TIME-VARYING AND FINITE FIELD FILTER BANKS

Thesis by

See-May Phoong

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Pasadena, California

1996

(Submitted May 20, 1996)

Acknowledgement

First of all, I would like to thank my advisor, Professor Vaidyanathan. Without him, this thesis would never have been possible. It has been my greatest fortune to work under his excellent guidance. He was always there to offer invaluable advice, constant support, and to unselfishly share his vast knowledge with us all. I have always admired his insight and persistence in attacking research problems. There is no way that I can describe my appreciation and gratitude towards him in one paragraph. Not only is he the best teacher that a student can ever dream of, he is also the best friend. It is this friendship that has made my stay at Caltech an enjoyable and unforgettable experience. I will never forget all the wonderful moments we have had at Caltech and during our conferences.

I also want to thank Professors Jehoshua Bruck, Michelle Effros, Robert J. McEliece, and Dr. Xiang-Gen Xia. I appreciate their interest in and comments on my research, and also their time serving on my examination committee.

I would also like to thank my colleagues and friends, Dr. Truong Nguyen, Dr. Tsuhan Chen, Dr. Anand Soman, Dr. Ming-Chieh Lee, Dr. Igor Djokovic, Yuan-Pei Lin, Ahmet Kirac and Jamal Tuqan. Their useful comments on my work are greatly appreciated. I want to thank them and other friends at Caltech for sharing many delightful moments with me. I also want to thank Robert Freeman, Lilian Porter, and Lavonne Martin in the Electrical Engineering department for their help.

Finally, I am most grateful to my parents. They provided me with the best education during my childhood, which I will never forget. I owe a lot to them, and my brothers and sisters for their undivided love, constant support and encouragement.

Abstract

Filter banks find many applications in signal processing. This thesis deals with four different problems in filter banks.

First we find a new application of filter banks: Filter bank convolver. We prove two filter bank convolution theorems which tell us how to do the convolution in the subbands. Applying the multirate technique to the problem of convolution, we obtain a significant improvement in the accuracy of the convolutional result when the computation is done with finite precision. The derivation also leads to a low sensitivity robust structure for FIR filters.

In the second part, a new class of two-channel biorthogonal filter banks is proposed. We successfully design IIR filter banks which achieve the following desired properties simultaneously: (i) Perfect reconstruction (PR); (ii) causality and stability; (iii) near linear-phase; (iv) frequency selectivity. Two classes of causal stable maximally flat IIR wavelets are derived and closed form formulas are given. We also provide a novel mapping of the proposed 1D framework into 2D. The mapping preserves: (i) PR; (ii) stability in the IIR case and linear phase in the FIR case; (iii) frequency selectivity; (iv) low complexity.

In the third part, the theory of paraunitary (PU) filter banks is extended to the case of $GF(q)$ with prime q. We show that finite field PU filter banks are very different from real or complex PU filter banks. Despite all the differences, we are able to prove a number of factorization theorems. All unitary matrices in $GF(q)$ are factorizable in terms of Householder-like matrices. The class of first-order PU matrices, the lapped orthogonal transform in finite fields, can always be expressed as a product of degree-one or degree-two building blocks.

Finally the theory of conventional LTI filter banks is extended to the time-varying case. We develop a polyphase representation method for time-varying filter bank (TVFB). Using the proposed polyphase approach, we are able to show some unusual properties which are not exhibited by the conventional LTI filter banks. For example, we can show that for a PR TVFB, the losslessness of analysis bank does not always imply that of the synthesis bank, and a PR TVFB in general will only generate a discrete-time frame, rather than a basis, for the class of finite energy signals. The class of lossless TVFB is studied in detail. We show that all lossless linear time-varying systems are invertible and provide explicit construction of the inverse. The interplay between invertibility, uniqueness and losslessness of the inverse is investigated. The factorizability of lossless TVFB is addressed and we show that there are factorizable and unfactorizable examples.

Table of Contents

List of Figures

1

Introduction

Multirate systems and filter banks find many applications in signal processing [Vai93]. For example, they have been applied successfully to the compression of video/image signals [Woo91, Tek95], to the problems in communication [Vet86, Xia95a], to the processing of audio/speech signals [Jay84], to the encryption of voice data [Cox87], to the problem of data rate conversion [Ans83], to the sampling of nonbandlimited signals [Wal92, Djo94a, Vai95a, Vai95b], and to the robust implementation of convolution [Vet88, Vai93a, Lin94, Pho95]. Some textbooks on the topic of filter banks are [Mal92, Vai93, Fli94, Vet95, Str96]. For a brief history of filter banks, the readers are referred to Chapter 1 of [Vai93].

Consider the M-channel filter bank in Fig. 1.1.1. In a filter bank, the input signal $x(n)$ is split into subband signals $y_k(n)$ by a set of filters $H_k(z)$ (called the analysis filters) at the analysis end. A typical magnitude response of these analysis filters is shown in Fig. 1.1.2. Hence $y_k(n)$ can often be regarded as the content of $x(n)$ at different frequency locations. Depending on the applications, these subband signals are first processed and then transmitted or stored. At the receiving end, the processed subband signals $\widehat{y}_k(n)$ are recombined by using a set of filters $F_k(z)$ (called the synthesis filters) to get the output signal $\widehat{x}(n)$. In the absence of subband processing (i.e., $\widehat{y}_k(n) = y_k(n)$), if the output $\widehat{x}(n) = x(n)$ for all possible input $x(n)$, then the filter bank is said to have *perfect reconstruction* (PR). All the filter banks considered in this thesis have the PR property.

To explain the advantages of multirate filter banks, we take subband coding as an example. Subband coding technique is widely used in the compression of audio, image and video signals. Some of the advantages are listed below:

1. The perceptual properties of human visual/auditory systems can be incorporated easily in the process of coding [Woo91, Tek95]. By exploiting the fact that human eyes/ears have different sensitivity at different frequencies, high quality coded signals can be obtained at a moderate bit rate.
2. The idea of multiresolution analysis [Mal89] can be easily implemented in the subband coding technique. In a filter bank, the subband signal at the lowpass channel can be viewed as a low resolution approximation of the original input while the subband signals in other bandpass channnels can be regarded as details of the original input. By successive splitting of the lowpass channel, one obtains a coarser approximation.

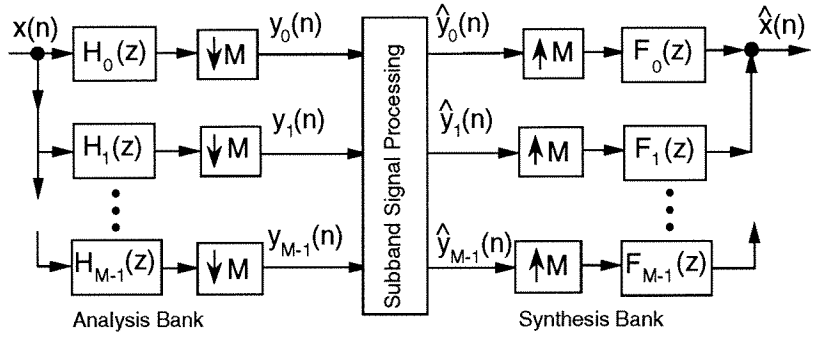

Fig. 1.1.1. M-channel filter bank.

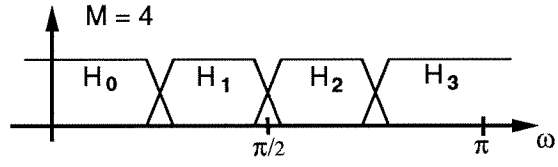

Fig. 1.1.2. Typical magnitude response of the analysis filters.

3. Compared to the discrete cosine transform, filter banks provide better frequency separation and hence can achieve a higher compression ratio for a given SNR criteria. Moreover blocking effect (which is very serious in discrete cosine transform coder) is in general reduced significantly in subband coding.

1.1. SCOPE AND OUTLINE

This thesis consists of five journal papers which deal with different aspects of filter banks. The five main results will be presented in Chapters 2–6. We shall give a summary of these results below.

1.1.1. Application of Filter Banks to Convolution Problem (Chapter 2)

Convolution plays a central role in digital signal processing. In some applications, due to the hardware limitation, we have to quantize the sequences to a low bit rate. Therefore it is important to develop a robust convolution algorithm for this purpose. The filter bank convolver is such an algorithm. It is well-known [Vai93] that in subband coding, we can exploit the energy distribution of a signal to obtain a more accurate representation of the signal with a fixed number of bits. In the problem of convolution, we have two sequences. If the convolution is done with finite precision, we will show how the energy distribution of *both* sequences can be exploited to obtain a more accurate convolution result compared to direct convolution. We will prove two filter bank convolution theorems which tell us how the convolution result can be obtained from the subband signals. Optimal bit allocation and coding gain over the direct convolution are derived. In the case of orthonormal filter banks, the convolutional coding gain is shown to be always greater than unity. Experiments demonstrate that coding gain of more than 20 dB can be achieved.

In most cases, one of the sequences is the impulse response of a digital filter. In this case, the ideas of filter bank convolvers can be used to implement FIR filters. As the filter bank convolvers produce a much more accurate result than the direct convolution, this implementation is robust against the coefficient quantization. Therefore the derivation of filter bank convolvers leads to a novel low sensitivity structure for FIR filters. The proposed structure is particularly attractive when the filter is frequency selective and has a long impulse response, or it has some special time-frequency relation. Examples show that the subband filter coefficients can be quantized to an averaged bit rate of 2 bits without degrading the frequency response. The results of Chapter 2 have been reported in [Pho93, Vai93a, Pho95].

1.1.2. Two-Channel PR Filter Banks with Causal Stable IIR Filters and Linear Phase FIR Filters (Chapter 3)

The special case of two-channel filter banks (Fig. 1.1.3) is probably the most studied case in the literature. What makes two-channel filter banks so attractive is the following reason. By cascading two-channel filter banks in a tree-structured manner, one can generate discrete-time wavelet transforms [Vai93, Som93]. These nonuniform filter banks have the advantage that the supports of the analysis/synthesis filters match the critical bands in the perceptual models of human visual and auditory systems. Thus they are widely applied in subband coding of audio, video/image signals [Woo91, Tek95].

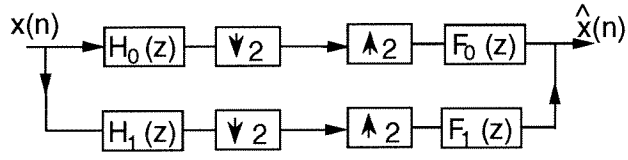

Fig. 1.1.3. Two-channel filter bank.

In Chapter 4, we shall restrict ourself to the special case of two-channel filter banks as shown in Fig. 1.1.3. In FIR filter banks, all the four filters H_0, H_1, F_0, and F_1, are FIR filters while in the case of IIR filter banks, some or all of these filters are IIR filters. The following are the desired properties of an IIR filter bank: (i) PR; (ii) stability; (iii) causality; (iv) frequency selectivity. The earliest good designs for the IIR case were such that the analysis bank was paraunitary and the polyphase components of $H_0(z)$ and $H_1(z)$ were allpass [Vai87b]. Some other design techniques for PR filter banks with non causal IIR filters have also been reported [Ram88, Mit92, Her93]. However, none of these technique achieved Properties (i)–(iv) simultaneously. Recently the authors in [Bas95] proposed an IIR PR technique providing causal stable solutions, but no satisfactory design method was given.

In Chapter 3, we propose a novel framework for a class of two-channel PR filter banks [Pho94, Pho95a]. The framework covers two useful subclasses: (i) *causal stable* IIR filter banks; (ii) *linear phase* FIR filter banks. The proposed IIR filter banks achieve Properties (i)–(iv) simultaneously. There exists a very efficient structurally PR implementation for such a class. PR is attained even when all the filter

coefficients are quantized. Filter banks of high frequency selectivity can be achieved by using the proposed framework with low complexity. The design of the four filters H_0, H_1, F_0, and F_1 reduces to the design of a single transfer function. Very simple design methods are given both for FIR and IIR cases. In addition to these advantages, a number of other useful properties will be elaborated in Chapter 3.

Moreover, our proposed construction can easily impose zeros of arbitrary multiplicity at the aliasing frequency, for the purpose of generating wavelets with regularity property. In the IIR case, two new classes of IIR *maximally flat* filters different from Butterworth filters are introduced. Closed form formulas are given for the coefficients of these maximally flat filters. We also generate the wavelet bases corresponding to the PR filter banks and show that smooth wavelets can be obtained.

We also introduce a new design method of 2D nonseparable filter banks in Chapter 3. We will provide a novel mapping of the proposed 1D framework into 2D. The mapping preserves all the following desired properties: (i) PR; (ii) stability in the IIR case; (iii) linear phase in the FIR case; (iv) zeros at aliasing frequency; (v) frequency characteristic of the filters. The results of Chapter 3 have been reported in [Pho94, Pho95a, Pho95d].

1.1.3. Theory of Paraunitary Filter Banks over Finite Fields (Chapter 4)

In real and complex fields, unitary and paraunitary (PU) matrices have found many applications in signal processing [Vai93]. What makes PU filter banks so attractive in the application of subband coding is that both the analysis and synthesis banks have the energy preservation property. Therefore any error introduced in the subbands will not be amplified. Moreover for the class of PU matrices, a useful *factorization theorem* can be proved [Vai88, Dog88, Vai93]. The theorem gives a *complete* and *minimal* characterization of PU matrices in terms of their delays and free parameters. The structure that is derived from the theorem is very valuable in both the design and implementation of PU filter banks [Vai93].

In Chapter 4, we will extend the theory of PU filter banks to the case of $GF(q)$ with q prime [Pho96a, Pho95f]. Various properties of unitary and PU matrices in finite fields will be studied. In particular, a number of factorization theorems will be given. We will show that: (i) All unitary matrices in $GF(q)$ are factorizable in terms of Householder-like matrices and permutation matrices; (ii) the class of first-order PU matrices, the lapped orthogonal transform in finite fields, can always be expressed as a product of degree-one or degree-two building blocks. If $q > 2$, we do not need degree-two building blocks. While many properties of PU matrices in finite fields are similar to those of PU matrices in real or complex fields, there are a number of differences. For example, unlike the conventional PU systems, in finite fields there are PU systems that are *unfactorizable* in terms of smaller building blocks. Even though the case of $GF(q)$ with prime $q > 2$ shares some similarities with the $GF(2)$ case, there are many differences. The results of Chapter 4 have been reported in [Pho96a, Pho95f].

