MULTIPLIERLESS MULTIRATE FIR FILTER DESIGN AND IMPLEMENTATION

YU YAJUN (*M. Eng.*)

A Thesis Submittted
for the Degree of Doctor of Philosophy
Department of Electrical & Computer Engineering
National University of Singapore
2003

Acknowledgements

The work leading to this thesis was done during my years as graduate student at the Signal Processing & VLSI Design Laboratory at the Department of Electrical & Computer Engineering. I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

The first person I would like to thank is my supervisor Professor Lim Yong Ching for his stimulating suggestions and constant encouragement throughout the entire course of this research. His enthusiasm and integral view on research and his mission for providing only high-quality work, have made a deep impression on me. I am most grateful to him for cultivating me into this attitude of doing research. Besides being an excellent supervisor, he is as close as a relative and a good friend to me. I am really glad that I am his student.

I also want to take the opportunity to thank Professor Tapio Saramäki and Dr. Robert Bregović, at the Institute of Signal Processing, Tampere University of Technology, for precious discussion, and to Professor Wu-Shen Lu, at the Department of Electrical Engineering, University of Victoria and Professor Teo Kok Lay of the Applied Mathematics Department, the Hong Kong Polytechnic University, for their advices on optimization techniques.

The pleasant research atmosphere in the lab is due to several factors. One of the most important factors are the people through the different stages of my own stay here: Mr. Shi Qian, Mr. Shen Ling, Mr. Guan Xiang, Dr. Ha Yajun, Mr. Anslem Yep, Mr. Zhu Haiqing, Dr. Goh Chee-Kiang, Ms. Zhang Xiwen, Mr. Francis Boey, Mr. Yu Wen, Mr. Wu Haijie, Ms. Xu Lianchun, Mr. Jiang Bin, Mr. Liu Xiaoyun, Mr. Yang Chunzhu, Mr. Yu Jianghong, Ms. Cui Jiqing, Mr. Luo Zhenyin, Mr. Zhou

Xiangdong, Mr. Liang Yunfeng, Ms. Zheng Huanqun, Ms. Sun Pinping, Mr. Wang Xiaofeng, Mr. Lee Jun Wei, Ms. Cen Lin, Mr. Xia Xiaojun.

Of these I want to give special thanks to Shi Qian, Shen Ling and Xia Xiaojun for the happy hours we played tennis together during the years, to Yang Chunzhu for his delicious food cooked for us, and to Yu Wen for his kindness in providing accommodations for me at one stage.

Finally, I would like to give my special thanks to my parents, Yu Qijia and Peng Wensen, and my sister, Yu Yachen, whose love and trust enabled me to complete this work. I also want to thank all of my friends for their invaluable support, patience and encouragement throughout my years of study.

Contents

Summary

Multirate systems and filter banks have found various applications in many areas, such as speech coding, image compression, adaptive signal processing as well as signal transmission. The function of a multirate filter bank is to separate the input signal into two or more frequency bands of signals, or combining two or more different frequency bands of signals into a single output signal. The two-channel filter bank is an important filter bank family. It can be used as a basic building block to construct an M-channel filter bank.

Multiplierless techniques have been successfully applied in the synthesis of linear phase FIR filters with very low complexity. Recently, much attention has been given to the design of multiplierless multirate filter banks. Among all the various types of this class of filter bank, the lattice-structure perfect-reconstruction (PR) filter bank presents a desirable feature that the PR property is preserved even under the lattice coefficient quantization.

In this thesis, the design of multiplierless two-channel lattice filter bank is discussed with respect to two aspects. First, several optimization techniques for the design of signed power-of-two (SPT) coefficient lattice filter bank are developed. The optimization techniques include the successive reoptimization technique, improved genetic algorithm, and width-recursive depth-first tree search algorithm. Based upon the new results obtained in this thesis and those reported in the previous literatures, it can be concluded that the tree search algorithm is more suitable than the other techniques for the design of the multiplierless two-channel lattice filter bank. Second, the statistical SPT rounding error distribution and the effects

of rounding the coefficient values to SPT values on the filter bank frequency responses are studied. Based on the knowledge of the SPT rounding error and its effects on the frequency response, an SPT term allocation scheme is developed. A tree search algorithm incorporating the SPT term allocation scheme is developed for the design of SPT coefficient filter banks with different number of SPT terms being allocated to each coefficient keeping the total number of SPT terms fixed; the stopband attenuation achieved is very much superior to the filters designed when each coefficient is allocated the same number of SPT terms.

In addition, a new polyphase implementation technique is introduced in the thesis. In this new technique, coefficient symmetry is preserved for each of the polyphase components. This results in a factor-of-two reduction in the multiplication rate.

List of Tables

List of Figures

Chapter 1

Introduction

FINITE IMPULSE RESPONSE (FIR) filters possess many virtues, such as exact linear phase property, guaranteed stability, free of limit cycle oscillations, and low coefficient sensitivity [61,63,64]. However, the order of an FIR filter is generally higher than that of a corresponding infinite impulse response (IIR) filter meeting the same magnitude response specifications. Thus, FIR filters require considerably more arithmetic operations and hardware components — delay, adder and multiplier. This makes the implementation of FIR filters, especially in applications demanding narrow transition bands, very costly. When implemented in VLSI (Very Large Scale Integration) technology, the coefficient multiplier is the most complex and the slowest component. The cost of implementation of an FIR filter can be reduced by decreasing the complexity of the coefficients [41,48,52,68]. Coefficient complexity reduction includes reducing the coefficient wordlength and coefficient representation using a limited number of signed power-of-two (SPT) terms.

Since the 60's, much attention has been put into the study of the effect of coefficient quantization on the frequency responses of FIR filters [11, 16, 39, 40] for implementation on general purpose digital computer or special purpose hardware. A statistical bound on the error due to coefficient quantization was developed. Subsequently, optimal finite wordlength FIR digital filters in the minimax sense were designed by using mixed integer linear programming (MILP) [12, 41]. It was reported that the computing resources required by running MILP algorithm were very

high. However, coefficient wordlength of the optimum solution obtained by using MILP is only a few bits shorter than that obtained by simple coefficient rounding. Almost concurrent with the use of MILP for the design of limited wordlength FIR filter was the use of MILP for the design of FIR filter with SPT coefficients [49,52]. Filters with SPT coefficients have the advantage that they can be implemented without multipliers, i.e., the filter's coefficient multipliers can be replaced by simple shift-and-add circuits. Thus, the computational complexity of the filter is reduced.

During the past decades, numerous algorithms have been proposed for the design of FIR filters with SPT coefficients. Besides the "optimal" technique employing MILP, there are other suboptimal techniques such as local search methods [67,86], tree searches with weighted least-squares criteria [45,53], stochastic optimization, for example, simulated annealing [5] and genetic algorithms [26,46], dynamic SPT terms allocation algorithms [47], quantization by coefficient sensitivity [10,72], and SPT terms allocation incorporating local search approach [15].

With increasing applications of multirate systems and filter banks in many areas [79], recently, much attention has been given to the design of multiplierless multirate filter banks [34,35]. Among the various types of this class of filter bank structures, the lattice-structure perfect-reconstruction (PR) filter bank [81] has attracted particular attention because it possesses the desirable feature that the PR property is preserved even under coefficient quantization.

1.1 Contributions

Filter banks have found applications in audio and video signal processing [24,79], especially for subband coding of speech and image signals. The main function of a multirate filter bank is to separate the input signal into two or more frequency bands of signals or for recombining two or more different frequency bands of signals into a single signal. The two-channel filter bank is an important member of the filter bank family. It can be used as a fundamental building block to construct an

M-channel filter bank in a tree structure.

The two-channel FIR filter banks can be classified into three types, viz., quadrature mirror filter banks, orthogonal filter banks, and biorthogonal filter banks [20]. During the last two decades, many techniques have been developed to optimize the two-channel filter banks [6–9,13,30,31,35,54,58,70,81–83,85]. The finite wordlength effects [71] and the design techniques [14, 34, 43, 57, 75, 76] for the finite wordlength coefficient filter banks have also been extensively studied.

The lattice orthogonal filter bank [81] has the property that the PR property is satisfied for any combination of the lattice coefficients. This property is very attractive for discrete coefficient optimization. The quantization of the lattice coefficients, however, still affects the frequency response of the filter bank. Several algorithms have been proposed to design the multiplierless lattice filter banks [34, 75]; however, these algorithms involved direct application of the conventional linear phase FIR filter design techniques without taking into consideration the properties of the filter bank. Furthermore, these existing algorithms are heuristic in nature and do not promise optimum solution. It is noted that there has been no report on the study of SPT rounding error distribution and its effects on the filter bank frequency response.

In this thesis, the design of multiplierless two-channel lattice filter bank is investigated in two aspects. First, several optimization techniques for the design of SPT coefficient lattice filter bank are developed with the consideration of the filter banks' property. Second, the statistical SPT rounding error distribution and the effects of rounding the coefficient to SPT values on the filter bank's frequency response are studied. Based on the knowledge of the SPT rounding error distribution and its effects on the filter bank, an SPT term allocation scheme is developed. The SPT term allocation scheme when incorporated into a suitable optimization algorithm is able to design the SPT coefficient filter banks with different number of SPT terms to each coefficient.

Under the conventional wisdom, coefficient symmetry is lost when a filter is

split into its polyphase components. In this thesis, a technique for preserving the coefficient symmetry under polyphase implementation is introduced. This results in a factor-of-two reduction in the multiplication rate required in the polyphase implementation.

For the multiplierless two-channel lattice orthogonal filter bank design and the polyphase implementation, the following is claimed to be original.

- A successive reoptimization approach is proposed for the design of the lattice filter bank. In this technique, the coefficient values are quantized sequentially one at a time. The order of selection of the coefficient for quantization is based on a coefficient sensitivity measure. It is observed that the lattice coefficient sensitivities differ greatly from coefficient to coefficient. The successive reoptimzation approach exploit this property by first quantizing the coefficient with the highest sensitivity measure and reoptimize the remaining coefficients to compensate for the frequency response deterioration caused by the coefficient quantization.

- An improved genetic algorithm is developed to optimize the lattice filter bank. A new coding scheme is introduced to code the SPT coefficients in such a way that the canonic property of the SPT values is preserved under genetic operation. Additionally, two new features which dramatically improve the genetic algorithm are introduced.

- A width-recursive depth-first tree search technique is developed to optimize the lattice filter bank. Compared with existing tree search methods, this technique has two advantages. First, it quickly yields a suboptimal discrete solution; second, it covers a large search space if the necessary computing resources are available. In this method, a frequency response deterioration measure is introduced to serve as a branching criterion for the search.

- SPT rounding error distribution is studied. A formula for the error probability density function is developed.

- The statistical effect of quantizing the lattice filter banks' coefficients to SPT values is studied. Based on this analysis, an SPT term allocation scheme is developed for the design of SPT coefficient lattice filter bank where each coefficient is allocated with a different number of SPT terms while keeping the total number of SPT terms allocated to the entire filter fixed.

- A polyphase implementation of the filter bank preserving the coefficient symmetry is presented.

Findings reported in this paper have been published or are being submitted for consideration for publication or are being prepared for publication in the following papers:

- Y.C. Lim and Y. J. Yu, "A successive reoptimization approach for the design of discrete coefficient perfect reconstruction lattice filter bank," in *Proc. IEEE. Int. Symp. Circuits and Syst.,* vol. 2, pp. 69-72, Switzerland, June 2000.

- Y. J. Yu and Y.C. Lim, "A sequential reoptimization approach for the design of signed power-of-two coefficient lattice QMF bank," in *Proc. IEEE. TENCON,* pp. 57-60, Singapore, Aug. 2001.

- Y. J. Yu and Y.C. Lim, "New natural selection process and chromosome encoding for the design of multiplierless lattice QMF using genetic algorithm," in *Proc. IEEE. Int. Conf. Elect. Compt. Syst.,* pp. 1273-1276, Malta, Sept. 2001.

- Y. J. Yu and Y.C. Lim, "A novel genetic algorithm for the design of a signed power-of-two coefficient quadrature mirror filter lattice filter bank," *Circuit Syst. Signal Process.,* vol. 21, pp. 263-276, May/June, 2002.

- Y.C. Lim and Y. J. Yu, "A width-recursive depth-first tree search approach for the design of discrete coefficient perfect reconstruction lattice filter bank," *IEEE Trans. Circuits, Syst. II,* vol. pp, 257-266, June 2003.

- Y. J. Yu, Y.C. Lim and T. Saramäki, "Restoring Coefficient Symmetry in Polyphase Implementation of Linear Phase FIR Filters," Submitted to *IEEE Trans. Circuits, Syst. I.*

- Y. J. Yu, Y.C. Lim and K.L. Teo, "An Analysis on Signed Power-of-Two Rounding Errors and Effects. I: Statistical Rounding Error Distributions," to be submitted to *IEEE Trans. Circuits, Syst. I.*

- Y. J. Yu, Y.C. Lim and K.L. Teo, "An Analysis on Singed Power-of-Two Rounding Errors and Effects. II: Statistical Rounding Error Effects and their Applications on the Design of Lattice Filter Banks with SPT coefficients," to be submitted to *IEEE Trans. Circuits, Syst. I.*

1.2 Thesis Outline

Chapter 1 gives an introduction to the problems considered and the contributions made in this thesis.

In Chapter 2, a literature review briefly describes the multirate systems and filter banks. Also presented in Chapter 2 are the property and necessary conditions for alias-free, perfect reconstruction two-channel filter banks. The signed power-of-two coefficient property and the existing SPT coefficient design techniques are also reviewed.

In Chapters 3, 4 and 5, the problems encountered in the optimization process of designing the two-channel lattice filter bank with SPT coefficients are discussed. Chapter 3 introduces a successive reoptimization approach, while Chapter 4 presents an improved genetic algorithm. A tree search algorithm for the design of SPT coefficient filter banks is proposed in Chapter 5. A comparison among the

techniques proposed in these three chapters and those reported in the previous literatures is also presented in Chapter 5.

Studies on the error distribution for quantizing a number to an SPT value are presented in Chapter 6. In Section 6.1, mathematical expressions of the error probability density function for representing a number by a given number of SPT terms and a given precision are deduced. Based on the error distributions, in Section 6.2, the statistical SPT quantization effects for the two-channel lattice orthogonal filter banks are discussed. An SPT term allocation scheme is developed in Section 6.3. This SPT term allocation scheme is incorporated into the width-cursive depth-first tree search algorithm in Section 6.4 to design the SPT coefficient lattice filter bank.

In Chapter 7, a new polyphase implementation technique is presented. In this technique, the coefficient symmetry of linear phase FIR filter is preserved for each polyphase component. A comparison among the proposed implementation, traditional polyphase implementation and direct form implementation is performed.

Chapter 8 contains a summary of the key results obtained in this research together with relevant conclusions drawn.

Chapter 2

Multirate Systems

TWO-CHANNEL FILTER BANKS operate at more than one sampling rate. Such systems are called multirate digital systems. In comparison with single rate digital system, a multirate digital system has two additional processes: the decimation process and interpolation process. The decimation process decreases the sampling rate, whereas the interpolation process increases the sampling rate.

This chapter reviews several basic topics on multirate systems and filter banks. First, the decimation and interpolation processes are introduced. Second, basic operation principles of a two-channel filter bank are discussed and the necessary conditions for aliasing-free and perfect-reconstruction (PR) filter banks are described. Last, the representation and properties of signed power-of-two (SPT) coefficients are described. Existing SPT coefficient design techniques are reviewed.

2.1 Decimation and Interpolation

The most basic operations in multirate digital signal processing are decimation and interpolation.

2.1.1 The Decimation Process

The decimation process reduces the sampling rate of a signal. It consists of an M-fold decimator, preceded by an anti-aliasing filter, $H(z)$, as shown in Fig. 2.1.

The M-fold decimator takes an input sequence $x(n)$ and produces one output sample in every M input samples. The relationship between the output sequence

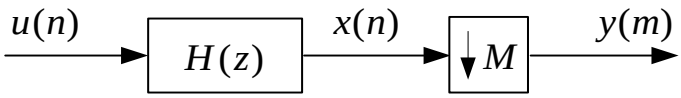

Fig. 2.1: The decimation process consisting of an anti-aliasing filter $H(z)$ and a decimator.

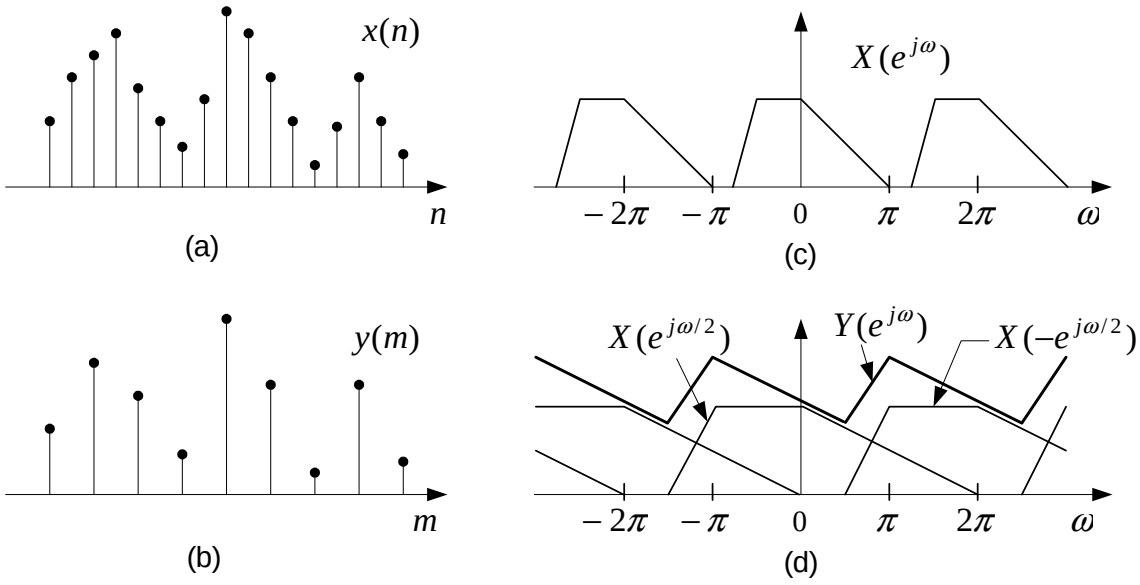

Fig. 2.2: A decimator for $M = 2$. (a) Input sequence $x(n)$, (b) Decimated output sequence $y(m)$, (c) Fourier transform of the input sequence, $X(e^{j\omega})$, and (d) Fourier transform of the decimated output sequence, $Y(e^{j\omega})$.

$y(m)$ and the input signal $x(n)$ is as follows:

$$y(m) = x(Mm), \tag{2.1}$$

where M is an integer. The sampling rate at the output of the M-fold decimator is M times slower than the sampling rate at the input of the M-fold decimator. An example of a 2-fold decimation process is shown in Fig. 2.2. Given an input sequence $x(n)$ as shown in Fig. 2.2(a), the output of the 2-fold decimator is illustrated in Fig. 2.2(b). Since the decimator retains only one in every M input samples, in general, it may not be possible to recover $x(n)$ from $y(m)$ because of loss of information.

Denote the z-transform of $x(n)$ as $X(z)$, and the z-transform of $y(m)$ as $Y(z)$.

$Y(z)$ can be expressed in terms of $X(z)$ as

$$Y(z) = \frac{1}{M} \sum_{k=0}^{M-1} X\left(z^{\frac{1}{M}} e^{-j\frac{2k\pi}{M}}\right). \tag{2.2}$$

By substituting z by $e^{j\omega}$ in (2.2), the Fourier transform of the decimator output is obtained as

$$Y(e^{j\omega}) = \frac{1}{M} \sum_{k=0}^{M-1} X\left(e^{\frac{j\omega - j2\pi k}{M}}\right). \tag{2.3}$$

It can be seen that $Y(e^{j\omega})$ is a sum of M stretched (by a factor of M) and shifted (uniformly in successive amount of 2π) versions of $X(e^{j\omega})$, followed by scaling the magnitude by a factor of M. Assume that the Fourier transform of the input sequence $x(n)$ in Fig. 2.2(a) is as shown in Fig. 2.2(c), the Fourier transform of its M decimated output, where $M = 2$, is illustrated in Fig. 2.2(d).

From Fig. 2.2(d), it can be seen that these M stretched and shifted versions of $X(e^{j\omega})$, in general, may overlap. This overlap effect is called aliasing. $x(n)$ cannot be recovered from the decimated version $y(m)$ if aliasing occurs. The aliasing, in general, can be avoided if $x(n)$ is a lowpass signal bandlimited to the region $|\omega| < \frac{\pi}{M}$. Therefore, in most applications, the decimator is preceded by a filter $H(z)$, as shown in Fig. 2.1, to ensure that the signal being decimated is bandlimited. Such a filter is called the decimation filter.

2.1.2 The Interpolation Process

In contrast to the decimation process which decreases the sampling rate, the interpolation process increases the sampling rate. It consists of an L-fold expander, followed by an anti-image filter, $H(z)$. The block diagram of an L-fold interpolation process is shown in Fig. 2.3.

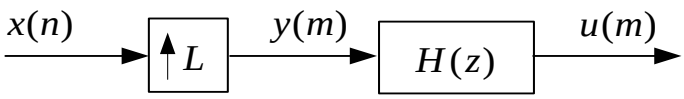

Fig. 2.3: The interpolation process consisting of an expander and an anti-image filter $H(z)$.

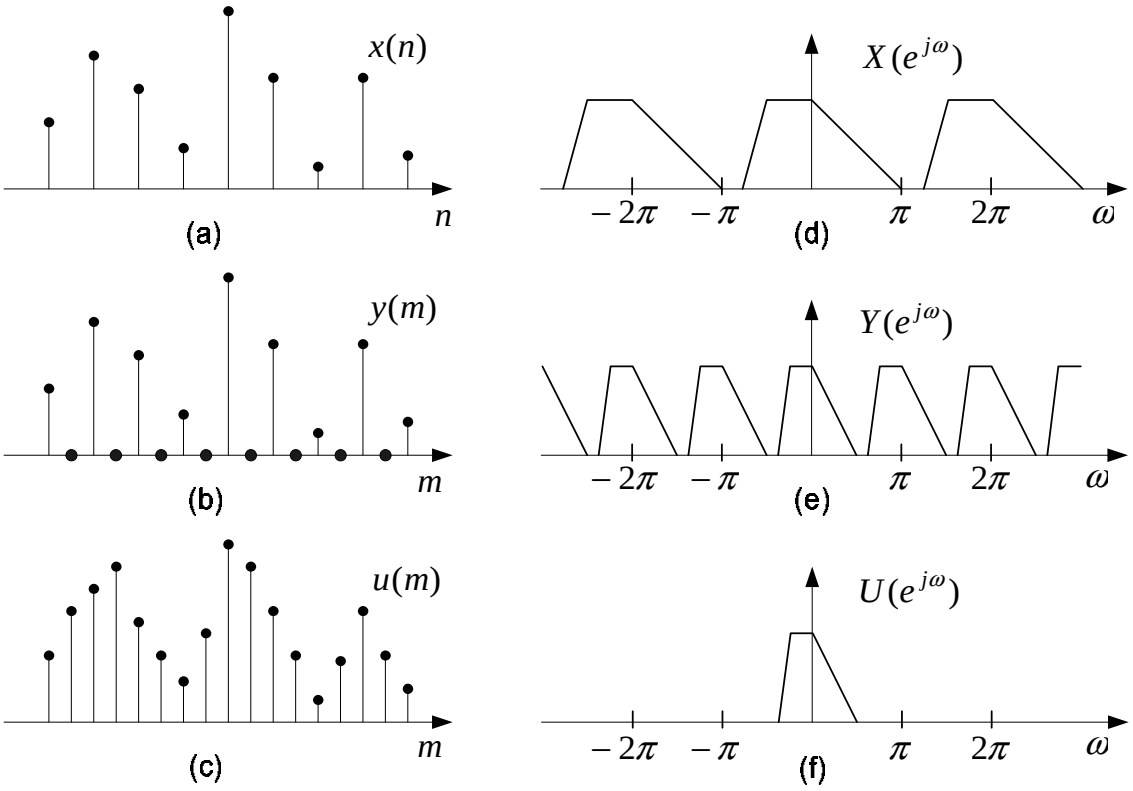

Fig. 2.4: An interpolation process for $L = 2$. (a) Input sequence $x(n)$, (b) expanded sequence, $y(m)$, (c) interpolated output, $u(m)$, (d) Fourier transform of the input sequence, $X(e^{j\omega})$, (e) Fourier transform of the expanded sequence, $Y(e^{j\omega})$, and (f) Fourier transform of the interpolated output sequence, $U(e^{j\omega})$.

The expander takes an input sequence $x(n)$ and produces an output sequence

$$y(m) = \begin{cases} x(\frac{m}{L}), & m = kL, \quad k \text{ is integer} \\ 0, & \text{otherwise,} \end{cases} \tag{2.4}$$

by placing $(L-1)$ equally spaced zeros between each pair of input samples. The sampling rate at the output of the L-fold expander is L times faster than that at the input. Fig. 2.4(a) and Fig. 2.4(b) demonstrate an input sequence, $x(n)$, and expanded sequence, $y(m)$, of an expander for $L = 2$. The expander does not cause any loss of information. The input sequence $x(n)$ can be recovered from $y(m)$ by an appropriate L-fold decimation.

Denoting the z-transform of $x(n)$ by $X(z)$, and the z-transform of $y(m)$ by

$Y(z)$. $Y(z)$ can be easily expressed in terms of $X(z)$ as

$$Y(z) = X(z^L). \tag{2.5}$$

The Fourier transform relationship between the input and output sequences of the expander is $Y(e^{j\omega}) = X(e^{j\omega L})$. This means that $Y(e^{j\omega})$ is an L compressed version of $X(e^{j\omega})$ as shown in Figs. 2.4(d) and Figs. 2.4(e). The expander introduces images in $Y(e^{j\omega})$ due to the periodicity of $X(e^{j\omega})$. To suppress all those images, the expander is followed by an interpolation filter, $H(z)$, as shown in Fig. 2.3. Typically, the interpolation filter is lowpass with cutoff frequency π/L. Thus, only the spectrum in Fig. 2.4(f) is retained. The effect in time domain, as shown in Fig. 2.4(c), is that the zero-valued samples introduced by the expander are interpolated.

2.1.3 Cascade Equivalences

As shown in Section 2.1.1 and Section 2.1.2, a multirate system is formed by an interconnection of a sampling rate change component and a digital filter. These components appear in a cascade form. An interchange of the components' positions may lead to a computationally efficient realization. Two important cascade equivalence relations are depicted in Fig. 2.5. The validity of these equivalences can be readily established by using (2.2) and (2.5).

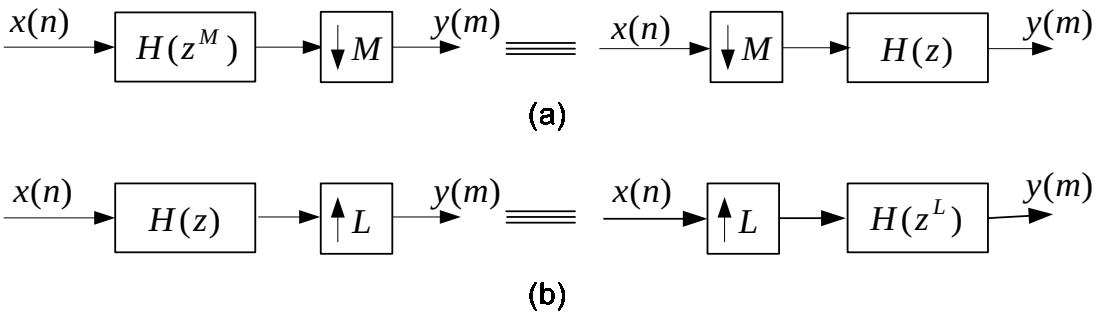

Fig. 2.5: Cascade equivalences: (a) the first equivalence, and (b) the second equivalence.

These two cascade equivalences enable us to move the basic sampling rate

change devices in multirate systems to more advantageous positions. They are extremely useful for efficient implementation of multirate systems.

2.1.4 Polyphase Decomposition

For decimation and interpolation processes, the computational complexity of the FIR filter may be reduced by using the polyphase decomposition [4] technique. The polyphase decomposition technique is reviewed in this section and its applications in the efficient realization of the decimation and interpolation processes are also illustrated.

Consider a filter $h(n)$ with z-transform $H(z)$:

$$H(z) = \sum_{n=-\infty}^{+\infty} h(n)z^{-n}. \tag{2.6}$$

$H(z)$ can be rewritten as

$$H(z) = \sum_{r=0}^{R-1} z^{-r}E_r(z^R) = \sum_{r=0}^{R-1} z^{-r} \sum_{k=-\infty}^{+\infty} h(kR+r)z^{-kR}, \tag{2.7}$$

where

$$E_r(z^R) = \sum_{k=-\infty}^{+\infty} e_r(k)z^{-kR} = \sum_{k=-\infty}^{+\infty} h(kR+r)z^{-kR}, \quad r = 0, 1, ..., R-1 \tag{2.8}$$

denotes the r-th polyphase component of $H(z)$.

Therefore, an M-fold decimation filter, as shown in Fig. 2.6(a), can be decomposed into its M polyphase components according to (2.7). The polyphase decomposition of the M-fold decimation filter is illustrated in Fig. 2.6(b).

Applying the first cascade equivalence shown in Fig. 2.5(a), Fig. 2.6(b) can be redrawn as shown in Fig. 2.6(c), which is computationally more efficient than the structure shown in Fig. 2.6(a). Each polyphase component in Fig. 2.6(c) operates at the output sampling rate, which is $\frac{1}{M}$ of the input rate. Therefore, the total computation rate in the system is reduced by a factor of M. By realizing each of the polyphase components in the structure shown in Fig. 2.6(c) as a transposed direct form FIR filter, as shown in Fig. 2.6(d), it can be observed that the same delay

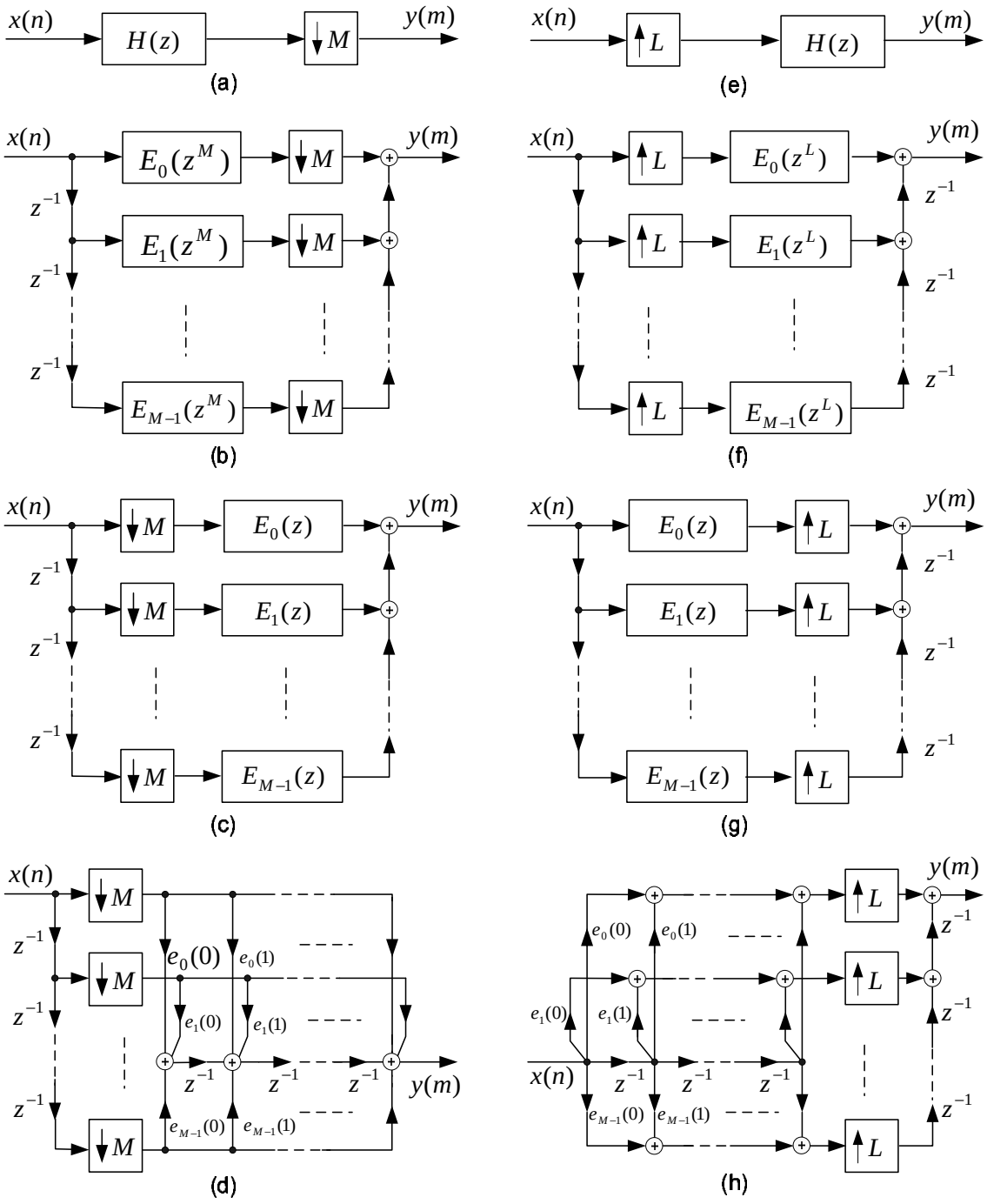

Fig. 2.6: M-fold decimation filter implemented based on (a) direct form, (b) polyphase decomposition, (c) polyphase decomposition applying the first cascade equivalence, (d) polyphase decomposition using shared delay elements. L-fold interpolation filter implemented based on (e) direct form, (f) polyphase decomposition, (g) polyphase decomposition applying the second cascade equivalence, (h) polyphase decomposition using shared delay elements.

elements can be shared among the polyphase components to hold the intermediate sum values. Therefore, the total storage requirement for data storage, as well as the computation rate, is reduced by a factor of M.

Transposing the structure of the polyphase M-fold decimation shown in Fig. 2.6(c), the L-fold interpolation structure is obtained as shown in Fig. 2.6(g), where M is replaced by L. Again the filtering operation of the polyphase components occurs at the lower-sampling rate side of the system. In comparison with the structure of Fig. 2.6(e), the computation rate is reduced by a factor of L. If each of the polyphase components is realized by a direct form FIR filter, as shown in Fig. 2.6(h), the same delay elements for holding the delayed values of $x(n)$ can be shared among the polyphase components. Therefore, the total data storage is also reduced by a factor of L.

2.2 Two-Channel Filter Banks

Decimating the signal gives rise to aliasing distortion. Bandlimiting the signal by a decimation filter may minimize aliasing distorting but leads to a loss in information content. Digital filter banks provide a way to get around this difficulty.

A digital filter bank is a set of digital bandpass filters with either a common input or a summed output. The filters are chosen such that a signal can be split into subband components and then decimated. During signal encoding, different bit rates are allocated to signals in different subbands depending on various criteria such as energy content, perceptual effects, etc. This is the basic principle of subband codding. The subband signals are then decoded and reconstructed to give the full band signal.

2.2.1 Basic Operation of a Two-Channel Filter Bank

The analysis/synthesis scheme used in most subband coding [21, 28, 74, 78] systems is maximal decimation, i.e., the decimation factor is equal to the number of bands

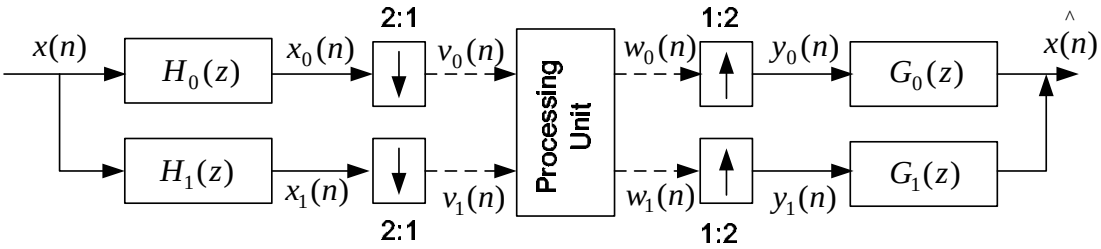

Fig. 2.7: Two-channel filter bank.

of the filter bank. Fig. 2.7 shows a two-channel filter bank. In subband processing, the input signal $x(n)$ is first filtered by two filters $H_0(z)$ and $H_1(z)$, which are the low-pass and high-pass filters, respectively. The subband signals are then decimated by a factor of two and encoded for transmission. At the receiver end, the subband signals are decoded, interpolated, and filtered by the filters $G_0(z)$ and $G_1(z)$ and then summed to produce the output signal $\hat{x}(n)$. $H_0(z)$ and $H_1(z)$ are called the analysis filters, whereas $G_0(z)$ and $G_1(z)$ are the synthesis filters. This analysis/synthesis system, however, may introduce three separate types of distortions: aliasing, amplitude distortion and phase distortion, which cause the reconstructed signal $\hat{x}(n)$ to differ from $x(n)$.

Consider the system shown in Fig. 2.7. Let the z-transforms of $x(n)$, $x_0(n)$, $x_1(n)$, $v_0(n)$, $v_1(n)$, $w_0(n)$, $w_1(n)$, $y_0(n)$, $y_1(n)$ and $\hat{x}(n)$ be $X(z)$, $X_0(z)$, $X_1(z)$, $V_0(z)$, $V_1(z)$, $W_0(z)$, $W_1(z)$, $Y_0(z)$, $Y_1(z)$ and $\hat{X}(z)$, respectively. Hence,

$$X_k(z) = H_k(z)X(z), \quad \text{for } k = 0, 1. \tag{2.9}$$

By using (2.2) and (2.9), $V_k(z)$'s, are expressible as

$$V_k(z) = \frac{1}{2}\left[H_k(z^{\frac{1}{2}})X(z^{\frac{1}{2}}) + H_k(-z^{\frac{1}{2}})X(-z^{\frac{1}{2}})\right], \quad \text{for } k = 0, 1. \tag{2.10}$$

Assume that the processing unit in Fig. 2.7 is lossless. By using (2.5) and (2.10), the output of the 2-fold expanders are given by

$$Y_k(z) = W_k(z^2) = V_k(z^2) = \frac{1}{2}\Big[X_k(z) + X_k(-z)\Big], \quad \text{for } k = 0, 1, \tag{2.11}$$

and the overall output is given by

$$\hat{X}(z) = G_0(z)Y_0(z) + G_1(z)Y_1(z). \tag{2.12}$$

The general relation between $\hat{X}(z)$ and $X(z)$, thus, is given by:

$$\begin{aligned}\hat{X}(z) &= \frac{1}{2}\Big[H_0(z)G_0(z) + H_1(z)G_1(z)\Big]X(z) \\ &\quad + \frac{1}{2}\Big[H_0(-z)G_0(z) + H_1(-z)G_1(z)\Big]X(-z) \\ &= \frac{1}{2}H(z)X(z) + \text{aliasing term},\end{aligned} \tag{2.13}$$

where

$$H(z) = H_0(z)G_0(z) + H_1(z)G_1(z). \tag{2.14}$$

2.2.2 Aliasing-Free QMF Banks

It was first shown by Croisier, et al [19] in the mid senventies that the aliasing problem in decimation-interpolation process can be completely eliminated by requiring that all of the analysis and synthesis filters involved be either scaled version or frequency shifted scaled versions of the same half-band lowpass filter. Such aliasing-free two-channel analysis/syntheis system is popularly called the Quadrature Mirror Filter (QMF) bank.

The second term of (2.13) represents the aliasing term. For aliasing free reconstruction, the second term of (2.13) must be zero, i.e.,

$$H_0(-z)G_0(z) + H_1(-z)G_1(z) = 0. \tag{2.15}$$

Choosing $G_0(z)$ and $G_1(z)$ as in (2.16) and (2.17) will satisfy (2.15).

$$G_0(z) = H_1(-z) \tag{2.16}$$

$$G_1(z) = -H_0(-z). \tag{2.17}$$

Thus, the overall transfer function $H(z)$ becomes

$$H(z) = H_0(z)H_1(-z) - H_0(-z)H_1(z). \tag{2.18}$$

In addition, (2.19) ensures that $H_1(z)$ is highpass if $H_0(z)$ is lowpass.

$$H_1(z) = H_0(-z). \tag{2.19}$$

Substituting (2.16), (2.17) and (2.19) into (2.14), the overall transfer function of the alias-free system is given by

$$H(z) = \frac{1}{2}\left[H_0^2(z) - H_0^2(-z)\right]. \tag{2.20}$$

2.2.3 Perfect Reconstruction Orthogonal Filter Banks

Design techniques for QMF bank were later developed by other authors to minimize the remaining distortions [3, 17, 22, 36–38]. It was independently observed by Mintzer [59] and Smith and Barnwell [73] that all the three distortions mentioned above can be eliminated and thus it results in exact reconstruction of the input signal.

For the above QMF class of analysis/synthesis system, perfect reconstruction requires that

$$H_0^2(z) - H_0^2(-z) = 2. \tag{2.21}$$

However, the perfect reconstruction condition of (2.21) leads to either the trivial case where $H_0(z) = 1 + z^{-1}$, $H_1(z) = 1 - z^{-1}$, or a pair of infinitely long, ideal half-band filters. Nevertheless, it will be shown that if (2.19) is relaxed, perfect reconstruction is possible without the above shortcomings.

From (2.13), it is clear that distortionless reconstruction is achieved for the class of filters which satisfies the condition

$$H_0(z)G_0(z) + H_1(z)G_1(z) = 2z^{-L}, \tag{2.22}$$

$$H_0(-z)G_0(z) + H_1(-z)G_1(z) = 0, \tag{2.23}$$

where L is a nonzero integer. Solving the simultaneous equations of (2.22) and (2.23) yields (2.24) and (2.25).

$$G_0(z) = \frac{2H_1(-z)z^{-L}}{H_0(z)H_1(-z) - H_0(-z)H_1(z)}, \tag{2.24}$$

$$G_1(z) = \frac{-2H_0(-z)z^{-L}}{H_0(z)H_1(-z) - H_0(-z)H_1(z)}. \tag{2.25}$$

Constraining both the analysis filters and the synthesis filters to be FIR filters, the denominator of (2.24) and (2.25) must satisfy (2.26).

$$H_0(z)H_1(-z) - H_0(-z)H_1(z) = Kz^{-N} \tag{2.26}$$

for some values of N and K. A solution meeting the requirement of (2.24), (2.25) and (2.26) is

$$G_0(z) = H_1(-z) \tag{2.27}$$

$$G_1(z) = -H_0(-z). \tag{2.28}$$

A new class of filters called conjugate quadrature filters (CQF) which satisfy (2.29) is introduced.

$$H_1(z) = -H_0(-z^{-1})z^{-N}. \tag{2.29}$$

Assuming that N is odd, substituting (2.27), (2.28) and (2.29) into (2.14), the overall transfer function of the analysis/synthesis system, which is free of aliasing, is given by

$$\begin{aligned} H(z) &= \frac{1}{2}\left[H_0(z)H_0(z^{-1}) + H_0(-z)H_0(-z^{-1})\right] z^{-N} \\ &= \frac{1}{2}\left[F_0(z) + F_0(-z)\right] z^{-N}, \end{aligned} \tag{2.30}$$

where $F_0(z)$ is called the product filter given by

$$F_0(z) = H_0(z)H_0(z^{-1}). \tag{2.31}$$

It is obvious from (2.30) that, for perfect reconstruction, the product filter $F_0(z)$ must meet two conditions. First, $F_0(z)$ must meet (2.32)

$$F_0(z) + F_0(-z) = 2, \tag{2.32}$$

i.e., $F_0(z)$ is a half-band filter. Second, $F_0(z)$ must be decomposable into analysis and synthesis filters in such a way that (2.31) is valid. Under these conditions, perfect reconstruction is achieved with a delay of N samples.

