On efficient digital filter structures with finite precision implementation consideration

Zhao, Zixue

2007

Zhao, Z. (2007). On efficient digital filter structures with finite precision implementation consideration. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/46972>

<https://doi.org/10.32657/10356/46972>

Nanyang Technological University

Downloaded on 26 Apr 2023 17:50:00 SGT

9477819

ON EFFICIENT DIGITAL FILTER STRUCTURES WITH FINITE PRECISION IMPLEMENTATION CONSIDERATION

Zhao Zixue

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy

2007

R
857
.D47
Z63
2007

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

04/11/2007

Date

Zhao Zixue

ZHAO Zixue

*To all in my family,
for their encouragement and love.*

Acknowledgements

First and foremost, I would like to express my deepest appreciation and sincere gratitude to my supervisor, Dr. Gang Li, for his great support. Three years ago, when I started this Ph.D program as a fresh Bachelor degree holder, I could not imagine that I can complete it three years later. It is Dr. Gang Li who gave me invaluable instruction and sincere encouragement throughout the tough time of the research. This work would not have been possible without his excellent guidance and persistent encouragement.

I would like to thank my parents, my grandma, and all the members in my family for teaching me to understand the value of life and providing me everything needed in life. Their love and support are always the greatest inspiration to me from the first day I came to the world.

I would like to acknowledge the School of EEE, Nanyang Technological University, Singapore, for awarding me the research scholarship and providing me with the excellent research facilities.

Finally, I would like to thank all of my friends, graduate students, the faculty and technicians at Information Systems Research Laboratory. All of my experiences and interactions with these people have shaped me and my work, and made my life during these three years worthwhile and unforgettable.

Summary

With the development of digital devices, digital signal processing has been considered as one of the key tools in modern technologies. Digital filters, as the essential elements in such processing systems, play a crucial role, and hence, digital filter design has become an active field. One of the important issues in digital filter design is the finite word length (FWL) implementation. In practice, the performance of a well-designed digital filter can be totally different from the desired one and hence a computationally efficient structure in infinite precision sense is no more efficient when FWL implementation is concerned. This leads to the optimal FWL structure problems. There are four types of FWL effects: parameter perturbation, roundoff noise, limit cycle, and overflow. The focus of this thesis is on the first two effects, which are measured by the transfer function sensitivity and roundoff noise gain, respectively.

Generally speaking, the objective for the research work in this thesis is to design and develop more robust digital filter structures which can minimize the FWL effects as well as are computationally efficient.

It is well known that for a digital filter of order K , the number of nontrivial parameters in the classical optimal state-space realizations is $K^2 + 2K + 1$, which is obviously not efficient for practical implementations. The conventional shift

operator based direct-forms (DF) such as direct form II (zDFII) and direct form II transposed (zDFIIt) structures are the most efficient structures, only having $3K + 1$ parameters to implement with the dynamic range constraint taken into account. The direct-forms in delta operator, such as δ DFIIt, have been studied in order to overcome the poor FWL performance of these DF structures. In this thesis, based on the concept of polynomial parametrization, a generalized structure, denoted as ρ DFIIt, is proposed, which possesses $3K + 1$ nontrivial parameters plus K free parameters at choice. One of the nice properties of the ρ DFIIt structure and its equivalent state-space realization is that for a given digital filter, they can be optimized with the K free parameters. Compared with the δ DFIIt which can only yield a good performance for the low-pass narrow band filters, these two structures are more robust in the sense that they can be used for a wider range of filters.

In addition, the ρ DFIIt structure is investigated for the implementation scheme of rounding before multiplication. The expression of roundoff noise gain is derived, based on which a revised ρ DFIIt structure, denoted as ρ RDFIIt, is obtained. It is shown that the ρ RDFIIt yields a smaller roundoff noise gain than the corresponding ρ DFIIt, while has the same implementation simplicity. The ρ RDFIIt structure with error feedback is also examined.

It is noted that the ρ DFIIt structure can not effectively reduce the roundoff noise gain for those filters whose poles are far away from the real axis of the z -plane. Given this observation, a novel efficient structure, denoted as ρ IDFIIt, was derived in [1]. Compared with the ρ DFIIt, this new structure has $K/2$ or $(K - 1)/2$ more free parameters, depending on K even or odd. However, in [1], the ρ IDFIIt structure was only analyzed for the situation that the K original

free parameters take values from three trivial parameters $\{-1, 0, +1\}$. In this thesis, this new structure is deeply discussed and its equivalent state-space realization, denoted as ρ ISS, is obtained as well, which is also a sparse structure. It is shown that the roundoff noise gain for the ρ DFIIt and ρ IDFIIt structures can be much reduced when the roots of the corresponding polynomials are close to the dominant poles of the filter (which are generally complex). We note that the polynomials in the ρ DFIIt have real roots only, while with a few more free parameters introduced in the ρ IDFIIt, the corresponding polynomials can have complex roots. Consequently, the roundoff noise gain can be reduced much more effectively. The equivalent state-space realization of the ρ IDFIIt yields another sparse structure and this structure is also studied without/with error feedback for the rounding before multiplication implementation.

Motivated by the results obtained in the case of one-dimensional digital filter structure design, an extension of this research work to two-dimensional filters (with separable-denominator or separable-numerator) is also presented. Firstly, based on the ρ DFIIt, an efficient structure, called the ZL, is obtained. This structure is a mixture of direct-forms for that its horizontal part is implemented with an DFII and its vertical part with an DFIIt. It is found that the polynomials to the vertical part of the ZL structure have more impact on the roundoff noise gain than the horizontal part and hence an improved ZL structure, denoted as the IZL for convenience, is developed. It is shown that the IZL structure and its state-space realization can outperform the corresponding ZL and its state-space realization.

The optimal structure problems are formed for each proposed structure in terms of minimizing roundoff noise gain or transfer function sensitivity, and are

efficiently attacked by genetic algorithms in this thesis.

Contents

Acknowledgements	i
Summary	ii
List of Figures	xi
List of Tables	xv
Notations	xvi
1 Introduction	1
1.1 Motivation	2
1.2 A history review	3
1.3 Objective and contributions	6
1.4 Outline of the thesis	13
2 Literature Review	15
2.1 Digital filter structures	15
2.2 FWL effects and measures	17
2.2.1 Sensitivity measure	19

2.2.2	Roundoff noise gain	22
2.3	Two-dimensional filters	26
2.4	Genetic algorithms for optimization	30
2.5	Summary	37
3	A Generalized DFII Structure and its State-Space Realization in Digital Filter Implementation	38
3.1	Introduction	38
3.2	Two new sparse structures	41
3.2.1	Polynomial parametrization	41
3.2.2	Equivalent state-space realization	43
3.2.3	Choice of $\{\Delta_k\}$ and l_2 -scaling	45
3.3	Performance analysis	46
3.3.1	Roundoff noise gain	47
3.3.1.1	ρ DFII structure	47
3.3.1.2	Equivalent state-space realization	51
3.3.2	L_2 -sensitivity measure	52
3.3.2.1	ρ DFII structure	52
3.3.2.2	Equivalent state-space realization	53
3.4	Optimal structures	54
3.5	Numerical examples	58
3.6	Conclusion	66
4	Revised ρDFII Structures without/with Error Feedback	69

<i>Contents</i>	viii
4.1 Introduction	69
4.2 Roundoff noise gain for the ρ DFIIt structure in RBM scheme . .	71
4.3 Revised ρ DFIIt structures	73
4.3.1 A revised ρ DFIIt structure	73
4.3.2 The ρ RDFIIt structure with error feedback	75
4.4 Structure optimization	77
4.5 Numerical examples	79
4.6 Conclusion	86
5 Improved ρDFIIt Structure and its State-Space Realization	87
5.1 Introduction	87
5.2 Two novel efficient structures	90
5.2.1 An improved ρ DFIIt structure	90
5.2.2 Equivalent state-space realization	94
5.3 Performance analysis and structure optimization for RAM scheme	96
5.3.1 Roundoff noise analysis	96
5.3.1.1 The ρ IDFIIt structures	96
5.3.1.2 The ρ ISS structures	99
5.3.2 L_2 -sensitivity measure	100
5.3.3 Structure optimization	101
5.3.4 Numerical examples	105
5.4 The ρ ISS structures for RBM scheme	115
5.4.1 Roundoff noise gain for the ρ ISS	115

<i>Contents</i>	ix
5.4.2 The ρ ISS structure with error feedback	116
5.4.3 Design examples	118
5.5 Conclusion	124
6 DFII Based Structures for 2-D Digital Filters	126
6.1 Introduction	126
6.2 2-D ZL structure	129
6.2.1 Transfer function re-parametrization	129
6.2.2 Equivalent state-space realization	134
6.3 Performance analysis	136
6.3.1 Roundoff noise gain	136
6.3.1.1 ZL structure	136
6.3.1.2 Equivalent state-space realization	139
6.3.2 L_2 -sensitivity measure	140
6.3.2.1 ZL structure	140
6.3.2.2 Equivalent state-space realization	141
6.3.3 Structure optimization	142
6.4 ZL structure for separable-numerator digital filters	143
6.5 An improved ZL structure	146
6.5.1 Two new structures	146
6.5.2 Performance analysis	149
6.6 Numerical Examples	150
6.7 Conclusion	159

<i>Contents</i>	<i>x</i>
7 Conclusions and Recommendations	160
7.1 Conclusions	160
7.2 Recommendations for future research work	162
Author's Publications	165
Bibliography	167

List of Figures

1.1	Magnitude responses for Example 1.1: i – ideal; ii – z DFIIt with $f_c(4, 11)$; iii – δ DFIIt with $f_c(1, 6)$; iv – R_G with $f_c(0, 7)$. . .	7
1.2	Magnitude responses for Example 1.2: i – ideal; ii – z DFIIt with $f_c(3, 8)$; iii – δ DFIIt with $f_c(4, 7)$; iv – R_G with $f_c(0, 7)$	8
2.1	Direct forms: (a) – DFII; (b) – DFIIIt	16
2.2	State-space realization for RBM scheme with ($D \neq 0$) and without ($D = 0$) error feedbacks	23
2.3	Flowchart of the GA to solve the optimization problems.	32
3.1	The ρ DFIIt structure	43
3.2	A realization of operator ρ_k^{-1}	43
3.3	Magnitude responses (in dB) and phase responses (in degree) for Example 3.1: i – ideal; ii – $\bar{\gamma}_z$ with $f_c(4, 11)$; iii – $\bar{\gamma}_\delta$ with $f_c(1, 6)$; iv – $\bar{\gamma}(G_\rho^{\text{opt}})$ with $f_c(1, 6)$	61
3.4	Magnitude responses (in dB) and phase responses (in degree) for Example 3.2: i – ideal; ii – $\bar{\gamma}_z$ with $f_c(3, 12)$; iii – $\bar{\gamma}_\delta$ with $f_c(4, 11)$; iv – $\bar{\gamma}(G_\rho^{\text{opt}})$ with $f_c(0, 7)$	62

3.5	Simulation for Example 3.1: (a) – ideal output; (b) – output error of ρ DFIIt with $\bar{\gamma}_z$ for $f_c(4, 11)$, $f_s(0, 15)$; (c) – output error of ρ DFIIt with $\bar{\gamma}_\delta$ for $f_c(0, 15)$, $f_s(0, 15)$; (d) – output error of ρ DFIIt with $\bar{\gamma}(G_\rho^{\text{opt}})$ for $f_c(0, 15)$, $f_s(0, 15)$	65
3.6	Simulation for Example 3.2: (a) – ideal output; (b) – output error of R_G for $f_c(0, 15)$, $f_s(0, 15)$; (c) – output error of ρ SS with $\bar{\gamma}_z$ for $f_c(3, 12)$, $f_s(3, 12)$; (d) – output error of ρ SS with $\bar{\gamma}(G_{ss}^{\text{opt}})$ for $f_c(0, 15)$, $f_s(0, 15)$	67
4.1	The ρ RDFIIt structure	73
4.2	A realization of operator ρ_k^{-1} with error feedback	76
4.3	Simulation for Example 4.1: (a) – ideal output; (b) – output error of R_G for $f_c(0, 15)$, $f_s(1, 14)$; (c) – output error of ρ DFIIt for $f_c(1, 14)$, $f_s(0, 15)$; (d) – output error of ρ RDFIIt for $f_c(0, 15)$, $f_s(0, 15)$	83
4.4	Structure performance vs filter order for Example 4.2: i — R_G ; ii — ρ DFIIt; iii — ρ RDFIIt; iv — R_G^g ; v — R_G^d ; vi — ρ REF . .	85
5.1	Block diagram of the proposed ρ IDFIIt structure	93
5.2	Implementation of the m th block in the ρ IDFIIt structure, where $\bar{m} \triangleq K - 2p + 2m$ and $w_{p+1}(n) = 0$	93
5.3	Magnitude responses (in dB) and phase responses (in degree) for Example 5.1: i – ideal; ii – z DFIIt with $f_c(4, 7)$; iii – δ DFIIt with $f_c(4, 7)$; iv – R_G with $f_c(0, 11)$; v – ρ DFIIt with $f_c(3, 8)$; vi – ρ IDFIIt with $f_c(0, 11)$	108

5.4	Simulation for Example 5.1: (a) – ideal output; (b) – output error of $zDFII_t$ with $f_s(2, 13)$; (c) – output error of $\delta DFII_t$ with $f_s(2, 13)$; (d) – output error of R_G with $f_s(0, 15)$; (e) – output error of $\rho DFII_t$ with $f_s(1, 14)$; (f) – output error of $\rho IDFII_t$ with $f_s(1, 14)$	110
5.5	Comparison results of $R(B_c)$ for the bandpass filter $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.35])$ for Example 5.4: i – zSS ; ii – δSS ; iii – R_G ; iv – ρSS ; v – ρISS	112
5.6	Comparison results of $R(B_c)$ for the low-pass filter $[V_b, V_a] = \text{butter}(8, 0.2)$ for Example 5.4: i – zSS ; ii – δSS ; iii – R_G ; iv – ρSS ; v – ρISS	113
5.7	Structure performance vs filter order for Example 5.7: i – R_G ; ii – ρSS ; iii – ρISS ; iv – R_G^g ; v – R_G^d ; vi – ρISS^g ; vii – ρISS^d	121
5.8	Roundoff noise performance vs normalized bandwidth for Example 5.8: i – R_G ; ii – ρSS ; iii – ρISS ; iv – R_G^g ; v – R_G^d ; vi – ρISS^g ; vii – ρISS^d	122
5.9	Comparison results of $V(B_s)$ for the bandpass filter $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.35])$ for Example 5.9: i – R_G ; ii – ρSS ; iii – ρISS	123
5.10	Comparison results of $V(B_s)$ for the low-pass filter $[V_b, V_a] = \text{butter}(6, 0.1)$ for Example 5.9: i – R_G ; ii – ρSS ; iii – ρISS . . .	124
6.1	2-D ZL structure for separable-denominator digital filters	133
6.2	Realizations of operators $h_k^{-1}(z_1)$ and $v_l^{-1}(z_2)$	134
6.3	2-D ZL structure for separable-numerator digital filters	145
6.4	Improved 2-D ZL structure	148

6.5	Magnitude responses for Example 6.1: (a) the ideal filter; (b) the ZL structure with $\bar{\gamma}_z$; (c) the ZL structure with $\bar{\gamma}(M_{ZL}^{\text{opt}})$	154
6.6	Magnitude responses for Example 6.3: (a) the ideal filter; (b) the ZL structure with shift-operators; (c) the optimized ZL structure; (d) the optimized IZL structure	158

List of Tables

1.1	Summary of 1-D Structures	11
2.1	Statistic History of the GA for Example 2.1	37
3.1	Comparison Results for Example 3.1	59
3.2	Comparison Results for Example 3.2	60
4.1	Comparison Results for Example 4.1	80
4.2	Statistical History of the GA for ρ REF in Example 4.1	81
4.3	Comparison Results for Example 4.3	86
5.1	Comparison Results for Example 5.1	107
5.2	Comparison Results for Example 5.2	111
5.3	Comparison Results for Example 5.3	111
5.4	Comparison Results for Example 5.5	114
5.5	Comparison Results for Example 5.6	119
6.1	Comparison Results for Example 6.1	152
6.2	Comparison Results for Example 6.2	155
6.3	Comparison Results for Example 6.3	157

Notations

- $x(n)$: a discrete-time signal x
- z^{-1} : the backward shift operator such that $z^{-1}x(n) = x(n - 1)$
- A^H : complex conjugate and transpose of matrix A
- A^{-1} , A^T : inverse and transpose of matrix A , respectively
- $tr(\cdot)$: trace operator
- e_m : the m th elementary (column) vector, whose elements are all zeros except the m th which is one
- I_K : the identity matrix of dimension K
- $\oplus$: the direct sum of two matrixes
- $V(i)$: the i th element in vector V
- $M(i, j)$: the element in the i th row and j th column of matrix M
- $A(i : j, k : l)$: the submatrix of matrix A formed with cross elements of the i th row to j th row and k th column to l th column
- $f_s(I_s, B_s)$: an implementation of coefficients where the integral and fractional parts of a coefficient are represented with I_s -bits and B_s -bits, respectively

- $f_c(I_c, B_c)$: an implementation of signal where the integral and fractional parts of the signal are represented with I_c -bits and B_c -bits, respectively
- $\bar{x}$: the mean (expected value) of the random signal $x(n)$
- σ_x^2 : the variance of the random signal $x(n)$

Chapter 1

Introduction

Digital signal processing (DSP) has been considered as one of the most important fields in engineering during the past three decades, and this will continue due to the rapid developments in digital computer technology. DSP could be traced back to the numerical procedure invented by Newton and Gauss in the 17th and 19th centuries, but it was during the decade of 1960-1970 that it became important. New technologies in the past 30 years have reduced the cost of digital hardware, and its speed has increased to such an extent that DSP has replaced a great deal of analog signal processing.

Digital filters are the essential and very important element in *any* DSP systems, as a result of which, digital filter design has received intensive attention during the last decades. Digital filter design is to approximate a desired frequency response with a stable transfer function $H(z)$ that can be actually implemented. A cost function is usually defined to measure the difference between the desired response and the one given by the transfer function. The optimal filter is obtained as the transfer function that minimizes this cost function.

Techniques for digital filter design can be found from many references [2] - [5].

In real-time applications, the well designed DSP systems, where digital filters are involved, have to be implemented with a digital device, say a micro-processor (for example, Motorola 68000 processor).

1.1 Motivation

One of the basic problems in DSP is the finite precision numbers and arithmetic, which was first considered in digital filter implementation [2], [6]. A linear filter is usually defined by its transfer function, which uniquely describes its input-output relationship. For a given digital filter, there exist a number of different structures (also called realizations)¹, which, in infinite precision, are all equivalent since they represent the same transfer function. However, when the parameters that describe the filter are implemented with finite word length (FWL), these realizations are no longer equivalent, and thus, the transfer function obtained from one realization can be different from another. Similarly, when the signals processed by the digital filter are encoded in finite word length, roundoff must occur at every arithmetic operation and the input-output behavior is again different from what it would be in infinite precision. The error caused by the roundoff operation is called roundoff noise. In practice, these two FWL effects occur jointly and both result in a performance degradation of the filter compared to its ideal behavior. Given this fundamental observation, it is then natural to formulate a range of optimal FWL structure problems, where optimality refers to objectives such as minimizing the sensitivity of the transfer

¹Here, a *structure* or *realization* means a way in which the output of a digital filter can be computed with an input given.

function with respect to parameter perturbation, minimizing the roundoff noise gain, etc. A lot of work has been done in this direction in the signal processing area during the past three decades.

In practice, it is desired that the actually implemented filter has a nice FWL performance as well as a *sparse* structure, which possesses fewer nontrivial² parameters. This entails fewer multipliers (and hence a greater computational speed) and more error-free computations. Therefore, designing suboptimal FWL structures, which are as sparse as possible, is another interested objective by many researchers.

1.2 A history review

Traditionally, finite precision analysis is separated into sensitivity and roundoff noise. In the early days, the digital filter was implemented with direct forms³ and the structure problem was not seriously considered. Later, the poor performance of a filter implemented in these forms made one realize the importance of the structure issue, which led to the so-called *optimal FWL structure design*. The key point is that for a fixed input-output relation, there exists an infinite class of realizations which are totally equivalent in infinite precision, while have different numerical properties when realized with FWL implementations. The optimal FWL structure design is to search for those structures that minimize the FWL effects in a certain FWL effects-related measure.

²By *trivial parameters*, we mean those that are 0 and ± 1 . Other parameters are, therefore, referred to *nontrivial parameters*.

³A direct form is one in which the coefficients of the numerator and denominator of the transfer function appear directly as parameters in the realizations.

State-space realization theory provides an infinite set of similar⁴ realizations for a fixed transfer function. It is well known that all these structures have the same dynamical input-output behavior [7], but it was not until 1968 that the problem of how to exploit this class of structures for the purpose of optimal FWL implementation of a digital system came under scrutiny. It turns out that the first significant results on the design of state-space realizations minimizing FWL effects are results on the minimization of the roundoff noise gain. In 1976, Mullis and Roberts [8] explored similarity transformations to determine the state-space realizations that minimize the output error variance due to roundoff errors. Independently, Hwang [9] considered the same problem and gave a constructive procedure to determine the optimal transformations or realizations. The corresponding realizations are called the optimal state-space realizations with minimum roundoff noise.

An alternative objective is to identify those state-space realizations that minimize a certain measure of sensitivity of the transfer function with respect to the coefficients in the matrices (A, B, C, d) of the realization. Since the numerical accuracy of a problem depends crucially on which particular state-space realization is used [10], [11], it is desirable to define appropriate measures that represent the sensitivity of the realized transfer function with respect to parameter perturbations. The first analytical attempt to synthesize minimum sensitivity state-space realizations was in [12], [14], where a rather sophisticated frequency independent sensitivity measure was defined, called L_1/L_2 sensitivity measure. The success of this method stems from the fact that an upper bound of this sensitivity measure can be obtained whose minimization with respect to all

⁴Similar realizations are state-space realizations whose states are related by a real nonsingular transformation matrix.

realizations is particularly simple. In [14], it is shown that not only the upper bound but also the sensitivity measure itself is minimized by all realizations whose controllability and observability Gramian matrices are identical. Later on, a pure L_2 based transfer function sensitivity measure was studied [15], and some properties of this measure were revealed in [16] - [18]. It should be noted that the L_2 sensitivity minimization is more natural and reasonable than the conventional L_1/L_2 mixed sensitivity minimization, but it is more challenging.

One observes that the optimal state-space realizations are usually fully parameterized. This means that for a digital filter of order K , the number of nontrivial parameters is $K^2 + 2K + 1$, which is obviously not efficient to implement. A lot of efforts have been made to search for realizations that not only yield a nice FWL performance but also have a sparse structure [19] - [27]. Since the degrees of freedom in the optimal roundoff noise realizations are very limited, sophisticated numerical algorithms were developed to obtain the so-called quasi-optimal realizations that have a very sparse structure [23], [24]. Besides the difficulty involved in those algorithms, the position of each trivial parameter in the structure is not predictable. In [27], based on a matrix factorization an interesting structure was derived, which possesses $5K - 1$ nontrivial parameters.

It is well known that though having poor numerical properties the conventional shift operator based direct forms such as direct-form II (DFII) and direct form II transposed (DFII_t) structures are the simplest structures. Recently, the direct-forms in delta operator have been studied by researchers [28] - [33]. In [28], the δ DFII_t structure was investigated for an arbitrary order IIR filter, where the concept of different coupling coefficients at different branch nodes was utilized to avoid overflow. It was found that one could make the transfer

function in delta operator have better numerical properties in the case where the poles of the transfer function are closer to $z = +1$ than $z = 0$ [34], [35]. This means that the delta operator based structures have a very good performance just for the type of *low-pass narrow band* and may not yield a satisfactory performance for other types of filters.

Error feedback [36] - [52] has been known as another effective technique for reducing the roundoff noise at the filter output. This is achieved by extracting the quantization error after multiplication and addition, and then feeding the error signal back through simple circuits. It can be applied to digital filters that are described by either external or internal models for roundoff noise minimization without affecting the filter's input-output relationship. In [46], a number of results for computing the optimal error feedback coefficients were presented for any given state-space realization.

All these results, which are obtained for the one-dimensional (1-D) digital filter implementation, can be extended to the problems of minimizing the FWL effects in two-dimensional (2-D) digital filters, see [4], [5], [47] - [70].

1.3 Objective and contributions

For a given digital filter, different structures have different behaviors towards the FWL effects. To make readers understand this fact more intuitively, let us see two examples.

Example 1.1: This is a fourth order low-pass Butterworth filter with a very narrow bandwidth, generated with MATLAB command $[V_b, V_a] = \text{butter}(4, 0.05)$.

Figure 1.1: Magnitude responses for Example 1.1: i – ideal; ii – $z\text{DFIIt}$ with $f_c(4, 11)$; iii – δDFIIt with $f_c(1, 6)$; iv – R_G with $f_c(0, 7)$

We compute and compare the magnitude responses for three popularly implemented realizations whose coefficients are implemented with I_c -bits for the integral parts and B_c -bits for the fractional parts, represented by $f_c(I_c, B_c)$ (see Notations). One realization is the shift operator based canonic DFII structure – $z\text{DFIIt}$, another one is the so-called δDFIIt structure [28], and the third one is the optimal state-space realization [8], [9], denoted as R_G . Figure 1.1 presents the magnitude responses (in dB) for the ideal filter and three FWL implementations.

Example 1.2: The second example is a sixth order bandpass filter, generated with MATLAB command $[V_b, V_a] = \text{butter}(3, [0.3 \ 0.4])$.

We also compute and compare the magnitude responses for the three FWL realizations. The corresponding results are shown in Figure 1.2.

Figure 1.2: Magnitude responses for Example 1.2: i – ideal; ii – $z\text{DFIIt}$ with $f_c(3, 8)$; iii – δDFIIt with $f_c(4, 7)$; iv – R_G with $f_c(0, 7)$

The two examples show clearly that the optimal realization R_G is significantly less sensitive to coefficient perturbations than $z\text{DFIIt}$ and δDFIIt , while it possesses much more nontrivial parameters which limit the efficiency of practical implementations. They also validate that the delta-operator based structures have a very good performance for the low-pass narrow band filters but may not be effective for other filter types.

Therefore, the objective for the research work in this thesis is to design more robust digital filter structures that can minimize the FWL effects, especially the roundoff noise gain and parameter sensitivity, for a wider range of digital filters. On the other hand, for the efficiency consideration of the practical implementations, the structures should be as sparse as possible.

The contributions of this thesis are summarized as follows:

- Motivated by [28] and based on the concept of *polynomial parametrization* [16], a set of special operators is derived, with which a generalized DFIIt

structure, denoted as ρDFIIt , is obtained. This structure, having $3K + 1$ non-trivial parameters plus K free parameters at choice, can be used for any types of digital filters. An equivalent state-space realization of the ρDFIIt structure, denoted as ρSS , is obtained, which possesses $4K$ nontrivial parameters. Expressions for evaluating the roundoff noise gain and L_2 -based sensitivity measure are derived for both structures in terms of the rounding after multiplication (RAM) implementations. It is interesting to note that for a given set of free parameters, the ρSS structure always yields a smaller roundoff noise gain than the corresponding ρDFIIt structure. One of the nice properties of these two structures is that for a given digital filter of any type, they can be optimized with a proper choice for the K free parameters in terms of minimizing the roundoff noise gain or sensitivity measure. Relationship between the free parameters and the poles of the filter is established in terms of roundoff noise gain, which implies that the optimal free parameters can be searched around the dominant poles of the filter. It is shown that the two proposed structures are computationally efficient, yield a better performance than the traditional $z\text{DFIIt}$ and the popular δDFIIt , and are very competitive to the fully parameterized optimal state-space realization R_G .

- The proposed ρDFIIt structure is analyzed for the implementation scheme of rounding before multiplication (RBM). Noting the fact that the roundoff error due to a certain state variable can make a significant contribution to the total roundoff noise gain, a revised ρDFIIt structure, denoted as ρRDFIIt , is obtained. It is shown that this new structure, having the same implementation simplicity, yields a smaller roundoff noise gain than the corresponding ρDFIIt . Moreover, the roundoff noise gain can be further reduced by applying error feedback to the ρRDFIIt structure. Each structure can be optimized with respect to the K

free parameters.

- Given the observation that the ρ DFIIt structure cannot effectively reduce the roundoff noise gain for those filters whose poles are far away from the real axis of the z -plane, a series of more robust polynomial operators are derived, based on which, an alternative structure, denoted as ρ IDFIIt, was developed in [1]. Compared with the ρ DFIIt, the proposed structure has $K/2$ or $(K-1)/2$ more free parameters, depending on K even or odd. However, in [1], the ρ IDFIIt structure was only analyzed for the situation that the K original free parameters take values from three trivial parameters $\{-1, 0, 1\}$. In this thesis, this new structure is deeply discussed and its equivalent state-space realization, denoted as ρ ISS, is obtained as well, which is also a sparse structure. Assuming the RAM implementations, the expressions of the roundoff noise gain and L_2 -based sensitivity measure for both structures are derived, from which, one finds that the ρ ISS always yields a smaller roundoff noise gain than its ρ IDFIIt counterpart. It is shown that the roundoff noise gain for the proposed structures can be much reduced when the roots of the corresponding polynomials are close to the dominant poles of the filter. We note that the polynomials used in the ρ DFIIt structure have real roots only while the filters generally have complex poles. With the extra degrees of freedom introduced, the corresponding polynomials to the ρ IDFIIt structure can have complex roots and consequently, the roundoff noise gain can be reduced much more effectively. Performance analysis is also carried out for the ρ ISS structure without/with error feedback in terms of RBM scheme. How to determine the optimal full-scale and diagonal error feedback matrices is explained.

For quick reference, Table 1.1 presents a brief summary of the proposed 1-

Table 1.1: Summary of 1-D Structures

Realization	Performance Description	Complexity
$z\text{DFIIt}$	the most efficient structure while having poor performance against FWL effects	$3K + 1$
δDFIIt	good FWL performance for low-pass narrow band filters	$3K + 1$
R_G	nice FWL performance for all the filter types while possessing the most nontrivial parameters	$K^2 + 2K + 1$
ρDFIIt	1. outperforming $z\text{DFIIt}$ and δDFIIt 2. being competitive to R_G 3. fitting for the filters whose poles are close to the real axis of the z -plane	$4K + 1$
ρSS	yielding a smaller roundoff noise gain than the corresponding ρDFIIt	$4K$
ρRDFIIt	1. yielding a much smaller roundoff noise gain than the corresponding ρDFIIt in terms of RBM implementations 2. having very good performance for low-pass narrow band filters	$4K + 1$
ρIDFIIt	1. over-performing R_G against roundoff noise as well as structure efficiency 2. fitting for all the filter types	$4K + 1 + K/2$
ρISS	yielding a smaller roundoff noise gain than the corresponding ρIDFIIt	$4K - 1 + K/2$

D structures in this thesis, as well as the traditional $z\text{DFIIt}$, δDFIIt structures and the optimal state-space realization R_G . The column of Complexity gives the number of nontrivial parameters for each structure, given a K th order digital filter⁵.

- The approach of deriving the ρDFIIt structure for 1-D digital filters is extended to the 2-D case (with separable denominator or numerator) and consequently, a new structure, denoted as 2-D ZL, is obtained. Taking a (M, N) th-order filter with separable denominator as example, the 2-D ZL structure, being a mixture of DFII for its horizontal part and DFIIt for its vertical part, possesses

⁵For the ρIDFIIt and ρISS structures, K being even is assumed here.

only $MN + 3(M + N) + 1$ nontrivial parameters plus $(M + N)$ free parameters. The equivalent state-space realization of the ZL structure is also obtained, which can be implemented with $MN + 4(M + N) - 1$ parameters. Performance analysis is carried out by deriving the expressions of roundoff noise gain and L_2 -norm based sensitivity measure for both structures. Like in the 1-D case, the free parameters can be used to minimize the FWL effects. Based on the observation that the roundoff noise performance of the proposed structures is strictly related to the polynomial operators in the vertical part, the more robust polynomials used in ρ IDFIIt are applied to the vertical part of the 2-D filters, resulting in another efficient structure, denoted as IZL for convenience. It is shown that the IZL structure and its state-space realization can outperform the corresponding ZL and its state-space realization.

- Genetic algorithms (GAs) [71] are utilized to find the optimal free parameters for each proposed structure. For the ρ DFIIt structure, since the free parameters take values from a finite space, they can be found using exhaustive search. While in practice, a good solution obtained in a reasonable time is better than the global optimum solution taking excessive time and/or computational resources. For that reason, GAs are utilized to efficiently solve the optimal structure problems. For the ρ IDFIIt structure, it contains more introduced free parameters which are continuous in a given space. This means that exhaustive search is no longer applicable, which is an unsolved problem in [1]. Hence, GAs further show its applicability and efficiency in solving our optimization problems.

1.4 Outline of the thesis

The rest of thesis is organized as follows.

Chapter 2: An overview of 1-D and 2-D digital filter structures is briefly described. FWL effects are analyzed, especially the measures for roundoff noise gain and transfer function sensitivity. Some existing results in the optimal FWL structure design are presented. As a standard optimization algorithm, GA is introduced, which is used to efficiently solve the optimal FWL structure problems in this thesis.

Chapter 3: Based on the concept of polynomial operators, a new efficient structure is proposed for digital filter implementation, denoted as ρDFIIt . Taken RAM scheme into consideration, expressions for evaluating the sensitivity measure and roundoff noise gain are derived for the ρDFIIt structure and its equivalent state-space realization that has the same structure simplicity. The relationship between the free parameters and the poles of the filter is also established in terms of roundoff noise gain, which implies that the optimal free parameters can be searched around the dominant poles of the filter.

Chapter 4: Two efficient structures are analyzed for the implementation scheme of RBM in this chapter. The first one is the proposed ρDFIIt . The expression of roundoff noise gain is derived, based on which a revised ρDFIIt structure, denoted as ρRDFIIt , is obtained, which yields a smaller roundoff noise gain than the corresponding ρDFIIt . The roundoff noise gain for the ρRDFIIt structure with error feedback is also derived.

Chapter 5: A novel efficient digital filter structure, denoted as ρIDFIIt , is derived. This structure can be block-diagramed as the proposed ρDFIIt ,

in which the first-order ρ -operators are replaced with a set of second-order operators. Assuming the RAM implementations, the expressions of the roundoff noise gain and L_2 -based sensitivity measure for the ρ IDFII and its equivalent state-space realization, ρ ISS, are derived. Performance analysis is also carried out for the ρ ISS structure without/with error feedback in terms of RBM scheme.

Chapter 6: Motivated by the results obtained in previous chapters, an extension of this research work to 2-D digital filters is discussed. Two 2-D structures, both of which are a mixture of DFII for their horizontal part and DFII for their vertical part, and their corresponding state-space realizations, are derived. Similar optimal structure problems are formulated and solved.

Chapter 7: The conclusions are presented and future research directions related to this work are recommended.

Chapter 2

Literature Review

In this chapter, we are going to provide some background that are needed in the sequel of this thesis. The review begins from presenting some traditional digital filter structures, then introducing the measures for FWL effects and some existing results in optimal FWL structure design, at last giving a brief description of genetic algorithm which is used to attack the optimization problems in this thesis.

2.1 Digital filter structures

A well-designed single-input-single-output time-invariant linear digital filter $H(z)$ is usually given by

$$H(z) = \frac{b_0 z^K + b_1 z^{K-1} + \cdots + b_{K-1} z + b_K}{z^K + a_1 z^{K-1} + \cdots + a_{K-1} z + a_K} \triangleq \frac{N(z)}{D(z)}. \quad (2.1)$$

Figure 2.1: Direct forms: (a) – DFII; (b) – DFIIIt

This filter can be implemented with its input-output relationship:

$$y(n) = - \sum_{k=1}^K a_k y(n-k) + \sum_{k=0}^K b_k u(n-k) \quad (2.2)$$

where $u(n)$ and $y(n)$ are the scalar input and output of the filter, respectively.

(2.2) is usually referred to the direct form I (DFI) structure.

The implementation complexity is usually measured with two factors. One is the number of multiplications and additions, needed for computing one output sample, and the other is the size of memory units, required for storing the coefficients and signals. The most efficient structures in infinite precision implementation include the DFII and DFIIIt structures, which are depicted in Figure

2.1, (a) and (b), respectively.