1.1.4. Basic Principles of Time-Varying Filter Banks (Chapter 5)

Filter banks have been successfully applied in the compression of video/image signals. In most applications, the conventional linear time-invariant (LTI) filter banks, where all the analysis and synthesis filters

are LTI, are used. The video/image signals coded by using such LTI filter banks suffer from ringing effect, and their edges become blurred, especially for low bit-rate coding. These problems can be solved by using a time-varying filter bank (TVFB) as shown in Fig. 1.1.4. A TVFB is a filter bank with time-varying analysis filters $h_k^i(n)$ and time-varying synthesis filters $f_k^i(n)$. TVFBs have the ability to adapt to the characteristics (such as edges, textures, smooth regions, etc.) of different regions of an image. Filters of different properties can be applied to different regions of an image in a time (or space) varying manner. By exploiting this flexibility of TVFBs, we can achieve good quality images (with sharp edges, little blocking effect, and small ringing effect) at a low bit rate [Smi95, Chu93]. As the performance of TVFBs is better than the conventional LTI filter banks, it is worthwhile studying their theory.

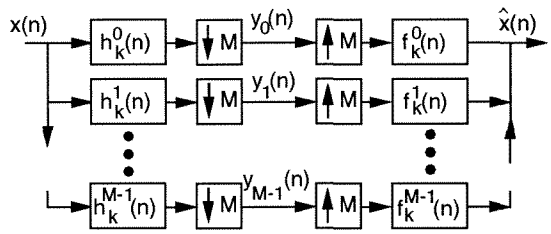

Fig. 1.1.4. Time-varying filter bank.

In Chapter 5, we study the fundamentals of TVFBs [Pho95b, Pho96]. As there is no z-transform in the linear time-varying (LTV) case, the study of LTV filters is in general based on time domain techniques. Here we introduce a transform domain description for LTV filters. We also develop a polyphase representation method for TVFBs. Using the proposed polyphase approach to TVFBs, we are able to show some unusual properties which are not exhibited by the conventional LTI filter banks. For example, we can show that for a PR TVFB, the losslessness of analysis bank does not always imply that of the synthesis bank, and replacing the delay z^{-1} in an implementation of a lossless LTV system with z^{-L} for integer L in general will result in a non lossless system. Moreover, we show that interchanging the analysis and synthesis filters of a PR TVFB will usually destroy the PR property, and a PR TVFB in general will *not* generate a discrete-time basis for the class of finite energy signals.

Furthermore we show that we can characterize all TVFBs by characterizing multi-input multi-output (MIMO) LTV systems. A useful subclass of LTV systems, namely the lossless systems, is studied in detail. All lossless LTV systems are invertible. Moreover the inverse is FIR if the original lossless system is FIR. Explicit construction of the inverses is given. However, unlike in the LTI case, we show that the inverse system is not necessarily unique or invertible. In fact, the inverse of a lossless LTV system is not necessarily lossless. Depending on the invertibility of their inverses, the lossless systems are divided into two groups: (i) Invertible inverse lossless (IIL) systems; (ii) non invertible inverse lossless (NIL) systems. We show that a NIL PR TVFB will only generate a discrete-time *tight frame* with unity frame bound. However if the PR FB is IIL, we will have an *orthonormal basis* for the class of finite energy signals. The results of Chapter 5 have been reported in [Pho95b, Pho95g, Pho96].

1.1.5. *Factorizability of Lossless Time-Varying Filters and Filter Banks* (Chapter 6)

In the LTI case, we know that all PU filter banks are factorizable [Vai93]. In Chapter 6, we shall study the factorizability of LTV lossless filters and filter banks [Pho95c, Pho95g]. We give a *complete* characterization of all degree-one lossless LTV systems. This is useful for both building and factorizing higher order lossless LTV systems. The traditional lapped orthogonal transform (LOT) [Mal92] is also generalized to the LTV case [Pho95e]. We identify two classes of time-varying LOT (TVLOT), namely the invertible inverse lossless (IIL) and non invertible inverse lossless (NIL) TVLOTs. We show that all IIL TVLOTs can be factorized *uniquely* into the proposed degree-one lossless building block. The factorization is *minimal* in terms of delay elements. For NIL TVLOTs, there are factorizable and unfactorizable examples. Both necessary conditions and sufficient conditions for factorizability of lossless LTV systems are given. These conditions lead to simple testing methods for the factorizability of lossless systems.

In control theory, the concepts of reachability, observability and minimality were proved to be very valuable in the study of LTI systems [Kai80, Deca89]. In Chapter 6, we will introduce the concepts of *strong eternal reachability* (SER) and *strong eternal observability* (SEO) of LTV systems. The SER and SEO of an implementation of LTV systems imply the *minimality* of the structure. Using these concepts, we are able to show that the cascade structure for a factorizable IIL LTV system is *minimal*. That implies that if an IIL LTV system is factorizable in terms of the lossless degree-one building blocks, the factorization is minimal in terms of delays as well as the number of building blocks. The results of Chapter 6 have been reported in [Pho95c, Pho95e, Pho95g].

1.2. NOTATIONS AND PRELIMINARIES

Notations: Throughout this thesis, we shall use the following notations:

1. Boldfaced lower case letters (such as $\mathbf{u}$, $\mathbf{v}$) represent vectors and boldfaced upper case letters (such as $\mathbf{U}$, $\mathbf{V}$) represent matrices. The symbol $\mathbf{I}$ is reserved for the identity matrix.
2. The notations $\mathbf{A}^T$ and $\mathbf{A}^*$ denote the transpose and the complex conjugate of $\mathbf{A}$ respectively. The notation $\mathbf{A}^\dagger$ denotes complex conjugation followed by transposition, i.e., $\mathbf{A}^\dagger = (\mathbf{A}^*)^T$.
3. A rational or polynomial matrix is denoted by $\mathbf{A}(z)$. One useful operation in the study of filter banks is the tilde operation. The tilde of $\mathbf{A}(z)$ is $\widetilde{\mathbf{A}}(z) = \mathbf{A}^\dagger(1/z^*)$.
4. The transfer functions $H_k(z)$ and $F_k(z)$ represent the k-th analysis and synthesis filters respectively.
5. The notations $(V(z))_{\downarrow M}$ and $(V(z))_{\uparrow M}$ denote the M-fold decimated and M-fold expanded versions of the signal $v(n)$ respectively.

Order and Degree: Consider a causal polynomial (i.e., FIR) matrix $\mathbf{E}(\mathbf{z}) = \sum_{k=0}^{N} \mathbf{e}(k)z^{-k}$ with $\mathbf{e}(N) \neq \mathbf{0}$. The order of $\mathbf{E}(\mathbf{z})$ is N, whereas the McMillan degree (often called just degree) is the smallest number of delays with which we can implement the system. For example, if $\mathbf{E}(\mathbf{z}) = \mathbf{e}(0) + z^{-1}\mathbf{e}(1)$ with $\mathbf{e}(1) \neq \mathbf{0}$, then its order $= 1$, whereas its degree is equal to the rank of the matrix $\mathbf{e}(1)$.

Polyphase Representations: The polyphase representations were first introduced in [Bel76]. It was later proved to be valuable in both the theory and design of filter banks [Vet86a, Swam86]. Consider a set of filters $H_k(z)$, $k = 0,\ 1,\ \ldots\ M-1$. They can be uniquely written in terms of their M polyphase components as $H_k(z) = \sum_{l=0}^{M-1} z^{-l} E_{kl}(z^M)$. This is known as Type 1 polyphase representation and $E_{kl}(z)$ is called the l-th polyphase component of $H_k(z)$. The $M \times M$ matrix $\mathbf{E}(z)$, with its kl-th element $[\mathbf{E}(z)]_{kl} = E_{kl}(z)$, is called the Type 1 polyphase matrix of the filters $H_k(z)$. Similarly, $F_k(z)$ can be written in terms of their Type 2 polyphase components as $F_k(z) = \sum_{l=0}^{M-1} z^{l} R_{lk}(z^M)$. The Type 2 polyphase matrix $\mathbf{R}(z)$ of the filters $F_k(z)$ is defined as $[\mathbf{R}(z)]_{lk} = R_{lk}(z)$.

Perfect Reconstruction (PR) Filter Banks: Consider Fig. 1.1.1. In the absence of subband processing, if the output $\hat{x}(n) = x(n)$ for all possible input $x(n)$, then the filter bank is said to have PR. All the filter banks considered in this thesis have the PR property. Using the polyphase representations introduced above, Fig. 1.1.1 can be redrawn as Fig. 1.2.1. So the PR condition reduces to

$$\mathbf{R}(z)\mathbf{E}(z) = \mathbf{I}. \tag{1.2.1}$$

In other words, the filter bank has PR if and only if $\mathbf{R}(z) = \mathbf{E}^{-1}(z)$. Given that $\mathbf{E}(z)$ is a FIR matrix, in general $\mathbf{R}(z)$ will be IIR. It is not difficult to see [Vai93] that $\mathbf{R}(z)$ is FIR if and only if $\det[\mathbf{E}(z)] = cz^{-n}$, for some nonzero constant c and integer n.

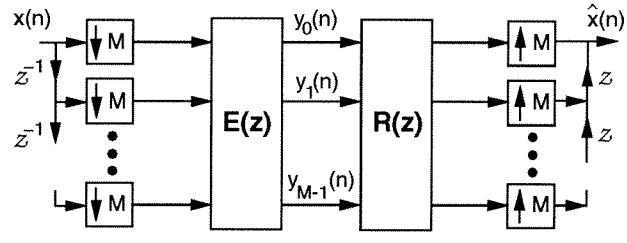

Fig. 1.2.1. Polyphase representation of the filter bank in Fig. 1.1.1.

Paraunitary (PU) Filter Banks: Within the class of PR filter banks, there is a useful subclass called PU filter banks [Vai93]. For a PU filter bank, the analysis polyphase matrix satisfies

$$\widetilde{\mathbf{E}}(z)\mathbf{E}(z) = \mathbf{I}, \tag{1.2.2}$$

where the tilde notation is defined above. In this case, PR can be obtained by simply taking $\mathbf{R}(z) = \widetilde{\mathbf{E}}(z)$. This in particular implies that: (i) The synthesis filters are FIR if the analysis filters are FIR; (ii) the synthesis and analysis filters are related as $F_k(z) = \widetilde{H}_k(z)$ (hence $|H_k(e^{j\omega})| = |F_k(e^{j\omega})|$). So in the design, we have to optimize one set of filters only. Another reason why PU filter banks became popular in various applications is because they enjoy the following energy conservation property [Vai93]:

$$\sum_n |x(n)|^2 = \sum_{k=0}^{M-1} |y_k(n)|^2, \tag{1.2.3a}$$

$$\sum_n |\widehat{x}(n)|^2 = \sum_{k=0}^{M-1} |\widehat{y}_k(n)|^2, \tag{1.2.3b}$$

for all finite energy signals. Hence PU filter banks are also known as *lossless* filter banks.

Biorthogonal and Orthonormal Bases: Consider Fig. 1.1.1. Assume that the filter bank has PR and there is no processing in the subbands so that $\widehat{x}(n) = x(n)$ and $\widehat{y}_k(n) = y_k(n)$. Then the output signal $x(n)$ and the subband signals $y_k(n)$ can respectively be expressed as

$$x(n) = \sum_{k=0}^{M-1} \sum_m y_k(m) f_k(n - Mm), \quad y_k(n) = \sum_l x(l) h_k(Mn - l). \tag{1.2.4}$$

If we define the families of the double index functions $\eta_{km}(n)$ and $\theta_{km}(n)$ respectively as

$$\eta_{km}(n) = f_k(n - Mm), \quad \theta^*_{km}(n) = h_k(Mm - n), \tag{1.2.5}$$

where $0 \leq k \leq M-1$ and $-\infty \leq m \leq \infty$, then (1.2.4) can be rewritten as

$$x(n) = \sum_{k=0}^{M-1} \sum_m \alpha_{km} \eta_{km}(n), \quad \text{where} \quad \alpha_{km} = \sum_n x(n) \theta^*_{km}(n). \tag{1.2.6}$$

The functions $\eta_{km}(n)$ and $\theta_{km}(n)$ are respectively called the synthesis and analysis functions in [Che94]. Hence we can view the subband splitting as the decomposition of signal $x(n)$ in terms of the basis functions $\eta_{km}(n)$. The coefficients α_{km} can be computed as the inner product defined in (1.2.6). However the basis functions are not arbitrary. All the analysis and synthesis functions, $\theta_{km}(n)$ and $\eta_{km}(n)$ for a fixed k, are respectively shifted versions of $\theta_{k0}(n)$ and $\eta_{k0}(n)$. That means for all m, $\theta_{km}(n)$ and $\eta_{km}(n)$ have the same shape as $\theta_{k0}(n)$ and $\eta_{k0}(n)$ respectively. By taking this signal decomposition viewpoint, it is known [Che94, Djo94] that a filter bank has PR if and only if the corresponding analysis/synthesis functions satisfy the following biorthogonality property:

$$\sum_n \eta_{k_1 m_1}(n) \theta^*_{k_2 m_2}(n) = \delta(k_1 - k_2)\delta(m_1 - m_2). \tag{1.2.7}$$

Therefore a PR filter bank is sometimes called a biorthogonal filter bank. In the special case of PU or lossless filter banks, the analysis and synthesis functions are identical [Che94].

$$\theta_{km}(n) = \eta_{km}(n). \tag{1.2.8}$$

The synthesis functions satisfy the following orthonormality property:

$$\sum_n \eta_{k_1 m_1}(n) \eta^*_{k_2 m_2}(n) = \delta(k_1 - k_2)\delta(m_1 - m_2). \tag{1.2.9}$$

The synthesis functions form an orthonormal basis for the class of finite energy signals. Therefore a PU filter bank is also known as an orthonormal filter bank. In this thesis, we will use the terms PU, lossless, orthonormal equivalently.

Factorization of PU Filter Banks: In addition to the advantage of energy conservation property, PU filter banks allow a factorization theorem. Every causal FIR PU matrix $\mathbf{E}(z)$ can be decomposed into a cascade of simple building blocks $\mathbf{D}(z)$, as shown in Fig. 1.2.2. The factorization has the following features [Vai93]:

1. Each building block $\mathbf{D}(z)$ has a simple form. PU property can be preserved by simply making $\mathbf{v}_k^\dagger \mathbf{v}_k = 1$ (which can be obtained by the implementation using the planar rotations). Hence PR is attained even when all the coefficients are quantized.
2. The factorization is *complete*. The cascade structure in Fig. 1.2.2 captures all FIR PU matrices.
3. The factorization is *minimal*. The number of delay used in the cascade structure is the McMillan degree of the PU matrix, which is the smallest number of delay required to implement the system.