The design procedure, therefore, is divided into two steps: first, a half band product filter $F_0(z)$ is designed to meet the condition (2.32); second, the product filter is decomposed into $H_0(z)$ and $H_0(z^{-1})$ as shown in (2.31).

Analysis/synthesis filters obtained by this procedure are no longer quadrature mirror symmetric. This class of analysis/synthesis systems is called orthogonal filter banks recently [20].

Later studies [79] showed that PR orthogonal filter banks are a special case of the PR two-channel filter banks. From (2.18), it is obvious that, for an aliasing-free two-channel filter bank, given:

$$H_0(z) = \sum_{n=0}^{N_0} h_0(n) z^{-n}, \quad \text{and} \quad H_1(z) = \sum_{n=0}^{N_1} h_1(n) z^{-n},$$

the PR condition, $H(z) = z^{-L}$, is met provided that the impulse response of

$$F(z) = H_0(z)H_1(-z) = \sum_{n=0}^{N_0+N_1} f(n) z^{-n}, \tag{2.33}$$

satisfies

$$f(n) = \begin{cases} \frac{1}{2} & \text{for } n = L \\ \\ 0 & \text{for } n \text{ odd and } n \neq L, \end{cases}$$

when L is odd and $(N_0 + N_1)$ is even [20, 79].

2.2.4 Perfect Reconstruction Lattice Orthogonal Filter Banks

The implementation of a perfect reconstruction filter bank using the tap delay line structure suffers from the disadvantage that the perfect reconstruction property is affected by coefficient quantization. A lattice analysis/systhesis system, which structurally ensures perfect reconstruction, was introduced by Vaidyanathan and Hoang in [81]. The analysis bank is shown in Fig. 2.8. An important virtue of the lattice structure filter bank is that the perfect reconstruction property is preserved even under severe coefficient quantization. Since the perfect reconstruction property is structurally ensured, it is only necessary to consider the frequency response when the coefficient values are optimized in discrete value space.

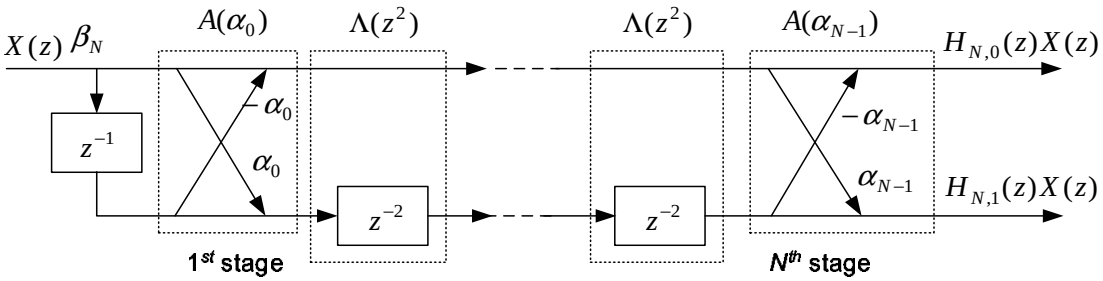

Fig. 2.8: Analysis bank of the perfect reconstruction lattice orthogonal filter bank.

Consider a $(2N-1)$-th order lattice structure (with N coefficients) implementing the analysis bank shown in Fig. 2.8. Let the z-transform transfer functions of the two channels be $H_{N,0}(z)$ and $H_{N,1}(z)$, respectively. Thus,

$$\begin{bmatrix} H_{N,0}(z) \\ \\ H_{N,1}(z) \end{bmatrix} = \beta_N \, A(\alpha_{N-1}) \, \Lambda \, A(\alpha_{N-2}) \, \Lambda \, \cdots \Lambda \, A(\alpha_k) \, \Lambda \, \cdots \Lambda \, A(\alpha_0) \begin{bmatrix} 1 \\ \\ z^{-1} \end{bmatrix}, \quad (2.34)$$

where

$$\beta_N^2 = \frac{1}{2} \prod_{k=0}^{N-1} \frac{1}{1+\alpha_k^2}, \qquad A(\alpha_k) = \begin{bmatrix} 1 & -\alpha_k \\ \\ \alpha_k & 1 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 1 & 0 \\ \\ 0 & z^{-2} \end{bmatrix}. \quad (2.35)$$

In Fig. 2.8, β_N appears as a scaling amplifier at the input. This is purely for the convenience of simplifying Fig. 2.8. In actual implementation, β_N may be factored and the factors distributed in between the lattice stages to optimize for roundoff noise performance.

It has been proved that the lattice structure of Fig. 2.8 satisfies the "power complementary property"

$$\left|H_{N,0}(e^{j\omega})\right|^2 + \left|H_{N,1}(e^{j\omega})\right|^2 = 1, \quad (2.36)$$

and the conjugate quadrature condition in equation (2.29). Conditions (2.29) and (2.36) ensure perfect reconstruction.

To determine the lattice coefficients of the filter bank, only the stopband energy of $H_{N,0}(e^{j\omega})$ should be considered, since the lattice structure ensures that the

stopband energy of $H_{N,1}(e^{j\omega})$ is equal to that of $H_{N,0}(e^{j\omega})$ and is automatically minimized. Moreover, a good stopband of $H_{N,0}(e^{j\omega})$ ensures a good passband of $H_{N,1}(e^{j\omega})$, and vice versa.

A minimax sense weighted least squares objective function is given by

$$f = \sum_{\omega \in [\omega_s, \pi]} B(\omega) |H_{N,0}(e^{j\omega})|^2. \tag{2.37}$$

In (2.37), $B(\omega)$ is the error weighting function and ω_s is the stopband edge. The stopband edge satisfies the constraint $\frac{\pi}{2} < \omega_s \leq \pi$.

2.3 Signed Power-of-Two Coefficient Design Issues

FIR digital filters designed over the signed power-of-two (SPT) discrete space was first proposed by Lim and Constantinides [49]. Extensive research has shown that the complexity of an FIR digital filter can be reduced by implementing its coefficients as sums of SPT terms. This section briefly describes the SPT number characteristics and existing optimization techniques for the design of digital filters subject to SPT coefficients.

2.3.1 Signed Power-of-Two Numbers

A number, Y, can be represented to a precision 2^Q by $L - Q$ trinary digits $y(i)$ according to

$$Y = \sum_{i=Q}^{L-1} y(i) 2^i, \quad y(i) \in \{\bar{1}, 0, 1\}, \quad Q \leq i \leq L-1, \tag{2.38}$$

where, $\bar{1}$ is equal to -1, L and Q are integers. A number represented in such a way is called an SPT number in this thesis. Each nonzero digit term , $y(i) \neq 0$, is counted as an SPT term. The wordlength of Y is $(L-Q)$-bit. Y is discrete values in increments of 2^Q in the range

$$-2^L + 2^Q \leq Y \leq 2^L - 2^Q, \tag{2.39}$$

in which there are $2^{L-Q+1}-1$ distinct values. However, with $L-Q$ digits each having 3 possibilities, there are 3^{L-Q} representations. For $L-1>Q$, 3^{L-Q} is larger than $2^{L-Q+1}-1$ and hence some numbers have more than one representation. A minimum representation refers to a representation requiring the minimum number of non-zero digits, i.e., minimum number of SPT terms, of which one number may also has more than one representation. A canonic representation is the unique minimum representation requiring $y(i)$ satisfying the constraints

$$y(i)y(i+1)=0, \tag{2.40}$$

i.e., there are no two SPT terms that are adjacent.

The canonic representation requirement imposes a further constraint on the representation; this will exclude the representation of some numbers that can be represented without the canonic representation requirement under the same wordlength. For example, with $L=4$ in (2.38), one cannot represent say 12 in the canonic form but it is possible in the non-canonic form. To represent 12 in a canonic form, it is necessary to increase L to 5 which is a drawback compared to the case with $L=4$ that allows non-canonic forms. However, because of the unique representation of the canonic SPT number which is very attractive for monitoring and ensuring the minimum representation of the number, many researches on the SPT coefficient design imposed the canonic condition on the SPT numbers for easy analysis and derivation [26, 34, 55, 67, 84], although there may exist other minimum representations.

For the particular condition where $Q=0$, the number Y is the set of all integers with magnitude less than $\frac{2^{L+1}}{3}$ when the canonic constraint is imposed on the number. For the particular condition where $L=0$, the number Y lies in the range $-1<Y<1$.

Since in canonic SPT representation, no two consecutive $y(i)$'s are non-zero, an R-bit Y can be represented using no more than $\frac{R+1}{2}$ SPT terms. Often, fewer terms are needed, and it has been shown in [65] that the expected number of SPT terms

in an R-bit canonic SPT number tends asymptotically to $(\frac{R}{3} + \frac{1}{9})$ as R increases.

A number represented in two's complement format can be easily converted to an equivalent canonic SPT representation as follows:

Let

$$X = (x_{R-1}, \cdots, x_1, x_0) \tag{2.41}$$

be an R-bit two's complement number and

$$Y = (y_{R-1}, \cdots, y_1, y_0) \tag{2.42}$$

be the equivalent SPT number, where $x_i \in \{0, 1\}$ and $y_i \in \{\bar{1}, 0, 1\}$ for $i = 0, \cdots, R-1$. The numerical value of X is the same as that for Y. For every digit x_i, y_i is generated using the following algorithm [65].

1. Initialize $i = 0$ and $\gamma_{-1} = x_{-1} = 0$. Arbitrarily define x_R as $x_R = x_{R-1}$.

2. Let $\theta_i = x_i \oplus x_{i-1}$.

3. Let $\gamma_i = \overline{\gamma_{i-1}}\theta_i$.

4. $y_i = (1 - 2x_{i+1})\gamma_i$.

5. If $i = R - 1$, stop; otherwise increment i and go to Step 2. □

In the above algorithm, the symbol $\oplus$ denotes exclusive OR and the overbar indicates complementation.

The following algorithm finds the best approximation $[x]_u$ for a number x using u SPT terms [50]:

1. Initialize $m = 1$ and $s_0 = x$.

2. Find $y(m)2^{g(m)}$ which minimizes $\left|s_{m-1} - y(m)2^{g(m)}\right|$.

3. If either $y(m) = 0$ or $m = u$, go to Step 6. Otherwise go to Step 4.

4. Update $s_m = s_{m-1} - y(m)2^{g(m)}$.

5. $m = m + 1$. Go to Step 2.

6. $[x]_u = \sum_{i=1}^{m} y(i)2^{g(i)}$. Stop. □

2.3.2 Existing Optimization Techniques

As analyzed in Section 2.3.1, when a number is represented as a sum of SPT terms, it has less or equal nonzero digits than when it is represented in two's complement. More interestingly, preliminary studies show that only a limited number of SPT terms are required to meet a respectable set of specifications if a good optimization technique exists. Hence, the coefficient multipliers can be replaced by a small number of add/subtract-shift operations. The hardware complexity as well as power consumption is therefore very much reduced.

Many methods have been developed for optimizing the frequency response of a digital filter subject to SPT constrains imposed on its coefficient values. These include the use of mixed-integer linear programming (MILP) [48, 49, 52], local search methods [67, 86], tree search with weighted least-squares criteria [45, 53], simulated annealing [5], genetic algorithm [26, 46], quantization guided by coefficient sensitivity analysis [10, 72], and optimization techniques incorporating SPT terms allocation strategies [15, 47, 55].

In MILP, linear programming is coupled with a suitable branch-and-bound search algorithm, such as the isocost search or depth-first search. The depth-first branch-and-bound search is often preferred for high-order filter design since the isocost search may not be able to produce a solution because of insufficient computing resources. The filters obtained by using MILP are optimized in the minimax sense. So far, MILP is the only known method which can guarantee global optimality in the minimax sense for a given SPT term allocation. Furthermore, MILP can minimize the total number of SPT terms if the problem is appropriately formulated [33], thus leading to a filter with minimal implementation cost. However, MILP requires excessive computing resources if the filter length is long. The computational cost required increases exponentially with the number of variables to be optimized.

Local search methods involve searching in the discrete space in the vicinity of the optimum continuous coefficient space. The solution obtained is a local optimum. Popular local search methods are the univariate or bivariate local search. Univariate local search employs the simplest neighborhood search technique by perturbing each coefficient one-at-a-time and increasing or decreasing one discrete step at a time. In the more powerful bivariate local search, coefficients are selected two-at-a-time (over all possible pairs). Each of the two selected coefficients is perturbed by one step leading to four combinations. The univariate and bivariate local search take trivial computing resources to arrive at a discrete solution in the immediate neighborhood of the continuous optimum solution but will require great computing resources to arrive at a good local optimum (if it ever happen) at a considerable distance away from the continuous optimum.

There are reports on replacing the linear programming algorithm in tree search method by a suitable weighted least-squares algorithm in the design of certain types of filters. In such algorithms, the filter's coefficient values are quantized one at a time. The remaining unquantized coefficients are optimized in the weighted least-squares sense. The computing time required is approximately proportional to the cube of the number of filter coefficients to be optimized but the optimal solution is not guaranteed.

Simulated annealing (SA) and genetic algorithm (GA) belong to the class of stochastic optimization techniques. SA is based on random moves and has some ability to overcome local optimums found on the way as it moves towards a better local optimum by accepting unfavourable move at a certain probability. Since SA is inherently suited for continuous space optimization, proper discretization of the variables' values is necessary. In contrast to SA, GA is inherently suited for discrete space optimization. It is a simulation of the evolution process of natural selection, where variables are encoded and the fitter ones have more probabilities to survive and produce offsprings. Both SA and GA are global optimization techniques in theory, but the computation cost is very expensive.

In the quantization guided by coefficient sensitivity analysis technique, each coefficient is first set to its nearest single-SPT-term number. The second-SPT-term is then allocated to the filters' coefficients one at a time in decreasing order of the coefficient sensitivity, until the frequency response meets the given specification. Coefficient sensitivity is defined as the sum of the increase in the peak passband ripple value and the increase in the peak stopband ripple value when the coefficient is set to its nearest single-SPT-term number. A modified sensitivity criterion considers the average ripple magnitude changes over all the frequency grid points in the passband and stopband.

Several authors reported techniques for optimizing FIR filters subject to a given total number of SPT terms; each coefficient value may have a different number of SPT terms. In such quantization scheme, the quantization step size is non-uniform. In [55], each coefficient of the filter is allocated a certain number of SPT terms according to the coefficient's statistical quantization step-size and sensitivity. After the assignment of the SPT terms, MILP is used to optimize the coefficient values. In [47], SPT terms are dynamically allocated to the currently most deserving coefficient, one at a time, to minimize the L^∞ distance between the SPT coefficients and their corresponding infinite wordlength values. In [15], each coefficient is first assigned SPT terms using the technique of [47]. Subsequently, a pool of SPT terms is created for each coefficient according to the coefficient's infinite precision value. A dynamic programming technique is used to allocate SPT terms taken from the coefficient's pool of SPT terms to each coefficient.

Among the above techniques, the local search approach and the GA has been applied to design the SPT coefficient lattice filter banks.

2.3.3 SPT term allocation

There are several schools of thoughts for the distribution of the SPT terms to the coefficients. Some researchers design filters where each coefficient is allocated with

the same number of SPT terms. Some researchers design filters where each coefficient is allocated with different number of SPT terms but the total number of SPT terms for the entire filter is fixed. It has been demonstrated in [55] that filters with different number of SPT terms allocated to each coefficient have significantly better frequency response performance than filters with the same number of total SPT terms but with all coefficients allocated with the same number of SPT terms. Both schools of thoughts have their own respective merits; it depends on the hardware platform used to implement the filters. For example, in a given implementation platform where only a fixed number of shifters are provided to each coefficients, the filter coefficient have to be optimized with the constraint that each coefficient is allocated with the number of SPT terms not more than the given number of shifters. If the implementation platform does not have such constraint on the number of shifters for each coefficient, the filter coefficients can be optimized under a total number of SPT terms to minimize the filter complexity.

Chapter 3

Successive Reoptimization Approach

OPTIMIZATION TECHNIQUES for the design of a transversal FIR filter with SPT coefficient values subject to a given frequency response requirement have been reviewed in Section 2.3.2. Unfortunately, many of the existing optimization techniques are not suitable for the design of lattice filters due to the lattice filters' special properties. For example, MILP cannot be used since the objective function for the lattice filter design is not a linear function of the lattice filter's coefficient values; in lattice orthogonal filter banks, the lattice coefficients cannot be scaled freely as those in transversal FIR filters due to the nonlinear property. The optimization techniques proposed, up to now, for the design of two-channel lattice orthogonal filter banks with SPT coefficients are mainly local search methods [34] and genetic algorithm [75].

In this chapter, as well as in the following two chapters, several methods for the design of SPT coefficient two-channel lattice orthogonal filter banks will be introduced. In this chapter, a successive reoptimization approach is proposed. Section 3.1 presents a weighted least squares algorithm for the design of the continuous coefficient filter banks, since generally a discrete coefficient design starts from a continuous optimum. The detailed successive reoptimization algorithm is introduced in Section 3.2.

3.1 Continuous Coefficient Filter Bank Design

A weighted least squares objective function which will produce minimax optimum results was given in (2.37). For easy reference, the objective function is reproduced in (3.1).

$$f = \sum_{\omega \in [\omega_s, \pi]} B(\omega) |H_{N,0}(e^{j\omega})|^2. \tag{3.1}$$

If the error weighting function $B(\omega)$ is uniform over all frequencies ω, minimizing (3.1) is a least squares optimization problem. A quasi-Newton approach for optimizing the least squares problem will be described in Section 3.1.1. An efficient line search algorithm which is necessary in the quasi-Newton approach is introduced in Section 3.1.2. The minimax sense minimum is achieved by Lim-Lee-Chen-Yang weighting function updating algorithm [29, 51], which will be reviewed in Section 3.1.3 for completeness.

3.1.1 The Least Squares Approach

The approach to minimize the least squares objective function is an iterative procedure. To design a $(2N-1)$-th order lattice filter, an N by 1 vector of lattice coefficients, $\boldsymbol{\alpha}$, at the p-th iteration is defined as

$$\boldsymbol{\alpha}^{(p)} = \left[\, \alpha_0^{(p)}, \alpha_1^{(p)}, \cdots, \alpha_{N-1}^{(p)} \, \right]^T. \tag{3.2}$$

The updating equation is

$$\boldsymbol{\alpha}^{(p+1)} = \boldsymbol{\alpha}^{(p)} + \gamma^{(p)} \boldsymbol{s}(\boldsymbol{\alpha}^{(p)}), \tag{3.3}$$

where, $\boldsymbol{s}(\boldsymbol{\alpha}^{(p)})$ is a search direction in the N dimensional space of $\boldsymbol{\alpha}$, and $\gamma^{(p)}$ (a positive scaler at the p-th iteration) is selected so as to minimize f in the $\boldsymbol{s}(\boldsymbol{\alpha}^{(p)})$ direction.

Let the objective function value f at the p-th iteration be denoted by $f^{(p)}$. In this proposed algorithm, a quasi-Newton method is employed [66]. In this method, the search direction, $\boldsymbol{s}(\boldsymbol{\alpha}^{(p)})$, is given by

$$\boldsymbol{s}(\boldsymbol{\alpha}^{(p)}) = -\mathbf{H}^{(p)} \nabla_{\boldsymbol{\alpha}} f^{(p)}, \tag{3.4}$$

where $\nabla_{\boldsymbol{\alpha}} f^{(p)}$ is the derivative of the objective function f at the p-th iteration with respect to $\boldsymbol{\alpha}^{(p)}$. $\mathbf{H}^{(p)}$ is an $N \times N$ matrix which is an approximation to the inverse of the Hessian, $\mathcal{H}^{-1} = \{\nabla^2 f_{opti}\}^{-1}$, where f_{opti} is the optimum value of f. The computation of the second order derivative of the objective function is, thus, avoided.

Denoting $\nabla_{\boldsymbol{\alpha}} f^{(p)}$ by $\boldsymbol{g}^{(p)}$, $\Delta\boldsymbol{\alpha}^{(p)}$ and $\Delta\boldsymbol{g}^{(p)}$ are given by

$$\Delta\boldsymbol{\alpha}^{(p)} = \boldsymbol{\alpha}^{(p+1)} - \boldsymbol{\alpha}^{(p)}, \tag{3.5}$$

$$\Delta\boldsymbol{g}^{(p)} = \boldsymbol{g}^{(p+1)} - \boldsymbol{g}^{(p)}. \tag{3.6}$$

The matrix $\mathbf{H}$ is updated by using BFGS method [23],

$$\begin{aligned}\mathbf{H}^{(p)} &= \mathbf{H}^{(p-1)} + \frac{\Delta\boldsymbol{\alpha}^{(p-1)}\left(\Delta\boldsymbol{\alpha}^{(p-1)}\right)^T}{\left(\Delta\boldsymbol{\alpha}^{(p-1)}\right)^T\Delta\boldsymbol{g}^{(p-1)}}\left(1 + \frac{\left(\Delta\boldsymbol{g}^{(p-1)}\right)^T\mathbf{H}^{(p-1)}\Delta\boldsymbol{g}^{(p-1)}}{\left(\Delta\boldsymbol{\alpha}^{(p-1)}\right)^T\Delta\boldsymbol{g}^{(p-1)}}\right) \\ &- \frac{\Delta\boldsymbol{\alpha}^{(p-1)}\left(\Delta\boldsymbol{g}^{(p-1)}\right)^T\mathbf{H}^{(p-1)} + \mathbf{H}^{(p-1)}\Delta\boldsymbol{g}^{(p-1)}\left(\Delta\boldsymbol{\alpha}^{(p-1)}\right)^T}{\left(\Delta\boldsymbol{\alpha}^{(p-1)}\right)^T\Delta\boldsymbol{g}^{(p-1)}},\end{aligned} \tag{3.7}$$

and $\mathbf{H}^{(0)}$ is initialized to the identity matrix, $\mathbf{I}$.

In the optimization of the lattice filter coefficients, the frequency response of the lowpass analysis filter $H_{N,0}^{(p)}(e^{j\omega})$ is obtained using (2.34). The derivative of the objective function, $\nabla_{\boldsymbol{\alpha}} f^{(p)}$, is given by

$$\nabla_{\boldsymbol{\alpha}} f^{(p)} = 2\sum_{\omega\in\Omega} \Re\Big\{B(\omega)H_{N,0}^{(p)}(e^{j\omega})\nabla_{\boldsymbol{\alpha}} H_{N,0}^{(p)}(e^{j\omega})\Big\}, \tag{3.8}$$

where $\Re$ denotes the real part of the complex value, $\nabla_{\boldsymbol{\alpha}} H_{N,0}^{(p)}(e^{j\omega})$ is the derivative of $H_{N,0}^{(p)}(e^{j\omega})$ with respect to $\boldsymbol{\alpha}^{(p)}$, and $\Omega = [\omega_s, \pi]$. For zero-phase frequency response, $\nabla_{\boldsymbol{\alpha}} H_{N,0}^{(p)}(e^{j\omega})$ is given by

$$\nabla_{\boldsymbol{\alpha}} H_{N,0}^{(p)}(e^{j\omega}) = \Big[\nabla_{\alpha_0} H_{N,0}^{(p)}(e^{j\omega}), \cdots, \nabla_{\alpha_k} H_{N,0}^{(p)}(e^{j\omega}), \cdots, \nabla_{\alpha_{N-1}} H_{N,0}^{(p)}(e^{j\omega})\Big]^T. \tag{3.9}$$

In (3.9), $\nabla_{\alpha_k} H_{N,0}^{(p)}(e^{j\omega})$ is given by

$$\begin{bmatrix} \nabla_{\alpha_k} H_{N,0}^{(p)}(e^{j\omega}) \\ \\ \nabla_{\alpha_k} H_{N,1}^{(p)}(e^{j\omega}) \end{bmatrix} = \beta_N^{(p)} A\left(\alpha_{N-1}^{(p)}\right) \Lambda A\left(\alpha_{N-2}^{(p)}\right) \Lambda \cdots \Lambda A\left(\alpha_{k+1}^{(p)}\right) \Lambda$$

$$\frac{1}{1+\left(\alpha_k^{(p)}\right)^2} A\left(\alpha_k^{(p)}\right) \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \Lambda A\left(\alpha_{k-1}^{(p)}\right) \Lambda \cdots \Lambda A\left(\alpha_0^{(p)}\right) \begin{bmatrix} 1 \\ e^{-j\omega} \end{bmatrix} \quad (3.10)$$

where

$$\beta_N^{(p)} = \left(\frac{1}{2} \prod_{k=0}^{N-1} \frac{1}{1+\left(\alpha_k^{(p)}\right)^2} \right)^{\frac{1}{2}}, \quad A\left(\alpha_k^{(p)}\right) = \begin{bmatrix} 1 & -\alpha_k^{(p)} \\ \alpha_k^{(p)} & 1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 1 & 0 \\ 0 & e^{-j2\omega} \end{bmatrix} \quad (3.11)$$

The iterative procedure starts with an arbitrary vector $\boldsymbol{\alpha}^0$ and $B(\omega) = 1$ for all ω. Subsequently, $\boldsymbol{\alpha}^{(p+1)}$ is obtained using (3.3). The value of $\boldsymbol{s}(\boldsymbol{\alpha}^{(p)})$ is obtained using (3.4) and $\gamma^{(p)}$ is obtained by using an efficient line search procedure described in Section 3.1.2. The iterative process terminates when

$$|\gamma^{(p)} \boldsymbol{s}(\boldsymbol{\alpha}^{(p)})| < \xi, \quad (3.12)$$

where ξ is a predefined error tolerance. After (3.12) is satisfied, $B(\omega)$ is updated by using the Lim-Lee-Chen-Yang algorithm [51], which will be reviewed in Section 3.1.3.

3.1.2 A Line Search Algorithm

In the proposed line search algorithm, given the lattice coefficients $\boldsymbol{\alpha}^{(p)}$ and search direction $\boldsymbol{s}(\boldsymbol{\alpha}^{(p)})$ at the p-th iteration, an initial guess of γ_1 is obtained. The procure is carried out as follows:

1. Evaluate f for $\boldsymbol{\alpha}^{(p)}$ and denote the obtained value by f_0.

   Set $\gamma_1 = 100 \times \min\left(\boldsymbol{\alpha}^{(p)} ./ \boldsymbol{s}(\boldsymbol{\alpha}^{(p)})\right)$, where './' denotes element by element division.

   If $\gamma_1 > 2^{-10}$, set $\gamma_1 = 2^{-10}$.

2. Evaluate f for γ_1 and denote the obtained value by f_1.

3. If $f_1 < f_0$, go to Step 4. Otherwise, go to Step 5.

4. Set $\gamma_0 = \gamma_1$, $f_0 = f_1$.

   Replace γ_1 by $2\gamma_1$.

   Evaluate f for γ_1 and denote the obtained value by f_1.

   If $f_1 < f_0$, go to Step 4. Otherwise, set $\gamma^{(p)} = \gamma_0$ and stop.

5. Set $\gamma_1 = \gamma_1/2$.

   Evaluate f for γ_1 and denote the obtained value by f_1.

   If $f_1 > f_0$, go to Step 5. Otherwise, set $\gamma^{(p)} = \gamma_1$ and stop. □

This line search algorithm only requires the objective function values, therefore, practically it is simple and fast. From the experiences, if the step size obtained using the above algorithm is larger than 2, for some examples, the results may deviate from the optimum solution. Therefore, in Step 4, a judgement of whether γ_0 is larger than 2 can be included to restrict the step size, $\gamma^{(p)}$, not to be larger than 2. A comparison of the proposed line search algorithm with Fletcher's line search algorithm [1, 23] is tabulated in Table 3.1. In Table 3.1, the two-channel lattice filter bank with stopband edges at $\omega_s = 0.52\pi, \omega_s = 0.54\pi$ and $\omega_s = 0.58\pi$ is optimized in the least squares sense. The filter length related parameter, N, ranges from 20 to 36.

From Table 3.1, it can be seen that the convergent time of the proposed algorithm is approximately one third of that of Fletcher's algorithm, whereas the stopband attenuation achieved by the two algorithms are consistent with the exception for the example for $\omega_s = 0.58\pi, N = 32$, where the stopband attenuation of the proposed algorithm is 1.82 dB inferior to that of the Fletcher's algorithm.

3.1.3 Lim-Lee-Chen-Yang Algorithm

In the previous two sections, iterative procedures have been presented to derive an optimal $\boldsymbol{\alpha}$ which minimizes the objective function defined in (3.1) for a given weighting function. An appropriate weighting function can be derived such that the optimal $\boldsymbol{\alpha}$ with respect to (3.1) is also optimal in the weighted minimax sense.

Table 3.1: A comparison of the proposed line search algorithm with Fletcher's line search algorithm.

	Time(sec)/Stopband attenuation(dB)					
	$\omega_s = 0.52\pi$		$\omega_s = 0.54\pi$		$\omega_s = 0.58\pi$	
N	Proposed	Fletcher's	Proposed	Fletcher's	Proposed	Fletcher's
20	$<$ 1/09.49	1/09.49	$<$ 1/18.96	2/18.96	1/40.89	3/40.49
24	1/11.18	2/11.18	1/23.03	3/23.01	3/49.72	8/49.72
28	1/12.96	5/12.96	2/27.19	7/27.18	4/58.62	9/58.14
32	3/14.81	9/14.81	5/31.43	14/31.43	7/65.76	22/67.58
36	5/16.73	14/16.73	8/35.72	24/35.72	16/76.59	50/76.59

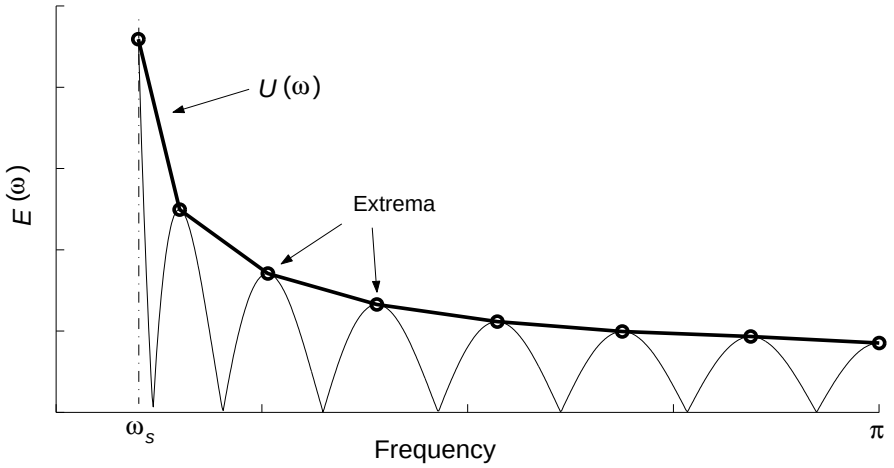

Fig. 3.1: An example of the error function.

This section reviews the Lim-Lee-Chen-Yang weighting function update algorithm.

The error function at the q-th iteration of updating the weighting function, $B(\omega)$, is given by

$$E_q(\omega) = H_{0,N}^2(e^{j\omega}), \quad \text{for } \omega_s \leq \omega \leq \pi, \tag{3.13}$$

where ω_s is the stopband edge of $H_{0,N}(e^{j\omega})$. An example of $E_q(\omega)$ is shown in Fig. 3.1.

Assume that the extrema of $E_q(\omega)$, indicated by symbol '$\circ$' in Fig. 3.1, occur at frequency grid points ω_i for $1 \leq i \leq N_e$, where N_e is the total number of extrema. These frequency grid points are labeled consecutively so that $\omega_1 < \omega_2 < \cdots < \omega_{N_e}$.

$E_q(\omega_s)$ and $E_q(\pi)$ are also considered as extrema. The values of the i-th extremum in $E_q(\omega)$ at the q-th iteration is denoted as

$$V_q(i) = E_q(\omega_i). \tag{3.14}$$

For any non-band-edge $V_q(i)$ less than 0.1 of $\min\Big(V_q(i-1), V_q(i+1)\Big)$, let $V_q(i) = 0.1 \times \min\Big(V_q(i-1), V_q(i+1)\Big)$. For $V_q(i)$ where ω_i is a band-edge frequency, let $V_q(i) = \max\Big(V_q(i), 0.1 \times V_q(i')\Big)$, where $V_q(i')$ is the neighboring extremum within the same frequency band, i.e., either $V_q(i+1)$ or $V_q(i-1)$ as appropriate.

Based on this set of extrema $V_q(i)$, an envelope function $U_q(\omega)$ is defined as follows:

$$U_q(\omega) = \frac{\omega - \omega_i}{\omega_{i+1} - \omega_i} V_q(i+1) + \frac{\omega_{i+1} - \omega}{\omega_{i+1} - \omega_i} V_q(i), \quad \omega_i \le \omega \le \omega_{i+1}, \tag{3.15}$$

where $1 \le i \le N_e - 1$. An example is shown in Fig. 3.1, where the envelope function is illustrated as the thickened line joining the extreme points.

The weighting function $B_q(\omega)$ is updated as follows:

$$B_{q+1}(\omega) = B_q(\omega) \left[\frac{U_q(\omega)}{\hat{U}_q}\right]^{\theta}. \tag{3.16}$$

In (3.16), $\hat{U}_q$ is given by

$$\hat{U}_q = \frac{\sum_{\omega \in [\omega_s, \pi]} U_q(\omega)}{N_\omega}, \tag{3.17}$$

where N_ω is the number of frequency grid points. The factor θ affects the convergence and convergent rate of this algorithm. An appropriate value is $\theta = 1.4$, as proposed in [54].

The weighting function update operation is terminated when

$$\frac{\max V_q(i) - \min V_q(i)}{\max V_q(i) + \min V_q(i)} < \epsilon, \tag{3.18}$$

where ϵ is a small constant which specifies the desired "flatness" of the envelope function, $U_q(\omega)$. Typically, (3.18) is achieved in about 6 to 8 iterations of updating the weighting function for the cases where $\epsilon = 0.01$.

The Lim-Lee-Chen-Yang algorithm updates the weighting function using the envelope of the ripple magnitude. It converges many times faster than Lawson's

algorithm [44] which updates the weighting function using the magnitude of the deviation.

The weighted least squares algorithm for the design of two-channel orthogonal lattice filter bank can now be summarized as follows:

1. Set $p = q = 0$. Initialize the lattice coefficient $\alpha_k^{(p)} = -(-1/2)^k$, for $0 \leq k \leq N-1$. Set $B_q(\omega) = 1$, $\omega \in [\omega_s, \pi]$.

2. Evaluate the search direction $\boldsymbol{s}(\boldsymbol{\alpha}^{(p)})$ according to (3.4).

3. Obtain $\gamma^{(p)}$ by using the line search procedure described in Section 3.1.2.

4. Update $\boldsymbol{\alpha}^{(p+1)}$ according to (3.3).

5. If (3.12) is satisfied, go to step 6, otherwise, set $p = p + 1$ and go to step 2.

6. Update $B_{q+1}(\omega)$ according to (3.16).

7. If (3.18) is satisfied, stop. Otherwise, set $\alpha^{(0)} = \alpha^{(p)}$, $p = 0$, $q = q + 1$ and go to step 2. □

3.2 Successive Reoptimization Approach

In this section, a lattice coefficient sensitivity analysis is first performed. It is shown that the coefficient sensitivities differ greatly from coefficient to coefficient. Based on this observation, in this technique, the coefficient values are quantized sequentially one at a time. After each coefficient is being quantized, the remaining unquantized coefficient values are reoptimized to partially compensate for the frequency response deviation caused by the quantization of that value. The order of selection of the coefficients for quantization is based on the coefficient sensitivity measure. Coefficients with higher sensitivity measures are quantized earlier than coefficients with lower sensitivity measures.

3.2.1 Coefficient Sensitivity Analysis

For the lattice structure PR orthogonal filter bank, the sensitivities of $H_{N,0}(z)$ and $H_{N,1}(z)$ with respect to the coefficient α_k denoted by $P_{N,k}(z)$ and $Q_{N,k}(z)$, respectively, are given by

$$\begin{bmatrix} P_{N,k}(z) \\ Q_{N,k}(z) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial H_{N,0}(z)}{\partial \alpha_k} \\ \dfrac{\partial H_{N,1}(z)}{\partial \alpha_k} \end{bmatrix} = \beta_N \, A(\alpha_{N-1}) \, \Lambda \, A(\alpha_{N-2}) \, \Lambda \cdots \Lambda A(\alpha_{k+1}) \Lambda$$
$$\frac{1}{1+\alpha_k^2} A(\alpha_k) \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \Lambda A(\alpha_{k-1}) \Lambda \cdots \Lambda A(\alpha_0) \begin{bmatrix} 1 \\ z^{-1} \end{bmatrix}, \tag{3.19}$$

and can be rewritten as

$$\begin{aligned} \begin{bmatrix} P_{N,k}(z) \\ Q_{N,k}(z) \end{bmatrix} &= \frac{\beta_N}{\beta_k} \frac{1}{1+\alpha_k^2} \, A(\alpha_{N-1}) \, \Lambda \cdots \Lambda A(\alpha_{k+1}) \Lambda A(\alpha_k) \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \Lambda \begin{bmatrix} H_{k,0}(z) \\ H_{k,1}(z) \end{bmatrix} \\ &= \frac{\beta_N}{\beta_k} \frac{1}{1+\alpha_k^2} A(\alpha_{N-1}) \Lambda \cdots \Lambda A(\alpha_{k+1}) \Lambda A(\alpha_k) \begin{bmatrix} -z^{-2} H_{k,1}(z) \\ H_{k,0}(z) \end{bmatrix}. \end{aligned} \tag{3.20}$$

From (3.20), the following can be obtained:

$$P_{N,k}(z)\widetilde{P}_{N,k}(z) + Q_{N,k}(z)\widetilde{Q}_{N,k}(z) = \frac{1}{(1+\alpha_k^2)^2} < 1, \qquad \text{for } \alpha_k \neq 0, \tag{3.21}$$

where $\widetilde{P}_{N,k}(z)$ and $\widetilde{Q}_{N,k}(z)$ denote the conjugate of $P_{N,k}(z)$ and $Q_{N,k}(z)$, respectively.

Proof: This is proved by mathematical induction. From (3.20), for $n = k+1$, $P_{k+1,k}(z)$ and $Q_{k+1,k}(z)$ are given by

$$\begin{bmatrix} P_{k+1,k}(z) \\ Q_{k+1,k}(z) \end{bmatrix} = \frac{\beta_{k+1}}{\beta_k} \frac{1}{1+\alpha_k^2} A(\alpha_k) \begin{bmatrix} -z^{-2} H_{k,1}(z) \\ H_{k,0}(z) \end{bmatrix}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{1+\alpha_k^2}} \cdot \frac{1}{1+\alpha_k^2} \begin{bmatrix} 1 & -\alpha_k \\ \alpha_k & 1 \end{bmatrix} \begin{bmatrix} -z^{-2}H_{k,1}(z) \\ H_{k,0}(z) \end{bmatrix} \\
&= \frac{1}{\sqrt{(1+\alpha_k)^3}} \begin{bmatrix} -z^{-2}H_{k,1}(z) - \alpha_k H_{k,0}(z) \\ -z^{-2}\alpha_k H_{k,1}(z) + H_{k,0}(z) \end{bmatrix}. \qquad (3.22)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& P_{n,k}(z)\widetilde{P}_{n,k}(z) + Q_{n,k}(z)\widetilde{Q}_{n,k}(z) \\
= \; & \frac{1}{(1+\alpha_k^2)^3}\Big[\Big(-z^{-2}H_{k,1}(z) - \alpha_k H_{k,0}(z)\Big)\cdot\Big(-z^2\widetilde{H}_{k,1}(z) - \alpha_k\widetilde{H}_{k,0}(z)\Big) \\
& + \Big(-z^{-2}\alpha_k H_{k,1}(z) + H_{k,0}(z)\Big)\Big(-z^2\alpha_k\widetilde{H}_{k,1}(z) + \widetilde{H}_{k,0}(z)\Big)\Big] \\
= \; & \frac{1+\alpha_k^2}{(1+\alpha_k^2)^3}\Big[H_{k,0}(z)\widetilde{H}_{k,0}(z) + H_{k,1}(z)\widetilde{H}_{k,1}(z)\Big]. \qquad (3.23)
\end{aligned}
$$

Since $H_{k,0}$ and $H_{k,1}$ satisfy the power-complement image condition [81], i.e.,

$$H_{k,0}(z)\widetilde{H}_{k,0}(z) + H_{k,1}(z)\widetilde{H}_{k,1}(z) = 1. \qquad (3.24)$$

Hence, (3.21) holds for $n = k+1$.

Suppose that (3.21) is true for $n = m$. Thus,

$$
\begin{bmatrix} P_{m,k}(z) \\ Q_{m,k}(z) \end{bmatrix} = \frac{\beta_m}{\beta_k}\frac{1}{1+\alpha_k^2}\, A(\alpha_{m-1})\,\Lambda\cdots\Lambda A(\alpha_{k+1})\Lambda A(\alpha_k) \begin{bmatrix} -z^{-2}H_{k,1}(z) \\ H_{k,0}(z) \end{bmatrix}, \qquad (3.25)
$$

and

$$P_{m,k}(z)\widetilde{P}_{m,k}(z) + Q_{m,k}(z)\widetilde{Q}_{m,k}(z) = \frac{1}{(1+\alpha_k^2)^2} < 1 \qquad \text{for } \alpha_k \neq 0. \qquad (3.26)$$

Then, for $n = m+1$,

$$
\begin{bmatrix} P_{m+1,k}(z) \\ Q_{m+1,k}(z) \end{bmatrix} = \frac{\beta_{m+1}}{\beta_k(1+\alpha_k^2)} A(\alpha_m)\Lambda\cdots\Lambda A(\alpha_k) \begin{bmatrix} -z^{-2}H_{k,1}(z) \\ H_{k,0}(z) \end{bmatrix}
$$

$$= \frac{1}{\sqrt{1+\alpha_m^2}} A(\alpha_m)\Lambda \begin{bmatrix} P_{m,k}(z) \\ \\ Q_{m,k}(z) \end{bmatrix}$$

$$= \frac{1}{\sqrt{1+\alpha_m^2}} \begin{bmatrix} P_{m,k}(z) - \alpha_m z^{-2} Q_{m,k}(z) \\ \\ \alpha_m P_{m,k}(z) + z^{-2} Q_{m,k}(z) \end{bmatrix}. \tag{3.27}$$

It can be shown that

$$\begin{aligned} & P_{m+1,k}(z)\widetilde{P}_{m+1,k}(z) + Q_{m+1,k}(z)\widetilde{Q}_{m+1,k}(z) \\ = \; & \frac{1}{1+\alpha_k^2}\Big[\Big(P_{m,k}(z) - \alpha_m z^{-2} Q_{m,k}(z) Bigg\Big) \cdot \Big(\widetilde{P}_{m,k}(z) - \alpha_m z^2 \widetilde{Q}_{m,k}(z)\Big) \\ & + \Big(\alpha_m P_{m,k}(z) + z^{-2} Q_{m,k}(z)\Big)\Big(\alpha_m \widetilde{P}_{m,k}(z) + z^{-2}\widetilde{Q}_{m,k}(z)\Big)\Big] \\ = \; & P_{m,k}(z)\widetilde{P}_{m,k}(z) + Q_{m,k}(z)\widetilde{Q}_{m,k}(z). \end{aligned} \tag{3.28}$$

Hence, if (3.21) is true for $n = m$, it is also true for $n = m+1$. Since (3.21) is true for $n = k+1$, it is true for all integer n, $k+1 \le n \le N$. ■

From (3.20) and (3.21), it can be shown that

$$|P_{N,k}(z)| = \left|\frac{\partial H_{N,0}(z)}{\partial \alpha_k}\right| \le \frac{1}{1+\alpha_k^2} < 1. \tag{3.29}$$

The coefficient sensitivity $|P_{N,k}(e^{j\omega})|$ is thus bounded by $\frac{1}{1+\alpha_k^2}$; in general, it is a function of frequency. To give an idea on the relative values of (a) the peak absolute value of $P_{N,k}(e^{j\omega})$, (b) the average of the absolute value of $P_{N,k}(e^{j\omega})$ in the stopband, (c) α_k, and (d) $\frac{1}{1+\alpha_k^2}$, these quantities for a particular 27-th order filter bank are tabulated in Table 3.2. It is interesting to note from Table 3.2 that the coefficient sensitivity increases with increasing k.

