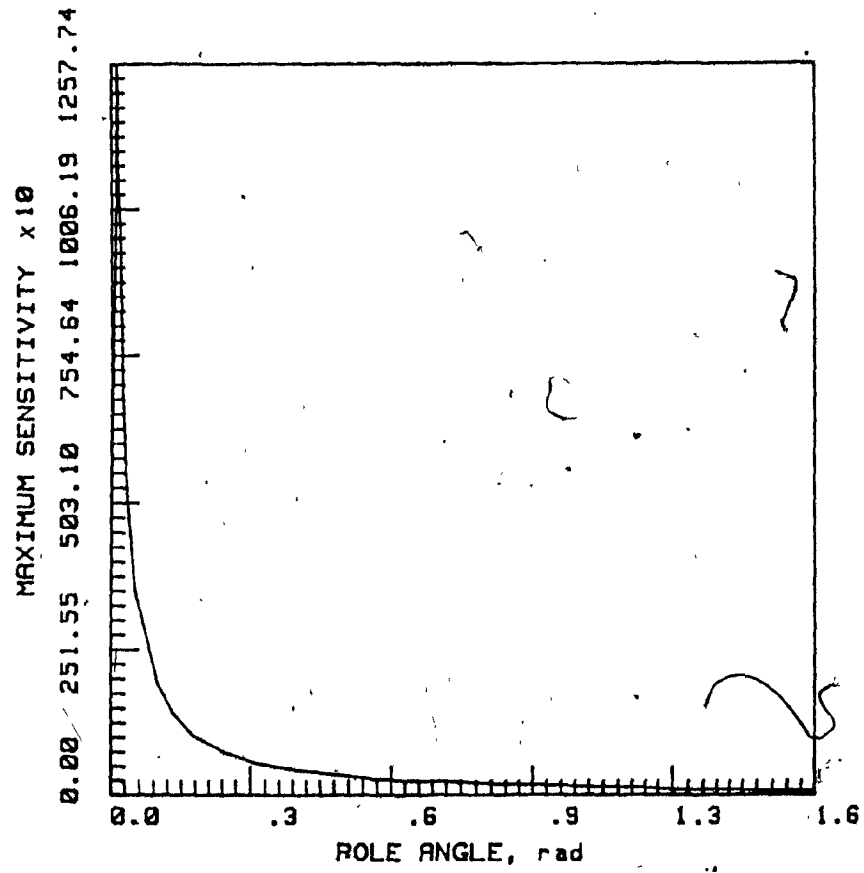




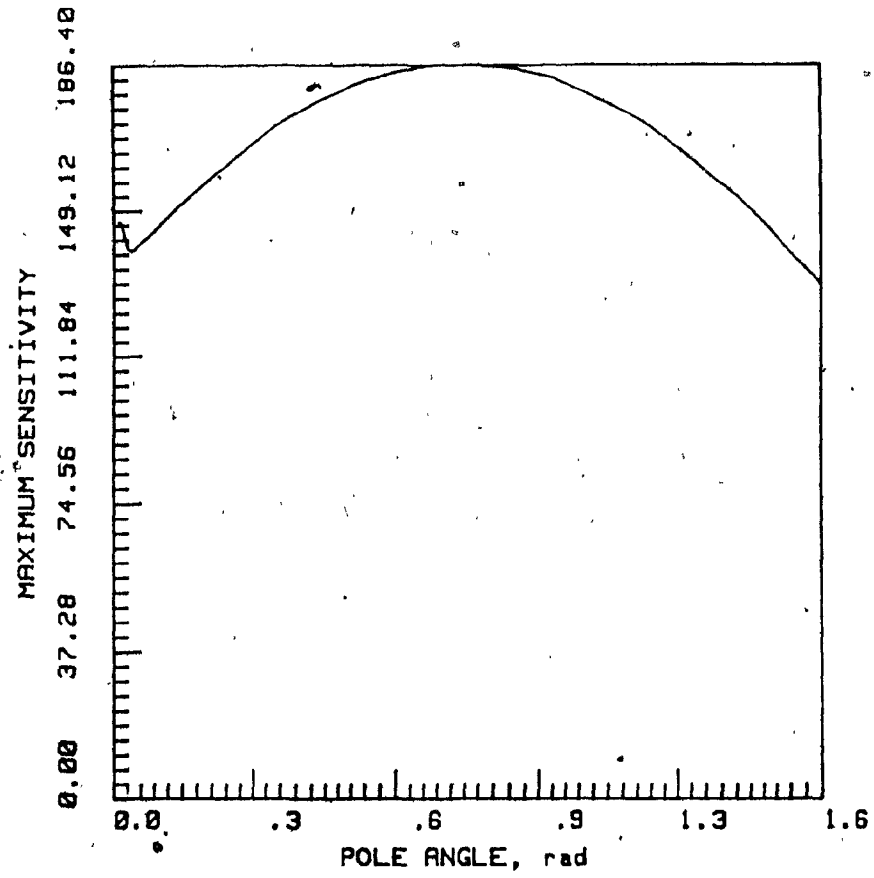
(a) TVGIC



(b) Improved TVGIC



(c) Direct Canonic



(d) Section-Optimal

Fig. 7.2: Maximum Sensitivity  $\hat{S}$  versus  $\omega_0$  with  $\alpha_2 = 0.985$

section-optimal structures are quite similar, while the improved TVGIC structure is the best for pole angles  $\omega_0$  smaller than  $\pi/3$ , as expected.

A factor that seriously affects the performance of the TVGIC structures lies in the fact that its zeros are formed by a linear combination of lowpass, bandpass and highpass transfer functions. As a result, the sensitivity to numerator multiplier coefficients can be as high as four times the sensitivity in the direct canonic structure. For example, if the zeros of a narrowband lowpass elliptic filter are realized by a combination of a lowpass transfer function with zeros at  $z=-1$  and a highpass transfer function with zeros at  $z=1$ , the sensitivity to variation in multiplier  $c_0$  is  $z^2 + 2z + 1$  ( $\cong 4$ , for  $z=1$ ) times a corresponding sensitivity in the direct canonic structure.

For the sake of comparison, coefficient quantization has been applied to the lowpass, bandpass, bandstop and highpass designs based on TVGIC, direct canonic and section-optimal structures (see Tables A.2 to A.5). Coefficients were assumed to be in fixed-point format and the quantization was by means of rounding. Then actual amplitude responses were obtained for the various designs as illustrated in Figs. 7.3 to 7.6. The coefficient wordlengths were assumed to be 11 bits in the lowpass designs, 9 bits in the bandpass designs, 7 bits in the bandstop designs and 8 bits in the highpass designs.

The section-optimal design performed better in the lowpass and highpass filters followed by the TVGIC. The direct canonic design was the best for the bandpass filter. In the bandstop case the section-optimal and direct canonic designs presented similar performance while

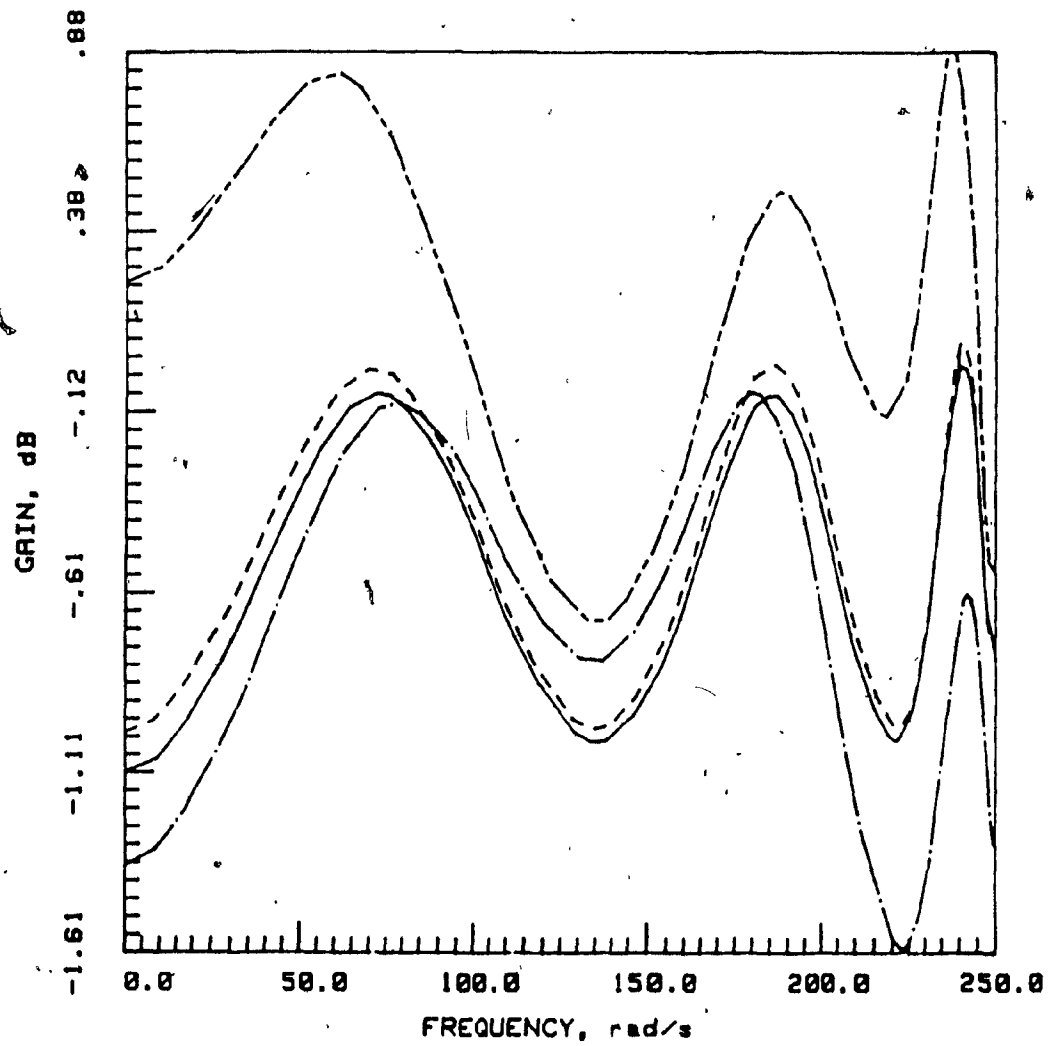


Fig. 7.3: Amplitude Response of the Lowpass Filters.  
(Coefficient Wordlength = 11 bits)

- Ideal
- Direct Canonic
- Section-Optimal
- . - . - TVGIC

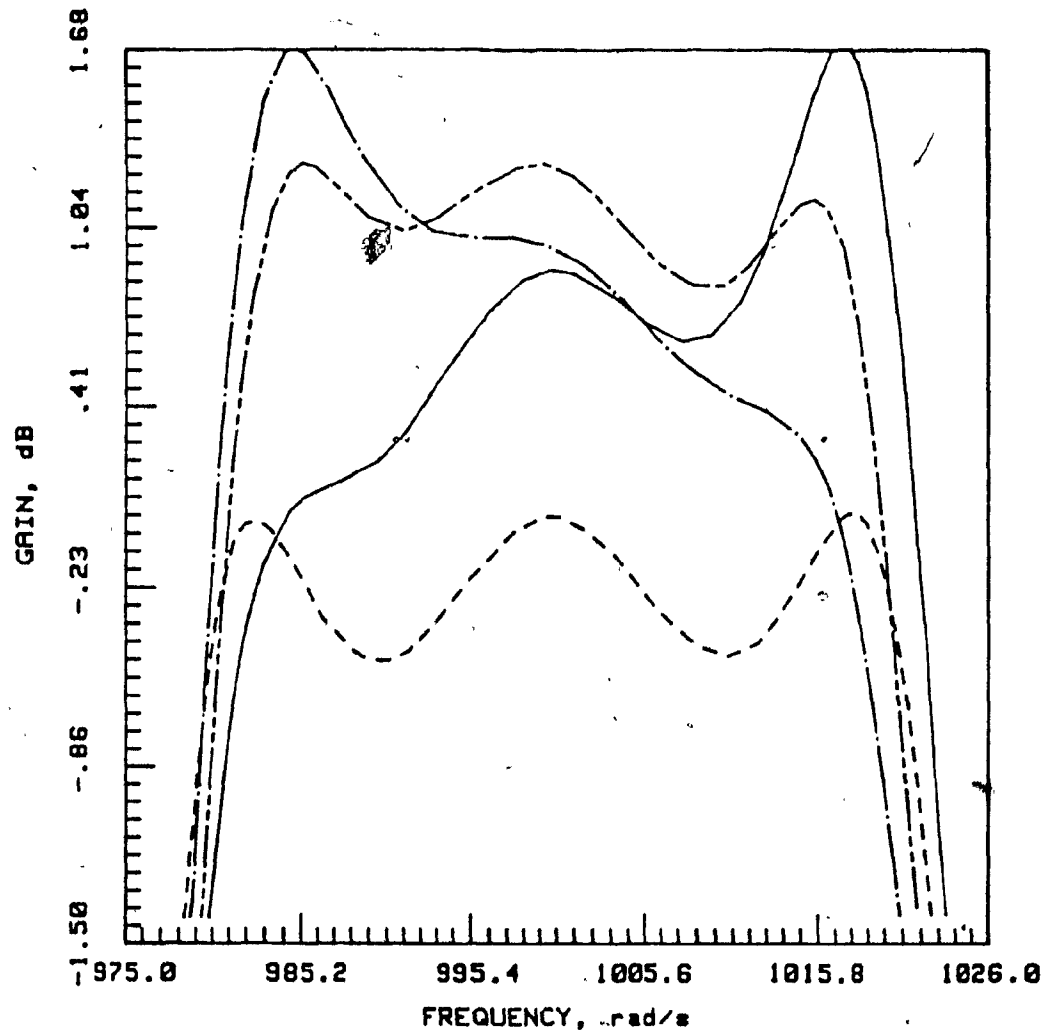


Fig. 7.4: Amplitude Response of the Bandpass Filters  
(Coefficient Wordlength = 9 bits)