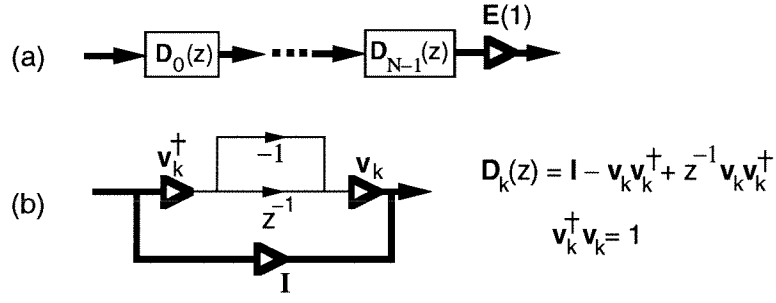

Fig. 1.2.2. Factorization of PU matrix and the degree-one building block.

Due to Feature 1, the PR condition is significantly simplified. Because of the above three features, the cascade structure in Fig. 1.2.2 is very useful in the design of filter banks [Vai93]. To summarize the results, we present the factorization theorem below [Vai93]:

Theorem 1.2.1. *Factorization of $M \times M$ FIR PU Matrices:* Let $\mathbf{E}(z)$ be a $M \times M$ causal FIR matrix. Then $\mathbf{E}(z)$ is PU with $\det[\mathbf{E}(z)] = cz^{-N}$ if and only if it can be expressed as:

$$\mathbf{E}(z) = \mathbf{E}(1)\ \mathbf{D}_{N-1}(z)\ \mathbf{D}_{N-2}(z) \ldots \mathbf{D}_0(z), \tag{1.2.10}$$

where $\mathbf{E}(1)$ is some unitary matrix and $\mathbf{D}_k(z)$ is a degree-one PU matrix of the form $\mathbf{I} - \mathbf{v}_k\mathbf{v}_k^\dagger + z^{-1}\mathbf{v}_k\mathbf{v}_k^\dagger$ with $\mathbf{v}_k^\dagger \mathbf{v}_k = 1$. The implementation of the cascade structure is given in Fig. 1.2.2. ■

Remark: Other types of factorization have also been reported in the literature. In [Ngu89], the authors derived a lattice structure for the class of two-channel linear phase filter banks. In [Som93b], a complete and minimal factorization theorem was derived for a subclass of M-channel linear phase filter banks.

2

One- and Two-Level Filter Bank Convolvers

2.1. INTRODUCTION

Convolution plays a central role in digital signal processing. Many well-known algorithms are proposed to reduce the computational complexity of convolution [Bla85]. In this chapter, our aim is not to find an algorithm that is faster than existing fast algorithms. Our goal is to find a more accurate way to compute the convolution when the convolution is implemented with finite precision. For this we use filter bank techniques.

2.1.1. Previous and Main Results of this Chapter

Consider Fig. 2.1.1, where a nonuniform filter bank (FB) is shown. Suppose that the filters $H_k(z)$ and $F_k(z)$ form a perfect reconstruction (PR) system. It was shown [Vai93a] that we can obtain the convolution of $x(n)$ and $g(n)$ by simply convolving $x_k(n)$ and $g_k^{(i)}(n)$ and adding the results. No *cross*-convolution between the subband signals is involved. When the computation is done with finite precision, it was also shown in [Vai93a] how the energy distribution in the subbands of $x(n)$ and $g(n)$ can be exploited to obtain a more accurate (compared to direct convolution) result. In this chapter, we further generalize the subband convolution theorem. We will also show that the coding gain for the generalized convolver is always greater than that derived in [Vai93a]. We will refer to the convolution theorem derived in [Vai93a] as *one-level* FB convolution theorem and the generalized theorem in this chapter as *two-level* FB convolution theorem.

In [Vai93a], only the quantization in the subbands of $x(n)$ was considered. In this chapter, we will address the case when the subband of filter $g(n)$ is quantized. In the process of quantization, the filter coefficients are treated as *deterministic* parameters instead of random variables as done in [Cha73]. Thus overflow of subband coefficients is completely avoided. The derivation leads to a novel low sensitivity structure for FIR filters. The new structure is particularly attractive when the filter $g(n)$ is frequency selective and has a long impulse response, or it has some special time-frequency relation, e.g. the matched filtering of a chirp signal in radar application [Ste91].

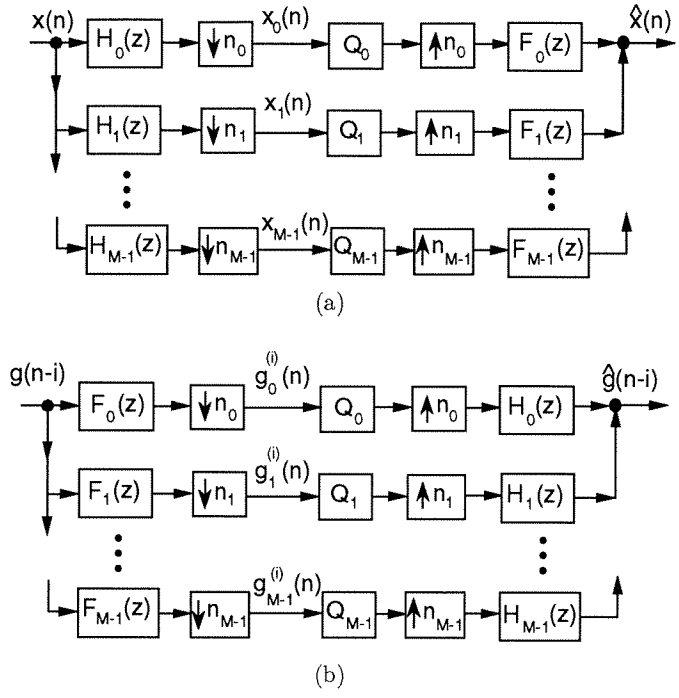

Fig. 2.1.1. Maximally decimated filter bank: (a) With input $x(n)$; (b) with input $g(n-i)$.

In this chapter, we also explore the relationship between the convolver and the digital block filtering [Bur71, Mit78, Bar80]. We show that both the one-level and two-level FB convolvers are generalizations of the conventional block filtering. The subband convolvers have both the advantages of coding gain and parallelism. In the view of generalized block filtering, the structure used in [Gil92] can be regarded as a simplified version of the two-level FB convolver introduced in this chapter.

The filter bank techniques have been used in [Vet88, Lin94] to implement FIR and IIR filters. In [Lin94], the authors applied the low cost nonmaximally decimated DFT filter banks and cosine modulated filter banks to the problem of FIR filtering. It was shown [Lin94] that the complexity can be reduced significantly but the result suffered from some minor aliasing error. A different subband convolution theorem which leads to computational saving is derived in [Vet88]. The subband convolution theorem proposed is applied to the digital pulse compression in radar application in [Ste91]. Our subband convolution theorems differ from that derived in [Vet88, Ste91, Lin94] in the sense that the convolution is *'perfect'* regardless of filter responses of the filter bank.

2.1.2. Chapter Outline

Our presentation will go as follows: In Section 2.2, we will generalize the subband convolution theorem. A pictorial proof of the theorem is provided to give a clearer insight into what is going on in the convolution theorem. In Section 2.3, we consider the quantization of the input signal $x(n)$. The optimal bit allocation and coding gain for the two-level FB convolver are presented. A low sensitivity structure is derived in

Section 2.4. Several numerical examples are included in Section 2.5 to demonstrate the usefulness of the low sensitivity structures. In Section 2.6, We will discuss the relationship among conventional block filtering, and one-level and two-level FB convolvers. In Section 2.7, we will give a low sensitivity structure for linear phase filters which can simultaneously exploit the advantage of the coefficient symmetry and coding gain of the FB convolvers. In the last section, we will relate the IFIR filter to the subband convolver and consider the problem of implementing IIR filters using FB convolvers.

2.1.3. Preliminaries

Quantizers: By b bit quantizer, we mean that the output signal of the quantizer is represented by b bits plus a sign bit. The weight on the most significant bit is fixed for a fixed quantizer.

Maximal Decimation: An M-channel nonuniform multirate system is said to be maximally decimated if $\sum_{i=0}^{M-1} \frac{1}{n_k} = 1$. In the uniform case where all n_k are equal, this translates to $n_k = M$ for all k.

Generalized Polyphase (GPP) Representations [Vai90, Som94]: GPP was introduced in [Vai90] and used in [Som94] to enhance the coding gain of subband coding. Instead of expressing a signal $v(n)$ in terms of the functions $\{z^{-i}\}$ as in the conventional polyphase representation (Section 1.2), we express $v(n)$ in terms of the functions $\{U_i(z)\}$ as follows:

$$V(z) = \sum_{i=0}^{M-1} V_i(z^M) U_i(z). \tag{2.1.1}$$

Eq. (2.1.1) is said to be a valid GPP representation if the functions $\{U_i(z)\}$ (called a "polyphase basis") satisfy the conditions [Som94]: (i) Every rational function $V(z)$ can be expressed as (2.1.1), where $V_i(z)$ are rational; (ii) $V(z)$ is FIR if and only if $V_i(z)$ are FIR. Let $\mathbf{U}(z)$ to be the conventional polyphase matrix of $\{U_i(z)\}$, these conditions reduce to $\det[\mathbf{U}(z)] = cz^k$, for $c \neq 0$ and integer k. We will call $V_i(z)$ the i-th GPP component of the signal $v(n)$ with respect to the polyphase basis $\{U_i(z)\}$.

Orthonormality and Biorthogonality [Vet86a, Vai93a, Som93, Djo94]: For a uniform filter bank, the biorthogonality and orthonormality are explained in Section 1.2. For a nonuniform filter bank as shown in Fig. 2.1.1(a), the z-transform of the output is

$$\widehat{X}(z) = \sum_{k=0}^{M-1} X_k(z^{n_k}) F_k(z). \tag{2.1.2}$$

If $\widehat{x}(n) = x(n)$ for all $x(n)$, then the system is called a biorthogonal or PR filter bank. The biorthogonality of the filter bank translates to the following condition on the filters $H_k(z)$ and $F_k(z)$:

$$\Big[H_k(z)F_m(z)\Big]_{\downarrow n_{k,m}} = \delta(k-m), \tag{2.1.3}$$

where $n_{k,m} = \gcd(n_k, n_m)$. The set of filters $\{F_k(z)\}$ is said to be orthonormal if $\big(F_k(z)F_m^*(1/z^*)\big)_{\downarrow n_{k,m}} = \delta(k-m)$.

2.2. ONE- AND TWO-LEVEL FB CONVOLUTION THEOREM

2.2.1. Review of One-Level FB Convolution Theorem

Consider the two maximally decimated filter banks as shown in Fig. 2.1.1 (ignore the quantizers in the discussion of this section). Assume that the system has PR. Then we have the following biorthogonal convolution theorem:

Theorem 2.2.1. *One-Level FB Convolver [Vai93a]:* Consider Fig. 2.1.1. Assume that the system has PR. Define the integer $p_k = L/n_k$, where $L = \text{lcm}\{n_k\}$. Let $x_k(n)$ and $g_k^{(i)}(n)$ be the subband signals defined in Fig. 2.1.1(a) and (b) respectively. Then the i-th polyphase component, $y_i(n)$ of $x(n) * g(n)$ can be written as

$$y_i(n) = \Big(x(n) * g(n-i)\Big)_{\downarrow L} = \sum_{k=0}^{M-1} \Big(x_k(n) * g_k^{(i)}(n)\Big)_{\downarrow p_k}. \tag{2.2.1}$$

■

The advantage of the subband convolution is that we can compute the result more accurately when the convolution is implemented with finite precision. It was shown in [Vai93a] how we can quantize the subband signals $x_k(n)$, and reduce the quantization noise by optimally allocating the bits in the subbands. By exploiting the subband energy distribution, the optimal bit allocation scheme and the coding gain over direct convolution were derived in [Vai93a].

Complexity: Notice that the subband convolution theorem holds even when the analysis and the synthesis filters are IIR filters. But if we consider computational cost, the FB convolver is useful only when $H_k(z)$ and $F_k(z)$ are FIR filters. Thus in this chapter, we will consider FB convolvers with FIR analysis and synthesis filters only. Also note that the computation of $x_k(n)$ involves filtering. Since $g(n)$ is a fixed filter, the subband signals $g_k^{(i)}(n)$ can always be precomputed and stored. Thus the complexity of the subband convolution is approximately equal to that of direct convolution plus the cost of implementing an analysis bank, assuming that no fast algorithm for convolution is used. If the complexity of the filter bank is low (compared to the length of the sequences $x(n)$ and $g(n)$), then the computational cost of $x_k(n)$ is negligible compared to that of the convolution. In this case the complexity of subband convolution and that of direct convolution are approximately the same.

2.2.2. Two-Level FB Convolution Theorem

Theorem 2.2.2. *Two-Level FB Convolver:* Let $H_k(z)$, $F_k(z)$, n_k, p_k and L be the same as in Theorem 2.2.1 and let $\{H'_k(z)\}$ and $\{F'_k(z)\}$ be respectively the analysis and synthesis filters of a "L-channel" uniform biorthogonal system. Let $x_k(n)$ and $g_k^{(i)}(n)$ be respectively the k-th subband signals defined in Fig. 2.1.1(a) and Fig. 2.2.1. Then the i-th GPP component $y_i(n)$ of $x(n) * g(n)$ with respect to the polyphase basis $\{F'_i(z),\ i = 0,\ 1,\ \ldots,\ L-1\}$ can be written as

$$y_i(n) = \sum_{k=0}^{M-1} \Big(x_k(n) * g_k^{(i)}(n)\Big)_{\downarrow p_k}. \tag{2.2.2}$$

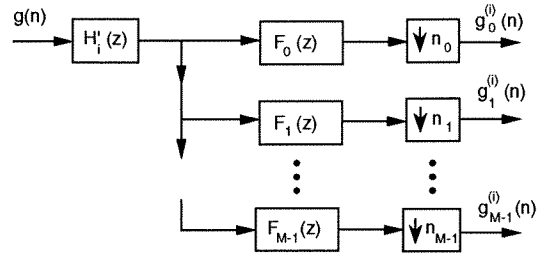

Fig. 2.2.1. Subband signals in two-level FB convolver.