3.2.2 Coefficient Quantization Algorithm

The technique starts with the design of the optimum continuous coefficient value minimax PR lattice orthogonal filter bank using the weighted least squares algorithm described in Section 3.1.

Table 3.2: Coefficient values and stopband coefficient sensitivities of a 27-th order PR orthogonal filter bank.

k	α_k	$\frac{1}{1+\alpha_k^2}$	Peak sensitivity	Average sensitivity
0	−4.5978691	0.0451663	0.0451663	0.0361323
1	1.5020457	0.3071120	0.3071118	0.2567841
2	−0.8649152	0.5720568	0.5720565	0.5041978
3	0.5793815	0.7486807	0.7486804	0.6888595
4	−0.4115350	0.8551678	0.8551675	0.8116573
5	0.2979295	0.9184744	0.9184740	0.8904599
6	−0.2146259	0.9559642	0.9559638	0.9397591
7	0.1509795	0.9777132	0.9777128	0.9693764
8	−0.1018520	0.9897327	0.9897322	0.9860167
9	0.0645909	0.9958454	0.9958449	0.9944699
10	−0.0375630	0.9985910	0.9985905	0.9981955
11	0.0193403	0.9996261	0.9996259	0.9995473
12	−0.0082989	0.9999311	0.9999307	0.9999225
13	0.0026803	0.9999928	0.9999928	0.9999926

After the continuous optimum solution is obtained, a coefficient sensitivity analysis is performed. The coefficient with the highest coefficient sensitivity measure is selected and rounded to its nearest discrete value. All the other coefficients are then reoptimized to partially compensate for the frequency response deterioration due to the quantization of the selected coefficient. The rationale for selecting the most sensitive coefficient for quantization is as follows. The most sensitive coefficient will cause the largest frequency response deviation. Selecting it to be the first coefficient to be quantized will have the advantage that there are many other coefficient values which can be reoptimized to compensate for the frequency response deterioration caused by its quantization. On the contrary, if the most sensitive coefficient is quantized after all the other coefficients are quantized, its effect cannot be compensated for by adjusting other coefficient values.

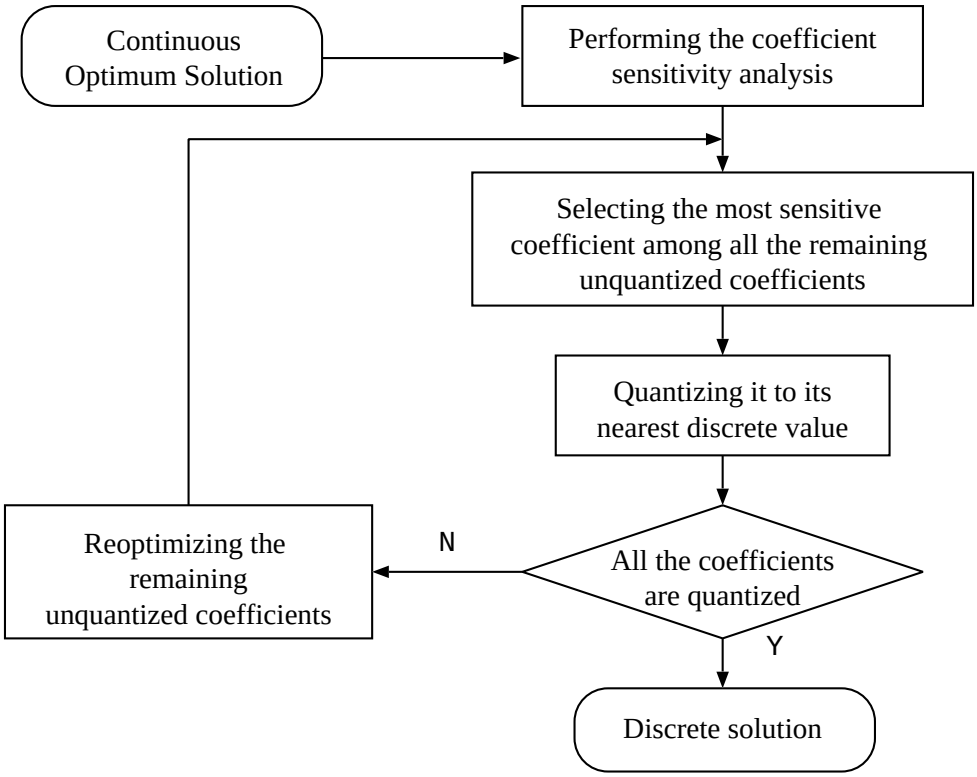

Fig. 3.2: A flowchart of the successive reoptimization procedure.

After fixing the selected coefficient and the remaining coefficient values reoptimized, the most sensitive coefficient among the remaining un-quantized coefficients is selected for quantization. The process of selecting the most sensitive coefficient among the unquantized coefficients for quantization and the reoptimization of all the unquantized coefficients after the quantization of each coefficient is repeated until all the coefficients are quantized. A flowchart of the successive reoptimization procedure is shown in Fig. 3.2.

In the proposed successive reoptimization approach, the procedure to reoptimize the remaining unquantized coefficients can be carried out as the procedure described in Section 3.1 by dropping out the quantized coefficients from the variable vector $\boldsymbol{\alpha}$.

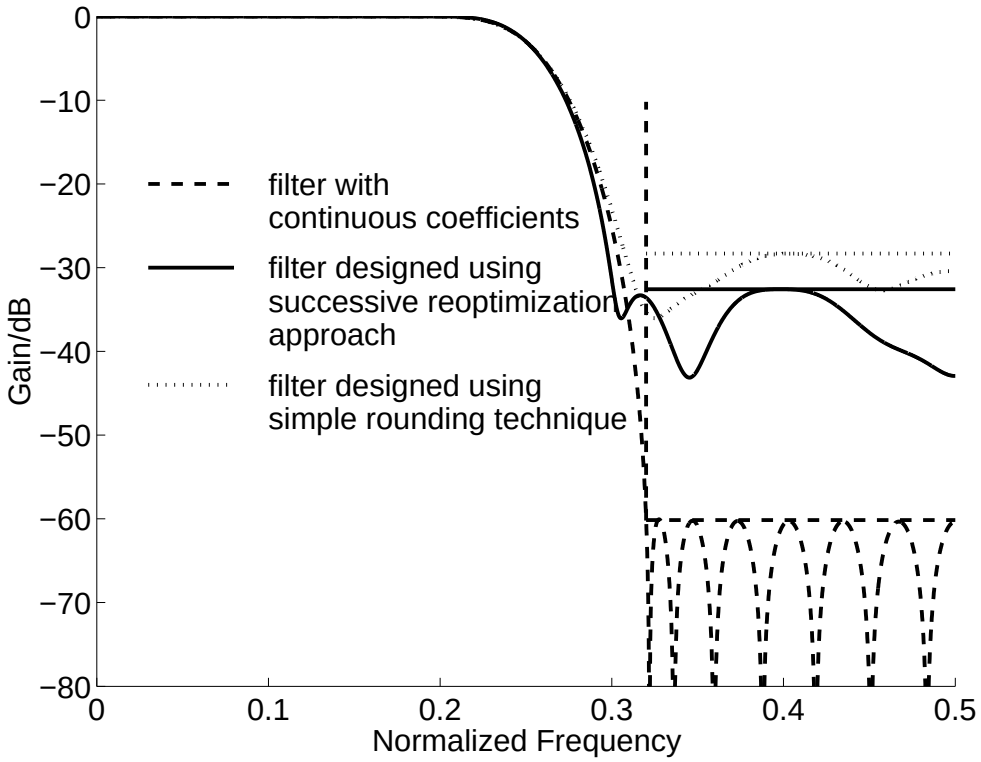

Fig. 3.3: Frequency response plots for the analysis lowpass filters. Each coefficient of the discrete coefficient design is represented by a sum of two signed power-of-two terms.

3.2.3 Design Example

A 27-th order (14 coefficients) filter bank is chosen as an example to illustrate the proposed technique. The lowpass filter's stopband edge is at 0.64π and its stopband frequency response is equiripple. Each coefficient value of the discrete coefficient is represented as a sum or difference of not more than two power-of-two terms; the smallest power-of-two term is 2^{-12}. The frequency responses of the lowpass filters for the (1) continuous coefficient optimum design, (2) SPT coefficient design obtained using the proposed successive reoptimization algorithm and (3) SPT coefficient solution obtained by simple rounding of coefficient values are shown in Fig. 3.3. The minimum stopband attenuation for the 3 designs in Fig. 3.3 are 60.20dB, 32.57dB and 28.32dB, respectively. The coefficient values of an SPT design obtained using the successive reoptimization approach are listed in Table. 3.3. From Fig. 3.3, it is obvious that the stopband attenuation of the discrete space

Table 3.3: Discrete coefficient values of a 27-th order PR orthogonal filter bank obtained by using the successive reoptimization approach. The stopband edge is at $\omega_s = 0.64\pi$.

k	SPT coefficients
0	$-2^{+3}+2^{+1}$
1	$2^{+1}-2^{-2}$
2	$-2^{+0}+2^{-3}$
3	$2^{-1}+2^{-5}$
4	$-2^{-1}+2^{-3}$
5	$2^{-2}+2^{-7}$
6	$-2^{-2}+2^{-4}$
7	$2^{-3}+2^{-6}$
8	$-2^{-3}+2^{-5}$
9	$2^{-4}-2^{-9}$
10	$-2^{-5}-2^{-8}$
11	$2^{-6}+2^{-9}$
12	-2^{-7}
13	$2^{-9}+2^{-11}$

design obtained using the successive reoptimization algorithm is superior to those obtained by simple rounding of coefficient values.

3.3 Conclusion

In this chapter, a successive reoptimization approach is presented for the design of SPT coefficient two-channel lattice orthogonal filter banks. A weighted least squares algorithm is employed to obtain the continuous coefficients; the continuous solution serves as the starting point for the SPT coefficient optimization process. The frequency responses of the filter banks obtained using the proposed successive reoptimization algorithm are significantly superior to those obtained by simple rounding of the coefficient values.

Chapter 4

Genetic Algorithm

RECENTLY, GENETIC ALGORITHMS [25,26,32,46,75,77] have emerged as a powerful and robust tool for the design of discrete coefficient digital filters. Genetic operations including reproduction, crossover and mutation are employed to minimize the frequency response error and the computational complexity [46,75]. In general, genetic operation will render the SPT representation of the coefficients non-canonic, i.e. the offsprings produced by the GA operations may no longer conform to the canonic SPT format. In [46], the offsprings are discarded if their coefficient values are not in the canonic SPT format. In [25] and [26], a technique was developed for restoring the canonic SPT numbers.

In this chapter, an improved genetic algorithm for the design of perfect reconstruction lattice orthogonal filter bank with canonic SPT coefficients is presented. First, an efficient encoding scheme for encoding the coefficient values of the filter is presented. In this encoding scheme, the canonic nature of the SPT coefficients is preserved during genetic operations. This is accomplished by encoding the canonic SPT numbers as the index of an SPT look-up table. Second, two new features which drastically improve the performance of GA are introduced. The new features are: (1) An additional level of natural selection is introduced to simulate the effect of natural selection when sperm cells compete to fertilize an ovule; this dramatically improves the offspring survival rate. Conventional GA is analogous to intracytoplasmic sperm injection and has an extremely low offspring survival rate resulting in very slow convergence. (2) The probability of mutation for each

codon of a chromosome is weighted by the reciprocal of its effect. The proposed GA approach proved to be highly effective and outperforms existing canonic SPT coefficient lattice orthogonal filter bank design algorithms.

4.1 The Genetic Algorithm

Genetic algorithms are optimization algorithms that simulate the evolution process of natural selection. When the genetic algorithm is used to optimize the coefficient values of a filter, the coefficients (which may be expressed in binary form) are concatenated to represent the chromosome of the filter. A prespecified number of filters are selected from the population pool and placed in a mating pool. Filters in the mating pool are paired up at random. The chromosomes of the two filters (the parents) in each pair are mixed at random to reproduce two new filters (the offsprings). Both the parent filters and either one or both of the offsprings are released into the population pool. The mixing of the chromosomes is called crossover. Mutation may be introduced into the reproduction process by randomly changing some of the binary bits forming the chromosomes. The size of the population is controlled by a natural selection process which has the tendency to reject inferior members of the population.

The breeding of one generation of filters is said to have completed when all the parents in the mating pool and their offsprings have been released into the population pool. After elimination by natural selection, the selection of a prespecified number of filters from the population pool to form members of the mating pool repeats. A second generation of offsprings is thus produced. The GA cycle is repeated until a desired termination criterion is reached. An example of a termination criterion is that a predefined number of generations is produced.

A detailed review of the genetic algorithms and their applications on the signal processing can be found in [77].

4.2 Filter Coefficient Encoding and Fitness Evaluation

One of the characteristics of a GA is the direct manipulation on the coded variables. This provides flexibility for solving different optimization problems. For an SPT coefficient, the most common encoding scheme is the ternary digit string [26,46,75] approach, i.e., using $0, 1\text{and} -1$ string to represent a coefficient value. However, such a digit string suffers from the problem that genetic operations may render the number of SPT terms for each coefficient to be non-minimum requiring more SPT terms than necessary. As the canonic representation ensures the use of the minimum number of SPT terms, the canonic representation is adopted. In this section, a binary digit encoding scheme which will ensure the canonic representation is developed to represent the SPT coefficients.

Let $S^+(L,K,Q)$ be a set of positive canonic SPT values that any $n \in S^+(L,K,Q)$ is a sum of no more than K canonic SPT terms and the largest power-of-two term is less than or equal to 2^{L-1}, whereas the smallest power-of-two term is larger than or equal to 2^Q, i.e.,

$$n = \sum_{i=0}^{m} y(i)2^{q(i)}, y(i) \in \{-1, 1\}, \tag{4.1}$$

where

$$Q \leq q(i) \leq L-1,$$

$$m = 0, 1, \ldots, K-1.$$

Furthermore, $q(i) \neq q(j)$ if $i \neq j$ and $q(i) \neq q(j)+1$ for any i, j. It is known that the number of elements of the set $S^+(L,K,Q)$, represented as $M^+(L,K,Q)$, is [55]

$$M^+(L,K,Q) = \sum_{m=1}^{K} \frac{2^{m-1}}{m!} \prod_{k=0}^{m-1} (L-Q-m+1-k). \tag{4.2}$$

Let $S(L,K,Q)$ be a set which is the union of $S^+(L,K,Q)$ and 0. Thus, $S(L,K,Q)$ represents the set of non-negative canonic SPT value with the same constraints as $S^+(L,K,Q)$. Let $M(L,K,Q)$ be the number of elements in the set $S(L,K,Q)$.

Table 4.1: Look-up table for $K = 2$, $Q = -2$ and $L = 2$.

Index	SPT value	Number of SPT terms
0	0	0
1	0.25	1
2	0.5	1
3	0.75	2
4	1	1
5	1.25	2
6	1.5	2
7	1.75	2
8	2	1
9	2.25	2
10	2.5	2

Therefore, $M(L, K, Q) = M^{+}(L, K, Q) + 1$. Thus, an $M(L, K, Q)$ element, three column look-up table can be established. In the look-up table, the elements in the first column are the set of integers [0, 1, 2, ..., $M(L, K, Q) - 1$] arranged in ascending order. The second column consists of the set of SPT values of $S(L, K, Q)$. The value in the third column is the number of SPT terms used in the corresponding SPT value. An example of a look-up table for $K = 2$, $Q = -2$ and $L = 2$ is listed in Table 4.1.

The proposed technique starts with the design of the optimum continuous coefficient value minimax lattice orthogonal filter bank using the weighted least squares algorithm described in Section 3.1. After the continuous optimum solution is obtained, each coefficient is allocated a number of SPT terms according to the SPT term allocation scheme reported in [55]. A continuous coefficient value, α_k, is quantized to the nearest discrete value of $|\alpha_k|$ whose number of SPT terms listed in the SPT coefficient look-up table is equal to the allocated number of SPT terms. The index, i, of the quantized $|\alpha_k|$, after attaching the sign of α_k, is expressed in two's complement form. This two's complement binary string forms a gene of

a coefficient. In order to avoid invalid genes, the number of bits used to encode the index and the sign should not be excessive. For example, in Table 4.1, only a maximum of 4 bits should be used. Concatenating all the genes together forms a row of binary bit stream which forms the chromosome in the proposed GA.

In the design of filters with SPT coefficients, besides the filter performance specifications, the smallest SPT term for the coefficient $\left(2^Q\right)$ and the total number of SPT terms, $\hat{K}$, for all the coefficients are usually pre-specified. The number of adders needed in the hardware implementation is equal to the total number of SPT terms minus one. It has been demonstrated [47,55] that significant advantage can be achieved if the coefficient values are allocated with different number of SPT terms while keeping the total number of SPT terms fixed.

The advantages of this look-up table binary encoding scheme are manifold. First, the SPT representation for each number is canonic. Second, it ensures that the magnitude of the smallest SPT term of the offspring of the genetic operations is not less than 2^Q. Third, it is easy to allocate the coefficients with different number of SPT terms since the number of SPT terms required by each value is already tabulated in the SPT look-up table. Last, the gene length is shorter than that using ternary digit string.

When evaluating the fitness of the chromosome, the signed integer index, i, represented as the encoded gene is recovered from its binary gene string. The $|i|$-th SPT value of the look-up table, after attaching the sign of i, is the coefficient value. During this decoding procedure, the number of SPT terms used in the coefficient is also obtained. The number of SPT terms will be used in the fitness measurement to control the total number of SPT terms.

The fitness of a chromosome is defined as

$$fitness = -20\log_{10}(f) - \mu \cdot p(\hat{k} - \hat{K}), \tag{4.3}$$

where

$$p(x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4.4}$$

and f is an objective function value of a minimax error criterion.

$$f = \max_{\omega \in [\omega_s, \pi]} \left| H_{N,0}(e^{jw}) \right|. \tag{4.5}$$

In (4.3), $\hat{K}$ is the pre-specified total number of SPT terms for all the coefficients, $\hat{k}$ is the total number of SPT terms actually used in all the coefficients, and μ is a positive weighting coefficient. From past experience, setting μ to be a value between 1 and 2 usually produces good results. The fitness measure of the population is defined as the best fitness measure among the members of the population.

4.3 Improved Genetic Operations

An initial population pool of filters is formed by perturbing the coefficient values of the filter whose coefficient values are obtained by rounding the optimum continuous coefficient values.

Filters are selected from the population pool using the Roulette Wheel selection procedure [32] to form members of the mating pool. Members of the mating pool are paired for mating to reproduce offsprings. The Roulette Wheel selection procedure ensures that fitter chromosomes have a higher chance of being selected as parents.

In the traditional GA approach, the chromosomes of two selected filters are mixed in the crossover process subject to a given probability of crossover to produce a pair of offsprings. The offsprings then compete with existing members of the population for survivor.

In the proposed algorithm, the two-point crossover process is arbitrarily adopted in forming the chromosomes of the offspring. In the two-point crossover approach, two points **A** and **B** are marked on each of the chromosomes involved in the crossover process as shown in Fig. 4.1. **A** and **B** are placed at random but they

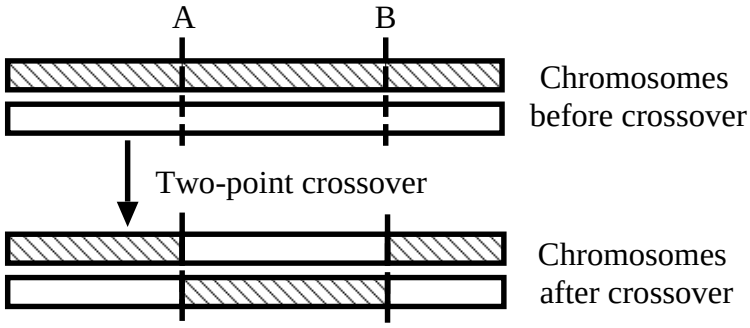

Fig. 4.1: Two-point crossover.

must be at corresponding codons of both chromosomes. They may be at both ends of the chromosomes. The sections of codons between **A** and **B** of the two chromosomes are swept to produce two new offsprings.

From past experience, the survivor rate of offsprings produced in the conventional GA is extremely low. This is because the crossover process is done by mixing the chromosomes at random and is analogous to intracytoplasmic sperm injection where the sperm cell is selected at random by a blindfolded gynaecologist. In a natural reproduction process, the sperm cells compete to fertilize an ovule; this competition introduces an additional level of natural selection to ensure the reproduction of healthy offsprings.

In the proposed algorithm, an additional natural selection process to ensure the competitiveness of the offspring produced is introduced. The chromosomes of the selected pair of filters are mixed at random (subject to a given probability of crossover) for a predefined number of times producing a large number of possible offsprings. The best two of these possible offsprings are selected as the legitimate offsprings of the mating; all other possible offsprings are discarded. Owing to this additional natural selection process, the survival rate of the offsprings is dramatically improved when competing with existing members of the population pool.

Accompanying the benefit of this additional natural selection process is the cost of an increase in the computation load. The bulk of this increase in computation load is the evaluation of the frequency responses of the possible offsprings. In

order to reduce this increase in computation load, the frequency responses of the possible offsprings are evaluated on a sparse frequency grid. The number of sparse grid points used is $\left\lceil (1 - \frac{\omega_s}{\pi}) \times 0.4 \times N \right\rceil + 1$, where N is the number of coefficients and $\lceil x \rceil$ is the smallest integer larger than or equal to x. (The number of grid points used to evaluate the frequency responses of the filters in the population pool is $\left\lceil (1 - \frac{\omega_s}{\pi}) \times 16 \times N \right\rceil + 1$). Although using the sparse frequency grid to evaluate the chromosome cannot ensure that the two offsprings selected are the fittest, it is a low cost approach to sift out those low performance candidates.

Mutation is an operator that introduces variations into the chromosome. The operation occurs with small probability. In the conventional GA, each bit has the same probability of mutation. This directly simulates the mutation of the living system. However, there are differences between a living organism and a filter system. In a living organism, a gene, which comprises the codons, determines or affects a single characteristic. All codons have equal importance and their function are not substitutable. In a filter, a gene, which comprises the digit bits, contributes to the frequency response in such a way that it is impossible to associate which gene contributes to which specific characteristic of the frequency response. All the genes work in coordination to determine the frequency response. Moreover, the importance of each bit in one gene is different from bit to bit and the effect due to mutating any bit can be partially compensated by mutating other bits.

In the proposed algorithm, when a chromosome is considered for mutation, the bits do not have equal mutation probability. The more significant bit has smaller mutation probability while the less significant bit has larger mutation probability. In the proposed approach, the mutation probability of the least significant bit of each gene is set to be $\frac{0.5}{N}$. The mutation probability of the second least significant bit of each gene is $\frac{(0.5)^2}{N}$ and so on such that the mutation probability for the n-th least significant bit is $\frac{(0.5)^n}{N}$. Thus, the product of mutation probability and the weight for each coefficient is kept constant.

4.4 Design Example

The design of the 27-th order (14 coefficients) filter bank specified in Section 3.2.3 is used as the example to illustrate the proposed algorithm. The specifications are repeated here, i.e. the lowpass filter's stopband edge is at 0.64π and its stopband frequency response is required to be equiripple; each coefficient value is represented as a sum of a limited number of signed power-of-two (SPT) terms; the smallest power-of-two term is 2^{-12}. The only difference between the specifications for this example and those in Section 3.2.3 is that in this example, the average number of SPT terms allocated to each coefficient is two, whereas not more than two SPT terms are allocated to each coefficient in Section 3.2.3.

In the GA operation, the population pool size and the mating pool size are set to be 1000 and 100, respectively. The crossover probability is set to be 0.8. The evolution process is terminated if the fitness measure of the population remains unchanged for 1000 generations. During each mating, P possible offsprings are evaluated and the best two are selected as the legitimate offsprings for introducing into the general population pool. The best results of the evolution process for $P = 40$, 20 and 2 as well as that for the conventional GA are plotted in Fig. 4.2. The conventional GA corresponds to unweighted probability of mutation for each bit and $P = 2$. The difference between the proposed algorithm for $P = 2$ and that for the conventional one is that for the proposed algorithm with $P = 2$, mutation probability for each bit is weighted whereas that for the conventional one it is unweighted. As can be seen from Fig. 4.2, the evolution processes for $P = 40, 20$ and 2 evolved to 45.97dB, 42.95dB and 42.40dB after 78, 76 and 612 generations, respectively. The conventional GA only evolved to 41.76dB after 340 generations.

The frequency responses of the lowpass filters for the (1) continuous coefficient optimum design, (2) SPT coefficient design obtained by the proposed GA and (3) SPT coefficient solution obtained by simple rounding of coefficient values are plotted in Fig. 4.3. The coefficients of the design obtained using the proposed GA is

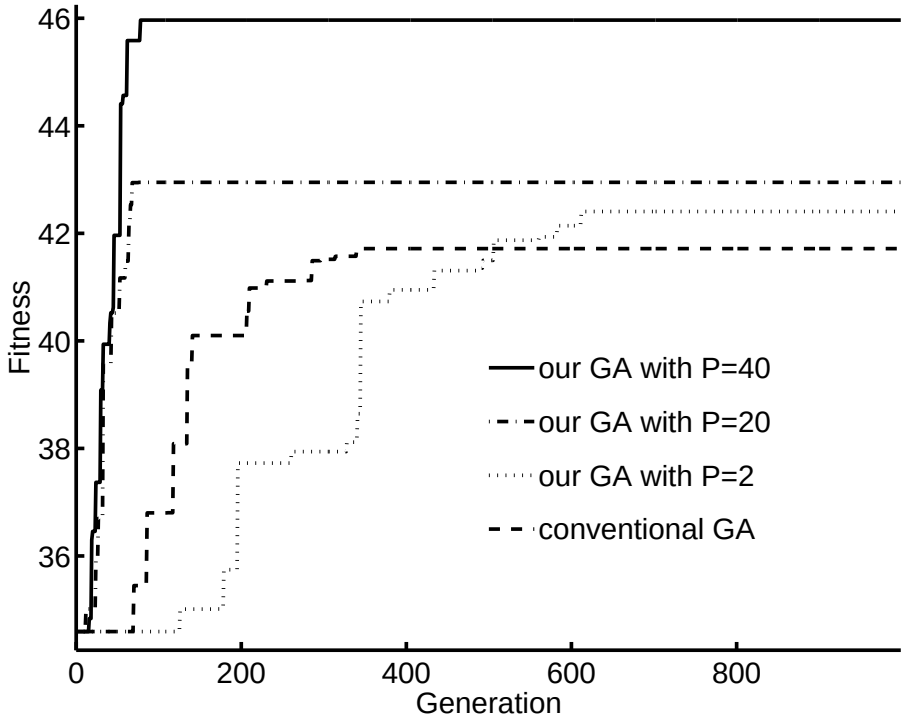

Fig. 4.2: The evolution process of the 27-th order example.

listed in Table 4.2. The superiority of the proposed new GA over the conventional GA is obvious from the results shown in Fig. 4.2.

Since GA's are essentially guided random search algorithms, the (sub)optimum solutions obtained and the number of generations needed to obtain the (sub)optimum solutions differ widely from run to run. For the 27-th order filter example, the average stopband attenuation achieved and the average number of generations averaging over ten runs for each GA are listed in Table 4.3. It can be seen that the average results are consistent with the best results shown in Fig. 4.2.

It should be noted that the frequency responses of the P offsprings produced from a selected pair of chromosomes are evaluated on a very sparse grid (one fortieth of the usual grid density). For $P = 40$, this increase in computation complexity is equivalent to that of the conventional GA whose population is scaled up by a factor of 1.5. For comparison, for the design of the 27-th order example, the best solution obtained using the proposed technique for $P = 40$ is 45.97dB in the stopband. Increasing the population by a factor of 1.5, the conventional

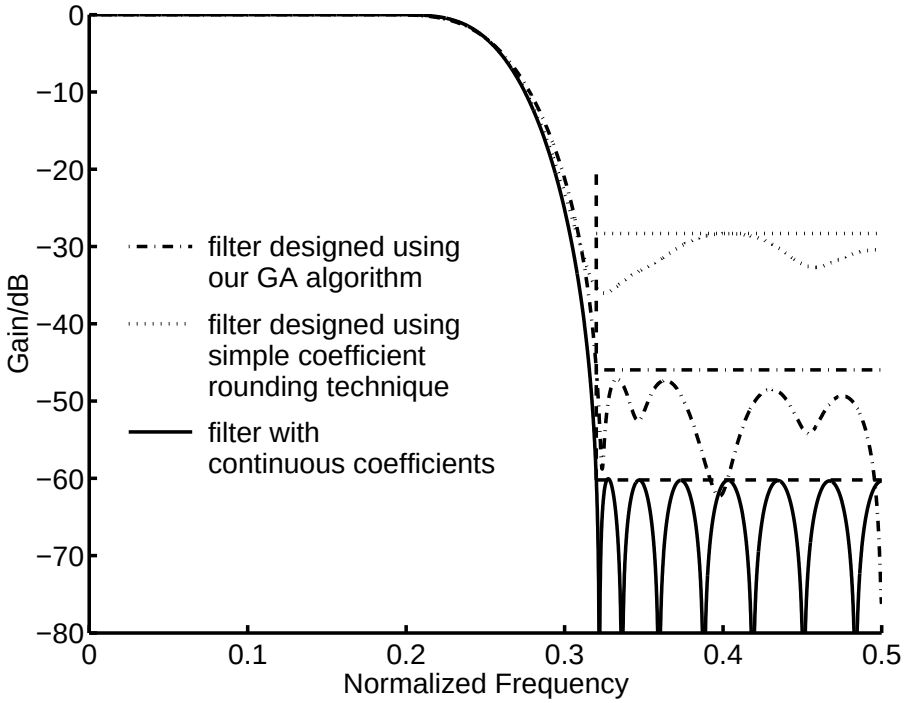

Fig. 4.3: The frequency response of the 27-th order example obtained using the improved GA, where the average number of SPT terms for each coefficient is two.

GA achieved a stopband attenuation of 42.95dB. The superiority of the proposed technique is evident.

4.5 Conclusion

In this chapter, an improved genetic algorithm for the design of SPT coefficient value lattice orthogonal filter bank has been developed. The design examples presented in Section 4.4 show that the speed of convergence for the proposed GA is faster than that for the conventional GA and the quality of the solution obtained using the proposed GA is also better than that produced by the conventional GA.

Table 4.2: Discrete coefficient values of the 27-th order orthogonal filter bank obtained using the proposed GA. The stopband edge is at $\omega_s = 0.64\pi$

k	SPT coefficients
0	-2^{+2}
1	$2^{+0}+2^{-2}+2^{-4}$
2	$-2^{+0}+2^{-2}$
3	$2^{-1}-2^{-7}$
4	$-2^{-1}+2^{-3}+2^{-5}+2^{-8}$
5	$2^{-2}-2^{-6}-2^{-10}$
6	$-2^{-3}-2^{-5}$
7	$2^{-3}-2^{-5}+2^{-7}-2^{-9}$
8	$-2^{-4}+2^{-8}$
9	$2^{-5}-2^{-9}$
10	$-2^{-6}+2^{-8}$
11	2^{-9}
12	0
13	0

Table 4.3: The average stopband attenuations and the number of generations needed by different GA's for the design of the 27-th order filter example.

GA's	Stopband attenuation	Generations
Conventional	39.09	671.7
$P = 2$	39.45	522.5
$P = 20$	39.85	137.7
$P = 40$	41.29	157.9

Chapter 5

Width-Recursive Depth-First Search

IN CHAPTER 3, a successive reoptimization approach was proposed for the design of discrete coefficient lattice orthogonal banks. However, there is still room for improvement. Two factors are considered in this chapter.

1) In the successive reoptimization approach, the order of selection of the coefficients for quantization is based on the coefficient sensitivity measure. Coefficients with higher sensitivity measures are quantized earlier than coefficients with lower sensitivity measures. It does not consider, however, the unevenly distributed grid space when the coefficients are optimized on the signed power-of-two space. Quantizing a coefficient with lower sensitivity, but located at sparse SPT value section may cause a larger frequency response deterioration. In order to overcome this difficulty, a new frequency response deterioration measure is proposed. The new frequency response deterioration measure includes the coefficient sensitivity as well as the grid density for the particular coefficient.

2) The selected coefficient, α_k, was rounded to its nearest discrete value. As the optimum value for α_k may be at a considerable distance from the infinite precision solution, the successive reoptimization technique may miss the optimum solution. It is necessary to assign several discrete values (in the vicinity of its continuous optimum value) to α_k. For each discrete value assigned to α_k, the remaining unquantized coefficients are reoptimized to partially compensate for the frequency

response deterioration due to the quantization of α_k. A tree search procedure should be developed to obtain the optimal (suboptimal) discrete coefficient values.

5.1 Frequency Response Deterioration Measure

The quantization of a coefficient α_k causes the value of α_k to be shifted by a small amount $\Delta\alpha_k$ from its continuous value. The frequency response deviation caused by the quantization of α_k is represented by $\Delta H^k_{N,0}(e^{j\omega})$, where $\Delta H^k_{N,0}(e^{j\omega})$ is given by

$$\Delta H^k_{N,0}(e^{j\omega}) = \frac{\partial H_{N,0}(e^{j\omega})}{\partial \alpha_k} \times \Delta\alpha_k = P_{N,k}(e^{j\omega}) \times \Delta\alpha_k, \tag{5.1}$$

provided that $\Delta\alpha_k$ is small.

The quantization of a coefficient with a higher sensitivity and a larger value of $|\Delta\alpha_k|$ causes a larger deviation to the frequency response. Thus, the product of coefficient sensitivity and quantization step size is an important measure on the frequency response deviation caused by quantizing a coefficient. A frequency response deterioration measure, $S_{N,k}$, given by

$$S_{N,k}(z) = P_{N,k}(z) \times G_k \tag{5.2}$$

is defined for the purpose of estimating the effect on the frequency response of the filter as a result of quantizing α_k. In (5.2), G_k is the grid spacing defined as the distance between the upper discrete level and the lower discrete level of a continuous coefficient α_k. The grid spacing is an indication of the quantization step size when the coefficient is actually quantized. A coefficient value changes from iteration to iteration as the optimization process proceeds. Thus, the grid spacing for a coefficient value changes from iteration to iteration in the case of a nonuniformly distributed coefficient space such as the power-of-two coefficient space.

The coefficient with the largest frequency response deterioration measure is selected to be quantized first. After the quantization of each coefficient, all the

other coefficients are then reoptimized to partially compensate for the frequency response deterioration due to the quantization of the selected coefficient.

If the discrete coefficient grid is evenly distributed such as that in the case of the integer grid where all coefficient values must be an integer after multiplying by a constant, the grid spacings are equal for all coefficients. In this case, $P_{N,k}(z)$ may be used instead of $S_{N,k}(z)$ for selecting a coefficient for quantization since the value of G_k for all k are equal.

5.2 Width-Recursive Depth-First Tree Search

The technique starts with the design of the optimum continuous coefficient value minimax PR orthogonal filter bank using the weighted least squares algorithm described in Section 3.1. After the continuous optimum solution is obtained, a coefficient α_k is selected for quantization. The method of selecting α_k has been discussed in Section 5.1. A straightforward method for assigning a discrete value to α_k is to round it to its nearest discrete value. As the optimum value for α_k may be at a considerable distance from the infinite precision solution, it is necessary to assign several discrete values (in the vicinity of its continuous optimum value) to α_k. For each discrete value assigned to α_k, the remaining unquantized coefficients are reoptimized to partially compensate for the frequency response deterioration due to the quantization of α_k. A tree search algorithm is then produced.

Before embarking on describing the novel tree search algorithm, two existing tree search algorithms are described briefly. The new algorithm is developed based on these two algorithms.

1) In the branch and bound depth first search algorithm [27], after the continuous optimum solution is obtained, a coefficient α_k is selected for branching. Suppose that α_2 is selected and that the integer space is the desired discrete coefficient space. Suppose also that the continuous optimum value of α_2 is 3.4. Two subproblem P_1 and P_2 are created by imposing the bounds $\alpha_2 \leq 3$ and $\alpha_2 \geq 4$,

respectively. See Fig. 5.1, problem P_2 is stored and P_1 is solved. Another coefficient (say α_1) is then selected for partitioning into two subproblems P_3 and P_4 by imposing bounds on the selected coefficient (say $\alpha_1 \leq 0$ and $\alpha_1 \geq 1$, respectively). P_4 is stored and P_3 is solved. P_3 is then further partitioned into P_5 and P_6. In a similar way, P_6 is stored and P_5 is solved. Suppose that P_5 yields a discrete solution. P_5 is then fathomed and P_6 is solved. In Fig. 5.1, a line underneath a node indicates that no further exploration from that node can be profitable. Such a node is said to be fathomed. If P_6 yields a discrete solution, P_6 is fathomed and the algorithm backtracks to P_3 and switches to solve P_4. The branching, backtracking and searching process continues until all the nodes are fathomed. The algorithm searches the tree in a depth-first manner and earns its name "depth-first" search.

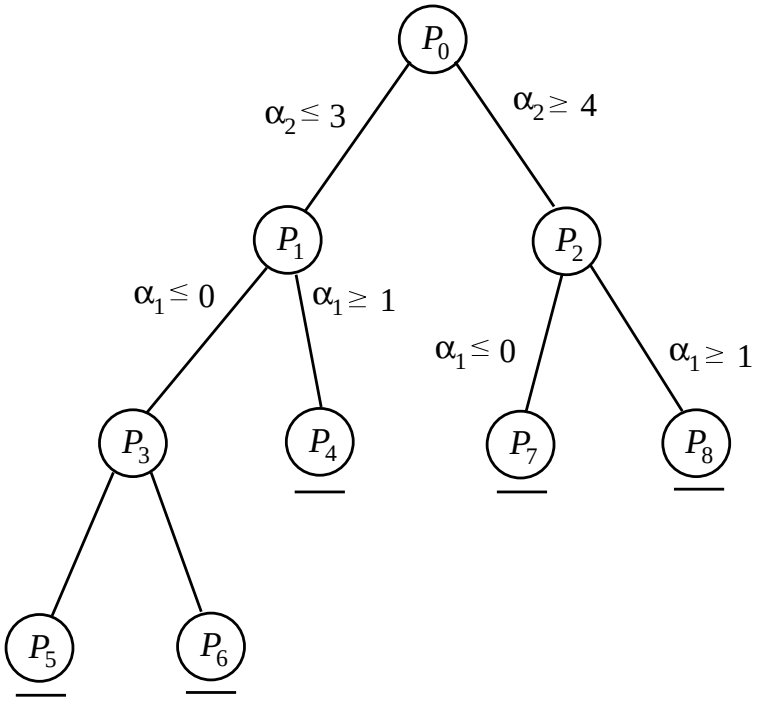

Fig. 5.1: An example of a Branch and Bound Tree.

2) A tree search technique which can yield a good suboptimal solution quickly was described in [53] for the design of filters subject to discrete coefficient constraint. In that technique, after obtaining the continuous coefficient value design, P_0, a coefficient α_k is selected and L discrete values are assigned to α_k. See Fig. 5.2.

This produces L optimization problems — an optimization problem for each discrete value of α_k. Fig. 5.2 shows the case where L is 3. After these L problems are solved, another coefficient is selected for quantization. Thus, each of the L problems produces L further optimization problems. Hence, there are L^2 problems when two coefficients are assigned discrete values. In order to limit the size of the tree, only L out of these L^2 problems are selected for further quantization of the coefficients; the rest are discarded. Each of the L problems selected from the L^2 problems produces L further optimization problems when a third coefficient is assigned discrete values. The process of selecting L problems from L^2 problems and the branching of each of the selected L problems into further L problems continues until all the coefficients are assigned discrete values. Increasing the value of L will increase the chance of obtaining the global optimum solution but will also increase the computer time requirement.

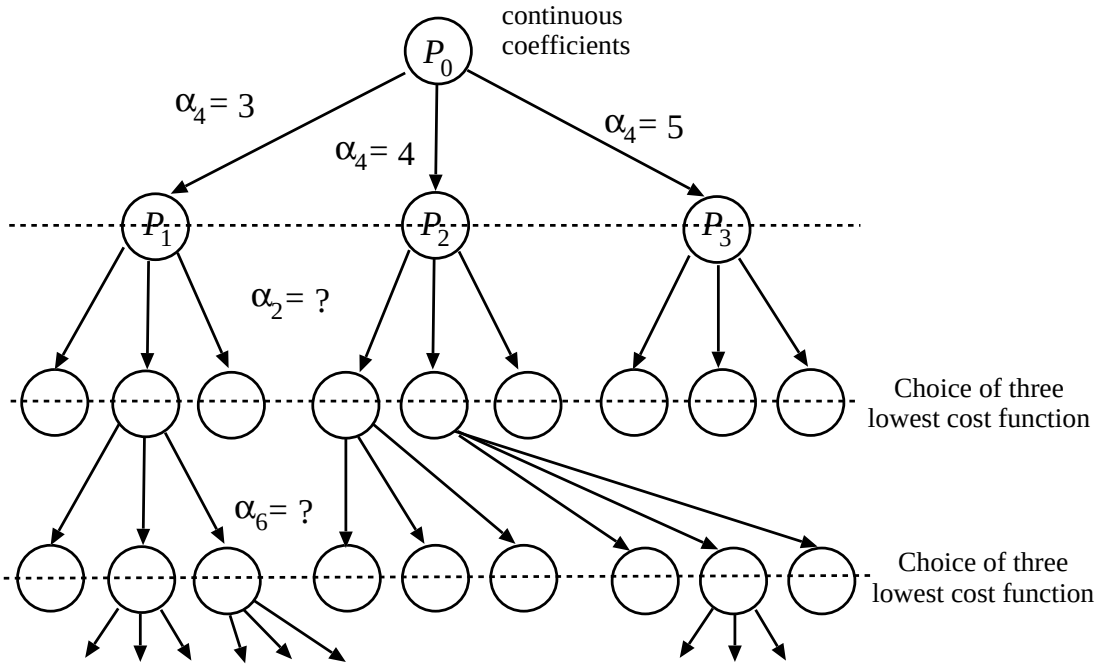

Fig. 5.2: An example of a hybrid of breadth-first and depth-first tree structure for the case where $L = 3$.

The ability of the branch and bound depth first search algorithm to produce a good suboptimal solution early in the search is particularly useful. The branch and

bound depth first search algorithm is indeed eminently suitable when an efficient constrained optimization algorithm capable of handling the bounds imposed on the variables is available. Unfortunately, in the case of designing the lattice PR orthogonal filter bank, such an efficient constrained optimization algorithm is not available. Although the tree search algorithm described in [53] does not impose bounds on the variables (because the variables are fixed at discrete values), it produces discrete solutions only at the final step of the algorithm. This will be a problem if there is insufficient computing resources to complete the execution of the algorithm. The optimization algorithm for optimizing linear phase FIR filters to meet a given frequency response requirement does not require large computing resources. Thus, choosing a fairly large value of L is not a problem in the case of [53]. In the design of lattice perfect reconstruction filter bank, the optimization algorithm requires long computer time. The type of tree search algorithm used in [53] is obviously not suitable. A suitable tree search technique should be one which will produce a good suboptimal solution within a reasonable time and will produce improved solutions as more time elapsed; that implies some form of depth-first search strategy. Taking the particular nature of the problem into consideration, in this section, a width-recursive depth-first tree search algorithm is developed.

The new algorithm is developed from that described in [53]. It starts with $L = 1$. At each node of the tree, the coefficient selected for quantization is the one with the largest performance deterioration measure discussed in Section 5.1. When the predefined maximum tree width is larger than 1, the above solution with $L = 1$ becomes the first suboptimal discrete solution.

Since the weighting function $B(\omega)$ in (2.37) is updated after every iteration as the optimization process proceeds, the objective function value f is not a good indicator of the optimality condition. In the proposed algorithm, the minimum weighted attenuation in the stopband of the analysis filter is used as a criterion for evaluating the quality of a solution.