- Ideal
- .-.-.-.- Direct Canonic
- Section-Optimal
- TVGIC

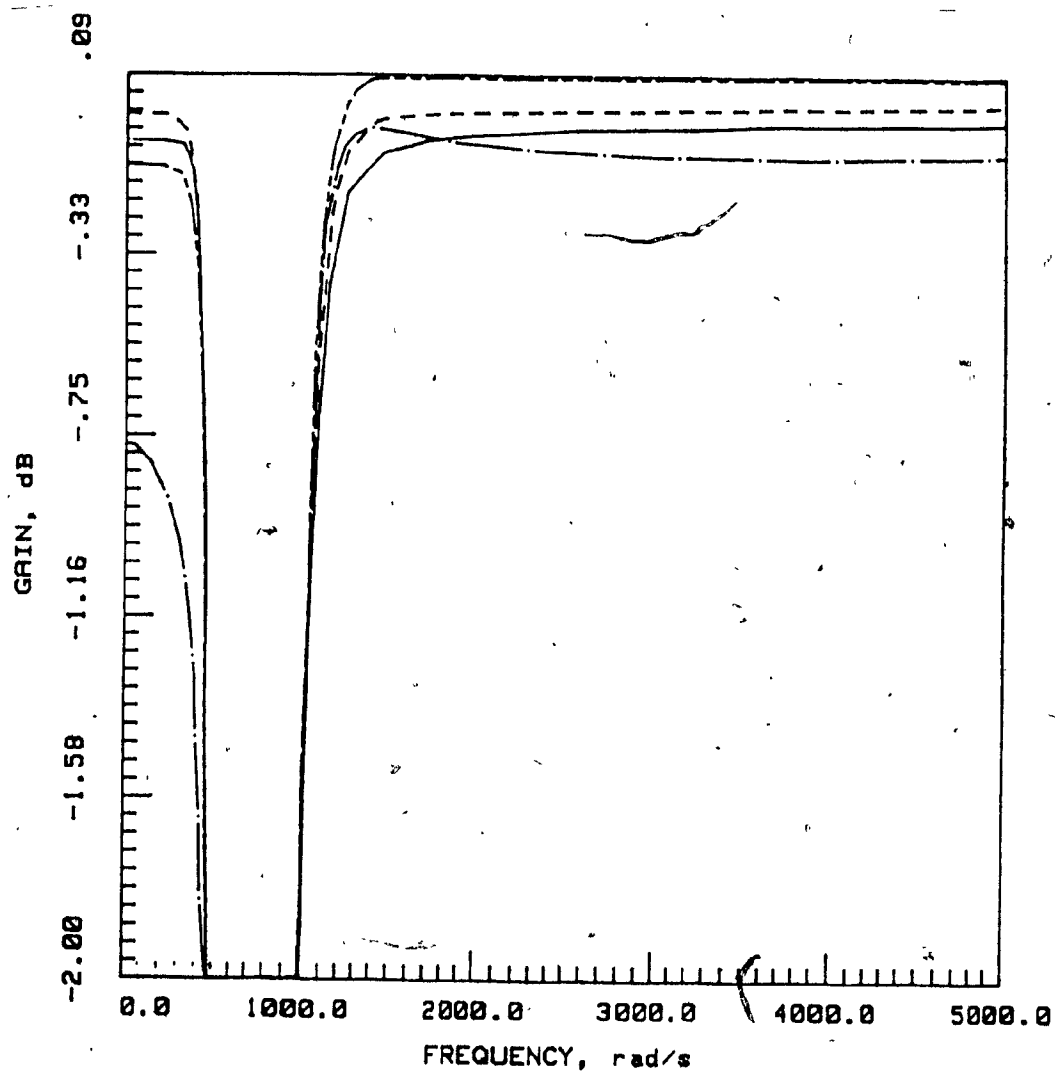


Fig. 7.5: Amplitude Response of the Bandstop Filters  
(Coefficient Wordlength = 7 bits)

- Ideal
- . - . - . Direct Canonic
- \_\_\_\_\_ Section-Optimal
- \_\_\_\_\_ TVGIC



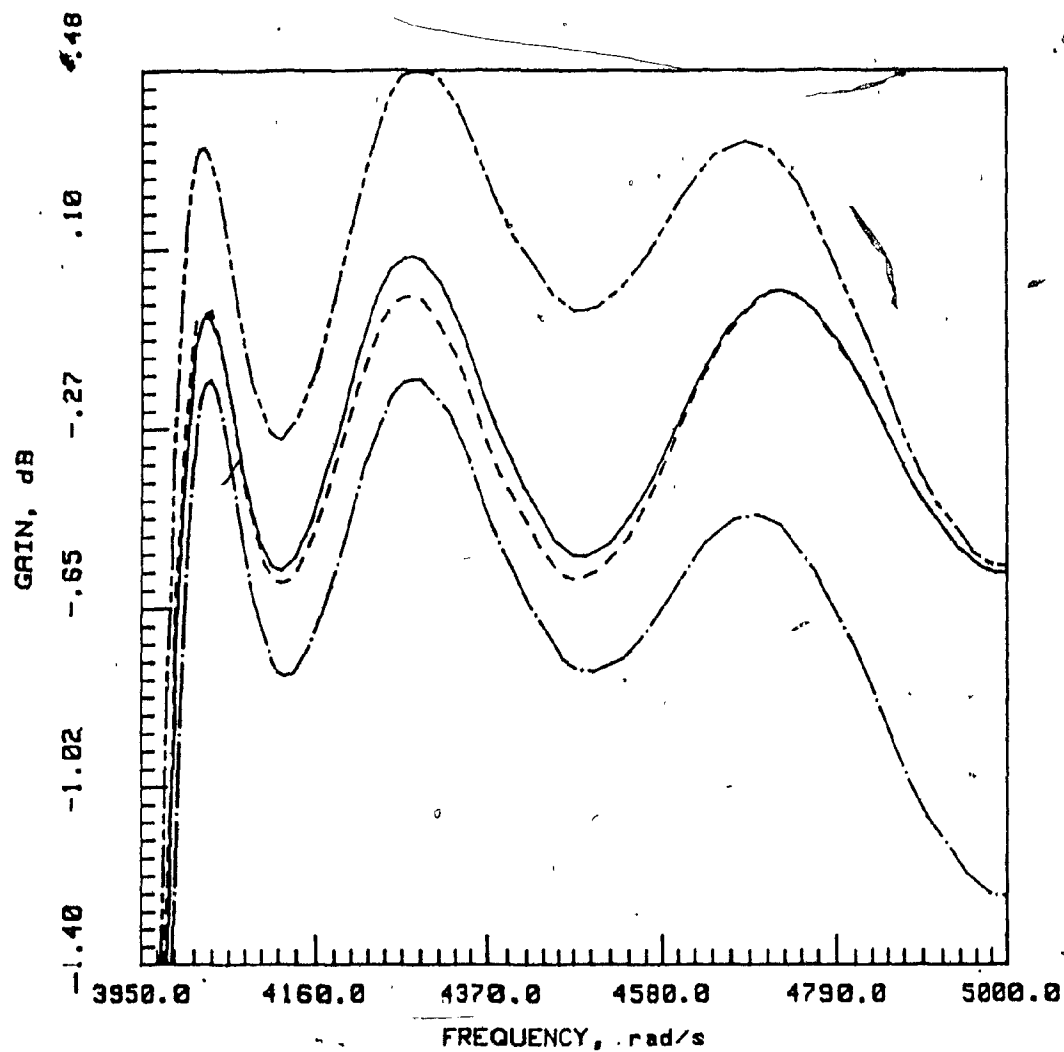


Fig. 7.6: Amplitude Response of the Highpass Filters  
(Coefficient Wordlength = 8 bits)

- Ideal
- Direct Canonic
- Section-Optimal
- . - . - . TVGIC

the TVGIC design presented a poor performance.

Very little improvement was obtained by using the improved TVGIC structure in the lowpass and highpass examples, while a poor performance was obtained in the bandpass and bandstop examples. Examples using the improved TVGIC were not included, since variations in multiplier constants  $c_0$ ,  $c_1$  and  $c_2$  were found to have more influence than  $m_1$ ,  $m_2$ .

The VGIC and TVGIC structures are quite similar, that is they have the same characteristic polynomial and the zeros are formed in the same way. These facts lead us to believe that their sensitivity properties are also very similar.

#### 7.4 Low-Sensitivity Structures of Chapter 3

In Chapter 3 several low-sensitivity structures have been developed in which ESS can efficiently be applied: For each permissible range of denominator coefficient  $\alpha_1$ , at least two distinct structures are possible as can be seen in Table 3.1 and 3.2. By using the maximum sensitivity  $\hat{S}$  as a sensitivity measure, it is possible to identify an optimal subset of these structures.

The maximum sensitivity  $\hat{S}$  is assumed to occur at  $\omega \equiv \omega_0$ , where  $\omega_0$  is the angle of the poles. It can be determined from Eqn. 7.1 by using the numerator polynomials  $S_{m_i}^H(z)$  given in Table 7.2.

For  $-1.5 < \alpha_1 < -0.5$ , the structures to be compared are I-5, I-6, and II-2. Assuming that  $\alpha_2$  is close to unity, Eqns. 7.1 and 7.5, and Table 7.2 yield the sensitivities of these structures as

$$\hat{S}_{I-5} \equiv \frac{N_{I-5}}{|D(e^{j\omega_0 T})|}$$

TABLE 7.2  
Numerator and Denominator Polynomials  
of Sensitivities

Case	$N_{m_1}(z)$	$N_{m_2}(z)$	$D_{m_1}(z)$
I	$m_1(z-D)$	$-m_2$	$D(z)$
II	$m_1(z-B)$	$-m_2Bz$	$D(z)$

$$\hat{S}_{I-6} = \frac{N_{I-6}}{|D(e^{j\omega_0 T})|}$$

$$\hat{S}_{II-2} = \frac{N_{II-2}}{|D(e^{j\omega_0 T})|}$$

where

$$N_{I-5} = 1 - \alpha_2 + |1 + \alpha_1|$$

$$N_{I-6} = |\alpha_1 + \alpha_2| + |1 + \alpha_1| \sqrt{(2 - 2\cos\omega_0)}$$

$$N_{II-2} = |\alpha_1 + \alpha_2| + |1 - \alpha_2| \sqrt{(2 - 2\cos\omega_0)}.$$

For  $-1.5 < \alpha_1 < -1$  and  $\alpha_2 < 1$

$$N_{I-5} = -(\alpha_2 + \alpha_1)$$

$$N_{I-6} = -(\alpha_1 + \alpha_2) - (1 + \alpha_1) \sqrt{(2 - 2\cos\omega_0)}$$

$$N_{II-2} = -(\alpha_1 + \alpha_2) + (1 - \alpha_2) \sqrt{(2 - 2\cos\omega_0)}$$

and hence

$$\hat{S}_{I-5} < \min \{ \hat{S}_{I-6}, \hat{S}_{II-2} \}.$$

For  $-1 < \alpha_1 < -\alpha_2$

$$N_{I-5} = 2 - \alpha_2 + \alpha_1 \quad (7.6)$$

$$N_{I-6} = -(\alpha_1 + \alpha_2) + (1 + \alpha_1) \sqrt{(2 - 2\cos\omega_0)} \quad (7.7)$$

$$N_{II-2} = -(\alpha_1 + \alpha_2) + (1 - \alpha_2) \sqrt{(2 - 2\cos\omega_0)}. \quad (7.8)$$

Evidently,

$$N_{I-6} < N_{II-2}$$

and, in addition

$$N_{I-6} < N_{I-5}$$

as can be shown by adding  $(\alpha_1 + \alpha_2)$  to each of Eqns. 7.6 and 7.7, and then dividing each by  $(1 + \alpha_1)$ . As a consequence

$$\hat{S}_{I-6} < \min \{ \hat{S}_{I-5}, \hat{S}_{II-2} \}.$$

Similarly, for  $-\alpha_2 < \alpha_1 < -0.5$

$$N_{I-5} = 2 - \alpha_2 + \alpha_1$$

$$N_{I-6} = (\alpha_1 + \alpha_2) + (1 - \alpha_1) \sqrt{(2 - 2\cos\omega_0)}$$

$$N_{II-2} = (\alpha_1 + \alpha_2) + (1 - \alpha_2) \sqrt{(2 - 2\cos\omega_0)}$$

and as above we can show that

$$\hat{S}_{II-2} < \min \{ \hat{S}_{I-5}, \hat{S}_{I-6} \}. \quad (7.9)$$

In effect, the choice of structure tends to depend heavily on the relative values of coefficients  $\alpha_1$  and  $\alpha_2$ . For example, if  $\alpha_1 = -0.9$  and  $\alpha_2 = 0.985$ , I-5, I-6, and II-2 are possible structures, and according to Eqn. 7.9 the optimal one is II-2. This result is confirmed by the sensitivity plots of Fig. 7.7.