This filter can also be implemented with alternative structures [3], such as cascade-form structures and lattice structures. The state-space realizations provide another class of structures for digital filter implementation:

$$\begin{cases} x(n+1) = Ax(n) + Bu(n) \\ y(n) = Cx(n) + du(n) \end{cases} \quad (2.3)$$

where $x(n) \in \mathcal{R}^{K \times 1}$ is the state vector, and $R \triangleq (A, B, C, d)$ with $A = \{a_{lk}\} \in \mathcal{R}^{K \times K}$, $B = \{b_k\} \in \mathcal{R}^{K \times 1}$, $C = \{c_k\} \in \mathcal{R}^{1 \times K}$, $d = b_0 \in \mathcal{R}$, is called a (state-space) realization of $H(z)$, satisfying

$$H(z) = d + C(zI_K - A)^{-1}B. \quad (2.4)$$

Denote $S_H \triangleq \{R : R \text{ satisfies (2.4)}\}$. In fact, S_H can be characterized with

$$A = T^{-1}A_0T, \quad B = T^{-1}B_0, \quad C = C_0T \quad (2.5)$$

where $(A_0, B_0, C_0, d) \in S_H$ is a given (initial) realization and $T \in \mathcal{R}^{K \times K}$ is any non-singular real valued matrix, called a similarity transformation.

2.2 FWL effects and measures

In practice, digital filters must be implemented with finite precision numbers and arithmetic. That means all the filter coefficients and signals should be quantized in a discrete form, which leads to four types of FWL effects [72]:

1. Quantization of the filter coefficients can make the actual filter response differ from the ideal one. This is referred to as *parameter perturbation*.
2. Quantization (rounding or truncation) of the filter arithmetic operations results in the error in the filter output, which is called *roundoff noise*.
3. Quantization of the filter arithmetic operations also renders the filter slightly nonlinear. For some kinds of filters, this nonlinearity can cause spurious oscillations called *limit cycles*.
4. When the computation results are out of the range of the hardware capacity, the *overflow* occurs, which can lead to strange effects, including instability of an otherwise stable filter.

Among these effects, the third type - limit cycle oscillations, is a potential problem which may arise from the nonlinear effects of overflow and signal quantization in any practical implementation of a recursive digital filter. It occurs when the quantization error sequences are highly correlated. Considerable work has been done with the aim of determining reasonable bounds for the amplitude of limit cycles [73] - [80]. However, many of the proposed bounds are poorly suited to actual use in some situations because they are difficult to apply to high-order filters, or apply only to certain filter structures [79], [80].

The fourth type - overflow, is related to the wordlength capacity of the application on which the computations are performed [81], [82]. Overflow results when the dynamic range of the signals exceeds the wordlength capacity. The proper way to avoid overflow in arithmetic computations is by scaling the variables so that the dynamic range of internal variables, obtained after multiplications, never exceeds the available wordlength. There are several ways of

performing a scaling operation, going from very simple input scaling to the more sophisticated dynamical scaling of the internal variables as the computations proceed. Usually, there exists a compromise between an excessively cautious scaling that results in an available dynamic range that far exceeds the likely range of the signals (this would totally prevent overflow but at the expense of high quantization noise), and an insufficient scaling that would result in likely overflow. In this thesis, we are using the l_2 -scaling scheme to avoid overflow for all the proposed structures (see Section 2.2.2).

In this thesis, assuming that the quantization noise sequences from different multipliers in the filter are statistically independent and uncorrelated with each other and with the processed signal, limit cycles are not investigated for the proposed structures. With the overflow taken into consideration, the outstanding task is to minimize the first two effects - parameter perturbation and roundoff noise, which are measured by the transfer function sensitivity and roundoff noise gain, respectively.

2.2.1 Sensitivity measure

Denote $\{\tau_k\}$ as the parameter set of a filter structure and

$$S_{FWL} \triangleq \{\tau_k\}$$

as the set of the parameters which can not be represented exactly with a given finite word length implementation (say, of 16 bits). For an actual implementation of this structure, these (ideal) parameters have to be truncated into an FWL format. Therefore, the actually implemented transfer function, denoted

by $\tilde{H}(z)$, is different from $H(z)$. In the traditional FWL analysis, the parameter errors are modelled as zero mean uniformly distributed independent random variables [83], [84]. Keeping this in mind, one can show [16], [61] that the variance of the degradation $\Delta H(z) \triangleq \tilde{H}(z) - H(z)$ is proportional to the following sensitivity measure:

$$M_{L_2} \triangleq \sum_{\tau_k \in S_{FWL}, \forall k} \left\| \frac{\partial H(z)}{\partial \tau_k} \right\|_2^2 \quad (2.6)$$

where $\|\cdot\|_2$ is the L_2 -norm defined below:

$$\begin{aligned} \|S(z)\|_2 &\triangleq \left[\frac{1}{2\pi} \int_0^{2\pi} \sum_{i=1}^l \sum_{k=1}^m |s_{ik}(e^{j\omega})|^2 d\omega \right]^{1/2} \\ &= \left[\text{tr} \left(\frac{1}{2\pi j} \oint_{|z|=1} S(z) S^H(z) z^{-1} dz \right) \right]^{1/2} \end{aligned} \quad (2.7)$$

with $S(z) = \{s_{ik}(z)\} \in \mathbb{C}^{l \times m}$ any complex matrix-valued function of the complex variable z .

Let $S(z) = D + L(zI - \Phi)^{-1}J$. It can be shown that¹

$$\|S(z)\|_2^2 = \text{tr}(DD^T + LW_c L^T) = \text{tr}(D^T D + J^T W_o J) \quad (2.9)$$

where W_c, W_o are the controllability and observability gramians of the realiza-

1

$$\|S(z)\|_2^2 = \text{tr} \left\{ \frac{1}{2\pi j} \oint_{|z|=1} [D + L(zI - \Phi)^{-1}J] [D^T + J^T (z^{-1}I - \Phi^T)^{-1}L^T] z^{-1} dz \right\} \quad (2.8)$$

Noting that

$$(zI - A)^{-1}B = z^{-1} \sum_{i=0}^{+\infty} A^i B z^{-i}$$

and 1-D Cauchy integral theorem

$$\frac{1}{2\pi j} \oint_{|z|=1} K z^{-i} dz = \begin{cases} K, & \text{if } i = 1 \\ 0, & \text{otherwise} \end{cases}$$

(2.9) can be derived from (2.8).

tion (Φ, J, L, D) , respectively, which are the solutions of the Lyapunov equations

$$W_c = \Phi W_c \Phi^T + J J^T, \quad W_o = \Phi^T W_o \Phi + L^T L. \quad (2.10)$$

When the filter is implemented with a realization (A, B, C, d) , with $A = \{a_{kl}\}$, $B = \{b_k\}$, $C = \{c_k\}$, it can be shown [18] that M_{L_2} can be evaluated with

$$\begin{aligned} M_{L_2} = & \sum_{l=1}^K \sum_{k=1}^K \phi(a_{kl}) \begin{bmatrix} C & \mathbf{0} \end{bmatrix} R_{kl} \begin{bmatrix} C & \mathbf{0} \end{bmatrix}^T \\ & + \sum_{k=1}^K [\phi(b_k) W_o(k, k) + \phi(c_k) W_c(k, k)] + \phi(d) \end{aligned} \quad (2.11)$$

where W_c and W_o are the controllability and observability gramians of the realization (A, B, C, d) , respectively, and $\phi(x)$ is a function defined as

$$\phi(x) \triangleq \begin{cases} 1, & x \in S_{FWL} \\ 0, & \text{otherwise} \end{cases} \quad (2.12)$$

and R_{kl} is the solution of the following Lyapunov equation

$$R_{kl} - \begin{bmatrix} A & e_k e_l^T \\ \mathbf{0} & A \end{bmatrix} R_{kl} \begin{bmatrix} A & e_k e_l^T \\ \mathbf{0} & A \end{bmatrix}^T = \begin{bmatrix} \mathbf{0} \\ B \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ B \end{bmatrix}^T, \quad \forall(k, l) \quad (2.13)$$

with e_k the k th elementary vector, see Notations.

Though for a given realization, one can compute this sensitivity measure, how to minimize M_{L_2} within the realization space S_H is very difficult. As far as we know, this is still an open problem. However, when the realizations are

fully parameterized, i.e., all the parameters are belong to S_{FWL} , one has

$$M_{L_2} = \left\| \frac{\partial H}{\partial A} \right\|_2^2 + \left\| \frac{\partial H}{\partial B} \right\|_2^2 + \left\| \frac{\partial H}{\partial C^T} \right\|_2^2 + \left\| \frac{\partial H}{\partial d} \right\|_2^2.$$

The corresponding optimal sensitivity problem has been solved in [16], [17], [61].

2.2.2 Roundoff noise gain

A rounding operation has to be applied in a fixed-point arithmetic implementation due to the FWL constraint. The rounding error, say $e(n)$, which is usually modelled as a white noise sequence and statistically independent [8], [9], generates the corresponding deviation at the filter output, say $\Delta y(n)$. Denote $F(z)$ as the transfer function between $e(n)$ and $\Delta y(n)$. It is well known [16] that $\Delta y(n)$ is a stationary process and the variance $E[(\Delta y(n))^2] = \|F(z)\|_2^2 E[(e(n))^2]$. The roundoff noise gain then can be given by

$$G \triangleq \frac{E[(\Delta y(n))^2]}{E[(e(n))^2]} = \|F(z)\|_2^2 \quad (2.14)$$

where $\|\cdot\|$ is the L_2 -norm defined in (2.7).

Depending on the hardware, two different roundoff versions are used in arithmetic operations: rounding performed before multiplication (RBM) and rounding after multiplication (RAM) [9], [85]. Taking the RBM scheme for example, let us see the roundoff noise analysis for a state-space realization. In this case, an FWL implementation of (2.3) can be expressed by

$$\begin{cases} x^*(n+1) &= A Q[x^*(n)] + Bu(n) \\ y^*(n) &= C Q[x^*(n)] + du(n) \end{cases} \quad (2.15)$$

Figure 2.2: State-space realization for RBM scheme with ($D \neq 0$) and without ($D = 0$) error feedbacks

whose corresponding block diagram is shown in Figure 2.2 with D a zero matrix. Each parameter in (A, B, C, d) assumes an exact fractional B_c -bit representation. All the FWL $x^*(n)$ and $y^*(n)$ have a B -bit representation, while $u(n)$ is a $(B - B_c)$ -bit fraction. The quantizer $Q[\cdot]$ in (2.15) rounds the B -bit fraction $x^*(n)$ into $(B - B_c)$ -bit, where the sign bit is not counted. Therefore, the roundoff error vector

$$e(n) \triangleq x^*(n) - Q[x^*(n)]$$

has a covariance $\sigma^2 I_K$ with $\sigma^2 = (1/12)2^{-2(B-B_c)}$.

Denote

$$\Delta y(n) \triangleq y(n) - y^*(n), \quad \Delta x(n) \triangleq x(n) - x^*(n).$$

Subtracting (2.15) from (2.3) yields

$$\begin{cases} \Delta x(n+1) &= A \Delta x(n) + A e(n) \\ \Delta y(n) &= C \Delta x(n) + C e(n) \end{cases} \quad (2.16)$$

from which, one can obtain the transfer function between $e(n)$ and $\Delta y(n)$,

denoted as $H_e(z)$, given by

$$H_e(z) = C(zI_K - A)^{-1}A + C. \quad (2.17)$$

It follows from (2.7) - (2.9) that the roundoff noise gain of the realization by (2.15) is

$$G = \|H_e(z)\|_2^2 = \text{tr}(W_o). \quad (2.18)$$

For the RAM implementations, roundoff noise gain is strongly related to the number of nontrivial parameters in (A, B, C, d) . With the similar approach, it can be shown [8], [9], [16] that the roundoff noise gain in this case is given by

$$G = \text{tr}(W_o Q) + m_{K+1} \quad (2.19)$$

where $Q = \text{diag}(m_1, \dots, m_k, \dots, m_K)$ with m_k the number of nontrivial parameters in the k th row vector of $[A \ B]$ for $k = 1, \dots, K$, and m_{K+1} the number of nontrivial parameters in the row vector $[C \ d]$.

It is well known [28], [86] that when two's complement and less-than-double precision fixed-point arithmetic are used, summation nodes are allowed to overflow except before multiplier. Under the assumption that the input $u(n)$ and the output $y(n)$ of the filter are properly pre-scaled, the only signals which may have overflow are the states $\{x_k(n)\}$, which, therefore, have to be scaled. There exist different scaling schemes for preventing variables from overflow. The popularly used ones are the l_2 - and l_∞ -scalings. In this thesis, we will concentrate on the l_2 -scaling scheme.

The l_2 -scaling means that each state variable should have a unit variance

when the input is a white noise with a unit variance. This can be achieved if the controllability gramian W_c of (A, B, C, d) satisfies [8], [9]

$$W_c(i, i) = 1, \quad \forall i. \quad (2.20)$$

Under this assumption, the optimal FWL realization problem in terms of round-off noise gain is to find those realizations that minimize G given by (2.18) or (2.19) subject to (2.20):

$$\min_{W_c(i, i)=1, \forall i} G. \quad (2.21)$$

With W_o the observability gramian of (A, B, C, d) , it can be shown [8], [9] that

$$\min_{W_c(i, i)=1, \forall i} \text{tr}(W_o) = \frac{1}{K} \left(\sum_{i=1}^K \sigma_i \right)^2 \quad (2.22)$$

where $\{\sigma_i^2\}$ are the eigenvalues of the matrix $W_{oc}W_{oo}$, corresponding to any realization (A_0, B_0, C_0, d) of the filter. The state-space realization satisfying (2.22) is called the optimal realization with minimum roundoff noise gain, denoted as R_G .

It is clear that for an RBM implementation described with (2.15), the optimal FWL realization problem (2.21) is just solved by (2.22). However, for an RAM implementation, it seems very difficult and is still an open problem, just as the optimal sensitivity realization problem. Meanwhile, one notes that if the realization is fully parameterized (with nontrivial parameters), that is $m_k = K + 1, \forall k$, one has

$$G = [\text{tr}(W_o) + 1](K + 1) \quad (2.23)$$

with the minimum $\text{tr}(W_o)$ given by (2.22).

Error feedback is another effective technique for minimizing roundoff noise gain for the RBM implementations. It is known [46] that the roundoff error vector $e(n)$ can be fed back through a matrix D in the FWL filter (2.15), shown in Figure 2.2. The filter now can be characterized by the state-space model

$$\begin{cases} x^*(n+1) = A Q[x^*(n)] + Bu(n) + De(n) \\ y^*(n) = C Q[x^*(n)] + du(n) \end{cases} \quad (2.24)$$

With the similar approach above, one can obtain the roundoff noise gain for (2.24), given by

$$G = \text{tr}[(A - D)^T W_o (A - D) + C^T C] \triangleq \text{tr}(W_D). \quad (2.25)$$

Therefore, the problem here is to find an appropriate error feedback matrix D so that $\text{tr}(W_D) < \text{tr}(W_o)$ is satisfied subject to all the diagonal elements of W_c equal to unity.

2.3 Two-dimensional filters

Consider a single-input-single-output linear time-invariant 2-D digital filter:

$$H(z_1, z_2) = \frac{\sum_{k=0}^M \sum_{l=0}^N b_{kl} z_1^{-k} z_2^{-l}}{\sum_{k=0}^M \sum_{l=0}^N a_{kl} z_1^{-k} z_2^{-l}} \quad (2.26)$$

where $a_{00} = 1$.

It is called a *canonic realization* of $H(z_1, z_2)$ of order (M, N) if the total number of delay elements needed in the realization is $M + N$. Although for the

1-D digital filters such realizations can readily be obtained, in the case of 2-D filters they may not exist [87], [88]. However, a canonic realization always exists if the numerator or denominator of the transfer function is separable, which is given by

$$H(z_1, z_2) = \frac{(\sum_{k=0}^M b_{1k} z_1^{-k})(\sum_{l=0}^N b_{2l} z_2^{-l})}{\sum_{k=0}^M \sum_{l=0}^N a_{kl} z_1^{-k} z_2^{-l}} \quad (2.27)$$

for the separable-numerator case, and

$$H(z_1, z_2) = \frac{\sum_{k=0}^M \sum_{l=0}^N b_{kl} z_1^{-k} z_2^{-l}}{\sum_{k=0}^M a_{1k} z_1^{-k} \sum_{l=0}^N a_{2l} z_2^{-l}} \quad (2.28)$$

where $a_{10} = 1$, $a_{20} = 1$ for the separable-denominator case.

A filter by (2.27) or (2.28) can be represented by the state-space model [89], [90] as follows:

$$\begin{cases} \begin{bmatrix} x^h(i+1, j) \\ x^v(i, j+1) \end{bmatrix} = \underbrace{\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}}_A \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix} + \underbrace{\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}}_B u(i, j) \\ y(i, j) = \underbrace{\begin{bmatrix} C_1 & C_2 \end{bmatrix}}_C \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix} + du(i, j) \end{cases} \quad (2.29)$$

where $x^h(i, j) \in \mathcal{R}^{M \times 1}$ and $x^v(i, j) \in \mathcal{R}^{N \times 1}$ are the horizontal and vertical local state vector, respectively, $u(i, j)$ and $y(i, j)$ are the scalar input and output of the filter. $R \triangleq (A, B, C, d)$ with A, B, C and d real constant matrices of appropriate dimensions, is called a realization of 2-D system $H(z_1, z_2)$, satisfying

$$H(z_1, z_2) = d + C(Z - A)^{-1}B \quad (2.30)$$

where $Z \triangleq z_1 I_M \oplus z_2 I_N$.

Like 1-D case, denote $S_{H_2} \triangleq \{R : R \text{ satisfies (2.30)}\}$. S_{H_2} is characterized by

$$A = T^{-1}A_0T, \quad B = T^{-1}B_0, \quad C = C_0T \quad (2.31)$$

where $R_0 \triangleq \{A_0, B_0, C_0, d\} \in S_{H_2}$, and $T = T_1 \oplus T_2$, with $T_1 \in \mathcal{R}^{M \times M}$ and $T_2 \in \mathcal{R}^{N \times N}$ any nonsingular (transformation) matrices.

For a given 2-D filter structure, the L_2 -sensitivity measure is computed using (2.6) with $H(z)$ replaced by $H(z_1, z_2)$, where $\|\cdot\|_2$ is defined below

$$\|S(z_1, z_2)\|_2 \triangleq \left\{ \text{tr} \left[\frac{1}{(2\pi j)^2} \oint_{|z_1|=1} \oint_{|z_2|=1} S(z_1, z_2) S^H(z_1, z_2) \frac{dz_1 dz_2}{z_1 z_2} \right] \right\}^{1/2} \quad (2.32)$$

with $S(z_1, z_2)$ being any complex matrix-valued function of the complex variables z_1 and z_2 .

If $S(z_1, z_2) = D + L(Z - \Phi)^{-1}J$, it can be shown that

$$\|S(z_1, z_2)\|_2^2 = \text{tr}(DD^T + LW_c L^T) = \text{tr}(D^T D + J^T W_o J) \quad (2.33)$$

where W_c and W_o are the controllability and observability gramians of the realization (Φ, J, L, D) , respectively, which are given by

$$\begin{aligned} W_o &= \frac{1}{(2\pi j)^2} \oint_{|z_1|=1} \oint_{|z_2|=1} (Z - \Phi^T)^{-1} L^T L (Z - \Phi^T)^{-H} \frac{dz_1 dz_2}{z_1 z_2} \triangleq \begin{bmatrix} W_{o1} & W_{o2} \\ W_{o2}^T & W_{o4} \end{bmatrix} \\ W_c &= \frac{1}{(2\pi j)^2} \oint_{|z_1|=1} \oint_{|z_2|=1} (Z - \Phi)^{-1} J J^T (Z - \Phi)^{-H} \frac{dz_1 dz_2}{z_1 z_2} \triangleq \begin{bmatrix} W_{c1} & W_{c2} \\ W_{c2}^T & W_{c4} \end{bmatrix} \end{aligned}$$

where $W_{o1}, W_{c1} \in \mathcal{R}^{M \times M}$, $W_{o4}, W_{c4} \in \mathcal{R}^{N \times N}$. The algorithms to compute 2-D gramians are available in [61], [91], [92].

When the filter is implemented with the state-space realization by (2.29), it is shown [18] that M_{L_2} can be evaluated with

$$M_{L_2} = \sum_{k,l=1}^{M+N} \phi(a_{kl}) \bar{C} \bar{W}_{c_{kl}} \bar{C}^T + \sum_{k=1}^{M+N} [\phi(b_k) W_o(k, k) + \phi(c_k) W_c(k, k)] + \phi(d) \quad (2.34)$$

where $\phi(x)$ is defined in (2.12), W_c and W_o are the controllability and observability gramians of the realization (A, B, C, d) , and $\bar{W}_{c_{kl}}$ is the controllability gramian of $(\bar{A}_{kl}, \bar{B})$ with

$$\bar{A}_{kl} \triangleq U \begin{bmatrix} A & e_k e_l^T \\ \mathbf{0} & A \end{bmatrix} U, \quad \bar{B} \triangleq U \begin{bmatrix} \mathbf{0} \\ I_{M+N} \end{bmatrix} B, \quad U \triangleq \text{diag}(I_{\tilde{p}}, \begin{bmatrix} \mathbf{0} & I_p \\ I_p & \mathbf{0} \end{bmatrix}, I_{\tilde{p}})$$

where $p \triangleq \max\{M, N\}$, and $\tilde{p} \triangleq M + N - p$.

The problem of minimizing M_{L_2} of 2-D digital filters is also an open problem as far as we know.

As to the roundoff noise gain of the realization (2.29), it can also be evaluated by (2.18) and (2.19) for the RBM implementations and RAM implementations, respectively. It was shown in [55] that for the optimal state-space realization R_G in terms of roundoff noise under the l_2 -scaling,

$$\min_{W_c(i, i)=1, \forall i} \text{tr}(W_o) = \frac{1}{M} \left(\sum_{m=1}^M \sigma_{hm}^2 \right) + \frac{1}{N} \left(\sum_{n=1}^N \sigma_{vn}^2 \right) \quad (2.35)$$

where $\{\sigma_{hm}^2\}$ and $\{\sigma_{vn}^2\}$ are the eigenvalues of $W_{o_1} W_{c_1}$ and $W_{o_4} W_{c_4}$, respectively.

2.4 Genetic algorithms for optimization

As it was mentioned in Chapter 1, for the proposed ρ DFIIt and ρ RDFIIt structures, since free parameters take values from a finite space, they can be decided by exhaustive search. However, in practice, a good solution obtained in a reasonable time is better than the global optimum solution taking excessive time and/or computational resources. Because of that reason, an optimization approach could be utilized to efficiently solve the optimal structure problems. As to the ρ IDFIIt structure, it contains more introduced free parameters which are continuous in a given space. This means that exhaustive search is no longer applicable, which remains an unsolved problem in [1]. Hence, it further shows that an optimization approach should be used to solve our optimization problems. There are a great number of optimization approaches which can be used to solve the optimal structure problems (by searching for those free parameters which can minimize the FWL effects), among which, genetic algorithm (GA) is adopted in this thesis.

GAs were first studied by Holland and his colleagues at University of Michigan in 1975 [93] and have been applied to a number of design issues regarding both digital and analogue filters in the past few years [94] - [101]. They are based on the genetic processes of biological organisms. Over many generations, natural populations evolve according to the principles of natural selection and “survival of the fittest”. By mimicking this process, genetic algorithms are able to “evolve” solutions to real world problems, if they have been suitably encoded.

GAs work with a population of “individuals”, each representing a possible solution to a given problem. Each individual is assigned a “fitness score” according to how good a solution to the problem it is. The highly-fit individuals

are given opportunities to “reproduce”, by “cross breeding” with other individuals in the population. This produces new individuals as “offspring”, which share some features taken from each “parent”. The least fit members of the population are less likely to get selected for reproduction, and so “die out”. A whole new population of possible solutions is thus produced by selecting the best individuals from the current “generation”, and mating them to produce a new set of individuals. This new generation contains a higher proportion of the characteristics possessed by the good members of the previous generation. In this way, over many generations, good characteristics are spread throughout the population. By favoring the mating of the more fit individuals, the most promising areas of the search space are explored. If the GA has been designed well, the population will converge to an optimal solution to the problem.

The advantages of the GA technique over the traditional solution techniques for solving the optimization problems in this work are summarized as below:

- GAs do not require the search space to be continuous and differentiable².

One can notice in the following chapters that all the free parameters for the proposed ρ DFIIIt and ρ RDFIIIt structures are taking values from a discrete space. For the ρ IDFIIIt structure, it has two sets of free parameters, one set is taking values from a discrete space and the other one is taking values from a continuous space. Meanwhile, the functions of roundoff noise gain or sensitivity measure for each proposed structure are high nonlinear of the free parameters. Therefore, it is difficult to compute the derivative of these functions.

²A number of conventional methods for optimizing functions have been developed which rely on using information about the gradient of the function to guide the direction of search. If the derivative of the function can not be computed, because it is discontinuous, for example, these methods often fail.

Figure 2.3: Flowchart of the GA to solve the optimization problems.

- Binary encoding [71] is the most common and simplest method for GAs to encode each individual solution, in which there are only '1's and '0's. It is with the same format that we are using to represent the discrete free parameters for the proposed structures.
- For the proposed ρ IDFIIt structure, two sets of free parameters are required to be optimized at the same time: one is taking values from a discrete space, and the other is from a continuous space. Most traditional optimization approach can not simultaneously optimize the discrete parameters as well as the continuous parameters, while, GAs can solve this problem.

In what follows, we take an example to illustrate the procedure of a simple GA, which is shown in Figure 2.3.

Example 2.1: Consider the following problem

$$\begin{aligned} \max \quad & f(x_1, x_2) = 21.5 + x_1 \sin(2\pi x_1) + x_2 \sin(10\pi x_2) \\ \text{subject to} \quad & -3.0 \leq x_1 \leq 12.1, \quad 4.1 \leq x_2 \leq 5.8. \end{aligned}$$

The procedure of a simple GA solving Example 2.1 is presented in the fol-

lowing.

Step 1: Initialization

Denote the solution vector $\vartheta \triangleq [x_1 \ x_2]$. In order to initialize a population, we need to set values for:

1. l_i - the number of bits to represent x_i , $i = 1, 2$.
2. P - the total number of solution candidates, i.e., $\vartheta_1, \vartheta_2, \dots, \vartheta_P$.

So that the initial population can be represented by a matrix in dimension $\mathcal{R}^{P \times (l_1 + l_2)}$, in which each string represents a solution. Random initialization is commonly used although there are several other alternatives. If discretization is needed, l_i can be computed depending on the required precision. The selection of P influences the computation cost and the diversity of the population.

In our problem, assuming 4 decimal point precision on x_i is required, l_1 is chosen as the smallest integer satisfying $10^4[12.1 - (-3.0)] \leq 2^{l_1} - 1$, resulting in $l_1 = 18$, and similarly, $l_2 = 15$.

Step 2: Evaluation

Each solution candidate in the population is suitably decoded subject to the constraints $-3.0 \leq x_1 \leq 12.1$, $4.1 \leq x_2 \leq 5.8$. Then the fitness of every candidate is evaluated, i.e., the value of $f(x_1, x_2)$ in this example.

Step 3: Reproduction

Reproduction operator copies solution strings from the current generation into a mating pool which is as same as the population size, taking into consideration the values of $f(x_1, x_2)$ of individual solution strings. Strings with higher fitness values will likely be represented in higher numbers in the mating pool,

thereby passing on their genetic information into future generations. This operator is an artificial version of the Darwinian process of *natural selection*. For the optimization problem in this thesis, we employ the population replacement in the reproduction operator. During each iteration, we virtually choose $P/2$ best solutions from the parents and offsprings, respectively, as the new parents in the next generation. Although this may lead to one or two fit strings dominating the population, still in most cases it improves the simple GA and gives better results.

Step 4: Crossover

Crossover operator is applied to the re-ordered mating pool generated in Step 3. It is performed between two randomly selected parent strings in the mating pool and resulting in 2 offsprings. This operation exchanges partial information in each string, and therefore the offsprings have distinct parts of both parents. Since the population size is P , it will perform $P/2$ times to get P offsprings. Crossover probability p_c may be used to determine if there will be crossover between 2 parent strings. As we will perform crossover $P/2$ times, we can generate a random vector with $P/2$ elements each between 0 and 1. If the generated random number is smaller than p_c , the crossover operation is performed. Otherwise, no crossover is performed and the selected parents are passed on as offsprings to be processed by the next step.

After determining whether crossover will take place, two-point crossover (although there are other alternatives) is performed in our problem as follows:

- Two integer positions k_1, k_2 are randomly selected along the string $1 \leq k_1, k_2 \leq L - 1$ where L is the length of the string, i.e., 33 in our problem.

- Two new strings (offsprings) are generated by breaking up the parent strings at k_1 , k_2 and exchanging the information.
- For example, assume $k_1 = 5$, $k_2 = 26$,

Parent 1: 110001010111101001101011001011110
 Parent 2: 010110010101111000011010101010111
 Offspring 1: 110000010101111000011010101011110
 Offspring 2: 010111010111101001101011001010111

Step 5: Mutation

Mutation in general plays a minor role in the GAs. It will introduce a degree of diversity to the population, prevent a premature convergence, help to sample unexplored regions of the search space. An appropriate mutation operation in our problem will randomly change the bit values in the generated offsprings from '0' to '1' or from '1' to '0'. The probability of mutation operation, p_m , is usually small. In our problem, there are total $L \times P$ bits in the generated offsprings. Thus we can generate $L \times P$ random numbers between 0 and 1. If a generated value is smaller or equal to p_m , we change the bit value in the solution string at that position.

Step 6: Evaluation

Step 6 will be carried in the same approach to Step 2. Step 3 - 6 are repeated, which makes the highly fit solutions propagate in a increasing number, until the termination conditions in Step 7 are satisfied.

Step 7: Termination

Termination is the criterion by which the genetic algorithm decides whether

to continue searching or stop the search. Each of the enabled termination criterion is checked after each generation to see if it is time to stop. There are several types of termination criterion, for example:

- **Generation Number** - A termination method that stops the evolution when the user-specified max/min number of generations has been run.
- **Evolution Time** - A termination method that stops the evolution when the elapsed evolution time exceeds the user-specified max/min evolution time.
- **Fitness Threshold** - A termination method that stops the evolution when the best fitness in the current population becomes less than the user-specified fitness threshold and the objective is set to minimize the fitness.
- **Fitness Convergence** - A termination method that stops the evolution when the fitness is deemed as converged.
- **Population Convergence** - A termination method that stops the evolution when the population is deemed as converged.

Usually, user can choose any one type of the termination criterions or combine some of them.

For this example, since it is a much simple optimization problem, we choose completion of a given number of generations (generation number), say *gens*, as the termination condition. However, for the optimization problem in this thesis, the termination condition is a combination of the generation number, fitness threshold, and fitness convergence.

Table 2.1: Statistic History of the GA for Example 2.1

Generation	Mean Fitness	Max Fitness	Optimal x	
0	20.0097	28.1121	5.24794	5.69229
1	20.5109	31.3676	11.1644	5.40165
2	25.0796	30.8422	7.39586	5.46536
3	24.8226	30.9746	7.39218	5.46536
4	29.1870	36.0553	11.1672	5.43527
5	26.8065	36.8255	11.1672	5.64778
6	29.2083	36.0418	11.2234	5.62122
7	33.8577	36.7590	11.1654	5.64778
8	32.1361	36.4867	11.2235	5.62454
9	32.8414	36.7665	11.1654	5.65193
10	33.8150	36.7590	11.1654	5.64778
11	35.2510	37.3038	11.1818	5.64778
12	33.6739	38.0023	11.2097	5.65027
13	36.5556	37.9906	11.2095	5.65193
14	28.8883	37.3645	10.2659	5.65027
15	32.7485	37.3582	10.2655	5.65193

Assume $P = 10$, $p_c = 0.5$, $p_m = 0.05$, $gens = 15$, Table 2.1 shows the statistic history of the GA for this example, where the best solution occurs at 12th generation.

2.5 Summary

It is well known that the optimal roundoff noise realizations R_G yield a very good performance in terms of reducing FWL effects. They, however, possess too many nontrivial parameters, which limit their applications in practical implementation. In what follows in this thesis, several novel structures will be developed. These structures are very efficient and can yield a FWL performance comparable to the fully parameterized realization R_G and some existing filter structures by former researchers.

Chapter 3

A Generalized DFII_t Structure and its State-Space Realization in Digital Filter Implementation

3.1 Introduction

For real-time implementation, an efficient filter structure is a trade-off between performance against the FWL effects and computational simplicity. It is well known [8], [9] that for a digital filter of order K the number of nontrivial parameters in the classical optimal state-space realizations is $K^2 + 2K + 1$. The conventional shift operator based direct-forms (DF) such as direct-form II and direct form II transposed (denoted as zDFII_t) structures have $2K + 1$ parameters and are the most efficient structures in infinite precision sense. With the dynamic range constraint taken into account, the corresponding structures have $3K + 1$ parameters to implement. The direct-forms in delta operator have been

studied by researchers [28] - [33] in order to overcome the poor FWL performance of these DF structures. An extensive comparative study of different direct forms in delta operator was given in [33], where the transfer function was cascaded into second order sections and each section was implemented with a direct form in delta operator. It was shown there that among all the DFs, the DFII structure, has the lowest quantization noise level at output. In [28], the δ DFII structure was investigated for an arbitrary order IIR filter, where the concept of different coupling coefficients at different branch nodes was utilized to avoid overflow. The use of delta operator, defined as $\delta = (z - 1)/T_s$ with T_s the sampling period, was first promoted by Peterka [102] and Middleton and Goodwin [103] in estimation and control applications. Later on, the numerical properties of the delta operator, where T_s was replaced by a positive factor Δ , were investigated in [16] from a pure algebraic point of view. It was found that one could make the transfer function in delta operator have better numerical properties in the case where the poles of the transfer function are closer to $z = +1$ than $z = 0$ [34], [35]. This means that the delta operator based structures, including the δ DFII structure, have a very good performance just for the type of low-pass narrow band filters and may not yield a satisfactory performance for other types of filters.

In this chapter, motivated by [28] and based on the concept of polynomial operators [16], a set of special operators is derived, with which a generalized DFII structure, denoted as ρ DFII, is obtained. This structure, having $3K + 1$ non-trivial parameters plus K free parameters at choice, can be used for any types of digital filters. An equivalent state-space realization of the ρ DFII structure is obtained. The digital filter can also be implemented with this sparse state-space realization of $4K$ parameters. It is interesting to note that these two

structures can be optimized by a proper choice for the K free parameters in terms of minimizing the roundoff noise gain or sensitivity measure.

Our main contributions in this chapter are found in Sections 3.2 and 3.3. Based on the polynomial operator concept, two sparse structures, the so-called ρ DFIIt structure and its equivalent state-space realization are derived in Section 3.2. Section 3.3 is devoted to analyzing the performance of the two structures in terms of the RAM implementations. Expressions for roundoff noise gain and L_2 -based sensitivity measure are derived for both structures. It is interesting to note that for a given set of free parameters, the equivalent state-space realization always yields a smaller roundoff noise gain than that of the ρ DFIIt structure. Noting the roundoff noise gains depend on the choice for the free parameters, the problems of optimal structures are formed in Section 3.4 in terms of minimizing roundoff noise gain. Taking into account the fact that all the K free parameters have to be in FWL format for actual implementation, one can solve the optimal structure problems using exhaustive search in a parameter space of finite elements. The relationship between the free parameters and the poles of the filter is also established in terms of roundoff noise gain in this section, which implies that the optimal free parameters can be searched around the dominant poles of the filter. Numerical examples are given in Section 3.5 to illustrate the design procedure and to compare the performance of the optimal proposed structures with that of the z DFIIt and δ DFIIt as well as the classical fully parameterized optimal realizations.

3.2 Two new sparse structures

3.2.1 Polynomial parametrization

Denote

$$V_a \triangleq \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_k \\ \vdots \\ a_K \end{bmatrix}, \quad V_b \triangleq \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \\ \vdots \\ b_K \end{bmatrix}, \quad \bar{z} \triangleq \begin{bmatrix} z^K \\ z^{K-1} \\ \vdots \\ z^{K-k} \\ \vdots \\ 1 \end{bmatrix}. \quad (3.1)$$

Let $\bar{T} \in \mathcal{R}^{(K+1) \times (K+1)}$ be a nonsingular matrix and

$$\bar{q}(z) \triangleq \begin{bmatrix} q_0(z) & q_1(z) & \cdots & q_{K-k}(z) & \cdots & q_K(z) \end{bmatrix}^T = \bar{T} \bar{z}.$$

It can be shown that (2.1) can be rewritten as¹

$$H(z) = \frac{\sum_{k=0}^K \beta_k q_k(z)}{q_0(z) + \sum_{k=1}^K \alpha_k q_k(z)} = \frac{\beta_0 + \sum_{k=1}^K \frac{\beta_k q_k(z)}{q_0(z)}}{1 + \sum_{k=1}^K \frac{\alpha_k q_k(z)}{q_0(z)}} \quad (3.2)$$

where

$$\begin{cases} V_\alpha \triangleq \begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_k & \cdots & \alpha_K \end{bmatrix}^T \\ V_\beta \triangleq \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_k & \cdots & \beta_K \end{bmatrix}^T \end{cases} = \kappa^{-1} \bar{T}^{-T} \begin{bmatrix} V_a \\ V_b \end{bmatrix} \quad (3.3)$$

with κ a certain constant such that $V_\alpha(1) = 1$.