■

Proof: From the definition of $X_k(z)$ and $G_k^{(i)}(z)$ and using the biorthogonality (2.1.2) of the filters $\{H_k(z)\}$ and $\{F_k(z)\}$, we get

$$X(z) = \sum_{k=0}^{M-1} X_k(z^{n_k})F_k(z), \quad \text{and} \quad G(z)H_i'(z) = \sum_{l=0}^{M-1} G_l^{(i)}(z^{n_l})H_l(z). \tag{2.2.3}$$

Multiplying the above two equations and decimating both sides by L, we have

$$\begin{aligned}\Big[X(z)G(z)H_i'(z)\Big]_{\downarrow L} &= \sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\Big[X_k(z^{n_k/n_{k,l}})G_l^{(i)}(z^{n_l/n_{k,l}})\Big]_{\downarrow p_{k,l}}\Big[F_k(z)H_l(z)\Big]_{\downarrow L} \\ &= \sum_{k=0}^{M-1}\Big[X_k(z)G_k^{(i)}(z)\Big]_{\downarrow p_k},\end{aligned} \tag{2.2.4}$$

where $n_{k,l} = \gcd(n_k, n_l)$ and the integer $p_{k,l} = L/n_{k,l}$. The second equality follows from (2.1.3) and the fact that $\big[V(z)\big]_{\downarrow L} = \big[\big(V(z)\big)_{\downarrow n_{k,l}}\big]_{\downarrow p_{k,l}}$. Applying the biorthogonality of the filters $\{H_i'(z)\}$ and $\{F_i(z)\}$ the left hand side of (2.2.4) is by definition the i-th GPP component of $X(z)G(z)$ with respect to the polyphase basis $\{F_i'(z)\}$. The proof is complete. ■

Eqn. (2.2.2) gives only the i-th GPP component of $x(n) * g(n)$. The convolution output $y(n)$ can be synthesized from the GPP components as follows:

$$Y(z) = \sum_{i=0}^{L-1} F_i'(z)Y_i(z^L) = \sum_{i=0}^{M-1} F_i'(z)\sum_{k=0}^{M-1} X_k(z^{n_k})G_k^{(i)}(z^{n_k}). \tag{2.2.5}$$

Remark: Notice that even for the nonuniform case, the second-level filter banks with filters $\{H_k'(z)\}$ and $\{F_k'(z)\}$ are constrained to be uniform filter banks with decimation ratio $L = \text{lcm}\{n_k\}$.

Comparison of One- and Two-Level FB Convolvers: Theorems 2.2.1 and 2.2.2 give us respectively the implementations of one- and two-level FB convolvers as Fig. 2.2.2(a) and (b). Since $g(n)$ passes through two levels of filter banks, we call the subband convolver in Fig. 2.2.2(b) two-level FB convolver. From these figures, the two-level FB convolver is clearly a generalization of the one-level FB convolver. By taking $H_k'(z) = z^{-i}$ and $F_k'(z) = z^i$, the two-level FB convolver reduces to the one-level FB convolver. The two-level FB convolver usually computes the convolution much more accurately than the one-level

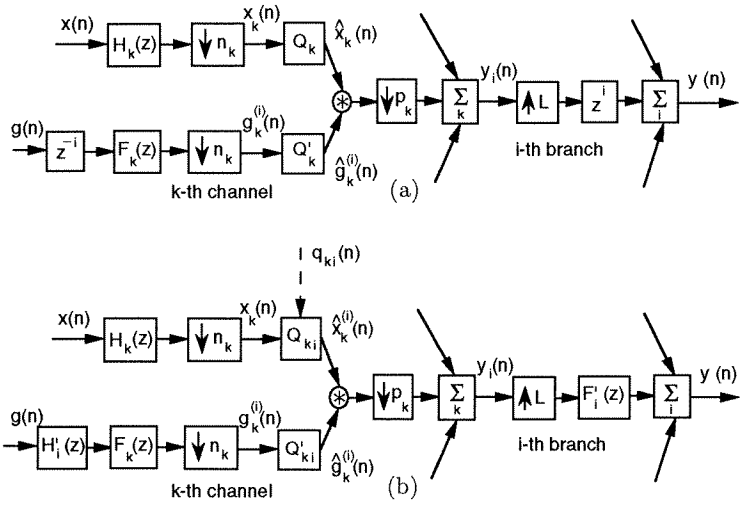

Fig. 2.2.2. Implementation of filter bank convolvers: (a) One-level FB convolver; (b) two-level FB convolver. Asterisk $*$ denotes convolution.

FB convolver, for the same average bit rate. The complexity of the former is that of the latter plus the cost of an additional synthesis bank $F_k'(z)$ (since $g_k^{(i)}(n)$ can be precomputed and stored). Thus if the complexity of the filter bank $\{F_k'(z)\}$ is low, then the complexity of the new subband convolution is comparable to that of direct convolution.

2.2.3. Pictorial Proof of the Subband Convolution Theorems

The above subband convolution theorems can be proved easily by using a sequence of figures. The pictorial proof of Theorem 2.2.1 leads us naturally to the two-level FB convolution theorem. It also gives a clear insight into what is going on in the subbands, and why perfect convolution is perserved when we pass from the one-level FB convolution theorem to the two-level FB convolution theorem. By using the same technique, the subband convolution theorems have been generalized to the most general case of the multidimensional nonuniform filter banks with rational decimation ratios [Che94].

Consider Fig. 2.2.3(a), where we want to compute $x(n) * g(n)$. Clearly, any two identity systems $\mathbf{I}_1$ and $\mathbf{I}_2$ can be inserted before and after the filter $G(z)$ without changing the convolution output, as shown in Fig. 2.2.3(b). If we choose the identity systems to be filter banks with PR, then we can utilize the frequency splitting property of the filter banks and quantize the subband signals according to the energy distribution in each subband. We may also select other identity systems, depending upon the task we want to perform. If we choose $\mathbf{I}_1$ to be the PR system shown in Fig. 2.1.1(a), and $\mathbf{I}_2$ to be an L-channel delay chain, then we can show that the equivalent system shown in Fig. 2.2.3(c) is the same as that depicted in Fig. 2.2.2(a). By using the fact that $L = n_k p_k$, i. e. an L-fold decimator is equivalent to an n_k-fold decimator followed by an p_k-fold decimator, the i-th branch of the system in Fig. 2.2.3(c) (i. e., the system from $x(n)$ to $y_i(n)$) can be redrawn as Fig. 2.2.3(d). Applying the identity in Fig. 2.2.3(e),

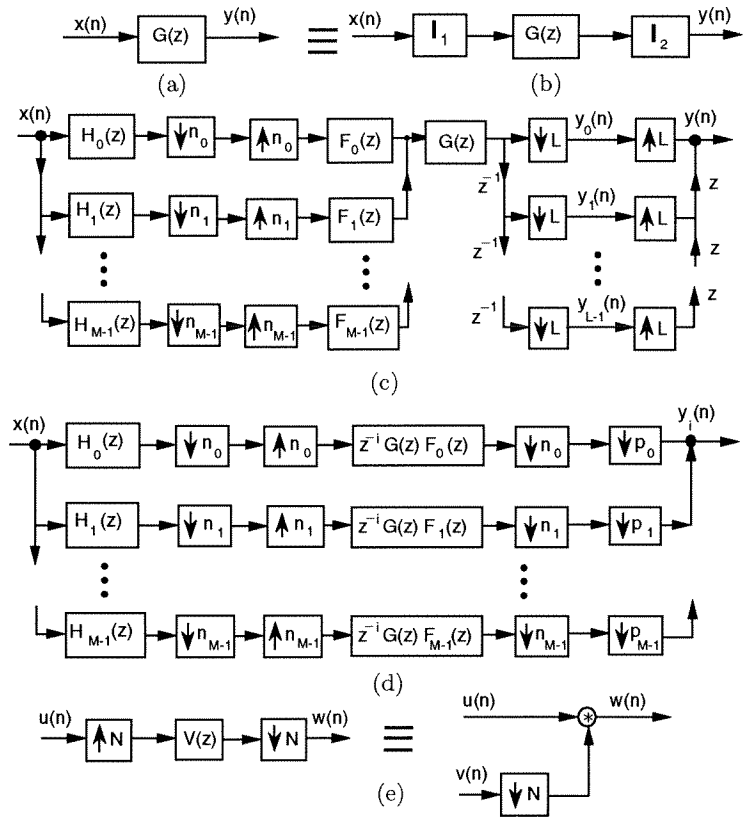

Fig. 2.2.3. Pictorial proof of Theorem 2.2.1: (a) Convolution of $x(n)$ and $g(n)$; (b) two identity systems inserted; (c) identity systems chosen to be two multirate systems; (d) i-th branch of (c); (e) an identity.

the system in Fig. 2.2.3(d) is equivalent to the i-th branch of the system in Fig. 2.2.2(a). This is the one-level FB convolver in Theorem 2.2.1.

Similarly, to prove Theorem 2.2.2, we select $\mathbf{I}_2$ to be an L-channel PR filter bank instead of a trivial delay chain. By carrying out exactly the same procedure as above, we can arrive at the result proved in Theorem 2.2.2.

Remark: In [Pho93a], the authors chose $\mathbf{I}_1$ to be a differential pulse coded modulator combined with error spectrum shaping and $\mathbf{I}_2$ to be the corresponding differential pulse coded demodulator. In this case, $\mathbf{I}_1$ and $\mathbf{I}_2$ are not identity systems but the product $\mathbf{I}_1\mathbf{I}_2$ is. *Perfect* convolution is achieved because all the operations involved are LTI. The authors also showed that high coding gain can be obtained by using such convolvers.

2.3. CODING GAIN OF TWO-LEVEL FB CONVOLVERS

In this section, we will consider the coding gain for the quantization of the input signal $x(n)$ only. The filter $g(n)$ is not quantized. For the case of one-level FB orthonormal convolver, the optimal bit allocation and coding gain were discussed in detail in Section 3.2 and 3.3 of [Vai93a] respectively. Since the uniform

convolver is strictly a special case of the nonuniform convolver, we will only derive the result for the nonuniform case.

Consider Fig. 2.2.2(b). Assume that all the signals and filter coefficients are real so that the quantizer operates on real inputs only. Let b_{ki} be the number of bits per sample of $x_k(n)$, allocated to Q_{ki}, the quantizer in the k-th channel in the i-th branch. Therefore the average bit rate is

$$b = \frac{1}{L}\sum_{i=0}^{L-1}\sum_{k=0}^{M-1}\frac{b_{ki}}{n_k}. \tag{2.3.1}$$

Since $g_k^{(i)}(n)$ usually have different energy for different i, b_{ki} vary greatly with respect to i (as we will see later). Therefore, we use double index for the bit rate.

2.3.1. The Noise Model

The error due to the quantizer Q_{ki} is defined as

$$q_{ki}(n) \triangleq \widehat{x}_k^{(i)}(n) - x_k(n), \tag{2.3.2}$$

where $\widehat{x}_k^{(i)}(n)$ is the quantized version of $x_k(n)$ in the i-th branch. The quantization error can be modeled as an additive noise source. Thus the quantizer Q_{ki} can be replaced by the broken line as shown in Fig. 2.2.2(b). To analyze the convolution error, we make the following assumptions:

1. $x(n)$ is zero mean wide sense stationary (WSS) with variance σ_x^2. Then $x_k(n)$ are also WSS, with variance
$$\sigma_{x_k}^2 = \int_0^{2\pi} S_{xx}(e^{j\omega})|H_k(e^{j\omega})|^2\frac{d\omega}{2\pi}, \tag{2.3.3}$$
where $S_{xx}(e^{j\omega})$ is the power spectrum of $x(n)$.
2. $g(n)$ is a deterministic sequence. We define a useful parameter α_{ki}^2 as
$$\alpha_{ki}^2 = M\sum_n |g_k^{(i)}(n)|^2, \tag{2.3.4}$$
where α_{ki}^2/M can be interpreted as the energy of the subband signal $g_k^{(i)}(n)$.
3. $q_{ki}(n)$ is zero mean white with variance $\sigma_{q_{ki}}^2$, where under certain conditions, $\sigma_{q_{ki}}^2$ is related to $\sigma_{x_k}^2$ as
$$\sigma_{q_{ki}}^2 = c\,\sigma_{x_k}^2 2^{-2b_{ki}}. \tag{2.3.5}$$
See Chapter 4 of [Jay84] or Appendix C of [Vai93]. Here c is a constant which depends only on the probability distribution of the subband signals $x_k(n)$. We have assumed that c is independent of k which is true only if all $x_k(n)$ have the same probability distribution.
4. The cross-correlation of $q_{ki}(n)$ is
$$E\{q_{ki}(n)q_{mi}(l)\} = \sigma_{q_{ki}}^2\delta(k-m)\delta(n-l), \tag{2.3.6a}$$

i. e., $q_{ki}(n)$ is uncorrelated to $q_{mi}(l)$ for $k \neq m$ and for all i, n, l. Notice that $E\{q_{ki}(n)q_{kj}(n)\}$ need not be zero for $i \neq j$. We also assume that $q_{ki}(n)$ is uncorrelated to the subband signals $x_k(l)$, that is

$$E\{x_k(l)q_{ki}(n)\} = 0. \tag{2.3.6b}$$

2.3.2. Optimal Bit Allocation and Coding Gain for the Two-Level FB Convolver

To derive the optimal bit allocation and coding gain formulas for the two-level FB convolver, we assume that the second set of synthesis filters $F_i'(z)$ are orthonormal. Consider Fig. 2.2.2(b). The error in the subband convolution output $y_i(n)$ is

$$q_{y_i}(n) = \sum_{k=0}^{M-1} \Big(q_{ki}(n) * g_k^{(i)}(n) \Big)_{\downarrow p_k}. \tag{2.3.7}$$

By using (2.3.4)-(2.3.6) and the fact that the decimator will not change the variance, the variance of $q_{y_i}(n)$ can be expressed as

$$\begin{aligned} \sigma_{q_{y_i}}^2 &= \sum_{k=0}^{M-1} \sigma_{q_{ki}}^2 \sum_l |g_k^{(i)}(l)|^2 \\ &= \frac{c}{M} \sum_{k=0}^{M-1} 2^{-2b_{ki}} \sigma_{x_k}^2 \alpha_{ki}^2, \quad \text{for} \quad 0 \le i \le L-1. \end{aligned} \tag{2.3.8}$$