The above approach has been applied for the remaining range of  $\alpha_1$ . The optimum structures identified are summarized in Table 7.3.

The sensitivity performance of the structures of Chapter 3 is compared with that of section-optimal and direct canonic structures in Figs. 7.8 to 7.11. These plots show actual amplitude responses with the multipliers coefficients quantized. As in previous examples, the coefficients are assumed to be in fixed-point representation and the quantization is assumed to be by rounding.

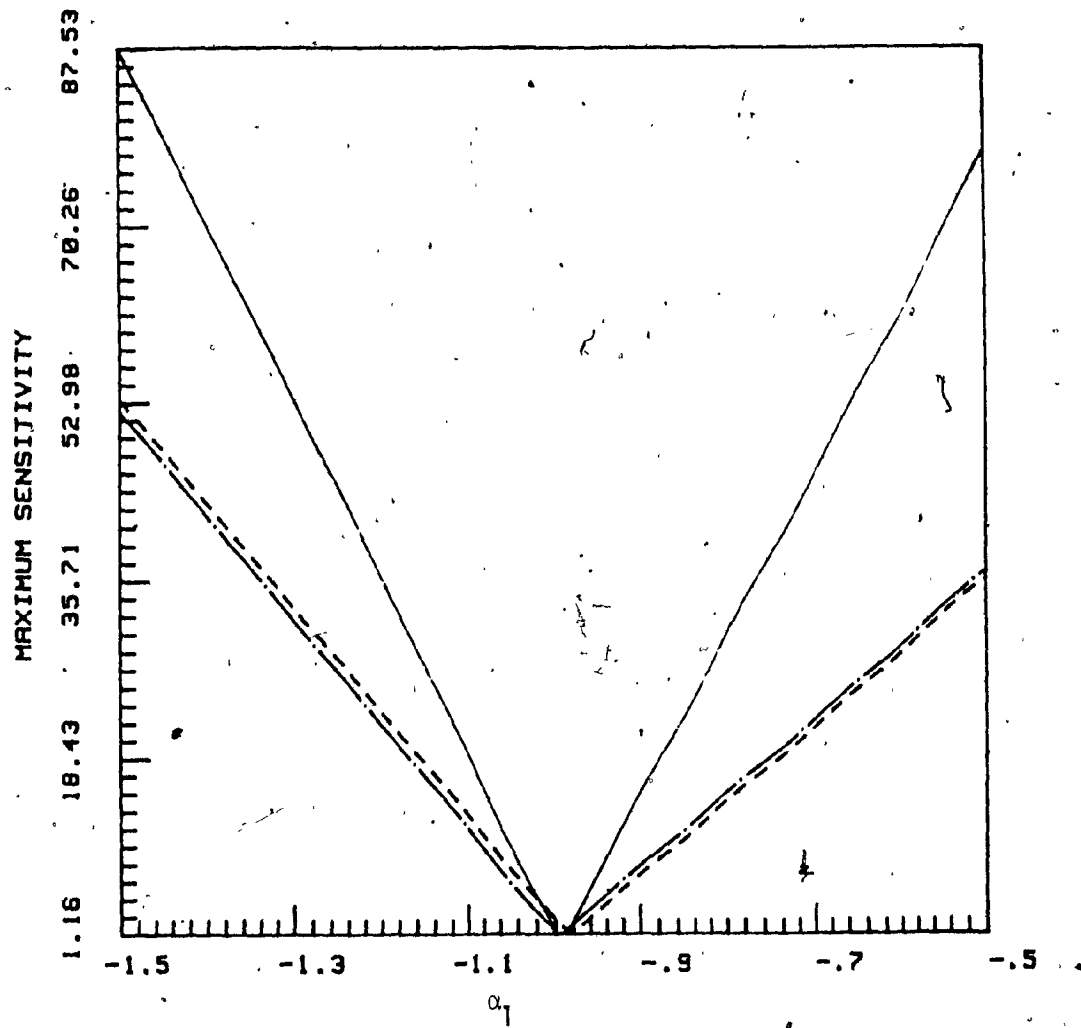


Fig. 7.7: Maximum Sensitivity  $\hat{S}$  versus  $\alpha_1$  ( $\alpha_2=0.985$ )

——— I-5  
 - - - I-6  
 - · - II-2

TABLE 7.3  
Optimum Structures

Range of $\alpha_1$	Structure	Condition for Optimality
$-2 < \alpha_1 < -1.5$	II-1	$-2 < \alpha_1 < -1.5$
$-1.5 \leq \alpha_1 < -0.5$	I-5	$-1.5 \leq \alpha_1 < -1.0$
	I-6	$-1.0 \leq \alpha_1 < -\alpha_2$
	II-2	$-\alpha_2 \leq \alpha_1 < -0.5$
$-0.5 \leq \alpha_1 < 0.5$	I-7	$-0.5 \leq \alpha_1 < 0.5$
$0.5 \leq \alpha_1 < 1.5$	I-10	$1.0 < \alpha_1 < 1.5$
	I-11	$\alpha_2 < \alpha_1 < 1.0$
	II-5	$0.5 \leq \alpha_1 \leq \alpha_2$
$1.5 \leq \alpha_1 < 2$	II-6	$1.5 \leq \alpha_1 < 2$

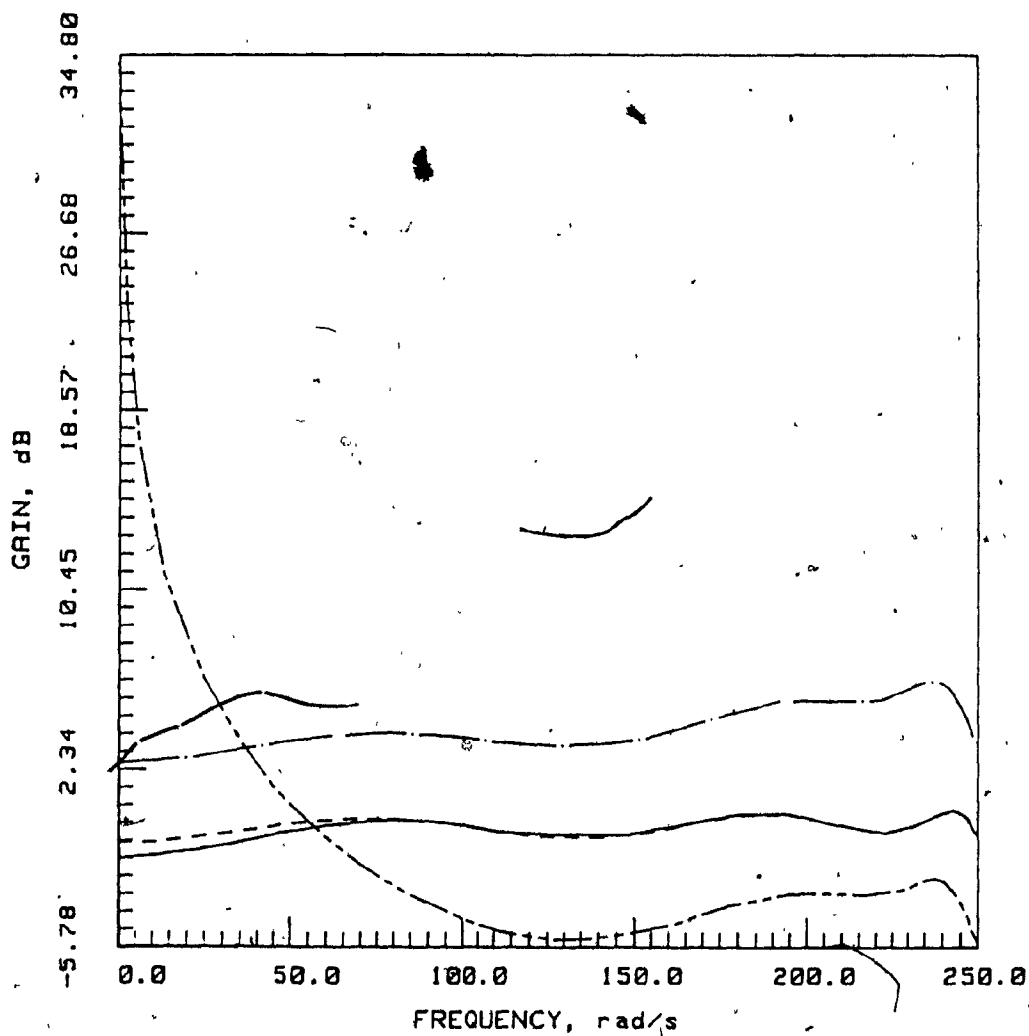


Fig. 7.8: Amplitude Response of the Lowpass Filters  
(Coefficient Wordlength = 8 bits)

- Ideal
- - - - - Direct Canonic
- Section-Optimal
- . - . - II-1



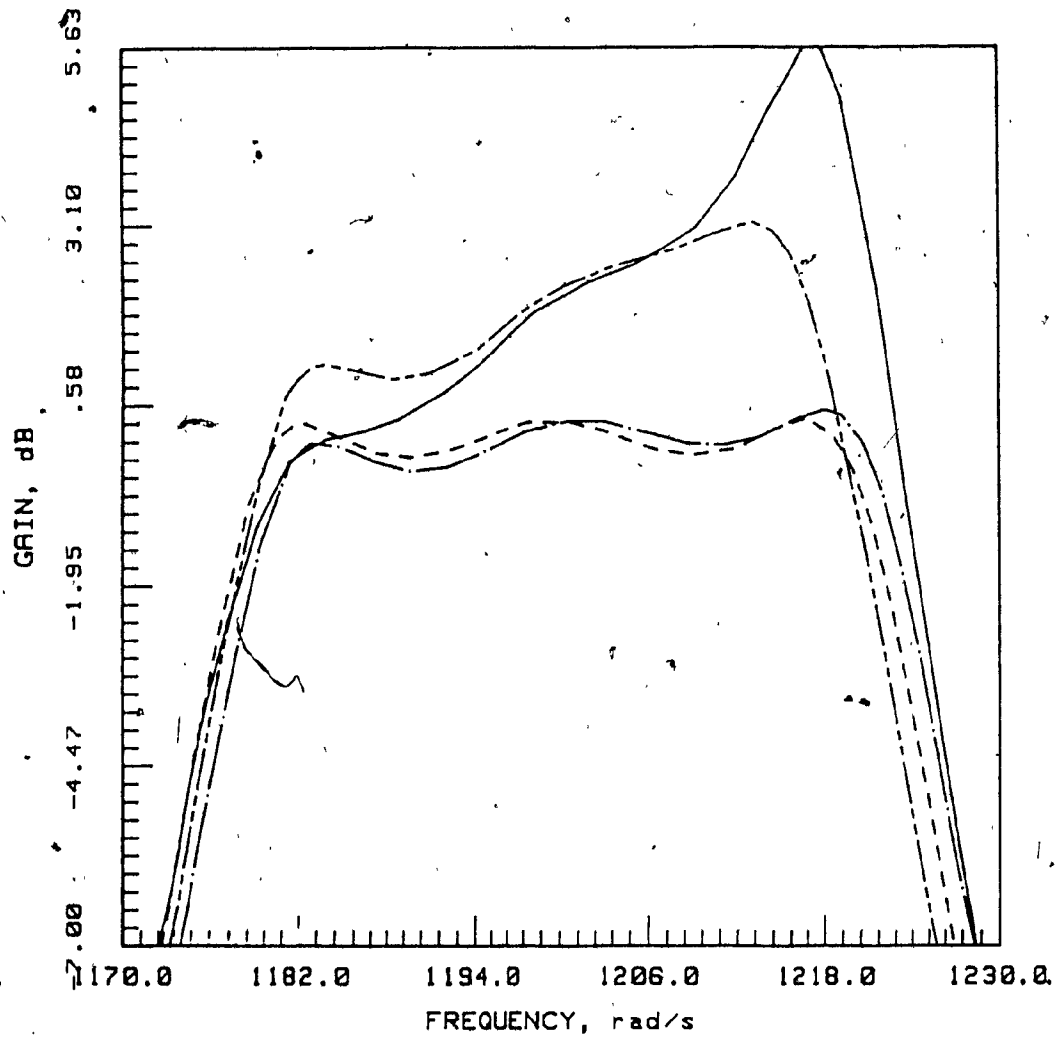


Fig. 7.9: Amplitude Response of the Bandpass Filters  
(Coefficient Wordlength = 8 bits)