After obtaining the first suboptimal discrete solution (i.e. node P_A in Fig. 5.3),

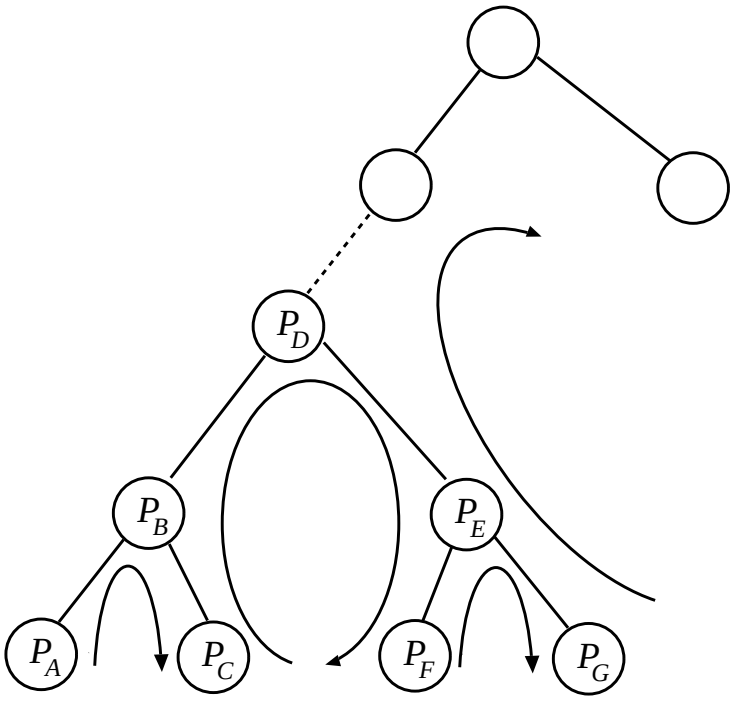

Fig. 5.3: A width-recursive depth-first tree.

the width of the tree is incremented by one. The search is backtracked to P_B and branched into P_C by fixing the last continuous coefficient value to its next nearest discrete value (It has been fixed at its nearest discrete value at P_A). Another discrete solution is obtained. If this solution is better than the previous one, it replaces the previous one as the best known discrete solution; otherwise, it is discarded. The search then backtracks to P_D and switches to search along P_E and P_F as shown in Fig. 5.3. The process of backtracking, switching, and searching forward is repeated until all nodes which have the possibility of yielding a better discrete solution than the best currently known discrete solution are searched. The search along a given path is terminated whenever the minimum weighted stopband attenuation is smaller than that of the best known discrete solution. For $L = 2$, the tree looks very much like a branch and bound depth first search tree with the exception that, in the width-recursive depth-first case, the branch length to a discrete solution is equal to the number of the discrete variables whereas, in the case of the branch and bound depth first search, the branch lengths leading to a

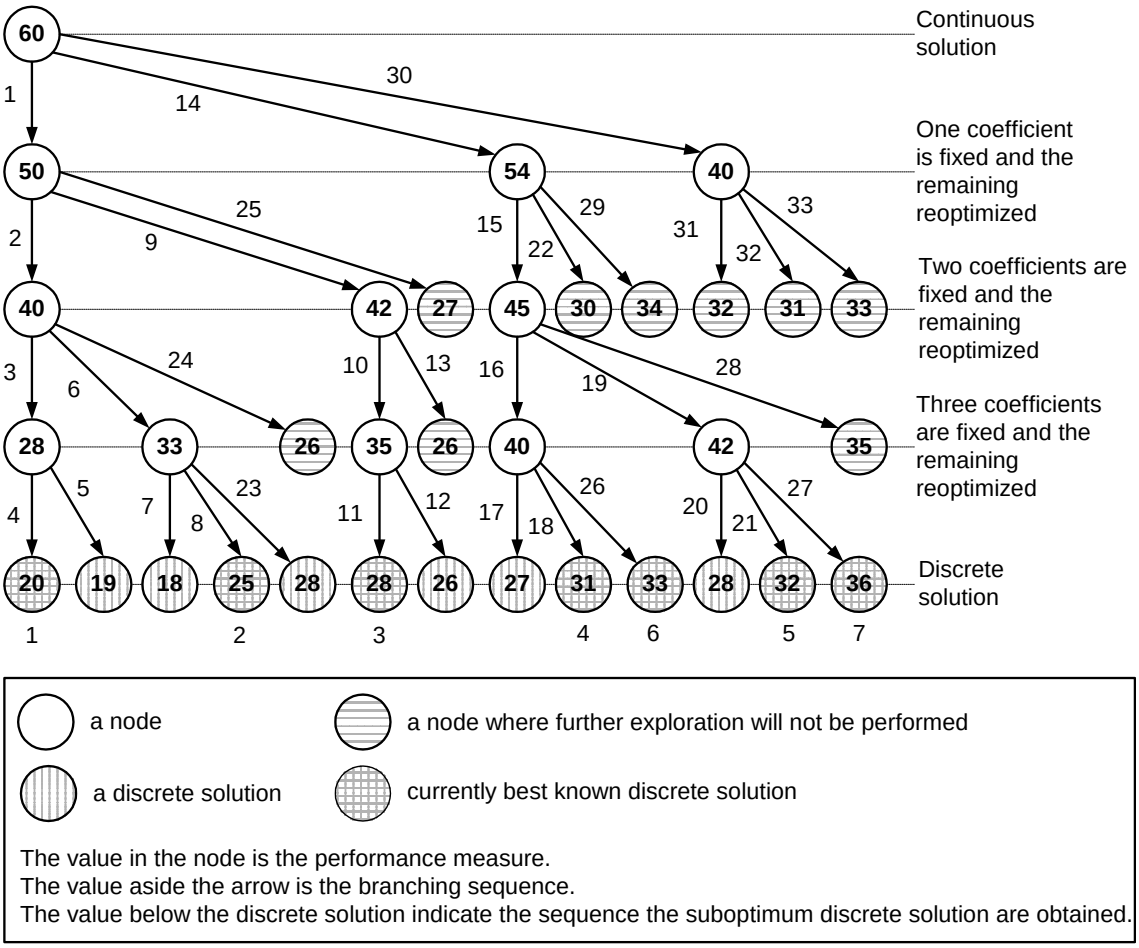

Fig. 5.4: An illustration for the proposed width-recursive depth-first tree search strategy for the case where $N = 4$ and $L = 3$.

discrete solution are usually larger than the number of discrete variables.

The width of the tree is increased recursively by one at a time until a predefined tree width, L, is reached. An example of the tree for $N = 4$ and $L = 3$ is shown in Fig. 5.4. The proposed new tree search strategy has the following advantages. First, it quickly yields a suboptimal discrete solution; second, it covers a large search space if the necessary computing resources are available.

5.3 Design Example

The 27-th order (14 coefficients) filter bank, specified in Section 3.2.3, is selected as an example to illustrate the tree search technique. The specifications are repeated here for easy reference. The lowpass filter's stopband edge is at 0.64π and its

stopband frequency response is equiripple.

The frequency responses of the lowpass filters for the (1) continuous coefficient optimum design, (2) SPT coefficient design obtained using the proposed tree search algorithm, (3) SPT coefficient design obtained using the successive reoptimization approach described in Section 3.2 and (4) SPT coefficient design obtained by simple rounding of coefficient values are shown in Fig. 5.5. In Fig. 5.5, each coefficient value is represented as a sum of two SPT terms; the smallest power-of-two term is 2^{-12}. The minimum stopband attenuations for the four designs in Fig. 5.5 are 60.20dB, 46.20dB, 32.57dB and 28.32dB, respectively. The SPT coefficient values of the design obtained using the proposed technique are listed in Table 5.1.

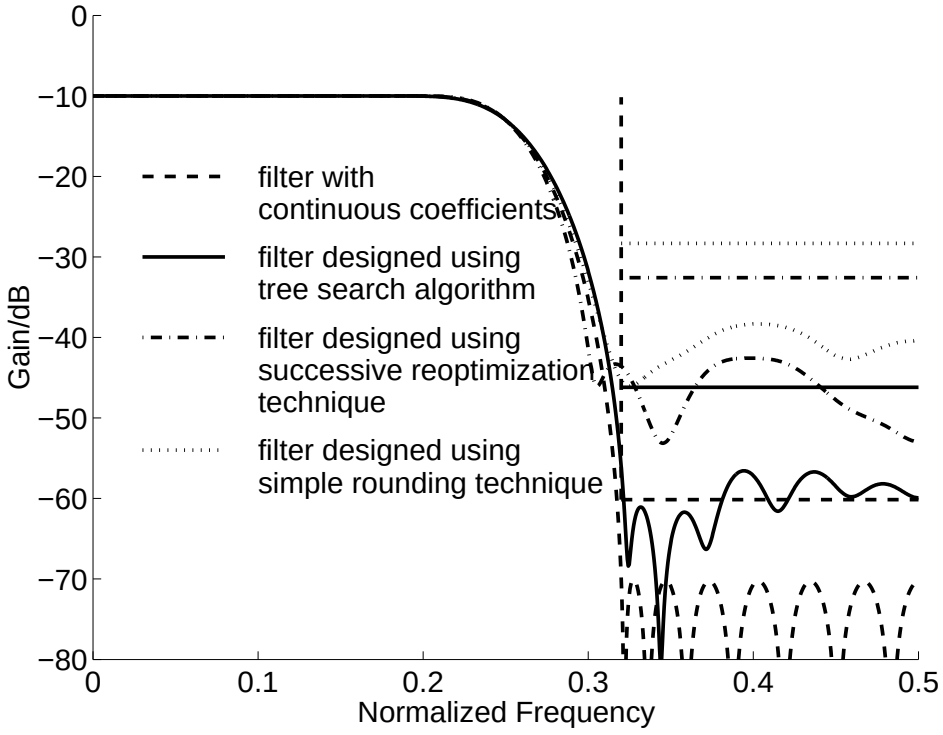

Fig. 5.5: Frequency response plots for the analysis lowpass filters. Each coefficient of the discrete coefficient design is represented by a sum of two signed power-of-two terms.

From Fig. 5.5, it is obvious that the stopband attenuation of the discrete space design obtained using the tree search algorithm is significantly superior to those obtained by simple rounding of coefficient values and successive reoptimization

Table 5.1: Discrete coefficient values of the 27-th order PR orthogonal filter bank obtained using the proposed tree search approach. The stopband edge is at $\omega_s = 0.64\pi$.

k	SPT coefficients
0	$-2^{+2}-2^{-2}$
1	$2^{+1}-2^{-1}$
2	$-2^{+0}+2^{-4}$
3	$2^{-1}+2^{-11}$
4	$2^{-8}-2^{-11}$
5	$-2^{-2}+2^{-5}$
6	$-2^{+1}+2^{-1}$
7	$2^{+1}-2^{-4}$
8	$2^{+0}-2^{-2}$
9	$-2^{+0}+2^{-9}$
10	$-2^{-4}+2^{-8}$
11	$2^{-2}+2^{-8}$
12	$-2^{-3}+2^{-6}$
13	$2^{-6}+2^{-9}$

approach.

The stopband attenuation obtained improves with increasing tree width associated with increasing computing cost. The stopband attenuation obtained for each tree width and its computing time are plotted in Fig. 5.6 for the 27-th order filter design. It can be seen from Fig. 5.6 that the computing time of the tree search algorithm is approximately proportional to L^2.

The width-recursive depth-first tree search technique does not necessarily result in the optimal solution. However, it does provide an efficient and reasonably good solution to the problem. In order to show the relationship between the filter length and the performance of a filter and that between the filter length and the computer time required, a set of PR orthogonal filter banks are designed. The stopband edge of the filter banks' lowpass filters is 0.56π. N ranges from 16 to 32, where $(2N-1)$ is the filter order. Each coefficient value of the discrete coefficient design is represented

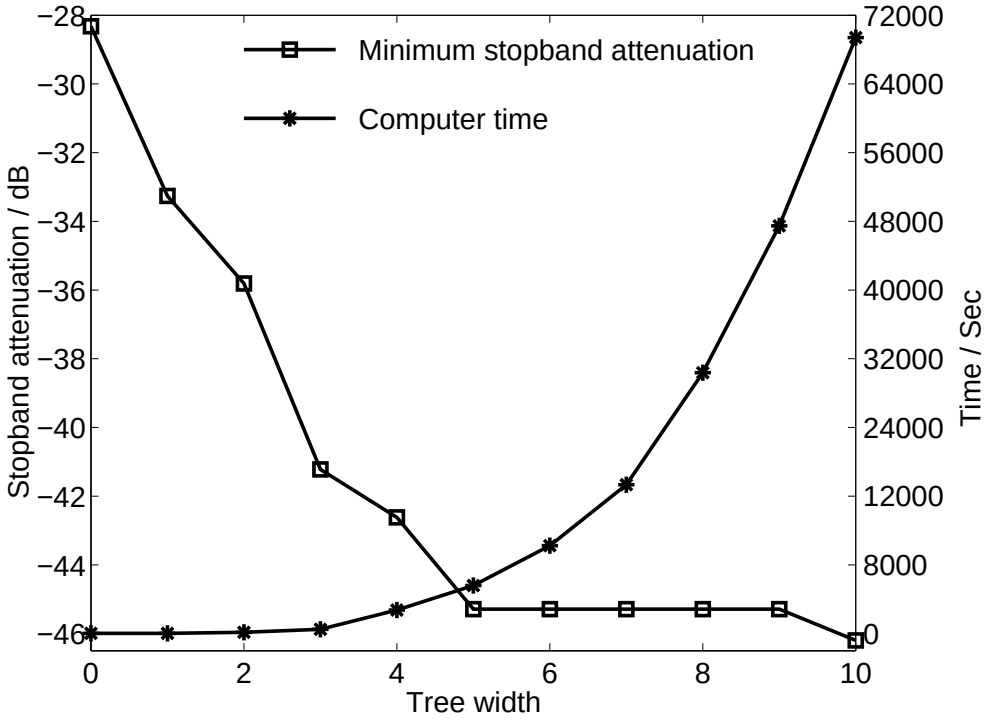

Fig. 5.6: Stopband attenuation and computing cost versus tree width plot for the example designed using width-recursive depth-first tree search technique, where each coefficient value is represented by a sum of two SPT terms.

as a sum of two SPT terms; the smallest SPT term is 2^{-10}. The minimum stopband attenuations of the lowpass filters for the (1) continuous coefficient optimum design, (2) SPT coefficient design obtained by simple rounding of coefficient values, and (3) SPT coefficient design obtained using the tree search technique are shown in Fig. 5.7. It can be seen from Fig. 5.7 that, for the same filter specifications, the minimum stopband attenuation of the infinite precision coefficient designs when expressed in dB is proportional to the filter length $2N$. For SPT coefficient designs, the minimum stopband attenuation is very close to the continuous coefficient design for small values of N. When N exceeds a certain limit, the peak stopband gain of the discrete coefficient design deviates away from the infinite precision coefficient design and finally remains fairly constant despite increasing filter length. As can be seen from Fig. 5.7, the peak stopband gain of the SPT coefficient filter designed using the proposed tree search method is significantly smaller than that of the

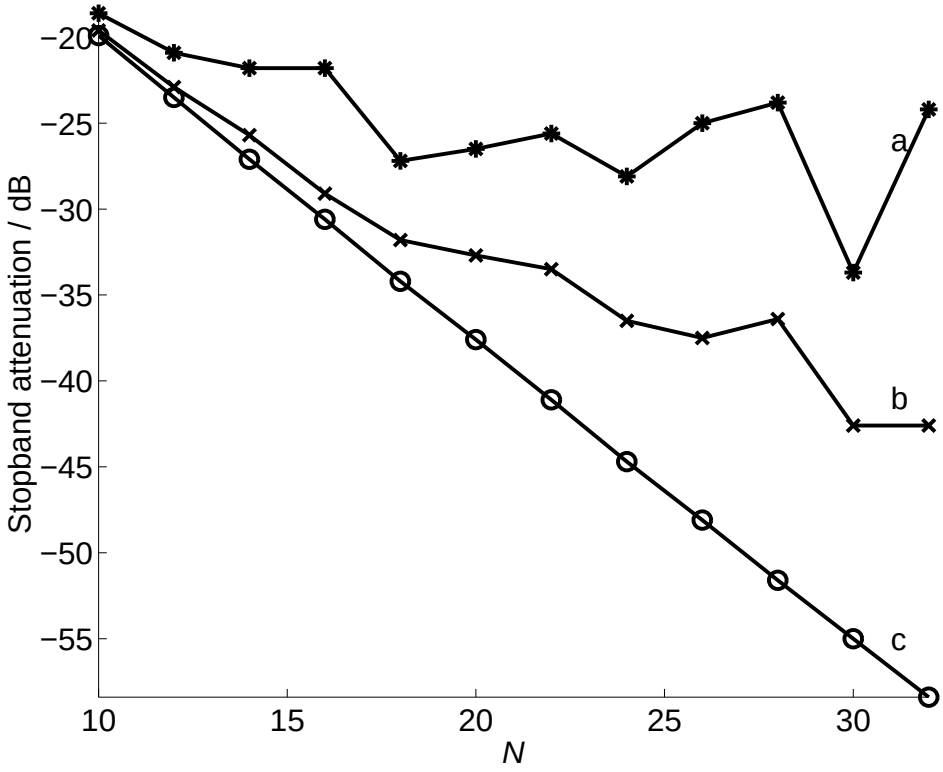

Fig. 5.7: The minimum stopband attenuations of the lowpass filters for the a) infinite precision coefficient designs; b) discrete coefficient designs obtained using tree search algorithm; c) discrete coefficient designs by simple coefficient rounding technique.

simple rounded coefficient design. The computing time required by the proposed algorithm for tree width equal to two for various filter lengths is plotted in Fig. 5.8. As can be seen from Fig. 5.8, the computing time increases exponentially with respect to filter length $2N$.

5.4 Discussion

In this chapter, Chapter 3 and Chapter 4, three methods for the design of SPT coefficient two-channel lattice orthogonal filter banks have been presented. The three methods are the successive reoptimization approach, the improved genetic algorithm and the width-recursive depth-first tree search algorithm. The width-recursive depth-first tree search algorithm is an extension of the successive reoptimization approach. Therefore, the tree search algorithm and the genetic algorithm

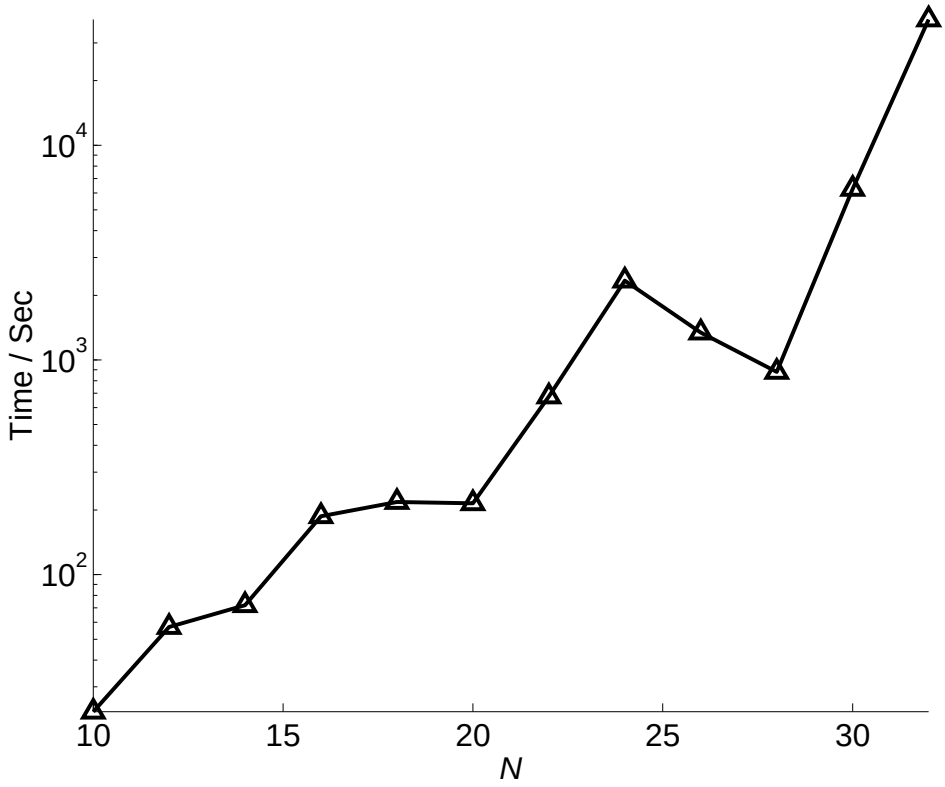

Fig. 5.8: The computing time of a set of discrete coefficient designs by using the proposed algorithm when the tree width is equal to 2.

are the two independent approaches.

As stated in Section 2.3.3, there are two schools of thoughts for the distribution of the SPT terms to the coefficients. One is that each coefficient is allocated with the same number of SPT terms. Another one is that each coefficient is allocated with different number of SPT terms but the total number of SPT terms for the entire filter is fixed. Both schools of thoughts have their own respective merits. Also, both the improved genetic algorithm and the tree search optimization technique are suitable for optimizing both these two cases. If it is desired to optimize filters with different number of SPT terms for each coefficient using the tree search algorithm, the SPT terms must be preallocated by using some other SPT term allocation scheme which will be discussed in Chapter 6.

When compared with the results obtained by using the improved GA presented in Chapter 4, the tree search algorithm presented in this chapter and the local

search methods [67, 86] for the design of SPT coefficient lattice filter banks, the following is observed:

1) The example of the 27-th order filter bank designed showed that the improved genetic algorithm presented in Chapter 4 is superior over the conventional GA. However, the result of 46.0dB obtained using the improved genetic algorithm is still slightly inferior to the result of 46.2dB designed using the tree search algorithm, although the genetic algorithm has taken the advantage that the coefficient values are allocated with different number of SPT terms while keeping the total number of SPT terms fixed; in the tree search algorithm, all the coefficients are allocated the same number of SPT terms.

2) The above phenomena is frequently observed. The example taken from references [34, 75] is selected for another illustration. The filter specifications are as follows: a 31-st order filter bank with stopband edge at 0.56π, the stopband frequency response is equiripple; a total number of 64 SPT terms are allocated to all the coefficients. The infinite precision optimum solution has a peak stopband gain of -30.7dB and the coefficient values are tabulated in column two of Table 5.2. Reference [34] used a local search algorithm and reference [75] used a genetic algorithm. Both methods reported solutions with 29.0dB attenuation in the stopband. The proposed tree search method produced a design with peak stopband ripple of -29.1dB even though a further constraint that all the coefficients must have the same number of SPT terms is imposed. The coefficients are tabulated in column three of Table 5.2.

3) Previous reports [42, 67, 86] showed that, for the design of SPT coefficient linear phase FIR filters, the local search may produce results close to those obtained by MILP. For the design of lattice filter banks, the results obtained using the local search method reported in [67, 86] are much inferior to those obtained using the tree search algorithm, especially when the filter order is high. The stopband attenuations for the cluster of filters with stopband edge at 0.56π and N ranges from 16 to 32 are plotted in Fig. 5.9.

Table 5.2: Coefficient values of the 31-th order design with the stopband edge at $\omega_s = 0.56\pi$.

k	Continuous coefficients	SPT coefficient. Each coefficient has two SPT terms
0	−2.6619195	$-2^{+2}+2^{+0}$
1	0.8784588	$2^{+0}+2^{-8}$
2	−0.5167097	$-2^{-1}-2^{-4}$
3	0.3580536	$2^{-1}-2^{-3}$
4	−0.2670765	$-2^{-2}-2^{-5}$
5	0.2072396	$2^{-2}-2^{-5}$
6	−0.1640125	$-2^{-3}-2^{-5}$
7	0.1310766	$2^{-3}+2^{-7}$
8	−0.1049166	$-2^{-3}+2^{-6}$
9	0.0835565	$2^{-4}+2^{-6}$
10	−0.0659682	$-2^{-4}-2^{-8}$
11	0.0510935	$2^{-4}-2^{-6}$
12	−0.0388140	$-2^{-5}-2^{-8}$
13	0.0286118	$2^{-5}-2^{-10}$
14	−0.0201751	$-2^{-6}-2^{-9}$
15	0.0228699	$2^{-6}+2^{-8}$

The reasons that the genetic algorithms and local search methods are not able to obtain good lattice filter but may obtain good linear phase FIR filter are due to the following facts:

1) As suggested by the name of the technique, local search methods search in the discrete space in the vicinity of the the continuous optimum coefficients. Therefore, such method will be successful only when good discrete solution exists around the continuous optimum.

2) The frequency response of linear phase FIR filters are linear functions of the filter coefficients. The shape of the frequency response is unaffected by multiplying all the coefficients by a constant scaling factor. This scaling factor has

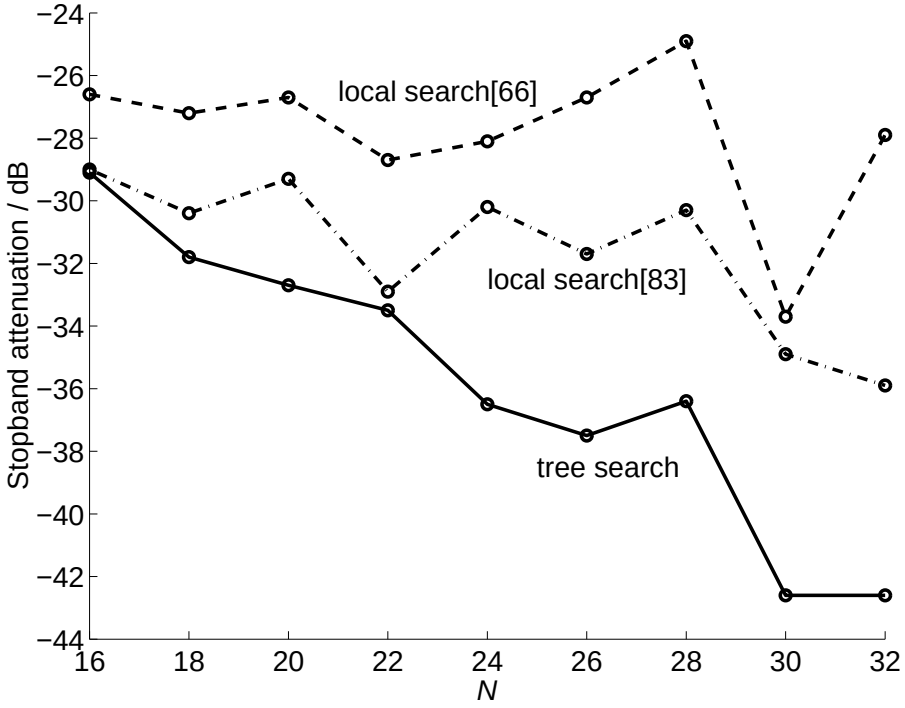

Fig. 5.9: The stopband attenuation for filter banks with stopband edge at 0.56π.

significant effect on the coefficient optimization process for SPT design [52] due to the nonuniformly distributed SPT values. Scaling the continuous coefficients prior to rounding the coefficients may improve the chances that good discrete solutions are located near the continuous ones. This is an important reason that local search may get good results for the design of linear phase FIR filters.

3) The frequency responses of the lattice filter banks are not linear functions of the lattice coefficients. Therefore, the scaling factor strategy is not adoptable.

4) As shown in Fig. 2 in [86], the SPT coefficients of filter designed by using MILP are almost the same as those of the continuous coefficients. However, it is not the case in the lattice filter bank design. For the 27-th order lattice filter bank with the stopband edge at 0.64π, using the local search method [86], 35.0dB stopband attenuation is achieved and the SPT coefficients are almost the same as the continuous ones, as shown in Fig. 5.10 by the symbols '$\circ$' and '$+$' for the continuous and discrete values, respectively. However, a much better solution obtained using the tree search algorithm achieved 46.2dB stopband attenuation, and some of its

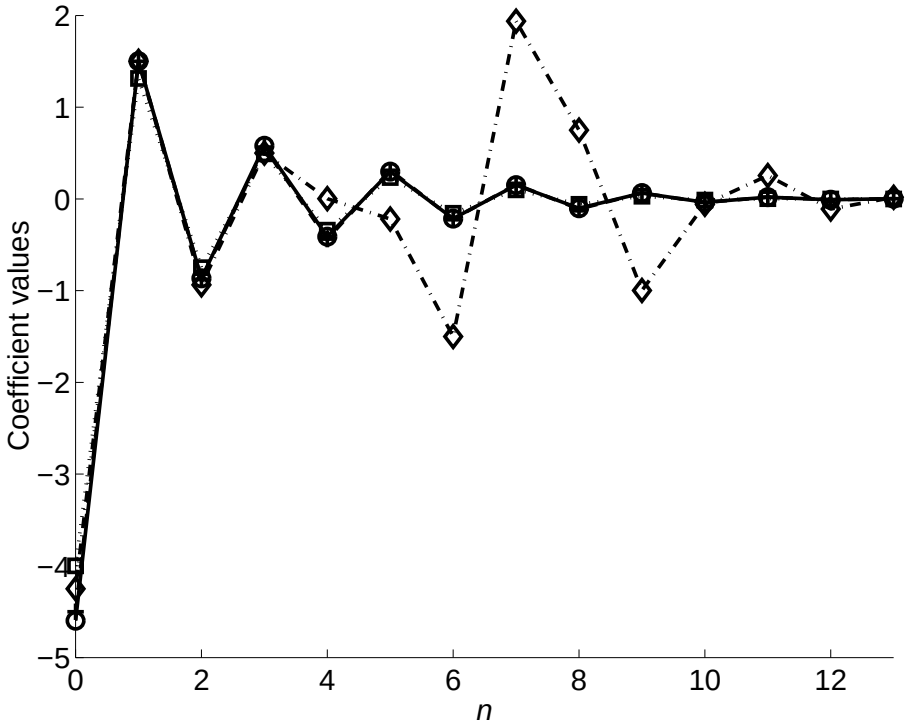

Fig. 5.10: The coefficient values for the 27-th order example with stopband edge at 0.64π. '$\circ$': Continuous coefficients, '+': SPT coefficients obtained by local search, '$\square$': SPT coefficients obtained by genetic algorithm, and '$\diamond$': SPT coefficients obtained by tree search.

SPT coefficients deviated far away from the continuous ones, as shown in Fig. 5.10 by the symbol '$\diamond$'.

5) Genetic algorithm also produces discrete solutions near the continuous optimum, as can be seen from the coefficient values shown in Fig. 5.10 by the symbol '$\square$'.

5.5 Conclusion

In this chapter, Chapter 3 and Chapter 4, two independent methods, improved genetic algorithm and width-recursive depth-first tree search algorithm, for the design of SPT coefficient lattice filter banks are presented. Although the new GA proposed in Chapter 4 is superior to the conventional GA, being GA in nature, it has the same limitations as those in local search methods in which the discrete solutions obtained are in the vicinity of the initial continuous coefficients. Local

search method as well as GA therefore have few chances to hop to discrete solutions located far away from the the continuous one. The global optimum discrete solution of the lattice filter bank may be located far away from the continuous optimum. The tree search algorithm proposed in this chapter overcame this difficulty very well in two aspects. First, the coefficient selected to be quantized may be fixed at a considerable distance away from the continuous values for large L. Second and more importantly, after each coefficient is fixed, the remaining unquantized coefficients are reoptimized; the reoptimization process may throw the coefficient values far away from the original continuous optimum coefficients.

.

Chapter 6

Analysis of SPT Number Effects

COEFFICIENT QUANTIZATION may cause the frequency response of an FIR filter to deteriorate to such an extent that it may no longer be acceptable. To analyze the effect of coefficient quantization, it is necessary to know the coefficient quantization error probability density distribution. The error probability density distribution for quantizing a number to a finite wordlength value is uniform, and its effects on the frequency response of an FIR filter had been extensively studied [39, 40]. For SPT values, some statistical analysis had been reported in [55, 84]. In [55], the statistically estimated number of SPT terms required to represent a coefficient was investigated. In [84], the distribution of the SPT terms is deduced. However, there is no report on the statistical distribution of SPT quantization error.

In this chapter, the distribution of quantization error in the SPT space is studied. Mathematical expressions for the quantization error distributions is established subject to a given smallest power-of-two term and a given number of SPT terms. The effects of quantizing the coefficients to SPT values on the frequency responses of the two-channel lattice orthogonal filter banks are investigated. A new SPT term allocation scheme is also developed.

Detailed proofs of the quantization error distributions are given in Appendices A and B at the end of this chapter. Appendix C presents several lemmas which are referred by the proofs in Appendices A and B.

Unless stated otherwise, in this chapter, a filter bank refers to a two-channel

lattice orthogonal filter bank.

6.1 Rounding Error Probability Density Function Analysis

The permitted discrete value of an integer that can be represented as a sum of not more than K SPT terms is unevenly distributed on the integer space. Therefore, its error is also unevenly distributed. In this section, a discussion on the error probability density function (PDF) when infinite precision numbers are represented using SPT integers is presented.

As the canonic SPT number ensures a unique minimum representation, in the following analysis, canonic condition is imposed on all SPT numbers to simplify the derivation.

As reviewed in Section 2.3.1, a number n can be represented to a precision 2^Q by a canonic SPT number with K SPT terms as

$$n = \sum_{i=0}^{K-1} y(i)2^{q(i)}, y(i) \in \{-1, 1\}, \tag{6.1}$$

where $Q \leq q(i) \leq L-1$. Furthermore, for any i and j, it satisfies the constraints that $q(i) \neq q(j)$ if $i \neq j$ and that $q(i) \neq q(j)+1$. Without loss of generality, assume that $q(i) < q(j)$ if $i > j$, i.e., $q(i)$ is a decreasing sequence. For the particular condition where $Q = 0$, n is an L-bit integer. In this section, this particular case for quantizing a number to an SPT number where $Q = 0$ is considered. The case where $Q \neq 0$ will be deduced from the case where $Q = 0$.

Let $\mathcal{Z}^+$ denote the set of all positive integers, and let $L, K \in \mathcal{Z}^+$. As introduced in Section 2.3.1, an L-bit canonic SPT number has at most $\left\lfloor \frac{L+1}{2} \right\rfloor$ SPT terms, where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x. Assume further that $L \geq 2K-1$. Let $T^+(L, K)$ be a subset of $\mathcal{Z}^+$ such that any $n \in T^+(L, K)$ is a sum of exactly K canonic SPT terms and the largest power-of-two term is less than or equal to 2^{L-1}, where n is given by (6.1) in which Q is equal to 0.

It is known that the number of elements of the set $T^+(L, K)$, represented as

$N^+(L,K)$, is [55]

$$N^+(L,K) = \frac{2^{K-1}}{K!}\prod_{k=0}^{K-1}(L-K+1-k). \tag{6.2}$$

Let $S^+(L,K)$ be a subset of $\mathcal{Z}^+$ such that $S^+(L,K) = \bigcup_{k=1}^{K} T^+(L,k)$, i.e., any $n \in S^+(L,K)$ is a sum of not more than K canonic SPT terms and the largest power-of-two term is less than or equal to 2^{L-1}. It is noticed that $S^+(L,K)$ does not include zero.

Let $M^+(L,K)$ be the number of elements of the set $S^+(L,K)$. It is straightforward to show that

$$M^+(L,K) = \sum_{k=1}^{K} N^+(L,k). \tag{6.3}$$

Therefore, the number of elements of $S^+\left(L, \left\lfloor \frac{L+1}{2} \right\rfloor\right)$ is $M^+\left(L, \left\lfloor \frac{L+1}{2} \right\rfloor\right)$. Denote $M^+\left(L, \left\lfloor \frac{L+1}{2} \right\rfloor\right)$ as M_L^+, we have

$$M_L^+ = \left\lfloor \frac{2^{L+1}}{3} \right\rfloor. \tag{6.4}$$

M_L^+ is also the largest number which can be represented by a canonic SPT interge where the largest power-of-two term is less than or equal to 2^{L-1}.

Let

$$M_{L\infty}^+ = \sum_{k=0}^{\infty} 2^{L-1-2k} = \frac{2^{L+1}}{3}. \tag{6.5}$$

$M_{L\infty}^+$ is the largest infinite precision number represented by SPT terms with the largest power-of-two term less than or equal to 2^{L-1}.

Furthermore, let $T(L,K)$ be the union of $T^+(L,K)$ and $-T^+(L,K)$, i.e., $T(L,K)$ is the integer set where each member is an L-bit SPT number with exactly K SPT terms, including both the positive and negative numbers. $S(L,K)$ is denoted as the union of $S^+(L,K)$, $-S^+(L,K)$ and the element 0, i.e., $S(L,K)$ is the integer set that can be represented by L-bit SPT number with not more than K SPT terms, including both the positive and negative numbers as well as the zero element. Obviously, the number of elements in $T(L,K)$ is $2N^+(L,K)$, and the number of elements in $S(L,K)$ is $2M^+(L,K)+1$.

6.1.1 Error Probability Density Function

Before embarking on the analysis of the error distribution for rounding an infinite precision number x to the $S(L,K)$ space, it is assumed that $x \in \{x| - M^+_{L\infty} \leq x \leq M^+_{L\infty}, x \in \mathcal{R}\}$, where $\mathcal{R}$ is the set of real numbers, and x is uniformly distributed in $[-M^+_{L\infty}, M^+_{L\infty}]$, where $[a,b]$ denotes all the infinite precision numbers in the range bounded by a and b inclusive, as shown in Fig. 6.1. The reason for making the above assumption is that the error incurred in a number x to be quantized to SPT form is related to the value of the number. Larger number may cause larger error for a given K. Therefore, the distribution of x affects the distribution of the rounding error. In quantizing the coefficient values of a filter, L is selected so that it is just large enough to accommodate the coefficient values, and assume that the coefficient values are uniformly distributed in $[-M^+_{L\infty}, M^+_{L\infty}]$.

Let $\bar{x}$ be the integer value nearest to x.

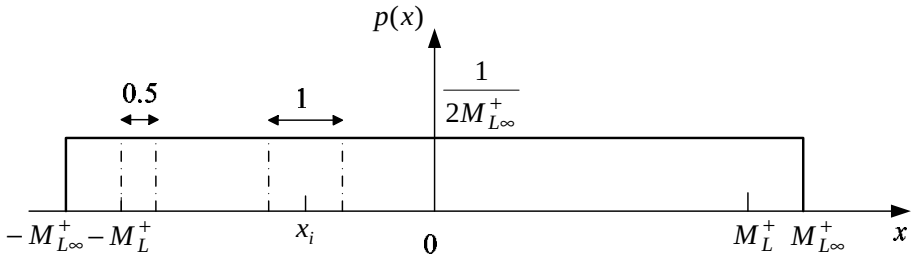

Fig. 6.1: A uniformly distributed random number x, $x \in \{x| - M^+_{L\infty} \leq x \leq M^+_{L\infty}, x \in \mathcal{R}\}$.

Property 6.1 For $L = 2K - 1$, the error PDF of rounding a number $x \in \{x| - M^+_{L\infty} \leq x \leq M^+_{L\infty}, x \in \mathcal{R}\}$ to the element in $S(L,K)$ nearest to it, denoted as

$p_{L,K}(e)$, is as follows:

$$p_{L,K}(e) = \begin{cases} \dfrac{2M_L^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in \left[-3^{-1}, 3^{-1}\right], \\ \dfrac{M_L^+}{M_{L\infty}^+}, & \text{for } e \in \pm\left[3^{-1}, 2^{-1}\right], \\ 0, & \text{otherwise.} \end{cases} \tag{6.6}$$

□

Proof: For $L = 2K - 1$, according to (6.4) and (6.5), it is obvious that $M_{L\infty}^+ = M_L^+ + 3^{-1}$. For any number $x \in \left\{x \middle| -M_L^+ \le x \le M_L^+, x \in \mathcal{R}\right\}$, we have $\bar{x} \in S(L,K)$. According to Lemma 6.1 in Appendix C, the rounding error is uniformly distributed in $[-2^{-1}, 2^{-1}]$ with unity PDF. Furthermore, the probability of $x \in \{x| -M_L^+ \le x \le M_L^+, x \in \mathcal{R}\}$ is $\frac{2M_L^+}{2M_{L\infty}^+}$. The probability of $x \in \left\{x \middle| M_L^+ \le x \le M_{L\infty}^+\right\}$ and $x \in \left\{x \middle| -M_{L\infty}^+ \le x \le -M_L^+\right\}$ are both $\frac{1}{6M_{L\infty}^+}$, and the rounding errors are distributed uniformly in $[0, 3^{-1}]$ and $[-3^{-1}, 0]$, respectively, with PDF equal to 3. Therefore,

$$p_{L,K}(e) = \begin{cases} \dfrac{2M_L^+}{2M_{L\infty}^+} \times 1, & \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\ \dfrac{1}{6M_{L\infty}^+} \times 3, & \text{for } e \in \left[0, 3^{-1}\right], \\ \dfrac{1}{6M_{L\infty}^+} \times 3, & \text{for } e \in \left[-3^{-1}, 0\right], \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \dfrac{2M_L^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in \left[-3^{-1}, 3^{-1}\right], \\ \dfrac{M_L^+}{M_{L\infty}^+}, & \text{for } e \in \pm\left[3^{-1}, 2^{-1}\right], \\ 0, & \text{otherwise.} \end{cases}$$

■

Property 6.2 The error PDF for rounding a number $x \in \{x| -M_{L\infty} \le x \le$

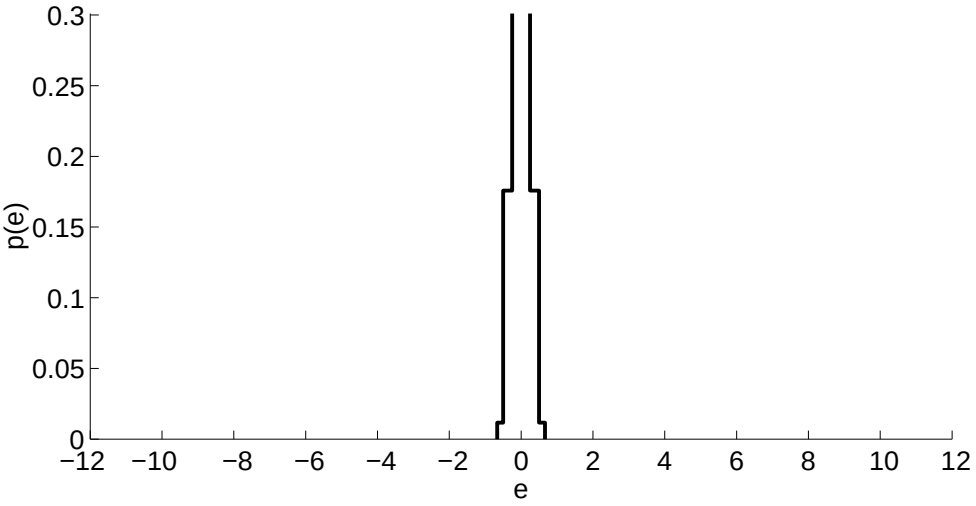

Fig. 6.2: The PDF for rounding a number to an L bit SPT integer with not more than K SPT terms, where $L = 8$ and K=2.

$M_{L\infty}, x \in \mathcal{R}\}$ to an element in $S(L,K)$ when $L \geq 2K$ is given by:

$$p_{L,K}(e) = \begin{cases} \dfrac{2M^+(L,K)+1}{2M^+_{L\infty}}, & \text{for } e \in [-2^{-1}, 2^{-1}], \\ \dfrac{4M^+(L-1-k,K) - 2M^+(L-k,K)+1}{2M^+_{L\infty}}, & \text{for } e \in \pm[2^{k-1}, 2^k], \\ & k = 0, \cdots, L-2K-1, \\ \dfrac{1}{2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{L-2K-1}, \dfrac{2^{L-2K+1}}{3}\right] \\ & \text{when } L-2K-1 \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6.7}$$

□

The proof for Property 6.2 is shown in Appendix A.

An example of the error PDF for rounding an infinite precision number to an element in $S(8,2)$ is shown in Fig.6.2.