It is clear that the filter is now parametrized by (V_α, V_β) under the *polyno-*

¹

$$H(z) = \frac{V_b^T \bar{z}}{V_a^T \bar{z}} = \frac{\kappa^{-1} V_b^T \bar{T}^{-1} \bar{T} \bar{z}}{\kappa^{-1} V_a^T \bar{T}^{-1} \bar{T} \bar{z}} \triangleq \frac{V_\beta^T \bar{q}(z)}{V_\alpha^T \bar{q}(z)}$$

mial operators $\{q_k(z)\}$. The choice of the polynomial operators depends on the applications. In [16], the problem of how to choose these polynomials was investigated for system identification and parameter estimation. In this chapter, we will consider a special set of the polynomial operators, which leads to a very interesting structure for digital filter implementation.

Define

$$\rho_k(z) \triangleq \frac{z - \gamma_k}{\Delta_k}, \quad k = 1, 2, \dots, K \quad (3.4)$$

where $\{\gamma_k\}$ and $\{\Delta_k > 0\}$ are two sets of constants to be discussed later. For simplicity, we will use the notation ρ_k for $\rho_k(z)$ in the sequel. Let us consider the following choice of polynomial operators:

$$\begin{cases} q_k(z) = \prod_{l=k+1}^K \rho_l(z), & k = 0, 1, \dots, K-1 \\ q_K(z) = 1 \end{cases} \quad (3.5)$$

With such a choice, one can specify the corresponding transformation $\bar{T}$ and it is easy to see $\kappa = \prod_{k=1}^K \Delta_k$. The corresponding transfer function can be obtained from (3.2), based on which it can be shown that the output of $H(z)$ can be computed with the following equations

$$\begin{cases} y(n) &= \beta_0 u(n) + w_1(n) \\ &\vdots \\ w_{k-1}(n) &= \rho_{k-1}^{-1} [\beta_{k-1} u(n) - \alpha_{k-1} y(n) + w_k(n)] \\ w_k(n) &= \rho_k^{-1} [\beta_k u(n) - \alpha_k y(n) + w_{k+1}(n)] \\ w_{k+1}(n) &= \rho_{k+1}^{-1} [\beta_{k+1} u(n) - \alpha_{k+1} y(n) + w_{k+2}(n)] \\ &\vdots \end{cases} \quad (3.6)$$

with $w_{K+1}(n) = 0$. Figure 3.1 shows the corresponding realization diagram to

Figure 3.1: The ρ DFIIt structure Figure 3.2: A realization of operator ρ_k^{-1}

(3.6), called the ρ DFIIt structure, where $w_k(n)$ is the output of ρ_k^{-1} .

This structure possesses $3K + 1$ nontrivial parameters $\{\alpha_k\}$, $\{\beta_k\}$ and $\{\Delta_k\}$ plus K free parameters $\{\gamma_k\}$. Clearly, when $\gamma_k = 0$, $\forall k$, Figure 3.1 is the traditional z DFIIt structure, while with $\gamma_k = 1$, $\forall k$, one gets the δ DFIIt structure studied in [28]. With the K free parameters $\{\gamma_k\}$, as to be seen in Section 3.4, one has more degrees of freedom to minimize the FWL effects in the ρ DFIIt structure for any given digital filter.

3.2.2 Equivalent state-space realization

ρ_k^{-1} is a first order transfer function. One can implement it with the realization depicted in Figure 3.2. We choose $\{x_k(n)\}$ indicated in Figure 3.2 as the state variables and denote $x(n)$ as the state vector. It can be shown that the proposed

structure is equivalent to the following state-space realization:

$$\begin{cases} x(n+1) = A_\rho x(n) + B_\rho u(n) \\ y(n) = C_\rho x(n) + \beta_0 u(n) \end{cases}$$

where

$$A_\rho = D_\gamma + A_o D_\Delta, \quad B_\rho = \bar{\beta} - \beta_0 \bar{\alpha}, \quad C_\rho = e_1^T D_\Delta, \quad \beta_0 = b_0 \quad (3.7)$$

with e_1 the first elementary vector defined in Notations. $D_\gamma \triangleq \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_K)$, $D_\Delta \triangleq \text{diag}(\Delta_1, \Delta_2, \dots, \Delta_K)$ are two diagonal matrices, and

$$A_o \triangleq \begin{bmatrix} -\alpha_1 & 1 & 0 & \cdots & 0 & 0 \\ -\alpha_2 & 0 & 1 & \cdots & 0 & 0 \\ & & & \ddots & & \\ -\alpha_{K-1} & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_K & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \bar{\beta} \triangleq \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{K-1} \\ \beta_K \end{bmatrix}, \quad \bar{\alpha} \triangleq \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{K-1} \\ \alpha_K \end{bmatrix}. \quad (3.8)$$

The transfer function from $u(n)$ to $y(n)$ is then

$$H(z) = \beta_0 + C_\rho (zI_K - A_\rho)^{-1} B_\rho. \quad (3.9)$$

Let $\bar{V}_\alpha, \bar{V}_\beta$ be the V_α, V_β obtained with (3.3) when $\Delta_k = 1, \forall k$ and $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$, the corresponding equivalent state-space realization. It can be shown that the realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ with any given $\{\Delta_k\}$ can be obtained from $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$ with a similarity transformation, denoted as T_{sc} ,

$$A_\rho = T_{sc} \bar{A}_\rho T_{sc}^{-1}, \quad B_\rho = T_{sc} \bar{B}_\rho, \quad C_\rho = \bar{C}_\rho T_{sc}^{-1} \quad (3.10)$$

where

$$T_{sc} = \text{diag}(t_1, t_2, \dots, t_K), \quad t_k = \prod_{m=1}^k \Delta_m^{-1}, \quad \forall k. \quad (3.11)$$

Such T_{sc}^{-1} is actually a diagonal scaling similarity transformation.

3.2.3 Choice of $\{\Delta_k\}$ and l_2 -scaling

The parameters $\{\Delta_k\}$, called *coupling coefficients* in [28], affect the dynamic range of the signals at the branch nodes, which are actually the state variables $\{x_k(n)\}$ in Figure 3.2. In order to avoid overflow, these coupling coefficients have to be used to scale the state variables. As mentioned in Section 2.2.2, the l_2 -scaling means that each state variable should have a unit variance when the input is a white noise with a unit variance. This can be achieved if $W_c^\rho(k, k) = 1, \forall k$, where W_c^ρ is the controllability gramian of $(A_\rho, B_\rho, C_\rho, \beta_0)$. See (2.20).

Denote $\bar{W}_c^\rho$, corresponding to $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$. It follows from (3.10) - (3.11) that $W_c^\rho = T_{sc} \bar{W}_c^\rho T_{sc}^T$. Therefore, the l_2 -scaling can be achieved if $t_k^2 \bar{W}_c^\rho(k, k) = 1, \forall k$, which leads to

$$\Delta_1 = \sqrt{\bar{W}_c^\rho(1, 1)}, \quad \Delta_k = \sqrt{\frac{\bar{W}_c^\rho(k, k)}{\bar{W}_c^\rho(k-1, k-1)}}, \quad k = 2, 3, \dots, K. \quad (3.12)$$

In the sequel, the structures under discussion are assumed to have been

l_2 -scaled. It is interesting to note that with (3.12), one has from (3.5)²

$$\begin{cases} \kappa = \prod_{k=1}^K \Delta_k = \sqrt{\bar{W}_c^\rho(K, K)} \\ q_0(z) = \sqrt{\frac{1}{\bar{W}_c^\rho(K, K)}} \prod_{l=1}^K (z - \gamma_l), \quad q_K(z) = 1 \\ q_k(z) = \sqrt{\frac{\bar{W}_c^\rho(k, k)}{\bar{W}_c^\rho(K, K)}} \prod_{l=k+1}^K (z - \gamma_l), \quad k = 1, 2, \dots, K-1 \end{cases} \quad (3.13)$$

For a given digital filter $H(z)$ and a specified set $\{\gamma_k\}$, one, on the one hand, can compute the l_2 -scaled coupling coefficients with (3.12) and hence implement $H(z)$ with the ρ DFIIt structure as shown in Figures 3.1 and 3.2. On the other hand, one can also implement $H(z)$ with the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ with (3.7). As seen, both structures are very sparse. In the next section, we will analyze the performance of the two structures.

3.3 Performance analysis

In this section, we investigate the performance of the ρ DFIIt structure and its equivalent state-space realization in terms of roundoff noise gain and sensitivity measure. Their performance depends on the choice of $\{\gamma_k\}$. These parameters, generally speaking, can be chosen from the set of the real numbers. Since we confine ourselves to *fixed-point* implementation for which the FWL effects are more serious, it is assumed in the sequel that γ_k is absolutely not bigger than one. For a fixed-point implementation of B_c -bits, every γ_k has to be truncated

2

$$\begin{aligned} q_0(z) &= \prod_{l=1}^K \rho_l(z) = \frac{\prod_{l=1}^K (z - \gamma_l)}{\prod_{l=1}^K \Delta_l} = \sqrt{\frac{1}{\bar{W}_c^\rho(K, K)}} \prod_{l=1}^K (z - \gamma_l) \\ q_k(z) &= \prod_{l=k+1}^K \rho_l(z) = \frac{\prod_{l=k+1}^K (z - \gamma_l)}{\prod_{l=k+1}^K \Delta_l} = \sqrt{\frac{\bar{W}_c^\rho(k, k)}{\bar{W}_c^\rho(K, K)}} \prod_{l=k+1}^K (z - \gamma_l), \quad k = 1, 2, \dots, K-1 \end{aligned}$$

or rounded into a B_γ -bit number (with $B_\gamma \leq B_c$), which is of the form

$$v = \pm \sum_{m=1}^{B_\gamma} b_m 2^{-m}, \quad b_m = 0, 1, \quad \forall m$$

except that it is equal to ± 1 . Taking these into account, it is easy to understand that $\{\gamma_k\}$ should take values within a *discrete* space, denoted as S_γ , which is the union of two sets containing $2^{B_\gamma+1} + 1$ elements:

$$S_\gamma \triangleq \{-1, 1\} \cup \left\{ \pm \sum_{m=1}^{B_\gamma} b_m 2^{-m}, \quad b_m = 0, 1, \quad \forall m \right\}. \quad (3.14)$$

In what follows, it is assumed that all the parameters $\gamma_k \in S_\gamma$. As a consequence of that, these parameters have no contribution to the overall sensitivity measure but they do cause roundoff noise if they are nontrivial.

3.3.1 Roundoff noise gain

In this chapter, the ρ DFII structure and its equivalent state-space realization are studied in terms of RAM scheme implementations.

3.3.1.1 ρ DFII structure

It has been shown that the roundoff noise gain due to a nontrivial multiplier parameter in the structure can be computed with (2.14). We note that in this structure, the parameters are $\{\alpha_k\}$, $\{\beta_k\}$, $\{\Delta_k\}$ and $\{\gamma_k\}$. Let us consider the effect of roundoff noise due to the term $\alpha_k y(n)$ on the output. In that case,

(3.6) becomes

$$\left\{ \begin{array}{lcl} y^*(n) & = & \beta_0 u(n) + w_1^*(n) \\ & \vdots & \\ w_{k-1}^*(n) & = & \rho_{k-1}^{-1} [\beta_{k-1} u(n) - \alpha_{k-1} y^*(n) + w_k^*(n)] \\ w_k^*(n) & = & \rho_k^{-1} \{ \beta_k u(n) - [\alpha_k y^*(n) + \psi(\alpha_k) e_{\alpha_k}(n)] + w_{k+1}^*(n) \} \\ w_{k+1}^*(n) & = & \rho_{k+1}^{-1} [\beta_{k+1} u(n) - \alpha_{k+1} y^*(n) + w_{k+2}^*(n)] \\ & \vdots & \end{array} \right. \quad (3.15)$$

where $\psi(\alpha_k) e_{\alpha_k}(n) \triangleq Q[\alpha_k y^*(n)] - \alpha_k y^*(n)$ is the roundoff noise due to the quantizer $Q[\cdot]$, with $\psi(x)$ defined as

$$\psi(x) \triangleq \begin{cases} 1, & x \text{ is nontrivial} \\ 0, & \text{otherwise} \end{cases} \quad (3.16)$$

Denote

$$\Delta y(n) \triangleq y^*(n) - y(n), \quad \Delta w_k(n) \triangleq w_k^*(n) - w_k(n)$$

it follows from (3.6) and (3.15) that

$$\left\{ \begin{array}{lcl} \Delta y(n) & = & \Delta w_1(n) \\ & \vdots & \\ \Delta w_{k-1}(n) & = & \rho_{k-1}^{-1} [-\alpha_{k-1} \Delta y(n) + \Delta w_k(n)] \\ \Delta w_k(n) & = & \rho_k^{-1} [-\psi(\alpha_k) e_{\alpha_k}(n) - \alpha_k \Delta y(n) + \Delta w_{k+1}(n)] \\ \Delta w_{k+1}(n) & = & \rho_{k+1}^{-1} [-\alpha_{k+1} \Delta y(n) + \Delta w_{k+2}(n)] \\ & \vdots & \end{array} \right. \quad (3.17)$$

Comparing above equation with (3.6), one can see that the transfer function from $-\psi(\alpha_k) e_{\alpha_k}(n)$ to $\Delta y(n)$, denoted as $H_k(z)$, is given by (3.9) with $\beta_0 = 0$

and $\bar{\beta}$ replaced by e_k , the k th elementary vector. Therefore,

$$H_k(z) = C_\rho(zI_K - A_\rho)^{-1}e_k, \quad \forall k. \quad (3.18)$$

It follows from (2.9) and (2.14) that the roundoff noise gain due to parameter α_k is given by

$$G_{\alpha_k} = \psi(\alpha_k) \|H_k(z)\|_2^2 = \psi(\alpha_k) e_k^T W_o^\rho e_k \triangleq \psi(\alpha_k) G_k, \quad \forall k$$

where W_o^ρ is the observability of the realization $(A_\rho, B_\rho, C_\rho, \beta_0)$.

Similarly, one can analyze the roundoff noise gain due to the parameter β_k , with $k = 1, 2, \dots, K$. Using the same procedure, one can show that

$$G_{\beta_k} = \psi(\beta_k) G_k, \quad \forall k.$$

Now, let us consider the term $\beta_0 u(n)$. With $\psi(\beta_0)e_{\beta_0}(n)$ denoting the corresponding roundoff noise, one has

$$\left\{ \begin{array}{l} \Delta y(n) = \psi(\beta_0)e_{\beta_0}(n) + \Delta w_1(n) \\ \vdots \\ \Delta w_{k-1}(n) = \rho_{k-1}^{-1}[-\alpha_{k-1}\Delta y(n) + \Delta w_k(n)] \\ \Delta w_k(n) = \rho_k^{-1}[-\alpha_k\Delta y(n) + \Delta w_{k+1}(n)] \\ \Delta w_{k+1}(n) = \rho_{k+1}^{-1}[-\alpha_{k+1}\Delta y(n) + \Delta w_{k+2}(n)] \\ \vdots \end{array} \right. \quad (3.19)$$

One observes via comparing the above equation with (3.6) that the transfer function from $\psi(\beta_0)e_{\beta_0}(n)$ to $\Delta y(n)$, denoted as $H_0(z)$, is given by (3.9) with

$\beta_0 = 1$ and $\bar{\beta} = \mathbf{0}$. Therefore, one has

$$H_0(z) = 1 - C_\rho(zI_K - A_\rho)^{-1}\bar{\alpha} \quad (3.20)$$

and hence

$$G_{\beta_0} = \psi(\beta_0)\|H_0(z)\|_2^2 = \psi(\beta_0)(1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) \triangleq \psi(\beta_0)G_0.$$

With $w_{k+1}(n)$ replaced by $\Delta_{k+1}x_{k+1}(n)$ in (3.6), one can see that the roundoff noise produced by Δ_{k+1} can be considered as that by β_k . Therefore, one readily has

$$G_{\Delta_k} = \psi(\Delta_k)G_{k-1}, \quad \forall k.$$

For parameter γ_k , it is easy to see that the transfer function from $\psi(\gamma_k)e_{\gamma_k}(n)$ to $\Delta y(n)$ is $H_k(z)$ defined above. In fact, this roundoff noise can be considered as the one produced by $\beta_k u(n)$. Therefore,

$$G_{\gamma_k} = \psi(\gamma_k)G_k, \quad \forall k.$$

Therefore, the total roundoff noise gain of the ρ DFII structure is then defined as

$$G_\rho \triangleq \sum_{k=1}^K G_{\alpha_k} + \sum_{k=0}^K G_{\beta_k} + \sum_{k=1}^K G_{\Delta_k} + \sum_{k=1}^K G_{\gamma_k} \triangleq \sum_{k=0}^K \xi_k G_k \quad (3.21)$$

where the coefficients ξ_k can be specified with the expressions obtained above

of roundoff noise gain for all the parameters:

$$\begin{cases} \xi_0 = \psi(\Delta_1) + \psi(\beta_0) \\ \xi_k = \psi(\alpha_k) + \psi(\beta_k) + \psi(\gamma_k) + \psi(\Delta_{k+1}), \quad k = 1, 2, \dots, K-1 \\ \xi_K = \psi(\alpha_K) + \psi(\beta_K) + \psi(\gamma_K) \end{cases} \quad (3.22)$$

3.3.1.2 Equivalent state-space realization

It follows from (3.7) - (3.8) that the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ has the following sparse structure:

$$\begin{aligned} A_\rho &= \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & \cdots & 0 & 0 & 0 \\ & & & \vdots & & & \\ a_{(K-1)1} & 0 & 0 & \cdots & 0 & a_{(K-1)(K-1)} & a_{(K-1)K} \\ a_{K1} & 0 & 0 & \cdots & 0 & 0 & a_{KK} \end{bmatrix} \\ B_\rho &= \begin{bmatrix} b_1 & b_2 & \cdots & b_{K-1} & b_K \end{bmatrix}^T, \quad C_\rho = \begin{bmatrix} c_1 & 0 & \cdots & 0 & 0 \end{bmatrix}. \end{aligned} \quad (3.23)$$

In fact, applying (2.19) to (3.23), the corresponding roundoff noise gain, denoted as G_{ss} , can be given by

$$G_{ss} = \sum_{k=1}^K \zeta_k G_k + \zeta_0 \quad (3.24)$$

where G_k is the same as defined before and

$$\begin{cases} \zeta_0 &= \psi(c_1) + \psi(\beta_0) \\ \zeta_1 &= \psi(a_{11}) + \psi(a_{12}) + \psi(b_1) \\ \zeta_k &= \psi(a_{k1}) + \psi(a_{kk}) + \psi(a_{k(k+1)}) + \psi(b_k), \quad k = 2, 3, \dots, K-1 \\ \zeta_K &= \psi(a_{K1}) + \psi(a_{KK}) + \psi(b_K) \end{cases} \quad (3.25)$$

Assuming the parameters in both structures are nontrivial, it is interesting to note that

$$G_\rho - G_{ss} = 2\bar{\alpha}^T W_o^\rho \bar{\alpha} + e_1^T W_o^\rho e_1 > 0$$

which implies that for a given set of $\{\gamma_k\}$, the equivalent state-space realization always yields a smaller roundoff noise gain than the ρ DFII structure.

3.3.2 L_2 -sensitivity measure

Here, we will analyze the transfer function sensitivity with respect to those parameters of non-FWL format for both the ρ DFII structure and its equivalent state-space realization. Unlike [28], where a mixture of L_1 and L_2 measures was used, a pure L_2 -based sensitivity measure will be derived for the ρ DFII structure.

3.3.2.1 ρ DFII structure

Since $\{\gamma_k\}$ are assumed of the FWL format, as indicated in (3.14), it follows from (2.6) that the L_2 -norm based sensitivity measure is given by

$$M_\rho = \sum_{\forall \tau_l, \tau_l \in \{\alpha_k, \beta_k, \Delta_k\} \text{ and } \tau_l \in S_{FWL}} \left\| \frac{\partial H(z)}{\partial \tau_l} \right\|_2^2. \quad (3.26)$$

With some manipulations, it follows from (3.7) - (3.9) that³

$$\begin{cases} \frac{\partial H(z)}{\partial \beta_0} = 1 - C_\rho(zI_K - A_\rho)^{-1}\bar{\alpha} \\ \frac{\partial H(z)}{\partial \beta_k} = C_\rho(zI_K - A_\rho)^{-1}e_k \\ \frac{\partial H(z)}{\partial \alpha_k} = [-\beta_0 - C_\rho(zI_K - A_\rho)^{-1}B_\rho]C_\rho(zI_K - A_\rho)^{-1}e_k \\ \frac{\partial H(z)}{\partial \Delta_k} = [e_1^T e_k + C_\rho(zI_K - A_\rho)^{-1}A_\rho e_k]e_k^T(zI_K - A_\rho)^{-1}B_\rho \end{cases} \quad (3.27)$$

One observes that all the sensitivity functions in (3.27) are of the form $S(z) = [D_2 + C_2(zI_2 - A_2)^{-1}B_2][D_1 + C_1(zI_1 - A_1)^{-1}B_1] \triangleq S_2(z)S_1(z)$, where I_k , $k = 1, 2$ denotes the identity matrix of a proper dimension. For $S(z) = \partial H(z)/\partial \beta_k$, for example, one has $A_1 = A_\rho$, $B_1 = e_k$, $C_1 = C_\rho$, $D_1 = 0$ and $D_2 = 1$, $C_2 = 0$ with A_2, B_2 arbitrary scalars. It is easy to verify that

$$\begin{aligned} S(z) &= D_2D_1 + \begin{bmatrix} D_2C_1 & C_2 \end{bmatrix} \left[z \begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix} - \begin{bmatrix} A_1 & 0 \\ B_2C_1 & A_2 \end{bmatrix} \right]^{-1} \begin{bmatrix} B_1 \\ B_2D_1 \end{bmatrix} \\ &\triangleq \bar{D} + \bar{C}(z\bar{I} - \bar{A})^{-1}\bar{B}. \end{aligned}$$

Therefore, M_ρ by (3.26) can be evaluated easily with (2.9).

3.3.2.2 Equivalent state-space realization

As to the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$, the corresponding L_2 -based sensitivity measure, denoted as M_{ss} , can be evaluated directly with

³Here, we take $\partial H(z)/\partial \alpha_k$ for example:

$$\begin{aligned} \frac{\partial H(z)}{\partial \alpha_k} &= C_\rho(zI - A_\rho)^{-1} \frac{\partial A_\rho}{\partial \alpha_k} (zI - A_\rho)^{-1} B_\rho + C_\rho(zI - A_\rho)^{-1} \frac{\partial B_\rho}{\partial \alpha_k} \\ &= -C_\rho(zI - A_\rho)^{-1} e_k e_1^T D_\Delta (zI - A_\rho)^{-1} B_\rho - C_\rho(zI - A_\rho)^{-1} \beta_0 e_k \\ &= C_\rho(zI - A_\rho)^{-1} e_k [-\beta_0 - e_1^T D_\Delta (zI - A_\rho)^{-1} B_\rho] \\ &= C_\rho(zI - A_\rho)^{-1} e_k [-\beta_0 - C_\rho(zI - A_\rho)^{-1} B_\rho] \end{aligned}$$

(2.11) - (2.13), where noting (3.23), one has $\phi(b_k) = 1, \forall k, \phi(\beta_0) = 1$ and

$$\begin{aligned} \phi(a_{kl}) &= \begin{cases} 1, & (k, l) = (m, 1), \text{ with } m = 1, 2, \dots, K \\ 1, & (k, l) = (m, m+1), \text{ with } m = 1, 2, \dots, K-1 \\ 0, & \text{otherwise} \end{cases} \\ \phi(c_k) &= \begin{cases} 1, & k = 1 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (3.28)$$

Unlike the roundoff noise gain, there exists no direct relationship between M_ρ and M_{ss} .

3.4 Optimal structures

So far, we have derived the expressions for evaluating the roundoff noise gain and the L_2 -based sensitivity measure for both the ρ DFIIt structure and its equivalent state-space realization of a digital filter initially parameterized with (V_a, V_b) . For a given set of $\{\gamma_k\}$, the detailed procedure can be summarized as below:

- Compute $\bar{V}_\alpha, \bar{V}_\beta$ using (3.3) with $\Delta_k = 1, \forall k$, the given $\{\gamma_k\}$ and (V_a, V_b) , then form the corresponding state-space realization $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$.
- Evaluate the controllability gramian $\bar{W}_c^\rho$, say using MATLAB command *dgram.m*, and then compute the l_2 -scaling factors $\{\Delta_k\}$ via (3.12), with which the l_2 -scaled (V_α, V_β) and hence the l_2 -scaled equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ can be obtained.
- Compute the observability gramian W_o^ρ of the $(A_\rho, B_\rho, C_\rho, \beta_0)$. The round-

off noise gains G_ρ and G_{ss} can then be evaluated with (3.21) and (3.24), respectively.

- With the obtained l_2 -scaled state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$, M_ρ and M_{ss} can be computed as discussed previously.

Denote

$$\bar{\gamma} \triangleq \begin{bmatrix} \gamma_1 & \cdots & \gamma_k & \cdots & \gamma_K \end{bmatrix} \quad (3.29)$$

and $\bar{S}_\gamma \in \mathcal{R}^{1 \times K}$ as the space within which $\bar{\gamma}$ takes values. Therefore, one has

$$\bar{S}_\gamma \triangleq \{\bar{\gamma} : \gamma_k \in S_\gamma, \forall k\} \quad (3.30)$$

where S_γ is defined in (3.14). For convenience, we denote $\bar{\gamma}_z$ and $\bar{\gamma}_\delta$ for $\gamma_k = 0, \forall k$ and $\gamma_k = 1, \forall k$, respectively.

It is easy to understand that both G_ρ and G_{ss} are functions of $\bar{\gamma}$. An interesting problem is to identify those $\bar{\gamma}$ in $\bar{S}_\gamma$ such that G_ρ or G_{ss} is minimized, which leads to the following optimal structure problems in terms of minimizing roundoff noise gain:

$$\begin{aligned} \min_{\bar{\gamma} \in \bar{S}_\gamma} G_\rho &\implies \bar{\gamma}(G_\rho^{\text{opt}}) \\ \min_{\bar{\gamma} \in \bar{S}_\gamma} G_{ss} &\implies \bar{\gamma}(G_{ss}^{\text{opt}}) \end{aligned} \quad (3.31)$$

where $\bar{\gamma}(G_\rho^{\text{opt}})$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$ are the optimal solution that can minimize G_ρ and G_{ss} , respectively.

Though it is very hard to find closed-form solutions to the above minimization problems since G_ρ and G_{ss} are highly nonlinear function of $\bar{\gamma}$, the problems,

however, can be solved easily using exhaustive search because the space $\bar{S}_\gamma$ given by (3.30) contains $(2^{B_\gamma+1} + 1)^K$ elements, where K is the order of the filter and B_γ is the number of bits for implementing $\{\gamma_k\}$, which is typically 4 or 8 bits⁴. Clearly, repeating the procedure summarized above a finite number of times, one can find the $\bar{\gamma}(G_\rho^{\text{opt}})$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$. It is true that this number becomes very large when K is big, but for off-line design that is not a big problem. In practice, a good solution obtained in a reasonable time is more appreciated than the global optimal solution taking excessive time. Hence, GAs [71], which have been introduced in Section 2.4, can be applied to attack (3.31) for this purpose.

It is interesting to note that the total roundoff noise gain G_ρ by (3.21) and G_{ss} by (3.24) can be expressed in terms of $\{G_k\}$, where $G_k = \|H_k(z)\|_2^2$ with $H_k(z)$ defined in the previous section. On the one hand, it follows from (3.27) that $\partial H(z)/\partial \beta_k = H_k(z)$, while, on the other hand, from (3.2) one has

$$\frac{\partial H(z)}{\partial \beta_k} = \frac{q_k(z)}{V_\alpha^T \bar{q}(z)} = \frac{\kappa q_k(z)}{D(z)}, \quad \forall k$$

where $D(z)$ is the denominator of $H(z)$, see (2.1). With (3.13), one can specify $\partial H(z)/\partial \beta_k$, $\forall k$:

$$\frac{\partial H(z)}{\partial \beta_k} = \begin{cases} \frac{\prod_{l=1}^K (z - \gamma_l)}{D(z)}, & k = 0 \\ \frac{\sqrt{\bar{W}_c^p(k, k)} \prod_{l=k+1}^K (z - \gamma_l)}{D(z)}, & k = 1, 2, \dots, K-1 \\ \frac{\sqrt{\bar{W}_c^p(K, K)}}{D(z)}, & k = K \end{cases}$$

Let $D(z) = \prod_{i=1}^K (z - \mu_i)$. To simplify the discussion, assume that the filter has no repeated poles, that is $\mu_l \neq \mu_i$, $\forall l \neq i$, and no poles at $z = 0$. According

⁴The optimal filter structure problem does not make much sense for high precision implementation since all the structures yield almost the same performance in such a situation.

to the *Residue Theorem*, it follows from (2.7) that

$$G_k = \begin{cases} \prod_{l=1}^K \frac{\gamma_l}{\mu_l} + \sum_{i=1}^K \frac{\prod_{l=1}^K (\mu_i - \gamma_l)(1 - \gamma_l \mu_i)}{\prod_{l \neq i}^K (\mu_i - \mu_l) \prod_{l=1}^K (1 - \mu_i \mu_l)} \mu_i^{-1}, & k = 0 \\ \bar{W}_c^\rho(k, k) \sum_{i=1}^K \frac{\prod_{l=k+1}^K (\mu_i - \gamma_l)(1 - \gamma_l \mu_i)}{\prod_{l \neq i}^K (\mu_i - \mu_l) \prod_{l=1}^K (1 - \mu_i \mu_l)} \mu_i^{k-1} \\ \quad \triangleq \sum_{i=1}^K \text{Res}_k(i), & k = 1, 2, \dots, K-1 \\ \bar{W}_c^\rho(K, K) \sum_{i=1}^K \frac{1}{\prod_{l \neq i}^K (\mu_i - \mu_l) \prod_{l=1}^K (1 - \mu_i \mu_l)} \mu_i^{K-1}, & k = K \end{cases} \quad (3.32)$$

Let $T_q = \bar{T}(2 : K+1, 2 : K+1) \in \mathcal{R}^{K \times K}$ be the submatrix of the transformation matrix $\bar{T}$ corresponding to (3.5) with $\Delta_k = 1, \forall k$, for a given $\{\gamma_k\}$. It can be shown that the normalized realization $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho)$ can be obtained from a realization, denoted as (A_z, B_z, C_z) , with the similarity transformation $T = T_q^T$, where (A_z, B_z, C_z) is the equivalent state-space realization, corresponding to the normalized z DFII structure, i.e., $\Delta_k = 1, \gamma_k = 0, \forall k$. Denote $\bar{W}_c^z$ as the corresponding controllability gramian of this realization. Therefore, one has

$$\bar{W}_c^\rho = T^{-1} \bar{W}_c^z T^{-T} = T_q^{-T} \bar{W}_c^z T_q^{-1}. \quad (3.33)$$

Note that T_q , an upper triangular matrix with all diagonal elements equal to one, is determined uniquely by $\gamma_k, k = 2, 3, \dots, K$. Equation (3.33) yields the relationship between $\{\gamma_k\}$ and $\{\bar{W}_c^\rho(k, k)\}$. Due to this complicated relationship, $\{G_k\}$ and hence G_ρ, G_{ss} are a highly nonlinear but well-behaved function of $\{\gamma_k\}$.

We point out that like G_ρ and G_{ss} , the sensitivity measures M_ρ and M_{ss} are also functions of $\bar{\gamma}$. Therefore, the optimal structure problem in terms of minimizing sensitivity can be formulated and solved using exactly the same

procedure as the one for solving (3.31). In addition, it is easy to see from (3.26) - (3.27) that the sensitivities with respect to the parameters $\{\beta_k\}$ are strictly related to $\{G_k\}$. Without giving the details, we mention that the sensitivities with respect to $\{\alpha_k\}$ and $\{\Delta_k\}$ can be rewritten in terms of the poles $\{\mu_i\}$ and the parameters $\{\gamma_k\}$, which are very similar to (3.32) but more complicated.

We observe that for a given $\{\gamma_k\}$, G_k could be very large if $H(z)$ has some poles clustered together and close to the unit circle. One can see from (3.32) that $\text{Res}_k(i) = 0$ if $\gamma_l = \mu_i$. It is our conjecture that G_k and hence G_ρ , G_{ss} can be much reduced if $\{\gamma_k\}$ are chosen close to the dominant poles. As a result of that, the exhaustive search for the optimal $\bar{\gamma}$ can be carried out efficiently within the neighborhood of these poles. The same applies to the sensitivity analysis. This suggests that the optimal $\bar{\gamma}$ can be searched within the neighborhood of those poles. These arguments have been confirmed by the use of δ -operator based structures in implementing narrow band low-pass filters, where the poles are all clustered around $z = +1$, and coincide with our simulation results to be presented in the next section. As a result of that, the procedure of searching for $\bar{\gamma}$ can be done efficiently.

3.5 Numerical examples

In this section, we present two design examples to confirm the theoretical results obtained in Section 3.4, illustrate the performance of the two proposed structures, and compare them with some existing structures. For convenience, denote the equivalent state-space realization of the ρ DFII structure as ρ SS.

In the two examples, $B_\gamma = 4$ is assumed. For no loss of generality, $\bar{\gamma}(G_\rho^{\text{opt}})$

Table 3.1: Comparison Results for Example 3.1

$\bar{\gamma}$	G_ρ	G_{ss}	M_ρ	M_{ss}	N_ρ	N_{ss}
$\bar{\gamma}_z$	8.3497×10^5	4.2017×10^5	4.2837×10^6	4.8823×10^6	13	13
$\bar{\gamma}_\delta$	1.1841×10^1	7.5808×10^0	1.9064×10^1	4.1020×10^1	13	13
$\bar{\gamma}(G_\rho^{\text{opt}})$	9.6240×10^0	5.2789×10^0	1.1639×10^1	2.0539×10^1	15	15
$\bar{\gamma}(G_{ss}^{\text{opt}})$	1.4615×10^1	4.2864×10^0	1.4143×10^1	1.6777×10^1	16	16

and $\bar{\gamma}(G_{ss}^{\text{opt}})$, i.e., the solutions to (3.31), are found using exhaustive search within the corresponding $\bar{S}_\gamma$ (3.30) for each example in this chapter. The ρ DFIIt and ρ SS structures of $\bar{\gamma}(G_\rho^{\text{opt}})$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$ are compared with those of $\bar{\gamma}_z$ and $\bar{\gamma}_\delta$ as well as the fully parameterized optimal roundoff noise realization R_G (see Chapter 2) in terms of roundoff noise gain and sensitivity measure.

Example 3.1: This is a fourth order Butterworth low-pass filter with very narrow bandwidth, generated with MATLAB command $[V_b, V_a] = \text{butter}(4, 0.05)$. This filter has a normalized bandwidth 0.025. The corresponding poles are $p_{1,2} = 0.9319 \pm j0.1364$ and $p_{3,4} = 0.8630 \pm j0.0523$, which are clustered around $z = +1$.