- Ideal
- Direct Canonic
- Section-Optimal
- . - . - . I-5

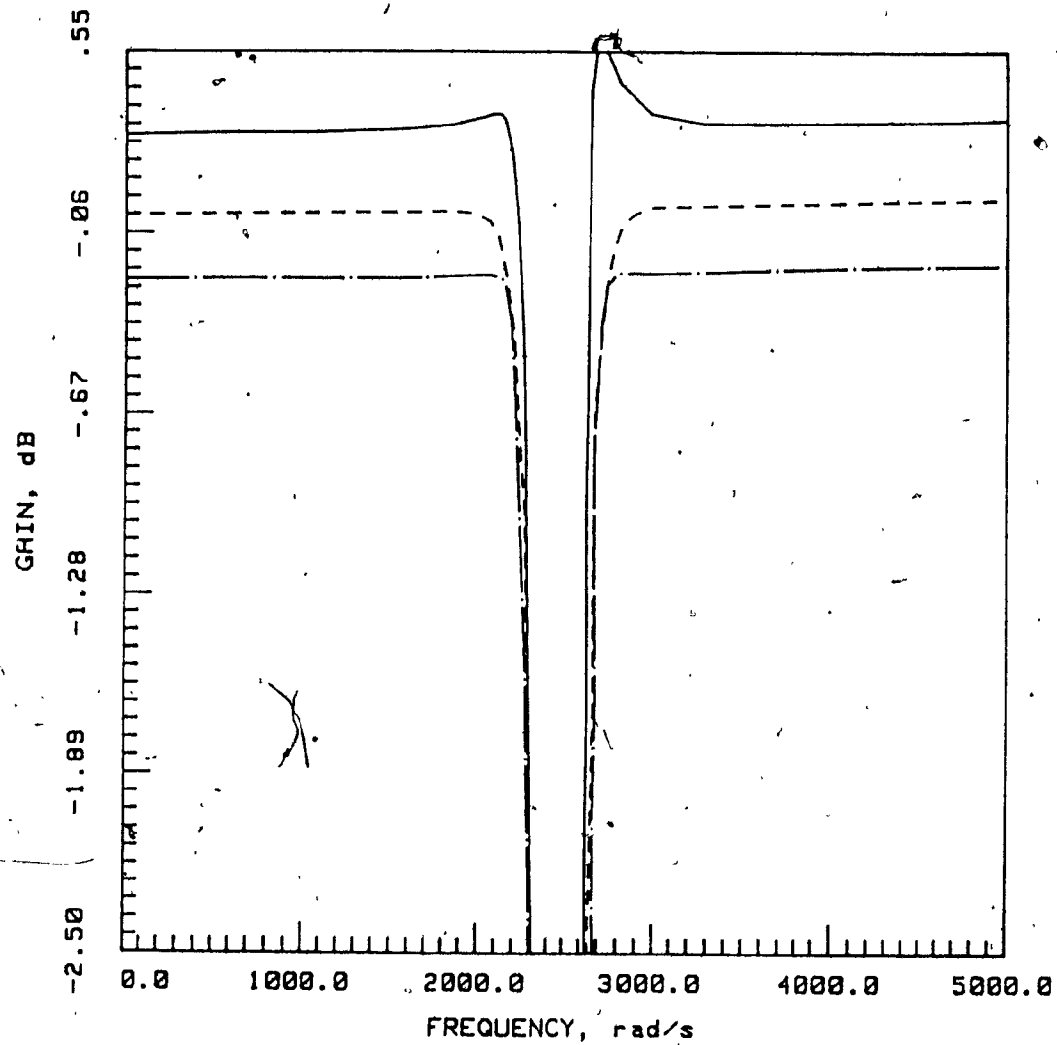


Fig. 7.10: Amplitude Response of the Bandstop Filters  
(Coefficient Wordlength = 8 bits)

- Ideal
- . - . - Direct Canonic and I-7
- \_\_\_\_\_ Section-Optimal

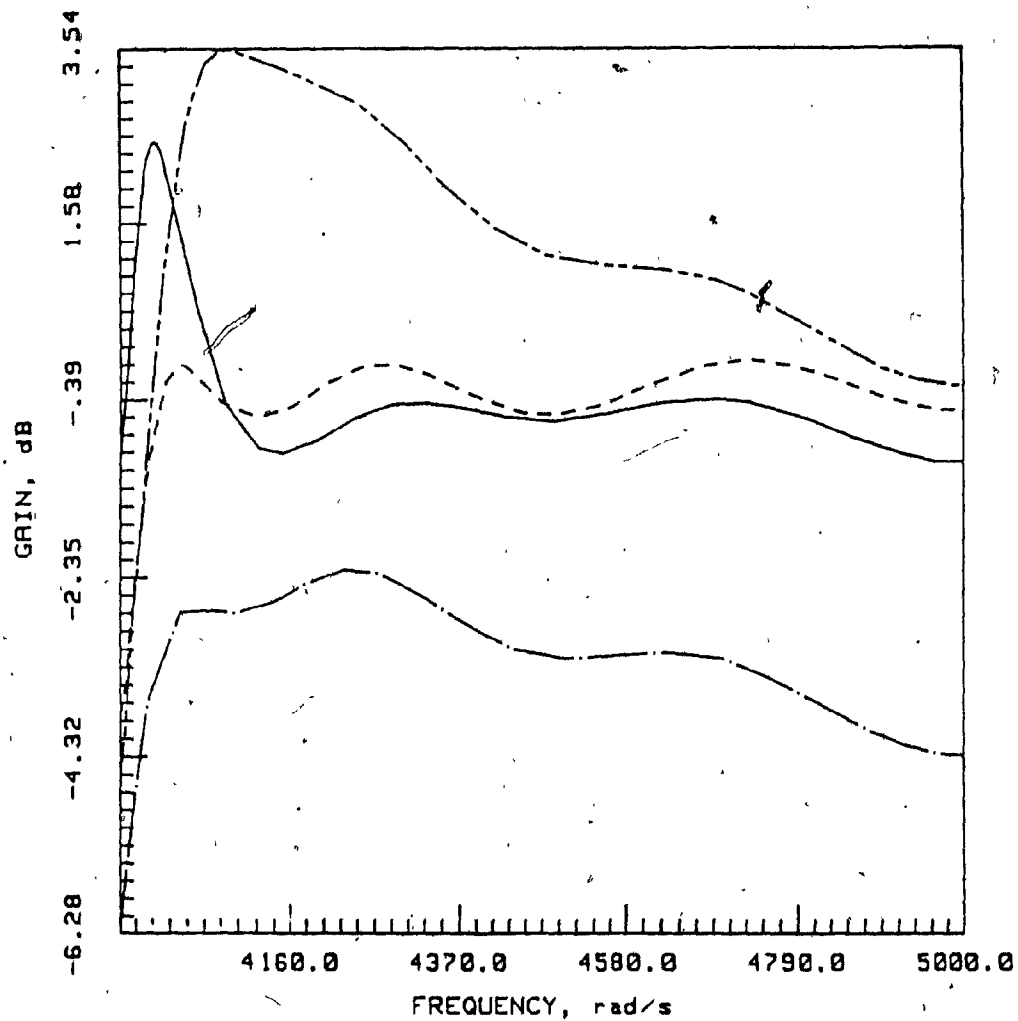


Fig. 7.11: Amplitude Response of the Highpass Filters  
(Coefficient Wordlength = 5 bits)

- Ideal
- · - · - Direct Canonic
- Section-Optimal
- II-6

Fig. 7.8 shows the amplitude responses of the lowpass designs described in Tables 6.2 and A.1. The coefficient wordlength was assumed to be 8 bits in each case. The amplitude response of the section-optimal design is close to the desired response. Similarly, the response of the design based on structure II-1 is close to the desired response except for a shift in the gain.

Fig. 7.9 shows the amplitude responses of the bandpass designs described in Tables 6.5 and A.6. The coefficient wordlength was assumed to be 8 bits. The better performance of the new structure is clearly noted.

Fig. 7.10 shows amplitude responses of the bandstop designs described in Tables 6.5 and A.6. The coefficient wordlength was assumed to be 5 bits. As can be seen structure I-7 and the direct canonic structure yield similar results, while the section-optimal structure is worse. The reason for the good performance of the direct canonic structure in this example can be explained by the fact that the poles of this particular bandstop filter are very close to  $z = \pm j$ , where the sensitivity of the direct canonic structure reaches a minimum value (see Fig. 7.2(c)).

Finally, the amplitude responses of the highpass designs described in Tables 6.2 and A.1 are depicted in Fig. 7.11. The coefficient wordlength was assumed to be 5 bits. The response of the design based on structure II-6 is the closest to the ideal response, except for a shift in the gain, which can easily be corrected.

### 7.5 State-Space Structures

The sensitivity  $S$  of a second-order state-space structure characterized by matrix  $A$  in the form given in Eqn. 5.1 can be obtained as

$$S \cong 2|s_a^{H(z)}| + 2|s_e^{H(z)}| \quad (7.10)$$

The use of this formula gives the maximum sensitivity  $\hat{S}$  as

$$\hat{S} = \{2e^2 + 2a\sqrt{1+a^2-2a\cos\omega_0}\} \frac{1}{\delta\sqrt{(2-\delta)^2+4\cos^2\omega_0(\delta-1)}} \quad (7.11)$$

where

$$\delta = 1 - r,$$

$$e = r \sin \omega_0$$

and

$$a = r \cos \omega_0.$$

Now if the poles of the transfer function are close to the unit circle, i.e.,  $\delta$  is very small, Eqn. 7.11 can be simplified as

$$\hat{S} \cong \frac{1}{\delta} (r^2 \sin \omega_0 + r \cos \omega_0). \quad (7.12)$$

This equation is very accurate when the value of  $\omega_0$  is much greater than  $\delta$  if the poles have positive real parts, or when  $(\pi - \omega_0)$  is much greater than  $\delta$  if the poles have negative real parts. The peak value of  $\hat{S}$  is reached when  $\tan \omega_0 = r$ , i.e., if the poles are close to the unit circle this peak occurs when  $\omega_0 \cong \pi/4$ . On the other hand, when the value of  $\omega_0$  is much smaller than  $\delta$ , the peak value of  $\hat{S}$  can be deduced as

$$\hat{S} \cong \frac{2a}{\delta}. \quad (7.13)$$

If  $\alpha_2 = 0.985$  and  $\omega_0$  is assumed to vary from 0 to 0.16 rad/s, Eqn. 7.11 shows  $\hat{S}$  to remain almost constant with a value of about 132 for  $\omega_0 \gg \delta$  whereas Eqn. 7.12 gives the peak value of  $\hat{S}$  as 265. These results are confirmed in Fig. 7.12 where the exact value of  $\hat{S}$  is plotted versus  $\omega_0$ .

In Figs. 7.13 to 7.16 the new state-space and section-optimal designs are compared on the basis of actual amplitude responses with the coefficients quantized. The filters considered are the lowpass, bandpass, bandstop, and highpass filters described in Tables 6.2 and A.1. The coefficient wordlength was assumed to be 8 bits for the lowpass, 9 bits for the bandpass and 7 bits for the bandstop and highpass designs. As can be seen, the performance of the new state-space structure is very similar to that of the section-optimal structure in all examples.

The new state-space structure is not as insensitive as the low-sensitivity structures of Chapter 3. Nevertheless, it may sometimes be preferred because limit cycles can easily be eliminated and the output roundoff noise can be kept low without the application of ESS.

## 7.6 Conclusions

Several sensitivity aspects pertaining to the VGIC and TVGIC structures of Chapter 2, the low-sensitivity structures of Chapter 3, and the new state-space structure of Chapter 5 have been considered in detail.

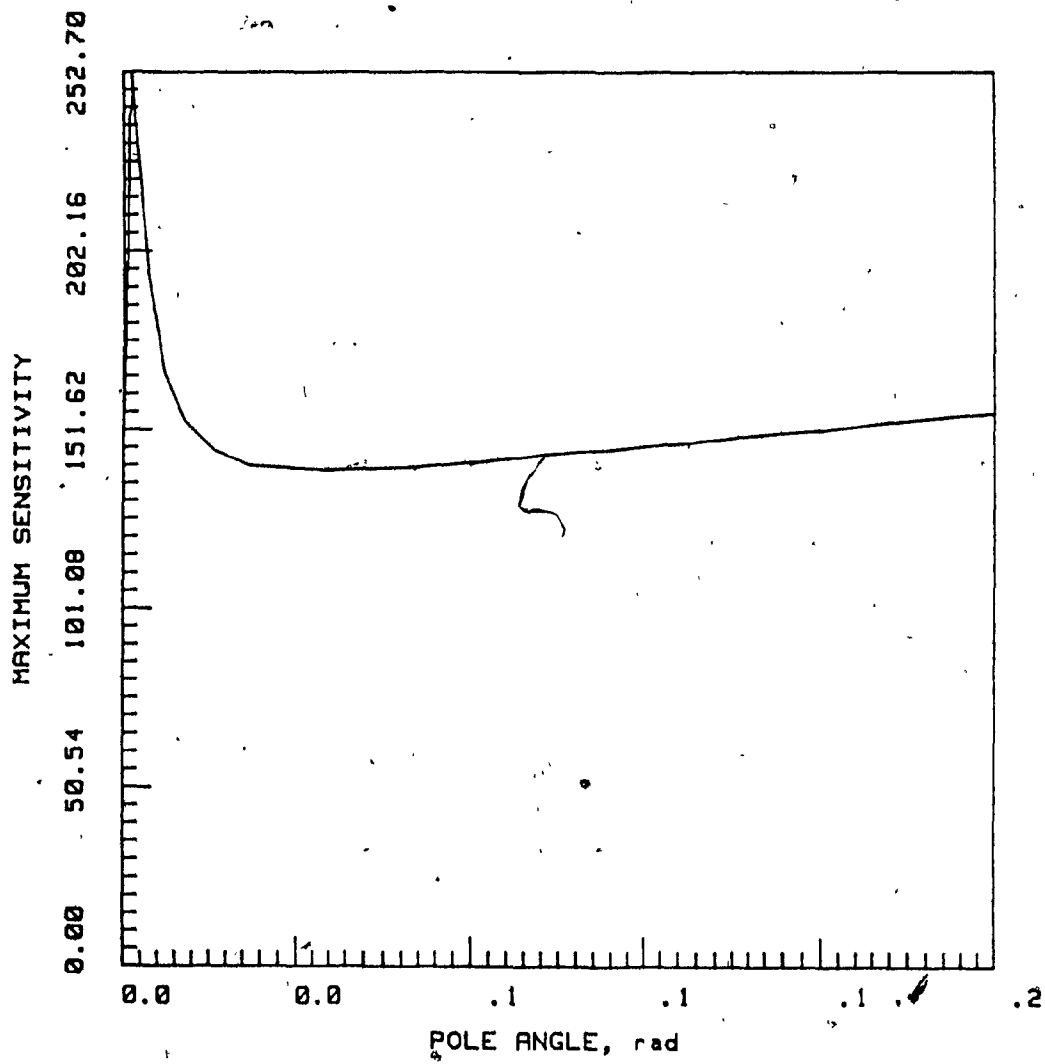


Fig. 7.12:  $\hat{S}$  versus  $\omega_0$  for the State-Space Structure  
( $\alpha_2 = 0.985$ )