6.1.2 Mean and Variance

Property 6.3 The mean of the error caused by rounding a number x to an element in $S(L,K)$, $E(e)$, is equal to 0. The variance of the error, $\sigma^2_{L,K}(e)$, is given by

$$\sigma^2_{L,K}(e) = \begin{cases} \dfrac{1}{12}\cdot\dfrac{27M_L^+ + 4}{27M_L^+ + 9}, & \text{for } L = 2K-1, \\ \dfrac{1}{12}\cdot\dfrac{27M_L^+ + 32}{27M_L^+ + 18}, & \text{for } L = 2K, \\ \dfrac{1}{2M_{L\infty}^+}\left[\left(\dfrac{7}{3}M^+(2K,K) + \dfrac{128}{81}\right)2^{3(L-2K-1)} - M^+(L,K)\right. & \\ \qquad \left. - \displaystyle\sum_{k=0}^{L-2K-2} 7M^+(L-1-k,K)2^{3k}\right], & \text{for } L \geq 2K+1. \end{cases} \tag{6.8}$$

□

The proof for Property 6.3 is shown in Appendix B.

The quantization process where the smallest power-of-two term is 2^0 has been considered. If the smallest power-of-two term is 2^Q instead of 2^0, where the largest power-of-two term remains at 2^{L-1}, the wordlength of the resulting SPT number becomes $L-Q$; the error PDF of this quantization process is a scaled version of $p_{L-Q,K}(e)$, where the value of the error PDF is scaled by $\frac{1}{2^Q}$ and the error range is scaled by 2^Q. Therefore, when $L-Q \geq 2K$, the error PDF for quantizing a number to an SPT number with K SPT terms, and the largest and smallest power-of-terms are 2^{L-1} and 2^Q, respectively, denoted as $p_{L,K,Q}(e)$, is given by

$$p_{L,K,Q}(e) \;=\; p_{L-Q,K}\left(2^{-Q}e\right)\cdot\frac{1}{2^Q}$$

$$
\begin{aligned}
= \begin{cases}
\dfrac{2M^+(L-Q,K)+1}{2M^+_{(L-Q)\infty}} \cdot \dfrac{1}{2^Q}, & \text{for } e \in [-2^{Q-1}, 2^{Q-1}], \\
\dfrac{4M^+(L-Q-1-k,K) - 2M^+(L-Q-k,K)+1}{2M^+_{(L-Q)\infty}} \cdot \dfrac{1}{2^Q}, & \\
& \text{for } e \in \pm[2^{Q+k-1}, 2^{Q+k}], \\
& k = 0, \cdots, L-Q-2K-1, \\
\dfrac{1}{2M^+_{(L-Q)\infty}} \cdot \dfrac{1}{2^Q}, & \text{for } e \in \pm\left[2^{L-2K-1}, \dfrac{2^{L-2K+1}}{3}\right] \\
& \text{when } L-Q-2K-1 \geq 0, \\
0, & \text{otherwise,}
\end{cases} \\
= \begin{cases}
\dfrac{2M^+(L-Q,K)+1}{2M^+_{L\infty}}, & \text{for } e \in [-2^{Q-1}, 2^{Q-1}], \\
\dfrac{4M^+(L-Q-1-k,K) - 2M^+(L-Q-k,K)+1}{2M^+_{L\infty}}, & \\
& \text{for } e \in \pm[2^{Q+k-1}, 2^{Q+k}], \\
& k = 0, \cdots, L-Q-2K-1, \\
\dfrac{1}{2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{L-2K-1}, \dfrac{2^{L-2K+1}}{3}\right] \\
& \text{when } L-Q-2K-1 \geq 0, \\
0, & \text{otherwise.}
\end{cases}
\end{aligned} \tag{6.9}
$$

Thus, the mean of the errors remains at 0, whereas the variance, $\sigma^2_{L,K,Q}(e)$, is given

Table 6.1: Some values of $\sigma_{L,K,Q}$ for $Q = -10$.

L \ K	2	3	4	5	6
4	1.9646E−1	3.6496E−2	6.8449E−3	1.3019E−3	3.4860E−4
3	9.8229E−2	1.8248E−2	3.4244E−3	6.7739E−4	2.9171E−4
2	4.9115E−2	9.1242E−3	1.7178E−3	3.9782E−4	2.8194E−4
1	2.4557E−2	4.5629E−3	8.7432E−4	3.0129E−4	2.8189E−4
0	1.2279E−2	2.2839E−3	4.7644E−4	2.8202E−4	–
−1	6.1396E−3	1.1497E−3	3.1988E−4	2.8183E−4	–
−2	3.0707E−3	5.9718E−4	2.8234E−4	–	–

by

$$\sigma^2_{L,K,Q}(e) = \begin{cases} \dfrac{2^{2Q}}{12} \cdot \dfrac{27M^+_{L-Q}+4}{27M^+_{L-Q}+9}, & \text{for } L-Q = 2K-1, \\ \dfrac{2^{2Q}}{12} \cdot \dfrac{27M^+_{L-Q}+32}{27M^+_{L-Q}+18}, & \text{for } L-Q = 2K, \\ \dfrac{2^{3Q}}{2M^+_{L\infty}} \Bigg[\left(\dfrac{7}{3}M^+(2K,K) + \dfrac{128}{81}\right) 2^{3(L-Q-2K-1)} & \\ \qquad -M^+(L-Q,K) - \displaystyle\sum_{k=0}^{L-Q-2K-2} 7M^+(L-Q-1-k,K)2^{3k} \Bigg], & \\ & \text{for } L-Q \geq 2K+1. \end{cases} \tag{6.10}$$

Several values of $\sigma_{L,K,Q}(e)$ for L ranges from 4 to -2 and K ranges from 2 to 6, corresponding to $Q = -10$ are listed in Table 6.1. In Table 6.1, each row of values corresponds to the range $\left[-\frac{2^{L+1}}{3}, \frac{2^{L+1}}{3}\right]$. It can be seen from Table 6.1 that the variance decreases with increasing K for a given L and decreases with decreasing L for a given K. Therefore, to achieve approximately the same variance, say 10^{-3}, when rounding a number to an SPT value, larger L requires larger K, i.e., more SPT terms. When $L = 4$, five SPT terms are required, whereas for $L = -1$, three SPT terms are required.

If the largest power-of-two term is fixed, i.e., L is fixed, the error variance for representing a random number to SPT value decreases with decreasing Q for a given K. In Fig. 6.3, the largest power-of-two term is fixed to be $2^{L-1} = 2^4$, thus

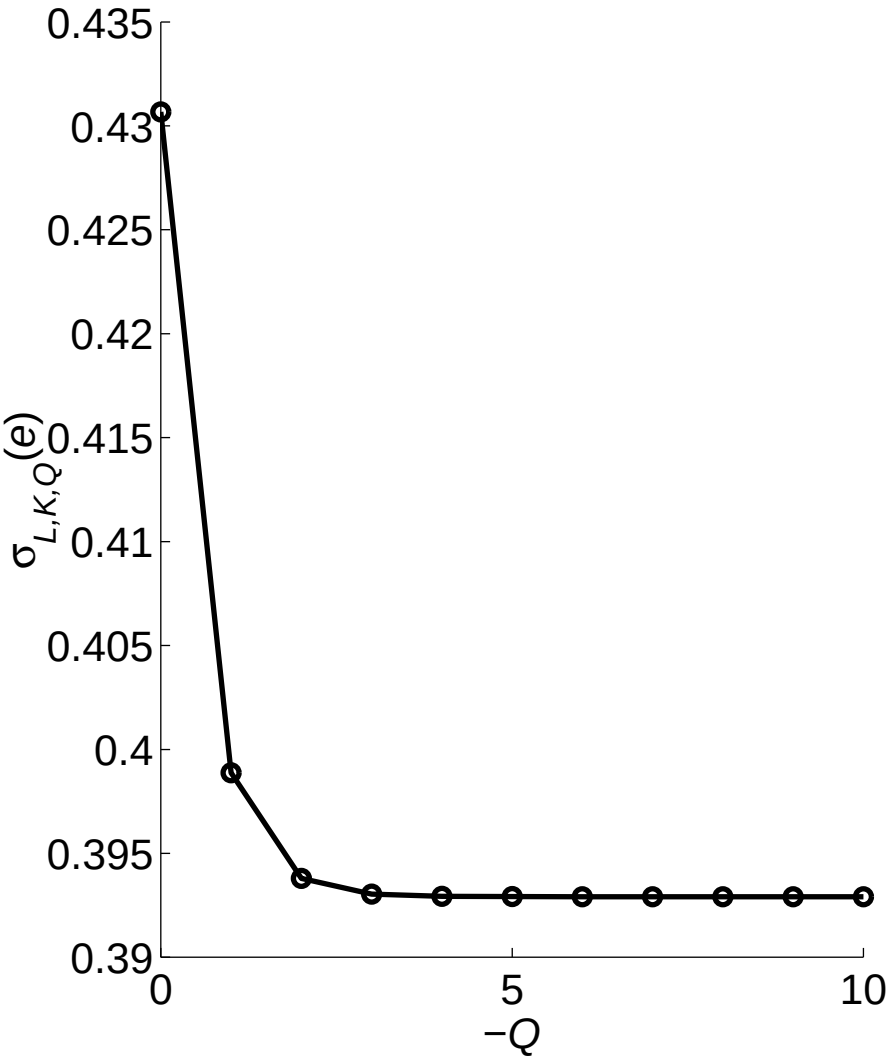

Fig. 6.3: $\sigma_{L,K,Q}(e)$ plot for $L = 5$ and $K = 2$. Note that the error variance decreases with decreasing Q for a given number range $\left[-\frac{2^{L+1}}{3}, \frac{2^{L+1}}{3}\right]$ and a given K.

the number representable is in $\left[-\frac{64}{3}, \frac{64}{3}\right]$, and $K = 2$. When Q decreases from 0 to -3, i.e., the smallest power-of-two term decreases from 2^0 to 2^{-3}, the error variance drops quickly from 0.431 to 0.393. However, for $Q < -3$, the decrease in the variance is almost not noticeable and the variance approaches a constant asymptotically as Q decreases. Increasing K may reduce the variance further, as shown in Fig. 6.4 where $K = 3$. However, the variance in Fig. 6.4 approaches asymptotically a smaller constant as Q decreases.

The remaining part of this section is devoted to present the error distribution

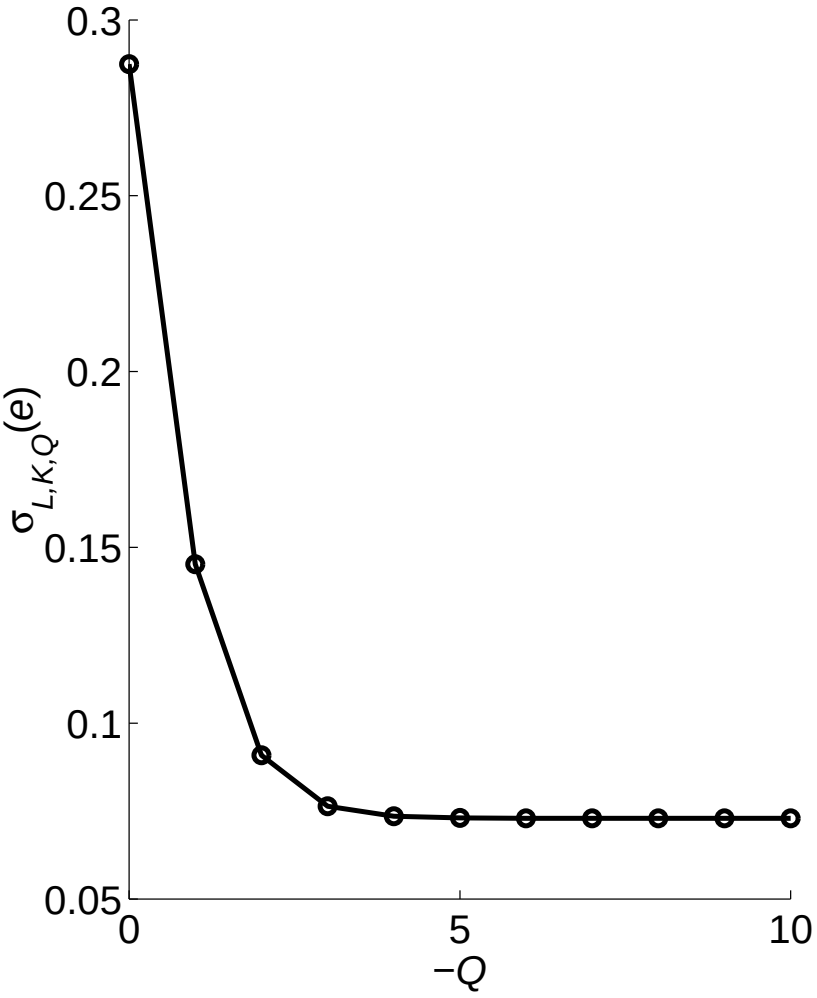

Fig. 6.4: $\sigma_{L,K,Q}(e)$ plot for $L = 5$ and $K = 3$. Note that the error variance decreases with decreasing Q for a given number range $\left[-\frac{2^{L+1}}{3}, \frac{2^{L+1}}{3}\right]$ and a given K.

and variances for several special cases.

The first case is $0 < L - Q < 2K - 1$, i.e., an $(L - Q)$-bit SPT number is represented with K SPT terms and $K > \left\lfloor \frac{L-Q+1}{2} \right\rfloor$. For canonic $(L - Q)$-bit SPT numbers, at most $\left\lfloor \frac{L-Q+1}{2} \right\rfloor$ SPT terms are used. Therefore, the redundant SPT terms do not contribute to the precision of the number. Thus,

$$p_{L,K,Q}(e) = p_{L,\lfloor \frac{L-Q+1}{2} \rfloor,Q}(e) \tag{6.11}$$

and

$$\sigma^2_{L,K,Q}(e) = \sigma^2_{L,\lfloor \frac{L-Q+1}{2} \rfloor,Q}(e) \tag{6.12}$$

for $0 < L - Q < 2K - 1$.

All the above expressions are true for $L-Q \geq 1$ and $K \geq 1$ since the wordlength $(L-Q)$ and the number of SPT terms (K) cannot be less than unity. Nevertheless, it is convenient to define several quantities corresponding to the trivial case where $K = 0$ or $L - Q \leq 0$.

$K = 0$ is the case that a number is represented with 0 SPT terms. Therefore, the rounding error is the number itself and the distribution is given as

$$p_{L,0,Q} = \begin{cases} \dfrac{1}{2M^+_{L\infty}}, & \text{for } e \in \left[-M^+_{L\infty}, M^+_{L\infty}\right], \\ 0, & \text{otherwise.} \end{cases} \tag{6.13}$$

According to (6.41), the variance is given by

$$\sigma^2_{L,0,Q}(e) = \frac{(M^+_{L\infty})^2}{3}. \tag{6.14}$$

The interpretation for $L - Q \leq 0$ will be explained in the next section. For $L - Q \leq 0$, define

$$\sigma^2_{L,K,Q}(e) = \begin{cases} \dfrac{\left(M^+_{L\infty}\right)^2}{3}, & \text{for } L - Q = 0, -1, \\ \dfrac{\left(M^+_{(L+2)\infty}\right)^2}{3}, & \text{for } L - Q \leq -2. \end{cases} \tag{6.15}$$

6.2 Statistical Effect of Coefficient Quantization

A statistical analysis on the effect of coefficient quantization on the frequency response of the filter bank is presented in this section.

The rounding of an infinite precision coefficient value to its nearest SPT value may be modeled as adding an error term to the coefficient value. Suppose that the error term associated with the rounding of the filter coefficient α_k is $\Delta\alpha_k$. Define $H_{N,0}(e^{j\omega})$ and $H^*_{N,0}(e^{j\omega})$ to be the frequency responses of a filter bank which has unquantized and quantized coefficients, respectively. Let the frequency response error of $H_{N,0}(z)$ in (2.34) due to the rounding of all its filter coefficients be denoted

by $E_{N,0}(e^{j\omega})$. Thus,

$$E_{N,0}(e^{j\omega}) = H^*_{N,0}(e^{j\omega}) - H_{N,0}(e^{j\omega}). \tag{6.16}$$

For small $|\Delta\alpha_k|$, $E_{N,0}(e^{j\omega})$ may be approximated as

$$E_{N,0}(e^{j\omega}) = \sum_{k=0}^{N-1} P_{N,k}(e^{j\omega}) \times \Delta\alpha_k, \tag{6.17}$$

where, $P_{N,k}(e^{j\omega})$ is given by (3.19). It is reasonable to assume that $\Delta\alpha_k$ for $k = 0, 1, \ldots, N-1$, are mutually independent. If each coefficient value is represented by an $L-Q$-bit SPT value with not more than K SPT terms, where 2^Q is the smallest power-of-two term, then the PDF of $\Delta\alpha_k$ is a piecewise constant function $p_{L,K,Q}(\Delta\alpha_k)$ as given in (6.9), and thus has zero mean and variance $\sigma^2_{L,K,Q}(\Delta\alpha_k)$, as given in (6.10). In the derivation, L is selected to be just large enough to represent the largest lattice coefficient. The statistical model in [11, 40] is used in the following analysis. From (3.29), (6.17) and the statistical property of $\Delta\alpha_k$, it can be shown that, for coefficient rounding, $E(e^{j\omega})$ has zero mean and variance $\sigma^2_{E_{N,0}}$ given by

$$\begin{aligned} \sigma^2_{E_{N,0}} &= \overline{E_{N,0}(e^{j\omega})^2} = \sum_{k=0}^{N-1} {P^2}_{N,k}(e^{j\omega}) \times \overline{\Delta\alpha_k^2} \\ &< \sum_{k=0}^{N-1} \overline{\Delta\alpha_k^2} = \sigma^2_{L,K,Q}(e)N \end{aligned} \tag{6.18}$$

Define,

$$\sigma_E = \sqrt{\sigma^2_{L,K,Q}(e)N} = \sigma_{L,K,Q}(e)\sqrt{N}. \tag{6.19}$$

It can be seen from (6.17) that $E_{N,0}(e^{j\omega})$ consists of a summation of N terms. With the operation of the central limit theorem, $E_{N,0}(e^{j\omega})$ becomes Gaussian distributed when N is large. Thus, for large N, $|E_{N,0}(e^{j\omega})|$ is less than $2\sigma_E$ with 98% chance. Thus,

$$\left|E_{N,0}(e^{j\omega})\right| \tilde{\leq} 2\sigma_E, \tag{6.20}$$

where $\tilde{\leq}$ denotes "is 98% chance less than or equal to".

6.2.1 Statistical Boundary of Stopband Attenuation Deterioration

Suppose that $H_{N,0}(e^{j\omega})$ has been designed to minimize the maximum stopband ripple so that

$$\max_{\omega\in[\omega_s,\pi]} |H_{N,0}(e^{j\omega})| = \delta, \tag{6.21}$$

where, δ is the maximum stopband ripple. Therefore, for all ω

$$\begin{aligned} |H^*_{N,0}(e^{j\omega})| &\leq |H^*_{N,0}(e^{j\omega}) - H_{N,0}(e^{j\omega})| + |H_{N,0}(e^{j\omega})| \\ &\leq \max_{\omega\in[\omega_s,\pi]} |E_{N,0}(e^{j\omega})| + \delta \\ &\widetilde{\leq} 2\sigma_{L,K,Q}(e)\sqrt{N} + \delta. \end{aligned} \tag{6.22}$$

Let D^* and D denote the minimum stopband attenuation in dB of the rounded coefficient filter and the infinite precision coefficient filter, respectively. Thus,

$$\begin{aligned} D &= -20\log_{10}\left(\max_{\omega\in[\omega_s,\pi]} |H(e^{j\omega})|\right) \\ &= -20\log_{10}(\delta), \end{aligned} \tag{6.23}$$

and

$$\begin{aligned} D^* &= -20\log_{10}\left(\max_{\omega\in[\omega_s,\pi]} \left|H^*(e^{j\omega})\right|\right) \\ &\widetilde{\geq} -20\log_{10}(2\sigma_{L,K,Q}(e)\sqrt{N} + \delta). \end{aligned} \tag{6.24}$$

From (6.23), δ can be written as

$$\delta = 10^{-D/20}. \tag{6.25}$$

Substitute δ into (6.24), it can be shown that

$$D^* \widetilde{\geq} -20\log_{10}\left(10^{-D/20} + 2\sigma_{L,K,Q}(e)\sqrt{N}\right). \tag{6.26}$$

From (6.23) and (6.26), we arrive at

$$D - D^* \widetilde{\leq} 20\log_{10}\left(1 + 2\cdot 10^{D/20}\sigma_{L,K,Q}(e)\sqrt{N}\right). \tag{6.27}$$

(6.27) gives a statistical bound on the increase in ripple magnitude due to filter coefficient quantization. However, this bound is conservative since it assumes

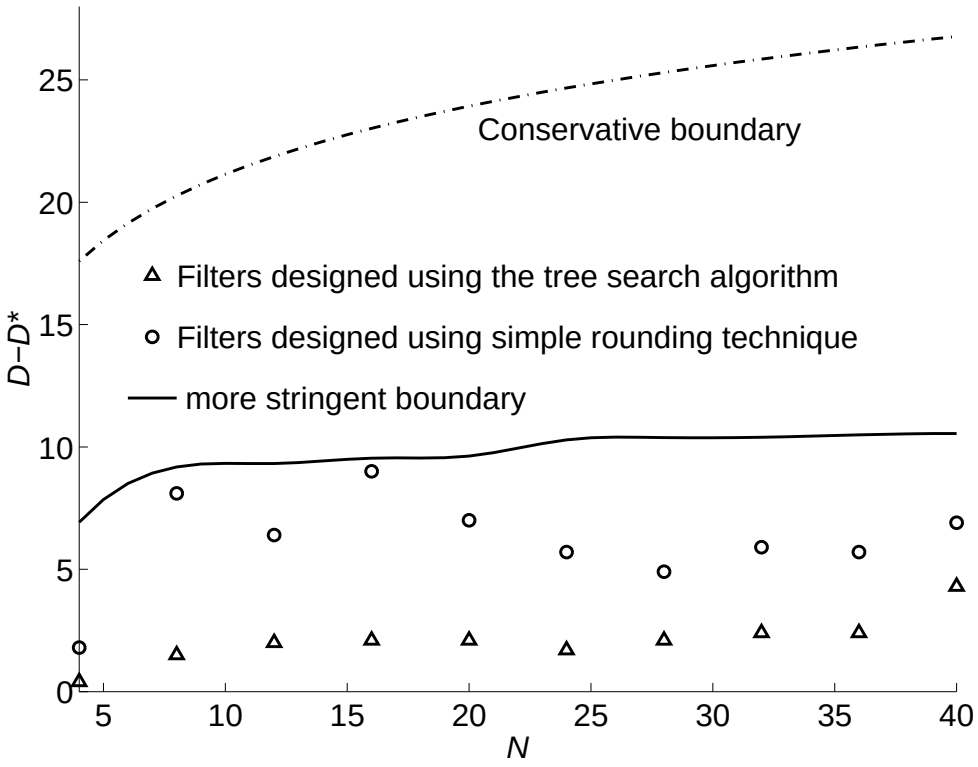

Fig. 6.5: Comparison between experimental data and predicted statistical bound for the stopband attenuation.

that the largest power-of-two terms of all the coefficients are 2^{L-1}. In fact, the magnitude of the coefficient values, $|\alpha_k|$, decreases with increasing k, and usually, the largest power-of-two terms of smaller coefficients are less than 2^{L-1}.

Fig. 6.5 shows a $(D - D^*)$ versus N plot for several examples of filter banks with $Q = -7$, $D = 30.5$dB and N ranging from 4 to 40. In Fig. 6.5, the filters obtained using simple coefficient rounding are denoted using 'o' and those obtained using the depth-first width-recursive algorithm are denoted using '△'. The function $20\log_{10}\left(1 + 2 \cdot 10^{D/20}\sigma_{L,K,Q}(e)\sqrt{N}\right)$, where $L = 2, K = 2$ and $Q = -7$ is also plotted (dash-dot curve) in Fig. 6.5. It can be seen from Fig. 6.5 that the examples have $(D - D^*)$ values significantly smaller than $20\log_{10}\left(1 + 2 \cdot 10^{D/20}\sigma_{L,K,Q}(e)\sqrt{N}\right)$. The $(D - D^*)$ values for those designed using the tree search algorithm are significantly smaller than the $(D - D^*)$ values for those obtained using simple coefficient rounding.

A more stringent bound can be obtained if the continuous design is known *a priori*. Supposing that the permitted smallest power-of-two term is not smaller than 2^Q, and L_k is selected such that 2^{L_k-1} is the largest power-of-two term for the coefficient α_k. Thus, α_k can be represented by an $L_k - Q$ bit SPT number. As it has been assumed that each coefficient is allocated with K SPT terms, for the coefficient α_k, the most significant bit 2^{L_k-1} must be 1 or -1 and it requires one SPT term. For the remaining $(L_k - Q - 2)$ bits, there are $(K - 1)$ SPT terms available for use. The error PDF for the rounding process is $p_{L_k-2,K-1,Q}(e)$ and the error variance is $\sigma^2_{L_k-2,K-1,Q}(e)$. Therefore, the variance of frequency response error can be written as

$$\sigma^2_{E_{N,0}} = \sum_{k=0}^{N-1} \frac{1}{(1+\alpha_k^2)^2} \times \sigma^2_{L_k-2,K-1,Q}(e). \tag{6.28}$$

It meets the requirement described in (6.15), i.e., $(L_k - Q - 2)$ may be less than or equal to 0 for $\sigma^2_{L_k-2,K-1,Q}$. If $L_k - 2 - Q$ is equal to 0 or -1, then the largest power-of-two term for α_k $\left(2^{L_k-1}\right)$ is 2^{Q+1} or 2^Q, respectively; those bits with weights less than or equal to 2^{L_k-3} constitute rounding errors. Therefore, the rounding error of the coefficient is uniformly distributed in the range of $\left[-M^+_{(L_k-2)\infty}, M^+_{(L_k-2)\infty}\right]$ and the variance, according to (6.41), is given by $\frac{\left(M^+_{(L_k-2)\infty}\right)^2}{3}$. If $L_k - 2 \leq -2$, the largest power-of-two term for α_k is less than 2^Q and cannot be represented by any SPT term. Therefore, the rounding error is distributed in the range of $\left[-M^+_{(L_k)\infty}, M^+_{(L_k)\infty}\right]$, i.e., $\left[-M^+_{(L_k-2+2)\infty}, M^+_{(L_k-2+2)\infty}\right]$. Thus, the variance is $\frac{M^2_{(L_k-2+2)\infty}}{3}$. This qualifies (6.15).

Hence, define

$$\sigma'_E = \sqrt{\sum_{k=0}^{N-1} \frac{1}{(1+\alpha_k^2)^2} \times \sigma^2_{L_k-2,K-1,Q}(e)}. \tag{6.29}$$

Substitute $\sigma_{L,K,Q}(\Delta\alpha_k)\sqrt{N}$ in (6.27) with (6.29), a more stringent bound is given by

$$D - D^* \widetilde{\leq} 20\log_{10}\left(1 + 2\cdot 10^{D/20}\sqrt{\sum_{k=0}^{N-1} \frac{\sigma^2_{L_k-2,K-1,Q}(e)}{(1+\alpha_k^2)^2}}\right). \tag{6.30}$$

It should be noted that the application of the central limit theorem requires that

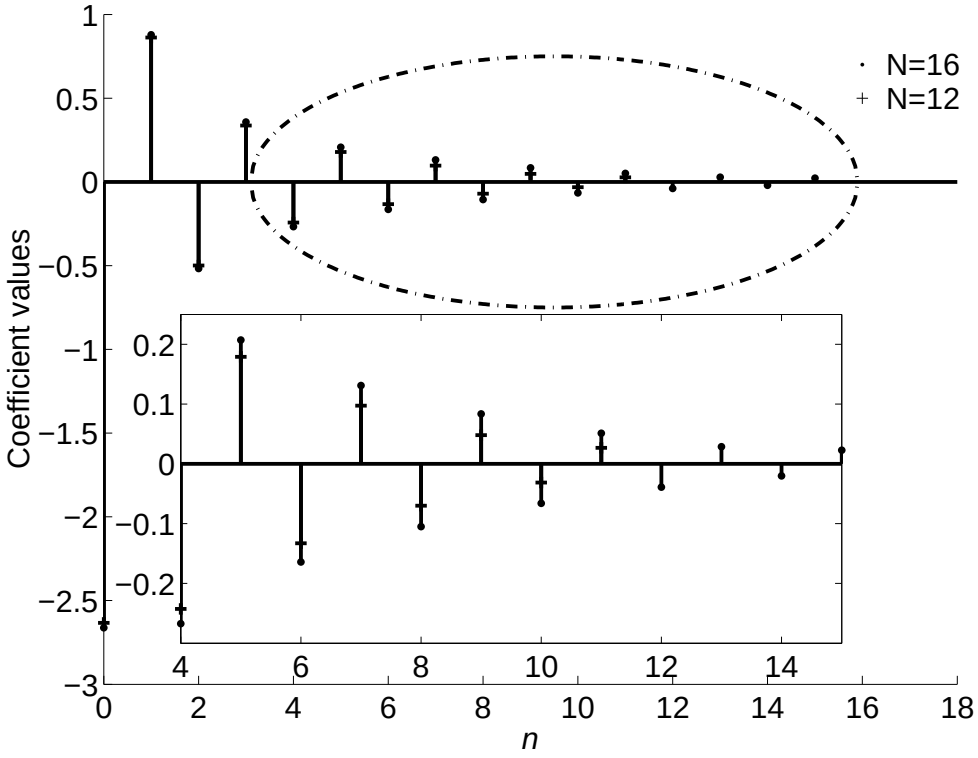

Fig. 6.6: The lattice coefficient values for $N = 12$ and $N = 16$ when $D = 30.5$dB.

the PDF of all the variables involved in the summation are to be the same. In (6.30), the PDF's of the variables are not the same and hence the central limit theorem does not apply in the strictest sense. Nevertheless, in order to facilitate theoretical analysis, the central limit theorem is employed.

The more stringent bound in (6.30) is also plotted in Fig. 6.5 in solid line, which is smoothed by piecewise cubic spline interpolation. No example has $(D-D^*)$ value larger than the boundary. It is interesting to note that this boundary is flat and increases very slowly with increasing N, which is consistent with observed results. The reason for this phenomenon is that for a given D, it has been observed that the magnitudes of the coefficients increase with increasing filter length. The coefficient values for an example with $D = 30.5$dB and $N = 12$ or 16 are shown in Fig. 6.6. If the increase in the magnitude of the coefficient does not result in more bits to represent the new coefficient value, the coefficient rounding error variance remains unchanged. Since the magnitude of α_k diminishes with increase k, the additional

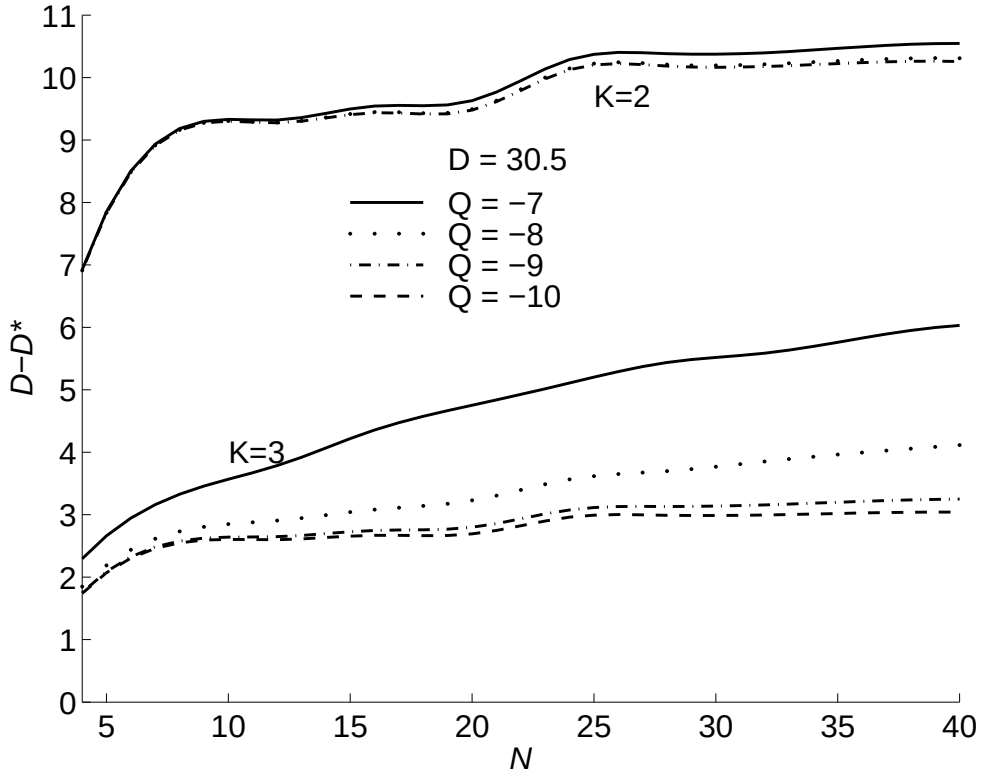

Fig. 6.7: $D - D^*$ versus N plot for $K = 2$ and 3 and $Q = -7, -8, -9$ and -10.

coefficients causes small additional quantization errors.

Decreasing Q from -7 to -9 while maintaining $K = 2$ only lowers the bound by a small amount as shown in Fig. 6.7, whereas increasing K by 1, causes a big drop on the statistical bound. It can be seen that when $K = 3$, the bounds for $Q = -7$ and $Q = -8$ increases more rapidly with increasing N than those for $Q = -9$ and $Q = -10$. This can be explained as follows: as N increases, the magnitudes of those additional coefficients are small and decrease with increases N. Some of these small magnitude coefficients have values smaller than 2^{-5} and so are represented using less than three SPT terms when Q=−7; the quantization step could not decrease as the coefficient magnitude decreases. Thus, $D - D^*$ increases rapidly as N increases. The story is different for $Q = -10$; the small magnitude coefficients will be represented by three SPT terms and so have significantly smaller quantization errors. In this case, as N increases, the increase in $D - D^*$ is very

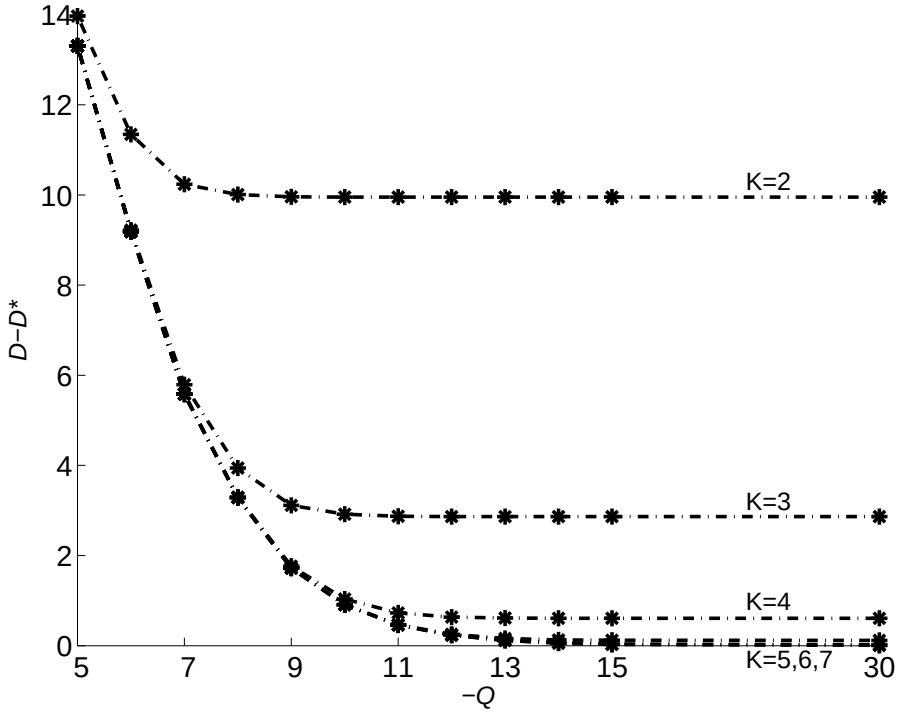

Fig. 6.8: $D - D^*$ versus $-Q$ plot. The minimum stopband attenuation of the infinite precision prototype is 30dB.

small. Hence, the curve for $Q = -10$ is flat.

6.2.2 Effective Selections of Q and K for Coefficient Rounding

From the above analysis, it is clear that if the coefficients of the filter banks are to be quantized to SPT values with fixed number of SPT terms by simple rounding, arbitrarily choosing Q and K values may cause wastage on hardware resources.

Fig. 6.8, Fig. 6.9 and Fig. 6.10 show $(D - D^*)$ values with various K's and Q's for D equal to 30dB, 45dB and 60dB, respectively.

It can be seen from these figures that for a given K, decreasing Q beyond certain value will not decrease $D - D^*$ much further if simple rounding is used as the quantization technique. Similarly, for a given Q, increasing K beyond certain value also will not decrease $D - D^*$ much further.

For a given K, let $d_K = D - D_K^*$ where D_K^* is obtained by setting $Q = -30$ (assuming that 2^{-30} is sufficiently close to zero for practical purpose). Thus, d_K may

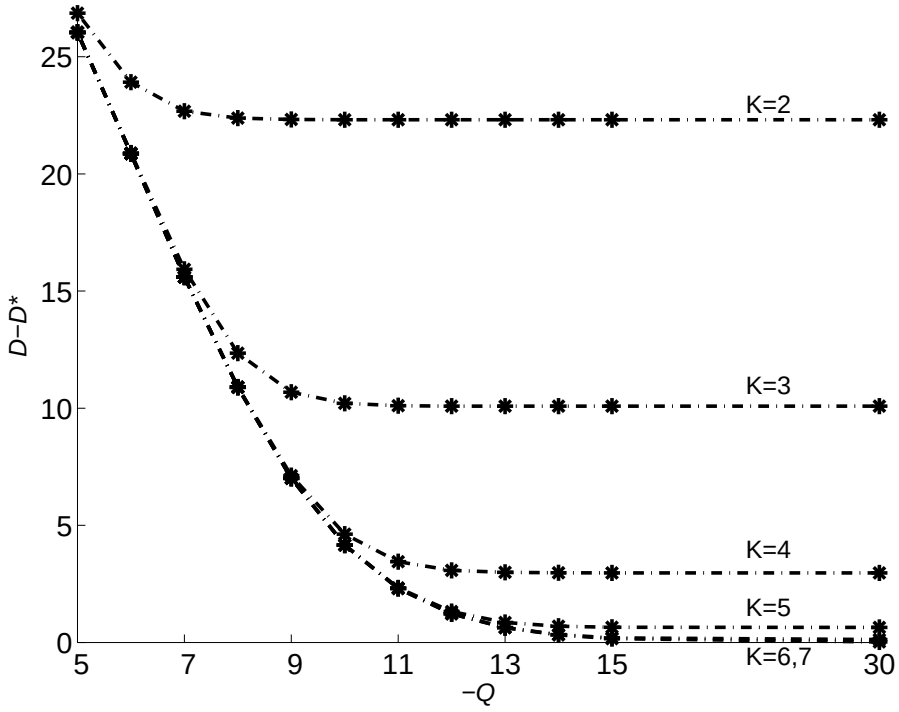

Fig. 6.9: $D - D^*$ versus $-Q$ plot. The minimum stopband attenuation of the infinite precision prototype is 45dB.

be considered as the minimum deterioration when quantizing the filter coefficients to a given K SPT terms. Graphs of Q versus K that will cause a deterioration less than $(d_K + 1)$dB with 98% chance are plotted in Fig. 6.11 for various values of K. Therefore, the chance of having a better than 1dB improvement in the stopband attenuation by using a value of Q smaller that shown in Fig. 6.11 is 2%.

Fig. 6.12 shows the values of K for a given Q, where the chance of achieving a better than 1dB improvement in the stopband attenuation by increasing K is 2%.

It should be noted that the above analysis is applicable for direct rounding of the coefficients of a lattice filter bank to SPT values and all the coefficients have the same number of SPT terms. For other type of filter structures, K and Q may be related in some other ways.

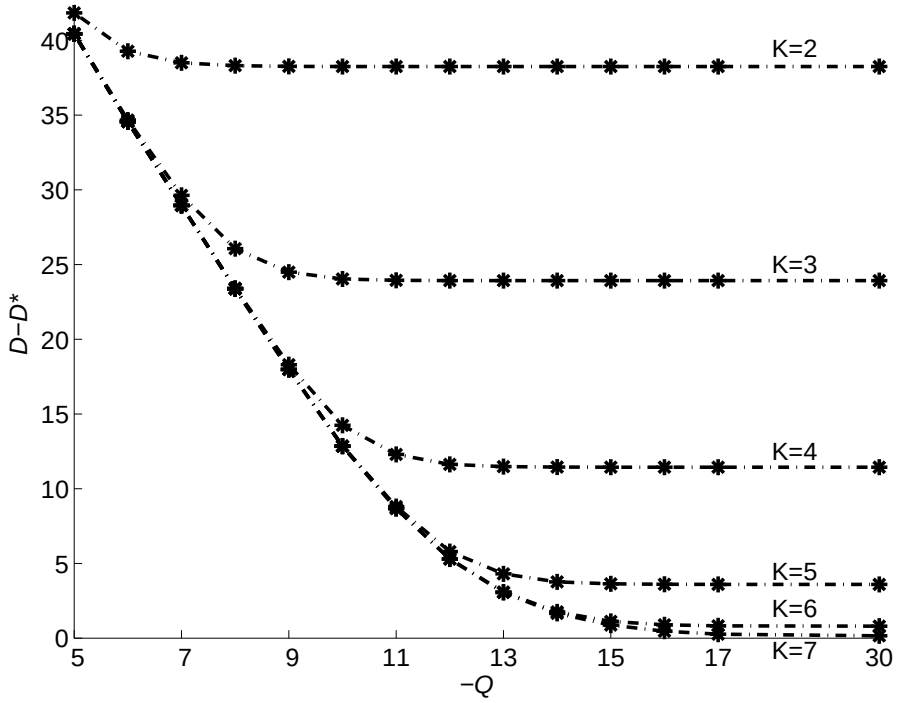

Fig. 6.10: $D - D^*$ versus $-Q$ plot. The minimum stopband attenuation of the infinite precision prototype is 60dB.

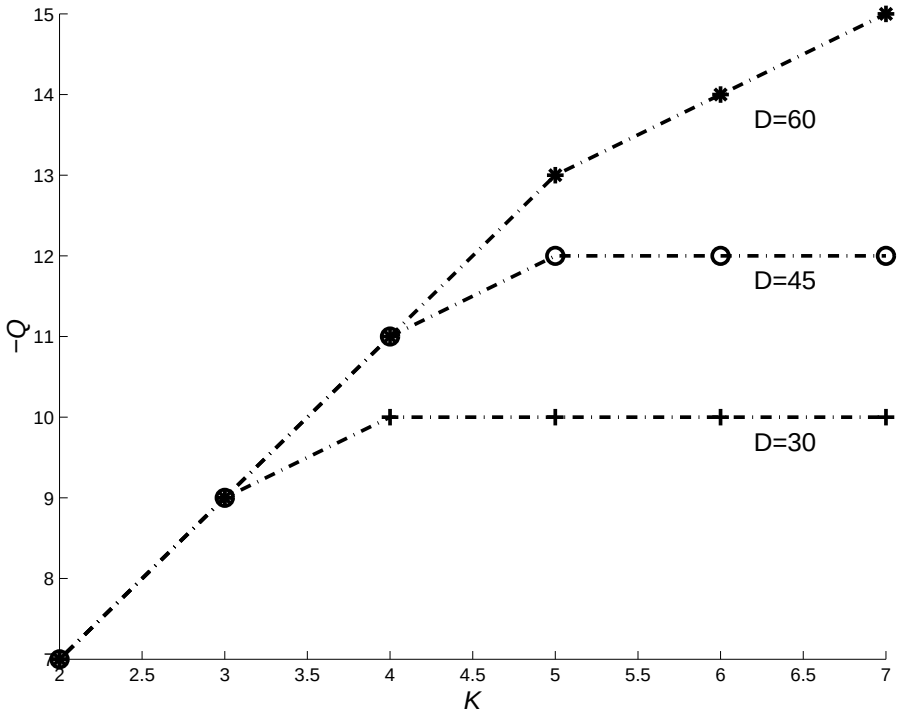

Fig. 6.11: $-Q$ versus K plots where the chance of having a better than 1dB improvement in the stopband attenuation by increasing Q is 2% for a given K.