For this example, R_G yields a roundoff noise gain 7.7777 and a L_2 -sensitivity measure 30.2693, which has 25 nontrivial parameters. Computations show that

$$\begin{aligned}\bar{\gamma}(G_\rho^{\text{opt}}) &= \begin{bmatrix} 1 & 1 & 0.9375 & 0.9375 \end{bmatrix} \\ \bar{\gamma}(G_{ss}^{\text{opt}}) &= \begin{bmatrix} 1 & 0.875 & 0.875 & 0.875 \end{bmatrix}\end{aligned}$$

Table 3.1 shows the statistics on the roundoff noise gain G_ρ , the L_2 -sensitivity measure M_ρ , and the number of nontrivial parameters N_ρ for the ρ DFIIt structure, and G_{ss} , M_{ss} and N_{ss} for the ρ SS structure with each of the four $\bar{\gamma}$ sets

Table 3.2: Comparison Results for Example 3.2

$\bar{\gamma}$	G_ρ	G_{ss}	M_ρ	M_{ss}	N_ρ	N_{ss}
$\bar{\gamma}_z$	2.8649×10^5	2.7042×10^5	5.3365×10^5	6.8851×10^5	19	19
$\bar{\gamma}_\delta$	5.1568×10^8	4.6017×10^8	1.7810×10^{10}	1.1435×10^9	19	19
$\bar{\gamma}(G_\rho^{\text{opt}})$	1.9910×10^1	1.2375×10^1	2.2463×10^1	2.9600×10^1	24	24
$\bar{\gamma}(G_{ss}^{\text{opt}})$	2.0121×10^1	1.2326×10^1	2.2675×10^1	2.9636×10^1	24	24

mentioned above.

Example 3.2: The second example to be considered is a Butterworth band-pass filter of sixth order with a normalized bandwidth 0.075, generated with MATLAB command $[V_b, V_a] = \text{butter}(3, [0.75 \ 0.90])$. The corresponding poles are $p_{1,2} = -0.6327 \pm j0.5747$, $p_{3,4} = -0.8809 \pm j0.2937$, $p_{5,6} = -0.7071 \pm j0.3359$. We note that the real parts of the poles are around $z = -0.75$.

The corresponding R_G yields a roundoff noise gain 13.5869 and a L_2 -sensitivity measure 29.0036, possessing 49 nontrivial parameters. It is shown that

$$\begin{aligned}\bar{\gamma}(G_\rho^{\text{opt}}) &= \begin{bmatrix} -1 & -0.75 & -0.6875 & -0.75 & -0.75 & -0.8125 \end{bmatrix} \\ \bar{\gamma}(G_{ss}^{\text{opt}}) &= \begin{bmatrix} -1 & -0.6875 & -0.6875 & -0.75 & -0.75 & -0.8125 \end{bmatrix}\end{aligned}$$

Table 3.2 shows the same kind of statistics.

Through two examples, we observe the following:

- Both examples confirm the argument that the optimal γ -parameters are close to the poles of the filters.
- For Example 3.1, as expected, the δDFII structure yields a much better

Figure 3.3: Magnitude responses (in dB) and phase responses (in degree) for Example 3.1: i – ideal; ii – $\bar{\gamma}_z$ with $f_c(4, 11)$; iii – $\bar{\gamma}_\delta$ with $f_c(1, 6)$; iv – $\bar{\gamma}(G_\rho^{\text{opt}})$ with $f_c(1, 6)$

performance than the traditional z DFIIt. In fact, the former has a performance very close to the optimal ρ DFIIt while this is no longer true for Example 3.2 since the poles of the filter are far away from $z = 1$.

- It is interesting to note that for both examples the ρ DFIIt and ρ SS structures of the $\bar{\gamma}(G_\rho^{\text{opt}})$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$ not only yield almost the same performance as that by the fully parameterized optimal roundoff noise realization R_G but also are much more efficient than the latter. Particularly, the ρ SS structure can outperform the R_G in terms of roundoff noise gain.

Figure 3.4: Magnitude responses (in dB) and phase responses (in degree) for Example 3.2: i – ideal; ii – $\bar{\gamma}_z$ with $f_c(3, 12)$; iii – $\bar{\gamma}_\delta$ with $f_c(4, 11)$; iv – $\bar{\gamma}(G_\rho^{\text{opt}})$ with $f_c(0, 7)$

Figures 3.3 and 3.4 show the magnitude responses and phase responses of the ideal filter and ρ DFIIt structures implemented with $f_c(I_c, B_c)$ for $\bar{\gamma}_z$, $\bar{\gamma}_\delta$, and $\bar{\gamma}(G_\rho^{\text{opt}})$ for Examples 3.1 and 3.2, respectively, which illustrate the effect of the coefficient quantization on frequency response of digital filters. One can observe that in these two figures, either the magnitude response or the phase response of the optimal ρ DFIIt is very close to the ideal ones.

In what follows, we simulate the filtering performance of different structures for Examples 3.1 and 3.2, and compare the output of each FWL structure with that of the ideal filter. The simulations are carried out with the quantization errors of both the signals and parameters taken into consideration. In both simulations, the input signal $u(n)$ is a random sequence which is normally distributed with mean $\bar{u} = 0$ and variance $\sigma_u^2 = 1$, generated with MATLAB command $u(n) = \text{randn}(1, 1200)$.

Simulation for Example 3.1: In this simulation, we compare the outputs of the ρ DFIIt structures with $\bar{\gamma}_z$, $\bar{\gamma}_\delta$, and $\bar{\gamma}(G_\rho^{\text{opt}})$ with that of the ideal filter. For the three FWL structures, the coefficients are implemented with $f_c(I_c, B_c)$ and the signals with $f_s(I_s, B_s)$ (see Notations).

Denote $y(n)$ as the output of the ideal filter, and $y_1(n)$, $y_2(n)$, $y_3(n)$ as that of the ρ DFIIt structures corresponding to $\bar{\gamma}_z$, $\bar{\gamma}_\delta$, and $\bar{\gamma}(G_\rho^{\text{opt}})$, respectively. Defining the error sequence as

$$e_x(n) \triangleq y(n) - y_x(n), \text{ for } x = 1, 2, 3. \quad (3.34)$$

It can be shown that the mean and the variance of each error sequence are

$$\begin{cases} \bar{e}_1 = -0.0472, & \bar{e}_2 = -3.2213 \times 10^{-5}, & \bar{e}_3 = -5.5508 \times 10^{-5} \\ \sigma_{e_1}^2 = 0.3016, & \sigma_{e_2}^2 = 6.0199 \times 10^{-6}, & \sigma_{e_3}^2 = 2.5711 \times 10^{-6} \end{cases}$$

Obviously, the outputs of the ρ DFIIt structures with $\bar{\gamma}_\delta$ and $\bar{\gamma}(G_\rho^{\text{opt}})$ are very close to the ideal one for this low-pass narrow band filter.

In Figure 3.5, (a) shows the output sequence $y(n)$ of the ideal filter, and (b) - (d) present the error sequence $e_x(n)$ with $x = 1, 2, 3$ for each FWL structure. One can observe that the outputs of ρ DFIIt structures with $\bar{\gamma}_\delta$ and $\bar{\gamma}(G_\rho^{\text{opt}})$ are much close to the ideal one, as their output error sequences are around zero, which are shown in (c) and (d), respectively. However, the output of the ρ DFIIt structure with $\bar{\gamma}_z$ is far away from the ideal output according to the error sequence shown in (b). This is consistent with our results for Example 3.1, showing that for low-pass narrow band filters, the ρ DFIIt structure with $\bar{\gamma}_\delta$ or $\bar{\gamma}(G_\rho^{\text{opt}})$ can yield a better FWL performance than the one with $\bar{\gamma}_z$.

Simulation for Example 3.2: The second simulation is used to illustrate the filtering performance of the ρ SS structure implemented with $\bar{\gamma}_z$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$, as well as the R_G for Example 3.2.

Denote $y(n)$ as the ideal output, $y_1(n)$ as the output of the R_G , and $y_2(n)$, $y_3(n)$ as that of the ρ SS with $\bar{\gamma}_z$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$, respectively. If the error sequence is denoted like in (3.34), with $x = 1, 2, 3$, one can compute that

$$\begin{cases} \bar{e}_1 = 6.7596 \times 10^{-5}, & \bar{e}_2 = -9.2572 \times 10^{-6}, & \bar{e}_3 = 8.6619 \times 10^{-5} \\ \sigma_{e_1}^2 = 3.1270 \times 10^{-5}, & \sigma_{e_2}^2 = 0.0143, & \sigma_{e_3}^2 = 1.8499 \times 10^{-5} \end{cases}$$

Figure 3.5: Simulation for Example 3.1: (a) – ideal output; (b) – output error of ρDFIIt with $\bar{\gamma}_z$ for $f_c(4, 11)$, $f_s(0, 15)$; (c) – output error of ρDFIIt with $\bar{\gamma}_\delta$ for $f_c(0, 15)$, $f_s(0, 15)$; (d) – output error of ρDFIIt with $\bar{\gamma}(G_\rho^{\text{opt}})$ for $f_c(0, 15)$, $f_s(0, 15)$

In Figure 3.6, (a) shows the ideal output $y(n)$, and (b) - (d) present the error sequence $e_x(n)$ with $x = 1, 2, 3$ for each FWL structure.

3.6 Conclusion

Our contribution in this chapter is three-fold. Firstly, based on the concept of polynomial operators a generalized DFII structure, called ρ DFII, has been developed. One of the nice properties of this sparse structure is that it provides more degrees of freedom to allow us to minimize the FWL effects for any types of digital filter. Secondly, the performance of this proposed structure and its equivalent state-space realization has been analyzed by deriving the expressions for the roundoff noise gain and the L_2 -sensitivity measure for RAM implementations. It has been shown that the equivalent state-space realization always yields a smaller roundoff noise gain than the ρ DFII. The last one is to investigate the optimal ρ DFII problem. Since the parameter space from which the free parameters take values is discrete and of finite elements, this problem can be solved using exhaustive search. The expressions for the roundoff noise gain G_ρ has been derived in terms of poles of the filter and the free parameters, which implies that the optimal free parameters can be searched around the dominant poles. This conclusion helps search for the optimal $\bar{\gamma}$ efficiently and also gives some insights for exploring more robust structures against FWL effects via polynomial operator concept. Two design examples have been given, with which it is shown that the two proposed structures not only outperform the traditional z DFII and the prevailing δ DFII but also yield almost the same performance as that of the fully parameterized optimal roundoff noise realiza-

Figure 3.6: Simulation for Example 3.2: (a) – ideal output; (b) – output error of R_G for $f_c(0, 15)$, $f_s(0, 15)$; (c) – output error of ρ_{SS} with $\bar{\gamma}_z$ for $f_c(3, 12)$, $f_s(3, 12)$; (d) – output error of ρ_{SS} with $\bar{\gamma}(G_{ss}^{\text{opt}})$ for $f_c(0, 15)$, $f_s(0, 15)$

tions. In fact, the optimized equivalent state-space realizations can beat the latter in both examples.

Chapter 4

Revised ρ DFIIt Structures without/with Error Feedback

4.1 Introduction

It is well known that there are two rounding schemes depending on the hardware: rounding before multiplication and rounding after multiplication. The classical optimal state-space realization R_G has been known as one of the most effective methods to reduce the roundoff noise gain under l_2 -norm dynamic-range scaling constraints on the state-variables for both rounding schemes, which is always fully parameterized with nontrivial parameters. In Chapter 3, based on the concept of polynomial operators [16], a set of special operators was derived, with which a generalized DFIIt structure, denoted as ρ DFIIt, was obtained. This structure, having only $3K + 1$ nontrivial parameters plus K free parameters at choice, can be used for any type of digital filters to minimize the roundoff noise gain. The FWL performance of this structure was analyzed for RAM scheme. As

known, error feedback [36], [38] - [46], [51], [52] is another effective technique for reducing roundoff noise in terms of RBM implementations. This is achieved by extracting the quantization error and feeding it back through simple circuits. It can be applied to digital filters that are described by either external or internal models¹ for roundoff noise minimization without affecting the filter's input-output characteristics. In [46], a number of results in computing the optimal error feedback coefficients were presented for any given state-space realization.

In this chapter, two efficient structures are analyzed for RBM scheme. The first one is the ρ DFIIt structure proposed in the previous chapter, for which the expression of roundoff noise gain is derived in Section 4.2. Noting the roundoff error due to the state variable $x_1(n)$ can make a much significant contribution to the total roundoff noise gain, a *revised* ρ DFIIt structure, denoted as ρ RDFIIt, is obtained in Section 4.3. It is shown that this new structure has the same implementation simplicity as the corresponding ρ DFIIt while yielding a smaller roundoff noise gain. The performance of this structure with error feedback is also analyzed. In Section 4.4, all these structures are optimized with the free parameters in terms of reducing roundoff noise gain. A genetic algorithm (GA) [71] is utilized to efficiently solve these optimization problems. Numerical examples are given in Section 4.5 to compare the performance of the optimized structures with that of the traditional optimal state-space realization without/with error feedback in terms of roundoff noise gain and structure simplicity, and illustrate the efficiency of GA.

¹The external model refers to an input-output relationship, such as equation (2.2); the internal model refers to an inside description of the filter structure, such as the state-space realization (2.5).

4.2 Roundoff noise gain for the ρ DFIIt structure in RBM scheme

Looking at Figures 3.1 and 3.2. The signals to be rounded under RBM scheme are $\{x_k(n)\}$ and $y(n)$. To make the problem easy to understand, assuming there are enough bits to implement the integral parts of the coefficients and signals (actually, from the examples in Chapter 3, we could find that for the optimized ρ DFIIt and ρ SS structures, the number of bits used to implement the integral parts of the coefficients or signals is always much less than that for the fractional parts). Hence, in what follows, we only consider the bits implemented for the fractional parts, i.e., B_c and B_s . For the fixed-point implementation of B_c -bits, the input signal $u(n)$ is a B_s -bit fraction, the FWL signals $\{x_k^*(n)\}$ and $y^*(n)$ then have a B -bit representation, where $B = B_c + B_s$.

First of all, let us consider the roundoff noise effect due to the quantization of the state variable $x_k^*(n)$, which has a multiplication with Δ_k and γ_k , respectively. As assumed, the structure parameters $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\Delta_k\}$ are all in a B_c -bit FWL format, while the free parameters $\{\gamma_k\}$ are in an exact fractional B_γ -bit format, just as shown in (3.14). Therefore, the B -bit state variable $x_k^*(n)$ is rounded into $(B - B_\gamma)$ -bits before multiplying with γ_k and the error variance is hence $\sigma^2 = (1/12)2^{-2(B-B_\gamma)}$. Noting the fact that usually B_γ is chosen to be much less than B_c , i.e., $(1/12)2^{-2(B-B_\gamma)} \ll (1/12)2^{-2(B-B_c)}$, the free parameters $\{\gamma_k\}$ produce either no roundoff noise at all or a much smaller one than that caused by the non-free parameters. As a result of that, the roundoff noise gain due to the multiplication with γ_k can be neglected, see [35],

[46]. Denote

$$\Delta y(n) \triangleq y^*(n) - y(n), \quad e_{x_k}(n) \triangleq x_k^*(n) - Q[x_k^*(n)], \quad \forall k$$

from which, one notes $\Delta_k Q[x_k^*(n)] = \Delta_k x_k^*(n) - \Delta_k e_{x_k}(n)$. It follows from the roundoff noise analysis in Section 3.3.1 that the transfer function between $-\Delta_k e_{x_k}(n)$ and $\Delta y(n)$, can be expressed by $H_{k-1}(z)$, with $H_k(z)$ denoted as

$$H_k(z) = \begin{cases} 1 - C_\rho(zI_K - A_\rho)^{-1}\bar{\alpha}, & k = 0 \\ C_\rho(zI_K - A_\rho)^{-1}e_k, & k = 1, 2, \dots, K \end{cases} \quad (4.1)$$

According to (2.14) and (2.9), the roundoff noise gain due to $e_{x_k}(n)$ is given by

$$G_{x_k} = \Delta_k^2 \|H_{k-1}(z)\|_2^2 = \begin{cases} \Delta_1^2(1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}), & k = 1 \\ \Delta_k^2 e_{k-1}^T W_o^\rho e_{k-1}, & k = 2, 3, \dots, K \end{cases} \quad (4.2)$$

Next, let us consider the roundoff noise effect due to the quantization on the filter output. Similarly, let

$$e_y(n) \triangleq y^*(n) - Q[y^*(n)].$$

Noting $-\alpha_k Q[y^*(n)] = -\alpha_k y^*(n) + \alpha_k e_y(n)$, $\forall k$, and according to Section 3.3.1, one can obtain the transfer function from $e_y(n)$ to the corresponding output error $\Delta y(n)$, denoted as $H_y(z)$,

$$H_y(z) = \sum_{k=1}^K \alpha_k H_k(z). \quad (4.3)$$

Figure 4.1: The ρ RDFIIt structure

Therefore, the roundoff noise gain due to $e_y(n)$ is obtained:

$$G_y = \|H_y(z)\|_2^2 = \bar{\alpha}^T W_o^\rho \bar{\alpha} = G_{x_1} / \Delta_1^2 - 1. \quad (4.4)$$

Hence, the total roundoff noise gain for the ρ DFIIt structure in terms of RBM scheme, defined as $G_\rho \triangleq \sum_{k=1}^K G_{x_k} + G_y$, is

$$G_\rho = \bar{\alpha}^T W_o^\rho \bar{\alpha} + \Delta_1^2 (1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) + \text{tr}(W_o^\rho Q_\Delta) \quad (4.5)$$

where $Q_\Delta = \text{diag}(\Delta_2^2, \Delta_3^2, \dots, \Delta_K^2, 0)$.

4.3 Revised ρ DFIIt structures

4.3.1 A revised ρ DFIIt structure

It has been found that the contribution of G_{x_1} to G_ρ can be very significant. Based on this observation, we propose a new structure depicted in Figure 4.1,

where

$$\beta_{r0} = \Delta_1^{-1} \beta_0, \quad \alpha_{rk} = \Delta_1 \alpha_k, \quad \forall k$$

and $\rho_k^{-1}, k = 2, 3, \dots, K$ is still implemented with Figure 3.2. We observe this structure, denoted as ρ RDFIIt for convenience, can be considered as a cascade implementation of the transfer function $H(z) = \Delta_1 \bar{H}(z)$ with $\bar{H}(z)$ realized by an ρ DFIIt in which the operator ρ_1^{-1} has its Δ -parameter equal to one. The corresponding state-space realization of the transfer function $\bar{H}(z)$ implemented with Figure 4.1, denoted as $(A_r, B_r, C_r, \beta_{r0})$, is clearly given by

$$A_r = A_\rho, \quad B_r = B_\rho, \quad C_r = \Delta_1^{-1} C_\rho. \quad (4.6)$$

Let $y_0(n)$ as the output of $\bar{H}(z)$ excited with the input $u(n)$, hence $y(n) = \Delta_1 y_0(n)$. It can be observed that the signals to be rounded in ρ RDFIIt are $y_0(n)$ and $\{x_k(n), k = 2, 3, \dots, K\}$. Denote

$$\Delta y_0(n) \triangleq y_0^*(n) - y_0(n), \quad e_{y_0}(n) \triangleq y_0^*(n) - Q[y_0^*(n)]$$

one can then observe that the input-output relationship between $e_{y_0}(n)$ and $\Delta y_0(n)$ is equivalent to that between $\Delta_1 e_{x_1}(n)$ and $\Delta y(n)$ in the corresponding ρ DFIIt structure. Following from (4.1) and noting that $y(n) = \Delta_1 y_0(n)$, the roundoff noise gain due to $e_{y_0}(n)$ can be given by

$$G_{y_0} = \Delta_1^2 (1 + \bar{\alpha}_r^T W_o^r \bar{\alpha}_r) = \Delta_1^2 (1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) \quad (4.7)$$

where $\bar{\alpha}_r \triangleq [\alpha_{r1} \ \alpha_{r2} \ \dots \ \alpha_{rK}]^T = \Delta_1 \bar{\alpha}$, $W_o^r \triangleq \Delta_1^{-2} W_o^\rho$ is the observability gramian of the realization $(A_r, B_r, C_r, \beta_{r0})$.

Next, for the roundoff noise gain due to $e_{x_k}(n)$, $k = 2, 3, \dots, K$, with the similar approach above one can show that it has the same value as the one in the corresponding ρ DFIIt, which is given in (4.2).

Therefore, the total roundoff noise gain for the ρ RDFIIt structure, defined as $G_r \triangleq \sum_{k=2}^K G_{x_k} + G_{y_0}$, is

$$G_r = \Delta_1^2(1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) + \text{tr}(W_o^\rho Q_\Delta) \quad (4.8)$$

where Q_Δ is the same as defined below (4.5).

Comparing (4.5) with (4.8), it is clear that $G_\rho = G_r + \bar{\alpha}^T W_o^\rho \bar{\alpha} > G_r$. The difference, $\bar{\alpha}^T W_o^\rho \bar{\alpha}$, depending on the filters, can be very large. This is the motivation to propose the ρ RDFIIt structure indeed.

4.3.2 The ρ RDFIIt structure with error feedback

The performance of the ρ RDFIIt structure can be further improved with error feedback. In what follows, we apply error feedback technique to the ρ RDFIIt structure, and the resultant realization, denoted as ρ REF for convenience, is depicted in Figures 4.1 and 4.2, where the state residue $e_{x_k}(n)$ is fed back through a constant coefficient d_k into the filter. The error feedback coefficients $\{d_k\}$ are chosen such that the roundoff noise gain is minimized. In order to avoid new roundoff noise sources caused by $\{d_k\}$, these coefficients are usually constrained to a B_d -bit format number with B_d much smaller than B_c [46], so that the roundoff noise due to the multiplication with d_k can be neglected.

For the roundoff noise effect of $y_0(n)$, it is the same as the one in the corresponding ρ RDFIIt, which is $\Delta_1^2(1 + \bar{\alpha}^T W_o^\rho \bar{\alpha})$. For $x_k(n)$, $k = 2, 3, \dots, K$, one

Figure 4.2: A realization of operator ρ_k^{-1} with error feedback

notes that the transfer functions from $d_k e_{x_k}(n)$ and $-\Delta_k e_{x_k}(n)$ to the corresponding output error $\Delta y(n)$ are $H_k(z)$ and $H_{k-1}(z)$, respectively, where $H_k(z)$ is defined in (4.1). Hence, the total roundoff noise gain for ρ REF can be expressed as

$$G_e = \Delta_1^2 (1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) + \sum_{k=2}^K V_k^T W_o^\rho V_k \quad (4.9)$$

where $V_k \triangleq d_k e_k - \Delta_k e_{k-1}$.

The problem now is how to choose $\{d_k\}$ to make G_e less than G_r . In [46], the roundoff noise gain was minimized by considering the error feedback coefficients as continuous variables and the optimal error feedback coefficients of B_d -bit format were then obtained by truncating the resultant optimal continuous coefficients into B_d -bit representations. With this approach, the optimal $\{d_k\}$ can be worked out. Taking the first derivative of (4.9) with respect to d_k yields

$$\frac{\partial G_e}{\partial d_k} = e_k^T W_o^\rho V_k + V_k^T W_o^\rho e_k.$$

Let $\partial G_e / \partial d_k$ equal to zero, the optimal continuous value of d_k is obtained by

$$d_k^{\text{opt}} = \Delta_k \frac{e_k^T W_o^\rho e_{k-1}}{e_k^T W_o^\rho e_k}, \quad k = 2, 3, \dots, K. \quad (4.10)$$

With such a choice of $\{d_k\}$, G_e achieves its minimum value, denoted as G_{eo} ,

$$G_{eo} = G_r - \sum_{k=2}^K \Delta_k^2 \frac{(e_k^T W_o^\rho e_{k-1})^2}{e_k^T W_o^\rho e_k} < G_r. \quad (4.11)$$

The price paid is $K - 1$ more nontrivial parameters.

4.4 Structure optimization

In the previous sections, we have derived the expressions of roundoff noise gain for the ρ DFIIt structure and its revised version ρ RDFIIt without/with error feedback in terms of RBM scheme. It is obvious that the roundoff noise performance of the proposed structures are actually depending on the choice of $\bar{\gamma}$, which is defined in (3.29). Consequently, one can form the optimal structure problems:

$$\begin{aligned} \min_{\bar{\gamma} \in \bar{S}_\gamma} G_\rho &\implies \bar{\gamma}(G_\rho^{\text{opt}}) \\ \min_{\bar{\gamma} \in \bar{S}_\gamma} G_r &\implies \bar{\gamma}(G_r^{\text{opt}}) \\ \min_{\bar{\gamma} \in \bar{S}_\gamma} G_{eo} &\implies \bar{\gamma}(G_{eo}^{\text{opt}}) \end{aligned} \quad (4.12)$$

where $\bar{\gamma}(G_\rho^{\text{opt}})$, $\bar{\gamma}(G_r^{\text{opt}})$, and $\bar{\gamma}(G_{eo}^{\text{opt}})$ are the optimal solution that can minimize G_ρ , G_r , and G_{eo} , respectively.

These minimization problems also can be solved using exhaustive search in $\bar{S}_\gamma$ given by (3.30). One can observe that G_ρ , G_r and G_{eo} defined in this chapter can be actually expressed in terms of $\{G_k\}$, which were analyzed in detail by (3.32). Therefore, discussions on how to speed up the exhaustive search can be found in Section 3.4. In this chapter, for the efficiency consideration, a standard

optimization algorithm – GA – is applied to tackle the problems in (4.12).

GAs attempt to mimic (in an over-simplified manner) the natural evolution of species in order to solve difficult optimization problems. Figure 2.3 shows the procedure of the GA used to find the optimal $\bar{\gamma}$ for the problems defined in (4.12). In Step 1, to randomly initialize a population, we need to set values for i) N_b - the number of bits to represent γ_k , ii) N_s - total number of solution candidates in each generation, i.e., $\bar{\gamma}_1, \bar{\gamma}_2, \dots, \bar{\gamma}_{N_s}$. So that the initial population can be represented by a matrix in dimension $\mathcal{R}^{N_s \times (N_b \times K)}$, in which each string represents a solution of $\bar{\gamma}$. In Step 2, each solution in the population is suitably decoded under the constrained conditions (3.14), and then G_ρ (G_r, G_{eo}) is evaluated for every candidate solution. After that, the evolutionary computation operators - reproduction, crossover, and mutation are applied in Step 3 to Step 6 to obtain the next generation of solution strings. These evolutionary operators make the solutions that minimize the roundoff noise gain more and more increase in each generation [71], until the termination conditions in Step 7 are satisfied. For the problem in this chapter, the termination condition is a combination of generation number and fitness convergence - i) generation number - a minimum generation number, *gens*, is set to avoid the pre-convergence of fitness, ii) fitness convergence - if we denote F_c, F_l as the best fitness of current and last generations, respectively, then $Range \triangleq \frac{|F_c - F_l|}{F_l}$ must be within a user-specified range for ten consecutive generations².

How to encode the variables is a key issue for GAs. In our problem, we choose standard binary encoding [71] and hence there are only '1's and '0's in the individual strings, which is just corresponding to the exact fractional rep-

²Here, we must point out *gens* and *range* could be diverse for different problems. For example, with the filter order increasing, the minimum generation number should be increased as well.

resentation for γ_k . This, besides efficiency consideration, is another motivation to use GA solving (4.12).

4.5 Numerical examples

In this section, we examine the performance of the three optimized structures (i.e., ρ DFIIt, ρ RDFIIt, ρ REF) and compare them with some existing structures such as the optimal fully parameterized state-space realization R_G [8], [9], and the structures based on R_G with a general error feedback matrix and a diagonal error feedback matrix [46], denoted as R_G^g and R_G^d , respectively. For all the examples, $B_\gamma = 4$ is assumed and the error feedback coefficients are implemented by 4-bit quantization.

Table 4.1: Comparison Results for Example 4.1

Realization	R_G	ρDFIIt	ρRDFIIt	R_G^g	R_G^d	ρREF
G	0.7121	0.2582	0.0839	0.0428	0.0462	0.0770
N_p	49	25	25	85	55	30

Example 4.1: Firstly, we consider a Butterworth low-pass filter generated with MATLAB command: $[V_b, V_a] = \text{butter}(6, 0.05)$, which yields a sixth order low-pass Butterworth digital filter with a normalized bandwidth 0.025. The corresponding poles of this filter are $\{0.9493 \pm j0.1452, 0.8893 \pm j0.0996, 0.8580 \pm j0.0352\}$.

Since the real part of each pole is positive, the optimal problem can be solved with $\gamma_k \in (0, 1]$. In that case, $N_b = 4$. Set $N_s = 10$ (i.e., there are 10 solution candidates of $\bar{\gamma}$ in each generation) and let the $\text{gens} = 30$, $\text{Range} = 0.1$. With the GA program, the optimal solution for each structure is obtained as the following

$$\begin{aligned}\bar{\gamma}(G_\rho^{\text{opt}}) &= \begin{bmatrix} 0.9375 & 0.8750 & 0.9375 & 0.8750 & 0.8750 & 0.8750 \end{bmatrix}, \\ \bar{\gamma}(G_r^{\text{opt}}) &= \begin{bmatrix} 0.9375 & 0.8750 & 0.9375 & 0.8750 & 0.8750 & 0.8750 \end{bmatrix}, \\ \bar{\gamma}(G_{eo}^{\text{opt}}) &= \begin{bmatrix} 0.9375 & 0.9375 & 0.8750 & 0.8750 & 0.8750 & 0.8750 \end{bmatrix}.\end{aligned}$$

The optimal continuous error feedback coefficients obtained for $\bar{\gamma}(G_{eo}^{\text{opt}})$ are

$$\{d_k^{\text{opt}}\} = \begin{bmatrix} 0.047530 & 0.114623 & 0.117027 & 0.118910 & 0.120435 \end{bmatrix}.$$

Table 4.1 shows the statistics of the roundoff noise gain G and number of non-trivial parameters N_p for each of the six structures.

Table 4.2: Statistical History of the GA for ρ REF in Example 4.1

Generation	Min G_{eo}	Optimal $\bar{\gamma}$					
0	181.4761	0.6250	0.1875	0.1250	0.8125	0.9375	0.6875
1	67.7165	0.6250	0.1875	0.4375	0.7500	0.7500	0.9375
2	125.7176	0.4375	0.5000	0.5000	0.5625	0.9375	0.7500
3	23.7047	0.6250	0.6875	0.3750	0.8125	0.9375	0.6875
4	36.6150	0.6250	0.1250	0.3750	0.9375	0.9375	0.6875
5	4.1782	0.3750	0.5625	0.8750	0.8125	0.9375	0.6875
$\vdots$	$\vdots$	$\vdots$					
84	0.7815	0.4375	0.9375	0.7500	0.8750	0.7500	0.8750
85	0.0833	0.8750	0.9375	0.8750	0.8750	0.9375	0.8750
86	0.0770	0.9375	0.9375	0.8750	0.8750	0.8750	0.8750
87	0.0780	0.8750	0.9375	0.8750	0.8750	0.8750	0.8750
88	0.1289	0.8750	0.6250	0.9375	0.8750	0.8750	0.8750
$\vdots$	$\vdots$	$\vdots$					
122	0.0815	0.9375	0.9375	0.9375	0.8750	0.8750	0.8750
123	0.0844	0.9375	0.8125	0.8750	0.8750	0.8750	0.8750
124	0.0775	0.9375	0.8750	0.8750	0.8750	0.8750	0.8750
125	0.0770	0.9375	0.9375	0.8750	0.8750	0.8750	0.8750
126	0.0844	0.9375	0.8125	0.8750	0.8750	0.8750	0.8750

To confirm the optimal solutions computed with the GA, we used the exhaustive search for computing the optimal ρ REF structure which yields almost the same solution. It, however, needs to compute G_{eo} $(2^4)^6 = 16777216$ times, while for the GA, it stops at the 126th generation and the corresponding number is only $10 \times 127 = 1270$. Running the programs that we have developed using MATLAB on a computer (Pentium 4 CPU, 1.90GHz and 256MB RAM), the runtime is 4.3319×10^3 mins and 0.3175 min for the exhaustive search and the GA, respectively. It should be pointed out that, besides the computer used to do the calculations, the runtime also depends on the efficiency of the program used to evaluate G_{eo} . Table 4.2 shows the statistical history of the GA, where the best solution firstly occurs at the 86th generation.

Now, let us see the filtering performance of the R_G , ρ DFIIt and ρ RDFIIt

structures for the low-pass filter in this example. The output of each of these three structures is compared with that of the ideal filter, where the input signal is generated by MATLAB command $u(n) = \text{randn}(1, 1000)$. For the three FWL structures, the coefficients are implemented with $f_c(I_c, B_c)$ and the input signal is implemented with $f_s(I_s, B_s)$.

Denote $y(n)$ as the output of the ideal filter, and $y_1(n)$, $y_2(n)$, $y_3(n)$ as that of the R_G , ρDFIIt , and ρRDFIIt , respectively. Defining the error sequence as

$$e_x(n) \triangleq y(n) - y_x(n), \text{ for } x = 1, 2, 3.$$

Computations show that the mean and the variance of each error sequence are

$$\begin{cases} \bar{e}_1 = 1.7279 \times 10^{-4}, & \bar{e}_2 = 7.9841 \times 10^{-5}, & \bar{e}_3 = 4.8078 \times 10^{-5} \\ \sigma_{e_1}^2 = 6.9793 \times 10^{-7}, & \sigma_{e_2}^2 = 1.9025 \times 10^{-7}, & \sigma_{e_3}^2 = 8.7461 \times 10^{-8} \end{cases}$$

In Figure 4.3, (a) plots the ideal output $y(n)$, and (b) - (d) present the error sequences $e_x(n)$ with $x = 1, 2, 3$ for each FWL structure.

From Figure 4.3 and the variance of the output error for each structure, we can find there is almost no difference between the output of each of these FWL structures and the ideal output. This further confirms that all the three structures studied in this simulation have a very good FWL performance, while, we should note that the R_G possesses many more nontrivial parameters than the other two structures.

Figure 4.3: Simulation for Example 4.1: (a) – ideal output; (b) – output error of R_G for $f_c(0, 15)$, $f_s(1, 14)$; (c) – output error of ρDFIIt for $f_c(1, 14)$, $f_s(0, 15)$; (d) – output error of ρRDFIIt for $f_c(0, 15)$, $f_s(0, 15)$

Example 4.2: Now, let us see a class of low-pass filters generated with:

$$[V_b, V_a] = \text{butter}(K, 0.05)$$

to consider the effect of filter order K on the performance of the six structures. In Figure 4.4, (a) and (b) show the roundoff noise gain in dB ($10\log_{10}G$) and the number of nontrivial parameters (N_p) in the structure versus filter order (K), respectively.

It is obvious that the ρDFIIt structure excels R_G in both roundoff noise performance and structure simplicity, while ρRDFIIt has even much lower roundoff noise gain than ρDFIIt without increasing structure complexity. Although R_G^g and R_G^d can yield a smaller roundoff noise gain, they have much more complicated structures than our proposed structures.

Example 4.3: In this example, we examine a sixth order Butterworth bandpass filter generated with MATLAB command $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.40])$, which has a normalized bandwidth of 0.075. The corresponding poles are $\{p_{1,2} = 0.3008 \pm j0.8244, p_{3,4} = 0.6267 \pm j0.6522, p_{5,6} = 0.4333 \pm j0.6519\}$.

Table 4.3 shows the statistics of the roundoff noise gain G and the number of nontrivial parameters N_p for each of the six structures, and also gives the runtime (minute) of GA, denoted as T_{GA} for ρDFIIt , ρRDFIIt and ρREF (Pentium 4 CPU, 1.90GHz and 256MB RAM). We can observe that the proposed structures in this section are not effective for the bandpass filters. Hence, it is necessary to explore more robust structures which can be used for a wider range of filters.

Figure 4.4: Structure performance vs filter order for Example 4.2: i — R_G ; ii — ρDFIIt ; iii — ρRDFIIt ; iv — R_G^g ; v — R_G^d ; vi — ρREF

Table 4.3: Comparison Results for Example 4.3

Realization	R_G	ρDFIIt	ρRDFIIt	R_G^g	R_G^d	ρREF
G	0.9410	26.7292	11.2680	0.1183	0.7174	11.1727
N_p	49	25	25	85	55	30
T_{GA}	—	0.7701	0.3331	—	—	0.1201

4.6 Conclusion

This chapter has investigated two efficient structures with consideration of the roundoff noise in terms of RBM implementations. The performance has been analyzed for each of these structures by deriving the expressions for the roundoff noise gain. These structures have been optimized in terms of minimizing round-off noise gain. Design examples have been given, which show that the proposed structures can outperform the optimal state-space realization in terms of round-off noise performance as well as the structure simplicity for the type of low-pass narrow band filters, but do not have the same good performance for the other types. Therefore, more robust structures that can be effective for a wider range of filters are demanded, which will be introduced in the next chapter.