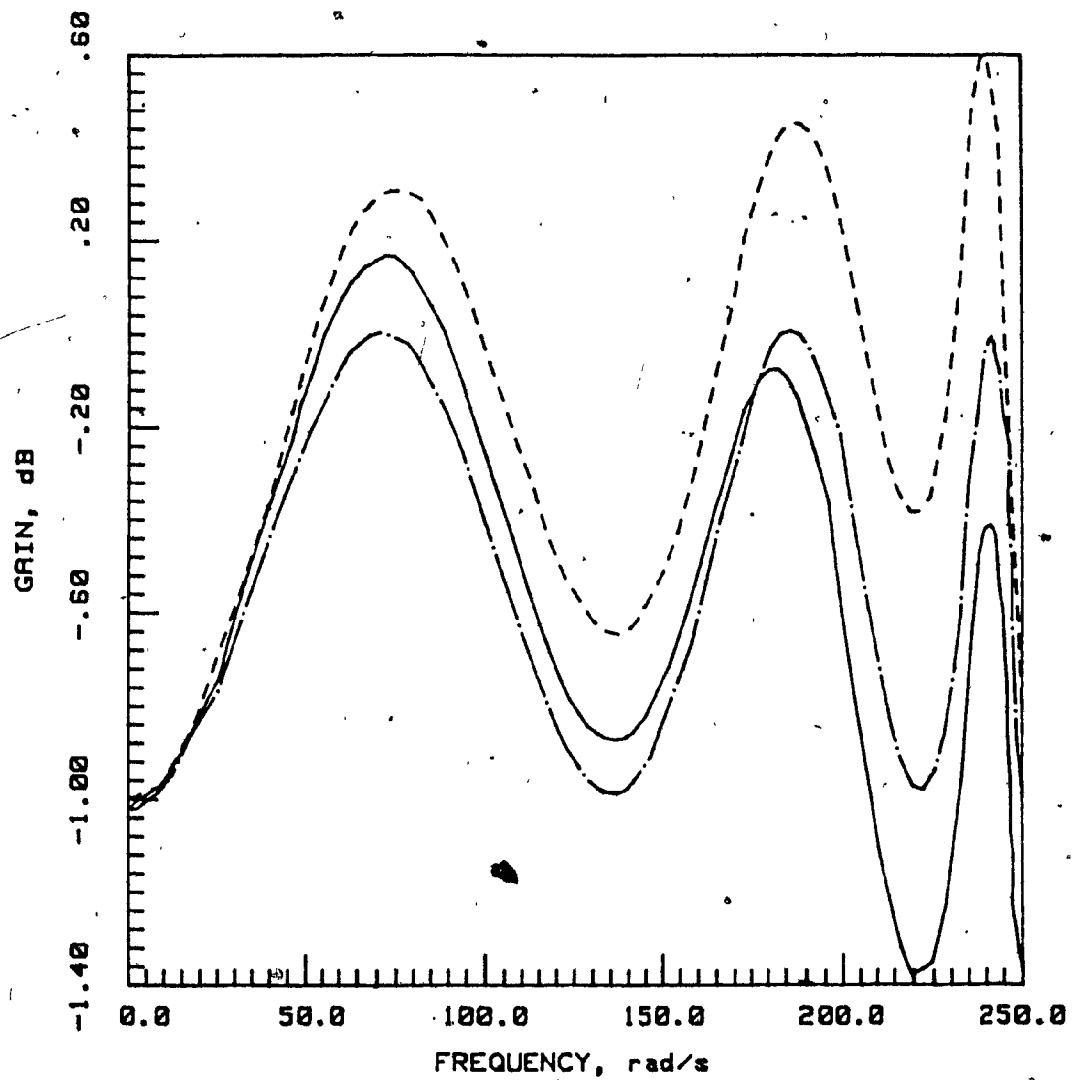


Fig. 7.13: Amplitude Response of the Lowpass Filters  
(Coefficient Wordlength = 8 bits)

- Ideal
- Section-Optimal
- .- New State-Space



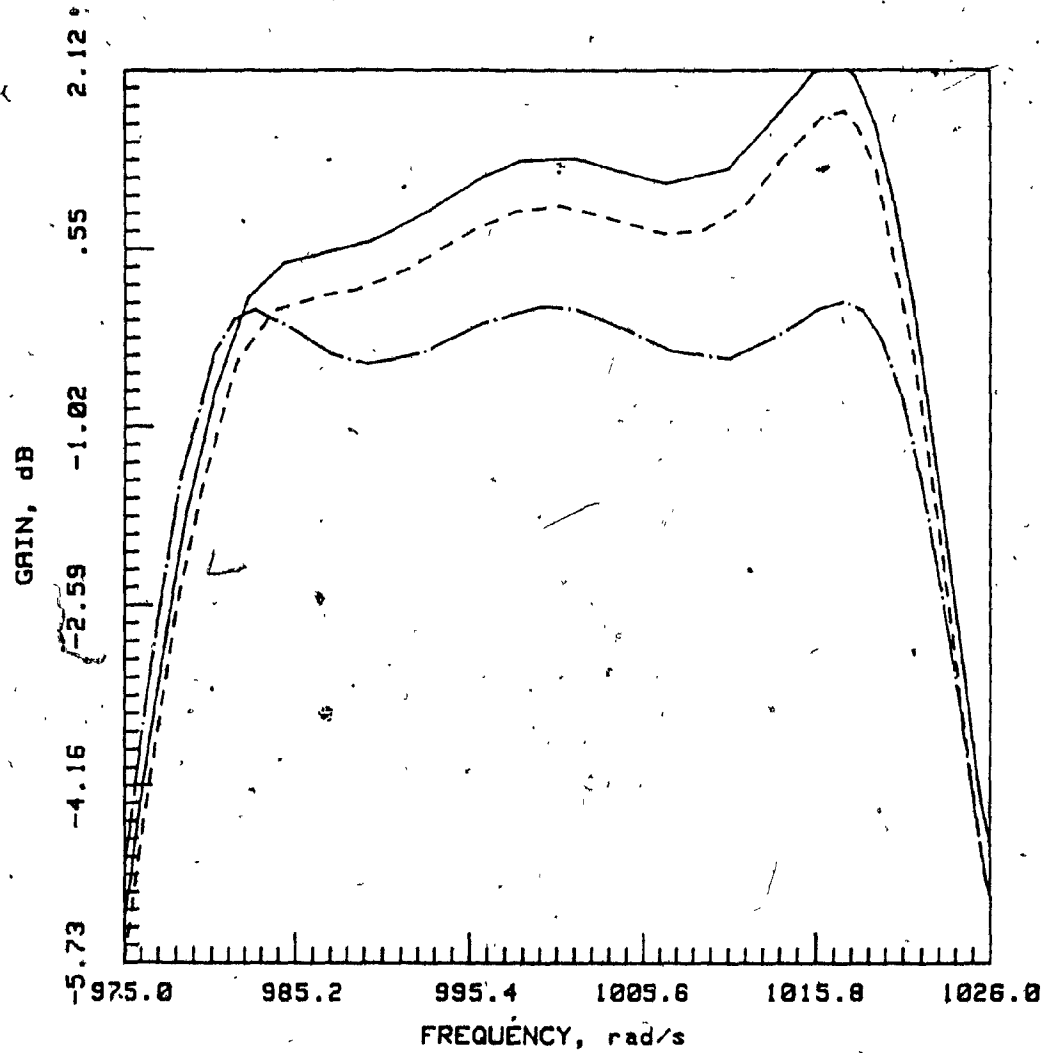


Fig. 7.14: Amplitude Response of the Bandpass Filters  
(Coefficient Wordlength = 9 bits)

— · — · — · — Ideal  
- - - - - Section-Optimal  
———— New State-Space

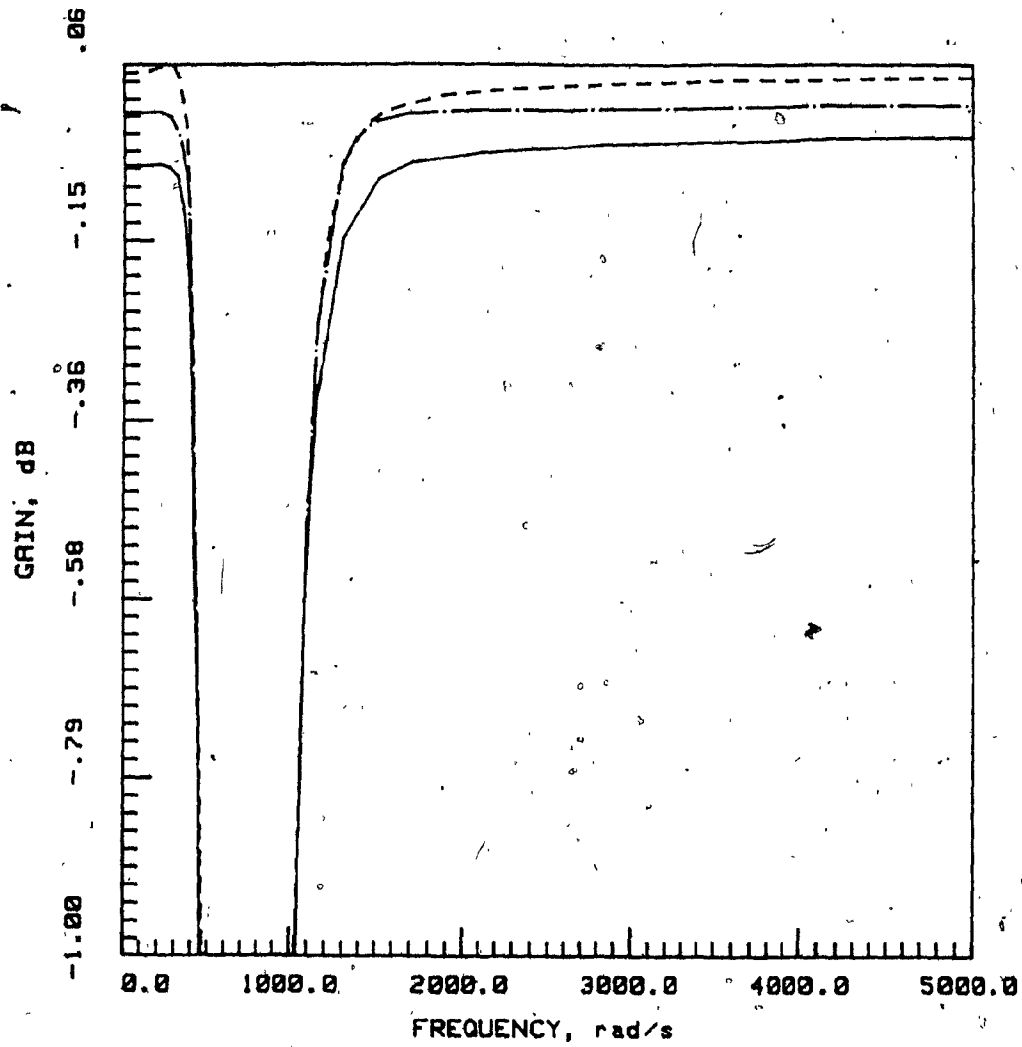


Fig. 7.15: Amplitude Response of the Bandstop Filters  
(Coefficient Wordlength = 7 bits)