Since the synthesis filters $F_i'(z)$ are orthonormal, the average variance over L samples, $\sigma_{q_y}^2$ of the output error is simply the average of $\sigma_{q_{y_i}}^2$, for $0 \le i \le L-1$, see Section C.4.2 of [Vai93] or [Som93a]. So we have

$$\sigma_{q_y}^2 = \frac{1}{L} \sum_{i=0}^{L-1} \sigma_{q_{y_i}}^2 = \frac{c}{ML} \sum_{i=0}^{L-1} \sum_{k=0}^{M-1} 2^{-2b_{ki}} \sigma_{x_k}^2 \alpha_{ki}^2. \tag{2.3.9}$$

To obtain the optimal bit allocation, we minimize the average output noise variance under the constraint (2.3.1). We form the Lagrangian

$$\phi = \sigma_{q_y}^2 - \lambda \Big(b - \frac{1}{L} \sum_{i=0}^{L-1} \sum_{k=0}^{M-1} \frac{b_{ki}}{n_k} \Big). \tag{2.3.10}$$

By setting $\partial\phi/\partial b_{ki} = 0$ for all $0 \le i \le L-1$, $0 \le k \le M-1$ and $\partial\phi/\partial\lambda = 0$, we get

$$n_k 2^{-2b_{ki}} \sigma_{x_k}^2 \alpha_{ki}^2 = D \quad \text{for} \quad 0 \le i \le L-1, \ 0 \le k \le M-1 \tag{2.3.11}$$

where D is a constant independent of i and k. Let γ_k^2 be the geometric mean of α_{ki}^2 over the index i, that is

$$\gamma_k^2 = \prod_{i=0}^{L-1} \left(\alpha_{ki}^2\right)^{1/L}. \tag{2.3.12}$$

By using (2.3.1), (2.3.11) and (2.3.12), we find that

$$D = 2^{-2b} \prod_{k=0}^{M-1} \left(n_k \sigma_{x_k}^2 \gamma_k^2 \right)^{1/n_k}. \tag{2.3.13}$$

Substituting (2.3.13) into (2.3.11), we find that the optimal number of bits allocated to the quantizer Q_{ki} at the k-th channel in the i-th branch is

$$b_{ki} = b + 0.5\log_2\left(n_k\sigma_{x_k}^2\alpha_{ki}^2\right) - 0.5\sum_{j=0}^{M-1}\log_2\left(n_j\sigma_{x_j}^2\gamma_j^2\right)^{1/n_j}. \tag{2.3.14}$$

Periodically Time-Varying Bit Allocation: Intuitively, we would assign more bits to those quantizers in branches where $g(n) * h_i'(n)$ has higher energy and in channels where $x_k(n)$ has higher energy. (2.3.14) tells exactly how this should be done according to the energy distribution. In the case of the one-level FB convolver, since $g_k^{(i)}(n)$ is simply obtained by time-shifting $g(n)$ (see Fig. 2.2.2(a)), we would expect that α_{ki}^2 will have very little dependency on i. In this case, b_{ki} are the same for all i and (2.3.14) reduces to (3.32) in [Vai93a]. However in the case of the two-level FB convolver, α_{ki}^2 may differ greatly for different i, especially when the filter $g(n)$ is a frequency selective filter (which is usually the case). Then b_{ki} may vary greatly with respect to i. In this case, not all branches are equally important as in the case of the one-level FB convolver, and the coding gain may increase significantly by using this "periodically time-varying" bit allocation scheme.

By using (2.3.11), (2.3.13) and the fact that the filter bank is maximally decimated ($\sum_{k=0}^{M-1}\frac{1}{n_k} = 1$), we find that the average output noise variance under optimal bit allocation is

$$\sigma_{q_y,opt}^2 = \frac{c\,D}{M} = \frac{c}{M}2^{-2b}\prod_{k=0}^{M-1}\left(n_k\sigma_{x_k}^2\gamma_k^2\right)^{1/n_k}. \tag{2.3.15}$$

If $x(n)$ is quantized to b bits, then in the direct convolution the output noise variance due the quantization is found to be

$$\sigma_{\text{direct}}^2 = c2^{-2b}\sigma_x^2\sum_l |g(l)|^2. \tag{2.3.16}$$

Under optimal bit allocation, the coding gain of the two-level FB convolver over the direct form is

$$\begin{aligned} G_{x,\text{ two}} &= \frac{\text{output variance}|_{\text{direct conv}}}{\text{output variance}|_{\text{subband conv}}} \\ &= \frac{\sigma_x^2}{\prod_{i=0}^{M-1}\left(\sigma_{x_i}^2\right)^{1/n_i}} \times \frac{\sum_n |g(n)|^2}{\frac{1}{M}\prod_{i=0}^{M-1}\left(n_i\gamma_i^2\right)^{1/n_i}}. \end{aligned} \tag{2.3.17}$$

The "x, two" in the subscript in (2.3.17) indicates that the coding gain is obtained by using the two-level FB convolver and quantizing the signal $x(n)$. This subscript is used to distinguish (2.3.17) from the deterministic coding gain which is obtained by quantizing $g(n)$ in the next section. From the right-hand side of (2.3.17), we see that the variation of subband energy of both $x(n)$ and $g(n)$ contributes to the coding gain. The first term is the gain contributed by $x(n)$ and the second term is the gain contributed by $g(n)$.

If the filters $\{F_k(z)\}$ are orthonormal, then we can prove that the coding gain for the two-level FB convolver is always greater than unity, regardless of the *quality* of the filters $\{H_k(z)\}$, $\{F_k(z)\}$, $\{H_k'(z)\}$

and $\{F_k'(z)\}$. Moreover, we can prove that this coding gain is never smaller than that of the one-level FB convolver derived in Section 3.3 in [Vai93a]:

Lemma 2.3.1. The coding gain $G_{x,\text{two}}$ of the two-level orthonormal FB convolver (i. e. FB in both levels are orthonormal) is never smaller than that of the one-level orthonormal FB convolver, regardless of the choice of PU filters $\{H_k'(z)\}$, provided that $x(n)$, $g(n)$, $\{H_k(z)\}$ in both cases are the same. Moreover, they are equal if and only if the sequence $g_k^{(i)}(n)$ has the same energy for all $0 \le i \le L-1$. ■

In [Vai93a], it was shown that under optimal bit allocation, the coding gain of the one-level FB convolver is

$$G_{x,\text{one}} = \frac{\sigma_x^2}{\prod_{i=0}^{M-1}\left(\sigma_{x_i}^2\right)^{1/n_i}} \times \frac{\sum_n |g(n)|^2}{\frac{1}{M}\prod_{i=0}^{M-1}\left(n_i\alpha_i^2\right)^{1/n_i}}, \tag{2.3.18}$$

where α_k^2 is defined as

$$\alpha_k^2 = \frac{M}{L}\sum_{i=0}^{L-1}\sum_n |g_{k,\text{one}}^{(i)}(n)|^2, \tag{2.3.19}$$

where the "one" in the subscript is used to denote that $g_{k,\text{one}}^{(i)}(n)$ are the subband filters of the one-level FB convolver (see Fig. 2.3.1). Comparing (2.3.18) with (2.3.17), we find that the coding gain formulas for both the one-level and two-level FB convolvers are very similar, except that α_k^2 is replaced by γ_k^2. Therefore in the following proof of Lemma 2.3.1, we need to establish the relation between α_k^2 and γ_k^2.

Proof of Lemma 2.3.1: By defining $\mathbf{h}'(z) = [H_0'(z)\ H_1'(z)\ \ldots\ H_{L-1}'(z)]^T$, and $\mathbf{e}(z) = [1\ z^{-1}\ \ldots\ z^{-(L-1)}]^T$, we have $\mathbf{h}'(z) = \mathbf{E}'(z^L)\,\mathbf{e}(z)$, where $\mathbf{E}'(z)$ is the $L \times L$ polyphase matrix of $\mathbf{h}'(z)$. From the definition of $g_k^{(i)}(n)$ and $g_{k,\text{one}}^{(i)}(n)$, it is clear that $g_k^{(i)}(n)$ can be obtained by passing $g_{k,\text{one}}^{(i)}(n)$ through $\mathbf{E}'(z^{p_k})$ as shown in Fig. 2.3.1. Since $\mathbf{E}'(z)$ is PU, we have [Som93a]

$$\sum_{i=0}^{L-1}\sum_n |g_k^{(i)}(n)|^2 = \sum_{i=0}^{L-1}\sum_n |g_{k,\text{one}}^{(i)}(n)|^2. \tag{2.3.20}$$

By using (2.3.4), (2.3.19) and (2.3.20), we find the following important equality

$$\alpha_k^2 = \frac{1}{L}\sum_{i=0}^{L-1}\alpha_{ki}^2 = \text{arithmetic mean of } \alpha_{ki}^2. \tag{2.3.21}$$

By taking the ratio of $G_{x,\text{two}}$ to $G_{x,\text{one}}$, we find that the ratio of the coding gain of the two-level FB convolver to that of the one-level FB convolver is

$$R_x = \prod_{i=0}^{M-1}\left(\frac{\alpha_k^2}{\gamma_k^2}\right)^{1/n_k}. \tag{2.3.22}$$

Using (2.3.12) and (2.3.21) and applying the AM-GM inequality, each term in the product in (2.3.22) is greater or equal to unity with equality if and only if $\alpha_{ki}^2 = \alpha_k^2$ for all i. So we conclude that $R_x \ge 1$, with equality if and only if $\alpha_{ki}^2 = \alpha_k^2 = \gamma_k^2$ for all i. Or equivalently, the sequences $g_k^{(i)}(n)$ have the same energy for all $0 \le i \le L-1$. ■

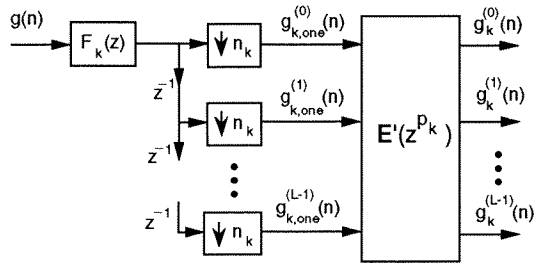

Fig. 2.3.1. Relationship between the subband signals of the one-level and two-level FB convolvers.

Corollary 2.3.1. $G_{x,\text{two}} \geq 1$ for the two-level orthonormal FB convolver, regardless of the choice of the orthonormal sets of filters $\{H_k(z)\}$ and $\{H'_k(z)\}$. Equality holds if and only if both $\sigma^2_{x_k}$ and $n_k\alpha^2_{ki}$ are independent of k and i. ■

2.4. LOW SENSITIVITY STRUCTURE FOR FIR FILTERS

Ignore the quantizers in the subbands of $x(n)$ for the discussion of this section. Very similar to the idea of quantizing $x(n)$, we can quantize the filter coefficients $g_k(n)$ in the subbands based on the input signal variance and maximum amplitude of the subband filter coefficients. However, the coefficients have to be treated as deterministic parameters so that overflow is avoided completely. In this implementation, the convolution error due to the coefficient quantization is much smaller than that in the direct form implementation. Let $\widehat{g}_k^{(i)}(n)$ be the quantized version of $g_k^{(i)}(n)$. Then we can redraw Fig. 2.2.2(a) and (b) as Fig. 2.4.1(a) and (b) respectively. The implementations in Fig. 2.4.1(a) and (b) can be regarded as low sensitivity implementations of the filter $g(n)$. For a preview of the advantage of the implementation, compare Fig. 2.5.2 and Fig. 2.5.3. When the same average number of bits is used to quantize the filter coefficients for direct convolution (Fig. 2.5.2) and subband convolution (Fig. 2.5.3), the improvement shown in these figures is significant. In the following, we will translate this improvement into a mathematical formula.

2.4.1. Low Sensitivity FIR Filter Structures Using the One-Level FB Convolver

With the quantizers inserted in the subbands of $g(n)$ as in Fig. 2.2.2(a), let b_k be the number of bits per sample of $g_k(n)$, allocated to the quantizers Q'_k. Then the average bit rate b is defined as:

$$b = \sum_{k=0}^{M-1} \frac{b_k}{n_k}. \tag{2.4.1}$$

The Noise Model: Define the deterministic quantization error to be

$$q_k^{(i)}(n) \stackrel{\triangle}{=} \widehat{g}_k^{(i)}(n) - g_k^{(i)}(n), \tag{2.4.2}$$

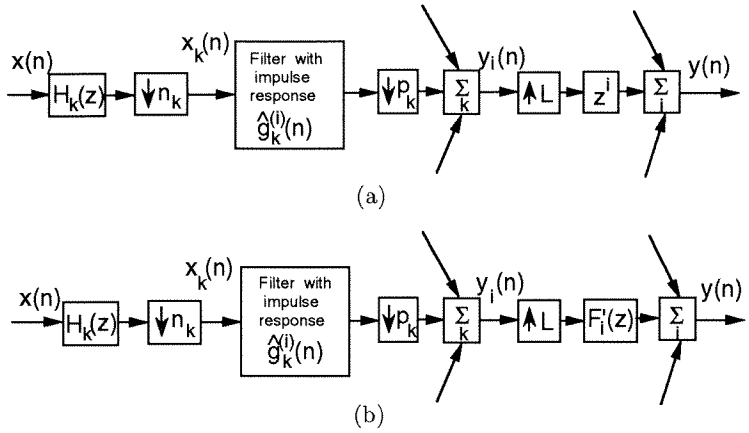

Fig. 2.4.1. Low sensitivity structures for FIR filters: (a) With one-level filter bank convolver; (b) with two-level filter bank convolver.

where $\widehat{g}_k^{(i)}(n)$ is the quantized version of $g_k^{(i)}(n)$. To avoid overflow in the filter coefficients, we assume that the weighting of the most significant bit assigned to the quantizer Q_k' is greater than $g_{k,max}$, where

$$g_{k,max} = \max_{i,n} |g_k^{(i)}(n)|. \tag{2.4.3}$$

Under this condition, the stepsize in the k-th quantizer would be $\triangle_k = c_1\, g_{k,max} 2^{-b_k}$ and the mean square value of the quantization error $q_k^{(i)}(n)$ is

$$\sigma^2_{q_k^{(i)}} = 1/L_{g_k} \sum_{n=0}^{L_{g_k}-1} |q_k^{(i)}(n)|^2 = c_2\, g^2_{k,max} 2^{-2b_k}, \tag{2.4.4}$$

where c_1 and c_2 are constants independent of k and i, L_{g_k} is the length of the subband filter $g_k^{(i)}(n)$. In practice, c_1 and c_2 will depend on $g_k^{(i)}(n)$, but the bit allocation and coding gain is insensitive to the variation of these constants. To carry on the analysis, we assume that they are constant. We further assume that:

1. $x(n)$ is WSS.
2. The deterministic cross-correlation of the quantization error $q_k^{(i)}(n)$, approximately satisfies

$$\frac{1}{L_{g_k}} \sum_{p=0}^{L_{g_k}-1} q_k^{(i)}(p)\, q_j^{(i)}(p+m) \approx \sigma^2_{q_k^{(i)}} \delta(k-j)\delta(m). \tag{2.4.5}$$

   This is of course never exact because $q_k^{(i)}(n)$ is FIR.
3. The length L_g of $g(n)$ is much greater than that of the analysis filters. So $L_{g_k} \approx L_g/n_k$. This is usually the case if the filter bank is of low complexity.