Chapter 5

Improved ρ DFIIt Structure and its State-Space Realization

5.1 Introduction

In Chapter 3, based on the concept of polynomial operators [16], the ρ DFIIt structure was proposed and its FWL performance was analyzed in terms of RAM scheme. This structure, compared with the prevailing δ DFIIt which only has an excellent performance for the low-pass narrow band filters, is more robust in the sense that it can be used for a wider range of filters. For a K th order filter, this structure contains $3K + 1$ nontrivial parameters plus K free parameters. In Chapter 4, the roundoff noise performance of the ρ DFIIt structure was investigated for RBM implementations. Moreover, a revised ρ DFIIt structure, denoted as ρ RDFIIt, was obtained, which is in the same structure simplicity to the corresponding ρ DFIIt. It was shown that the ρ RDFIIt structure without/with error feedback is very competitive to the classical optimal state-space

realization without/with error feedback in terms of minimizing roundoff noise gain as well as possessing an efficient structure. However, it has been observed that all the structures proposed in Chapters 3 and 4 cannot effectively reduce the roundoff noise gain for those filters whose poles are far away from the real axis of the z -plane. Based on this observation, an improved ρ DFIIt structure, denoted as ρ IDFIIt, was developed in [1]. Compared with the corresponding ρ DFIIt, this proposed structure has $K/2$ or $(K - 1)/2$ more free parameters, depending on K even or odd. However, in [1], the ρ IDFIIt structure was only discussed on the situation that the K original free parameters take values from three trivial parameters $\{-1, 0, 1\}$. Meanwhile, it is an open problem in [1] that how to efficiently determine the newly introduced free parameters for ρ IDFIIt against FWL effects.

In this chapter, the ρ IDFIIt structure is deeply discussed and the optimization problem for the newly introduced parameters is solved. The equivalent state-space realization of the ρ IDFIIt structure, denoted as ρ ISS, is obtained as well, which has about the same structure simplicity. Assuming the RAM implementations, the expressions of the roundoff noise gain and L_2 -based sensitivity measure for both structures are derived, from which, one finds that the ρ ISS always yields a smaller roundoff noise gain than its counterpart ρ IDFIIt. It is shown that the roundoff noise gain for the proposed structures can be much reduced when the roots of the corresponding polynomials are close to the dominant poles of the filter. We note that the polynomials used in the ρ DFIIt structure have real roots only while the filters generally have complex poles. With the extra degrees of freedom introduced, the corresponding polynomials to the ρ IDFIIt structure can have complex roots and consequently, the roundoff noise gain can be reduced much more effectively. Performance analysis is also

carried out for the ρ ISS structure without/with error feedback in terms of RBM scheme.

An outline of this chapter is given as follows. Based on the polynomial operator concept, two novel efficient structures, the ρ IDFIIt and ρ ISS, are derived in Section 5.2. Section 5.3 is devoted to analyzing the performance of the two structures in terms of RAM implementations. Expressions of roundoff noise gain and L_2 -sensitivity are derived for both structures. It is interesting to note that these two structures contain about $3K/2$ free parameters and for a given set of the free parameters, the ρ ISS structure always yields a smaller roundoff noise gain than that of the corresponding ρ IDFIIt. Noting the roundoff noise gains depend on the choice for the free parameters, the problems of optimal structures are formed in this section in terms of minimizing roundoff noise gain. The relationship between the free parameters and the poles of the filter is also established in terms of roundoff noise gain, with which the GA can effectively find the solutions to the problems. Extensive numerical examples are given to illustrate the design procedure. Simulations show that the proposed structures can outperform the z DFIIt, δ DFIIt, and ρ DFIIt structures as well as the classical fully parameterized optimal realizations in terms of FWL performance and structure efficiency. In Section 5.4, by taking RBM scheme into account, the roundoff noise expression for the ρ ISS structure is also derived. After that, with the approach of evaluating the optimal error feedback coefficients for the state-space realization [46], the roundoff noise gain of the ρ ISS can be further reduced. Design examples for the structures in RBM scheme are also presented in this section, which show the advantage of the optimized ρ ISS structure without/with error feedback and support the theoretical analysis.

5.2 Two novel efficient structures

5.2.1 An improved ρ DFIIt structure

Defining $p \triangleq K/2$ if K is even, $p \triangleq (K+1)/2$ if K is odd and

$$\rho_k(z) \triangleq \frac{z - \gamma_k}{\Delta_k}, \quad k = 1, 2, \dots, K-1, K \quad (5.1)$$

where $\{\gamma_k\}$ and $\{\Delta_k > 0\}$ are two sets of parameters to be discussed later, one then has a set of polynomials $\{q_k(z), k = 0, 1, \dots, K\}$ obtained with the following recursive equations

$$\begin{cases} q_K(z) = 1, & q_{K-1}(z) = \rho_K(z) \\ q_{K-2m}(z) = \bar{\rho}_{p-m+1}(z)q_{K-2m+2}(z) \\ q_{K-2m-1}(z) = \rho_{K-2m}(z)q_{K-2m}(z) \end{cases} \quad (5.2)$$

for all $m = 1, 2, \dots, p$, where

$$\begin{cases} \bar{\rho}_{p-m+1}(z) \triangleq \rho_{K-2m+1}(z)\rho_{K-2m+2}(z) + \eta_{p-m+1} \\ \eta_{p-m+1} \triangleq \frac{\bar{\eta}_{p-m+1}}{\Delta_{K-2m+1}\Delta_{K-2m+2}} \end{cases} \quad (5.3)$$

with $\{\bar{\eta}_m\}$ a set of free parameters. It should be pointed out that when K is odd, $\rho_0(z) \triangleq 1$, $\bar{\eta}_1 = 0$ are assumed, which leads to $\bar{\rho}_1(z) = \rho_1(z)$ and hence $q_0(z) = \bar{\rho}_1(z)q_1(z)$.

It is interesting to note that

$$q_k(z) = \left[\prod_{l=k+1}^K \Delta_l^{-1} \right] \bar{q}_k(z), \quad \forall k \quad (5.4)$$

where all $\bar{q}_k(z)$ are obtained using (5.2) with $\Delta_k = 1, \forall k$. Clearly, these polynomials are uniquely determined by the two sets of free parameters $\{\gamma_k\}$ and $\{\bar{\eta}_m\}$. In the next section, we will discuss how to choose these parameters to minimize the roundoff noise.

Denoting

$$V_a \triangleq \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_k \\ \vdots \\ a_K \end{bmatrix}, \quad V_b \triangleq \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \\ \vdots \\ b_K \end{bmatrix}, \quad \bar{z} \triangleq \begin{bmatrix} z^K \\ z^{K-1} \\ \vdots \\ z^{K-k} \\ \vdots \\ 1 \end{bmatrix} \quad (5.5)$$

one then has

$$\bar{q}(z) \triangleq \begin{bmatrix} q_0(z) & q_1(z) & \cdots & q_{K-k}(z) & \cdots & q_K(z) \end{bmatrix}^T = T_q \bar{z}$$

where $T_q \in \mathcal{R}^{(K+1) \times (K+1)}$ an upper triangular matrix with its k th row determined by the coefficients of $q_k(z)$ for $k = 1, 2, \dots, K$ and $T_q(K+1, K+1) = 1$. It follows from (2.1) and (5.2) that

$$H(z) = \frac{V_b^T \bar{z}}{V_a^T \bar{z}} = \frac{V_\beta^T \bar{q}(z)}{V_\alpha^T \bar{q}(z)} = \frac{\sum_{k=0}^K \beta_k q_k(z)}{q_0(z) + \sum_{k=1}^K \alpha_k q_k(z)}$$

where

$$\begin{cases} V_\alpha \triangleq \begin{bmatrix} 1 & \alpha_1 & \cdots & \alpha_k & \cdots & \alpha_K \end{bmatrix}^T = \kappa^{-1} T_q^{-T} V_a \\ V_\beta \triangleq \begin{bmatrix} \beta_0 & \beta_1 & \cdots & \beta_k & \cdots & \beta_K \end{bmatrix}^T = \kappa^{-1} T_q^{-T} V_b \end{cases} \quad (5.6)$$

with $\kappa \triangleq \prod_{k=1}^K \Delta_k$ leading to $V_\alpha(1) = 1$.

Define

$$\begin{cases} B_{p-m}(z) \triangleq \beta_{K-2m-1}\rho_{K-2m}(z) + \beta_{K-2m} \\ A_{p-m}(z) \triangleq \alpha_{K-2m-1}\rho_{K-2m}(z) + \alpha_{K-2m} \end{cases}$$

for all possible m except $m = p - 1$ when K is odd, for which, $A_1(z) = \alpha_1$, $B_1(z) = \beta_1$. The transfer function can finally be rewritten as

$$H(z) = \frac{\beta_0 q_0(z) + \sum_{m=0}^{p-1} B_{p-m}(z) q_{K-2m}(z)}{q_0(z) + \sum_{m=0}^{p-1} A_{p-m}(z) q_{K-2m}(z)}. \quad (5.7)$$

Now, let us consider how to implement the filter using the polynomial-parameterized transfer function (5.7). First of all, it follows from (5.7) that

$$\begin{aligned} y(n) &= \beta_0 u(n) + \sum_{m=0}^{p-1} [B_{p-m}(z) \frac{q_{K-2m}(z)}{q_0(z)} u(n) - A_{p-m}(z) \frac{q_{K-2m}(z)}{q_0(z)} y(n)] \\ &\triangleq \beta_0 u(n) + w_1(n). \end{aligned} \quad (5.8)$$

Noting $q_{K-2m}(z) = \bar{\rho}_{p-m+1}(z) q_{K-2m+2}$ and $q_0(z) = \prod_{m=1}^p \bar{\rho}_m(z)$ (for any K), one can show that

$$\begin{aligned} w_1(z) &= \bar{\rho}_1^{-1}(z) [B_1(z) u(n) - A_1(z) y(n) + \sum_{m=0}^{p-2} [B_{p-m}(z) \frac{q_{K-2m}(z)}{\prod_{m=2}^p \bar{\rho}_m(z)} u(n) \\ &\quad - A_{p-m}(z) \frac{q_{K-2m}(z)}{\prod_{m=2}^p \bar{\rho}_m(z)} y(n)] \\ &\triangleq \bar{\rho}_1^{-1}(z) [B_1(z) u(n) - A_1(z) y(n) + w_2(n)]. \end{aligned}$$

Using the same procedure, one can show that with $w_{p+1}(n) = 0$,

$$w_m(z) = \bar{\rho}_m^{-1}(z) [B_m(z) u(n) - A_m(z) y(n) + w_{m+1}(n)], \quad \forall m. \quad (5.9)$$

Figure 5.1 shows the corresponding block diagram to (5.8) - (5.9), where

Figure 5.1: Block diagram of the proposed ρ IDFIIt structure

Figure 5.2: Implementation of the m th block in the ρ IDFIIt structure, where $\bar{m} \triangleq K - 2p + 2m$ and $w_{p+1}(n) = 0$

the output of the m th block, i.e., $w_m(n)$, can be computed with the structure depicted in Figure 5.2. The structure shown in Figures 5.1 and 5.2 is called an improved ρ DFIIt structure, denoted as ρ IDFIIt, and possesses $\{\alpha_k, \beta_k, \Delta_k\}$ plus two sets of free parameters $\{\gamma_k\}, \{\eta_m\}$. It is interesting to note that when $\bar{\eta}_m = 0, \forall m$, the ρ IDFIIt structure returns to the ρ DFIIt structure proposed in Chapter 3.

5.2.2 Equivalent state-space realization

Choose $\{x_k(n)\}$ indicated in Figure 5.2 as the state variables and denote $x(n)$ as the state vector. It can be shown that a ρ IDFIIt structure is equivalent to the following state-space model, denoted as ρ ISS:

$$\begin{cases} x(n+1) = A_\rho x(n) + B_\rho u(n) \\ y(n) = C_\rho x(n) + \beta_0 u(n) \end{cases} \quad (5.10)$$

where $(A_\rho, B_\rho, C_\rho, \beta_0)$ is given by

$$A_\rho = D_\gamma + (A_o - A_\eta)D_\Delta, \quad B_\rho = \bar{\beta} - \beta_0\bar{\alpha}, \quad C_\rho = e_1^T D_\Delta \quad (5.11)$$

with $D_\gamma \triangleq \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_K)$ and $D_\Delta \triangleq \text{diag}(\Delta_1, \Delta_2, \dots, \Delta_K)$ two diagonal matrices, and¹

$$\left\{ \begin{array}{l} A_o \triangleq \begin{bmatrix} -\alpha_1 & 1 & 0 & \cdots & 0 & 0 \\ -\alpha_2 & 0 & 1 & \cdots & 0 & 0 \\ -\alpha_3 & 0 & 0 & \cdots & 0 & 0 \\ -\alpha_4 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & \\ -\alpha_{K-1} & 0 & 0 & \cdots & 0 & 1 \\ -\alpha_K & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad A_\eta \triangleq \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \eta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \eta_2 & \cdots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & \eta_p & 0 \end{bmatrix} \\ \bar{\alpha} \triangleq \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_K \end{bmatrix}^T, \quad \bar{\beta} \triangleq \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \end{bmatrix}^T \end{array} \right. \quad (5.12)$$

¹The A_η presented here is for K being even, that is $K = 2p$. When K is odd, $A_\eta(2m - 1, 2m - 2) = \eta_m$ for $m = 2, 3, \dots, p$ with $\eta_1 = 0$.

The transfer function from $u(n)$ to $y(n)$ is then

$$H(z) = \beta_0 + C_\rho(zI_K - A_\rho)^{-1}B_\rho. \quad (5.13)$$

One notes that $(A_\rho, B_\rho, C_\rho, \beta_0)$ contains only $4K + p - 1$ nontrivial parameters, which is more efficient than the classical optimal state-space realization.

With (5.4), the state-space realization can be obtained with

$$A_\rho = T_{sc}\bar{A}_\rho T_{sc}^{-1}, \quad B_\rho = T_{sc}\bar{B}_\rho, \quad C_\rho = \bar{C}_\rho T_{sc}^{-1} \quad (5.14)$$

where $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$ is the equivalent realization corresponding to $\Delta_k = 1, \forall k$, and

$$T_{sc} = \text{diag}(t_1, t_2, \dots, t_K), \quad t_k = \prod_{l=1}^k \Delta_l^{-1}, \quad \forall k. \quad (5.15)$$

Such T_{sc}^{-1} is actually a diagonal scaling transformation.

It has been noted in Chapter 2 that any structure has to be l_2 -scaled. This can be achieved if $W_c^\rho(k, k) = 1, \forall k$, where W_c^ρ is the controllability gramian of $(A_\rho, B_\rho, C_\rho, \beta_0)$. Denote $\bar{W}_c^\rho$ as the controllability gramian corresponding to $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$. It follows from (5.14) - (5.15) that the l_2 -scaling can be achieved if $t_k^2 \bar{W}_c^\rho(k, k) = 1, \forall k$, which leads to

$$\Delta_1 = \sqrt{\bar{W}_c^\rho(1, 1)}, \quad \Delta_k = \sqrt{\frac{\bar{W}_c^\rho(k, k)}{\bar{W}_c^\rho(k-1, k-1)}}, \quad k = 2, 3, \dots, K.$$

In the sequel, the structures under discussion are assumed to have been

l_2 -scaled. Therefore, for the ρ IDFIIIt,

$$\begin{cases} \kappa &= \sqrt{W_c^\rho(K, K)} \\ q_0(z) &= \sqrt{\frac{1}{W_c^\rho(K, K)}} \bar{q}_0(z), \quad q_K(z) = 1 \\ q_k(z) &= \sqrt{\frac{W_c^\rho(k, k)}{W_c^\rho(K, K)}} \bar{q}_k(z), \quad k = 1, \dots, K-1 \end{cases}$$

Clearly, these polynomials are uniquely determined by the two sets of free parameters $\{\gamma_k\}$ and $\{\bar{\eta}_m\}$. In Section 5.3, we will discuss how to choose these parameters to minimize roundoff noise gain.

5.3 Performance analysis and structure optimization for RAM scheme

5.3.1 Roundoff noise analysis

In what follows, we will analyze the roundoff noise performance of the proposed structures in terms of RAM scheme.

5.3.1.1 The ρ IDFIIIt structures

We note that the parameters in a ρ IDFIIIt structure are $\{\alpha_k\}$, $\{\beta_k\}$, $\{\Delta_k\}$, $\{\gamma_k\}$ and $\{\eta_k\}$. First of all, let us consider the roundoff noise by $\{\beta_k\}$. It follows from (5.7) that

$$y(n) = \beta_0 u(n) + \sum_{l=1}^K \left[\frac{q_l(z)}{q_0(z)} \beta_l u(n) - \frac{q_l(z)}{q_0(z)} \alpha_l y(n) \right]. \quad (5.16)$$

When $\beta_0 u(n)$ is replaced with $\beta_0 u(n) + \psi(\beta_0) e_{\beta_0}(n)$ in a ρ IDFIIIt structure, where $\psi(x)$ was defined in (3.16), the corresponding output is $y(n) + \Delta y(n)$.

With (5.16), one can show that $\Delta y(n) = H_0(z)\psi(\beta_0)e_{\beta_0}(n)$, where

$$H_0(z) = \frac{q_0(z)}{q_0(z) + \sum_{l=1}^K \alpha_l q_l(z)}.$$

Comparing $H_0(z)$ with (5.7), it then follows from (5.11) and (5.13) that

$$H_0(z) = [\beta_0 + C_\rho(zI_K - A_\rho)^{-1}B_\rho]_{|\beta_0=1, \bar{\beta}=0} = 1 - C_\rho(zI_K - A_\rho)^{-1}\bar{\alpha}.$$

According to (2.14) and (2.9), we have

$$G_{\beta_0} = \psi(\beta_0)\|H_0(z)\|_2^2 = \psi(\beta_0)(1 + \bar{\alpha}^T W_o^\rho \bar{\alpha}) \triangleq \psi(\beta_0)G_0$$

where W_o^ρ is the observability gramian of the realization $(A_\rho, B_\rho, C_\rho, \beta_0)$.

Using the same procedure, one can show that the transfer function from $\psi(\beta_k)e_{\beta_k}(n)$ to $\Delta y(n)$, denoted as $H_k(z)$, is

$$H_k(z) = \frac{q_k(z)}{q_0(z) + \sum_{l=1}^K \alpha_l q_l(z)} = C_\rho(zI - A_\rho)^{-1}e_k$$

for $k = 1, 2, \dots, K$. Therefore,

$$G_{\beta_k} = \psi(\beta_k)e_k^T W_o^\rho e_k \triangleq \psi(\beta_k)G_k, \quad \forall k.$$

Comparing the positions of α_k, γ_k and Δ_{k+1} with that of β_k in Figure 5.2, one can see easily that

$$G_{\alpha_k} = \psi(\alpha_k)G_k, \quad G_{\gamma_k} = \psi(\gamma_k)G_k, \quad G_{\Delta_k} = \psi(\Delta_k)G_{k-1}$$

for $k = 1, 2, \dots, K$. Similarly, the positions of $\beta_{K-2p+2m}$ and η_m in Figure 5.2 suggest that for $m = 1, 2, \dots, p$

$$G_{\eta_m} = \psi(\eta_m)G_{K-2p+2m}.$$

Denote $G_x \triangleq \sum_{\forall k} G_{x_k}$ for $x = \alpha, \beta, \gamma, \Delta, \eta$. The total roundoff noise gain of the ρ IDFIIt structure is then defined as $G_\rho \triangleq \sum_{\forall x} G_x$, which is equal to

$$G_\rho = \sum_{k=0}^K \xi_k G_k, \quad (5.17)$$

where the coefficients ξ_k can be specified easily with the expressions obtained above of roundoff noise for all the parameters:

$$\left\{ \begin{array}{l} \xi_0 = \psi(\Delta_1) + \psi(\beta_0) \\ \xi_K = \psi(\alpha_K) + \psi(\beta_K) + \psi(\gamma_K) + \psi(\eta_p) \\ \xi_{K-2m} = \psi(\alpha_{K-2m}) + \psi(\beta_{K-2m}) + \psi(\gamma_{K-2m}) + \psi(\Delta_{K-2m+1}) + \psi(\eta_{p-m}) \\ \xi_{K-2m+1} = \psi(\alpha_{K-2m+1}) + \psi(\beta_{K-2m+1}) + \psi(\gamma_{K-2m+1}) + \psi(\Delta_{K-2m+2}) \\ \forall m \in \{1, 2, \dots, p-1\} \end{array} \right. \quad (5.18)$$

when K is odd. For K even, (5.18) still holds but with $\xi_1 = \psi(\alpha_1) + \psi(\beta_1) + \psi(\gamma_1) + \psi(\Delta_2)$.

5.3.1.2 The ρ ISS structures

It follows from (5.11) and (5.12) that the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ has the following sparse structure (for K even):

$$A_\rho = \begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & \cdots & 0 & 0 & 0 \\ a_{31} & 0 & a_{33} & \cdots & 0 & 0 & 0 \\ a_{41} & 0 & a_{43} & \cdots & 0 & 0 & 0 \\ & & & \vdots & & & \\ a_{(K-1)1} & 0 & 0 & \cdots & 0 & a_{(K-1)(K-1)} & a_{(K-1)K} \\ a_{K1} & 0 & 0 & \cdots & 0 & a_{K(K-1)} & a_{KK} \end{bmatrix} \quad (5.19)$$

$$B_\rho = \begin{bmatrix} b_1 & b_2 & \cdots & b_{K-1} & b_K \end{bmatrix}^T, \quad C_\rho = \begin{bmatrix} c_1 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

which generally has $\frac{9K}{2} - 1$ or $\frac{9K+1}{2} - 1$ parameters, depending on K even or odd, compared with $\frac{9K}{2} + 1$ or $\frac{9K+1}{2}$ for the ρ IDFIIt structure.

The relationship between these structure parameters and $\{\alpha_k, \beta_k, \Delta_k, \gamma_k, \eta_k\}$ can be specified with (5.11) and (5.12). With the same approach as the one used for ρ IDFIIt, one can analyze the roundoff noise performance of the filter when implemented with this state-space realization and obtain the total roundoff noise gain (see (2.19)), denoted as G_{ss} :

$$G_{ss} = \sum_{k=1}^K \zeta_k G_k + \zeta_0, \quad (5.20)$$

where G_k is the same as defined before and

$$\left\{ \begin{array}{l} \zeta_0 = \psi(c_1) + \psi(\beta_0), \quad \zeta_1 = \psi(a_{11}) + \psi(a_{12}) + \psi(b_1) \\ \zeta_2 = \psi(a_{21}) + \psi(a_{22}) + \psi(a_{23}) + \psi(b_2) \\ \zeta_K = \psi(a_{K1}) + \psi(a_{K(K-1)}) + \psi(a_{KK}) + \psi(b_K) \\ \zeta_l = \psi(a_{l1}) + \psi(a_{l(l-1)}) + \psi(a_{ll}) + \psi(a_{l(l+1)}) + \psi(b_l) \\ \zeta_{l+1} = \psi(a_{(l+1)1}) + \psi(a_{(l+1)(l+1)}) + \psi(a_{(l+1)(l+2)}) + \psi(b_{l+1}) \\ l = K - 2m, \quad \forall m \in \{1, 2, \dots, p-2\} \end{array} \right. \quad (5.21)$$

when K is odd. For K even, (5.21) still holds but with $\zeta_3 = \psi(a_{31}) + \psi(a_{33}) + \psi(a_{34}) + \psi(b_3)$.

Generally, the parameters in both structures are nontrivial. In that case, one has

$$G_\rho - G_{ss} = 2\bar{\alpha}^T W_o^\rho \bar{\alpha} + \begin{cases} G_1 + G_2, & K \text{ is even} \\ G_1, & K \text{ is odd} \end{cases}$$

which means that the ρ ISS structure always yields a smaller roundoff noise gain than the corresponding ρ IDFIIt. As to be discussed in Section 5.3.3, one can set the free parameters $\{\gamma_k\}$ to be trivial and $\{\eta_m\}$ to be zero. It is easy to verify that the superiority of the equivalent state-space realization still holds in such a situation.

5.3.2 L_2 -sensitivity measure

For the ρ DFIIt structure, there are two sets of free parameters $\{\gamma_k\}$ and $\{\eta_m\}$. Noting that $\{\gamma_k\}$ are assumed of a FWL format, while $\{\eta_m\}$ not (this will be explained in detail in Section 5.3.3), it follows from (2.6) that the L_2 -norm based

sensitivity measure is

$$M_\rho = \sum_{\forall \tau_l, \tau_l \in \{\alpha_k, \beta_k, \Delta_k, \eta_m\} \text{ and } \tau_l \in S_{FWL}} \left\| \frac{\partial H(z)}{\partial \tau_l} \right\|_2^2. \quad (5.22)$$

The expressions of the derivatives $\partial H(z)/\partial \beta_0$, $\partial H(z)/\partial \beta_k$, $\partial H(z)/\partial \alpha_k$ and $\partial H(z)/\partial \Delta_k$ are indicated in (3.27), and

$$\frac{\partial H}{\partial \eta_m} = \begin{cases} C_\rho(zI_K - A_\rho)^{-1} e_{2m} e_{2m-1}^\tau D_\Delta(zI_K - A_\rho)^{-1} B_\rho, & K \text{ is even} \\ C_\rho(zI_K - A_\rho)^{-1} e_{2m-1} e_{2m-2}^\tau D_\Delta(zI_K - A_\rho)^{-1} B_\rho, & K \text{ is odd} \end{cases} \quad (5.23)$$

with noting that when K is odd, $m \neq 1$, and D_Δ defined above (5.12).

For the state-space realization $(A_\rho, B_\rho, C_\rho, \beta_0)$, noting (5.19), one can evaluate its L_2 -sensitivity measure, denoted as M_{ss} , directly by (2.11) - (2.13).

5.3.3 Structure optimization

We have derived the expressions of the roundoff noise gain and L_2 -sensitivity measure for the ρ IDFII structure and its equivalent state-space realization with any given $\{\gamma_k\}$, $\{\bar{\eta}_m\}$. As mentioned before, for a given digital filter $H(z)$, there exists a class of ρ IDFII structures, which is determined by two spaces, denoted as S_γ and $S_{\bar{\eta}}$, from which the free parameters $\{\gamma_k\}$ and $\{\bar{\eta}_m\}$ take value, respectively. An interesting problem is to identify those ρ IDFII structures which minimize G_ρ . This leads to the following optimized ρ IDFII structure problem:

$$\min_{\gamma_k \in S_\gamma, \bar{\eta}_m \in S_{\bar{\eta}}, \forall k, m} G_\rho. \quad (5.24)$$

Similarly, one can formulate the following optimal equivalent state-space realization problem:

$$\min_{\gamma_k \in S_\gamma, \bar{\eta}_m \in S_{\bar{\eta}}, \forall k, m} G_{ss}. \quad (5.25)$$

One notes that both G_ρ and G_{ss} can be expressed in a linear combination of all $G_k = \|H_k(z)\|_2^2$ with

$$H_k(z) = \frac{q_k(z)}{q_0(z) + \sum_{l=1}^K \alpha_l q_l(z)} = \frac{\kappa q_k(z)}{D(z)}, \quad \forall k$$

where $D(z)$ is the denominator of $H(z)$, see (2.1). It follows from (5.16) that

$$H_k(z) = \begin{cases} \frac{\bar{q}_0(z)}{D(z)}, & k = 0 \\ \sqrt{\bar{W}_c^\rho(k, k)} \frac{\bar{q}_k(z)}{D(z)}, & k = 1, 2, \dots, K-1 \\ \frac{\sqrt{\bar{W}_c^\rho(K, K)}}{D(z)}, & k = K \end{cases}$$

Let $D(z) = \prod_{l=1}^K (z - \mu_l)$. To simplify the discussion, assume that the filter has no repeated poles, that is $\mu_l \neq \mu_i, \forall l \neq i$, and no poles at $z = 0$. According to the Residue Theorem, it follows from (2.7) that

$$G_k = \begin{cases} \frac{\bar{q}_0(0)}{D(0)} + \sum_{i=1}^K \frac{\bar{q}_0(\mu_i) \bar{q}_0(\mu_i^{-1})}{D(\mu_i^{-1}) \prod_{l \neq i}^K (\mu_i - \mu_l)} \mu_i^{-1}, & k = 0 \\ \bar{W}_c^\rho(k, k) \sum_{i=1}^K \frac{1}{D(\mu_i^{-1}) \prod_{l \neq i}^K (\mu_i - \mu_l)} \mu_i^{-1}, & k = K \\ \bar{W}_c^\rho(k, k) \sum_{i=1}^K \frac{\bar{q}_k(\mu_i) \bar{q}_k(\mu_i^{-1})}{D(\mu_i^{-1}) \prod_{l \neq i}^K (\mu_i - \mu_l)} \mu_i^{-1}, & k \neq 0, K \end{cases} \quad (5.26)$$

It is easy to see from (5.26) that G_k could be very large if $H(z)$ has some poles clustered together and close to the unit circle. More importantly, one can observe that G_k can be much reduced if $\bar{q}_k(\mu_i) \approx 0$. This explains the fact that

the FWL effects are very serious on the class of low-pass narrow-band filters for which the poles are all clustered around $z = +1$ and that the δ DFIIt structure [28] (a special ρ IDFIIt structure, corresponding to $\gamma_k = 1, \forall k$, $\bar{\eta}_m = 0, \forall m$) is very efficient in reducing the roundoff noise gain in such a situation. Though the proposed ρ DFIIt structure can be used for a wider range of digital filters than the δ DFIIt, it may not be able to reduce the FWL effects effectively for the filters whose dominant poles are far away from the real axis of the z -plane. This is the main motivation to propose the ρ IDFIIt structure. In fact, for the ρ IDFIIt structure

$$\begin{cases} \bar{q}_{K-2m}(z) = [(z - \gamma_{K-2m+1})(z - \gamma_{K-2m+2}) + \bar{\eta}_{p-m+1}] \bar{q}_{K-2m+2}(z) \\ \bar{q}_{K-2m-1}(z) = [z - \gamma_{K-2m}] \bar{q}_{K-2m}(z) \end{cases}$$

Clearly, with the three free parameters γ_{K-2m+1} , γ_{K-2m+2} and $\bar{\eta}_{p-m+1}$, $\bar{q}_{K-2m}(z)$ and $\bar{q}_{K-2m-1}(z)$ can have any assigned roots. Let $z = re^{\pm j\theta}$ be a pair of poles to be assigned to $\bar{q}_{K-2m}(z)$. Clearly, this can be achieved if

$$\begin{cases} \gamma_{K-2m+1} + \gamma_{K-2m+2} = 2r\cos\theta \\ \bar{\eta}_{p-m+1} + \gamma_{K-2m+1}\gamma_{K-2m+2} = r^2 \end{cases} \quad (5.27)$$

One of the solutions to above equations is

$$\begin{cases} \gamma_{K-2m+1} = \gamma_{K-2m+2} = r\cos\theta \\ \bar{\eta}_{p-m+1} = r^2\sin^2\theta \end{cases}$$

It is our conjecture that G_k and hence G_ρ, G_{ss} can be much reduced if $\{\gamma_k\}, \{\bar{\eta}_m\}$ are chosen such that the roots of the polynomials $\bar{q}_k(z)$ are close to those poles that dominate the roundoff noise gain. This suggests that the

optimal $\{\gamma_k\}$, $\{\bar{\eta}_m\}$ can be searched within the space for which the roots of the corresponding $\bar{q}_k(z)$ are around the neighborhood of those dominant filter poles. This has been supported by many numerical examples.

It is known that γ_k takes value from a finite space, say S_γ , which is rewritten in the following

$$S_\gamma \triangleq \{-1, 1\} \cup \left\{ \pm \sum_{l=1}^{B_\gamma} b_l 2^{-l}, b_l = 0, 1, \forall l \right\}. \quad (5.28)$$

As to $\bar{\eta}_m$, they are not the structure parameters and hence do not have to be in an FWL format. Since it is found that γ_{K-2m+1} and γ_{K-2m+2} are better in the same sign² and for a stable digital filter, $0 \leq r < 1$, it follows from (5.27) that

$$S_{\bar{\eta}} = \{\bar{\eta}_m : -1 \leq \bar{\eta}_m < 1\}. \quad (5.29)$$

One notes that $S_{\bar{\eta}}$ is not a *discrete* space, which means that the exhaustive search mentioned in the previous chapters can not solve the optimal structure problems any more. Therefore, we should try to find an appropriate optimization algorithm to attack our problems. This, besides the efficiency consideration, is the more important motivation to utilize GA in our research work.

Denote ³

$$\theta \triangleq \begin{bmatrix} \gamma_1 & \cdots & \gamma_k & \cdots & \gamma_K & \bar{\eta}_1 & \cdots & \bar{\eta}_m & \cdots & \bar{\eta}_p \end{bmatrix}. \quad (5.30)$$

²One can see from (5.27), if γ_{K-2m+1} and γ_{K-2m+2} are not in the same sign, one of them should be bigger or smaller than $2r \cos \theta$, depending on $\theta \in [0, \pi]$ or $\theta \in [-\pi, 0]$, respectively, which may violate the assumption that $|\gamma_k| \leq 1$.

³Here, K is assumed to be even. When K is odd, $\bar{\eta}_1 = 0$ and hence the parameter vector θ contains $K + p - 1$ elements.

To randomly initialize a population, we need to set values for i) N_b^γ, N_b^η - the number of bits to represent γ_k and $\bar{\eta}_m$, respectively, $\forall k, m$ (usually, $N_b^\eta \gg N_b^\gamma$, for that $\bar{\eta}_m$ approximates a continuous number), ii) N_s - total number of solution candidates, i.e., $\theta_1, \theta_2, \dots, \theta_{N_s}$. So that the initial population can be represented by a matrix in dimension $\mathcal{R}^{N_s \times (N_b^\gamma K + N_b^\eta p)}$, in which each string represents a solution of θ . The following procedure of the GA used to find the optimal θ for the problems (5.24) and (5.25) is very similar to that finding the optimal $\bar{\gamma}$ for the problems in (4.12), while in the GA program in this chapter, the crossover operator is performed among the bits representing γ_k and the bits representing $\bar{\eta}_m$, respectively, and the termination criterion is a combination of generation numbers, fitness threshold, and fitness convergence.

It is well known that a good initial population will speed up the GA to find a solution. With the arguments on the relationship between the optimal polynomial operators (equivalent to the optimal $\{\gamma_k\}, \{\bar{\eta}_m\}$) and the poles of the filter to be implemented, the initial population can be chosen such that the corresponding polynomial operators have their roots close to the dominant poles of the given filter. This, as supported by many numerical examples, can make the GA find a good solution very efficiently.

5.3.4 Numerical examples

In this subsection, we examine the performance of the two proposed structures (i.e., ρ IDFIIt, ρ ISS) and compare them with some existing structures such as the optimal fully parameterized state-space realization R_G [8], [9], z DFIIt, and δ DFIIt [28] as well as ρ DFIIt and its equivalent state-space realization ρ SS in Chapter 3. For all the examples, $B_\gamma = 4$ is assumed, leading to $N_s^\gamma = 5$.

$N_s^\eta = 19$ is chosen such that a 4 decimal point precision is achieved for $\bar{\eta}_m$, $\forall m$.

Example 5.1 : First, we consider a sixth order Butterworth bandpass filter generated by the MATLAB command $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.35])$, which has a normalized bandwidth $BW = (0.35 - 0.25)/2 = 0.05$ with a normalized center frequency $f_0 = 0.5 \times (0.25 + 0.35)/2 = 0.15$. The corresponding poles are located at $\{0.4364 \pm j0.8074, 0.6489 \pm j0.6695, 0.5137 \pm j0.6802\}$.

For this example, R_G has a roundoff noise gain of 13.5869, which needs 49 nontrivial parameters, while the ρDFIIt and ρSS yield a roundoff noise gain of 365.8191 and 222.3884, respectively.

With the GA program, the optimal parameters for the ρIDFIIt structure are obtained:

$$\begin{aligned}\{\gamma_k\} &= \{ 0.5 \ 0.5625 \ 0.5 \ 0.5 \ 0.625 \ 0.5625 \} \\ \{\bar{\eta}_m\} &= \{ 0.495601 \ 0.554839 \ 0.436950 \}\end{aligned}$$

The corresponding ρIDFIIt has a roundoff noise gain of 6.6739, just 1.82% of that of the ρDFIIt . The optimized ρISS is obtained with

$$\begin{aligned}\{\gamma_k\} &= \{ 0.5 \ 0.5 \ 0.75 \ 0.25 \ 0.5 \ 0.625 \} \\ \{\bar{\eta}_m\} &= \{ 0.567059 \ 0.605882 \ 0.424706 \}\end{aligned}$$

which yields a roundoff noise gain of 6.1454, i.e., 2.76% of that of the ρSS . Table 5.1 provides the statistics of the roundoff noise gain G , the L_2 -sensitivity measure M , and the number of nontrivial parameters N_p for each structure as well as the runtime (minute) of GA, denoted as T_{GA} for ρDFIIt , ρSS , ρIDFIIt ,

Table 5.1: Comparison Results for Example 5.1

Realization	G	M	N_p	T_{GA}
$zDFIIt$	1.0514×10^4	4.3974×10^4	16	–
$\delta DFIIIt$	4.7877×10^3	2.0015×10^4	19	–
R_G	1.3587×10^1	3.6482×10^1	49	–
$\rho DFIIIt$	3.6582×10^2	6.8501×10^2	25	0.0855
ρSS	2.2239×10^2	1.0332×10^3	24	0.2302
$\rho IDFIIIt$	6.6739×10^0	9.9385×10^0	28	2.3201
ρISS	6.1454×10^0	1.4757×10^1	26	1.9167

and ρISS (Pentium 4 CPU, 1.90GHz and 256RAM).