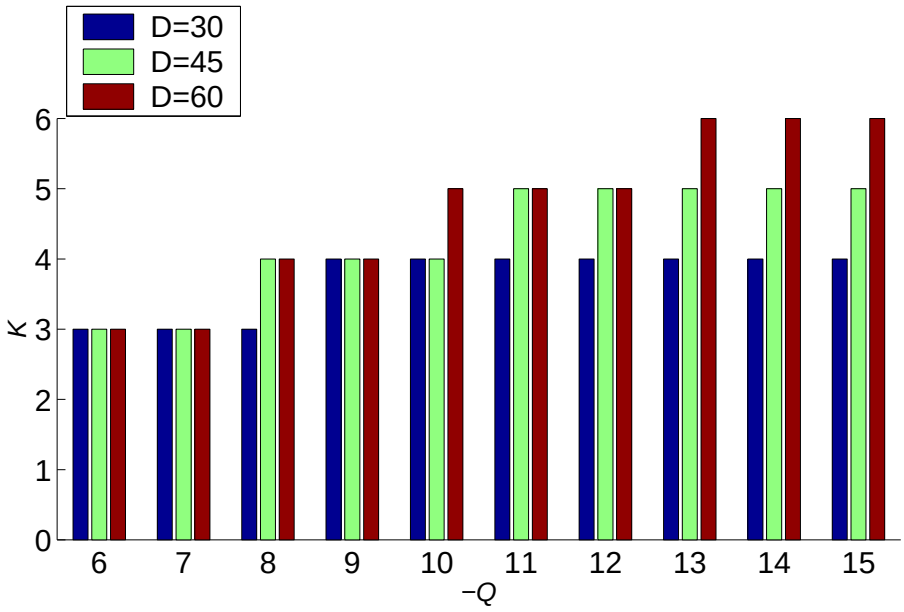

Fig. 6.12: Bar graphs of K versus $-Q$ plots where the chance of achieving a better than 1dB improvement in the stopband attenuation by increasing K is 2% for a given Q.

6.3 SPT Term Allocation Scheme Based on Statistical Analysis

In many applications, it is not necessary to make all the coefficient values to have the same number of SPT terms. In this case, it is desirable to minimize the total number of SPT terms for the entire filter. It has been demonstrated in [47,55] that significant advantage can be achieved if the coefficient values are allocated with different number of SPT terms while keeping the total number of SPT terms fixed. In [55], the SPT terms are assigned based on a statistical analysis on the number of SPT terms required to represent an integer, whereas in [47], the SPT terms are allocated to the coefficients one by one to the currently most deserving coefficient to minimize the L^∞ distance between the SPT coefficients and their corresponding infinite precision values. In this section, an SPT term-allocation scheme based on the statistical analysis on the rounding error distribution is presented.

From the coefficient quantization analysis presented in the previous section, it can be seen that the frequency response deviation of a filter bank caused by

coefficient quantization is bounded by (6.31)

$$D - D^* \lesssim 20 \log_{10} \left(1 + 2 \cdot 10^{D/20} \sigma_{E_{N,0}}\right). \tag{6.31}$$

where,

$$\sigma^2_{E_{N,0}} = \sum_{k=0}^{N-1} \frac{1}{(1+\alpha_k^2)^2} \times f(L_k, K_k, Q). \tag{6.32}$$

In (6.32), the function $f(L_k, K_k, Q)$ is given by

$$f(L_k, K_k, Q) = \begin{cases} \sigma^2_{L_k-2, K_k-1, Q}(e), & \text{for } K_k > 0, \\ \sigma^2_{L_k, K_k, Q}(e), & \text{for } K_k = 0, \end{cases} \tag{6.33}$$

where L_k is selected such that 2^{L_k-1} is the largest power-of-two term of α_k, K_k is the number of SPT terms allocated to α_k, and 2^Q is the smallest power-of-two term allowed for all the coefficients. Equation (6.33) is to be interpreted as follows: When α_k for any given k is allocated at least one SPT term, its rounding error variance is given by $\sigma^2_{L_k-2,K_k-1,Q}(e)$, whereas if no SPT term is allocated to α_k, its rounding error variance is given by $\sigma^2_{L_k,0,Q}(e)$.

For any given set of infinite precision coefficient values, the largest power-of-two term for each coefficient is known, and thus L_k for all k are known. Thus, for a given Q and $\hat{K}$ (the total number of SPT terms for all the coefficients), the problem is to devise a method for determining K_k to minimize the frequency response deterioration. Since $\sigma^2_{E_{N,0}}$ determines the lower bound of the frequency response deterioration as shown in (6.31), the frequency response deviation is minimized when $\sigma^2_{E_{N,0}}$ is minimized.

Since an analytic solution for SPT term allocation that will result in the minimum $\sigma^2_{E_{N,0}}$ is not available, an iterative scheme that assigns one SPT term at a time to the coefficients is proposed. The coefficient to receive an SPT term is the one with the largest value of

$$E^2(\Delta\alpha_k) = \frac{1}{(1+\alpha_k^2)^2} \times f(L_k, K_k, Q). \tag{6.34}$$

The SPT term allocation scheme runs as follow:

1. Let $\hat{K}$ be the total number of SPT terms to be allocated. Let K_k be the number of SPT terms to be assigned to α_k. Initialize $K_k = 0$ for all k. Obtain the infinite precision α_k and select L_k in such a way that the largest power-of-two term of α_k is 2^{L_k-1}, and 2^Q is the permitted smallest power-of-two term.

2. Evaluate $E^2(\Delta\alpha_k)$ according to (6.34) for all k.

3. Let $E^2(\Delta\alpha_i)$ be the largest $E^2(\Delta\alpha_k)$ for all k, i.e., $E^2(\Delta\alpha_i) \geq E^2(\Delta\alpha_j)$ for all $j \neq i$.

4. $K_i = K_i + 1$ and $\hat{K} = \hat{K} - 1$.

5. If $\hat{K} = 0$, stop; otherwise, go to Step 2. □

In the above SPT term allocation scheme, SPT terms are assigned to the coefficient with the largest value for the product of coefficient sensitivity and rounding error variance. For a given Q, the error variance $f(L_k, K_k, Q)$ obtained in (6.33) is determined by L_k, the largest power-of-two term of the coefficient, and K_k, the number of SPT terms which have been assigned to the coefficient. As the rounding error variance was deduced by the assumption that the number under consideration is uniformly distributed in a range determined by L_k, a more accurate estimation of L_k leads to a more accurate error PDF. Therefore, once a coefficient α_k with the largest power-of-two term 2^{L_k-1} is assigned with an SPT term, if the coefficient value is updated by $\alpha_k = \alpha_k - 2^{L_k-1}$ (or $\alpha_k = \alpha_k + 2^{L_k-1}$ depending on the sign of α_k), a new L_k is produced for the new α_k. The error variance estimated by using the new α_k and L_k will be closer to the actual error distribution. A modified SPT term allocation scheme is derived as follows:

1. Let $\hat{K}$ be the total number of SPT terms to be allocated. Let K_k be the number of SPT terms to be assigned to α_k. Initialize $K_k = 0$ for all k. Obtain the infinite precision α_k and select L_k in such a way that the largest

power-of-two term for α_k is 2^{L_k-1}, and 2^Q is the permitted smallest power-of-two term. Evaluate the coefficient sensitivity $S_k^2 = \frac{1}{(1+\alpha_k^2)^2}$.

2. Evaluate $E^2(\Delta\alpha_k) = S_k^2 \times f(L_k, K_k, Q)$ for all k.

3. Let $E^2(\Delta\alpha_i)$ be the largest $E^2(\Delta\alpha_k)$ for all k, i.e., $E^2(\Delta\alpha_i) \geq E^2(\Delta\alpha_j)$ for all $j \neq i$.

4. $K_i = K_i + 1$, $\hat{K} = \hat{K} - 1$.

5. $\alpha_i = \alpha_i - \text{sign}(\alpha_i) \times 2^{L_i-1}$.

6. L_i is updated such that 2^{L_i-1} is the largest power-of-two term of the new α_i.

7. If $\hat{K} = 0$, stop; otherwise, go to Step 2. □

A series of examples based on the following specifications are selected to illustrate the advantage that can be gained from the proposed allocation scheme over that where all the coefficient are allocated with the same number of SPT terms. Comparisons between the proposed technique and those reported in [47, 55] will also be made. The specifications are:

1) Filter order: $2N - 1$ for N ranging from 16 to 40.
2) Stopband edge: $\omega_s = 0.56\pi$.
3) The allowed smallest power-of-two term: $2^Q = 2^{-10}$.

When each coefficient value is allocated with two SPT terms, the total number of SPT terms allocated is $2N$. Alternatively, these $2N$ SPT terms are allocated to the coefficients using some SPT term-allocation techniques in such a way that different coefficient may have a different number of SPT terms. After obtaining the infinite precision coefficients, the coefficients are rounded using the proposed SPT term allocation scheme. The stopband attenuation deterioration, $(D - D^*)$, are plotted in Fig. 6.13 for N ranging from 16 to 40 for the following cases:

1) The number of SPT terms allocated to each coefficient is not more than two.

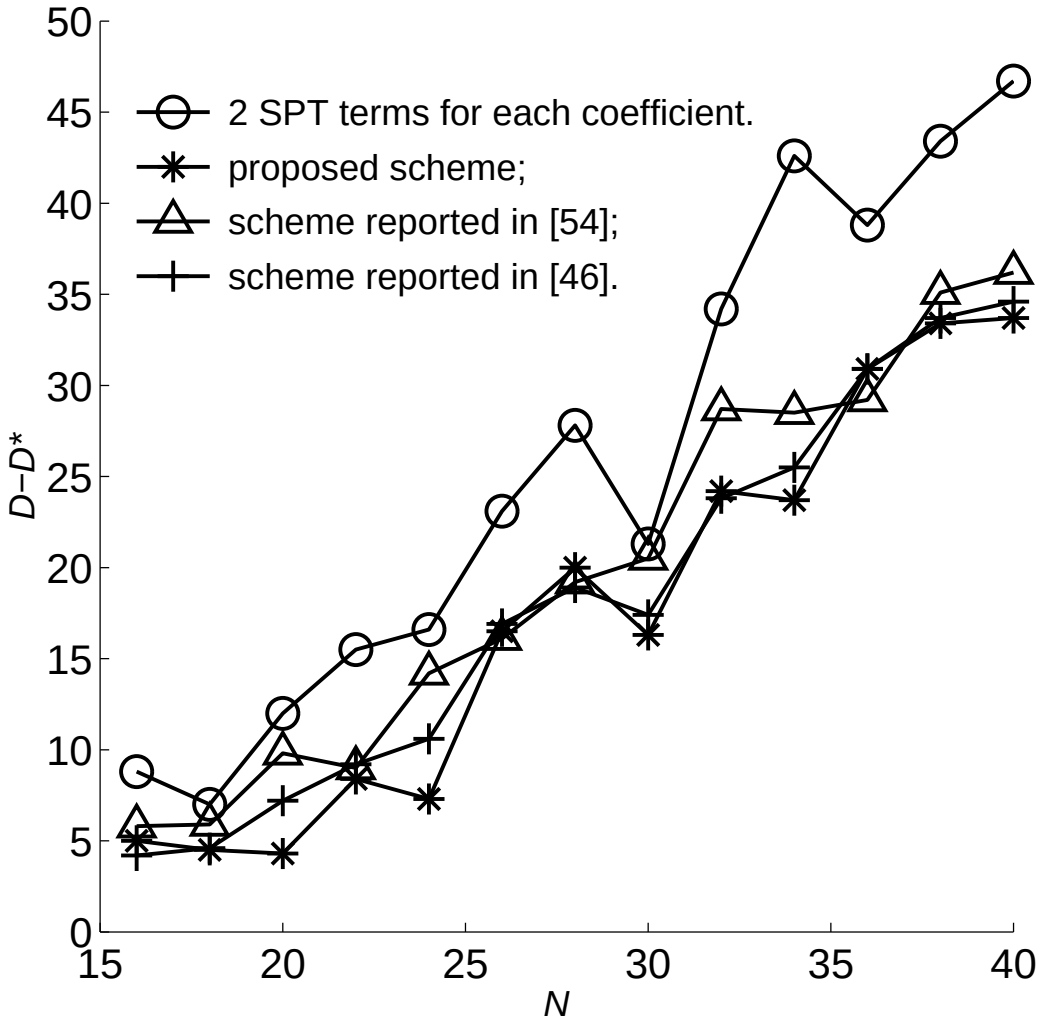

Fig. 6.13: In the proposed scheme and those schemes reported in [55] and [47], each coefficient values is allocated with a different number of SPT terms such that the average number of SPT terms per coefficient is two.

2) The average number of SPT terms allocated to each coefficient is not more than two. The SPT terms are allocated using the proposed allocation scheme.
3) The average number of SPT terms allocated to each coefficient is not more than two. The SPT terms are allocated using the allocation scheme reported in [55].
4) The average number of SPT terms allocated to each coefficient is not more than two. The SPT terms are allocated using the allocation scheme reported in [47].

Fig. 6.13 shows that the proposed allocation scheme produced filters with very much smaller deterioration when compared with the allocation scheme where each coefficient is allocated with two SPT terms. The improvement in the reduction of the stopband attenuation deterioration ranges from a few dB to approximately 20dB. The improvement increases with increasing N although the increase is not monotonous. In Fig. 6.13, the largest stopband attenuation improvement is 18.9dB occurring at $N = 34$. Furthermore, compared with the allocation schemes reported in [47, 55], the proposed allocation scheme produced the smallest deterioration for the majority of the cases.

6.4 Incorporating the SPT Allocation Scheme with the Tree Search Algorithm

In Chapter 5, a depth-first width-recursive tree search algorithm was proposed to design SPT coefficient filter banks. In that algorithm, each coefficient was allocated with the same number of SPT terms. The tree search algorithm was also suitable for optimizing the lattice filter bank with different number of SPT terms for each coefficient by incorporating an SPT term allocation scheme. In this section, the tree search algorithm incorporating the SPT term allocation scheme reported in Section 6.3 is discussed. Examples are included to show the efficiency of the combination of these two techniques.

In the algorithm described in Chapter 5, when a node of the tree is created, the deterioration measures of the coefficients which have not been quantized are evaluated. The coefficient with the largest deterioration measure is fixed to its nearest discrete value with a pre-determined number of SPT terms. The remaining infinite precision coefficients are reoptimized. This particular coefficient may be fixed to other discrete values in subsequent search until a pre-determined tree width is achieved.

To incorporate the proposed SPT term allocation scheme to the tree search algorithm, after a node of the tree is created and the deterioration measures of all

the unquantized coefficients are evaluated, the SPT allocation procedure described in Section 6.3 is performed to all the infinite precision coefficients subject to a total number of SPT terms. This total number of SPT terms is the predetermined total number of SPT terms less the number of SPT terms which have been assigned to the quantized coefficients at the current node. The coefficient with the largest deterioration measure is fixed to the nearest discrete value with the allocated number of SPT terms. The remaining infinite precision coefficients are reoptimized. This particular coefficient may also be fixed to other discrete values with the allocated number of SPT terms in subsequent search until a pre-determined tree width is achieved.

The tree search algorithm incorporating the SPT term allocation scheme is used to design the set of filter banks which have been designed using the tree search algorithm alone as reported in Section 5.3. The specifications are repeated here:

1) Filter order: $2N - 1$ for N ranging from 16 to 32.
2) Stopband edge: $\omega_s = 0.56\pi$.
3) Average number of SPT terms allocated to each coefficient: not more than two.
4) The allowed smallest power-of-two term: $2^Q = 2^{-10}$.

The stopband attenuations versus N are plotted for N ranging from 16 to 32 for the following cases, as shown in Fig. 6.14:

a) Each coefficient value is represented to infinite precision.
b) Tree search algorithm incorporating the proposed SPT term allocation scheme. The average number of SPT terms allocated to each coefficient is not more than two.
c) Tree search algorithm incorporating the proposed SPT term allocation scheme. The number of SPT terms allocated to each coefficient is not more than two.

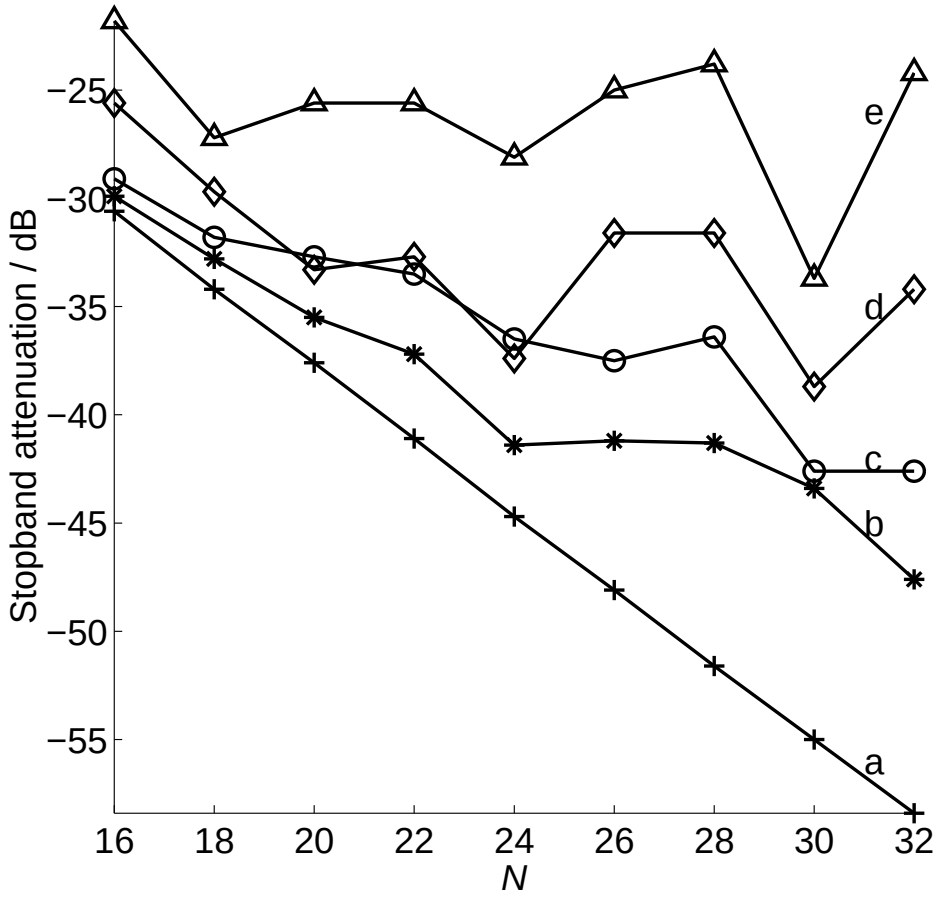

Fig. 6.14: Stopband attenuations. a) Infinite precision design; b) Tree search design where the average number of SPT terms is not more than two per coefficient; c) Tree search design where the number of SPT terms for each coefficient is not more than two; d) Simple rounding result where the average number of SPT terms is not more than two per coefficient; e) Simple rounding result where the number of SPT terms for each coefficient is not more than two.

d) Simple rounding. The average number of SPT terms allocated to each coefficient is not more than two.

e) Simple rounding. The number of SPT terms allocated to each coefficient is not more than two.

It can be seen from Fig. 6.14 that compared to the tree search design where each coefficient is allocated with not more than two SPT terms, the tree search algorithm incorporating the proposed SPT allocation scheme produces filters with very much larger stopband attenuation.

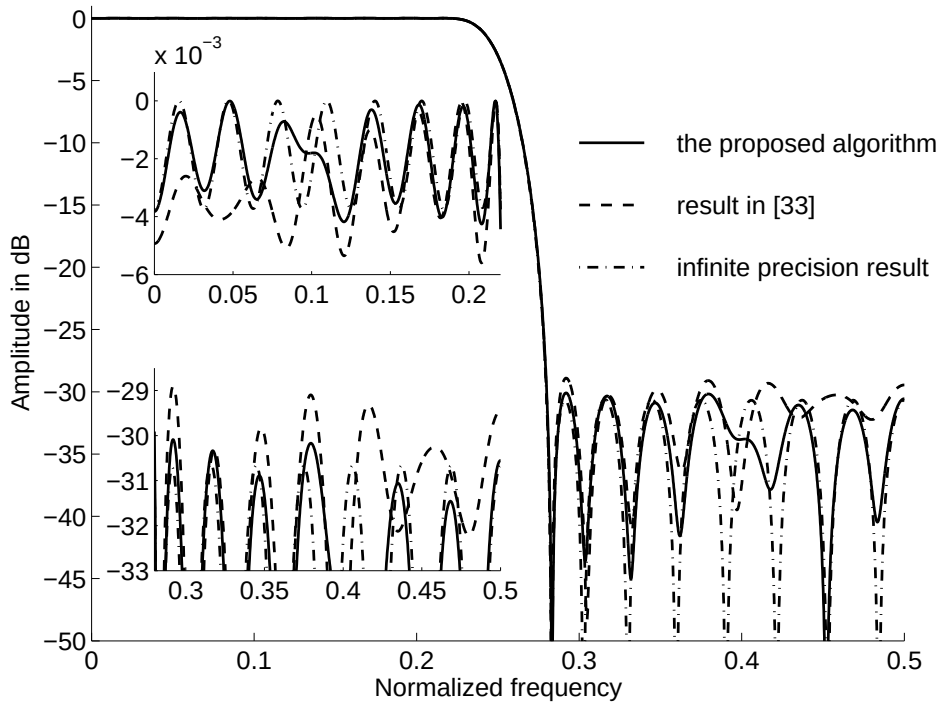

Fig. 6.15: Frequency responses of the analysis filters of the 31-th order filter bank with stopband edge at 0.56π.

Fig. 6.15 shows the frequency responses of the analysis filters of the 31-th order filter bank. The infinite precision optimum solution has a peak stopband gain of -30.7dB and the coefficient values are tabulated in column two of Table 6.2. This example has been designed in Section 5.4 using the tree search algorithm and each of the coefficients is allocated with two SPT terms. The stopband attenuation achieved is 29.1dB. In this section, the filter bank is designed using the tree search algorithm incorporating the proposed SPT allocation scheme and not more than 32 SPT terms are allowed for the entire filter. Reference [34] reported solutions with 28.9dB attenuation in the stopband. The proposed technique produces a design with a peak stopband gain of -29.9dB. The coefficient values for the design obtained using the proposed algorithm were tabulated in column three of Table 6.2. The frequency responses of the filters are shown in Fig. 6.15.

Another example taken from reference [34] is also designed. It is a 47-th order filter bank with stopband edge at $\omega_s = 0.605\pi$. The infinite precision optimum

Table 6.2: Coefficient values of the 31-th order filter bank, whose stopband edge is $\omega_s = 0.56\pi$.

k	Continuous coefficients	SPT coefficient. Each coefficient has on average two SPT terms.
0	−2.6619195	$-2^{+1}-2^{-1}-2^{-3}$
1	0.8784588	$2^{+0}-2^{-3}-2^{-5}+2^{-8}$
2	−0.5167097	-2^{-1}
3	0.3580536	$2^{-1}-2^{-3}-2^{-5}$
4	−0.2670765	$-2^{-2}-2^{-8}$
5	0.2072396	$2^{-2}-2^{-4}+2^{-7}+2^{-9}$
6	−0.1640125	$-2^{-3}-2^{-5}$
7	0.1310766	2^{-3}
8	−0.1049166	$-2^{-3}+2^{-5}-2^{-8}$
9	0.0835565	$2^{-4}+2^{-6}$
10	−0.0659682	-2^{-4}
11	0.0510935	$2^{-4}-2^{-6}$
12	−0.0388140	-2^{-5}
13	0.0286118	2^{-5}
14	−0.0201751	-2^{-6}
15	0.0228699	2^{-6}

solution has a peak stopband gain of −74.7dB and the coefficient values are tabulated in column two of Table 6.3. For the SPT coefficient design, not more than 100 SPT terms are allocated to the entire filter coefficients. Reference [34] reported a solution with 71.8dB attenuation in the stopband. The proposed technique produces a design with a peak stopband gain of −74.0dB. The frequency responses of the filters are plotted in Fig. 6.16. The coefficient values for the design using the proposed algorithm were tabulated in Table 6.3.

Table 6.3: Coefficient values of the 47-th order filter bank, whose stopband edge is $\omega_s = 0.605\pi$.

k	Continuous coefficients	SPT coefficient. Entirely 100 SPT terms
0	−6.2685829	$-2^{+3} + 2^{+1} - 2^{-1} - 2^{-3} + 2^{-5} + 2^{-8}$
1	2.0661994	$2^{+1} + 2^{-3} + 2^{-5} + 2^{-7} + 2^{-14}$
2	−1.2148287	$-2^{+0} - 2^{-2} - 2^{-6}$
3	0.8438989	$2^{+0} - 2^{-3} - 2^{-11} - 2^{-14}$
4	−0.6336645	$-2^{-1} - 2^{-3} - 2^{-5} + 2^{-8} - 2^{-10} - 2^{-12}$
5	0.4965643	$2^{-1} + 2^{-7} + 2^{-9} + 2^{-11}$
6	−0.3988496	$-2^{-1} + 2^{-3} - 2^{-5} - 2^{-9} - 2^{-11}$
7	0.3247923	$2^{-2} + 2^{-4} + 2^{-6} + 2^{-8} + 2^{-14}$
8	−0.2661209	$-2^{-2} - 2^{-5} + 2^{-7} + 2^{-9} - 2^{-12}$
9	0.2181214	$2^{-2} - 2^{-5} + 2^{-8} - 2^{-13}$
10	−0.1779513	$-2^{-2} + 2^{-4} + 2^{-7} - 2^{-9} + 2^{-13}$
11	0.1438487	$2^{-3} + 2^{-5} - 2^{-7} - 2^{-9} + 2^{-12} + 2^{-15}$
12	−0.1146933	$-2^{-3} + 2^{-7} + 2^{-13} - 2^{-15}$
13	0.0897830	$2^{-3} - 2^{-5} - 2^{-9}$
14	−0.0686552	$-2^{-4} - 2^{-7}$
15	0.0510080	$2^{-4} - 2^{-7} - 2^{-9} - 2^{-11} + 2^{-13}$
16	−0.0365852	$-2^{-5} - 2^{-7} + 2^{-9} - 2^{-11} - 2^{-14}$
17	0.0251540	$2^{-5} - 2^{-7} + 2^{-9} + 2^{-11} + 2^{-13}$
18	−0.0164216	$-2^{-6} - 2^{-9} + 2^{-11} + 2^{-14}$
19	0.0100694	$2^{-6} - 2^{-8} - 2^{-10} - 2^{-12}$
20	−0.0056912	$-2^{-7} + 2^{-9} - 2^{-13}$
21	0.0029051	$2^{-8} - 2^{-10} + 2^{-13}$
22	−0.0012505	$-2^{-9} + 2^{-11} + 2^{-13}$
23	0.0004620	2^{-11}

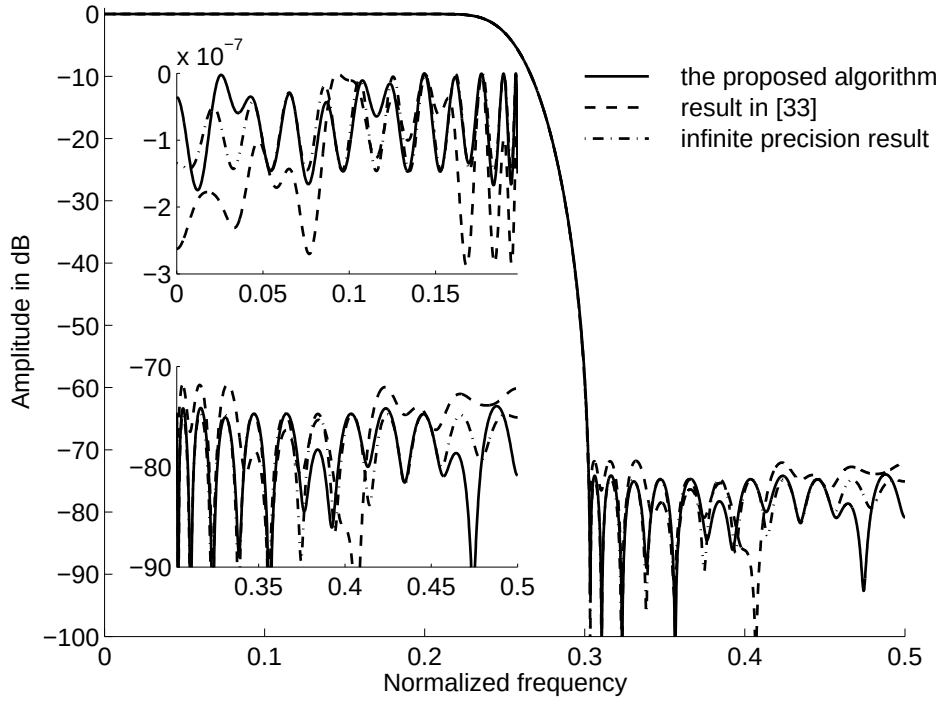

Fig. 6.16: Frequency responses of the analysis filters of the 47-th order filter bank.

6.5 Conclusion

In this chapter, the error distribution for quantizing infinite precision numbers to SPT values for a given L, K, Q is deduced, where K is the number of SPT terms, 2^{L-1} and 2^Q are the largest and smallest power-of-two terms, respectively. The error probability density function is an even symmetrical piecewise constant function. It has zero mean value. The variance of the error is also derived. The effect of quantizing the coefficients to SPT values for the filter bank is analyzed statistically based on the knowledge of the quantization error PDF. Statistical bounds on the stopband attenuation deterioration for quantizing the coefficient values to SPT values are developed. Guidelines for the selection of proper K and Q are developed for simple rounding technique.

Based on the statistical analysis, an SPT term allocation scheme is also developed to design filters with different number of SPT terms. This allocation scheme

shows superiority compared with those reported in previous literatures. The width-recursive depth-first tree search algorithm incorporating the SPT term allocation scheme produces excellent results when it is employed to design the examples taken from published literatures.

Appendix A: Proof for Property 6.2

Proof: According to Lemma 6.2 in Appendix C, it is noted that (6.7) is true for $L = 2K$. For $L \geq 2K + 1$, (6.7) is proved by mathematical induction. According to (6.3), when $L = 2K + 1$, $M^+(L, K + 1) = 2M^+(L - 1, K) + 1$. Since $M^+(L, K + 1) = M^+(L, K) + N^+(L, K + 1)$, therefore,

$$2N^+(L, K + 1) - 1 = 4M^+(L - 1, K) - 2M^+(L, K) + 1. \tag{6.35}$$

Furthermore, according to Lemma 6.4 in Appendix C, (6.53) is true when $L = 2K + 1$. Substituting (6.35) into (6.53), we have

$$p_{L,K}(e|e \in \pm[2^{-1}, 2^0]) = \frac{4M^+(L - 1, K) - 2M^+(L, K) + 1}{2M^+_{L\infty}}. \tag{6.36}$$

Therefore, (6.53) can be written as

$$p_{L,K}(e) = \begin{cases} \dfrac{2M^+(L, K) + 1}{2M^+_{L\infty}}, & \text{for } e \in [-2^{-1}, 2^{-1}], \\ \dfrac{4M^+(L - 1, K) - 2M^+(L, K) + 1}{2M^+_{L\infty}}, & \text{for } e \in \pm[2^{-1}, 2^0], \\ \dfrac{1}{2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^0, \dfrac{2^2}{3}\right], \\ 0, & \text{otherwise.} \end{cases} \tag{6.37}$$

Hence, (6.7) is true for $L = 2K + 1$.

Assume that the error PDF for rounding a number x to an element in $S(L, K)$ for $L \geq 2K + 1$ is given by (6.7), where $x \in \{x| - M^+_{L\infty} \leq x \leq M^+_{L\infty}, x \in \mathcal{R}\}$.

According to Lemma (6.5) in Appendix C,

$$
p_{L+1,K}(e) = \begin{cases}
\dfrac{2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\
\dfrac{4M^+(L,K)-2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \pm\left[2^{-1}, 2^{0}\right], \\
\dfrac{p_{L,K}\left(e \middle| e \in \pm\left[2^{k-1}, 2^{k}\right]\right)}{2}, & \text{for } e \in \pm\left[2^{k}, 2^{k+1}\right], \\
 & k = 0, \cdots, L+1-2K-2, \\
\dfrac{p_{L,K}\left(e \middle| e \in \pm\left[2^{L+1-2K-2}, \frac{2^{L+1-2K}}{3}\right]\right)}{2}, & \\
 & \text{for } e \in \pm\left[2^{L+1-2K-1}, \dfrac{2^{L+1-2K+1}}{3}\right] \\
 & \text{when } L+1-2K-1 \geq 0, \\
0, & \text{otherwise.}
\end{cases}
$$

$$
= \begin{cases}
\dfrac{2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\
\dfrac{4M^+(L,K)-2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \pm\left[2^{-1}, 2^{0}\right], \\
\dfrac{4M^+(L-1-k,K)-2M^+(L-k,K)+1}{2\cdot 2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{k}, 2^{k+1}\right], \\
 & k = 0, \cdots, L+1-2K-2, \\
\dfrac{1}{2\cdot 2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{L+1-2K-1}, \dfrac{2^{L+1-2K+1}}{3}\right] \\
 & \text{when } L+1-2K-1 \geq 0, \\
0, & \text{otherwise.}
\end{cases} \tag{6.38}
$$

Since $2M^+_{L\infty} = M^+_{(L+1)\infty}$, (6.38) can be written as

$$p_{L+1,K}(e) = \begin{cases} \dfrac{2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\ \dfrac{4M^+(L,K)-2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \pm\left[2^{-1}, 2^{0}\right], \\ \dfrac{4M^+(L-k,K)-2M^+(L+1-k,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \pm\left[2^{k-1}, 2^{k}\right], \\ & k = 1, \cdots, L+1-2K-1, \\ \dfrac{1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \pm\left[2^{L+1-2K-1}, \dfrac{2^{L+1-2K+1}}{3}\right] \\ & \text{when } L+1-2K-1 \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

$$= \begin{cases} \dfrac{2M^+(L+1,K)+1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\ \dfrac{4M^+(L+1-1-k,K)-2M^+(L+1-k,K)+1}{2M^+_{(L+1)\infty}}, & \\ & \text{for } e \in \pm\left[2^{k-1}, 2^{k}\right], \\ & k = 0, \cdots, L+1-2K-1, \\ \dfrac{1}{2M^+_{(L+1)\infty}}, & \text{for } e \in \pm\left[2^{L+1-2K-1}, \dfrac{2^{L+1-2K+1}}{3}\right] \\ & \text{when } L+1-2K-1 \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, if (6.7) is true for L where $L \geq 2K+1$, it is also true for $L+1$. Since (6.7) is true for $L = 2K+1$, it is true for all integer $L \geq 2K+1$.

Therefore, the error probability density function for $L \geq 2K$ can be written as (6.7) and thus, Property 6.2 is proved. ■

Appendix B: Proof for Property 6.3

Proof: The error PDF is symmetric with respect to $e = 0$. Therefore, it is obvious that the mean of the errors, denoted as $E(e)$, caused by rounding a number x to an element in $S(L, K)$, is equal to 0, i.e.,

$$E(e) = 0. \tag{6.39}$$

The variance of the error is defined as:

$$\sigma^2(e) = \int_{-\infty}^{\infty} \Big(e - E(e)\Big)^2 p(e)de = \int_{-\infty}^{\infty} e^2 p(e)de. \tag{6.40}$$

The error PDF $p(e)$ is a piecewise constant function. Given zero mean, the variance of any piece of the piecewise constant function, as shown in Fig. 6.17, is

$$\int_a^b ce^2 de = \frac{c}{3}e^3\Big|_a^b = \frac{c}{3}(a^3 - b^3). \tag{6.41}$$

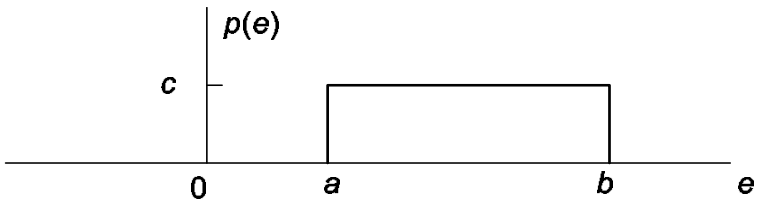

Fig. 6.17: A piece of the error PDF.

When $L = 2K - 1$, it is seen from Property 6.1 that there are three pieces of piecewise constant values in the error PDF, where two of the three pieces have the same value, i.e., $p_{L,K}\left(e\Big|e \in [-3^{-1}, 3^{-1}]\right)$ and $p_{L,K}\left(e\Big|e \in \pm[3^{-1}, 2^{-1}]\right)$. Thus,

$$\begin{aligned}\sigma^2_{L,K}(e) &= \frac{2M_L^+ + 1}{2M_{L\infty}^+} \times \frac{1}{3}\left(\left(\frac{1}{3}\right)^3 - \left(-\frac{1}{3}\right)^3\right) + \frac{M_L^+}{M_{L\infty}^+} \times \frac{2}{3}\left(\left(\frac{1}{2}\right)^3 - \left(\frac{1}{3}\right)^3\right) \\ &= \frac{1}{12} \cdot \frac{27M_L^+ + 4}{27M_{L\infty}^+} = \frac{1}{12} \cdot \frac{27M_L^+ + 4}{27M_L^+ + 9}.\end{aligned} \tag{6.42}$$

When $L = 2K$, it is seen from Lemma 6.2 in Appendix C that there are three pieces of piecewise constant values in the PDF, i.e., $p_{L,K}\left(e\Big|e \in [-2^{-1}, 2^{-1}]\right)$ and

$p_{L,K}\left(e\middle|e \in \pm\left[2^{-1}, \frac{2}{3}\right]\right)$. Thus,

$$\begin{aligned}\sigma^2_{L,K}(e) &= \frac{2M_L^+ + 1}{2M_{L\infty}^+} \times \frac{1}{3}\left(\left(\frac{1}{2}\right)^3 - \left(-\frac{1}{2}\right)^3\right) + \frac{1}{2M_{L\infty}^+} \times \frac{2}{3}\left(\left(\frac{2}{3}\right)^3 - \left(\frac{1}{2}\right)^3\right) \\ &= \frac{1}{12} \cdot \frac{27M_L^+ + 32}{27M_{L\infty}^+} = \frac{1}{12} \cdot \frac{27M_L^+ + 32}{27M_L^+ + 18}. \end{aligned} \tag{6.43}$$

When $L \geq 2K + 1$, from Property 6.2, we have

$$\begin{aligned}\sigma^2_{L,K}(e) &= \frac{1}{2M_{L\infty}^+}\left[\frac{2M^+(L,K)+1}{3} \cdot \left(\left(\frac{1}{2}\right)^3 - \left(-\frac{1}{2}\right)^3\right)\right. \\ &\quad + \sum_{k=0}^{L-2K-1} \frac{2\left(4M^+(L-1-k,K) - 2M^+(L-k,K)+1\right)}{3} \cdot \left(2^{3k} - 2^{3(k-1)}\right) \\ &\quad \left. + \frac{2}{3}\left(\left(\frac{2^{L-2K+1}}{3}\right)^3 - \left(2^{L-2K-1}\right)^3\right)\right] \\ &= \frac{1}{2M_{L\infty}^+}\left[\frac{M^+(L,K)}{6} + \frac{1}{12} + \frac{74}{81} \cdot 2^{3(L-2K-1)}\right. \\ &\quad \left. + \sum_{k=0}^{L-2K-1} \left(4M^+(L-1-k,K) - 2M^+(L-k,K)+1\right) \cdot \frac{7}{12} \cdot 2^{3k}\right]. \end{aligned} \tag{6.44}$$

Therefore, for $L \geq 2K + 1$, from (6.44), we arrive at

$$\begin{aligned}\sigma^2_{L,K}(e) &= \frac{1}{2M_{L\infty}^+}\left[\left(\frac{7}{3}M^+(2K,K) + \frac{128}{81}\right)2^{3(L-2K-1)} - M^+(L,K)\right. \\ &\quad \left. - \sum_{k=0}^{L-2K-2} 7M^+(L-1-k,K)2^{3k}\right]. \end{aligned}$$

■

Appendix C

Lemma 6.1 Given a random number $x \in \{x| - M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ and given $\bar{x} \in S(L,K)$, the rounding error caused by quantizing x to the element in $S(L,K)$ nearest to it is statistically uniformly distributed in $[-2^{-1}, 2^{-1}]$. Let the error PDF be $p_{L,K}\left(e, x\middle| -2^{-1} \leq e \leq 2^{-1}, \bar{x} \in S(L,K)\right)$. The error PDF is unity for $-2^{-1} \leq e \leq 2^{-1}$. □

Proof: For any number $x \in \{x|-M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$, since $\bar{x} \in S(L,K)$, if $\bar{x} \neq \pm M_L^+$, the rounding error PDF is unity in $[-2^{-1}, 2^{-1}]$. The probability of

$\bar{x} \neq \pm M_L^+$ given that $\bar{x} \in S(L,K)$ is $\frac{2M_L^+-1}{2M_L^+}$. The probabilities of $\bar{x} = M_L^+$ or $\bar{x} = -M_L^+$ given that $\bar{x} \in S(L,K)$ are $\frac{0.5}{2M_L^+}$, respectively. The rounding error PDF in each of $[0, 0.5]$ and $[-0.5, 0]$ is 2. Therefore,

$$p_{L,K}\Big(e, x\Big| -2^{-1} \leq e \leq 2^{-1}, \bar{x} \in S(L,K)\Big)$$

$$= \begin{cases} \dfrac{2M_L^+ - 1}{2M_L^+} \times 1 & \text{for } -2^{-1} \leq e \leq 2^{-1}, \\ \dfrac{0.5}{2M_L^+} \times 2 & \text{for } 0 \leq e \leq 2^{-1}, \\ \dfrac{0.5}{2M_L^+} \times 2 & \text{for } -2^{-1} \leq e \leq 0, \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} 1 & \text{for } -2^{-1} \leq e \leq 2^{-1}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.45}$$

■

Lemma 6.2 For $L = 2K$, the error PDF for rounding a number $x \in \{x| - M_{L\infty}^+ \leq x \leq M_{L\infty}^+, x \in \mathcal{R}\}$ to the member in $S(L,K)$ nearest to it, denoted as $p_{L,K}(e)$, is as follows:

$$p_{L,K}(e) = \begin{cases} \dfrac{2M_L^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in \left[-\dfrac{1}{2}, \dfrac{1}{2}\right], \\ \dfrac{1}{2M_{L\infty}^+}, & \text{for } e \in \pm\left[\dfrac{1}{2}, \dfrac{2}{3}\right], \\ 0, & \text{otherwise.} \end{cases} \tag{6.46}$$

□

Proof: The proof is similar to that for Property 6.1. For $L = 2K$, $M_{L\infty}^+ = M_L^+ + \frac{2}{3}$. For any number $x \in \{x| - M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$, we have $\bar{x} \in S(L,K)$. According to Lemma 6.1, the rounding error is uniformly distributed in $[-2^{-1}, 2^{-1}]$ with unity PDF; the probability of $x \in \{x| - M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ is $\frac{2M_L^+}{2M_{L\infty}^+}$. The probabilities of $x \in \{x|M_L^+ \leq x \leq M_{L\infty}^+\}$ and $x \in \{x| - M_{L\infty}^+ \leq x \leq -M_L^+\}$ are both $\frac{1}{3M_{L\infty}^+}$, and the rounding errors are uniformly distributed in $\left[0, \frac{2}{3}\right]$ and $\left[-\frac{2}{3}, 0\right]$,

respectively, with PDF equal to $\frac{3}{2}$. Therefore,

$$p_{L,K}(e) = \begin{cases} \frac{2M_L^+}{2M_{L\infty}^+} \times 1, & \text{for } e \in [-2^{-1}, 2^{-1}], \\ \frac{1}{3M_{L\infty}^+} \times \frac{3}{2}, & \text{for } e \in \left[0, \frac{2}{3}\right], \\ \frac{1}{3M_{L\infty}^+} \times \frac{3}{2}, & \text{for } e \in \left[-\frac{2}{3}, 0\right], \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{2M_L^+ + 1}{2M_{L\infty}^+}, & \text{for } e \in \left[-\frac{1}{2}, \frac{1}{2}\right], \\ \frac{1}{2M_{L\infty}^+}, & \text{for } e \in \pm\left[\frac{1}{2}, \frac{2}{3}\right], \\ 0, & \text{otherwise.} \end{cases}$$

■

Lemma 6.3 Given a random number $x \in \{x | -M_{L\infty}^+ \le x \le M_{L\infty}^+, x \in \mathcal{R}\}$, and given $\bar{x} \in T(L, K+1)$, according to (6.1), $\bar{x}$ can be written as $\bar{x} = \sum_{i=0}^{K} y(i)2^{q(i)}$. Given the least significant SPT term of $\bar{x}$ is $y(K)2^{q(K)}$, where $q(K) = L-2K-1 \ge 0$ and $y(K) = 1$ or -1, the error caused by rounding x to the member in $S(L, K)$ nearest to it is uniformly distributed with unity PDF in $\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right] \cup \left[-2^{q(K)}, -2^{q(K)} + 2^{-1}\right]$. □

Proof: According to the given conditions, $\bar{x}$ is an L-bit integer with $K+1$ SPT terms; these $K+1$ SPT terms are occurring at the most significant $2K+1$ bits. Since in canonic representations, there are no two nonzero adjacent bits, the $K+1$ bits corresponding to the $K+1$ SPT terms must occur at every other bit locations.