The Optimal Bit Allocation and the Deterministic Coding Gain: Consider Fig. 2.4.1(a). The error of the subband convolution output $y_i(n)$, due to quantization of $g_k^{(i)}(n)$, can be expressed as

$$q_{y_i}(n) = \sum_{k=0}^{M-1} \Big(x_k(n) * q_k^{(i)}(n) \Big)_{\downarrow p_k}. \tag{2.4.6}$$

Using the noise model and carrying out the same procedure in Section 2.3, we find that the optimal number of bits used to quantize the subband filter $g_k^{(i)}(n)$ is

$$b_k = b + 0.5 \log_2 \sigma_{x_k}^2 g_{k,max}^2 - 0.5 \sum_{i=0}^{M-1} \log_2 \left(\sigma_{x_i}^2 g_{i,max}^2\right)^{1/n_i}. \tag{2.4.7}$$

Under this optimal bit allocation, the average output variance is

$$\sigma_{q_y,opt}^2 = c\, L_g 2^{-2b} \prod_{k=0}^{M-1} \left(\sigma_{x_k}^2 g_{k,max}^2\right)^{1/n_k}. \tag{2.4.8}$$

In contrast, suppose we have convolved directly (i. e., without any filter bank). If $g(n)$ is quantized to b bits and without coefficient overflow, then the output error variance is

$$\sigma_{\text{direct}}^2 = c\, L_g 2^{-2b} g_{max}^2 \sigma_x^2, \tag{2.4.9}$$

where $g_{max} = \max_n |g(n)|$. Therefore, from (2.4.8) and (2.4.9), we find that the deterministic coding gain of the one-level FB convolver over the direct form is

$$G_{g,\text{ one}} = \frac{\sigma_x^2}{\prod_{k=0}^{M-1} \left(\sigma_{x_k}^2\right)^{1/n_k}} \times \frac{g_{max}^2}{\prod_{k=0}^{M-1} \left(g_{k,max}^2\right)^{1/n_k}}. \tag{2.4.10}$$

A Lower Bound for the Coding Gain: In the above derivation, orthonormality property of the filter bank is not required, biorthogonality is sufficient. However the deterministic coding gain cannot be proved to be always greater than unity. The likelihood that the deterministic coding gain is less than unity is very low. In fact, in all the examples we encountered in numerical experiments, the coding gain is quite large. However, if the analysis and synthesis filters have unit energy, we can obtain a (very pessimistic) lower bound for the coding gain. From Appendix 2.A, we have:

$$g_{k,max} \leq \sqrt{L_{H_k}}\, g_{max}, \tag{2.4.11}$$

where L_{H_k} is the length of the filter $H_k(z)$. Substituting (2.4.11) into (2.4.10), we find that the coding gain is lower bounded as

$$G_{g,\text{ one}} \geq \frac{\sigma_x^2}{\prod_{k=0}^{M-1} \left(L_{H_k} \sigma_{x_k}^2\right)^{1/n_k}}. \tag{2.4.12}$$

2.4.2. Low Sensitivity FIR Filter Structures Using the Two-Level FB Convolver

We can implement FIR filters using the two-level FB convolver instead of the one-level FB convolver. This will give a lower sensitivity. Or equivalently, we can afford to quantize the subband filters $g_k^{(i)}(n)$ to a much lower bit rate for a fixed accuracy. Again for a preview of the advantage of the two-level FB convolver over the one-level FB convolver, compare Fig. 2.5.3 and Fig. 2.5.4. The equivalent filter responses for both the cases are comparable even though $b = 2$ in the two-level FB convolver and $b = 4$ in one-level FB convolver.

Consider Fig. 2.4.1(b). Let b_{ki} be the number of bits used to quantize the subband filter $g_k^{(i)}(n)$. Then the average bit rate b is defined as in (2.3.1). The noise model assumed here is the same as that in Section 2.4.A except that (2.4.4) is replaced with

$$\sigma^2_{q_k^{(i)}} = 1/L_{g_k} \sum_{n=0}^{L_{g_k}-1} |q_k^{(i)}(n)|^2 = c_2\, g^2_{ki,max} 2^{-2b_{ki}}, \tag{2.4.13}$$

where

$$g_{ki,max} \triangleq \max_n |g_k^{(i)}(n)|. \tag{2.4.14}$$

Optimal Bit Allocation and Deterministic Coding Gain: The error at the location $y_i(n)$ in Fig. 2.4.1(b) can be expressed as (2.4.6). To carry on the analysis, we will assume that the filter bank $\{F_i'(z)\}$ is PU. By using the same technique as in the previous section, we find that the optimal bit used to quantize $g_k^{(i)}(n)$ is

$$b_{ki} = b + 0.5 \log_2 \sigma^2_{x_k} g^2_{ki,max} - 0.5 \sum_{j=0}^{M-1} \log_2 \left(\sigma^2_{x_j} \beta_j^2\right)^{1/n_j}, \tag{2.4.15}$$

where

$$\beta_k^2 = \prod_{i=0}^{L-1} \left(g^2_{ki,max}\right)^{1/L} = \text{geometric mean of } g^2_{ki,max}. \tag{2.4.16}$$

The average output noise variance under optimal bit allocation is

$$\sigma^2_{q_y,opt} = c\, L_g 2^{-2b} \prod_{k=0}^{M-1} (\sigma^2_{x_k} \beta_k^2). \tag{2.4.17}$$

From (2.4.9) and (2.4.17), we find that the deterministic coding gain of the two-level FB convolver over the direct form is

$$G_{g,\text{ two}} = \frac{\sigma_x^2}{\prod_{k=0}^{M-1} \left(\sigma^2_{x_k}\right)^{1/n_k}} \times \frac{g^2_{max}}{\prod_{k=0}^{M-1} \left(\beta_k^2\right)^{1/n_k}}. \tag{2.4.18}$$

By taking the ratio of (2.4.18) to (2.4.10), we find that the ratio of the deterministic coding gain of the two-level FB convolver to that of the one-level FB convolver is

$$R_g = \prod_{k=0}^{M-1} \left(\frac{g^2_{k,max}}{\beta_k^2}\right)^{1/n_k}. \tag{2.4.19}$$

A Lower Bound for Coding Gain: Again we cannot show that the coding gain is always greater than unity. From Appendix 2.A (with $h_k(n)$ replaced with $h_k(n) * h_i'(n)$), we can obtain the following (very pessimistic) lower bound:

$$G_{g,\text{ two}} \geq \frac{\sigma_x^2}{\prod_{k=0}^{M-1} \left((L_{H_k} + L_{H'} - 1)\sigma^2_{x_k}\right)^{1/n_k}} \tag{2.4.20}$$

where $L_{H'}$ is the length of the analysis filter $H_i'(z)$, assumed to be the same for all i.

Comparisons of Results: Comparing the coding gain formulas in all the cases ($G_{x,\text{one}}$, $G_{x,\text{two}}$, $G_{g,\text{one}}$ and $G_{g,\text{two}}$), we find that all of them have the following form:

$$G = \frac{\sigma_x^2}{\prod_{i=0}^{M-1} \left(\sigma^2_{x_i}\right)^{1/n_i}} \times \frac{A^2}{\prod_{i=0}^{M-1} \left(A_i^2\right)^{1/n_i}}. \tag{2.4.21}$$

All of them have a common first factor which is always greater than unity when the filter bank is orthonormal. They differ only in the second factor. All of them can be obtained by substituting A^2 and A_i^2 with the corresponding parameters. The only difference is that unlike in the case of the statistical coding gain in Section 2.3, for the deterministic coding gain we *cannot* prove a result similar to (2.3.21). That is, we cannot prove that $g_{k,max}^2$ is the arithmetic mean of $g_{ki,max}^2$, even if the filter $H_i'(z)$ is PU. So the ratio of the deterministic coding gain, R_g in (2.4.19), cannot be proved to be always greater than one. Nevertheless, in practice, we will find that β_k^2 is usually much smaller than $g_{k,max}^2$ for a frequency selective filter $g(n)$. The reason is that under usual situations the arithmetic mean of $g_{ki,max}^2$ would not differ much from $g_{k,max}^2$. But $g_{ki,max}^2$ may vary considerably with respect to i if $g(n)$ is frequency selective. Thus, we may expect that the coding in (2.4.18) would be much larger than that in (2.4.10) as we will see in the numerical examples.

Coding Gain When Both Input Signal $x(n)$ ***and Filter*** $g(n)$ ***are Quantized:*** When quantizers are inserted in both the subbands of $x(n)$ and $g(n)$, the coding gain is not the product of G_x and G_g. To obtain the coding gain, we apply the optimal bit allocation formulas in (2.3.14) and (2.4.15) respectively to the quantization of $x_k(n)$ and $g_k^{(i)}(n)$, and ignore the second order effect. The coding gain is

$$G = \frac{\{\sigma_{q_y,opt}^2\}_x + \{\sigma_{q_y,opt}^2\}_g}{\{\sigma_{\text{direct}}^2\}_x + \{\sigma_{\text{direct}}^2\}_g}, \tag{2.4.22}$$

where the subscript "x" is used to denote the case when only $x(n)$ is quantized, and "g" is used to denote the case when only $g(n)$ is quantized. We see that the largest error term in (2.4.22) will dominate the coding gain.

2.5. NUMERICAL EXAMPLES

In this section, only $g_k(n)$ are quantized, but not $x_k(n)$. In the presence of quantizers in the subbands of $g(n)$, the LTI system with impulse response $g(n)$ is effectively replaced with a periodically time varying system (LPTV) with period L (see next section for the discussion). To describe the system, we have to characterize all L transfer functions $T_k(z)$ as shown in Fig. 2.5.1. In all the following examples, we therefore show all transfer functions.

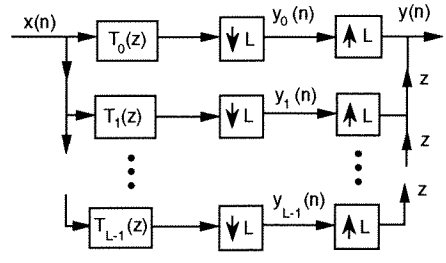

Fig. 2.5.1. Representation of an LPTV system.

In the first four examples, $g(n)$ is an equiripple lowpass filter with $L_g = 132$. The stopband attenuation $\delta_s = -60$ dB and the passpand ripple size $\delta_p = 0.010$. The frequency responses of $g(n)$ with direct quantization to 4 bits and without quantization are shown in Fig. 2.5.2, the stopband attenuation reduces to -17 dB and the passband ripple size increases to 0.049 after quantization.

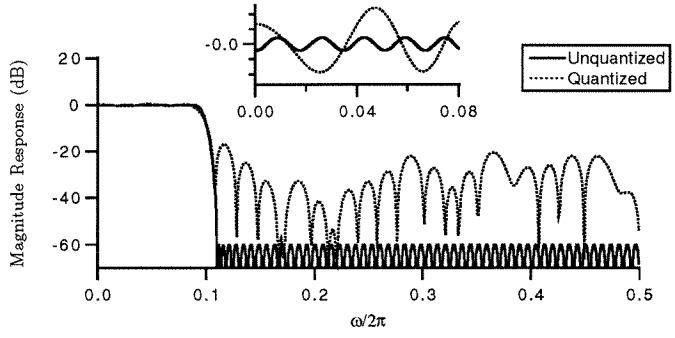

Fig. 2.5.2. Magnitude response (and passband detail) of $g(n)$ without quantization and with direct quantization to 4 bits.

Example 2.5.1. *Four-Channel PU Filter Bank (One-Level FB Convolver):* $L = M = 4$, and $b = 4$ bits. The 4 channel filter bank in Fig. 2.4.1(a) is taken to be a tree-structured PU filter bank obtained by using two-channel PU filter bank in a symmetric tree. The two-channel PU system uses Filter 8A in [Vai88]. If we implement the analysis bank $\{H_k(z)\}$ in lattice form, we need only 8 multiplications per input sample. The corresponding optimal bit allocation is $b_0 = 10$, $b_1 = 5$, $b_2 = 1$, $b_3 = 0$ bits. As shown in Fig. 2.5.3, the stopband attenuations of all the 4 filters $T_i(z)$ are more than 42 dB, i. e., more than 25 dB better than that of the direct quantization. The passband ripple $\delta_p = 0.013$. The effect of quantization on the ripple size is negligible. To visualize the effect of the quantization on the phase response, we show the phase responses of $z^i T_i(z)$ in Fig. 2.5.4. From the plots, we see that the phase distortion in the passband is negligible. ■

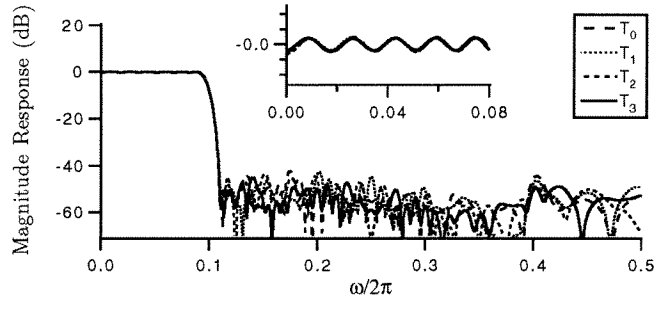

Fig. 2.5.3. Example 2.5.1–Magnitude response (and passband detail) of $g(n)$, with subband quantization to 4 bits by using one-level FB convolver.