Computations show that for the $\rho IDFIIIt$ the union of the roots of $\bar{q}_k(z)$, $\forall k$, is $\{0.5000 \pm j0.7449, 0.5937 \pm j0.6603, 0.5313 \pm j0.7033, 0.5625, 0.5000, 0.5625\}$, and for the ρISS , $\{0.5000 \pm j0.7371, 0.5625 \pm j0.6487, 0.5000 \pm j0.7530, 0.5000, 0.2500, 0.6250\}$. As seen, they are around the poles of the filter.

In what follows, let us consider two simulations to compare the numerical performance of different FWL realizations for the band-pass filter in this example. Firstly, to illustrate the effect of the parameter quantization, we compute and compare the magnitude responses for several realizations whose coefficients are implemented with $f_c(I_c, B_c)$. Figure 5.3 shows the magnitude responses and phase responses of the ideal filter and the $zDFIIt$, $\delta DFIIIt$, R_G , $\rho DFIIIt$, and $\rho IDFIIIt$ structures. One can see that the magnitude responses and phase responses of R_G and $\rho IDFIIIt$ are almost overlapped with the ideal response.

Secondly, to illustrate the effect of the roundoff noise, we compute and compare the output for each of the five FWL realizations in the above simulation with that of the ideal filter, shown in Figure 5.4, where the input signal is generated by MATLAB command $u(n) = \text{randn}(1, 1000)$. Since in this simulation, the effect of parameter quantization is not taken into account, assume that the

Figure 5.3: Magnitude responses (in dB) and phase responses (in degree) for Example 5.1: i – ideal; ii – z DFIIt with $f_c(4, 7)$; iii – δ DFIIt with $f_c(4, 7)$; iv – R_G with $f_c(0, 11)$; v – ρ DFIIt with $f_c(3, 8)$; vi – ρ IDFIIt with $f_c(0, 11)$

coefficients are implemented with $f_c(I_c, B_c)$, which satisfies that $I_c + B_c = 63$, and the signals are implemented with $f_s(I_s, B_s)$. Denote $y(n)$ as the output of the ideal filter, and $y_1(n)$, $y_2(n)$, $y_3(n)$, $y_4(n)$, $y_5(n)$ as that of the z DFIIt, δ DFIIt, R_G , ρ DFIIt, and ρ IDFIIt, respectively. Defining

$$e_x(n) \triangleq y(n) - y_x(n), \text{ for } x = 1, 2, 3, 4, 5.$$

Computations show that the mean and variance of each error sequence $e_x(n)$ are

$$\begin{cases} \bar{e}_1 = 1.9859 \times 10^{-4}, & \sigma_{e_1}^2 = 2.4951 \times 10^{-4} \\ \bar{e}_2 = 2.0347 \times 10^{-5}, & \sigma_{e_2}^2 = 1.3355 \times 10^{-4} \\ \bar{e}_3 = -7.9730 \times 10^{-7}, & \sigma_{e_3}^2 = 5.5413 \times 10^{-7} \\ \bar{e}_4 = -1.8695 \times 10^{-6}, & \sigma_{e_4}^2 = 1.1966 \times 10^{-5} \\ \bar{e}_5 = -1.3527 \times 10^{-5}, & \sigma_{e_5}^2 = 4.7824 \times 10^{-7} \end{cases}$$

Example 5.2: In this example, we use the following class of low-pass filters:

$$[V_b, V_a] = \text{butter}(K, 0.20)$$

to examine the effect of filter order on the roundoff noise gain for each structure. As the equivalent state-space realization always yields a smaller roundoff noise gain than its DFIIt counterpart, Table 5.2 shows the roundoff noise gain and number of nontrivial parameters for the five state-space realizations.

Example 5.3: We now consider the effect of filter bandwidth on the roundoff

Figure 5.4: Simulation for Example 5.1: (a) – ideal output; (b) – output error of $zDFIIt$ with $f_s(2, 13)$; (c) – output error of $\delta DFIIIt$ with $f_s(2, 13)$; (d) – output error of R_G with $f_s(0, 15)$; (e) – output error of $\rho DFIIIt$ with $f_s(1, 14)$; (f) – output error of $\rho IDFIIt$ with $f_s(1, 14)$

Table 5.2: Comparison Results for Example 5.2

Realization	$K = 5$		$K = 6$		$K = 7$		$K = 8$	
	G	N_p	G	N_p	G	N_p	G	N_p
zSS	8.3685×10^2	16	7.2582×10^3	19	6.4809×10^4	22	5.8917×10^5	25
δSS	5.1864×10^1	16	1.4836×10^2	19	5.1197×10^2	22	1.8364×10^3	25
R_G	9.8125×10^0	36	1.1985×10^2	49	1.4294×10^1	64	1.6738×10^1	81
ρSS	5.8519×10^0	20	7.7542×10^0	24	1.0753×10^1	28	1.5675×10^1	32
ρISS	4.7963×10^0	22	5.2639×10^0	26	5.8684×10^0	31	6.2954×10^0	35

Table 5.3: Comparison Results for Example 5.3

Realization	$BW = 0.050$	$BW = 0.075$	$BW = 0.100$	$BW = 0.125$	$BW = 0.150$
zSS	1.0334×10^4	2.5111×10^3	1.1011×10^3	7.2433×10^2	6.6823×10^2
δSS	4.1908×10^3	1.1274×10^3	4.8459×10^2	2.7031×10^2	1.8197×10^2
R_G	1.3587×10^1	1.3587×10^1	1.3587×10^1	1.3587×10^1	1.3587×10^1
ρSS	2.3016×10^2	4.8598×10^1	1.8254×10^1	1.0203×10^1	7.8869×10^0
ρISS	5.9982×10^0	6.0990×10^0	6.2300×10^0	6.3432×10^0	6.3905×10^0

noise gain for the following class of filters:

$$[V_b, V_a] = \text{butter}(3, [0.30 - BW \ 0.30 + BW])$$

where BW is the normalized bandwidth varying from 0.050 to 0.150. Once again, the five state-space realizations are examined and the corresponding results are given in Table 5.3.

Example 5.4: In this last example, we use two filters to compare the frequency responses of the five state-space realizations. Denote $H_{FWL}^{B_c}(\omega)$ as the frequency response for each of the FWL realizations whose coefficients are implemented in fixed point with $f_c(4, B_c)$, where B_c ranges from 4 to 30. We compare $H_{FWL}^{B_c}(\omega)$ with the ideal frequency response $H(\omega)$ implemented with infinite precision by

Figure 5.5: Comparison results of $R(B_c)$ for the bandpass filter $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.35])$ for Example 5.4: i – zSS ; ii – δSS ; iii – R_G ; iv – ρSS ; v – ρISS

computing the *worst deviation* over the frequency range:

$$R(B_c) \triangleq \log\left(\max_{\omega \in (0, 2\pi)} |H(\omega) - H_{FWL}^{B_c}(\omega)|\right).$$

One of the filters examined is the bandpass filter considered in Example 5.1, and the other is the low-pass filter considered in Example 5.2 with $K = 8$. Figures 5.5 and 5.6 show the comparison results of $R(B_c)$ for each FWL state-space realization for the two filters, respectively, from which, we can clearly observe the superiority of the ρISS structure over the other four realizations whatever B_c is.

Example 5.5: In this example, we consider a sixth order bandpass Chebyshev Type I filter with 0.5 dB of ripple in the passband, generated by MATLAB

Figure 5.6: Comparison results of $R(B_c)$ for the low-pass filter $[V_b, V_a] = \text{butter}(8, 0.2)$ for Example 5.4: i – $z\text{SS}$; ii – δSS ; iii – R_G ; iv – ρSS ; v – ρISS

command $[V_b, V_a] = \text{cheby1}(3, 0.5, [0.25 \ 0.35])$. The corresponding poles are located at $\{0.4272 \pm j0.8455, 0.6804 \pm j0.6747, 0.5414 \pm j0.7255\}$.

For this example, R_G has a roundoff noise gain of 16.1360, which needs 49 nontrivial parameters, while the ρDFIIt and ρSS yield a roundoff noise gain of 808.8448 and 568.5645, respectively.

With the GA program, the optimal parameters for the ρIDFIIt structure are obtained:

$$\begin{aligned} \{\gamma_k\} &= \{0.875 \ 0.25 \ 0.6875 \ 0.5 \ 0.1875 \ 0.875\} \\ \{\bar{\eta}_m\} &= \{0.651711 \ 0.528543 \ 0.682991\}. \end{aligned}$$

The corresponding ρIDFIIt has a roundoff noise gain of 11.0055. The optimized

Table 5.4: Comparison Results for Example 5.5

Realization	G	M	N_p
$zDFIIt$	2.3903×10^4	8.1624×10^4	16
$\delta DFIIIt$	9.1494×10^3	2.8545×10^4	19
R_G	1.6136×10^1	4.8940×10^1	49
$\rho DFIIIt$	8.0884×10^2	1.5002×10^3	25
ρSS	5.6856×10^2	2.2899×10^3	24
$\rho IDFIIIt$	1.1006×10^1	1.9833×10^1	28
ρISS	8.8404×10^0	2.3604×10^1	26

ρISS is obtained with

$$\begin{aligned} \{\gamma_k\} &= \{ 0.5 \quad 0.75 \quad 0.875 \quad 0.125 \quad 0.6875 \quad 0.5 \} \\ \{\bar{\eta}_m\} &= \{ 0.607722 \quad 0.745552 \quad 0.527566 \}. \end{aligned}$$

Table 5.4 provides the statistics of the roundoff noise gain G , the L_2 -sensitivity measure M and the number of nontrivial parameters N_p for seven different structures.

The results are self-explanatory. One observes that in all the examples, the proposed two structures can beat the traditional optimal realizations R_G in terms of minimizing roundoff noise, L_2 -sensitivity, and structure efficiency. Compared with the recently proposed $\rho DFIIIt$ and its corresponding ρSS , the two proposed structures always yield a smaller roundoff noise gain and sensitivity than their counterpart, as expected, and the improvement can be very significant especially for the type of bandpass filters with narrow bandwidth. Numerical examples have been carried out for high-pass filters and all these conclusions hold.

5.4 The ρ ISS structures for RBM scheme

So far, we have analyzed the FWL performance for the proposed ρ IDFIIt structure and its equivalent state-space realization ρ ISS in terms of RAM scheme. In what follows, assuming the RBM implementations, we will investigate the ρ ISS structure without/with error feedback.

5.4.1 Roundoff noise gain for the ρ ISS

It follows from (5.11) that $A_\rho = D_\gamma + (A_o - A_\eta)D_\Delta \triangleq D_\gamma + A_K$. Assuming the B_c -bit fixed point implementation, it was known that the free parameters γ_k are in a B_γ -bit FWL format, with B_γ always much less than B_c . Hence, the roundoff noise gain due to γ_k can be neglected. Therefore, the FWL implementation of the ρ ISS structure given by (5.10) becomes

$$\begin{cases} x^*(n+1) &= D_\gamma x^*(n) + A_K Q[x^*(n)] + B_\rho u(n) \\ &= A_\rho Q[x^*(n)] + B_\rho u(n) + D_\gamma e(n) \\ y^*(n) &= C_\rho Q[x^*(n)] + \beta_0 u(n) \end{cases} \quad (5.31)$$

with $e(n) \triangleq x^*(n) - Q[x^*(n)]$ the roundoff error vector.

It can be found that (5.31) has the similar format to the state-space model with error feedback shown in (2.24). Hence the roundoff noise gain for the ρ ISS structure, in terms of RBM scheme, can be computed by substituting $(A_\rho, B_\rho, C_\rho, \beta_0)$ and D_γ into (2.25), i.e.,

$$G_{ss} = \text{tr}[(A_\rho - D_\gamma)^T W_o^\rho (A_\rho - D_\gamma) + C_\rho^T C_\rho] = \text{tr}(A_K^T W_o^\rho A_K + C_\rho^T C_\rho). \quad (5.32)$$

5.4.2 The ρ ISS structure with error feedback

Looking at Figure 2.1, suppose that the roundoff error vector $e(n)$ is fed back through a matrix D into the FWL model (5.31). It follows from (2.24) - (2.25) that the roundoff noise gain for the ρ ISS structure with error feedback, denoted as G_D , can be given by

$$G_D = \text{tr}[(A_K - D)^T W_o^\rho (A_K - D) + C_\rho^T C_\rho]. \quad (5.33)$$

In the following, let's see how to choose the feedback matrix D to make G_D less than G_{ss} . It was discussed in [46] that D could be chosen under three cases: a scalar, diagonal, or general matrix, which represented an increase of structure complexity and, contrarily, a decrease of total roundoff noise gain. In this thesis, we consider the ρ ISS structure with a general or diagonal feedback matrix.

1) D is a general matrix

Rewriting (5.33) with noting (5.32), we obtain

$$G_D = G_{ss} + \text{tr}(D^T W_o^\rho D) - 2\text{tr}(D^T W_o^\rho A_K). \quad (5.34)$$

Taking the first derivative of (5.34) with respect to the error feedback matrix D yields

$$\frac{\partial G_D}{\partial D} = 2W_o^\rho (D - A_K).$$

By choosing the error feedback matrix as $D^{\text{opt}} = A_K$, G_D achieves its minimum

value

$$G_D^{\min} = \text{tr}(C_\rho^T C_\rho) = \Delta_1^2 < G_{ss}. \quad (5.35)$$

2) D is a diagonal matrix

In this case, $D = \text{diag}(d_1, d_2, \dots, d_K)$, and

$$G_D = G_{ss} + \sum_{k=1}^K \{d_k^2 W_o^\rho(k, k) - 2d_k [(W_o^\rho A_K)(k, k)]\}. \quad (5.36)$$

From (5.36), it is clear that $G_D < G_{ss}$ holds if $\{d_k\}$ satisfy

$$d_k [d_k - 2 \frac{(W_o^\rho A_K)(k, k)}{W_o^\rho(k, k)}] < 0, \quad k = 1, 2, \dots, K.$$

Taking the first derivative of (5.36) with respect to d_k and setting it to zero, the optimal d_k is obtained

$$d_k^{\text{opt}} = \frac{(W_o^\rho A_K)(k, k)}{W_o^\rho(k, k)} \quad (5.37)$$

which leads G_D to

$$G_D^{\min} = G_{ss} - \sum_{k=1}^K \frac{[(W_o^\rho A_K)(k, k)]^2}{W_o^\rho(k, k)} < G_{ss}. \quad (5.38)$$

One can see from (5.32) that $G_{ss} = \text{tr}(W_o^\rho) + \text{tr}(D_\gamma^T W_o^\rho D_\gamma) - 2\text{tr}(D_\gamma^T W_o^\rho A_\rho)$.

It means that G_{ss} and hence G_D are closely related to the diagonal elements of W_o^ρ , which can be computed by

$$W_o^\rho(k, k) = \|C_\rho(zI - A_\rho)^{-1} e_k\|_2^2 = \|H_k(z)\|_2^2$$

with $H_k(z)$ defined and analyzed in Section 5.3. This means that the relationship

between the optimal polynomial operators and the filter poles is still applicable for minimizing the roundoff noise gain of the ρ ISS structure without/with error feedback in RBM scheme.

5.4.3 Design examples

In the following, we examine the performance of the optimized ρ ISS structures without/with error feedback and compare them to some existing structures such as the state-space realization of the ρ DFIIt structure (denoted as ρ SS) as well as the optimal fully parameterized state-space realization R_G [8], [9] without/with error feedback [46]. For convenience, denote the ρ ISS structure with a general and a diagonal error feedback matrix as ρ ISS^g and ρ ISS^d, respectively, and similarly the R_G in two cases of error feedback matrices as R_G^g and R_G^d . For all the examples, $B_\gamma = 4$ is assumed and the error feedback coefficients are implemented by 4-bit quantization.

Example 5.6: First, we consider the Butterworth bandpass filter used in the Example 5.1 of Section 5.3.4: $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.35])$.

For this example, R_G has a roundoff noise gain of 0.9410. With the GA program, the optimal free parameters for the ρ ISS structure are obtained:

$$\begin{aligned} \{\gamma_k\} &= \{ 0.4375 \ 0.5625 \ 0.5 \ 0.5 \ 0.5 \ 0.625 \}, \\ \{\bar{\eta}_m\} &= \{ 0.463265 \ 0.541790 \ 0.478061 \}. \end{aligned}$$

Table 5.5: Comparison Results for Example 5.6

Realization	R_G	ρ SS	ρ ISS	R_G^g	R_G^d	ρ ISS ^g	ρ ISS ^d
G	0.9410	39.6150	0.7115	0.0863	0.6462	0.1070	0.7084
N_p	49	24	26	85	55	45	32

⁴Next, we apply error feedback to above optimized ρ ISS structure. In one case where D is a general matrix, taking $D = A_K$, computations show that

$$D^{\text{opt}} = \begin{bmatrix} 0.073189 & 0.816485 & 0 & 0 & 0 & 0 \\ -0.644640 & 0 & 0.270386 & 0 & 0 & 0 \\ 0.164257 & 0 & 0 & 0.854592 & 0 & 0 \\ -0.053558 & 0 & -0.633975 & 0 & 0.296205 & 0 \\ -0.036084 & 0 & 0 & 0 & 0 & 0.592821 \\ -0.052308 & 0 & 0 & 0 & -0.806418 & 0 \end{bmatrix}.$$

In the other case, where D is a diagonal matrix, it is calculated by (5.37) that

$$D^{\text{opt}} = \text{diag}(0.018179 \quad 0.036148 \quad 0.069597 \quad 0.096714 \quad 0.109220 \quad -0.000304).$$

Table 5.5 provides the statistics of the roundoff noise gain G and the number of nontrivial parameters N_p for each of the seven structures. One can see that the ρ ISS structure can yield a roundoff noise gain which is just 1.8% of the one by the ρ SS at the price of only 2 more nontrivial parameters, and it excels the R_G at both roundoff noise performance and structure simplicity. For the comparison of the realizations with error feedback, there exists a trade-

⁴Since the optimized ρ ISS structure with error feedback can yield a roundoff noise gain which is very close to the structure optimizing the free parameters and feedback coefficients simultaneously, we apply the error feedback to the optimized ρ ISS structure directly due to the space and efficiency consideration.

off between roundoff noise gain and structure efficiency. One should note that ρISS^g and ρISS^d possess even less parameters than R_G . Computations show that for ρISS , the union of the roots of $\bar{q}_k(z)$, $\forall k$ is $\{0.5000 \pm j0.7361, 0.5625 \pm j0.6886, 0.5000 \pm j0.6778, 0.5625, 0.5000, 0.6250\}$. As seen, they are around the poles of the given filter.

Example 5.7: Now, we use the following class of low-pass filters:

$$[V_b, V_a] = \text{butter}(K, 0.10)$$

to examine the effect of the filter order K on the performance for each structure. In Figure 5.7, (a) and (b) show the roundoff noise gain in dB ($10\log_{10}G$) and the number of nontrivial parameters (N_p) in each structure versus filter order (K), respectively.

Example 5.8: This example is to consider the effect of filter bandwidth on the roundoff noise gain for the following class of filters:

$$[V_b, V_a] = \text{butter}(3, [0.30 - BW \ 0.30 + BW])$$

where the normalized bandwidth BW varies from 0.050 to 0.150. Since the filter is fixed at sixth order, the comparison of the number of nontrivial parameters for each structure can be referred to Figure 5.7 (b). For the roundoff noise performance, the seven realizations are examined and the corresponding results are shown in Figure 5.8.

Figure 5.7: Structure performance vs filter order for Example 5.7: i — R_G ; ii — ρ SS; iii — ρ ISS; iv — R_G^g ; v — R_G^d ; vi — ρ ISS g ; vii — ρ ISS d

Figure 5.8: Roundoff noise performance vs normalized bandwidth for Example 5.8: i — R_G ; ii — ρ SS; iii — ρ ISS; iv — R_G^g ; v — R_G^d ; vi — ρ ISS g ; vii — ρ ISS d

Example 5.9: Finally, let us see the filtering performance of the R_G , ρ SS and ρ ISS structures for two filters. We compute and compare the filter output for each of these three FWL realizations with that of the ideal filter, with the input generated by MATLAB command $u(n) = \text{randn}(1, 500)$. Since the frequency response can reflect the effect of parameter quantization in a certain extent, which was studied and compared for these three structures in Example 5.4 of Section 5.3.4, the main objective here is to examine the effect of signal rounding errors. For no loss of generality, assume the coefficients are implemented with $f_c(I_c, B_c)$, which satisfies that $I_c + B_c = 15$. Denote $y_{\text{FWL}}^{B_s}(n)$ as the output for each of the FWL structures whose signals are implemented with $f_s(0, B_s)$, with B_s ranging from 4 to 30. We compare $y_{\text{FWL}}^{B_s}(n)$ with the ideal output $y(n)$

Figure 5.9: Comparison results of $V(B_s)$ for the bandpass filter $[V_b, V_a] = \text{butter}(3, [0.25 \ 0.35])$ for Example 5.9: i – R_G ; ii – ρSS ; iii – ρISS

by computing the *output error deviation*:

$$V(B_s) \triangleq \log \{ \text{var} [y(n) - y_{FWL}^{B_s}(n)] \}$$

where $\text{var}[x(n)]$ means the variance of signal $x(n)$.

One of the filters examined is a bandpass filter considered in Example 5.5, and the other is a low-pass filter considered in Example 5.6 with $K = 6$. Figures 5.9 and 5.10 plot the $V(B_s)$ for each of the three FWL structures of the two filters, respectively. We can find that the proposed ρISS structure has the obvious superiority over the other two structures whatever the filter type or the number of bits for B_s is.

One observes that for all the examples, the ρISS structure can over-perform the traditional optimal realization R_G in terms of minimizing roundoff noise and

Figure 5.10: Comparison results of $V(B_s)$ for the low-pass filter $[V_b, V_a] = \text{butter}(6, 0.1)$ for Example 5.9: i – R_G ; ii – ρSS ; iii – ρISS

structure efficiency. Compared with the recently proposed ρSS , the ρISS always yields a smaller roundoff noise gain than its counterpart. For the realizations with error feedback, there exists a tradeoff between roundoff noise and structure efficiency. It should be noted that with the increase of the filter order, the advantage of the proposed structures in this thesis in terms of structure simplicity becomes more and more obvious. Numerical examples have been carried out for high-pass filters and all the conclusions still hold.

5.5 Conclusion

The contribution in this chapter is fourfold. Firstly, based on the polynomial parametrization concept, two efficient structures have been exploited, which have more degrees of freedom than the one proposed in Chapter 3. Secondly, the FWL performance of each of the structures has been analyzed by deriving

the expressions for roundoff noise gain in terms of RAM scheme and L_2 -norm based sensitivity. It is shown that the ρ ISS structure always yields a smaller roundoff noise gain than its IDFIIt counterpart. Thirdly, the optimal structure problems have been formulated, which, can be effectively attacked using the genetic algorithm with the established relationship among the roundoff noise gain, the poles of the filter and the roots of the polynomials. Fourthly, taking RBM scheme into account, roundoff noise performance for the ρ ISS structure has been investigated. By applying the error feedback technique, the roundoff noise gain of the state-space model can be further reduced. Extensive examples have been carried out to illustrate the design procedure and compare the performance of the proposed structures with some existing structures.

Chapter 6

DFII Based Structures for 2-D Digital Filters

In the previous chapters, based on the concept of polynomial parametrization and error feedback, several robust digital filter structures have been proposed for 1-D digital filters. The most attractive features about these structures are that they are very efficient and can provide more degrees of freedom allowing us to minimize the FWL effects at the same time. One intuitive question to be asked is that when the methods for deriving these structures are extended to the case of 2-D digital filters if similar results can be obtained, which is the main concern in this chapter.

6.1 Introduction

It is well known that the general transfer function of a (M, N) th order 2-D digital filter, say (2.26), can not be realized with a canonic realization (see Section

2.3). Therefore, in many applications, 2-D digital filters are usually designed with either separable-numerator or separable-denominator. Alternatively, the class of 2-D separable-denominator digital filters is considered to be the most important in numerous multidimensional digital signal processing problems. This is partially due to the ease of stability test, symmetric properties in the frequency domain, computational tractability, and so on. Many studies have been reported on the various topics of 2-D separable-denominator digital filters [65] - [70].

The optimal 2-D state-space realization with minimum roundoff noise subject to l_2 -norm dynamic range constraints has been investigated in [4], [54] - [56]. However, this structure, just like its counterpart in 1-D field [8], [9], is always fully parameterized, possessing $(M + N + 1)^2$ nontrivial parameters. Although in recent years, delta-operator, which is well applicable for the type of narrow band low-pass filters, is very popular for 1-D case, it is seldom studied for 2-D filters [57]. Besides roundoff noise consideration, many research results obtained for 1-D case have also been extended to the problem of minimizing the coefficient sensitivity in 2-D state-space digital filters [58] - [64]. Some of them evaluate the sensitivity by using a mixture of L_1/L_2 norms [58] - [60]. The others rely on the use of a pure L_2 -norm [61] - [64]. In [62], the weighted-sensitivity minimization of 2-D state-space digital filters has been considered in both cases of a mixture of L_1/L_2 norms and a pure L_2 -norm. It should be noted that the L_2 -sensitivity minimization is more natural and reasonable than the conventional L_1/L_2 mixed sensitivity minimization, but it is more challenging. In [64], an iterative procedure is presented for synthesizing the optimal state-space filter structures that minimize the L_2 -sensitivity. However, all these minimizations of L_2 -sensitivity do not consider the dynamic-range constraints of the state-variables.

The main objective in this chapter is to investigate if the methods for deriving 1-D efficient digital filter structures in this thesis are effective for the case of 2-D filters. Firstly, motivated by the ρ DFIIt structure in Chapter 3, a new structure for 2-D digital filters with separable-denominator is derived based on a set of polynomial operators, denoted as the ZL structure. This structure, being a mixture of DFII for its horizontal part and DFIIt for its vertical part, possesses only $MN + 3(M + N) + 1$ nontrivial parameters plus $M + N$ free parameters. The equivalent state-space realization of the 2-D ZL yields another structure, which is of the similar structure simplicity to the ZL. It is shown that the roundoff noise performance of the ZL structure is closely related to the polynomial operators for the vertical part, i.e., when the free parameters in the vertical polynomials are chosen closely to the real part of the filter vertical poles, the roundoff noise gain can be much reduced. Based on this observation, the more robust polynomial operators used in the ρ IDFIIt structure are applied to the vertical part of the ZL structure, resulting in another efficient structure, called IZL for convenience. Since in the IZL, the roots of the vertical polynomials can be complex and when they are around the neighborhood of the vertical poles, the roundoff noise gain can be further reduced compared to its ZL counterpart.

An outline of this chapter is as follows. In Section 6.2, starting with the separable-denominator 2-D filters, we re-parameterize the transfer function (2.28) by applying the polynomial operators used in ρ DFIIt to both the horizontal and vertical parts. Based on this new transfer function, the 2-D ZL structure is derived, and hence its equivalent state-space realization. In Section 6.3, performance analysis is carried out for both structures by deriving the expressions of roundoff noise gain and L_2 -norm based sensitivity measure. It is shown, like

in the case of 1-D filters, that the free parameters can be used to minimize the FWL effects. In Section 6.4, the ZL structure for separable-numerator filters is given. Since its derivation procedure and performance analysis are very similar to that for separable-denominator filters, they are not presented in detail in this section. In Section 6.5, based on the observation that the roundoff noise performance of the ZL structure is closely related to the polynomial operators for the vertical part, the IZL structure (for separable-denominator digital filters) is developed which provides more degrees of freedom to minimize FWL effects. Three examples are given in Section 6.6, which show that the proposed structures can yield a better FWL performance than the shift-operator and delta-operator based structures and even beat the optimal fully parameterized state-space realization.

6.2 2-D ZL structure

6.2.1 Transfer function re-parametrization

Denote

$$\rho_{hk}(z_1) \triangleq \frac{z_1 - \gamma_{hk}}{\Delta_{hk}} \triangleq h_k(z_1), \quad \rho_{vl}(z_2) \triangleq \frac{z_2 - \gamma_{vl}}{\Delta_{vl}} \triangleq v_l(z_2), \quad \forall k, l \quad (6.1)$$

where $\{\gamma_{hk}\}$, $\{\gamma_{vl}\}$ are two sets of free parameters and $\{\Delta_{hk}\}$, $\{\Delta_{vl}\}$ are coupling coefficients. Consider the following polynomial operators:

$$\begin{cases} p_k(z_1) \triangleq h_k^{-1}(z_1)p_{k-1}(z_1), & p_0(z_1) \triangleq \prod_{i=1}^M h_i(z_1), \quad \forall k \\ q_l(z_2) \triangleq v_l^{-1}(z_2)q_{l-1}(z_2), & q_0(z_2) \triangleq \prod_{j=1}^N v_j(z_2), \quad \forall l \end{cases} \quad (6.2)$$

Let $\bar{T}_1 \in \mathcal{R}^{(M+1) \times (M+1)}$ and $\bar{T}_2 \in \mathcal{R}^{(N+1) \times (N+1)}$ be two nonsingular matrices and define

$$\begin{cases} \bar{z}_1 \triangleq \begin{bmatrix} z_1^M & z_1^{M-1} & \cdots & z_1^{M-k} & \cdots & 1 \end{bmatrix}^T \\ \bar{z}_2 \triangleq \begin{bmatrix} z_2^N & z_2^{N-1} & \cdots & z_2^{N-l} & \cdots & 1 \end{bmatrix}^T \end{cases}$$

one then has

$$\begin{cases} \bar{p}(z_1) \triangleq \begin{bmatrix} p_0(z_1) & p_1(z_1) & \cdots & p_{M-k}(z_1) & \cdots & p_M(z_1) \end{bmatrix}^T = \bar{T}_1 \bar{z}_1 \\ \bar{q}(z_2) \triangleq \begin{bmatrix} q_0(z_2) & q_1(z_2) & \cdots & q_{N-l}(z_2) & \cdots & q_N(z_2) \end{bmatrix}^T = \bar{T}_2 \bar{z}_2 \end{cases} \quad (6.3)$$

Hence, the transfer function (2.28) can be rewritten as

$$\begin{aligned} H(z_1, z_2) &= \frac{\bar{z}_1^T M_b \bar{z}_2}{(V_{a1}^T \bar{z}_1)(V_{a2}^T \bar{z}_2)} = \frac{\bar{z}_1^T \bar{T}_1^T \kappa_1^{-1} \bar{T}_1^{-T} M_b \bar{T}_2^{-1} \kappa_2^{-1} \bar{T}_2 \bar{z}_2}{(\kappa_1^{-1} V_{a1}^T \bar{T}_1^{-1} \bar{T}_1 \bar{z}_1)(\kappa_2^{-1} V_{a2}^T \bar{T}_2^{-1} \bar{T}_2 \bar{z}_2)} \\ &\triangleq \frac{\bar{p}^T(z_1) M_\beta \bar{q}(z_2)}{F(z_1) G(z_2)} \end{aligned} \quad (6.4)$$

where $M_b = \{b_{kl}\} \in \mathcal{R}^{(M+1) \times (N+1)}$, $V_{a1} = \{a_{1k}\} \in \mathcal{R}^{(M+1) \times 1}$, $V_{a2} = \{a_{2l}\} \in \mathcal{R}^{(N+1) \times 1}$, and

$$\begin{aligned} F(z_1) &\triangleq \kappa_1^{-1} V_{a1}^T \bar{T}_1^{-1} \bar{T}_1 \bar{z}_1 \triangleq V_f^T \bar{p}(z_1), \quad G(z_2) \triangleq \kappa_2^{-1} V_{a2}^T \bar{T}_2^{-1} \bar{T}_2 \bar{z}_2 \triangleq V_g^T \bar{q}(z_2) \\ M_\beta &\triangleq \begin{bmatrix} \beta_{00} & \cdots & \beta_{0l} & \cdots & \beta_{0N} \\ & & \vdots & & \\ \beta_{k0} & \cdots & \beta_{kl} & \cdots & \beta_{kN} \\ & & \vdots & & \\ \beta_{M0} & \cdots & \beta_{Ml} & \cdots & \beta_{MN} \end{bmatrix} \triangleq \begin{bmatrix} \bar{\beta}_0 & \cdots & \bar{\beta}_l & \cdots & \bar{\beta}_N \end{bmatrix} \end{aligned}$$

with $\kappa_1 = \prod_{k=1}^M \Delta_{hk}$, $\kappa_2 = \prod_{l=1}^N \Delta_{vl}$ such that $V_f(1) = f_0 = 1$, $V_g(1) = g_0 = 1$.

It is clear that $H(z_1, z_2)$ is now re-parameterized with $V_f = \{f_k\}$, $V_g = \{g_l\}$ and M_β under the *polynomial operators* $\{p_k(z_1), q_l(z_2)\}$. In the following, let us consider how to implement the filter using the polynomial-parameterized transfer function (6.4). One can see that

$$H(z_1, z_2) = \frac{\bar{p}^T(z_1)}{F(z_1)} \cdot \frac{\sum_{l=0}^N \bar{\beta}_l q_l(z_2)}{G(z_2)}.$$

Define

$$\begin{aligned} \bar{w}_h(i, j) &= \begin{bmatrix} w_{h0}(i, j) & w_{h1}(i, j) & \cdots & w_{hk}(i, j) & \cdots & w_{hM}(i, j) \end{bmatrix} \\ &\triangleq \frac{\bar{p}^T(z_1)}{F(z_1)} u(i, j) = \frac{\bar{p}^T(z_1)/p_0(z_1)}{1 + \sum_{k=1}^M f_k p_k(z_1)/p_0(z_1)} u(i, j) \end{aligned}$$

one then has

$$w_{hk}(i, j) = \frac{p_k(z_1)}{p_0(z_1)} \cdot \frac{1}{1 + \sum_{k=1}^M f_k p_k(z_1)/p_0(z_1)} u(i, j), \quad \forall k.$$

With some manipulations, $w_{h0}(i, j)$ can be written as

$$w_{h0}(i, j) = u(i, j) - \sum_{k=1}^M f_k w_{hk}(i, j) \quad (6.5)$$

which can be implemented with an DFII structure, depicted in the horizontal part of Figure 6.1. In the same way, one can show that

$$\begin{aligned} y(i, j) &= \bar{w}_h(i, j) \frac{\sum_{l=0}^N \bar{\beta}_l q_l(z_2)/q_0(z_2)}{1 + \sum_{l=1}^N g_l q_l(z_2)/q_0(z_2)} \\ &= \sum_{l=0}^N \frac{q_l(z_2)}{q_0(z_2)} \bar{w}_h(i, j) \bar{\beta}_l - \sum_{l=1}^N \frac{q_l(z_2)}{q_0(z_2)} g_l y(i, j) \end{aligned} \quad (6.6)$$

which can be implemented with an DFII structure, depicted in the vertical

part of Figure 6.1.

Therefore, Figure 6.1 shows the corresponding block diagram to (6.5) and (6.6), which can be described in time-domain by the following equations:

$$\left\{ \begin{array}{l} y(i, j) = w_{v1}(i, j) + \sum_{i=0}^M \beta_{i0} w_{hi}(i, j) \\ w_{h0}(i, j) = u(i, j) - \sum_{i=1}^M f_i w_{hi}(i, j) \\ w_{h1}(i, j) = h_1^{-1}(z_1) w_{h0}(i, j) \\ \vdots \\ w_{hk}(i, j) = h_k^{-1}(z_1) w_{h(k-1)}(i, j) \\ \vdots \\ w_{hM}(i, j) = h_M^{-1}(z_1) w_{h(M-1)}(i, j) \\ w_{v1}(i, j) = v_1^{-1}(z_2) [w_{v2}(i, j) + \sum_{i=0}^M \beta_{i1} w_{hi}(i, j) - g_1 y(i, j)] \\ \vdots \\ w_{vl}(i, j) = v_l^{-1}(z_2) [w_{v(l+1)}(i, j) + \sum_{i=0}^M \beta_{il} w_{hi}(i, j) - g_l y(i, j)] \\ \vdots \\ w_{vN}(i, j) = v_N^{-1}(z_2) [\sum_{i=0}^M \beta_{iN} w_{hi}(i, j) - g_N y(i, j)] \end{array} \right. \quad (6.7)$$

where the outputs of operators $h_k^{-1}(z_1)$ and $v_l^{-1}(z_2)$, i.e., $w_{hk}(i, j)$ and $w_{vl}(i, j)$, respectively, can be computed with the realizations depicted in Figure 6.2. The structure shown in Figures 6.1 and 6.2 is called the 2-D ZL structure (for separable-denominator digital filters), which possesses $MN + 3(M + N) + 1$ nontrivial parameters plus $M + N$ free parameters $\{\gamma_{hk}\}$ and $\{\gamma_{vl}\}$.