First, consider the case where x is positive. We have $\bar{x} - 2^{-1} \le x < \bar{x} + 2^{-1}$ and

$$\bar{x} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} + y(K)2^{L-2K-1}. \tag{6.47}$$

From (6.47), it can be shown that

$$\bar{x} - y(K)2^{L-2K-1} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} \in S(L, K), \tag{6.48}$$

and that

$$\bar{x} + y(K)2^{L-2K-1} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} + y(K)2^{L-2K}. \tag{6.49}$$

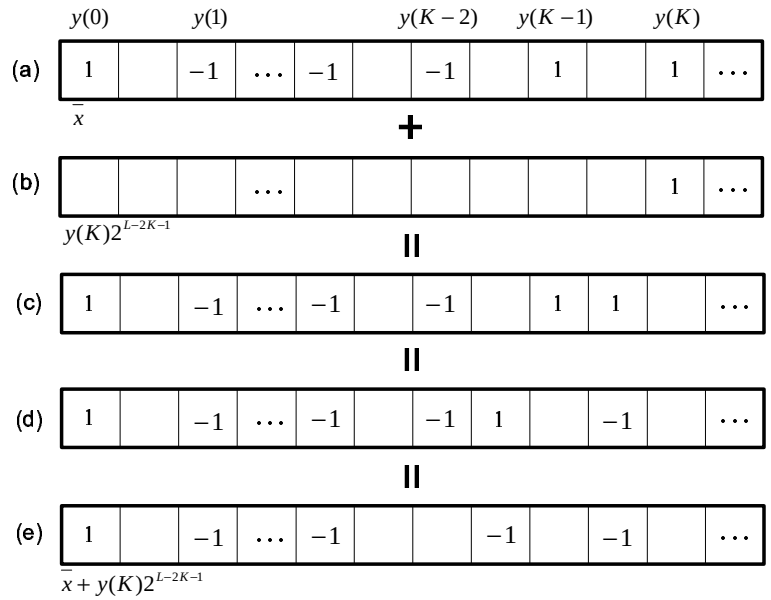

Fig. 6.18: For $\bar{x} = \sum_{i=0}^{K-1} y(i)2^{L-2i-1} + y(K)2^{L-2K-1}$, we have $\bar{x} + y(K)2^{L-2K-1} \in S(L, K)$.

When $\bar{x} \neq \sum_{i=0}^{K} 2^{L-2i-1}$, at least one of $y(i)$ for $i = 0, 1, \cdots, K-1$ is not equal to $y(K)$, i.e., they have different signs. An example of $\bar{x}$ is shown in Fig. 6.18(a), where $y(K) = 1$ and $y(K-2)$ has a different sign from $y(K)$. Thus, $\bar{x} + y(K)2^{L-2K-1}$ has $K+1$ SPT terms as shown in Fig. 6.18(c), where two nonzero bits occur adjacently. From Fig. 6.18(c) to Fig. 6.18(e), it is shown that this number (with $K+1$ SPT terms), $\bar{x} + y(K)2^{L-2K-1}$ as shown in Fig. 6.18(c), can be represented using K SPT terms canonically, as shown in Fig. 6.18(e), provided that there exists an integer $i > 0$ and for all j where $0 \leq j < i$ such that $y(K-i) = -y(K)$ and $y(K-j) = y(K)$. In this example, i equals to 2. Therefore, under the given conditions, $\bar{x} + y(K)2^{L-2K-1}$ is an element in $S(L, K)$, i.e.,

$$\bar{x} + y(K)2^{L-2K-1} \in S(L, K). \tag{6.50}$$

From (6.48) and (6.50), it is shown that

$$\bar{x} \pm y(K)2^{L-2K-1} \in S(L, K). \tag{6.51}$$

Moreover, it is obvious that $\bar{x} \pm 2^{L-2K-j} \notin S(L, K)$ for any $j > 1$. That means any number in $\left(\bar{x} - 2^{L-2K-1}, \bar{x} + 2^{L-2K-1}\right)$ is not in $S(L, K)$, where (a, b)

denotes all the infinite precision numbers in the range bounded by a and b exclusive. Therefore, the result of rounding x to an element in $S(L,K)$ is equal to $\bar{x} \pm 2^{q(K)}$. Since x is uniformly distributed in $\left[\bar{x} - 2^{-1}, \bar{x} + 2^{-1}\right]$, so that rounding x to the element in $S(L,K)$ nearest to it will lead to an error distributed uniformly in $\pm\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right]$, with PDF equal to $\frac{1}{2\left[2^{q(K)} - (2^{q(K)} - 2^{-1})\right]} = 1$. For a negative number x and $\bar{x} \neq -\sum_{i=0}^{K} 2^{L-2i-1}$, the same conclusion can be obtained. The probability of $\bar{x} \neq \pm\sum_{i=0}^{K} 2^{L-2i-1}$ for the given $\bar{x}$ is $\frac{2N^+(L,K+1)-1}{J}$, where $J = 2N^+(L,K+1)+1$ if $\sum_{i=0}^{K} 2^{L-2i-1} + 0.5 \leq M_{L\infty}^+$; otherwise, $J = 2N^+(L,K+1) + 2\left(M_{L\infty}^+ - \sum_{i=0}^{K} 2^{L-2i-1}\right)$.

When $\bar{x} = \sum_{i=0}^{K} 2^{L-2i-1}$ $\left(\text{occurring with probability of } \frac{J-\left(2N^+(L,K+1)-1\right)}{2J}\right)$, x is uniformly distributed in $\left[\bar{x} - 2^{-1}, \bar{x}\right]$. Furthermore, $\bar{x} - 2^{L-2K-1} \in S(L,K)$ and $\bar{x} - 2^{L-2K-j} \notin S(L,K)$ for any $j > 1$. Therefore, the error is uniformly distributed in $\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right]$, with PDF equal to $\frac{1}{2^{q(K)} - (2^{q(K)} - 2^{-1})} = 2$. Similarly, when $\bar{x} = -\sum_{i=0}^{K} 2^{L-2i-1}$ $\left(\text{occurring with probability of } \frac{J-\left(2N^+(L,K+1)-1\right)}{2J}\right)$, the rounding error is uniformly distributed in $\left[-2^{q(K)}, -2^{q(K)} + 2^{-1}\right]$ with PDF equal to 2. Therefore,

$$
\begin{aligned}
& p_{L,K}\left(e, x \middle| e \in \pm\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right], \bar{x} \in T(L,K+1), q(K) = L - 2K - 1\right) \\
&= \begin{cases}
\dfrac{2N^+(L,K+1)-1}{J} \times 1 & \text{for } e \in \pm\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right] \\
\dfrac{J - (2N^+(L,K+1)-1)}{2J} \times 2 & \text{for } e \in \left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right] \\
\dfrac{J - (2N^+(L,K+1)-1)}{2J} \times 2 & \text{for } e \in -\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right] \\
0 & \text{otherwise}
\end{cases} \\
&= \begin{cases}
1 \quad \text{for } e \in \pm\left[2^{q(K)} - 2^{-1}, 2^{q(K)}\right], \\
0 \quad \text{otherwise.}
\end{cases}
\end{aligned} \tag{6.52}
$$

■

Lemma 6.4 For $L = 2K + 1$, the error PDF of rounding a number $x \in \{x | -M_{L\infty}^+ \leq x \leq M_{L\infty}^+, x \in \mathcal{R}\}$ to be an element in $S(L,K)$, denoted as $p_{L,K}(e)$, is as

follows:

$$p_{L,K}(e) = \begin{cases} \dfrac{2M^+(L,K)+1}{2M^+_{L\infty}}, & \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\ \dfrac{2N^+(L,K+1)-1}{2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{-1}, 2^{0}\right], \\ \dfrac{1}{2M^+_{L\infty}}, & \text{for } e \in \pm\left[2^{0}, \dfrac{4}{3}\right], \\ 0, & \text{otherwise.} \end{cases} \tag{6.53}$$

□

Proof: First, $x \in \{x| - M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ is considered. In this case, since $L = 2K+1$, we have $M_L^+ = M^+(L, K+1)$ and $\bar{x} \in S(L,K) \cup T(L,K+1)$. The probability for $x \in \{x| - M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ and $\bar{x} \in S(L,K)$ is $\frac{2M^+(L,K)+1}{2M^+_{L\infty}}$ and is denoted as $P_L\Big(x\Big|\bar{x} \in S(L,K)\Big)$. According to Lemma 6.1, rounding x for $\bar{x} \in S(L,K)$ to an element in $S(L,K)$ leads to an error in $[-2^{-1}, 2^{-1}]$. Therefore,

$$\begin{aligned} & p_{L,K}\left(e\Big|e \in \left[-2^{-1}, 2^{-1}\right]\right) \\ = \; & P_L\Big(x\Big|\bar{x} \in S(L,K)\Big) \times p_{L,K}\left(e, x\Big|e \in \left[-2^{-1}, 2^{-1}\right], \bar{x} \in S(L,K)\right) \\ = \; & \frac{2M^+(L,K)+1}{2M^+_{L\infty}}. \end{aligned} \tag{6.54}$$

The probability for $x \in \{x| - M_L^+ \leq x \leq M_L^+, x \in \mathcal{R}\}$ and $\bar{x} \in T(L, K+1)$, denoted as $P_L\Big(x\Big|\bar{x} \in T(L,K+1)\Big)$, is $\frac{2N^+(L,K+1)-1}{2M^+_{L\infty}}$. For $L = 2K+1$ and $\bar{x} \in T(L,K+1)$, we have $\bar{x} = \sum_{i=0}^{K} y(i) 2^{q(i)}$ where $q(K) = 0$. According to Lemma 6.3, rounding x for $\bar{x} \in T(L,K+1)$ to an element in $S(L,K)$ nearest to it leads to an error distributed in $\pm\,[2^{-1}, 2^{0}]$. Therefore, the error PDF for $e \in \pm\,[2^{-1}, 2^{0}]$ is:

$$\begin{aligned} & p_{L,K}\left(e\Big|e \in \pm\left[2^{-1}, 2^{0}\right]\right) \\ = \; & P_L\Big(x\Big|\bar{x} \in T(L,K+1)\Big) \times p_{L,K}\left(e, x|e \in \pm\left[2^{-1}, 2^{0}\right], \bar{x} \in T(L,K+1)\right) \\ = \; & \frac{2N^+(L,K+1)-1}{2M^+_{L\infty}}. \end{aligned} \tag{6.55}$$

Finally, the probabilities for $x \in \{x|M_L^+ \leq x \leq M_{L\infty}^+, x \in \mathcal{R}\}$ and $x \in \{x| - M_{L\infty}^+ \leq x \leq -M_L^+, x \in \mathcal{R}\}$ are both $\frac{M^+_{L\infty} - M^+_L}{2M^+_{L\infty}}$; rounding x to an element in $S(L,K)$

leads to an error which is uniformly distributed in $\pm\left[2^0, \sum_{k=0}^{\infty} 2^{-2k}\right] = \pm\left[1, \frac{4}{3}\right]$ with PDF equal to 3. Therefore,

$$p_{L,K}\left(e\middle|e \in \pm\left[1,\frac{4}{3}\right]\right) = \begin{cases} \frac{M^+_{L\infty} - M^+_L}{2M^+_{L\infty}} \times 3, & \text{for } e \in \left[1,\frac{4}{3}\right] \\ \frac{M^+_{L\infty} - M^+_L}{2M^+_{L\infty}} \times 3, & \text{for } e \in \left[-\frac{4}{3},-1\right] \end{cases} = \frac{1}{2M^+_{L\infty}}. \quad (6.56)$$

Combining (6.54), (6.55) and (6.56), (6.4) is proved. ■

Lemma 6.5 For $L > 2K+1$, the error PDF for rounding a number $x \in \{x| - M^+_{L\infty} \le x \le M^+_{L\infty}, x \in \mathcal{R}\}$ to an element in $S(L,K)$, denoted as $p_{L,K}(e)$, is as follows:

$$p_{L,K}(e) = \begin{cases} \frac{2M^+(L,K)+1}{2M^+_{L\infty}}, & \text{for } e \in [-2^{-1}, 2^{-1}], \\ \frac{4M^+(L-1,K) - 2M^+(L,K)+1}{2M^+_{L\infty}}, & \text{for } e \in \pm[2^{-1}, 2^0], \\ \frac{p_{L-1,K}(e|e \in \pm[2^{k-1}, 2^k])}{2}, & \text{for } e \in \pm[2^k, 2^{k+1}], \\ & k = 0, \cdots, L-2K-2, \\ \frac{p_{L-1,K}\left(e\middle|e \in \pm\left[2^{L-2K-2}, \frac{2^{L-2K}}{3}\right]\right)}{2}, & \text{for } e \in \pm\left[2^{L-2K-1}, \frac{2^{L-2K+1}}{3}\right] \\ & \text{when } L-2K-1 \ge 0, \\ 0, & \text{otherwise.} \end{cases} \quad (6.57)$$

□

Proof: Assume that when $L \ge 2K+1$, the error PDF for rounding a number x', $x' \in \{x'| - M^+_{L\infty} \le x' \le M^+_{L\infty}, x' \in \mathcal{R}\}$ to an element in $S(L,K)$ is as follows:

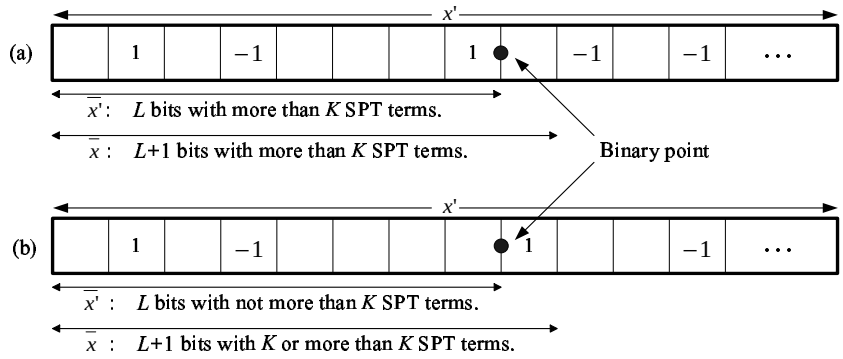

Fig. 6.19: A number x', $x' \in \{x' | -M_L^+ \leq x' \leq M_L^+, x' \in \mathcal{R}\}$ is represented in SPT form. (a) The integer part of x' has more than K SPT terms; (b) the integer part of x' has not more than K SPT terms, where $K = 2$.

$$p_{L,K}(e) = \begin{cases} p_{L,K}\left(e \middle| e \in [-2^{-1}, 2^{-1}]\right), \text{for } e \in \left[-2^{-1}, 2^{-1}\right], \\ p_{L,K}\left(e \middle| e \in \pm[2^{k-1}, 2^{k}]\right), \text{for } e \in \pm\left[2^{k-1}, 2^{k}\right], \\ \qquad\qquad k = 0, \cdots, L-2K-1, \\ p_{L,K}\left(e \middle| e \in \pm\left[2^{L-2K-1}, \dfrac{2^{L-2K+1}}{3}\right]\right), \\ \qquad\qquad \text{for } e \in \pm\left[2^{L-2K-1}, \dfrac{2^{L-2K+1}}{3}\right] \\ \qquad\qquad \text{when } L-2K-1 \geq 0, \\ 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{6.58}$$

In (6.58), the error $e \in [-2^{-1}, 2^{-1}]$ is produced when the integer part of $\bar{x}$ has not more than K SPT terms, whereas the errors in the other ranges in (6.58) are produced when the integer part of x' has more than K SPT terms. From Lemma 6.4, it is shown that this assumption is true for $L = 2K + 1$.

Since $M_{(L+1)\infty}^+ = 2M_{L\infty}^+$, the number $x = 2x'$ is obtained by shifting the binary point for x' by one bit to the right.

i) When the integer part of x' has more than K SPT terms, as shown in

Fig. 6.19(a), it is obvious that the error caused by rounding $x = 2x'$ to an element in $S(L+1,K)$ is distributed in $\pm[2^k, 2^{k+1}]$ for $k = 0, \cdots, L-2K-1$, or an error in $\pm\left[2^{L-2K}, \frac{2^{L-2K+2}}{3}\right]$. Therefore,

$$\begin{aligned} & p_{L+1,K}\left(e\middle|e \in \pm\left[2^k, 2^{k+1}\right]\right) \times 2\left(2^{k+1}-2^k\right) \\ = \; & p_{L,K}\left(e\middle|e \in \pm\left[2^{k-1}, 2^k\right]\right) \times 2\left(2^k-2^{k-1}\right) \\ & \qquad \text{for } k = 0, \cdots, L-2K-1, \end{aligned} \tag{6.59}$$

and

$$\begin{aligned} & p_{L+1,K}\left(e\middle|e \in \pm\left[2^{L-2K}, \frac{2^{L-2K+2}}{3}\right]\right) \times \frac{2^{L-2K}}{3} \\ = \; & p_{L,K}\left(e\middle|e \in \pm\left[2^{L-2K-1}, \frac{2^{L-2K+1}}{3}\right]\right) \times \frac{2^{L-2K-1}}{3}, \\ & \qquad \text{for } L-2K-1 \geq 0, \end{aligned} \tag{6.60}$$

where $p_{L+1,K}\left(e\middle|e \in \pm\left[2^k, 2^{k+1}\right]\right)$ and $p_{L+1,K}\left(e\middle|e \in \pm\left[2^{L-2K}, \frac{2^{L-2K+2}}{3}\right]\right)$ are the error PDF's for e in $\pm[2^k, 2^{k+1}]$ and $\pm\left[2^{L-2K}, \frac{2^{L-2K+2}}{3}\right]$, respectively, when rounding x to an element in $S(L+1,K)$. From (6.59) and (6.60), we have

$$p_{L+1,K}\left(e\middle|e \in \pm\left[2^k, 2^{k+1}\right]\right) = \frac{1}{2} \times p_{L,K}\left(e\middle|e \in \pm\left[2^{k-1}, 2^k\right]\right), \tag{6.61}$$

and

$$\begin{aligned} & p_{L+1,K}\left(e\middle|e \in \pm\left[2^{(L+1)-2K-1}, \frac{2^{(L+1)-2K+1}}{3}\right]\right) \\ = \; & \frac{1}{2} \times p_{L,K}\left(e\middle|e \in \pm\left[2^{L-2K-1}, \frac{2^{L-2K+1}}{3}\right]\right). \end{aligned} \tag{6.62}$$

ii) When the integer part of x' has not more than K SPT terms, rounding x' to an element in $S(L,K)$ leads to an error uniformly distributed in $[-2^{-1}, 2^{-1}]$ and occurring with probability

$$P_L\left(x'\middle|\bar{x'} \in S(L,K)\right) = \frac{2M^+(L,K)+1}{2M^+_{L\infty}}. \tag{6.63}$$

Therefore, rounding $2x'$ to an element in $S(L+1,K)$ leads to an error in $[-2^0, 2^0]$.

For all $\bar{x'} \in S(L,K)$, some of the $2\bar{x'}$ belong to $S(L+1,K)$, whereas the others belong to $S(L+1,K+1)$. For example, the number $2\bar{x}$, where $\bar{x}$ is shown in

Fig. 6.19(b), belongs to $S(L+1,K+1)$. For $L+1>2K+1$,

$$M^{+}_{(L+1)\infty}-M^{+}(L+1,K)\geq\sum_{k=K}^{\infty}2^{L-2k}>1. \tag{6.64}$$

Therefore, the probability of $\bar{x}=2\bar{x}'\in S(L+1,K)$, i.e, $\bar{x}$ can be represented exactly by an element in $S(L+1,K)$, is $\frac{2M^{+}(L+1,K)+1}{2M^{+}_{(L+1)\infty}}$ and is denoted as $P_{L+1}\Big(x\Big|\bar{x}\in S(L+1,K)\Big)$, so that these x will lead to errors in $[-2^{-1},2^{-1}]$. The error PDF for $e\in[-2^{-1},2^{-1}]$ is thus given by

$$\begin{aligned}&p_{L+1,K}(e|-2^{-1}\leq e\leq 2^{-1})\\ =\;& P_{L+1}\Big(x\Big|\bar{x}\in S(L+1,K)\Big)\times p_{L+1,K}\left(e,x\Big|-2^{-1}\leq e\leq 2^{-1},\bar{x}\in S(L+1,K)\right),\end{aligned} \tag{6.65}$$

where $p_{L+1,K}\left(e,x\Big|-2^{-1}\leq e\leq 2^{-1},\bar{x}\in S(L+1,K)\right)$ is equal to 1 according to Lemma 6.1. Therefore, we have

$$\begin{aligned}p_{L+1,K}(e|-2^{-1}\leq e\leq 2^{-1}) &= \frac{2M^{+}(L+1,K)+1}{2M^{+}_{(L+1)\infty}}\times 1\\ &= \frac{2M^{+}(L+1,K)+1}{2M^{+}_{(L+1)\infty}}.\end{aligned} \tag{6.66}$$

Rounding the remaining x, where $\bar{x}\in S(L+1,K+1)$, will lead to errors distributed in $\pm[2^{-1},2^{0}]$. We have

$$\begin{aligned}&p_{L+1,K}(e|e\in\pm[2^{-1},2^{0}])\\ =\;&\left[P_L\left(x'\Big|\bar{x}'\in S(L,K)\right)-P_{L+1}\left(x\Big|\bar{x}\in S(L+1,K)\right)\right]\\ \times\;& p_{L+1,K}\left(e,x\Big|x\in\pm[2^{-1},2^{0}],\bar{x}\in S(L+1,K+1)\right),\end{aligned} \tag{6.67}$$

where $p_{L+1,K}\left(e,x\Big|x\in\pm[2^{-1},2^{0}],\bar{x}\in S(L+1,K+1)\right)$ is equal to 1 according to Lemma 6.5. Therefore, we have

$$\begin{aligned}p_{L+1,K}(e|e\in\pm[2^{-1},2^{0}]) &= \left(\frac{2M^{+}(L,K)+1}{2M^{+}_{L\infty}}-\frac{2M^{+}(L+1,K)+1}{2M^{+}_{(L+1)\infty}}\right)\times 1\\ &= \frac{4M^{+}(L,K)-2M^{+}(L+1,K)+1}{2M^{+}_{(L+1)\infty}}.\end{aligned} \tag{6.68}$$

Combining (6.66), (6.68), (6.61) and (6.62), it can be seen if the assumption in (6.58) is true for $p_{L,K}(e)$ where $L\geq 2K+1$, (6.57) is true for $p_{L+1,K}(e)$. As the

assumption in (6.58) is true for $p_{L,K}(e)$ where $L = 2K + 1$ as stated in Lemma 6.4 and it is also true for the deduced $p_{L+1,K}(e)$, (6.5) is true for $p_{L,K}(e)$ for all $L \geq 2K + 1$. ■

Chapter 7

Symmetrical Polyphase Implementation

POLYPHASE STRUCTURES are widely adopted in multirate system such as interpolators, decimators, and filter banks, to effectively reduce the multiplication rate and data storage [18, 79]. Applications are also found in the implementation of frequency-response masking (FRM) techniques [56] to reduce the memory accesses. While the number of multipliers can be reduced by a factor of two in direct form implementation by exploiting the coefficient symmetry, it is, in general, not the case in polyphase implementation because the impulse response of each polyphase component is no longer symmetrical.

Over the past decades, there has been much effort spent in obtaining polyphase components with symmetrical impulse responses. A generalized polyphase decomposition approach [62, 80] was reported to realize an FIR filter as a parallel connection of several branches. Each branch is of the form of a narrow-band FRM (also known as IFIR) filter and optimized with coefficient symmetry imposed. The generalized polyphase structure is effective only when the filter length is even and the number of branches is an integer power of two.

A multiple branch FIR filter structure was presented in [2, 69]. In the multiple

branch FIR filter, an additional common stage was extracted from the generalized polyphase decomposition structure to reduce the delay elements required in the implementation. By applying the specially developed optimization technique presented in [2], the delay elements as well as the multipliers can be minimized.

A computationally efficient polyphase structure was also reported in [60]. It made use of the fact that the impulse response of non-symmetrical polyphase components exist in mirror image pairs, and a polyphase structure for decimator was presented. In [60], the two signals, which will be multiplied by the same coefficient in the two filters with mirror image impulse responses, were summed first. A transposed structure for interpolator can also be developed in a similar way, where after the input signal is multiplied by the same coefficient in the two polyphase components, they are summed through separate delay lines to their respective polyphase components. Compared with the traditional polyphase implementation, the multipliers used and the multiplication rate are reduced by a factor of two. However, the method reported in [60] requires a large amount of delay elements and memory accesses.

In this chapter, a technique is introduced to exploit the coefficient symmetry when a linear phase FIR filter is implemented in its polyphase components. In this technique, each pair of the time reversed polyphase components are synthesized from a pair of symmetrical and anti-symmetrical impulse response filters. Compared with the conventional polyphase implementation, there is a factor of two saving in the number of multipliers. Under certain conditions, the number of adders is slightly (and insignificantly) increased, whereas, under other conditions, the number of adders is reduced. The number of delay elements and memory accesses remain approximately the same as those of the conventional polyphase structure.

The remaining of this chapter is organized as follows. In section 7.1, FIR filter's polyphase expression is reviewed. The proposed new technique for restoring the

coefficient symmetry is presented in section 7.2. Comparison between the complexities of the filters implemented using the proposed new technique and previous techniques is presented in Section 7.3.

7.1 Polyphase Expression

The even and odd order filters are considered separately. First, an even order linear phase FIR filter is considered. Let the z-transform transfer function of a $2N$-th order FIR filter, $H(z)$, be given by

$$H(z) = \sum_{n=0}^{2N} h(n) z^{-n}, \tag{7.1}$$

where

$$h(n) = h(2N - n), \quad 0 \leq n \leq 2N. \tag{7.2}$$

$H(z)$ may be expressed in its R polyphase components as

$$H(z) = \sum_{k=0}^{2\lceil \frac{N}{R} \rceil} h(kR) z^{-kR} + \sum_{r=1}^{R-1} z^{-r} \sum_{k=0}^{2\lceil \frac{N}{R} \rceil - 1} h(kR + r) z^{-kR}, \tag{7.3}$$

where, $\lceil x \rceil$ is the smallest integer larger than or equal to x. If N is divisible by R, $\frac{N}{R}$ is an integer and $\left\lceil \frac{N}{R} \right\rceil = \frac{N}{R}$. If N is not divisible by R, N may be arbitrarily increased to an integer multiple of R by padding the filter with zero valued coefficients.

Let

$$h_r(k) = h(kR + r), r = 0, 1, ..., R - 1 \tag{7.4}$$

be the k-th impulse response of the r-th polyphase component. Let

$$H_0(z) = \sum_{k=0}^{2\lceil \frac{N}{R} \rceil} h_0(k) z^{-k} \tag{7.5}$$

and

$$H_r(z) = \sum_{k=0}^{2\lceil \frac{N}{R} \rceil - 1} h_r(k) z^{-k}, \quad r = 1, 2, \cdots, R - 1. \tag{7.6}$$

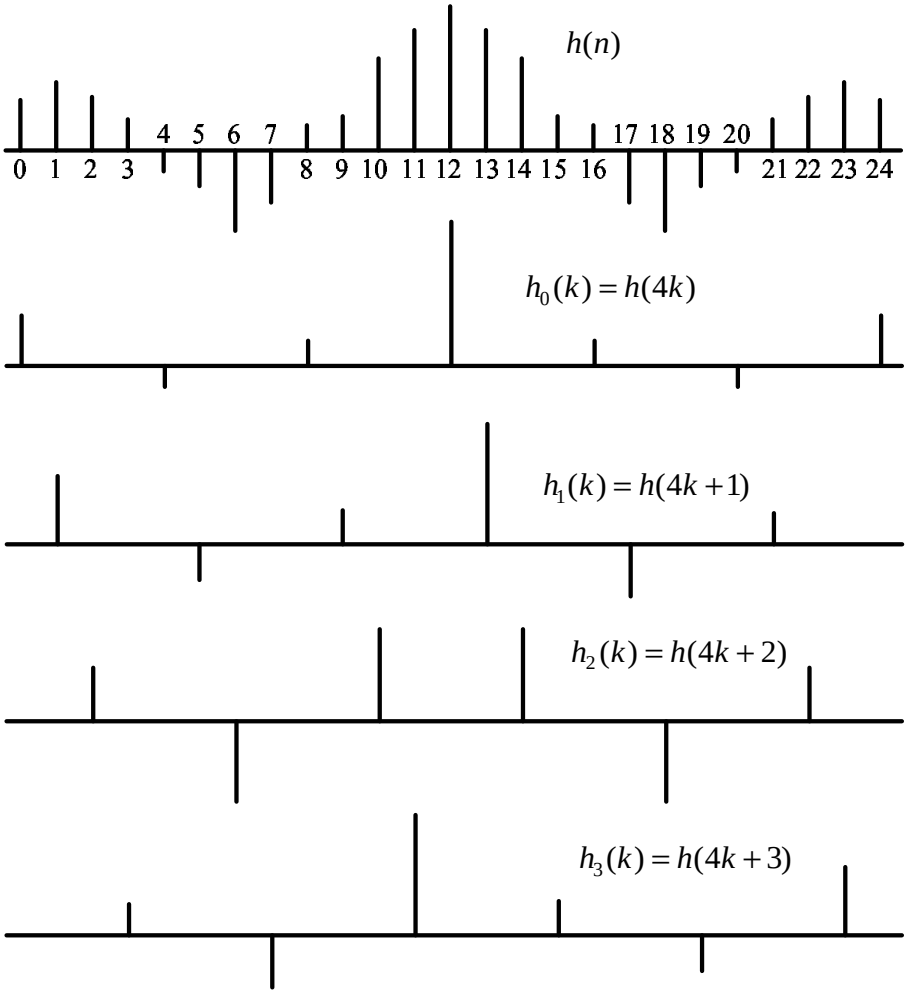

Fig. 7.1: A $2N$th-order filter and its R polyphase components, where $N = 12$ and $R = 4$.

From (7.2) and (7.4), it can be seen that when R is even, $H_0(z)$ is odd symmetrical, $H_{R/2}(z)$ is even symmetrical and the other polyphase filters are not symmetrical. When R is odd, only $H_0(z)$ is odd symmetrical. An example of the impulse responses of a $2N$-th order linear phase FIR filter and its R polyphase filters for $N = 12, R = 4$ is shown in Fig. 7.1.

7.2 Polyphase Implementation Exploiting Coefficient Symmetry

The case where R is even is considered in this section. The case where R is odd can be deduced in a similar way.

Although the polyphase filters, $H_r(z)$ for all r except for $r = 0$ and $r = \frac{R}{2}$, are not symmetrical, it is observed from (7.2) and (7.4) that $H_r(z)$ and $H_{R-r}(z)$ for $r = 1, ..., \frac{R}{2} - 1$ are mirror image filters, i.e. $h_r(k) = h_{R-r}\left(2\left\lceil\frac{N}{R}\right\rceil - 1 - k\right)$. Furthermore, we have

$$\begin{aligned} & z^{-r}H_r(z^R) + z^{-(R-r)}H_{R-r}(z^R) \\ = \; & z^{-r}\sum_{k=0}^{2\left\lceil\frac{N}{R}\right\rceil-1} h_r(k)z^{-kR} + z^{-(R-r)}\sum_{k=0}^{2\left\lceil\frac{N}{R}\right\rceil-1} h_{R-r}(k)z^{-kR}. \end{aligned} \tag{7.7}$$

Define

$$\begin{aligned} h'_r(k) &= \frac{1}{2}[h_r(k) + h_{R-r}(k)], \\ h'_{R-r}(k) &= \frac{1}{2}[h_r(k) - h_{R-r}(k)], \end{aligned} \tag{7.8}$$

for $r = 1, 2, ..., \frac{R}{2} - 1$. Thus,

$$\begin{aligned} h_r(k) &= h'_r(k) + h'_{R-r}(k), \\ h_{R-r}(k) &= h'_r(k) - h'_{R-r}(k), \end{aligned} \tag{7.9}$$

for $r = 1, 2, ..., \frac{R}{2} - 1$. Further define

$$H'_r(z) = \sum_{k=0}^{2\left\lceil\frac{N}{R}\right\rceil-1} h'_r(k)z^{-k}, \quad r = 1, 2, \cdots, R-1. \tag{7.10}$$

It is obvious that $H'_r(z)$ and $H'_{R-r}(z)$ are even symmetrical and even antisymmetrical $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$-tap filters, respectively, i.e.

$$h'_r(k) = h'_r\left(2\left\lceil\frac{N}{R}\right\rceil - 1 - k\right), \tag{7.11}$$

$$h'_{R-r}(k) = -h'_{R-r}\left(2\left\lceil\frac{N}{R}\right\rceil - 1 - k\right), \tag{7.12}$$

for $r = 1, 2, ..., R-1$ and $k = 0, 1, ..., \left\lceil \frac{N}{R} \right\rceil - 1$. Therefore, (7.7) can be expressed as

$$\begin{aligned} & z^{-r} H_r(z^R) + z^{-(R-r)} H_{R-r}(z^R) \\ = \; & z^{-r} \left[\left(1 + z^{-(R-2r)}\right) \sum_{k=0}^{2\left\lceil \frac{N}{R} \right\rceil - 1} h'_r(k) z^{-kR} + \left(1 - z^{-(R-2r)}\right) \sum_{k=0}^{2\left\lceil \frac{N}{R} \right\rceil - 1} h'_{R-r}(k) z^{-kR} \right] \\ = \; & z^{-r} \left[\left(1 + z^{-(R-2r)}\right) H'_r(z^R) + \left(1 - z^{-(R-2r)}\right) H'_{R-r}(z^R) \right]. \end{aligned} \tag{7.13}$$

Therefore, the overall filter $H(z)$ is expressed as

$$\begin{aligned} H(z) &= \sum_{r=0}^{R-1} z^{-r} H_r(z^R) \\ &= \sum_{r=1}^{\frac{R}{2}-1} z^{-r} \left[\left(1 + z^{-(R-2r)}\right) H'_r(z^R) + \left(1 - z^{-(R-2r)}\right) H'_{R-r}(z^R) \right] \\ &\quad + H_0(z^R) + z^{-\frac{R}{2}} H_{R/2}(z^R). \end{aligned} \tag{7.14}$$

Thus, $z^{-r} H_r(z^R) + z^{-(R-r)} H_{R-r}(z^R)$ can be implemented using (7.13) involving $H'_r(z)$ and $H'_{R-r}(z)$ whose coefficients are either symmetrical or anti-symmetrical. Incorporating sampling rate change, the overall filter implementations for decimators and interpolators, referred to as Type I and Type II symmetrical polyphase structures, respectively, are shown in Fig. 7.2. It can be seen from Fig. 7.2 that all the filters are operating at either input or output rate whichever is lower and all the filters have either symmetrical or antisymmetrical coefficients.

Fig. 7.3 shows the implementation for one pair of mirror image pair in the form of $z^{-r} H_r(z^R) + z^{-(R-r)} H_{R-r}(z^R)$ with down sampling rate R for the decimators. From Fig. 7.3, it can be seen that the signal $x(n)$ passes through a tapped delay line marked as "side delay chain". The delayed versions of $x(n)$, $x(n-r)$ and $x(n-R+r)$, are down sampled by R and summed and subtracted before being filtered by the symmetrical and anti-symmetrical filters, $H'_r(z)$ and $H'_{R-r}(z)$. $H'_r(z)$ and $H'_{R-r}(z)$ share the same delay line marked as "main delay chain" in Fig. 7.3. A total of $\left(2\left\lceil \frac{N}{R} \right\rceil\right)$ multiplications and $\left(4\left\lceil \frac{N}{R} \right\rceil + 2\right)$ additions are required to produce the output $y_r(m) + y_{R-r}(m)$, where the overall output of the decimator is produced by $y(m) = \sum_{r=0}^{R-1} y_r(m)$.

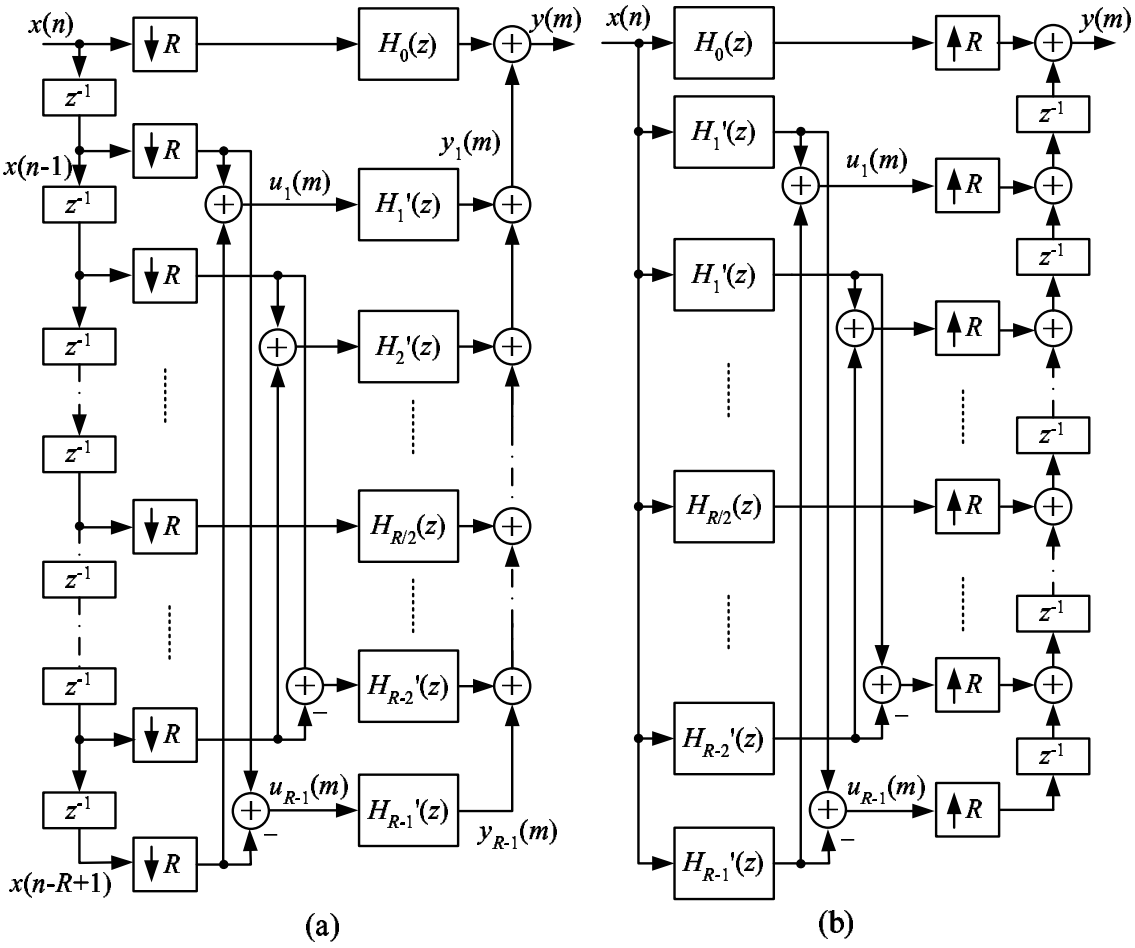

Fig. 7.2: Symmetrical polyphase structures. (a) Type I for decimator; (b) Type II for interpolator.

Unless stated otherwise, in this chapter, the arithmetic complexities are counted at either input or output sampling rate whichever is lower.

In the R polyphase implementation, there are $\left(\frac{R}{2}-1\right)$ such mirror image filter pairs. Therefore, each pair of mirror image filters need $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$ multiplications and $\left(4\left\lceil\frac{N}{R}\right\rceil+2\right)$ additions. Only one set each of the "main delay chain" and the "side delay chain" is needed because all the mirror image pairs share the same set of "main delay chain" and "side delay chain". The length of the "side delay chain" is $(R-1)$, whereas the length of the "main delay chain" is $\left(2\left\lceil\frac{N}{R}\right\rceil-1\right)$. The odd symmetrical and even symmetrical polyphase components, $H_0(z)$ and $H_{R/2}(z)$ also share the same set of "main delay chain" and "side delay chain", as shown in Fig. 7.4. It can be seen from Fig. 7.4 that, when compared with Fig. 7.3, the "main

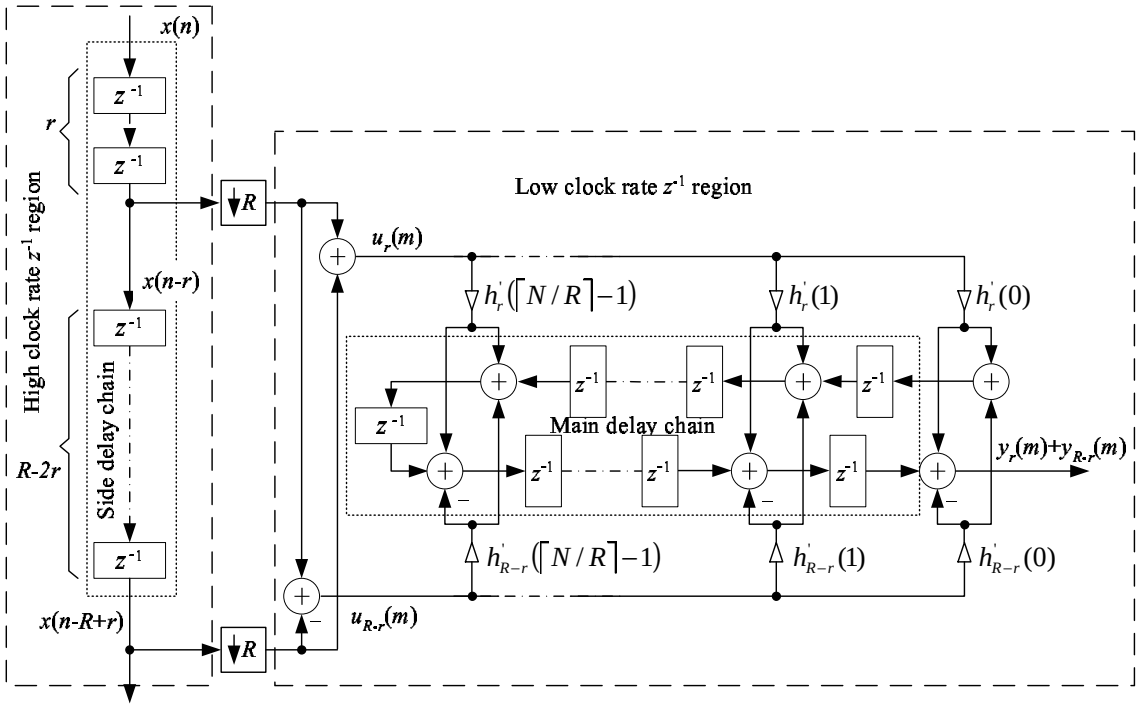

Fig. 7.3: The implementation of $H_r(z)$ and $H_{R-r}(z)$ mirror image filter pair by exploiting the coefficient symmetry of $H'_r(z)$ and $H'_{R-r}(z)$ for Type I symmetrical polyphase structure.