— — — — — Ideal  
————— Section-Optimal  
- - - - - New State-Space

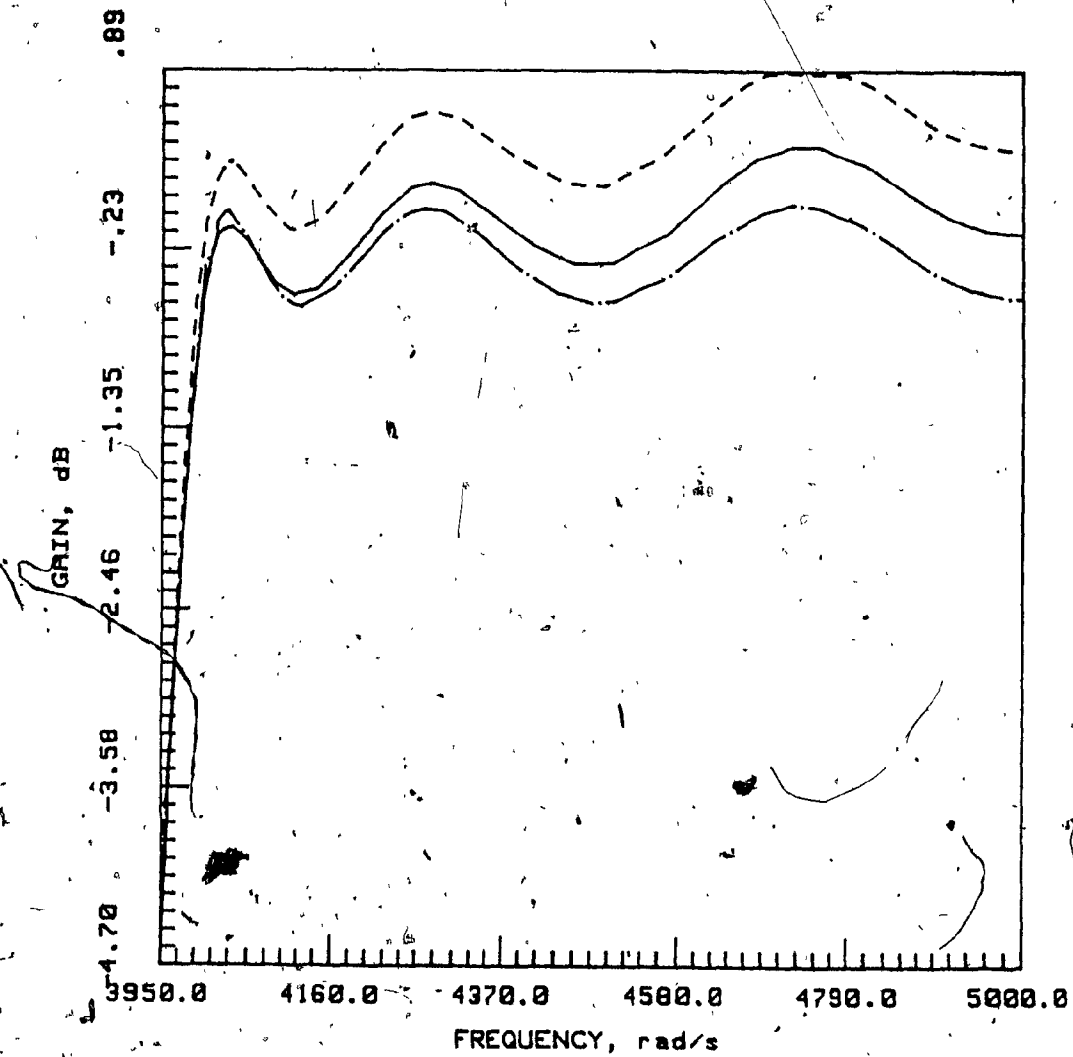


Fig. 7.16: Amplitude Response of the Highpass Filters  
(Coefficient Wordlength = 7 bits)

----- Ideal  
 \_\_\_\_\_ Section-Optimal  
 - . - . - . New State-Space

It has been shown that the technique of Chapter 3 can also be used to reduce the sensitivity in VGIC and TVGIC structures.

By using the maximum sensitivity  $\hat{S}$  as a sensitivity measure, an optimal subset of the low-sensitivity structures of Chapter 3 has been identified. These structures can collectively realize any second-order transfer-function.

Finally, a sensitivity comparison of the various structures has been undertaken. The results show that the low-sensitivity structures of Chapter 3 are almost always superior relative to corresponding direct canonic structures and also relative to the section-optimal structure. The new state-space structure has been shown to have similar performance as compared with the section-optimal structure while the TVGIC structure presented higher sensitivity relative to the remaining structures proposed in this thesis.

## CHAPTER 8

### CONCLUSIONS

#### 8.1 Introduction

The main objective of this thesis has been the generation of high-performance recursive digital filter structures. The emphasis has been placed on generating second-order structures which can be used in cascade or in parallel for the realization of high-order transfer functions. This objective has largely been met and several alternatives have been proposed for the design of low-noise, low-sensitivity, and limit-cycle-free digital filters.

In this chapter, the main contributions of this thesis are summarized and suggestions for future research are pointed out.

#### 8.2 Results of the Thesis

In Chapter 2 new second-order digital-filter structures have been developed by applying the concept of wave characterization to an analog configuration. This configuration realizes a continuous-time biquadratic transfer function by means of voltage-conversion type generalized-immittance converters (VGIC's). The new VGIC-based structures are canonical with respect to the number of multipliers and delays while having a reduced number of adders. In addition, through the use of transposition a digital structure (TVGIC) has been obtained which realizes simultaneously a lowpass, a bandpass and a highpass transfer function. This structure also realizes a transfer function

with zeros on the unit circle using the minimum number of multipliers. In addition, a universal digital biquad was derived which realizes simultaneously all the standard second-order transfer functions.

A special case of the TVGIC structure has been shown to be amenable to the application of error spectrum shaping (ESS). ESS brings about a dramatic reduction in the output roundoff noise.

In Chapter 3 a systematic procedure has been described which can be used for the generation of low-sensitivity digital-filter structures which are amenable to ESS. The procedure was then applied for the generation of several new as well as some known structures.

Chapter 4 showed that zero-input and overflow limit cycles can be eliminated in the VGIC and TVGIC structures. Next a theorem was proved which establishes sufficient conditions that will ensure freedom from constant-input limit cycles in a general digital-filter structure in which zero-input limit cycles can be eliminated. By applying this theorem to the TVGIC structure, to a sub-class of the VGIC structures, and to the universal digital biquad, constant-input limit cycles can efficiently be eliminated.

In Chapter 5 conditions have been derived which lead to the elimination of constant-input limit cycles in second-order state-space structures. From these conditions new and more economical state-space structures were generated. Specifically, a new state-space structure was derived in which the number of multipliers is reduced by two relative to that in the section-optimal structure. A step-by-step procedure was then given for the design of parallel filters.

In Chapter 6 the effect of product quantization in the various types of structures has been considered and several comparisons were undertaken. The VGIC and TVGIC structures were compared with the direct canonic and section-optimal structures, by evaluating the relative power spectral density of the output noise in several designs. The VGIC and TVGIC structures were shown to be comparable to the direct canonic structure in terms of output roundoff noise. The TVGIC structure with second-order ESS was shown to give better results than the section-optimal structure, while the TVGIC structure with first-order ESS was found to be comparable to the section-optimal structure. In addition, the low-sensitivity structures proposed in Chapter 3 have been shown to have improved noise performance as compared to the section-optimal and TVGIC designs. The only type of transfer function in which the section-optimal design outperformed the TVGIC and the low-sensitivity designs is in the bandstop design, because ESS is not effective in filters with wide passband(s). A comparison of the new state-space structure with the conventional section-optimal structure has shown the performance of the two types of structures to be very similar.

In Chapter 7 several sensitivity aspects pertaining to the VGIC and TVGIC structures of Chapter 2, the low-sensitivity structures of Chapter 3, and the new state-space structure of Chapter 4 have been considered in detail. In addition, the techniques of Chapter 3 have been used to reduce the sensitivity in VGIC and TVGIC structures. By using the maximum sensitivity  $\hat{S}$  as a sensitivity measure, an optimal subset of the low-sensitivity structures of Chapter 3 has been

identified. These structures can collectively realize any second-order transfer function. Finally, a sensitivity comparison of the various structures has been undertaken. The results have shown that the low-sensitivity structures of Chapter 3 are almost always superior relative to corresponding direct canonic structures and also relative to the section-optimal structure. The new state-space structure has been shown to have similar performance as the section-optimal structure. The comparison has also shown that the TVGIC structure presented higher sensitivity relative to the other structures proposed in this thesis.

### 8.3 General Comparisons

In the class of medium noise structures, the VGIC and TVGIC structures, have the advantage over the direct canonic structure that zero-input, constant-input and overflow limit cycles can be eliminated. Furthermore, in the TVGIC structure simultaneous realization of several transfer functions is possible. In terms of roundoff noise the VGIC structure is very similar to the direct canonic structure but the TVGIC structure is somewhat noisier. In terms of cost the VGIC and TVGIC structures, like the direct canonic structure, are both very economical.

In the class of low noise structures, the TVGIC and the low-sensitivity structures of Chapter 3 both using ESS present lower output noise as compared with the section-optimal structure, except for bandstop filters. The low-sensitivity structures of Chapter 3



[47] present the best sensitivity and noise performance among all structures presented in this thesis. The new state-space structure presents sensitivity and output noise which are comparable to the corresponding quantities in the section-optimal structure. However, the new structure allows elimination of constant-input limit cycles, and reduces the number of multipliers by two per second-order section. This structure is unique in the sense that it is the only low-noise structure in which all known types of limit cycles can be eliminated.

The structures with ESS are expected to have low-amplitude limit cycles [36] - [37]. Zero-input limit cycles can be eliminated for some choices of the ESS parameters [36], but the optimum noise performance may not be assured.

In terms of cost the TVGIC, the low-sensitivity and state-space structures require fewer multipliers than the section-optimal structure. However, the most economical structure will depend on the type of hardware available for the implementation of the filter.

The properties of the various structures described in this thesis as well as those of the direct canonic and section-optimal structures are summarized in Table 8.1.

#### 8.4 Suggestions for Future Research

Although some significant results have been obtained in this thesis, several outstanding problems remain.

The TVGIC and low-sensitivity structures of Chapter 2 and 3, respectively, seem to be suitable for implementation using

TABLE 8.1

General Comparison

	VGIC	TVGIC	TVGIC with ESS	Low Sensitivity Structures	New State- Space	Direct Canonic	Section- Optimal
noise	medium	medium	low	low	low	medium	low
sensitivity	medium	medium	medium	low	medium	medium	medium
Elimination of zero-input limit cycles	yes	yes	no, expected low amplitude bound	no, expected low amplitude bound	yes	no	yes
Elimination of constant- input limit cycles	no	yes	no	no	yes	no	no
Simultaneous realization of different types of transfer functions	no	yes	yes	no	no	no	no
Maximum number of multipliers	5	5	7	7	7	5	9

ROM-accumulator and stored-product techniques such as those described in [40] - [42]. The applicability and implementation of these techniques should be investigated.

By using different adaptors in the VGIC structure proposed in Chapter 2, new second-order structures can be obtained which would have the desirable features presented by the proposed VGIC structures. The generation of these structures is an interesting topic for future research.

The limit-cycle behavior of the TVGIC structures of Chapter 2 and the low-sensitivity structures of Chapter 3 should be investigated for the case where ESS is incorporated. Although limit cycles may occur in these structures, the amplitude of the limit cycles is expected to be low and presumably it can be reduced or eliminated by choosing the ESS coefficients properly.

The generation procedure of low-sensitivity structures which are amenable to ESS as presented in Chapter 3 can be applied for the generation of low-sensitivity structures of higher-order, or to generate second-order structures having more than three nodes in their recursive part. These problems should be investigated further.

Another topic for future research would be the application of the internal scaling procedures described in [19], [20] and [41], to the structures proposed in Chapter 3. This method can be used as an alternative to ESS in order to reduce the output noise.

The sensitivity performance of the structures of Chapter 3 when the filters are implemented by using floating-point arithmetic should be investigated. It seems that more significant improvements in the sensitivity performance can be obtained in this case.

The theorem of Chapter 4 for the elimination of constant-input limit cycles should be applied to some other important structures in which zero-input limit cycles can be eliminated such as lattices and orthogonal structures.

Another interesting topic for future investigation is the design of  $n$ th-order state-space structures having a sparse matrix  $\underline{A}$ , such that zero-input limit cycles are eliminated. In this approach, it might be possible to obtain a machine representable vector  $\underline{p}$  by forcing vector  $\underline{B}$  to have a desirable form. Then it might be possible to choose vector  $\underline{C}$  so as to place the zeros of the transfer function at desired positions. Since the choice of matrix  $\underline{A}$  would allow some free parameters, it should be possible to use an optimization algorithm to determine matrix  $\underline{A}$  such that zero-input limit cycles are eliminated and the output noise is reduced at the same time.