Figure 6.1: 2-D ZL structure for separable-denominator digital filters

Figure 6.2: Realizations of operators $h_k^{-1}(z_1)$ and $v_l^{-1}(z_2)$

6.2.2 Equivalent state-space realization

Choose $\{x_k^h(i, j)\}$, $\{x_l^v(i, j)\}$ indicated in Figure 6.2 as the state variables and denote $x^h(i, j)$ and $x^v(i, j)$ as the state vectors. It can be shown that the ZL structure is equivalent to the following state-space realization

$$\left\{ \begin{array}{l} \begin{bmatrix} x^h(i+1, j) \\ x^v(i, j+1) \end{bmatrix} = \underbrace{\begin{bmatrix} A_{\rho 1} & A_{\rho 2} \\ A_{\rho 3} & A_{\rho 4} \end{bmatrix}}_{A_\rho} \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix} + \underbrace{\begin{bmatrix} B_{\rho 1} \\ B_{\rho 2} \end{bmatrix}}_{B_\rho} u(i, j) \\ y(i, j) = \underbrace{\begin{bmatrix} C_{\rho 1} & C_{\rho 2} \end{bmatrix}}_{C_\rho} \begin{bmatrix} x^h(i, j) \\ x^v(i, j) \end{bmatrix} + du(i, j) \end{array} \right. \quad (6.8)$$

where

$$\left\{ \begin{array}{l} A_{\rho 1} = \begin{bmatrix} \gamma_{h1} - f_1 \Delta_{h1} & -f_2 \Delta_{h1} & \cdots & -f_M \Delta_{h1} \\ \Delta_{h2} & \gamma_{h2} & & \\ & \ddots & \ddots & \\ & & \Delta_{hM} & \gamma_{hM} \end{bmatrix} \\ A_{\rho 4} = \begin{bmatrix} \gamma_{v1} - g_1 \Delta_{v1} & \Delta_{v2} & & \\ -g_2 \Delta_{v1} & \gamma_{v2} & \ddots & \\ \vdots & & \ddots & \Delta_{vN} \\ -g_N \Delta_{v1} & & & \gamma_{vN} \end{bmatrix} \\ B_{\rho 1} = \begin{bmatrix} \Delta_{h1} & 0 & \cdots & 0 \end{bmatrix}^T, \quad B_{\rho 2} = \begin{bmatrix} \tilde{g}_1 & \tilde{g}_2 & \cdots & \tilde{g}_N \end{bmatrix}^T \\ C_{\rho 1} = \begin{bmatrix} \tilde{f}_1 & \tilde{f}_2 & \cdots & \tilde{f}_M \end{bmatrix}, \quad C_{\rho 2} = \begin{bmatrix} \Delta_{v1} & 0 & \cdots & 0 \end{bmatrix}, \quad d = \beta_{00} \end{array} \right. \quad (6.9)$$

and $A_{\rho 3} = \{\tilde{\beta}_{lk}\} \in \mathcal{R}^{N \times M}$, with $\tilde{f}_k = \beta_{k0} - \beta_{00}f_k$, $\tilde{g}_l = \beta_{0l} - \beta_{00}g_l$, $\tilde{\beta}_{lk} = \beta_{kl} - \beta_{k0}g_l - \beta_{0l}f_k + \beta_{00}f_k g_l$, and all the other elements in A_{ρ} are zero.

The transfer function from $u(i, j)$ to $y(i, j)$ is then

$$H(z_1, z_2) = \beta_{00} + C_{\rho}(Z - A_{\rho})^{-1}B_{\rho}. \quad (6.10)$$

with $Z \triangleq z_1 I_M \oplus z_2 I_N$.

Let $(\bar{A}_{\rho}, \bar{B}_{\rho}, \bar{C}_{\rho}, \beta_{00})$ be the equivalent state-space realization, corresponding to $\Delta_{hk} = 1, \Delta_{vl} = 1, \forall k, l$. It can be shown that the realization $(A_{\rho}, B_{\rho}, C_{\rho}, \beta_{00})$ with any given $\{\Delta_{hk}\}$ and $\{\Delta_{vl}\}$ can be obtained with a diagonal similarity transformation matrix, denoted as T_{sc} , by

$$A_{\rho} = T_{sc} \bar{A}_{\rho} T_{sc}^{-1}, \quad B_{\rho} = T_{sc} \bar{B}_{\rho}, \quad C_{\rho} = \bar{C}_{\rho} T_{sc}^{-1} \quad (6.11)$$

where $T_{sc} = T_{s1} \oplus T_{s2}$, with

$$\begin{cases} T_{s1} = \text{diag}(t_{h1}, t_{h2}, \dots, t_{hM}), & t_{hk} = \prod_{i=1}^k \Delta_{hi}, \quad \forall k \\ T_{s2} = \text{diag}(t_{v1}, t_{v2}, \dots, t_{vN}), & t_{vl} = \prod_{j=1}^l \Delta_{vj}^{-1}, \quad \forall l \end{cases} \quad (6.12)$$

It follows from (6.11) - (6.12) that the l_2 -scaling constraints on state variables are achieved with

$$\begin{cases} \Delta_{h1} = \frac{1}{\sqrt{\bar{W}_c^p(1,1)}}, & \Delta_{hk} = \frac{\sqrt{\bar{W}_c^p(k-1,k-1)}}{\sqrt{\bar{W}_c^p(k,k)}}, & \forall k \\ \Delta_{v1} = \sqrt{\bar{W}_c^p(M+1,M+1)}, & \Delta_{vl} = \frac{\sqrt{\bar{W}_c^p(M+l,M+l)}}{\sqrt{\bar{W}_c^p(M+l-1,M+l-1)}}, & \forall l \end{cases} \quad (6.13)$$

In the sequel, the structures under discussion are assumed to have been l_2 -scaled.

6.3 Performance analysis

6.3.1 Roundoff noise gain

For 2-D digital filters, we only analyze its roundoff noise performance in terms of RAM scheme in detail. For the RBM implementations, the analysis method is similar to that used for 1-D filter structures in the previous chapters.

6.3.1.1 ZL structure

In this structure, the parameters that may produce rounding errors are $\{\beta_{kl}\}$, $\{f_k\}$, $\{g_l\}$, $\{\Delta_{hk}\}$, $\{\Delta_{vl}\}$, and $\{\gamma_{hk}\}$, $\{\gamma_{vl}\}$. First of all, let us consider the roundoff noise by g_l . Looking at Figure 6.1, when $g_l y(i, j)$ is replaced by

$g_l y(i, j) + \psi(g_l) e_{g_l}(i, j)$ with $\psi(x)$ defined in (3.16), (6.7) becomes

$$\left\{ \begin{array}{l} y^*(i, j) = w_{v1}^*(i, j) + \sum_{i=0}^M \beta_{i0} w_{hi}(i, j) \\ w_{h0}(i, j) = u(i, j) - \sum_{i=1}^M f_i w_{hi}(i, j) \\ w_{h1}(i, j) = h_1^{-1}(z_1) w_{h0}(i, j) \\ \vdots \\ w_{hk}(i, j) = h_k^{-1}(z_1) w_{h(k-1)}(i, j) \\ \vdots \\ w_{hM}(i, j) = h_M^{-1}(z_1) w_{h(M-1)}(i, j) \\ w_{v1}^*(i, j) = v_1^{-1}(z_2) [w_{v2}^*(i, j) + \sum_{i=0}^M \beta_{i1} w_{hi}(i, j) - g_1 y^*(i, j)] \\ \vdots \\ w_{vl}^*(i, j) = v_l^{-1}(z_2) \{ w_{v(l+1)}^*(i, j) + \sum_{i=0}^M \beta_{il} w_{hi}(i, j) \\ \quad - [g_l y^*(i, j) + \psi(g_l) e_{g_l}(i, j)] \} \\ \vdots \\ w_{vN}^*(i, j) = v_N^{-1}(z_2) [\sum_{i=0}^M \beta_{iN} w_{hi}(i, j) - g_N y^*(i, j)] \end{array} \right. \quad (6.14)$$

from which, one can find that the roundoff noise by g_l does not affect the horizontal part of the filter. Denote $\Delta y(i, j) \triangleq y^*(i, j) - y(i, j)$, $\Delta w_{vl}(i, j) \triangleq w_{vl}^*(i, j) - w_{vl}(i, j)$. It follows from (6.7) and (6.14) that

$$\left\{ \begin{array}{l} \Delta y(i, j) = \Delta w_{v1}(i, j) \\ \Delta w_{v1}(i, j) = v_1^{-1}(z_2) [\Delta w_{v2}(i, j) - g_1 \Delta y(i, j)] \\ \vdots \\ \Delta w_{vl}(i, j) = v_l^{-1}(z_2) [\Delta w_{v(l+1)}(i, j) - g_l \Delta y(i, j) - \psi(g_l) e_{g_l}(i, j)] \\ \vdots \\ \Delta w_{vN}(i, j) = v_N^{-1}(z_2) [-g_N \Delta y(i, j)] \end{array} \right. \quad (6.15)$$

Comparing (6.15) with (6.7) and noting that the outputs of the horizontal part, i.e., $\{w_{hk}(i, j)\}$, are all zeros, the transfer function from $-\psi(g_l)e_{g_l}(i, j)$ to $\Delta y(i, j)$ can be written as

$$H_{g_l}(z_2) = C_{\rho 2}(z_2 I_N - A_{\rho 4})^{-1} e_l \quad (6.16)$$

for $l = 1, 2, \dots, N$. Therefore, the roundoff noise gain due to g_l is given by

$$G_{g_l} = \psi(g_l) \|H_{g_l}(z_2)\|_2^2 = \psi(g_l) e_l^T W_{ov} e_l \triangleq \psi(g_l) G_l \quad (6.17)$$

where W_{ov} is the gramian of the pair $(A_{\rho 4}^T, C_{\rho 2}^T)$.

With similar procedure, one can show that the transfer function from $\psi(\beta_{00})e_{\beta_{00}}(i, j)$ to $\Delta y(i, j)$, denoted as $H_{\beta_{00}}(z_2)$, is

$$H_{\beta_{00}}(z_2) = 1 - C_{\rho 2}(z_2 I_N - A_{\rho 4})^{-1} \bar{g} \quad (6.18)$$

where $\bar{g} \triangleq [g_1 \ g_2 \ \dots \ g_N]^T$. Hence,

$$G_{\beta_{00}} = \psi(\beta_{00})(1 + \bar{g}^T W_{ov} \bar{g}) \triangleq \psi(\beta_{00}) G_0. \quad (6.19)$$

Next, from Figure 6.1, it is obvious that the input-output relationship between the rounding error produced by f_k to the corresponding output error is the same as that between $u(i, j)$ and $y(i, j)$. The roundoff noise gain due to f_k , is then

$$G_f = \psi(f_k) \|H(z_1, z_2)\|_2^2 = B_\rho^T W_o^\rho B_\rho + \beta_{00}^2. \quad (6.20)$$

For the roundoff noise by Δ_{hk} , the procedure to obtain its roundoff noise gain is similar but much lengthier than above, which is not presented in detail in this thesis. Denoting $G_{\Delta_{hk}}$ as the roundoff noise gain due to Δ_{hk} , it can be given by

$$\begin{aligned} G_{\Delta_{hk}} &= \psi(\Delta_{hk}) \|H_{\Delta_{hk}}(z_1, z_2)\|_2^2 = \psi(\Delta_{hk}) \|C_\rho(Z - A_\rho)^{-1} E_k\|_2^2 \\ &= \psi(\Delta_{hk}) E_k^T W_o^\rho B_k \end{aligned} \quad (6.21)$$

where $E_k \triangleq [e_k^T \ B_{\rho 2}^T]^T$.

Comparing the positions of Δ_{v1} and $\{\beta_{i0}, i = 1, 2, \dots, M\}$ with that of β_{00} , the positions of γ_{vl} , $\Delta_{v(l+1)}$, and $\{\beta_{il}, i = 0, 1, \dots, M\}$ with that of g_l , and the position of γ_{hk} with that of Δ_{hk} in Figures 6.1 and 6.2, one can literally obtain the total roundoff noise gain for the ZL structure, denoted as G_{ZL} , given by

$$G_{ZL} = \sum_{l=0}^N \xi_{vl} G_l + \sum_{k=1}^M \xi_{hk} G_{\Delta_{hk}} + \xi_f G_f \quad (6.22)$$

where the coefficients ξ_{vl} , ξ_{hk} , ξ_f can be specified as

$$\left\{ \begin{array}{l} \xi_{v0} = \sum_{i=0}^M \psi(\beta_{i0}) + \psi(\Delta_{v1}) \\ \xi_{vl} = \sum_{i=0}^M \psi(\beta_{il}) + \psi(g_l) + \psi(\gamma_{vl}) + \psi(\Delta_{v(l+1)}), \quad l = 1, 2, \dots, N-1 \\ \xi_{vN} = \sum_{i=0}^M \psi(\beta_{iN}) + \psi(g_N) + \psi(\gamma_{vN}) \\ \xi_{hk} = \psi(\gamma_{hk}) + \psi(\Delta_{hk}), \quad \xi_f = \sum_{i=1}^M \psi(f_i) \end{array} \right. \quad (6.23)$$

6.3.1.2 Equivalent state-space realization

When the filter is implemented with the state-space realization (6.8) - (6.9), with $A_\rho = \{a_{ij}\}$, $B_\rho = \{b_i\}$, $C_\rho = \{c_i\}$, it follows from (2.19) that the total

roundoff noise gain, denoted as G_{ss} , is

$$G_{ss} = \text{tr}(W_o^\rho Q_{ss}) + \zeta_0 \quad (6.24)$$

where $Q_{ss} = \text{diag}(\zeta_1, \dots, \zeta_k, \dots, \zeta_{M+N})$, with

$$\left\{ \begin{array}{l} \zeta_0 = \sum_{i=1}^{M+1} \psi(c_i) + \psi(\beta_{00}), \quad \zeta_1 = \sum_{i=1}^M \psi(a_{1i}) + \psi(b_1) \\ \zeta_{M+1} = \sum_{i=1}^{M+2} \psi(a_{(M+1)i}) + \psi(b_{M+1}) \\ \zeta_{M+N} = \sum_{i=1}^{M+1} \psi(a_{(M+N)i}) + \psi(a_{(M+N)(M+N)}) + \psi(b_{M+N}) \\ \zeta_k = \psi(a_{k(k-1)}) + \psi(a_{kk}), \quad k = 2, 3, \dots, M \\ \zeta_k = \sum_{i=1}^{M+1} \psi(a_{ki}) + \psi(a_{kk}) + \psi(a_{k(k+1)}) + \psi(b_k), \\ \quad k = M+2, M+3, \dots, M+N-1 \end{array} \right. \quad (6.25)$$

6.3.2 L_2 -sensitivity measure

In this subsection, we analyze the transfer function sensitivity with respect to those parameters of non-FWL format for the proposed structures.

6.3.2.1 ZL structure

Noting that $\{\gamma_{hk}\}, \{\gamma_{vl}\}$ have no contribution to the overall sensitivity measure and denoting

$$S_p \triangleq \{f_k, g_l, \beta_{kl}, \Delta_{hk}, \Delta_{vl}, \forall k, l\}$$

it follows from (2.6) that the L_2 -norm based sensitivity measure, denoted as M_{ZL} , is given by

$$M_{ZL} = \sum_{\forall \tau_i, \tau_i \in S_p \text{ and } \tau_i \in S_{FWL}} \left\| \frac{\partial H(z_1, z_2)}{\partial \tau_i} \right\|_2^2. \quad (6.26)$$

With some manipulations, it follows from (6.9) and (6.10) that

$$\left\{ \begin{array}{l} \frac{\partial H}{\partial f_i} = [-\beta_{00} + C_\rho(Z - A_\rho)^{-1}(-B_\rho)][e_i^T(Z - A_\rho)^{-1}B_\rho] \\ \frac{\partial H}{\partial g_j} = [-\beta_{00} + (-C_\rho)(Z - A_\rho)^{-1}B_\rho][C_\rho(Z - A_\rho)^{-1}e_{M+j}] \\ \frac{\partial H}{\partial \beta_{i0}} = [1 + C_\rho(Z - A_\rho)^{-1}B_g][e_i^T(Z - A_\rho)^{-1}B_\rho] \\ \frac{\partial H}{\partial \beta_{0j}} = [1 + C_f(Z - A_\rho)^{-1}B_\rho][C_\rho(Z - A_\rho)^{-1}e_{M+j}] \\ \frac{\partial H}{\partial \beta_{ij}} = [C_\rho(Z - A_\rho)^{-1}e_{M+j}][e_i^T(Z - A_\rho)^{-1}B_\rho] \\ \frac{\partial H}{\partial \beta_{00}} = [1 + C_\rho(Z - A_\rho)^{-1}B_g][1 + C_f(Z - A_\rho)^{-1}B_\rho] \\ \frac{\partial H}{\partial \Delta_{hi}} = [e_i^T e_1 + e_i^T A_{\Delta_h}(Z - A_\rho)^{-1}B_\rho][C_\rho(Z - A_\rho)^{-1}e_i] \\ \frac{\partial H}{\partial \Delta_{vj}} = [e_{M+1}^T e_{M+j} + C_\rho(Z - A_\rho)^{-1}A_{\Delta_v}e_{M+j}][e_{M+j}^T(Z - A_\rho)^{-1}B_\rho] \end{array} \right. \quad (6.27)$$

for that $i = 1, 2, \dots, M$, $j = 1, 2, \dots, N$, and e_k is the k th elementary (column) vector of length $M + N$, and

$$C_f \triangleq \begin{bmatrix} -f_1 & \cdots & -f_M & 0 & \cdots & 0 \end{bmatrix}, \quad B_g \triangleq \begin{bmatrix} 0 & \cdots & 0 & -g_1 & \cdots & -g_N \end{bmatrix}^T$$

$$A_{\Delta_h} \triangleq \begin{bmatrix} -f_1 & -f_2 & \cdots & -f_{M-1} & -f_M \\ 1 & 0 & \cdots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad A_{\Delta_v} \triangleq \begin{bmatrix} -g_1 & 1 & 0 & \cdots & 0 \\ -g_2 & 0 & 1 & \cdots & 0 \\ & & \ddots & & \\ -g_{N-1} & 0 & 0 & \cdots & 1 \\ -g_N & 0 & 0 & \cdots & 0 \end{bmatrix}$$

6.3.2.2 Equivalent state-space realization

For the equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_{00})$ by (6.9) - (6.10), the corresponding L_2 -norm based sensitivity measure, denoted as M_{ss} , can be evaluated directly with (2.34).

6.3.3 Structure optimization

By now, we have known that the FWL performance of the proposed structures is dependent on the choice of the free parameters $\{\gamma_{hk}\}$, $\{\gamma_{vl}\}$. Denote

$$\bar{\gamma} \triangleq \begin{bmatrix} \gamma_{h1} & \gamma_{h2} & \cdots & \gamma_{hM} & \gamma_{v1} & \gamma_{v2} & \cdots & \gamma_{vN} \end{bmatrix}. \quad (6.28)$$

The problem below is to identify those sets of $\bar{\gamma}$ which can minimize G_{ZL} and G_{ss} , leading to the optimal ZL structure and its state-space realization problems in terms of roundoff noise gain:

$$\min_{\bar{\gamma}(i) \in S_{\gamma}, \forall i} G_{ZL} \Rightarrow \bar{\gamma}(G_{ZL}^{\text{opt}}), \quad \min_{\bar{\gamma}(i) \in S_{\gamma}, \forall i} G_{ss} \Rightarrow \bar{\gamma}(G_{ss}^{\text{opt}}) \quad (6.29)$$

where $\bar{\gamma}(G_{ZL}^{\text{opt}})$ and $\bar{\gamma}(G_{ss}^{\text{opt}})$ are the optimal solution that can minimize G_{ZL} and G_{ss} , respectively, and S_{γ} is defined in (3.14). Similarly, we can formulate the optimal structure problems in terms of L_2 -sensitivity:

$$\min_{\bar{\gamma}(i) \in S_{\gamma}, \forall i} M_{ZL} \Rightarrow \bar{\gamma}(M_{ZL}^{\text{opt}}), \quad \min_{\bar{\gamma}(i) \in S_{\gamma}, \forall i} M_{ss} \Rightarrow \bar{\gamma}(M_{ss}^{\text{opt}}). \quad (6.30)$$

Now, let us consider further the expression of G_{ZL} by (6.22). Noting that G_l is computed by $H_{g_l}(z_2)$ or $H_{\beta_{00}}(z_2)$, which can be written in the format of

$$\begin{aligned} H_{g_l}(z_2) &= \frac{q_l(z_2)}{G(z_2)} = \frac{\kappa_2 q_l(z_2)}{D(z_2)}, \quad l = 1, 2, \dots, N \\ H_{\beta_{00}}(z_2) &= \frac{q_0(z_2)}{G(z_2)} = \frac{\kappa_2 q_0(z_2)}{D(z_2)} \end{aligned}$$

where $D(z_2) \triangleq V_{a2}^T \bar{z}_2 = \prod_{j=1}^N (z_2 - \mu_j)$ and $\kappa_2 = \prod_{j=1}^N \Delta_{vj}$ given before. With $q_l(z_2)$ defined in (6.2), it follows from the same analysis in Section 3.4 that G_l ,

and hence G_{ZL} , can be much reduced if $\{\gamma_{vl}\}$ are chosen such that they are very close to the real part of $\{\mu_j\}$. This suggests that $\{\gamma_{vl}\}$ can be searched within the neighborhood of those poles.

6.4 ZL structure for separable-numerator digital filters

By now, we have derived the ZL structure for the separable-denominator digital filters given by (2.28) and analyzed its FWL performance. Although the case of a transfer function with separable-numerator, say (2.27), is not considered as much as that with separable-denominator, it is still under concern of some researchers [53]. For the completeness of our research work, we develop the ZL structure for the separable-numerator filters in this section. Since its deriving procedure and performance analysis are very similar to the separable-denominator case, we only present the final results.

With $\bar{p}(z_1)$, $\bar{q}(z_2)$ denoted in (6.2), (2.27) can be written as

$$H(z_1, z_2) = \frac{(V_{b1}\bar{z}_1)(V_{b2}\bar{z}_2)}{\bar{z}_1^T M_a \bar{z}_2} \triangleq \frac{[V_f^T \bar{p}(z_1)][V_g^T \bar{q}(z_2)]}{\bar{p}(z_1) M_\alpha \bar{q}(z_2)} \quad (6.31)$$

where $V_{b1} = \{b_{1k}\} \in \mathcal{R}^{(M+1) \times 1}$, $V_{b2} = \{b_{2l}\} \in \mathcal{R}^{(N+1) \times 1}$, $M_a = \{a_{kl}\} \in \mathcal{R}^{(M+1) \times (N+1)}$, and

$$\begin{aligned} V_f &\triangleq \kappa_1^{-1} \bar{T}_1^{-T} V_{b1}, \quad V_g \triangleq \kappa_2^{-1} \bar{T}_2^{-T} V_{b2} \\ M_\alpha &\triangleq \kappa_1^{-1} \bar{T}_1^{-T} M_a \bar{T}_2^{-1} \kappa_2^{-1} \end{aligned}$$

with $\kappa_1 = \prod_{k=1}^M \Delta_{hk}$, $\kappa_2 = \prod_{l=1}^N \Delta_{vl}$ such that $M_\alpha(1, 1) = 1$.

It is shown that the re-parameterized transfer function (6.31) can be realized with the ZL structure given by Figure 6.3, where the outputs of operators $h_k^{-1}(z_1)$ and $v_l^{-1}(z_2)$ are still computed with the realizations in Figure 6.2.

The equivalent state-space realization of the ZL structure for the separable-numerator case is given by (6.8), where

$$\left\{ \begin{array}{l} A_{\rho 1} = \begin{bmatrix} \gamma_{h1} - \alpha_{10}\Delta_{h1} & -\alpha_{20}\Delta_{h1} & \cdots & -\alpha_{M0}\Delta_{h1} \\ \Delta_{h2} & \gamma_{h2} & & \\ & \ddots & \ddots & \\ & & \Delta_{hM} & \gamma_{hM} \end{bmatrix} \\ A_{\rho 4} = \begin{bmatrix} \gamma_{v1} - \alpha_{01}\Delta_{v1} & \Delta_{v2} \\ -\alpha_{02}\Delta_{v1} & \gamma_{v2} & \ddots \\ \vdots & & \ddots & \Delta_{vN} \\ -\alpha_{0N}\Delta_{v1} & & & \gamma_{vN} \end{bmatrix} \\ B_{\rho 1} = \begin{bmatrix} g_0\Delta_{h1} & 0 & \cdots & 0 \end{bmatrix}^T, \quad B_{\rho 2} = \begin{bmatrix} \tilde{g}_1 & \tilde{g}_2 & \cdots & \tilde{g}_N \end{bmatrix}^T \\ C_{\rho 1} = \begin{bmatrix} \tilde{f}_1 & \tilde{f}_2 & \cdots & \tilde{f}_M \end{bmatrix}, \quad C_{\rho 2} = \begin{bmatrix} f_0\Delta_{v1} & 0 & \cdots & 0 \end{bmatrix}, \quad d = f_0g_0 \end{array} \right. \quad (6.32)$$

and $A_{\rho 3} = \{\tilde{\alpha}_{lk}\} \in \mathcal{R}^{N \times M}$, with $\tilde{f}_k = f_k - f_0\alpha_{k0}$, $\tilde{g}_l = g_l - g_0\alpha_{0l}$, $\tilde{\alpha}_{lk} = \alpha_{kl} - \alpha_{k0}\alpha_{0l}$, and all the other elements in A_{ρ} are zero.

Based on the performance analysis in Section 6.3, one can easily derive the expressions of roundoff noise gain and L_2 -sensitivity measure for the ZL structure for the separable-numerator case and its equivalent state-space realization.

Figure 6.3: 2-D ZL structure for separable-numerator digital filters

6.5 An improved ZL structure

In the case of 1-D digital filter design, based on the polynomial operator concept [16], several efficient structures, ρ DFIIt and ρ IDFIIt, and their related realizations, were proposed in previous chapters. We note that the polynomials used in the ρ DFIIt structure have only real roots, while, with more degrees of freedom introduced in the ρ IDFIIt, the corresponding polynomials can have complex roots. It is found that when the polynomial operators have their roots closer to the filter poles, which are generally complex, the roundoff noise gain can be much reduced. In this chapter, starting with the 2-D separable-denominator filter (2.28) and applying the polynomial operators used in ρ DFIIt to both horizontal part and vertical part, the 2-D ZL structure was derived. From the expressions of G_{ZL} in (6.22) - (6.23), one finds that the roundoff noise gain of the ZL structure is strongly related to the polynomials for the vertical part. To that effect, we apply the more robust polynomial operators used in ρ IDFIIt to the vertical part of 2-D separable-denominator filters and an alternative structure can be obtained. For the separable-numerator case, similar structure can also be developed, which is not presented in this thesis.

6.5.1 Two new structures

Defining $s \triangleq N/2$ if N is even, $s \triangleq (N+1)/2$ if N is odd and noting $\{v_l(z_2)\}$ denoted in (6.1), consider the following polynomial operators for the vertical part:

$$\begin{cases} q_N(z_2) = 1, & q_{N-1}(z_2) = v_N(z_2) \\ q_{N-2n}(z_2) = \bar{v}_{s-n+1}(z_2)v_{N-2n+2}(z_2) \\ q_{N-2n-1}(z_2) = v_{N-2n}(z_2)q_{N-2n}(z_2) \end{cases} \quad (6.33)$$

for all $n = 1, 2, \dots, s$, where

$$\begin{cases} \bar{v}_{s-n+1}(z_2) \triangleq v_{N-2n+1}(z_2)v_{N-2n+2}(z_2) + \eta_{s-n+1} \\ \eta_{s-n+1} \triangleq \frac{\bar{\eta}_{s-n+1}}{\Delta_{v(N-2n+1)}\Delta_{v(N-2n+2)}} \end{cases} \quad (6.34)$$

with $\{\bar{\eta}_n\}$ a set of free parameters. When N is odd, $v_0(z_2) \triangleq 1$, $\bar{\eta}_1 = 0$ are assumed, which lead to $\bar{v}_1(z_2) = v_1(z_2)$ and hence $q_0(z_2) = \bar{v}_1(z_2)q_1(z_2)$.

One then has $\bar{p}(z_1)$, $\bar{q}(z_2)$ defined by (6.3), with $\{p_k(z_1)\}$ obtained from (6.2) and $\{q_l(z_2)\}$ obtained from (6.33). The transfer function (2.28) therefore can be re-parameterized like in (6.4), which can be implemented with the structure shown in Figure 6.4¹, with $h_k^{-1}(z_1)$ and $v_l^{-1}(z_2)$ also realized in Figure 6.2. One can observe that the structure by Figure 6.4 is different from the 2-D ZL structure by Figure 6.1 only in its vertical part and hence is called an improved ZL structure, denoted as IZL for convenience, which introduces more degrees of freedom than the ZL by a set of free parameters $\{\eta_n\}$.

The equivalent state-space realization $(A_\rho, B_\rho, C_\rho, \beta_{00})$ of the IZL structure can be expressed by (6.8) - (6.9), but with $A_{\rho 4}$ replaced by²

$$A_{\rho 4} = \begin{bmatrix} \gamma_{v1} - g_1\Delta_{v1} & \Delta_{v2} & 0 & \cdots & 0 & 0 & 0 \\ -\eta_1\Delta_{v1} - g_2\Delta_{v2} & \gamma_{v2} & \Delta_{v3} & \cdots & 0 & 0 & 0 \\ -g_3\Delta_{v1} & 0 & \gamma_{v3} & \cdots & 0 & 0 & 0 \\ -g_4\Delta_{v1} & \eta_2\Delta_{v3} & 0 & \cdots & 0 & 0 & 0 \\ & & \vdots & & & & \\ -g_{N-1}\Delta_{v1} & 0 & 0 & \cdots & 0 & \gamma_{v(N-1)} & \Delta_{vN} \\ -g_N\Delta_{v1} & 0 & 0 & \cdots & 0 & \eta_s\Delta_{v(N-1)} & \gamma_{vN} \end{bmatrix} \quad (6.35)$$

¹The structure shown here is for N is even.

² $A_{\rho 4}$ presented here is only for N is even.

Figure 6.4: Improved 2-D ZL structure

and the coupling coefficients $\{\Delta_{hk}\}$, $\{\Delta_{vl}\}$ are also computed by (6.13).

6.5.2 Performance analysis

Firstly, let us consider the roundoff noise performance of the IZL structure and its equivalent state-space realization. For the IZL structure, comparing the position of η_n with that of $g_{N-2s+2n}$ in Figure 6.4 (the same to the case that N is odd), one can easily obtain the total roundoff noise gain for the IZL structure, denoted as G_{IZL} , given by

$$G_{IZL} = G_{ZL} + \sum_{n=1}^s \psi(\eta_n) G_{N-2s+2n} \quad (6.36)$$

with G_{ZL} computed by (6.22) - (6.23).

For the state-space realization, noting the changes in $A_{\rho 4}$, the roundoff noise gain, denoted as G_{Iss} , can be expressed by

$$G_{Iss} = G_{ss} + \text{tr}(W_o^{\rho} Q_{\eta}) \quad (6.37)$$

where G_{ss} is computed by (6.24) - (6.25) and $Q_{\eta} = \text{diag}(\varsigma_1, \dots, \varsigma_k, \dots, \varsigma_M)$, with

$$\varsigma_k = \begin{cases} \psi(a_{k(k-1)}), & k = M + N - 2n \text{ with } n = 0, 1, \dots, s - 2 \\ 0, & \text{otherwise} \end{cases} \quad (6.38)$$

The L_2 -sensitivity measures of both structures, denoted as M_{IZL} and M_{Iss} ,

can be obtained by (6.26) and (2.34), respectively, with denoting

$$S_p \triangleq \{f_k, g_l, \beta_{kl}, \Delta_{hk}, \Delta_{vl}, \eta_n, \forall k, l, n\}$$

for the IZL structure and its corresponding state-space realization.

By now, we have known the FWL performance of the improved ZL structure and its equivalent state-space realization. It is easy to understand that the roundoff noise performance of the IZL structure is better than the corresponding ZL just because its vertical polynomials $\{q_l(z_2)\}$ can have complex roots, and when these roots are distributed around the poles of the filter's vertical part, G_{IZL} is reduced. Hence, the optimal structure problems is to find appropriate free parameters $\{\gamma_{hk}\}$, $\{\gamma_{vl}\}$, and $\{\bar{\eta}_n\}$ to minimize the FWL performance of the proposed structures.

6.6 Numerical Examples

In this section, we examine the performance of the proposed structures. For convenience, denote $\bar{\gamma}_z$ and $\bar{\gamma}_\delta$ for $\bar{\gamma}(k) = 0, \forall k$ and $\bar{\gamma}(k) = 1, \forall k$, respectively. The optimized proposed structures are compared with the ZL structures in shift-operators (corresponding to $\bar{\gamma}_z$) and delta-operators (corresponding to $\bar{\gamma}_\delta$), as well as the optimal fully parameterized state-space realization R_G [55]. In both examples, $B_\gamma = 4$ bits is assumed.

Example 6.1: Consider a 2-D low-pass narrow band filter with separable-denominator, whose transfer function is given by (6.4) with

$$\begin{aligned} V_{a1} &= \begin{bmatrix} 1 & -2.173645 & 1.836929 & -0.599655 \end{bmatrix}^T \\ V_{a2} &= \begin{bmatrix} 1 & -2.280029 & 1.887939 & -0.564961 \end{bmatrix}^T \\ M_b &= \begin{bmatrix} 0.019421 & -0.027724 & 0.011468 & -0.000086 \\ 0.004839 & 0.017545 & -0.050267 & 0.033061 \\ -0.004328 & -0.008847 & 0.096260 & -0.083801 \\ -0.000138 & 0.007979 & -0.052927 & 0.062607 \end{bmatrix} \end{aligned}$$

which is obtained from the initial state-space realization R_0 in [64]. The corresponding poles for the horizontal part are $\{0.6707 \pm j0.5203, 0.8322\}$, while $\{0.7343 \pm j0.3963, 0.8113\}$ for the vertical part.

For this example, R_G yields a roundoff noise gain of 62.5557, and needs 49 nontrivial parameters. With the GA program, the corresponding solutions to (6.29) are

$$\begin{aligned} \bar{\gamma}(G_{ZL}^{\text{opt}}) &= \begin{bmatrix} 0 & 0 & 0.6875 & 0.875 & 0.75 & 0.75 \end{bmatrix} \\ \bar{\gamma}(G_{ss}^{\text{opt}}) &= \begin{bmatrix} 0.6875 & 0.75 & 0.6875 & 0.6875 & 0.75 & 0.75 \end{bmatrix} \end{aligned}$$

and the corresponding solutions to (6.30) are

$$\begin{aligned} \bar{\gamma}(M_{ZL}^{\text{opt}}) &= \begin{bmatrix} 0.625 & 0.6875 & 0.8125 & 0.75 & 0.75 & 0.75 \end{bmatrix} \\ \bar{\gamma}(M_{ss}^{\text{opt}}) &= \begin{bmatrix} 1 & 0.75 & 0.8125 & 0.875 & 0.75 & 0.75 \end{bmatrix} \end{aligned}$$

Table 6.1 provides the statistics of G_{ZL} , M_{ZL} , N_{ZL} (the number of nontrivial parameters) for the ZL structure, and similarly, G_{ss} , M_{ss} , N_{ss} for the state-space

Table 6.1: Comparison Results for Example 6.1

$\bar{\gamma}$	G_{ZL}	G_{ss}	M_{ZL}	M_{ss}	N_{ZL}	N_{ss}
$\bar{\gamma}_z$	1.7503×10^3	1.3740×10^3	3.1400×10^4	2.2132×10^3	28	28
$\bar{\gamma}_\delta$	1.1989×10^2	7.7880×10^1	1.3306×10^2	1.2387×10^2	28	28
$\bar{\gamma}(G_{ZL}^{opt})$	6.2756×10	6.4505×10^1	2.4323×10^2	1.3783×10^2	33	31
$\bar{\gamma}(G_{ss}^{opt})$	7.5871×10^1	<u>4.6621×10</u>	7.0016×10^1	7.7830×10^1	34	32
$\bar{\gamma}(M_{ZL}^{opt})$	7.8000×10^1	4.7275×10^1	<u>6.8643×10</u>	7.5356×10^1	34	32
$\bar{\gamma}(M_{ss}^{opt})$	7.8830×10^1	4.7440×10^1	7.0866×10^1	<u>7.4877×10</u>	33	32

realization with different $\bar{\gamma}$.