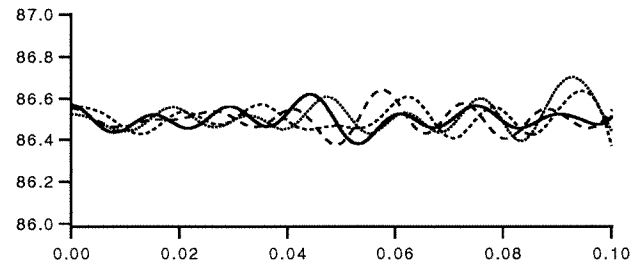

Fig. 2.5.4. Example 2.5.1–Group delay of $g(n)$, with subband quantization to 4 bits by using the one-level FB convolver.

Example 2.5.2. *Four-Channel PU Filter Bank (Two-Level FB Convolver):* $L = M = 4$, and $b = 2$ bits. Both the filter banks formed by $\{H_k(z)\}$ and $\{H_i'(z)\}$ are taken to be the filter bank used in Example 2.5.1. The corresponding bit allocation is shown in Table 2.5.1. As we would expect, b_{ki} are large for $i = 0$ because most of the energy of $G(z)$ is in the first branch. As shown in Fig. 2.5.5, the stopband attenuations (44 dB) are comparable to that obtained in Example 2.5.1 but the average bit rate b is reduced to half. The passband ripple $\delta_p = 0.015$. ■

$k =$	0	1	2	3
$i = 0$	11	7	2	0
$i = 1$	6	4	0	0
$i = 2$	2	0	0	0
$i = 3$	0	0	0	0

Table 2.5.1. Example 2.5.2.–The number of bits b_{ki} allocated to Q_{ki}'

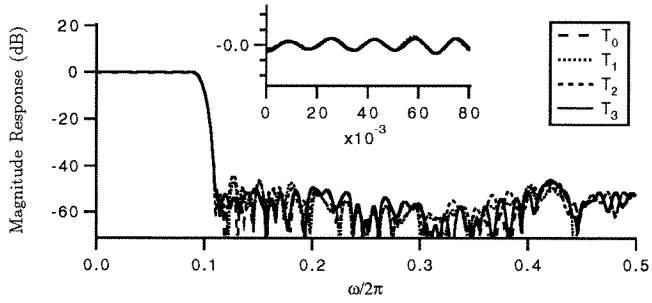

Fig. 2.5.5. Example 2.5.2–Magnitude response (and passband detail) of $g(n)$, with subband quantization to 2 bits by using two-level FB convolver.

Example 2.5.3. *Four- and Eight-Channel DCT Coders (One-Level FB Convolver):* $b = 4$ bits, we use the DCT filter bank, shown in Fig. 2.4.1 in [Vai93a]. In a transform coder filter bank, the polyphase matrix $\mathbf{E}(z)$ of the analysis filters is a constant matrix $\mathbf{T}$. In this example, two cases of $\mathbf{T}$ are considered: (i)

4×4 DCT matrix and (ii) 8×8 DCT matrix as defined in Eq. (12.157) [Jay84]. DCT has the advantage that the analysis filters have linear phase and there exists fast algorithm for the computation of DCT. The corresponding bit allocations are shown in Table 2.5.2. For each case, we show only one transfer function $T_0(z)$ in Fig. 2.5.6 for simplicity. We see that for $M = 4$, the stopband attenuation is 32 dB and $\delta_p = 0.022$. For $M = 8$, the stopband attenuation is 38 dB and $\delta_p = 0.012$. ■

$k =$	0	1	2	3	4	5	6	7
4×4 DCT	7	6	3	0	–	–	–	–
8×8 DCT	9	9	6	3	3	1	1	0

Table 2.5.2. Example 2.5.3–The number of bits b_k allocated to Q'_k

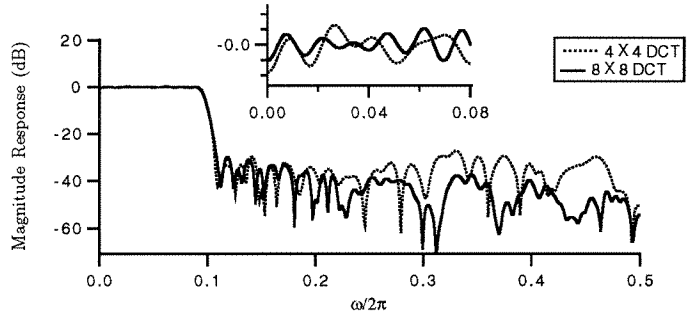

Fig. 2.5.6. Example 2.5.3–Magnitude response (and passband detail) of $g(n)$, with subband quantization to 4 bits by using one-level FB convolver.

Example 2.5.4. *Four- and Eight-Channel DCT Coders (Two-Level FB Convolver):* $b = 2$ bits. The filter bank used here is the same as Example 2.5.3. And $\{H'_k(z)\}$ is identical to $\{H_k(z)\}$. The optimal bit allocation for 4×4 DCT is shown in Table 2.5.3. The corresponding optimal bit allocation for 8×8 DCT is shown in Table 2.5.4. For simplicity, we show only $T_0(z)$ in Fig. 2.5.7. The stopband attenuations for $M = 4$ and $M = 8$ are 27 dB and 30 dB respectively. The passband ripple size increases to 0.035 and 0.026 respectively for $M = 4$ and $M = 8$. ■

$k =$	0	1	2	3
$i = 0$	7	6	3	0
$i = 1$	5	5	2	0
$i = 2$	2	2	0	0
$i = 3$	0	0	0	0

Table 2.5.3. Example 2.5.4(i)–The number of bits b_{ki} allocated to Q'_{ki}

$k=$	0	1	2	3	4	5	6	7
$i=0$	9	7	8	5	5	2	2	0
$i=1$	7	9	7	4	4	2	1	0
$i=2$	7	6	6	3	3	1	0	0
$i=3$	3	4	3	1	0	0	0	0
$i=4$	4	3	3	0	0	0	0	0
$i=5$	2	1	1	0	0	0	0	0
$i=6$	2	1	2	0	0	0	0	0
$i=7$	0	0	0	0	6	5	4	1

Table 2.5.4. Example 2.5.4(ii)–The number of bits b_{ki} allocated to Q'_{ki}

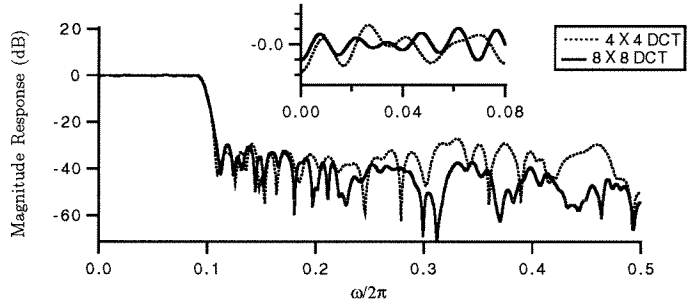

Fig. 2.5.7. Example 2.5.4–Magnitude response (and passband detail) of $g(n)$, with subband quantization to 2 bits by using two-level FB convolver.

Example 2.5.5. *Coding Gain (One-Level FB Convolver):* $M = 4$ and $b = 8$ bits. The filter bank used here is the same as that used in Example 2.5.1. The input signal $x(n)$ is taken to be an AR(5) process with autocorrelation coefficients $R(k)$ obtained from Table 2.2 of [Jay84] (lowpass speech source). The first two rows of Table 2.5.5 show respectively the coding gain obtained from (2.4.10) ($G_{g,\text{one}}$) and that obtained from experiment ($G_{g,\ expt,\text{one}}$) for five different filters $g(n)$ (Filter 1 is the $g(n)$ used in the previous 4 examples). In most cases the theoretical value obtained from (2.4.10) is very close to the experimental result, in spite of the many statistical assumptions used. ■

Example 2.5.6. *Coding Gain (Two-Level FB Convolver):* We take $H'_k(z) = H_k(z)$ and other conditions are the same as those in Example 2.5.5. The coding gain obtained from (2.4.18) ($G_{g,\text{two}}$) and that obtained from experiment ($G_{g\ expt,\text{two}}$) for the same set of five different filters $g(n)$ are shown in the third and fourth rows of Table 2.5.5 respectively. Again the theoretical values are very close to the experiment results. The performance of the two-level FB convolvers is much better (8.7–17.4 dB or equivalently 1.5–3 bits approximately) than that of one-level FB convolvers for all the five cases. The ratios of the coding gain for the two-level FB convolver to that of the one-level FB convolver, R_g (theoretical) and $R_{g,expt}$ (experimental) are shown in the last two rows of Table 2.5.5. ■

Table 2.5.6 summarizes all the results of Examples 2.5.1–4. From the table, we notice that the

Filter No.	1	2	3	4	5
$G_{f,\text{one}}$ (dB)	33.2	19.3	17.6	26.4	36.6
$G_{f,expt\ \text{one}}$ (dB)	33.5	19.8	15.5	20.9	34.5
$G_{f,\text{two}}$ (dB)	47.6	28.0	26.5	38.7	54.3
$G_{f,expt\ \text{two}}$ (dB)	49.5	28.5	25.8	36.7	51.9
R_g (dB)	14.4	8.7	8.9	12.3	17.7
$R_{g,expt}$ (dB)	16.0	8.7	10.3	15.8	17.4

Table 2.5.5. Example 2.5.5 and 2.5.6–Comparison of coding gain.

	b	δ_s (dB)	δ_p
No Quantization	–	−60	0.010
Direct Quantization	4	−17	0.049
4ch PU Bank: 1-level	4	−42	0.013
4ch PU Bank: 2-level	2	−44	0.015
4 × 4 DCT: 1-level	4	−32	0.022
8 × 8 DCT: 1-level	4	−38	0.012
4 × 4 DCT: 2-level	2	−27	0.035
8 × 8 DCT: 2-level	2	−30	0.026

Table 2.5.6. Summary of Examples 2.5.1–4.

performance of the DCT coder is not as good as that of the PU FB convolvers in Examples 2.5.1 and 2.5.2. The reason is that the analysis filters of the DCT coder have a smaller stopband attenuation. The leakage from the adjacent band is quite large. In the last two examples, we see that the deterministic coding gain for the two-level FB convolver is much larger than that of the one-level FB convolver although we cannot prove theoretically that this is always true. By using the two-level FB convolvers, we get a much higher accuracy at the expense of the cost of one filter bank.

2.6. RELATION TO BLOCK FILTER AND ALIASING EFFECT

2.6.1. Convolvers in the View of Block Filter

It is well-known [Bur71, Mit78, Bar80, Vai90] that block filtering is a technique to implement a scalar filter $G(z)$ in such a way as to increase the parallelism. In this section, we will explore the relationship between the FB convolver and the conventional block filtering technique. It was shown in [Hoa89] that the nonuniform system of Fig. 2.1.1 can be expanded as an L-channel uniform system. Therefore we will discuss the uniform case only.

Conventional Block Filtering: Given any scalar filter $G(z)$, we can implement it by using block filtering

technique as shown in Fig. 2.6.1(a). The matrix $\mathbf{G}(z)$ in Fig. 2.6.1(a) is a pseudocirculant matrix:

$$\mathbf{G}(z) = \begin{pmatrix} G_0(z) & G_1(z) & \dots & G_{M-1}(z) \\ z^{-1}G_{M-1}(z) & G_0(z) & \dots & G_{M-2}(z) \\ \vdots & \vdots & \ddots & \vdots \\ z^{-1}G_1(z) & z^{-1}G_2(z) & \dots & G_0(z) \end{pmatrix}, \tag{2.6.1}$$

where $G_i(z)$ is the i-th polyphase component of the scalar filter $G(z)$. In fact, the multirate system in Fig. 2.6.1(a) is a LTI system if and only if $\mathbf{G}(z)$ is a pseudocirculant matrix [Vai88a]. From (2.6.1), we have the following relationship between $[\mathbf{G}(z)]_{ik}$, the elements of the matrix $\mathbf{G}(z)$ and the filter $G(z)$:

$$z^{-i}G(z) = \sum_{k=0}^{M-1} z^{-k}\left[\mathbf{G}(z^M)\right]_{ik}. \tag{2.6.2}$$

Moreover, $\mathbf{G}(z)$ is PU if and only if the filter $G(z)$ is an allpass filter [Vai93]. When $G(z)$ is FIR, this is impossible unless $G(z)$ is a delay.

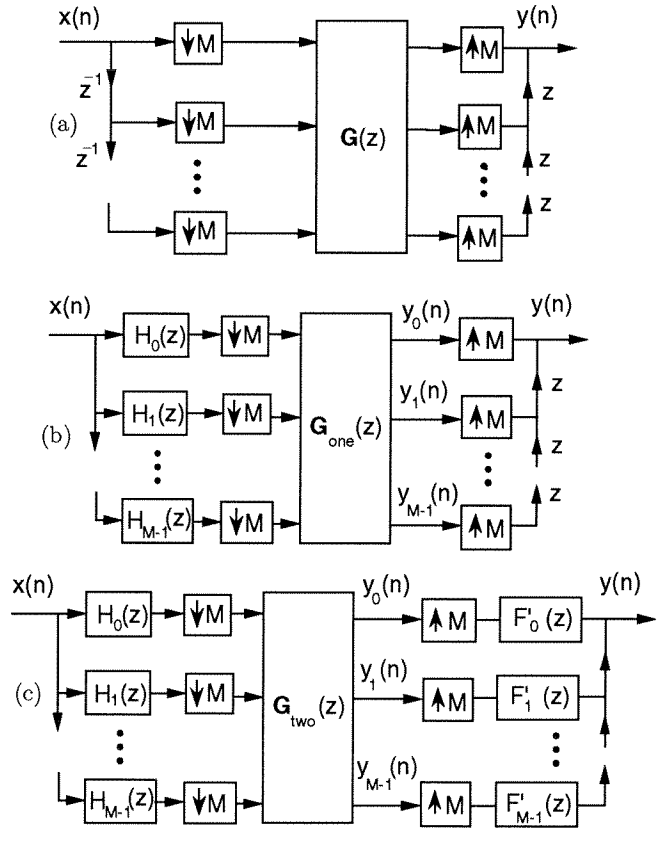

Fig. 2.6.1. Unified view of block filtering and FB convolvers: (a) Conventional block filtering; (b) one-level FB convolver; (c) two-level FB convolver.