REFERENCES

- [1] Jackson, L.B., "An Analysis of Roundoff Noise in Digital Filters", Dr. Sc. Thesis, Stevens Institute of Technology, Hoboken, New Jersey, 1969.
- [2] Antoniou, A., "Digital Filters: Analysis and Design", MacGraw-Hill Co., New York, 1979.
- [3] Oppenheim, A.V., and Schafer, R.W., "Digital Signal Processing", Prentice-Hall Book Co., 1975.
- [4] Butterweck, H.J., "Suppression of Parasitic Oscillations in Second-Order Digital Filters by Means of a Controlled-Rounding Arithmetic", Archiv für Elektronik und Übertragungstechnik, Vol. 29, pp. 371-374, September 1975.
- [5] Büttner, M., "Elimination of Limit Cycles in Digital Filters with Very Low Increase in the Quantization Noise", IEEE Trans. on Circuits and Systems, Vol. CAS-24, pp. 300-304, June 1977.
- [6] Willems, J.L., "Stability Theory of Dynamical Systems", Thomas Nelson and Sons LTD., London, 1970.
- [7] Fettweis, A., and Meerkötter, K., "Suppression of Parasitic Oscillations in Wave Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-22, pp. 239-246, March 1975.
- [8] Meerkötter, K., "Realization of Limit Cycle-Free Second-Order Digital Filters", Proc. IEEE Intern. Symp. on Circuits and Systems, pp. 295-298, 1976.
- [9] Meerkötter, K., and Wegener, W., "A new Second-Order Digital Filter Without Parasitic Oscillations", Archiv für Elektronik und Übertragungstechnik, Vol. 29, pp. 312-314, July/August 1975.

- [10] Verkroost, G., and Butterweck, H.J., "Suppression of Parasitic Oscillations in Wave Digital Filters and Related Structures by Means of Controlled Rounding"; Archiv für Elektronik und Übertragungstechnik, Vol. 30, pp. 181-186, May 1976.
- [11] Verkroost, G., "A General Second-Order Digital Filter with Controlled Rounding to Exclude Limit Cycles for Constant Input Signals", IEEE Trans. on Circuits and Systems, Vol. CAS-24, pp. 428-431, August 1977.
- [12] Liu, K.S., and Turner, L.E., "Stability Dynamic Range and Roundoff Noise in a New Second-Order Recursive Digital Filter", IEEE Trans. on Circuits and Systems, Vol. CAS-30, pp. 815-821, November 1983.
- [13] Claasen, T.A.C.M., Mecklenbraüker, W.F.G., and Peek, J.B.H., "On the Stability of the Forced Response of Digital Filters with Overflow Nonlinearities", IEEE Trans. on Circuits and Systems, Vol. CAS-22, pp. 692-696, August 1976.
- [14] Jackson, L.B., "Roundoff-Noise Analysis for Fixed-Point Digital Filters Realized in Cascade or Parallel Form", IEEE Trans. on Audio and Electroacoust., Vol. AU-18, pp. 107-122, June 1970.
- [15] Antoniou, A., and Rezk, M.G., "Digital Filter Synthesis Using Concept of Generalized-Immitance Convertor", IEE J. Electron. Circuits Syst., Vol. 1, pp. 207-216, November 1977. (see Vol. 2, p. 88, May 1978, for errata).
- [16] Antoniou, A., and Rezk, M.G., "A Comparison of Cascade and Wave Fixed-Point Digital-Filter Structures", IEEE Trans. on Circuits and Systems, Vol. CAS-27, pp. 1184-1194, December 1980.

- [17] Fettweis, A., "Digital-Filter Structures Related to Classical Filter Networks", Archiv für Elektronik und Übertragungstechnik, Vol. 25, pp. 79-89, February 1971.
- [18] Sedlmeyer, A., and Fettweis, A., "Digital Filters with True Ladder Configuration", Int. J. Circuit Theory and Appl., Vol. 1, pp. 5-10, March 1973.
- [19] Agarwal, R.C., and Burrus, C.S., "New Recursive Digital Filter Structures Having Very Low Sensitivity and Roundoff Noise", IEEE Trans. on Circuits and Systems, Vol. CAS-22, pp. 921-927, December 1975.
- [20] Nishimura, S., Hirano, K., and Pal, R.N., "A New Class of Very Low Sensitivity and Low Roundoff Noise Recursive Digital Filter Structures", IEEE Trans. on Circuits and Systems, Vol. CAS-28, pp. 1152-1158, December 1981.
- [21] Jackson, L.B., Lindgren, A.G., and Kim Y., "Optimal Synthesis of Second-Order State-Space Structures for Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 149-153, March 1979.
- [22] Fettweis, A., and Meerkötter, K., "On Adaptors for Wave Digital Filters", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, pp. 516-525, December 1975.
- [23] Mullis, C.T., and Roberts, R.A., "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-23, pp. 551-562, September 1976.

- [24] Mullis, C.T., and Roberts, R.A., "Roundoff Noise in Digital Filters: Frequency Transformations and Invariants", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. CAS-24, pp. 538-550, December 1976.
- [25] Hwang, S.Y., "Minimum Uncorrelated Unit Noise in State-Space Digital Filtering", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, pp. 273-281, August 1977.
- [26] Mills, W.L., Mullis, C.T., and Roberts, R.A., "Low Roundoff Noise and Normal Realizations of Fixed Point IIR Digital Filters", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-29, pp. 893-903, August 1981.
- [27] Barnes, C.W., "Roundoff Noise and Overflow in Normal Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 154-159, March 1979.
- [28] Mills, W.L., Mullis, C.T., and Roberts, R.A., "Digital Filter Realizations Without Overflow Oscillations", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, pp. 334-338, August 1978.
- [29] Jackson, L.B., "Limit Cycles in State-Space Structures for Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-26, pp. 67-68, January 1979.
- [30] Thong, T., and Liu, B., "Error Spectrum Shaping in Narrow-Band Recursive Filters", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-25, pp. 200-203, April 1977.
- [31] Chang, T.L., "Error Spectrum Shaping Structures for Digital Filters", Proc. 13th Asilomar Conf. Circuits Syst. Comput., Pacific Grove, CA, pp. 279-283, November 1979.



- [32] Munson, D.C., and Liu, B., "Narrow-Band Recursive Filters with Error Spectrum Shaping", IEEE Trans. on Circuits and Systems, Vol. CAS-28, pp. 160-163, February 1981.
- [33] Higgins, W.E., and Munson, D.C., "Noise Reduction Strategies for Digital Filters: Error Spectrum Shaping Versus the Optimal Linear State-Space Formulation", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-30, pp. 963-973, December 1982.
- [34] Higgins, W.E., and Munson, D.C., "Optimal Error Spectrum Shaping for Cascade-Form Digital Filters", Proc. IEEE Intern. Symp. on Circuits and Systems, pp. 1029-1032, 1982.
- [35] Abu-El-Haija, A.I., "Determining Coefficients of Error-Feedback Digital Filters to Obtain Minimum Roundoff Errors with Minimum Complexity", Proc. IEEE Intern. Symp. on Circuits and Systems, pp. 819-822, 1983.
- [36] Chang, T.L., "Suppression of Limit Cycles in Digital Filters Designed with One Magnitude-Truncation Quantizer", IEEE Trans. on Circuits and System, Vol. CAS-28, pp. 107-111, February 1981.
- [37] Bateman, M.R., and Liu, B., "Limit Cycle Bounds for Digital Filters with Error Spectrum Shaping", Proc. Asilomar Conf. Circuits Syst. Comput., Pacific Grove, CA, pp. 215-218, 1980.
- [38] Antoniou, A., "Realization of Gyration Using Operational Amplifiers and Their Use in RC-Active Network Synthesis", Proc. IEE, Vol. 116, pp. 1838-1850, November 1969.

- [39] Haug, K., and Lüder, E., "Determination of All Equivalent and Canonic Second Order Digital Filter Structures", Archiv für Elektronik und Übertragungstechnik, Vol. 36, pp. 436-442, November/December 1982.
- [40] Munson, D.C., and Liu, B., "ROM/ACC Realization of Digital Filters for Poles Near the Unit Circle", IEEE Trans. on Circuits and Systems, Vol. CAS-27, pp. 147-151, February 1980.
- [41] Munson, D.C., and Liu, B., "Low-Noise Realization for Narrow-Band Recursive Digital Filters", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, pp. 41-54, February 1980.
- [42] Monkewich, O., and Steenaart, W., "Stored Product Digital Filtering With Nonlinear Quantization", Proc. IEEE Intern. Symp. on Circuits and Systems, pp. 157-160, 1976.
- [43] Diniz, P.S.R., and Antoniou, A., "On the Elimination of Constant-Input Limit Cycles in Digital Filters", IEEE Trans. on Circuits and Systems, Vol. CAS-31, pp. 670-671, July 1984.
- [44] Turner, L.E., and Bruton, L.T., "Elimination of Limit Cycles in Recursive Digital Filters Using a Generalized Minimum Norm", IEEE Proc. Intern. Symp. on Circuits and Systems, pp. 817-820, 1981.
- [45] Higgins, W.E., and Munson, D.C., "A Section Ordering Strategy for Cascade-Form Digital Filters Using Error Spectrum Shaping", Proc. IEEE Intern. Symp. on Circuits and Systems, pp. 835-838, 1983.

- [46] Kim, Y., "State-Space Structures for Digital Filters", Ph.D. Thesis, University of Rhode Island, Kingston, Rhode Island, 1980.
- [47] Diniz, P.S.R., and Antoniou, A., "Low-Sensitivity Digital-Filter Structures which are Amenable to Error-Spectrum Shaping", Proc. Intern. Conference on Digital Signal Processing, Florence, 1984.

APPENDIX A  
DIGITAL-FILTER DESIGNS

This appendix includes several tables which give the transfer function coefficients, the multiplier constants, the error-spectrum shaping constants, and the scaling constants for the various filter designs considered in Chapter 6 and 7.

TABLE A.1  
Transfer Function Coefficients  
of Filters Described in Table 6.2

	j	$\gamma_{1j}$	$\gamma_{2j}$	$\alpha_{1j}$	$\alpha_{2j}$
Lowpass	1	-1.3586005	1.0	-1.918829	0.9225768
	2	-1.8893746	1.0	-1.936830	0.9518912
	3	-1.9337187	1.0	-1.960636	0.9849357
	$H_0 = 2.5992172 \times 10^{-4}$				
Bandpass	1	0.0	-1.0	-1.605156	0.9841072
	2	-1.4798344	1.0	-1.596076	0.9920225
	3	-1.7224138	1.0	-1.626727	0.9923016
	$H_0 = 1.391269 \times 10^{-4}$				
Bandstop	1	-1.8042262	1.0	-1.566901	0.7369230
	2	-1.8042264	1.0	-1.512319	0.8278297
	3	-1.8042257	1.0	-1.792957	0.8963944
	$H_0 = 0.73949244$				
Highpass	1	-2.0	1.0	1.648927	0.7011856
	2	-2.0	1.0	1.575104	0.7808699
	3	-2.0	1.0	1.546503	0.9172177
	$H_0 = 5.8128711 \times 10^{-5}$				

TABLE A.2  
Multiplier Coefficients for Lowpass Design

A.2 (a) VGIC Structure of Fig. 2.6

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.00753	1.94436	.00170	.00000	.05978
2	.00187	1.92070	.00194	.00000	.01015
3	.01215	1.97279	.00582	.00000	.34515
$\lambda_1 = 1.0$					

A.2 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.00753	1.94436	1.16390	.00000	40.92057
2	.00187	1.92070	.32069	.00000	1.67926
3	.01215	1.97279	1.48963	.00000	88.40768
$\lambda_1 = 3.43522 \times 10^{-8}$					

A.2 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.01215	1.97279	.00385	.00000	.22843
2	.00753	1.94436	.00676	.00000	.23764
3	.00187	1.92070	1.01547	.00000	5.31738
ESS Parameters					
j	$b_{1j}$		$b_{0j}$		
1	-1.98749		.99616		
2	-1.98423		.98854		
3	-1.41480		.41841		
$\lambda_1 = 7.22898 \times 10^{-4}$					

A.2 (d) TVGIC Structure with First-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.00753	1.94436	.00153	.00000	.05365
2	.01215	1.97279	.00817	.00000	.48488
3	.00187	1.92070	1.01547	.00000	5.31738
ESS Parameters					
j	$b_{1j}$				
1	-.99566				
2	-.99628				
3	-.99746				
$\lambda_1 = 1.50859 \times 10^{-3}$					

A.2 (e) Direct Canonic Structure

j	$m_{1j}$	$m_{2j}$	$f_{0j}$	$c_{1j}$	$c_{2j}$
1	-1.93683	.95189	.07864	-.14858	.07864
2	-1.91883	.92258	.02443	-.03319	.02443
3	-1.96064	.98494	44.69215	-86.42205	44.69215
$\lambda_1 = 3.0722 \times 10^{-3}$					



A.2 (f) Section-Optimal Structure

j	$A_{\sim j}$		$B_{\sim j}$	$C_{\sim j}$	$D_{\sim j}$
1	.96842	-.11195	.04757	.03067	.0614832
	.12562	.96842	.00151	.96870	
2	.95941	-.02870	.10382	.03262	.0120918
	.07320	.95941	.00346	.97885	
3	.98032	-.15979	.10297	.04570	.349620
	.14965	.98032	.00499	.94292	

TABLE A.3

Multiplier Coefficients for Bandpass Design

A.3 (a) VGIC Structure of Fig. 2.6

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.18948	1.79463	.00000	.00795	.00000
2	.19797	1.79405	.00938	.00000	.06277
3	.18279	1.80951	.01684	.00000	.22580
$\lambda_1 = 1.0$					

A.3 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.18948	1.79463	.00000	.02036	.00000
2	.18279	1.80951	.09501	.00000	1.27407
3	.19797	1.79405	3.65186	.00000	24.43041
$\lambda_1 = 1.777347 \times 10^{-4}$					