For the roundoff noise gain, one can observe that

- It confirms the argument in Section 6.3.3 that the roundoff noise gain can be much reduced if $\{\gamma_{vl}\}$ are chosen to be close to the poles of the vertical part of the filter.
- The structures in delta-operators have much better performance than those in shift-operators, while this is no longer true for the low-pass filters without narrow bandwidth whose poles are far away from $z = +1$ (see Example 6.2).
- As we can see, both of the proposed structures are more efficient than R_G . The optimized ZL structure yields almost the same performance as R_G and the optimized state-space realization even over-performs R_G in terms of roundoff noise gain.

And for the L_2 -sensitivity measure,

- The L_2 -sensitivity measure for R_0 is

$$M_1 = 2.2423893 \times 10^4.$$

Next, for the balanced realization [64], which can minimize FWL effects effectively, the L_2 -sensitivity measure can be obtained as

$$M_2 = 101.304839.$$

Then, utilizing the iterative procedure proposed in [64] to synthesize the optimal filter structure that minimize the L_2 -sensitivity, the final result is computed as

$$M_3 = 91.336902.$$

- All the minimizations above of the L_2 -sensitivity do not consider the dynamic-range of the state-variables. From Table 6.1, it can be seen that the optimized proposed structures yield smaller sensitivity measure than the structures proposed by former researchers, and more importantly, they are very efficient and subject to the l_2 -scaling constraints.

Figure 6.5 shows the ideal magnitude response of the filter prototype given in Example 6.1, and the ZL structures with shift-operators $\bar{\gamma}_z$ as well as the optimal polynomial operators $\bar{\gamma}(M_{ZL}^{\text{opt}})$, respectively, where the coefficients are implemented with $B_c = 8$ bits for the fractional parts.

Example 6.2: This is a (4, 4)th order low-pass filter with quadrantal symmetry, whose normalized bandwidth is $BW = 0.3/2 = 0.15$. The transfer function is

Figure 6.5: Magnitude responses for Example 6.1: (a) the ideal filter; (b) the ZL structure with $\bar{\gamma}_z$; (c) the ZL structure with $\bar{\gamma}(M_{ZL}^{\text{opt}})$

Table 6.2: Comparison Results for Example 6.2

$\bar{\gamma}$	G_{ZL}	G_{ss}	M_{ZL}	M_{ss}	N_{ZL}	N_{ss}
$\bar{\gamma}_z$	4.0743×10^1	1.5985×10^1	8.5274×10^1	2.1612×10^1	41	41
$\bar{\gamma}_\delta$	1.0059×10^2	3.3301×10^1	1.7236×10^2	3.4675×10^1	41	41
$\bar{\gamma}(G_{ZL}^{\text{opt}})$	1.1118×10	1.1066×10^1	2.5135×10^1	1.3938×10^1	45	44
$\bar{\gamma}(G_{ss}^{\text{opt}})$	1.3695×10^1	8.9900×10^0	1.2169×10^1	9.9091×10^0	49	47

given by (6.4) with

$$V_{a1} = V_{a2} = \begin{bmatrix} 1 & -1.5704 & 1.2756 & -0.4844 & 0.0762 \end{bmatrix}^T$$

$$M_b = \begin{bmatrix} 0.0003 & 0.0014 & 0.0021 & 0.0014 & 0.0003 \\ 0.0014 & 0.0055 & 0.0083 & 0.0055 & 0.0014 \\ 0.0021 & 0.0083 & 0.0124 & 0.0083 & 0.0021 \\ 0.0014 & 0.0055 & 0.0083 & 0.0055 & 0.0014 \\ 0.0003 & 0.0014 & 0.0021 & 0.0014 & 0.0003 \end{bmatrix}$$

The corresponding poles for either the horizontal or vertical part are $\{0.4488 \pm j0.5707, 0.3364 \pm j0.1772\}$.

For this filter, R_G yields a roundoff noise gain of 12.0369 and needs 81 nontrivial parameters. Table 6.2 provides the corresponding statistics. One notes that for this example, the poles of the filter are far away from $z = +1$ and the structures in delta-operators are no longer better than the ones in shift-operators. This is coincident with the theoretical results in Section 6.3.3. With the GA program, it shows that

$$\bar{\gamma}(G_{ZL}^{\text{opt}}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.375 & 0.375 & 0.375 & 0.375 \end{bmatrix}$$

$$\bar{\gamma}(G_{ss}^{\text{opt}}) = \begin{bmatrix} 0.9375 & 0.4375 & 0.25 & 0.4375 & 0.4375 & 0.25 & 0.25 & 0.4375 \end{bmatrix}$$

both of which yield a better performance and much efficient than R_G .

Example 6.3: Consider a (4, 4)th order low-pass filter with quadrantal symmetry, whose normalized bandwidth is $BW = 0.2/2 = 0.1$. The transfer function is given by (6.4) with

$$V_{a1} = V_{a2} = \begin{bmatrix} 1 & -2.3695 & 2.3140 & -1.0547 & 0.1874 \end{bmatrix}^T$$

$$M_b = 10^{-3} \times \begin{bmatrix} 0.0233 & 0.0931 & 0.1396 & 0.0931 & 0.0233 \\ 0.0931 & 0.3724 & 0.5586 & 0.3724 & 0.0931 \\ 0.1396 & 0.5586 & 0.8379 & 0.5586 & 0.1396 \\ 0.0931 & 0.3724 & 0.5586 & 0.3724 & 0.0931 \\ 0.0233 & 0.0931 & 0.1396 & 0.0931 & 0.0233 \end{bmatrix}$$

For this filter, R_G yields a roundoff noise gain of 11.0381 and needs 81 nontrivial parameters. The GA programs show that for the optimized IZL structure, the free parameters are:

$$\bar{\gamma} = \begin{bmatrix} 0.75 & 0.3125 & 0.3125 & 0.4375 & 0.625 & 0.5 & 0.5625 & 0.6875 \end{bmatrix}$$

$$\{\bar{\eta}_m\} = \{ 0.105983 \quad 0.140757 \}$$

And for the optimized state-space realization of the IZL structure,

$$\bar{\gamma} = \begin{bmatrix} 0.8125 & 0.625 & 0.5625 & 0.625 & 0.5 & 0.5 & 0.5625 & 0.6875 \end{bmatrix}$$

$$\{\bar{\eta}_m\} = \{ 0.311995 \quad 0.140366 \}$$

Table 6.3: Comparison Results for Example 6.3

Realization	ZL	ZL _{ss}	IZL	IZL _{ss}
G	9.7085	7.7095	8.6870	7.5335
M	21.6292	8.8363	9.5216	8.4969

Similarly, for the optimized ZL structure,

$$\bar{\gamma} = \begin{bmatrix} 0 & 0.375 & 0 & 0.5625 & 0.5625 & 0.625 & 0.5625 & 0.5625 \end{bmatrix}$$

and for the optimized state-space realization of the ZL structure,

$$\bar{\gamma} = \begin{bmatrix} 0.3125 & 0.625 & 0.5625 & 0.5625 & 0.5625 & 0.5625 & 0.5 & 0.5625 \end{bmatrix}$$

Table 6.3 presents the roundoff noise gain G and L_2 -sensitivity M for the ZL structure and its equivalent state-space realization (called ZL_{ss} for convenience), the IZL structure and its equivalent state-space realization (called IZL_{ss}). We can see the performance improvement of the IZL and IZL_{ss} structures. It should be pointed out that this improvement will be more obvious with the filter order increasing. Figure 6.6 shows the magnitude responses for the ideal filter, the ZL structures with shift-operators $\bar{\gamma}_z$, the optimized ZL structure and the optimized IZL structures, where the coefficients are implemented with $B_c = 8$ bits for the fractional parts. The results are self-explanatory.

Figure 6.6: Magnitude responses for Example 6.3: (a) the ideal filter; (b) the ZL structure with shift-operators; (c) the optimized ZL structure; (d) the optimized IZL structure

6.7 Conclusion

In this section, the minimization problems of FWL effects for 2-D digital filters have been discussed. Firstly, the ZL structure for the separable-denominator digital filters has been derived based on a set of special polynomial operators. The equivalent state-space realization of this sparse structure is also obtained, which is of the same structure simplicity. The performance of the two proposed structures have been analyzed by deriving the expressions for the roundoff noise gain and L_2 -sensitivity. The optimal structure problems have been formulated and solved with the genetic algorithms. Secondly, the ZL structure for the separable-numerator filters and its equivalent state-space realization are presented briefly, since they are very similar to the case of separable-denominator. Thirdly, based on the obtained relationship between the free parameters for the vertical part and the filter vertical poles in terms of roundoff noise gain for the ZL structure, a more robust structure, named the IZL, is developed. The vertical polynomials of this structure can have complex roots, and when those roots are close to the filter vertical poles which are generally complex, the roundoff noise gain for the IZL structure and the state-space realization (IZL_{ss}) can be lower than its corresponding ZL structure and the equivalent state-space realization (ZL_{ss}).

Chapter 7

Conclusions and Recommendations

7.1 Conclusions

This thesis deals with the problem of finite precision for digital signal processing. With the objective to explore efficient digital filter structures that can minimize the FWL effects, several sparse structures have been derived and discussed for both one-dimensional and two-dimensional fields.

✓ Based on the concept of polynomial operators a generalized direct-form II transposed structure, call ρ DFIIt, has been developed. One of the nice properties of this sparse structure is that it provides more degrees of freedom to allow us to minimize the FWL effects for any types of digital filters. The performance of this proposed structure and its equivalent state-space realization have been analyzed by deriving the expressions for the roundoff noise gain and L_2 -sensitivity measure. After that the problem of the optimal ρ DFIIt polynomial

operators has been investigated. The expression for the roundoff noise gain of the ρ DFIIt structure has been derived in terms of the poles of the filter and the free parameters, which implies that the optimal free parameters can be searched around the dominant poles. It is shown with many design examples that the two proposed structures not only outperform the traditional z DFIIt and the prevailing δ DFIIt but are also very competitive to the fully parameterized optimal roundoff noise realizations. In fact, the optimized equivalent state-space realization can beat the latter in both examples.

✓ The ρ DFIIt structure and its revised version, the ρ RDFIIt, are investigated in terms of RBM implementation scheme. The performance has been analyzed for each of these two structures by deriving the expressions of the roundoff noise gain. It is shown that the ρ RDFIIt structure can always yield a smaller roundoff noise gain than the corresponding ρ DFIIt, while having the same structure simplicity. With the technique of error feedback taken into consideration, the roundoff noise gain for the ρ RDFIIt structure can be further reduced.

✓ It has been noted that the roundoff noise gain for the ρ DFIIt structure can be reduced when the roots of the corresponding polynomials are close to the dominant poles of the filter. This observation gives insights for exploring more robust structures against FWL effects via polynomial operators concept. Hence, the ρ IDFIIt structure and its equivalent state-space realization have been exploited, which have more degrees of freedom than the proposed ρ DFIIt. Assuming the RAM scheme, the performance of each of the structures has been analyzed by deriving the expressions for roundoff noise gain and L_2 -sensitivity measure. It is shown that the equivalent state-space realization always yields a

smaller roundoff noise gain than its DFII_t counterpart. Moreover, this equivalent state-space realization is discussed in terms of RBM scheme without/with error feedback. The approach for determining the optimal full-scale and diagonal error-feedback matrices are explained. Extensive examples have been carried out to illustrate the design procedure and to compare the performance of the proposed structures with some existing structures.

✓ This research work in the case of 1-D digital filter structure design has been extended to the 2-D field. Two 2-D efficient structures, both of which are a mixture of DFII for their horizontal part and DFII_t for their vertical part, and their corresponding state-space realizations have been derived. The FWL performance of the proposed structures has been analyzed and the optimal structure problems have been formulated. Design examples have been given, which illustrate the advantages of the proposed structures over some existing structures.

7.2 Recommendations for future research work

This thesis has derived several efficient digital filter structures with consideration of finite precision implementation. The research and development in this area has been and will continue to be one of the most important branches for digital signal processing. And so far there are still many unsolved problems in this topic which need further investigation, and much more work should be done to develop better structures or algorithms to deal with the FWL effects. Based on the present work, future research can be recommended as follows.

- The proposed structures in this thesis, either in the case of 1-D digital

filters or 2-D field, are mainly based on direct-forms and their equivalent state-space realizations. It is well known that cascade form of digital filters (in the shift operators) is considered as one of the most efficient ways to minimize the FWL effects. It would be interesting to study this kind of structures with the shift operators replaced by the proposed first-order ρ -operators. Especially, if the transfer function is factored and expressed as the product of second-order sections, it would be worth investigating the structure's numerical properties when applying the second-order operators used in the ρ IDFIIt to each section of the cascade form.

- The design of optimum filters to perform signal estimation is a problem that frequently arises in the design of communication systems, control systems, in geophysics, and in many other applications and disciplines. In [104], the issue of numerical accuracy was addressed in parameter estimation and filtering problems, and it was showed that better accuracy can be obtained if a δ -operator parametrization was used instead of the usual shift operator parametrization. Based on this observation, further research can be conducted on the optimal filters in this area (estimation) with a ρ -operator parametrization.

- In Section 2.2, it was pointed out that limit cycles are not investigated in this thesis. Hence, optimizing the proposed structures and developing more robust structures, against limit cycles could be a very meaningful topic in the next stage.

- The extended state-space (e-state) model proposed in [105], compared to the classical state-space realization, requires less storage and less computations. However, the reduction in computational requirements for e-state model may be associated with an increase of product roundoff noise in the output. Considering

that the delay replacement technique proposed in this work is very effective in reducing the roundoff noise gain, it is a potential study direction to obtain a more general state-space realization combining these ρ -operators with the e-state model.

- GAs have been applied to the design issues regarding both digital and analogue filters in the past few years and are used to solve the optimization problems for the free parameters in our research work. This indicates the success of the GAs as powerful tools in digital filter design, however, the full power of the GAs in this area has not yet been explored. Hence, it is also an interesting and important future research direction to apply more efficient genetic algorithms to solve the similar optimization problems in the area of digital filter design.

- Though this thesis is about a fundamental research, the results have a lot of applications to commercial products. This is particularly true for real-time applications, where the cost and speed of processing are the essential requirements. For example, frequency equalizers, which consist of 10 - 30 filter banks, are a crucial part in many systems such as Hi-Fi instruments (for music) and media players. How to reduce the FWL effects as well as lower the cost is a big block for the products to have a place in the market. That may be achieved with the Hi-Fi implemented with an efficient structure. Two-dimensional digital filters find a lot of applications in many different areas, ranging from digital processing of satellite photographs, radar maps, and medical x-ray images to planning the power systems, so it is illuminating to find a practical environment in which the newly developed 2-D structures have light use.

Author's Publications

Published and accepted:

1. Zixue Zhao and Gang Li, "Roundoff noise analysis of two efficient digital filter structures," *IEEE Transactions on Signal Processing*, vol. 54, pp. 790-795, Feb. 2006.
2. Zixue Zhao, "Two efficient structures for 2-D digital filter implementation," *IEE Proceedings of Circuits, Devices, and Systems*, vol. 152, pp. 641-648, Dec. 2005.
3. Gang Li and Zixue Zhao, "On the generalized DFII structure and its state-space realization in digital filter implementation," *IEEE Transactions on Circuits and System I*, vol. 51, pp. 769-778, Apr. 2004.
4. Zixue Zhao and Gang Li, "Efficient structure for 2-D digital filters with minimum roundoff noise gain," *IEE Electronics Letter*, vol. 39, pp. 1424-1425, Sep. 2003.
5. Zixue Zhao, Gang Li, and Jiong Zhou, "Efficient digital filter structures with minimum roundoff noise gain," in *Proceedings of IEEE International Symposium on Circuits and System*, pp. 2587-2590, May 2005.
6. Zixue Zhao, Gang Li, and Jiong Zhou, "Roundoff noise analysis of two efficient digital filter structures," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 333-336, Mar. 2005.
7. Zixue Zhao and Gang Li, "An DFII based structure for 2-D separable-denominator digital filters with very low L_2 -sensitivity measure," in *Proceedings of IEEE International Symposium on Circuits and System*, vol. 3, pp. 485-488, May 2004.

8. Zixue Zhao and Gang Li, "DFII based structures for 2-D separable-numerator digital filters with minimal roundoff noise gain," in *Proceedings of International Conference on Information and Communications Security*, pp. 1889-1893, Dec. 2003.
9. Gang Li, Zixue Zhao and Jinxin Hao, "A generalized direct-form II transposed structure for IIR filter implementation with minimal roundoff noise gain," in *Proceedings of IEEE International Symposium on Circuits and System*, vol. 4, pp. 217-220, Jun. 2003.
10. Zixue Zhao and Gang Li, "Comparative study of the generalized DFII structure and its equivalent state-space realization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 557-560, Apr. 2003.

Under review:

1. Zixue Zhao, "An efficient state-space realization with minimum roundoff noise gain," revised and resubmitted to *IEEE Transactions on Circuits and System I*.

Bibliography

- [1] G. Li, C. R. Wan, and G. A. Bi, "An improved ρ DFIIt structure for digital filters with minimum roundoff noise," *IEEE Trans. Circuits Syst. II*, vol. 52, pp. 199-203, April 2005.
- [2] L.R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Prentice-Hall, 1975.
- [3] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing - Principles, Algorithms, and Applications*. Prentice-Hall, 3rd ed., 1999.
- [4] M. Kawamata and T. Higuchi, "Synthesis of 2-D separable denominator digital filters with minimum roundoff noise and no overflow oscillations," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 365-372, Apr. 1986.
- [5] W. S. Lu and A. Antoniou, *Two-Dimensional Digital Filters*. Marcel-Dekker, 1992.
- [6] L. R. Rabiner and C. M. Rader, *Digital Signal Processing*. IEEE Press, The Institute of Electrical and Electronics Engineers Inc., 1972.
- [7] T. Kailath, *Linear Systems*. Prentice-Hall, 1980.
- [8] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits and Syst.*, vol. 23, pp. 551-562, Sept. 1976.
- [9] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 25, pp. 273-281, Aug. 1977.
- [10] P. V. Dooren, "Numerical linear algebra: An increasing interest in linear system theory," in *Proc. European Conf. Theory Design*, The Hague, Netherlands, pp. 243-251, 1981.

- [11] P. V. Dooren and P. Dewilde, "Minimal cascade factorization of real and complex rational transfer matrices," *IEEE Trans. Circuits and Syst.*, vol. 28, pp. 390-400, May 1981.
- [12] V. Tavsanoğlu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. 31, pp. 884-888, Oct. 1984.
- [13] L. Thiele, "Design of sensitivity and roundoff noise optimal state-space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39-46, 1984.
- [14] L. Thiele, "On the sensitivity of linear state-space systems," *IEEE Trans. Circuits and Syst.*, vol. 33, pp. 502-510, May 1986.
- [15] J. E. Perkins, U. Helmke, and J. B. Moore, "Balanced realizations via gradient flow techniques," *Systems and Control Letters*, vol. 14, pp. 369-380, Jun. 1990.
- [16] M. Gevers and G. Li, *Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. Springer-Verlag, 1993.
- [17] W. Y. Yan and J. B. Moore, "On L_2 -sensitivity minimization of linear state-space systems," *IEEE Trans. Circuits and Syst. I*, vol. 39, pp. 641-648, Aug. 1992.
- [18] C. Xiao, "Improved L_2 -sensitivity for state-space digital system," *IEEE Trans. Signal processing*, vol. 45, pp. 837-840, Apr. 1997.
- [19] L. B. Jackson, A. G. Lindgren, and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. 23, pp. 179-153, Mar. 1979.
- [20] C. W. Barnes, "On the design of optimal state-space realizations of second-order digital filters," *IEEE Trans. Circuits Syst.*, vol. 31, pp. 602-608, Jul. 1984.
- [21] B. W. Bomar, "New second-order state-space structures for realizing low roundoff noise digital filters," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 33, pp. 106-110, Feb. 1985.
- [22] K. M. Anspach, B. W. Bomar, R. C. Engels, and R. D. Joseph, "Minimization of fixed-point roundoff noise in extended state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 43, pp. 193-206, Mar. 1996.

- [23] G. Amit and U. Shaked, "Small roundoff noise realization of fixed-point digital filters and controllers," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 36, pp. 880-891, Jun. 1988.
- [24] B. W. Bomar and J. C. Hung, "Minimum roundoff noise digital filters with some power-of-two coefficients," *IEEE Trans. Circuits and Syst.*, vol. 31, pp. 833-840, Oct. 1984.
- [25] W. J. Lutz and S. L. Hakimi, "Design of multi-input multi-output systems with minimum sensitivity," *IEEE Trans. Circuits and Syst.*, vol. 35, pp. 1114-1122, Sept. 1988.
- [26] G. Li, B. D. O. Anderson, M. Gevers, and J. E. Perkins, "Optimal FWL design of state-space digital systems with weighted sensitivity minimization and sparseness consideration," *IEEE Trans. Circuits and Syst.*, vol. 39, pp. 365-377, May 1992.
- [27] G. Li, M. Gevers, and Y. X. Sun, "Performance analysis of a new structure for digital filter implementations," *IEEE Trans. Circuits and Syst. I*, vol. 47, pp. 474-482, Apr. 2000.
- [28] N. Wong and T. S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," *IEEE Trans. Circuits and Syst. II*, vol. 48, pp. 425-431, Apr. 2001.
- [29] N. Wong and T. S. Ng, "Improved roundoff noise performance in a direct-form IIR filter using a modified delta operator," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 773-776, May 2001.
- [30] N. Wong and T. S. Ng, "Roundoff noise minimization in a modified direct form delta operator IIR structure," *IEEE Trans. Circuits and Syst. II*, vol. 47, pp. 1533-1536, Dec. 2000.
- [31] J. Kauraniemi and T. I. Laakso, "Roundoff noise analysis of modified delta operator direct form structures," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, pp. 2365-2368, Jun. 9-12, 1997.
- [32] J. Kauraniemi, T. I. Laakso, I. Harttimo, and S. J. Ovaska, "Roundoff noise minimization in a direct form delta operator structure," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 3, pp. 1371-1374, May 7-10, 1996.

- [33] J. Kauraniemi, T. I. Laakso, I. Harttimo, and S. J. Ovaska, "Delta operator realizations of direct-form IIR filters," *IEEE Trans. Circuits and Syst. II*, vol. 45, pp. 41-52, Jan. 1998.
- [34] G. Li and M. Gevers, "Comparative study of finite wordlength effects in shift and delta operator parametrizations," *IEEE Trans. Automatic Control*, vol. 38, pp. 803-807, May, 1993.
- [35] G. Li and M. Gevers, "Roundoff noise minimization using delta-operator realizations," *IEEE Trans. Signal Processing*, vol. 41, pp. 629-637, Feb. 1993.
- [36] H. A. Spang III and P. M. Shultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. 10, pp. 373-380, Dec. 1962.
- [37] L. B. Jackson, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust.*, vol. 18, pp. 107-122, Jun. 1970.
- [38] T. Thong and B. Liu, "Eoor spectrum shaping in narrowband recursive digital filters," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 25, pp. 200-203, Apr. 1977.
- [39] T. L. Chang and S. A. White, "An error cancellation digital filter structure and its distributed-arithmetic implementation," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 339-342, Apr. 1981.
- [40] D. C. Munson and D. Liu, "Narrowband recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 160-163, Feb. 1981.
- [41] W. E. Higgins and D. C. Munson, "Noise reduction strategies for digital filters: error spectrum shaping versus the optimal linear state-space formulation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 30, pp. 963-973, Dec. 1982.
- [42] M. Renfors, "Roundoff noise in error-feedback state-space filters," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 619-622, Apr. 1983.
- [43] W. E. Higgins and D. C. Munson, "Optimal and suboptimal error spectrum shaping for cascade-form digital filters," *IEEE Trans. Circuits and Syst.*, vol. 5, pp. 429-437, May 1984.
- [44] P. P. Vaidyanathan, "On error-spectrum shaping in state-space digital filters," *IEEE Trans. Circuits and Syst.*, vol. 32, pp. 88-92, Jan. 1985.

- [45] T. I. Laakso and I. O. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Trans. Signal Processing*, vol. 40, pp. 1096-1107, May. 1992.
- [46] T. Hinamoto, H. Ohnishi, and W. S. Lu, "Roundoff noise minimization of state-space digital filters using separate and joint error feedback/coordinate transformation optimizations," *IEEE Trans. Circuits Syst. I*, vol. 50, pp. 23-33, Jan. 2003.
- [47] P. Agathoklis and C. S. Xiao, "Low roundoff noise structure for 2-D filters," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 352-355, May 12-15, 1996.
- [48] P. Agathoklis and A. Kanellakis, "Roundoff noise 2-D digital filters using residue feedback," in *Proc. IEEE Southeastcon 'Creative Technology Transfer - A Global Affair'*, pp. 12-15, Apr. 10-13, 1994.
- [49] T. Hinamoto, S. Karino, and N. Kuroda, "2-D state-space digital filters with error spectrum shaping," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 2, pp. 766-769, May 1996.
- [50] T. Hinamoto, N. Kuroda, and T. Kuma, "Error feedback for noise reduction in 2-D digital filters with quadrantly symmetric or antisymmetric coefficients," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, pp. 2461-2464, Jun. 1997.
- [51] T. Hinamoto, S. Karino, N. Kuroda, and T. Kuma, "Error spectrum shaping in two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 46, pp. 1203-1215, Oct. 1999.
- [52] T. Hinamoto, K. Higashi, and W. S. Lu, "Separate/joint optimization of error feedback and coordinate transformation for roundoff noise minimization in two-dimensional state-space digital filters," *IEEE Trans. Signal Processing*, vol. 51, pp. 2436-2445, Sept. 2003.
- [53] L. M. Smith and B. W. Bomar, "An algorithm for constrained roundoff noise minimization in digital filters with application to two-dimensional filters," *IEEE Trans. Circuits and Syst.*, vol. 35, pp. 1359-1368, Nov. 1988.
- [54] T. Lin, M. Kawamata, and T. Higuchi, "A unified study on the roundoff noise in 2-D state-space digital filters," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 724-730, Jul. 1986.

- [55] W. S. Lu and A. Antoniou, "Synthesis of 2-D state-space fixed-point digital-filter structures with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 965-973, Oct. 1986.
- [56] T. Hinamoto, T. Hamanaka, and S. Maekawa, "A generalized study on the synthesis of 2-D state-space digital filters with minimum roundoff noise," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 1037-1042, Aug. 1988.
- [57] K. Premaratne, M. M. Ekanayake, J. I. Suarez, and P. H. Bauer, "Two-dimensional delta-operator formulated discrete-time systems: state-space realization and its coefficient sensitivity properties," *IEEE Trans. Signal Processing*, vol. 46, pp. 3445-3449, Dec. 1998.
- [58] A. Zilouchian and R. L. Carroll, "A coefficient sensitivity bound in 2-D state-space digital filtering," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 665-667, Jul. 1986.
- [59] M. Kawamata, T. Lin, and T. Higuchi, "Minimization of sensitivity of 2-D state-space digital filters and its relation to 2-D balanced realizations," in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 710-713, 1987.
- [60] T. Hinamoto and T. Takao, "Synthesis of 2-D state-space filter structures with low frequency-weighted sensitivity," *IEEE Trans. Circuits Syst. II*, vol. 39, pp. 646-651, Sept. 1992.
- [61] G. Li, "Two dimensional system optimal realizations with L_2 -sensitivity minimization," *IEEE Trans. Signal Processing*, vol. 46, pp. 809-813, Mar. 1998.
- [62] T. Hinamoto, Y. Zempo, Y. Nishina, and W. S. Lu, "An analytical approach for the synthesis of two-dimensional state-space filter structure with minimum weighted sensitivity," *IEEE Trans. Circuits and Syst. I*, vol. 46, pp. 1172-1183, Oct. 1999.
- [63] T. Hinamoto, S. Yokoyama, T. Inoue, W. Zeng, and W. S. Lu, "Analysis and minimization of L_2 -sensitivity for linear systems and two-dimensional state-space filters using general controllability and observability gramians," *IEEE Trans. Circuits Syst. I*, vol. 49, pp. 1279-1289, Sept. 2002.
- [64] T. Hinamoto and Y. Sugie, " L_2 -sensitivity analysis and minimization of 2-D separable-denominator state-space digital filters," *IEEE Trans. Signal Processing*, vol. 50, pp. 3107-3114, Dec. 2002.

- [65] T. Hinamoto and F. W. Fairman, "Separable-denominator state-space realization of two-dimensional filters using a canonic form," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 29, pp. 846-853, Aug. 1981.
- [66] P. K. Rajan and M. N. S. Swamy, "Design of separable denominator 2-dimensional digital filters possessing real circularly symmetric frequency responses," in *Proc. Inst. Elect. Eng. G*, vol. 129, pp. 235-240, Oct. 1982.
- [67] B. Lashgari, L. M. Silverman, and J. F. Abramatic, "Approximation of 2-D separable in denominator filters," *IEEE Trans. Circuits Syst.*, vol. 30, pp. 107-121, Feb. 1983.
- [68] H. K. Kwan and C. L. Chan, "Design of linear phase circularly symmetric two-dimensional recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1023-1029, Jul. 1989.
- [69] W. S. Lu, H. P. Wang, and A. Antoniou, "Design of two-dimensional digital filters using singular-value decomposition and balanced approximation method," *IEEE Trans. Signal Processing*, vol. 39, pp. 2253-2262, Oct. 1991.
- [70] C. Xiao and P. Agathoklis, "Design and implementation of approximately linear phase two-dimensional IIR filters," *IEEE Trans. Circuits and Syst. II*, vol. 45, pp. 1279-1288, Sept. 1998.
- [71] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.
- [72] B. W. Bomar, "Finite Wordlength Effects", in *The Circuits and Filters Handbook* (W. K. Chen, ed.), pp. 2739-2757, Boca Raton, FL: CRC Press, 2nd ed., 2003.
- [73] S. R. Parker and S. F. Hess, "Limit-cycle oscillations in digital filters," *IEEE Trans. Circuits Theory*, vol. 19, pp. 687-697, Nov. 1971.
- [74] P. H. Bauer, "Limit cycle bounds for floating-point implementations of second-order recursive digital filters," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 493-501, Aug. 1993.
- [75] B. D. Green and L. E. Turner, "New limit cycle bounds for digital filters," *IEEE Trans. Circuits Syst.*, vol. 35, pp. 365-374, Apr. 1988.
- [76] M. Buttner, "A novel approach to eliminate limit cycles in digital filters with minimum increase in the quantization noise," in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 291-294, Apr. 1976.

- [77] P. S. R. Diniz and A. Antoniou, "More economical state-space digital filter structures which are free of constant-input limit cycles," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 807-815, Aug. 1986.
- [78] B. W. Bomar, "Low-roundoff-noise limit-cycle-free implementation of recursive transfer functions on a fixed-point digital signal processor," *IEEE Trans. Industr. Electron.*, vol. 41, pp. 70-78, Feb. 1994.
- [79] A. N. Willson, Jr., "Limit cycle due to adder overflow in digital filters," *IEEE Trans. Circuits Theory*, vol. 19, pp. 342-346, July. 1972.
- [80] D. Mitra, "A bound on limit cycles in digital filters which exploits a particular structural property of the quantization," *IEEE Trans. Circuits Syst.*, vol. 24, pp. 581-589, Nov. 1977.
- [81] J. H. F. Ritzerfield, "a condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures using saturation," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 1049-1057, Aug. 1989.
- [82] W. T. Mills, C. T. Mullis, and R. A. Roberts, "Digital filter realizations without overflow oscillations," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 26, pp. 334-338, Aug. 1978.
- [83] J. B. Knowles and E.M. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuits Theory*, vol. 15, pp. 31-41, Mar. 1968.
- [84] C. Xiao, "A new statistical approach to the coefficient word length problem for digital filters," *IEEE Trans. Circuits Syst.*, vol. 22, pp. 190-196, Mar. 1975.
- [85] D. Williamson, "Roundoff noise minimization and pole-zero sensitivity in fixed-point digital filters using resider feedback," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 34, pp. 1210-1220, Oct. 1986.
- [86] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*. Addison-Wesley, 1987.
- [87] E. D. Sontag, "On first-order equations for multidimensional filters," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 26, pp. 480-482, Oct. 1978.
- [88] S. H. Zak, E. B. Lee, and W. S. Lu, "Realization of 2-D filters and time-delay systems," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 1241-1244, Dec. 1986.
- [89] R. P. Roesser, "A discrete state-space model for linear image processing," *IEEE Trans. Automat. Control*, vol. 20, pp. 1-10, Feb. 1975.

- [90] T. Hinamoto, "Realization of a state-space model from two-dimensional input-output map," *IEEE Trans. Circuits Syst.*, vol. 27, pp. 36-44, Jan. 1980.
- [91] K. Premaratne, E. I. Jury, and M. Mansour, "An algorithm for model reduction of 2-D discrete time systems," *IEEE Trans. Circuits Syst.*, vol. 37, pp. 1116-1131, Sept. 1980.
- [92] W. S. Lu, H. P. Wang, and A. Antoniou, "An efficient method for the evaluation of the controllability and observability gramians of 2-D digital filters and systems," *IEEE Trans. Circuits Syst.*, vol. 39, pp. 695-704, Oct. 1992.
- [93] J. H. Holland, *Adaptation in natural and artificial system*. Ann Arbor, University of Michigan Press, 1975.
- [94] D. H. Horrocks and Y. M. Khaliffa, "Genetically derived filter circuits using preferred value components," in *Proc. IEE Colloquium on Analogue Signal Processing*, (Oxford, UK), pp. 4/1-4/5, Oct. 1994.
- [95] D. Suckley, "Genetic algorithms in the design of FIR filters," *Circuits, Devices and Systems, IEE Proceedings G*, vol. 138, pp. 234-238, Apr. 1991.
- [96] G. Wade, P. E. Van, and H. Darwen, "Synthesis of efficient low-order FIR filters from primitive sections," *Circuits, Devices and Systems, IEE Proceedings G*, vol. 137, pp. 367-372, Oct. 1990.
- [97] A. Roberts and G. Wade, "A structured GA for FIR filter design," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, (Chelmsford, UK), vol. 1, pp. 16/1-16/8, Nov. 1993.
- [98] P. B. Wilson and M. D. Macleod, "Low implementation cost IIR digital filter design using genetic algorithms," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, (Chelmsford, UK), vol. 1, pp. 4/1-4/8, Nov. 1993.
- [99] M. S. White and S. J. Flockton, "A comparative study of natural algorithms for adaptive IIR filtering," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, (Chelmsford, UK), vol. 1, pp. 22/1-22/8, Nov. 1993.
- [100] T. Wicks and S. Lawson, "Genetic algorithm design of wave digital filters with a restricted coefficient set," in *Proc. IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, (Chelmsford, UK), vol. 1, pp. 17/1-17/7, Nov. 1993.

- [101] D. M. Etter, M. J. Hicks, and K. H. Cho, "Recursive adaptive filter design using an adaptive genetic algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 635-638, May 1982.
- [102] V. Peterka, "Control of uncertain processes: Applied theory and algorithms," *Kybernetika*, vol. 22, pp. 1-102, 1986.
- [103] R. H. Middleton and G. C. Goodwin, *Digital Control and Estimation: A Unified Approach*. Prentice-Hall, 1990.
- [104] G. C. Goodwin, "Some observations on robust stochastic estimation," in *Proc. 8th IFAC/IFORS Symp. Identification and System Parameter Estimation*, pp. 22-32, 1988.
- [105] J. Fadavi-Ardekani, S. K. Mitra, and B. D. O. Anderson, "Extended state-space model of discrete-time dynamical systems," in *IEEE Trans. Circuits Syst.*, vol. 29, pp. 547-556, Aug. 1982.