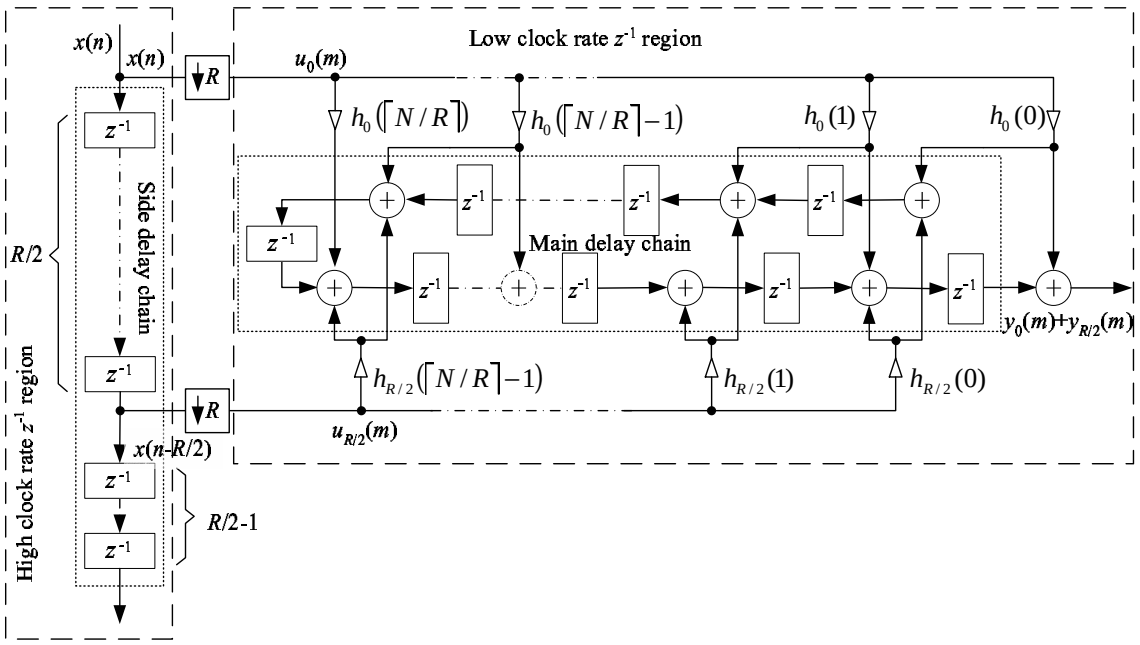

Fig. 7.4: The implementation of $H_0(z)$ and $H_{R/2}(z)$ for Type I symmetrical polyphase structure. The "main delay chain" is the same as that shown in Fig. 7.3 with the exception that an additional delay has been appended to its output end. The "side delay chain" is the same as that shown in Fig. 7.3.

	$H_r(z)$ and $H_{R-r}(z)^a$	$H_0(z)$	$H_{R/2}(z)$	Total
Multiplication	$2\left\lceil\frac{N}{R}\right\rceil\left(\frac{R}{2}-1\right)$	$\left\lceil\frac{N}{R}\right\rceil+1$	$\left\lceil\frac{N}{R}\right\rceil$	$\left\lceil\frac{N}{R}\right\rceil R+1$
Addition	$\left(4\left\lceil\frac{N}{R}\right\rceil+2\right)\left(\frac{R}{2}-1\right)-1$	$2\left\lceil\frac{N}{R}\right\rceil+1$	$2\left\lceil\frac{N}{R}\right\rceil$	$2R\left\lceil\frac{N}{R}\right\rceil+R-2$
Delays	$\left(2\left\lceil\frac{N}{R}\right\rceil-1\right)+R-1$	1^b	0^b	$2\left\lceil\frac{N}{R}\right\rceil+R-1$
MR	$2\left\lceil\frac{N}{R}\right\rceil+R-2$	1^c	0^c	$2\left\lceil\frac{N}{R}\right\rceil+R-1$
MW	$2\left\lceil\frac{N}{R}\right\rceil+R-1$	1^c	0^c	$2\left\lceil\frac{N}{R}\right\rceil+R$

[a] for $r=1,2,...,\frac{R}{2}-1$.

[b] $\left[\left(2\left\lceil\frac{N}{R}\right\rceil-1\right)\right]$ delays shared by mirror image filter pairs are counted once only.

[c] $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$ shared memory read/write operations are counted once only.

Table 7.1: Computation and storage complexities for Type I symmetrical R polyphase structure for a $2N$th-order linear phase FIR filter, where R is an even integer greater than two.

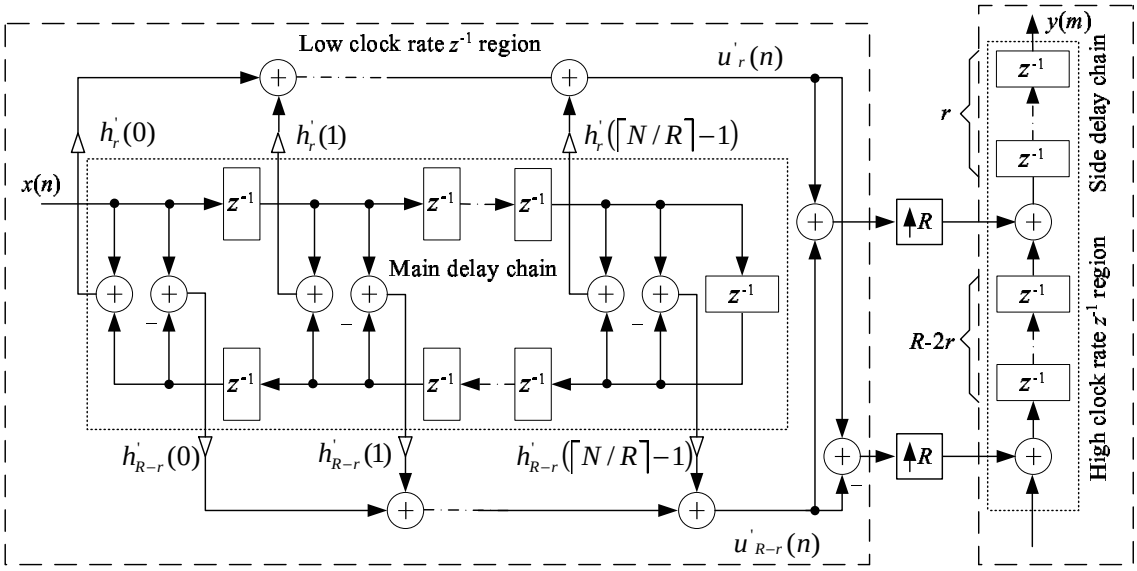

Fig. 7.5: The transposed structure of Fig. 7.3 for implementing the mirror image pairs for Type II symmetrical polyphase structure.

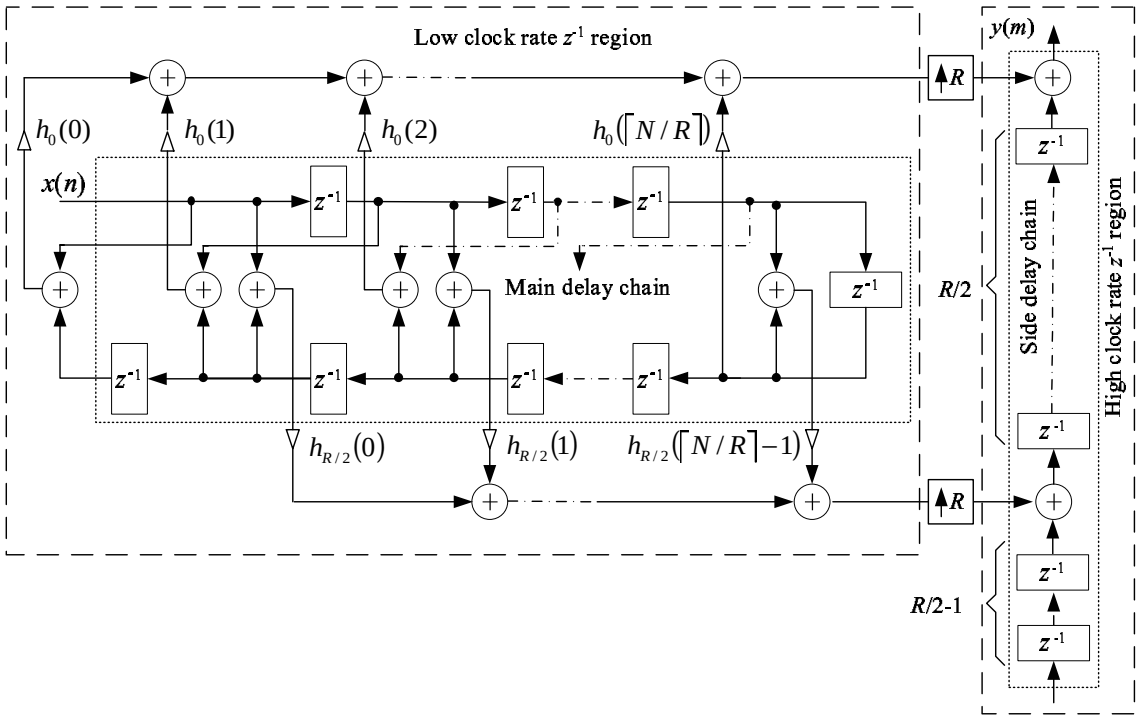

Fig. 7.6: The implementation of $H_0(z)$ and $H_{R/2}(z)$ for Type II symmetrical polyphase structure. The "main delay chain" is the same as that shown in Fig. 7.5 with the exception that an additional delay has been appended to the main delay chain's output end. The "side delay chain" is the same as that shown in Fig. 7.5.

delay chain" has been appended with an additional delay at the output. Therefore, the total length of the "main delay chain" becomes $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$. Furthermore, $H_0(z)$ requires $\left(\left\lceil\frac{N}{R}\right\rceil+1\right)$ multiplications, $\left(2\left\lceil\frac{N}{R}\right\rceil+1\right)$ additions, and $H_{R/2}(z)$ requires $\left\lceil\frac{N}{R}\right\rceil$ multiplications and $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$ additions.

The computation and storage cost for implementing the $\left(\frac{R}{2}-1\right)$ pairs of mirror image filters are listed in the 2nd column of Table 7.1. In certain implementation platform [56], it is necessary to reduce the number of data fetches from memory. Therefore, for the purpose of comparison, the number of memory accesses, designated as memory read (MR) and memory write (MW) are also listed in Table 7.1.

The implementation complexities for $H_0(z)$ and $H_{R/2}(z)$ are listed in the 3rd and 4th columns of Table 7.1. Listed in the 5th column of Table 7.1 are the total complexities to implement the R polyphase structure.

Fig. 7.3 and Fig. 7.4 are eminently suited for implementing the mirror image pairs of decimators. Their transposes as shown in Fig. 7.5 and Fig. 7.6, are eminently suited for implementing the mirror image pairs of interpolators.

The number of multipliers, delays and memory read operations for the Type II symmetrical polyphase structure are the same as those for the Type I, however, the number of adders and memory write cycles differ slightly from those for Type I. As can be seen from Fig.7.5, the signal $x(n)$ passes through the "main delay chain" and those delayed versions of $x(n)$ which will be multiplied by the symmetrical and anti-symmetrical coefficients are summed and subtracted before the multiplication processes. All the mirror image pairs of the polyphase components share the same set of "main delay chain" and "side delay chain". The sum and difference of $x(n-k)$ and $x\left(n-2\left\lceil\frac{N}{R}\right\rceil+k+1\right)$ are fanned out to $h'_r(k)$ and $h'_{R-r}(k)$ for $k=0,1,\cdots,\left\lceil\frac{N}{R}\right\rceil-1$ and $r=1,2,\cdots,\frac{R}{2}-1$. In addition to the $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$ adders enclosed in the "main delay chain" whose sums are fanned out to all the mirror image polyphase pairs, each pair of the polyphase components require an additional

$\left(2\left\lceil\frac{N}{R}\right\rceil - 2\right)$ additions to produce $u'_r(n)$ and $u'_{R-r}(n)$, and 4 additions to sum $u'_r(n)$ and $u'_{R-r}(n)$ to form the final output.

The odd symmetrical polyphase component, $H_0(z)$, is implemented by appending the "main delay chain" by an additional delay, as shown in Fig. 7.6. $H_0(z)$ is not able to use the sum obtained from the adders enclosed in the "main delay chain" in Fig. 7.5, therefore, it needs $\left(2\left\lceil\frac{N}{R}\right\rceil\right)$ additions. The even symmetrical polyphase component, $H_{R/2}(z)$, is able to share both the "main delay chain" and "side delay chain" and use the sum obtained from the adders enclosed in the "main delay chain" in Fig. 7.5. The implementation of $H_{R/2}(z)$ is also shown in Fig. 7.6. $H_{R/2}(z)$ needs $\left\lceil\frac{N}{R}\right\rceil$ additions. The numbers of additions for Type II symmetrical polyphase structure are listed in Table 7.2. Also listed in Table 7.2 are the numbers of memory write cycles.

7.3 Comparison and Discussion

The computation and storage requirements for a $2N$-th order linear phase FIR filter implemented using the: 1) direct form structure exploiting the coefficient symmetry; 2) conventional polyphase structure without exploiting the coefficient symmetry; 3) polyphase structure reported in [60] and 4) new technique proposed in this chapter, are listed in Table 7.3.

It can be seen from Table 7.3 that when compared with the conventional polyphase structure, the proposed structure achieves a 50% reduction in the number of multipliers and the multiplication rate, whereas the delay elements used and memory access remain approximately the same. When compared with the polyphase structure reported in [60], the proposed structure achieves approximately a factor of R reduction in the number of delay elements and a factor of $\frac{R}{2}$ reduction in the number of memory access, whereas the number of multipliers and multiplication rate remain the same.

	$H_r(z)$ and $H_{R-r}(z)^a$	$H_0(z)$	$H_{R/2}(z)$	Total
Addition	$2\left\lceil\frac{N}{R}\right\rceil + \left(2\left\lceil\frac{N}{R}\right\rceil + 2\right)\left(\frac{R}{2} - 1\right)$	$2\left\lceil\frac{N}{R}\right\rceil$	$\left\lceil\frac{N}{R}\right\rceil$	$(R+3)\left\lceil\frac{N}{R}\right\rceil + R - 2$
MW	R	0	0	R

[a] for $r = 1, 2, ..., \frac{R}{2} - 1$.

Table 7.2: Addition rate and memory write cycles for Type II symmetrical R polyphase structure for a $2N$th-order linear phase FIR filter, where R is even.

	Direct form	Conventional R Polyphase	R Polyphase proposed in [60]	R Polyphase proposed
Multiplication	$(N+1)R$	$\left\lceil \frac{2N+1}{R} \right\rceil R$	$\left\lceil \frac{N}{R} \right\rceil R+1$	$\left\lceil \frac{N}{R} \right\rceil R+1$
Add/sub (Type I)	$2NR$	$\left\lceil \frac{2N+1}{R} \right\rceil R-1$	$\left\lceil \frac{N}{R} \right\rceil R+N$	$2R\left\lceil \frac{N}{R} \right\rceil +R-2$
Add/sub (Type II)	$2NR$	$\left\lceil \frac{2N+1}{R} \right\rceil R-1$	$2\left\lceil \frac{N}{R} \right\rceil R$	$(R+3)\left\lceil \frac{N}{R} \right\rceil +R-2$
Delay elements	$2N$	$\left\lceil \frac{2N+1}{R} \right\rceil +R-2$	$2\left\lceil \frac{N}{R} \right\rceil R$	$2\left\lceil \frac{N}{R} \right\rceil +R-1$
MR	$2N$	$\left\lceil \frac{2N+1}{R} \right\rceil +R-2$	$2\left\lceil \frac{N}{R} \right\rceil R$	$2\left\lceil \frac{N}{R} \right\rceil +R-1$
MW (Type I)	R	$\left\lceil \frac{2N+1}{R} \right\rceil +R-1$	R	$2\left\lceil \frac{N}{R} \right\rceil +R$
MW (Type II)	$2N$	R	$2\left\lceil \frac{N}{R} \right\rceil R$	R

Table 7.3: Comparison for operation rate for implementing a $2N$-th order linear phase FIR filter, where R is even.

	$2N$-th order R is even	$2N$-th order R is odd	$(2N-1)$-th order R is even	$(2N-1)$-th order R is odd
Multiplication	$\left\lceil \frac{N}{R} \right\rceil R+1$	$\left\lceil \frac{N}{R} \right\rceil R+1$	$\left\lceil \frac{N}{R} \right\rceil R$	$\left\lceil \frac{N}{R} \right\rceil R$
Add/sub (Type I)	$2R\left\lceil \frac{N}{R} \right\rceil + R - 2$	$2R\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2R\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2R\left\lceil \frac{N}{R} \right\rceil + R - 2$
Add/sub (Type II)	$(R+3)\left\lceil \frac{N}{R} \right\rceil + R - 2$	$(R+3)\left\lceil \frac{N}{R} \right\rceil + R - 1$	$(R+2)\left\lceil \frac{N}{R} \right\rceil + R - 1$	$(R+2)\left\lceil \frac{N}{R} \right\rceil + R - 2$
Delay elements	$2\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2\left\lceil \frac{N}{R} \right\rceil + R - 2$	$2\left\lceil \frac{N}{R} \right\rceil + R - 2$
MR	$2\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2\left\lceil \frac{N}{R} \right\rceil + R - 2$	$2\left\lceil \frac{N}{R} \right\rceil + R - 2$
MW (Type I)	$2\left\lceil \frac{N}{R} \right\rceil + R$	$2\left\lceil \frac{N}{R} \right\rceil + R$	$2\left\lceil \frac{N}{R} \right\rceil + R - 1$	$2\left\lceil \frac{N}{R} \right\rceil + R - 1$
MW (Type II)	R	R	R	R

Table 7.4: Operation rate for implementing a linear phase FIR filter in its R polyphase components by using the proposed new technique.

The addition rate of the conventional polyphase structure and the structure reported in [60] are approximately $2N$, which is a factor of R less than that of the direct form structure. For Type I symmetrical polyphase structure with N divisible by R, the proposed technique also produces the same statistics as that in the conventional polyphase structure. When N is not divisible by R, Type I structure needs an additional $(R-1)$ additions. On the other hand, in Type II structure, the addition rate may be less than $2N$ under certain circumstances. The addition rate is reduced if (7.15) is satisfied.

$$(R+3)\left\lceil \frac{N}{R} \right\rceil + R - 2 \leq 2N. \tag{7.15}$$

Let $L = \frac{N}{R} > 1$. In this case, (7.15) reduces to

$$R \geq \frac{3L-2}{L-1}. \tag{7.16}$$

It can be easily verified by substituting positive integer values of L that

$$3 < \frac{3L-2}{L-1} = 3 + \frac{1}{L-1} \leq 4. \tag{7.17}$$

Therefore, when $3 < R < N$, the addition rate is reduced. The maximum reduction is $\left(N - 2\sqrt{3N} + 2\right)$ and occurs at $R = \sqrt{3N}$ if N is divisible by R. When $R = 2$, only $H_0(z)$ and $H_{R/2}(z)$ exist and there are no other mirror image filter pairs. The addition rate listed in Table 7.1 is not applicable; the addition rate remains at $2N$, the same as the case in the conventional polyphase implementation.

The above derivations are developed for even R. It is straightforward to derive similar expressions for odd R. When R is odd, $H_0(z)$ is odd symmetrical and there are $\frac{R-1}{2}$ pairs of mirror image filters, i.e. $H_r(z)$ and $H_{R-r}(z)$ for $r = 1, 2, ..., \frac{R-1}{2}$. The mirror image filter pairs as well as $H_0(z)$ can be implemented in the same way as those for even R. The computational complexities are listed in Table 7.4. It is shown that, besides the reduction in multiplication rate (compared with the conventional polyphase structure) and delay elements used (compared with the structure reported in [60]), when $3 < R < N$, the addition rate is reduced compared

with that in the other methods for Type II case and the maximum reduction occurs when $R = \sqrt{3N}$.

For a $(2N-1)$-th order linear phase FIR filter, its transfer function, $H(z)$, is given by

$$H(z) = \sum_{n=0}^{2N-1} h(n) z^{-n} \tag{7.18}$$

where

$$h(n) = h(2N-1-n), \quad 0 \leq n \leq 2N-1. \tag{7.19}$$

$H(z)$ may be expressed in its R polyphase components as

$$H(z) = \sum_{r=0}^{R-1} z^{-r} \sum_{k=0}^{2\left\lceil \frac{N}{R} \right\rceil - 1} h(kR+r) z^{-kR}. \tag{7.20}$$

The k-th impulse response of the r-th polyphase component, $h_r(k)$ for $r = 0, 1, ..., R-1$, is expressed as shown in (7.4). From (7.19) and (7.4), it can be seen that when R is even, the R polyphase components consist of $\frac{R}{2}$ pairs of mirror image filters, i.e. $h_r(k) = h_{R-r-1}\left(2\left\lceil \frac{N}{R} \right\rceil - 1 - k\right)$ for $r = 0, 1, ..., \frac{R}{2} - 1$. Therefore, the polyphase structure can be implemented by $\frac{R}{2}$ pairs of mirror image filters in the way described in Section 7.2, and the complexities are listed in Table 7.4. Similarly, if R is odd, $H_{(R-1)/2}(z)$ is even symmetrical and $H_r(z)$ and $H_{R-r-1}(z)$ for $r = 0, 1, ..., \frac{R-3}{2}$, are $\frac{R-1}{2}$ pairs of mirror image filters. The complexities of its implementation by using the proposed technique are listed in Table 7.4. It also shows that, for both the even R and odd R cases, when $2 < R < N$, the number of adders is fewer than the number of adders used in the other implementations for Type II case and the maximum reduction occurs when $R = \sqrt{2N}$.

For completeness, the implementation complexities for $2N$-th order filter with even R are also listed in Table 7.4.

Before concluding this chapter, the proposed symmetrical polyphase structure is compared with the polyphase decomposition in [62]. It is obvious from (7.14) that the proposed symmetrical polyphase structure may also be expressed using

the generalized polyphase decomposition form as

$$H(z) = \sum_{r=0}^{R} F_r(z) H'_r(z) \tag{7.21}$$

where $F_r(z)$ is usually a multiplier free polynomial and $H'_r(z)$ is symmetrical or antisymmetrical filters.

The generalized polyphase decomposition reported in [62] resulted in $F_r(z)$ becoming an R-term polynomial. $H'_r(z)$ is optimized one by one to approximate the original frequency response. The decomposition in [62] is effective only when the overall filter length is even and R is an integer power-of-two.

For the proposed symmetrical polyphase structure, $F_r(z)$ is a two-term polynomial and $H'_r(z)$ is transformed from the original polyphase component $H_r(z)$. The transformation is an identity transformation; there is no approximation involved. The proposed structure is effective for both even and odd length filters and effective for any $R < N$.

7.4 Conclusion

In this chapter, a technique to implement linear phase FIR filters in polyphase structures while restoring the coefficient symmetry property is presented. In the proposed new technique, each non-symmetrical but mirror image polyphase component pair are synthesized as a sum or difference of two symmetrical and antisymmetrical filters. Thus, a linear phase FIR filter can be implemented in its polyphase components using symmetrical and antisymmetrical filters. Two types of the structures are proposed to implement, respectively, decimators and interpolators. There is a 50% saving in the multiplication rate compared with the conventional polyphase structure. The proposed new technique may result in a slight increase (less than R) in the additions rate for the decimator structure and for the interpolator structure under certain circumstances. For the implementation of interpolators, under most

circumstances, the proposed technique results in a reduction in the addition rate; the maximum possible reduction is $\left(N - 2\sqrt{2N} + 2\right)$. The storage elements used and memory accesses remain approximately the same.

Chapter 8

Conclusion

IN THIS THESIS , design techniques for the SPT coefficient lattice filter bank are developed. The SPT coefficient optimum solution may be located very far away from the infinite precision coefficient solution. Local search and GA can only obtain SPT design near the infinite precision coefficient solution with reasonable high probability. For the linear phase FIR filter design, the chance that a very good SPT coefficient solution may be located near the infinite precision coefficient solution can be improved tremendously by simply scaling the passband gain of the filter. However, this is not the case for the lattice filter bank. Therefore, the scaling strategy which shows great potential for linear phase FIR filters is not suitable for the lattice filter bank. This leads to the consequence that the local search approach and GA are not very efficient for the design of SPT coefficient lattice filter bank.

The width-recursive depth-first tree search algorithm proposed in this thesis quantizes the coefficients one at a time and reoptimizes the remaining unquantized coefficients. The tree is developed in the so called width-recursive and depth-first manner. The order of the coefficients selected to be quantized is based on a frequency response deterioration measure, which is the product of the coefficient sensitivity of the frequency response and the grid spacing of the infinite precision coefficients in the SPT space. The maximum value of the coefficient sensitivity of

each coefficient is proved to be inversely proportional to one plus the square of the corresponding coefficient. The tree search algorithm is very suitable for the SPT coefficient lattice filter design and it overcomes the difficulty that the SPT optimum may be located far away from the infinite precision design values in two aspects. First, the coefficient selected to be quantized is fixed at discrete values step by step and further and further away from its continuous value with the increasing tree width. Second and more importantly, after each coefficient is fixed, the remaining unquantized coefficients are reoptimized; the reoptimization process may throw the coefficients far away from the original continuous optimum values.

The SPT numbers are unevenly distributed for a given wordlength $L-Q$, precision 2^Q and the number of SPT terms K. Therefore, the rounding error is also unevenly distributed. A mathematical representation of the SPT rounding error density function is developed; it is a piecewise constant staircase function symmetrical about zero. The error probability density function has larger magnitude for errors closed to zero. Its magnitude decreases with increasing error magnitude. The variance of the error probability density decreases with decreasing Q for given L and K. The variance approaches asymptotically to a constant as Q approaches minus infinity ($-\infty$). Increasing K may significantly reduce the variance, however, it will approach asymptotically to another constant as Q approaches $-\infty$ for given L and K.

The effects of quantizing the coefficient values to SPT values are analyzed using the SPT rounding error probability density function. The analysis showed that when directly rounding the coefficients to SPT values with a given K and a sufficiently small Q, for a given infinite precision stopband attenuation, the stopband attenuation deterioration increases very slowly with increasing filter length, i.e., the stopband attenuation deterioration versus the filter length plot is a flat line with small up slope. Analysis also showed that when directly rounding the coefficient values to SPT values, once one of K and Q is determined, the statistical bound of the other value is also determined, i.e., using K larger than a value or using Q

smaller than a value is not beneficial. It is very useful when direct rounding of the lattice filter coefficients to SPT values is considered.

Based on the statistical analysis of the SPT value and the effects on the frequency response of filter bank, an SPT term allocation scheme is also presented for the design of the SPT coefficient lattice filter bank. The tree search algorithm incorporating the SPT term allocation scheme is able to design SPT coefficient filter banks with different number of SPT terms efficiently.

Finally, a polyphase implementation of multirate system is presented. In the proposed implementation, the filter coefficients' symmetry which has been destroyed by the conventional polyphase implementation is restored.

Bibliography

[1] A. Antoniou, Algorithm 7.4 of notes "Optimization: theory and practice."

[2] P. Arian, T. Saramäki and S.K. Mitra, "A systematic technique for optimization multiple branch FIR filters for smapling rate conversion," in *Proc. of IEEE International Conference on Circuit, Syst.* vol. IV, pp.1-4, 2002.

[3] T.P. Barnwell "Subband coder design incorporating recursive quadrature mirror filters and optimum ADPCM coders, " *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-30, pp. 751-765, Oct. 1982.

[4] M.G. Bellanger, G. Bonnerot and M. Coudreuse, "Digital filtering by polyphase network: Application to sample rate alteration and filter banks," *IEEE Trans. Acoust. Speech Signal Process.,* vol. ASSP-24, pp. 109-114, April 1976.

[5] N. Benvenuto, M. Marchesi and A. Uncini, "Applications of simulated annealing for the design of special digital filters," *IEEE Trans. Signal Processing,* vol. 40, pp. 323-332, Feb. 1992.

[6] T. Blu, "A new design algorithm for two-band orthonormal rational filter banks and orthonormal ratioinal wavelets," *IEEE Trans. Signal Process.,* 46, pp. 1494-1504, June 1998

[7] R. Bregović and T. Saramäki, "Two-channel FIR filter banks - A tutorial review and new results," in *Proc. Second Int. Workshop on Transforms and Filter Banks,* Brandenburg, Germany, TICSP #4, 507-558, March 1999.

[8] ——, "An iterative method for designing orthogonal two-channel FIR filter banks," in *Proc. IEEE Int. Conf. Acout., Speech, Signal Process.,* Istanbul, Turkey, vol. I, pp. 484-487, June. 2000.

[9] ——, "A general-purpose optimization technique for designing two-channel FIR filter banks," in *Proc. Tenth Europea Signal Process. Conf.,* Tampere, Finland, vol. I, pp. 369-372, Sept. 2000.

[10] D. Ait-Boudaoud and R. Cemes, "Modified sensitivity criterion for the design of powers-of-two FIR filters," *Electron. Lett.,* vol. 29, pp. 1467-1469, Aug. 1993.

[11] D.S.K. Chan and L.R. Rabiner, "Analysis of quantization errors in the direct form for finite impulse response digital filters," *IEEE Trans. Audio, Electroacoustics,* vol. AU-21, pp.354-66, Aug. 1973.

[12] Y. Chen, S.M. Kand and T.G. Marshall, "The optimal design of CCD transversal filter using mixed integer programming technique," in *Proc. IEEE Int. Symp. Circuits and Syst.,* 1978

[13] C.-K. Chen and J.-H. Lee, "Design of quadrature mirror filters with linear phase in the frequency domain," *IEEE Trans. Circuits Syst. II,* vol. 39, pp. 593-605, Sept. 1992.

[14] ——, "Design of linear-phase quadrature mirror filters with powers-of-two coefficients," *IEEE Trans. Circuits Syst. II,* vol. 41, pp. 445 -456, Jul. 1994.

[15] C.-L. Chen, A.N. Willson Jr., "A Trellis search algorithm for the design of FIR fiters with signed-powers-of-two coefficients," *IEEE Trans. circuits syst. II,* vol. 46, pp. 29-39, Jan. 1999.

[16] R.E. Crochiere, "A new statistical approach to the coefficient wordlength problem for digital filters," *IEEE Trans. Circuits Syst.,* vol CAS-22, pp.190-196, Mar. 1975.

[17] ——, "A novel approach for implementing pitch prediction in sub-band coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp. 526-529, Washington, DC, Apr. 1979.

[18] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing,* Prentice-Hall, 1983.

[19] A. Croisier, D. Esteban, and C. Galand, "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques," in *Proc. IEEE Int. Conf. Inform. Sci. Syst.,* Patras, Greece. 1976.

[20] G.J. Dolecek, *Multirate System and Filter Banks,* Hershey PA: Idea Group Publishing, 2002, ch. 2.

[21] D. Esteban and C. Galnad, "Application of quandrature mirror filters to split band voice coding scheme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1977, pp. 191-195.

[22] A. Fettweis, J.A. Nossek and K. Meerkotter, "Reconstruction of signals after filtering and sampling rate reduction, " *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 893-902, Aug. 1985.

[23] R. Fletcher, *Practical methods of optimization,* John Wiley & Son, 1987.

[24] N.J. Fliege, *Multirate Digital Signal Processing,* Chicester, NY: John Wile and Sons.

[25] A. Fuller and B. Nowrouzian, Finite-pricision characterization of a class of bode-type variable-amplitude digital equalizers. in *Proc. of the 1997 Midwest Symposium on Circuits and Systems,* pp.429-32, Aug. 1997.

[26] A. Fuller, B. Nowrouzian and F. Ashrafzadeh, "Optimization of FIR digital filters over the canonical Signed-digit coefficient space using genetic algorithms," in *Proc. of the 1998 Midwest Symposium on Circuits and Systems,* pp.456-59, Aug. 1998.

[27] R.S. Garfinkel and G.L.Nemhauser, *Integer Programming,* New York: Wiley, 1972.

[28] H. Gharavi and A. J. Tabatabaik, "Application of quadrature mirror fitlering to the coding of monochrome and color images," in *Proc. IEEE Int. conf. Acoust., Speech, Signal process.*, 1987, pp. 2384-2387.

[29] C.K. Goh, "Weighted least squares techniques for the design of multirate filter banks," Ph.D. dissertation, National Univ. of Singapore, Singapore, 2001.

[30] C.K. Goh and Y.C. Lim, "A efficient algorithm to design weighted minimax perfect reconstruction quadraute mirror filters," *IEEE Trans. Signal Process.,* vol. 47, pp. 3303-3314, Dec. 1999.

[31] C.K. Goh, Y.C. Lim and C.S. Ng, "Improved weighted least squares algorithm for the design of quadraute mirror filters," *IEEE Trans. Signal Process.,* vol. 47, pp. 1866-1877, July 1999.

[32] D.E. Goldberg, Genetic algorithms in search and optimisation, Addison-Wesley, 1989.

[33] O. Gustafsson, H. Johansson and L. Wanhammar, "An MILP approach for the design of linear-phase FIR filters with minimum number of signed-power-of-two terms," in *Proc. European Conf. Circuit Theory Design,* vol. 2, pp. 217-220, Espoo, Finland, Aug. 2001.

[34] B.-R. Horng, H. Samueli, and A.N. Willson, Jr., "The design of two-channel lattice structure perfect-reconstruction filter banks using power-of-two coefficients," *IEEE Trans. Circuits Syst. I,* vol. 40, pp.497-9, July 1993.

[35] B.-R. Horng and A.N. Willson, "Lagrange multiplier approaches to the design of two-channel perfect-reconstruction linear-phase FIR filter banks," *IEEE Trans. Signal Process.,* vol. 40, pp. 364-374, Feb. 1992.

[36] V.K. Jain and R.E. Crochiere, "A novel approach to the design of analysis/synthesis filter banks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Boston, MA, Apr. 1983.

[37] ——, "Quadrature mirror filter design in the time domain, " *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-32, pp. 353-361, Apr. 1984.

[38] J.D, Johnston, "A filter family designed for use in quadrature mirror filter banks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Denver, CO, Apr. 1980.

[39] J.B. Knowles and R. Edwards, "Effects of a finite-word-length computer in a sampled data feedback system," in *Proc. Inst. Elec. Eng.,* vol. 112, no. 6, pp. 1197-1207, June, 1965.

[40] J.B. Knowles and E.M. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuit Theory,* vol. CT-15, pp.31-41, March 1968.

[41] D.M. Kodek, "Design of optimal finite wordlength FIR digital filters using integer programming techniques," *IEEE Trans. Acoust., Speech, Signal Process.,* vol ASSP-28, pp. 304-308, June 1980.

[42] D.M. Kodek and K. Steiglitz, "Comparison of optimal and local search methods for designing finite wordlength FIR digital filters," *IEEE Trans. Circuits Syst.,* vol. CAS-28, pp.28-32, Jan. 1981.

[43] C.W. Kok and T.Q. Nguyen, "Discrete coefficients filter banks and applications in image coding," in *IEEE Proc. Int. Conf. Acoust., Speech, Signal Process.,* vol. 3, pp. 1550 -1553, May 1996.

[44] C.L. Lawson, "Contribution to the theory of linear least maximum approximations," Ph.D. dissertation, Univ. California, Los Angeles, 1961.

[45] J.-H. Lee, C.-K. Chen and Y.C. Lim, "Design of discrete coefficient FIR digital filters with arbitary amplitude and phase responses," *IEEE Trans. Circuits Syst.,* vol. 40, pp. 444-448, July, 1993.

[46] A. Lee, M. Ahmadi, G.A. Jullien, W.C. Miller and R.S. Lashkari, "Digital filter design using genetic algorithm, in *Proc. IEEE Symp. Advances in Digital Filtering and Signal Processing,"* 1998.

[47] D.N. Li, Y.C. Lim, Y. Lian and J.J. Song, "A polynomial-time algorithm for design digital filters with power-of-two coefficients", *IEEE Trans. Signal Processing,* vol. 50, pp. 1935-1941, August 2002.

[48] Y.C. Lim, "Design of discrete-coefficient-value linear phase FIR filters with optimum normalized peak ripple magnitude," *IEEE Trans. Circuits Syst.,* vol. 37, pp. 1480-1486, Dec. 1990

[49] Y.C. Lim and A.G. Constantinides, "Linear phase FIR digital filter without multipliers," in *Proc. IEEE Int. Symp. Circuits and Syst.,* pp. 185-189, 1979

[50] Y.C. Lim, J.B. Evans and B. Liu, "Decomposition of binary integers into signed power-of-two terms," *IEEE Trans. Circuits Syst.,* vol. 38, pp. 667-672, June 1991.

[51] Y.C. Lim, J.H. Lee, C.K. Chen and R.H. Yang, "A weighted least squares algorithm for quasi-equiripple FIR and IIR digital filter design," *IEEE Trans. Signal Processing,* vol.40, pp.551-8, Mar. 1992.

[52] Y.C. Lim and S.R. Parker, "FIR filter design over a discrete power-of-two coefficient space", *IEEE Trans. Acoust., Speech, and Signal Processing,* vol. ASSP-31, pp.583-591, June 1983.

[53] ——, "Discrete coefficient FIR digital filter design based upon an LMS criteria", *IEEE Trans. Circuits and Systems,* vol. CAS-30, pp.723-739, Oct. 1983.

[54] Y.C. Lim, R.H. Yang and S.-N. Koh, "The design of weighted minimax quadrature mirror filters," *IEEE Trans.Signal Process.,* vol. 41, pp. 1780-1789, May 1993.

[55] Y.C. Lim, R. Yang, D.N. Li and J.J. Song, "Signed power-of-two term allocation scheme for the design of digital filters," *IEEE Trans. Circuits Syst. II,* vol. 46, pp.577-84, May 1999.

[56] Y.C. Lim, Y.J. Yu, H.Q. Zheng and S.W. Foo, "FPGA implementation of digital filters synthesized using frequency-response masking technique," in *Proc. of IEEE International Conference on Circuit, Syst.* vol.II, pp.173-176, 2001.

[57] P. Löwenborg, E. Elias, H. Johansson and L. Wanhammar, "Two-channel IIR/FIR filter banks with very low-complexity analysis or synthesis filters: finite wordlength effects," in *Proc. IEEE Midwest Symp. Circuits and Syst.,* vol. 1, pp. 126-129, 2001.

[58] W.-S. Lu, H. Xu and A. Antoniou, "A new method for the design of FIR quadrature mirror-image filter banks," *IEEE Trans. Circuits Syst. II,* vol. 45, pp 922-926, July 1998.

[59] F. Mintzer, "Filters for distortion-free two-band multirate filter banks," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 626-630, June 1985.

[60] S.K. Mitra, *Digital Signal Processing, A Computer-Based Approach,* pp.686-689. McGraw-Hill, 1998.

[61] S.K. Mitra and J.F. Kaiser, *Handbook for Digital Signal Processing,* John Wiley & Sons, 1993, ch. 4,8.

[62] S.K. Mitra, A. Mahalanobis and T. Saramäki, "Generalized structural subband decomposition of FIR filters and its application in efficient FIR filter design and implementation," *IEEE Trans. Circuits Syst. II,* vol. 40, pp.363-74, June 1993.

[63] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1989, ch. 7.

[64] L.R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1975, ch. 3.

[65] G.W. Reitwiesner, "Binary arithmetic," *Advances in Comoputers,* vol. 1, pp.231-308, 1960.

[66] G.V. Reklaitis, A. Ravindran and K.M. Ragsdell, *Engineering Optimization: Methods and Applications,* A Wiley-Interscience Publication, 1983.

[67] H. Samueli, "An improved search algorithm for the design of multiplierless FIR filters with powers-of-two coefficients," *IEEE Trans. Circuits Syst.,* vol. 36, pp. 1044-1047, July 1989.

[68] T. Saramäki, "A systematic technique for designing highly selective multiplier-free FIR filters," in *Proc. IEEE Int. Conf. Circuits Syst.,* Singapore, pp. 484-487, 1991.

[69] T. Saramäki and S.K. Mitra, "Multiple branch FIR filters for sampling rate conversion,"in *Proc. of IEEE International Conference on Circuit, Syst.* pp.1007-10, 1992.

[70] T. Saramäki and J. Yli-Kaakien, "Design of digital filters and filter banks by optimization: applications,"in *Proc. X European Signal Processing Conference,* Tampere, Finland, Sept. 2000.

[71] M. Schusdziarra, N.J. Fliege and U. Zölzer; "Finite wordlength effects in quadrature mirror filter banks," in *Proc. IEEE Int. Symp. Circuits Syst.,* vol. 3, pp. 1340 -1343, May, 1992.

[72] H. Shaffeu, M.M. Jones, H.D. Griffiths and J.T. Taylor, "Improved design procedure for multiplierless FIR digital filters," *Electron. Lett.,* vol. 27, pp. 1142-1144, June 20, 1991.

[73] M.J.T. Smith and T.P. Barnwell, "Exact reconstruction techniques for tree-structured subband coders, " *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-34, pp. 434-41, June 1986.

[74] M.J.T. Smith and S.L. Eddins, "Analysis-synthesis techniques for subband image coding, " *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 38, pp. 1446-56, Aug. 1990.

[75] S. Sriranganathan, D.R. Bull and D.W. Redmill, "The design of low complexity two-channel lattice-structure perfect-reconstruction filter banks using genetic algorithms," in *Proc. ISCAS.* IEEE, June 1997.

[76] J.I. Suarez and C.S. Lindquist, "Coefficient quantization error recovery," in *Proc. IEEE Int. Symp. Circuits Syst.,*vol. 3, pp. 435 -438, Jul. 1999.

[77] K.S. Tang, K.F. Man, S. Kwong and Q. He, Genetic algorithms and their applications, *IEEE Signal Processing Magazine,* vol. 136, pp.22-37, Nov. 1996.

[78] D. Taubman and A.Zakhor, "Multi-rate 3-d subband coding of video," *IEEE Trans. Image Processing*, vol. 3, pp. 572-588, Sept. 1994.

[79] P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Prentice-Hall, 1993.

[80] P.P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial," *Proc. IEEE*, vol. 78, Jan. pp. 56-93, 1990.

[81] P.P. Vaidyanathan and P.Q. Hoang, "Lattice Structures for Optimal Design and Robust Implementation of Two-Channel Perfect-Reconstruction QMF Banks, " *IEEE Trans., Acoust., Speech, Signal Processing,* vol. 36, pp.81-94, Jan. 1988.

[82] H. Xu, W.-S. Lu and A. Antoniou, "An improved method for the design of FIR quadrature mirror-image filter banks," *IEEE Trans. Signal Porcess.,* vol. 46, pp. 1275-1281, May 1998.

[83] S.-J. Yang, J.-H. Lee and B.-C. Chieu, "Perfect-reconstruction filter banks having linear-phase FIR filters with equiripple response," *IEEE Trans. Signal Porcess.,* vol. 46, pp. 3246-3255, Dec. 1998.

[84] C.-Y Yao, "A study of SPT-term distribution of CSD numbers and its application for designing fixed-point linear phase FIR filters," in *Proc. IEEE Int. Symp. Circuits and Syst., ISCAS2001*, vol. II, pp. 301-304.

[85] J. Yli-Kaakien and T. Saramäki, "Design of very low-sensitivity and low-noise recursive filters using a cascade of low order lattice wave digital filters," *IEEE Trans. Circits Syst. II,* vol. 46, pp. 906-914, July 1999.

[86] Q. Zhao and Y. Tadakoro, "A simple design of FIR filters with power-of-two coefficients," *IEEE Trans. Circits Syst.,* vol. 35, pp. 566-570, May 1988.