A.3 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.19797	1.79405	.02379	.00000	.15912
2	.18279	1.80951	.02441	.00000	.32733
3	.18948	1.79463	.00000	1.61805	.00000
ESS Parameters					
j	$b_{1j}$	$b_{0j}$			
1	-1.61757	.99978			
2	-1.60762	.99975			
3	-1.58832	.97845			
$\lambda_1 = 1.336472 \times 10^{-3}$					

A.3 (d) TVGIC Structure with First-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.18279	1.80951	.01908	.00000	.25586
2	.19797	1.79405	.03313	.00000	.22166
3	.18948	1.79463	.00000	1.61805	.00000
ESS Parameters					
j	$b_{1j}$				
1	-.80887				
2	-.81376				
3	-.80281				
$\lambda_1 = 1.227436 \times 10^{-3}$					

A.3 (e) Direct Canonic Structure

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	-1.60516	.98411	.00555	.00000	-.00555
2	-1.59608	.99202	.18367	-.27180	.18367
3	-1.62673	.99230	17.17636	-29.58480	17.17636
$\lambda_1 = 7.946004 \times 10^{-3}$					

A.3 (f) Section-Optimal Structure

j	$A_j$		$B_j$	$C_j$	$D_j$
1	.80258	-.59107	.00721	.88394	.0079464
	.57518	.80258	.01416	.45038	
2	.79804	-.59633	.01424	.29455	.072156
	.59557	.79804	.00442	.94881	
3	.81336	-.57522	.01304	-.88822	.242643
	.57498	.81332	-.03782	.30695	

TABLE A.4  
Multiplier Coefficients for Bandstop Design

A.4 (a) VGIC Structure of Fig. 2.6

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.08501	1.65191	.04251	.00000	.82596
2	.05172	1.84468	.02098	.00900	.40767
3	.15776	1.67007	.09722	.00000	1.88923
$\lambda_1 = 1.0$					

A.4 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.08501	1.65191	.38363	.00000	7.45455
2	.15776	1.67007	.63930	.00000	12.42269
3	.05172	1.84468	1.55069	.00000	30.13252
$\lambda_1 = 2.279712 \times 10^{-4}$					

A.4 (c) TVGIC Structure with Second-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.08501	1.65191	.06052	.00000	1.17600
2	.15776	1.67007	.03306	.00000	.64249
3	.05172	1.84468	1.55069	.00000	30.13252
ESS Parameters					
j	$b_{1j}$	$b_{0j}$			
1	.10504	.07280			
2	.03443	.04585			
3	-.20511	-.23729			
$\lambda_1 = 2.794126 \times 10^{-2}$					

A.4 (d) TVGIC Structure with First-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	.08501	1.65191	.06052	.00000	1.17600
2	.15776	1.67007	.03306	.00000	.64249
3	.05172	1.84468	1.55069	.00000	30.13252
ESS Parameters					
j	$b_{1j}$				
1	.09791				
2	.03292				
3	.26892				
$\lambda_1 = 2.794126 \times 10^{-2}$					



A.4 (e) Direct Canonic Structure

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	-1.56690	.73692	1.31591	-2.37419	1.31591
2	-1.51232	.82783	.62338	-1.12471	.62338
3	-1.79296	.89639	15.50610	-27.97651	15.50610
$\lambda_1 = 5.813693 \times 10^{-2}$					

A.4 (f) Section-Optimal Structure

j	$A_j$		$B_j$	$C_j$	$D_j$
1	.78345	-.24320	.21968	-.46912	.868461
	.50628	.78345	-.13844	.74441	
2	.75616	-.47955	.15194	-.68389	.711927
	.53395	.75616	-.22147	.46918	
3	.89648	-.30535	.48871	-.01379	1.19605
	.30366	.89648	-.00894	.75385	

TABLE A.5  
Multiplier Coefficients for Highpass Design

A.5 (a) VGIC Structure of Fig. 2.6

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	1.67799	.10288	.00000	.00000	.02617
2	1.67506	.02613	.00000	.00000	.02515
3	1.73186	.18536	.00000	.00000	.08833
$\lambda_1 = 1.0$					

A.5 (b) Limit-Cycle-Free TVGIC Structure of Fig. 2.8

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	1.67799	.10288	.00000	.00000	.01501
2	1.67506	.02613	.00000	.00000	.05100
3	1.73186	.18536	.00000	.00000	.99999
$\lambda_1 = 7.593438 \times 10^{-2}$					

A.5 (c) TVGIC Structure With Second-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	1.67799	.10288	.00000	.00000	.02604
2	1.73186	.18536	.00000	.00000	.08529
3	1.67506	.02613	.00000	.00000	.99999
ESS Parameters					
j	$b_{1j}$	$b_{0j}$			
1	1.80314	.94108			
2	1.78231	.91163			
3	1.84261	.88821			
$\lambda_1 = 2.617315 \times 10^{-2}$					

A.5 (d) TVGIC Structure With First-Order ESS of Fig. 2.10

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	1.67799	.10288	.00000	.00000	.02604
2	1.73186	.18536	.00000	.00000	.08529
3	1.67506	.02613	.00000	.00000	.99999
ESS Parameters					
j	$b_{1j}$				
1	.92894				
2	.93235				
3	.97585				
$\lambda_1 = 2.617315 \times 10^{-2}$					

A.5 (e) Direct Canonic Structure

j	$m_{1j}$	$m_{2j}$	$c_{0j}$	$c_{1j}$	$c_{2j}$
1	1.57510	.78087	.02515	-.05029	.02515
2	1.64893	.70119	.08833	-.17667	.08833
3	1.54650	.91722	.99999	-1.99998	.99999
$\lambda_1 = 2.616669 \times 10^{-2}$					

A.5 (f) Section-Optimal Structure

j	$A_j$		$B_j$	$C_j$	$D_j$
1	<del>-.78755</del> .41582	-.38630	.21927	-.21334 .91714	.0261701
2	-.82446 .24012	-.08931	.38151 -.05021	-.12025 .91369	.0251457
3	-.77325 .55052	-.58000	.54475 -.16912	-.28754 .92617	.0883327

TABLE A.6  
Transfer Functions Coefficients  
of Filters Described in Table 6.5

	j	$\gamma_{1j}$	$\gamma_{2j}$	$\alpha_{1j}$	$\alpha_{2j}$
Bandpass	1	0.	-1.0	-1.446678	0.9842731
	2	-1.2999169	1.0	-1.434637	0.9921319
	3	-1.5851062	1.0	-1.469855	0.9923436
	$H_0 = 1.35997 \times 10^{-4}$				
Bandstop	1	$8.68097 \times 10^{-8}$	1.0	$7.78194 \times 10^{-8}$	.7928745
	2	$8.68096 \times 10^{-8}$	1.0	.1886429	.8918090
	3	$8.68096 \times 10^{-8}$	1.0	-.1886427	.8918090
	$H_0 = 0.7940989$				

TABLE A.7  
Multiplier Coefficients for  
Lowpass Design (Structure II-1)

j	$m_{1j}$	$m_{2j}$	$\delta_{0j}$	$\delta_{1j}$	$\delta_{2j}$
1	-.04811	.01506	.05503	-0.5503	.00609
2	-.01506	.02430	.49306	-.49306	.03268
3	-.07742	.00375	1.59225	-1.59225	1.02127
ESS Parameters:      2 <sup>nd</sup> -order      1 <sup>st</sup> -order					
j	$b_{1j}$	$b_{0j}$	$b_{1j}$		
1	-1.98749	.99616	-.99566		
2	-1.95175	.95904	-.99628		
3	-1.41480	.41841	-.99746		
$\lambda_1 = 6.016012 \times 10^{-3}$					

TABLE A.8  
Multiplier Coefficients for Bandpass  
Design Described in Table 6.5

A.8 (a) Structure I-5

j	$m_{1j}$	$m_{2j}$	$\delta_{0j}$	$\delta_{1j}$	$\delta_{2j}$
1	.43464	-.00787	.18980	-.24673	.18980
2	.46986	-.00766	.17331	-.27471	.17331
3	.44668	-.01573	.75627	.00000	-.75627
ESS Parameters      2 <sup>nd</sup> order      1 <sup>st</sup> order					
j	$b_{1j}$		$b_{0j}$	$b'_{1j}$	
1	-1.45780		.99979	-.72898	
2	-1.46903		.99966	-.73464	
3	-1.43550		.98412	-.72354	
$\lambda_1 = 5.466735 \times 10^{-3}$					

A.9 (b) Section-Optimal Structure

j	$\tilde{A}_j$		$\tilde{B}_j$	$\tilde{C}_j$	$\tilde{D}_j$
1	.72334	-.68692	.00625	.91043	.00786488
	.67119	.72334	.01444	.39406	
2	.71732	-.69113	.01388	.35468	.0730954
	.69102	.71732	.00529	.93139	
3	.73493	-.67266	.01511	-.90244	.236563
	.67229	.73493	-.03631	.37548	



TABLE A.9

Multiplier Coefficients for Bandstop  
Design Described in Table 6.5

A.9 (a) Structure I-7

j	$m_{1j}$	$m_{2j}$	$\delta_{0j}$	$\delta_{1j}$	$\delta_{2j}$
1	.00000	-.20713	.72349	.00000	.72349
2	-.18864	-.10819	1.57708	.00000	1.57708
3	.18864	-.10819	4.75400	.00000	4.75400
ESS Parameters:                  2 <sup>nd</sup> order                                  1 <sup>st</sup> order					
j	$b_{1j}$	$b_{0j}$	$b_{1j}$		
1	.00000	-.08179	.00000		
2	.00000	-.02242	.00000		
3	-.19637	.09540	-.17927		
$\lambda_1 = .1463953$					

A.9 (b) Section-Optimal Structure

j	$A_j$		$B_j$	$C_j$	$D_j$
1	.00000	5.07820	.20713	.00000	.896437
	1.00000	.00000	.00000	.89644	
2	-.09432	-.92357	.14338	-.39449	.599687
	.95598	-.09432	-.07529	.75130	
3	.09432	-.83717	.38308	.36371	1.47717
	1.05464	.09432	.22190	.62788	

TABLE A.10  
Multiplier Coefficients for Highpass  
Design (Structure II-6)

j	$m_{1j}$	$m_{2j}$	$\delta_{0j}$	$\delta_{1j}$	$\delta_{2j}$
1	.21913	.20577	.02490	.02490	.09962
2	.08278	.37071	.08585	.08585	.34338
3	.29881	.05226	.27350	.27350	1.09398
ESS Parameters:      2 <sup>nd</sup> order      1 <sup>st</sup> order					
j	$b_{1j}$	$b_{0j}$	$b_{1j}$		
1	1.80314	.94108	.92894		
2	1.78231	.91163	.93235		
3	1.84261	.88821	.97585		
$\lambda_1 = .994154 \times 10^{-1}$					

TABLE A.11  
Multiplier Coefficients for  
Lowpass Design (New State-Space Structure)

j	$A_{\sim j}$		$B_{\sim j}$	$C_{\sim j}$	$D_{\sim j}$
1	.96842	-.11574	.03159	1.86061	.293238
	.12214	.96842	-.12214	.36722	
2	.95941	-.03650	.04059	.69867	.000410644
	.05755	.95941	-.05755	.45276	
3	.98052	-.14851	.01968	.94723	.548276
	.16101	.98032	-.16101	.02413	
$\lambda_1 = .393693$					

TABLE A.12  
Multiplier Coefficients for  
Bandpass Design (New State-Space Structure)

j	$\tilde{A}_j$		$\tilde{B}_j$	$\tilde{C}_j$	$\tilde{D}_j$
1	.80258	-.58295	.19742	-.00726	.0075870
	.58320	.80258	-.58320	-.02244	
2	.79804	-.59537	.20196	.06330	.201773
	.59653	.79804	-.59653	-.01789	
3	.81336	-.57536	.18664	-.88216	3.66522
	.57484	.81336	-.57484	.32369	
$\lambda_1 = 2.591733 \times 10^{-2}$					

TABLE A.13  
Multiplier Coefficients for  
Bandstop Design (New State-Space Structure)

j	$\tilde{A}_j$		$\tilde{B}_j$	$\tilde{C}_j$	$\tilde{D}_j$
1	.78345	-.29575	.21655	.10133	.669042
	.41632	.78345	.41632	.43410	
2	.75616	-.50931	.24384	-.81343	2.14340
	.50275	.75616	-.50275	.84999	
3	.89648	-.28032	.10352	.74690	.836687
	.33077	.89648	-.33077	.26226	
$\lambda_1 = .6163300$					

TABLE A.14  
Multiplier Coefficients for  
Highpass Design (New State-Space Structure)

j	$\tilde{A}_j$		$\tilde{B}_j$	$\tilde{C}_j$	$\tilde{D}_j$
1	-.78755	-.38630	1.78755	-.04461	.0446118
	.41582	-.78755	-.41582	.19178	
2	-.82466	-.08931	1.82446	-.03694	.0369422
	.24012	-.82446	-.24012	.28069	
3	-.77325	-.58000	1.77325	-.28754	.287538
	.55052	-.77325	-.55052	.92617	
$\lambda_1 = .122666$					