Relation of One-Level FB Convolver to Conventional Block Filtering: In the case of one-level FB convolver with uniform decimation ratios, the i-th Type 2 polyphase component of $x(n) * g(n)$ can be written as

(2.2.1) with $L = M$ and all $p_k = 1$. We reproduce the equation here for convenience:

$$Y_i(z) = \left(z^{-i}X(z)G(z)\right)_{\downarrow M} = \sum_{k=0}^{M-1} X_k(z)G_k^{(i)}(z), \quad 0 \le i \le M-1, \tag{2.6.3}$$

where $G_k^{(i)}(z)$ is defined in Fig. 2.1.1(b). By writing (2.6.3) for all i, we obtain:

$$\begin{pmatrix} Y_0(z) \\ Y_1(z) \\ \vdots \\ Y_{M-1}(z) \end{pmatrix} = \mathbf{G}_{\text{one}}(z)\mathbf{x}(z), \tag{2.6.4a}$$

where the column vector

$$\mathbf{x}(z) = \begin{bmatrix} X_0(z) & X_1(z) & \dots & X_{M-1}(z) \end{bmatrix}^T, \tag{2.6.4b}$$

and the matrix $\mathbf{G}_{\text{one}}(z)$ is defined as:

$$\mathbf{G}_{\text{one}}(z) = \begin{pmatrix} G_0^{(0)}(z) & G_1^{(0)}(z) & \dots & G_{M-1}^{(0)}(z) \\ G_0^{(1)}(z) & G_1^{(1)}(z) & \dots & G_{M-1}^{(1)}(z) \\ \vdots & \vdots & \ddots & \vdots \\ G_0^{(M-1)}(z) & G_1^{(M-1)}(z) & \dots & G_{M-1}^{(M-1)}(z) \end{pmatrix}. \tag{2.6.5}$$

Using Type 2 polyphase representation (Section 1.2), $y(n)$ can be written as:

$$Y(z) = \widetilde{\mathbf{e}}(z) \begin{pmatrix} Y_0(z^M) \\ Y_1(z^M) \\ \vdots \\ Y_{M-1}(z^M) \end{pmatrix} = \widetilde{\mathbf{e}}(z)\mathbf{G}_{\text{one}}(z^M)\mathbf{x}(z^M), \tag{2.6.6}$$

where $\widetilde{\mathbf{e}}(z)$ is the row vector $\begin{bmatrix} 1 & z & \dots & z^{M-1} \end{bmatrix}$. From (2.6.6), we immediately get the implementation in Fig. 2.6.1(b), by using the fact that $X_k(z) = \left[X(z)H_k(z)\right]_{\downarrow M}$.

Comparison Between One-Level FB Convolver and Conventional Block Filtering: Comparing Fig. 2.6.1(a) and (b), we discover that the one-level FB convolver is a generalized version of block filtering. Instead of decomposing $x(n)$ and $g(n)$ into their conventional polyphase components, we decompose $x(n)$ and $g(n)$ into their GPP components respectively with respect to two separate sets of polyphase basis, namely $\{H_i(z)\}$ and $\{F_i(z)\}$. The delay chain before the block filter is replaced by a more general analysis bank with filters $\{H_k(z)\}$. Therefore, we can view the convolver as a generalized block filtering technique, which provides not only the advantage of parallelism, but also the advantage of coding gain when implemented in finite precision. Of course, the coding gain is obtained at the expense of the cost of one filter bank. This generalized block filtering technique provides a good tradeoff between the coding gain and the complexity. By using GPP representation, a relationship similar to (2.6.2) between $\left[\mathbf{G}_{\text{one}}(z)\right]_{ik}$ and $G(z)$ can be interpreted nicely as:

$$z^{-i}G(z) = \sum_{k=0}^{M-1} \left[\mathbf{G}_{\text{one}}(z^M)\right]_{ik} H_k(z). \tag{2.6.7}$$

It also can be proved (see Appendix 2.B) that the matrix $\mathbf{G}_{\text{one}}(z)$ is PU if and only if the filter $G(z)$ is an allpass function, provided that the set of filters $\{H_k(z)\}$ is PU.

Relation of Two-Level FB Convolver to Conventional Block Filtering: For two-level FB convolver with uniform decimation ratios, the i-th GPP component of $x(n) * g(n)$ with respect to the polyphase basis $\{F_i'(z)\}$ is:

$$Y_i(z) = \sum_{k=0}^{M-1} X_k(z) G_k^{(i)}(z), \quad 0 \le i \le M-1, \tag{2.6.8}$$

where $G_k^{(i)}(z)$ are defined in Fig. 2.2.1. By writing (2.6.8) for all values of k, we get the equations similar to (2.6.4) and (2.6.5), except that the matrix $\mathbf{G}_{\text{one}}(z)$ is replaced by $\mathbf{G}_{\text{two}}(z)$, where $\left[\mathbf{G}_{\text{two}}(z)\right]_{ki} = G_k^{(i)}(z)$. By defining the row vector $\mathbf{f}'(z) = \left[F_0'(z) \;\; F_1'(z) \;\; \ldots \;\; F_{M-1}'(z)\right]$, the output of the convolution $y(n)$ can be reconstructed from the GPP components $Y_i(z)$ (as defined in (2.6.8)) as:

$$Y(z) = \sum_{k=0}^{M-1} Y_i(z^M) F_k'(z) = \mathbf{f}'(z) \mathbf{G}_{\text{two}}(z^M) \mathbf{x}(z^M), \tag{2.6.9}$$

where the column vector $\mathbf{x}(z)$ is as defined in (2.6.4b). From (2.6.9), we get the implementation of the two-level FB convolver as in Fig. 2.6.1(c).

Comparison of One- and Two-Level FB Convolver in the Light of Block Filtering: Comparing Fig. 2.6.1(b) and (c), clearly the two-level FB convolver is a generalization of one-level FB convolver. In two-level FB convolvers, the "advance chain" in one-level FB convolvers after the block filter is replaced by a more general synthesis bank with $\{F_i'(z)\}$. The relationship between $\left[\mathbf{G}_{\text{two}}(z)\right]_{ik}$ and $G(z)$ can be written as:

$$H_i'(z) G(z) = \sum_{k=0}^{M-1} \left[\mathbf{G}_{\text{two}}(z^M)\right]_{ik} H_k(z). \tag{2.6.10}$$

Similarly, we can prove (Appendix 2.B) that the matrix $\mathbf{G}_{\text{two}}(z)$ is PU if and only if the filter $G(z)$ is an allpass function, provided that the sets of filters $\{H_k(z)\}$ and $\{H_k'(z)\}$ are PU.

2.6.2. Aliasing Effects and the Equivalent LPTV Filter in the Presence of Quantizers

In the presence of quantizers in the subband of $g(n)$, the equivalent system is no longer a LTI system. It is a LPTV system. Let $Q_k^{(i)}(z)$ be the z transform of $q_k^{(i)}(n)$, where $q_k^{(i)}(n)$ is defined in (2.4.2). Define the matrix $\mathbf{Q}(z)$

$$\mathbf{Q}(z) = \begin{pmatrix} Q_0^{(0)}(z) & Q_1^{(0)}(z) & \ldots & Q_{M-1}^{(0)}(z) \\ Q_0^{(1)}(z) & Q_1^{(1)}(z) & \ldots & Q_{M-1}^{(1)}(z) \\ \vdots & \vdots & \ddots & \vdots \\ Q_0^{(M-1)}(z) & Q_1^{(M-1)}(z) & \ldots & Q_{M-1}^{(M-1)}(z) \end{pmatrix}. \tag{2.6.11}$$

Let $\widehat{\mathbf{G}}_{\text{one}}(z)$ be the quantized version of $\mathbf{G}_{\text{one}}(z)$. Then $\widehat{\mathbf{G}}_{\text{one}}(z) = \mathbf{G}_{\text{one}}(z) + \mathbf{Q}(z)$. The system in Fig. 2.6.1(b) can be drawn equivalently as that in Fig. 2.6.2(a). The upper path gives the desired output and the lower path represents the error. By using the polyphase representation, Fig. 2.6.2(a) can be redrawn as Fig. 2.6.1(b) where

$$\mathbf{P}(z) = \mathbf{Q}(z)\, \mathbf{E}(z), \tag{2.6.12}$$

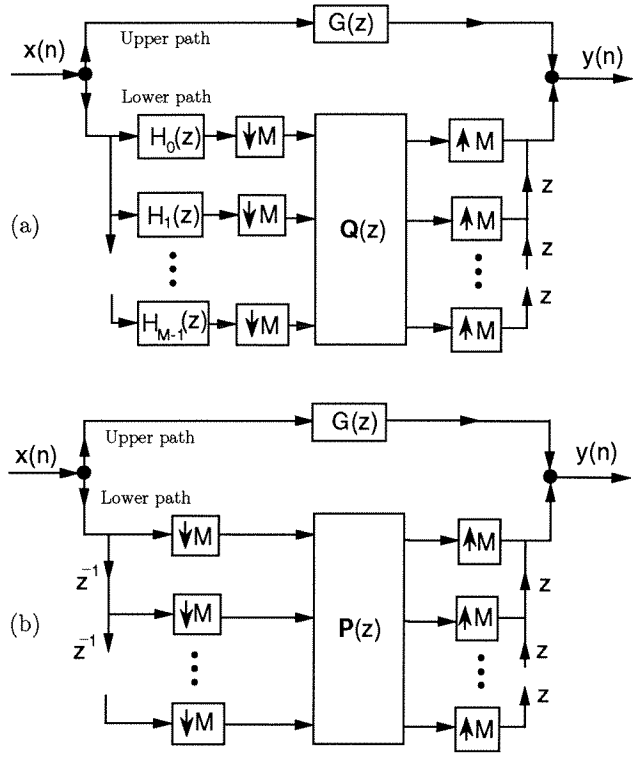

Fig. 2.6.2. (a) Equivalent representation of Fig. 2.6.1(a), and (b) block filter representation of (a).

and $\mathbf{E}(z)$ is the polyphase matrix of the analysis filters $H_k(z)$. From Fig. 2.6.2, we see that the lower path is an LPTV filter and it is an LTI filter if and only if $\mathbf{P}(z)$ is pseudocirculant [Vai93]. For the case of the two-level FB convolver as in Fig. 2.6.1(c), a similar result holds except that the matrix $\mathbf{P}(z)$ is replaced by

$$\mathbf{P}(z) = \mathbf{R}'(z)\,\mathbf{Q}(z)\,\mathbf{E}(z), \tag{2.6.13}$$

where $\mathbf{R}'(z)$ is the Type 2 polyphase matrix of the synthesis filters $F_i'(z)$.

Let $\mathbf{d}(z) = [\, d_0(z) \;\; d_1(z) \;\; \ldots \;\; d_{M-1}(z)\,]^T = \mathbf{P}(z^M)\,\mathbf{e}(z)$, where $\mathbf{e}(z) = [1 \;\; z^{-1} \;\; \ldots \;\; z^{-(M-1)}]^T$ and let $T_i(z) = z^{-i}G(z) + d_i(z)$. Then the system in Fig. 2.6.2 can be redrawn as Fig. 2.5.1. The aliasing components $A_i(z)$ (see Eq. (5.4.7) of [Vai93]) can be expressed as:

$$A_i(z) = \frac{1}{M}\sum_{k=0}^{M-1} z^k d_k(zW^i), \quad \text{for} \;\; 1 \le i \le M-1 \tag{2.6.14}$$

and

$$A_0(z) = G(z) + \frac{1}{M}\sum_{k=0}^{M-1} d_k(z). \tag{2.6.15}$$

$G(z)$ is the desired response, $\frac{1}{M}\sum_{k=0}^{M-1} d_k(z)$ represents the distortion and for $1 \le i \le M-1$, $A_i(z)$ are the aliasing components. The error due to the aliasing and distortion can be written as

$$\varepsilon^2 = \sum_{i=1}^{M-1} \left|A_i(e^{j\omega})\right|^2 + \left|A_0(e^{j\omega}) - G(e^{j\omega})\right|^2$$

$$= \sum_{i=0}^{M-1} \left| \frac{1}{M} \sum_{k=0}^{M-1} e^{jk\omega} d_k(e^{j(\omega - 2\pi i/M)}) \right|^2$$

$$\leq \frac{1}{M^2} \sum_{i=0}^{M-1} \sum_{k=0}^{M-1} |d_k(e^{j(\omega - 2\pi i/M)})|^2 \qquad (2.6.16)$$

The magnitude responses of $d_k(z)$ for Example 2.5.1 are shown in Fig. 2.6.3. All the magnitude responses are under 40 dB even though the coefficients are quantized to $b = 4$ bits only.

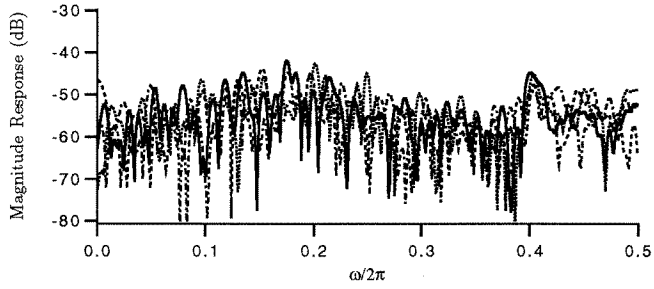

Fig. 2.6.3. Magnitude responses of the aliasing components.

2.6.3. Subband Implementation of LPTV Filters

From the earlier discussion, it is natural to ask if the subband convolver can be modified to implement an LPTV filter. The answer is in the affirmative. The implemenation leads to a low sensitivity structure for LPTV filters.

Given an LPTV filter with period L, we can characterize the filter by a set of L transfer functions $\{T_i(z)\}$ as shown in Fig. 2.5.1. Notice from the figure that the i-th polyphase component $y_i(n)$ of the output of the LPTV filter is completely determined by the transfer function $T_i(z)$. By Theorem 2.2.1, $y_i(n)$ can be obtained as:

$$y_i(n) = \Big(x(n) * t_i(n)\Big)_{\downarrow L} = \sum_{k=0}^{M-1} \Big(x_k(n) * t_{ik}(n)\Big)_{\downarrow p_k}, \qquad (2.6.17)$$

where $x_k(n)$ are defined in Fig. 2.1.1(a), and $t_{ik}(n)$ are the subband signals obtained by replacing $g(n-i)$ in Fig. 2.1.1(b) with $t_i(n)$. The periodically time varying bit allocation can be employed to achieve a low sensitivity structure for LPTV filters.

2.7. FB CONVOLVERS FOR LINEAR-PHASE FILTERS

Suppose that $g(n)$ has linear phase. In the direct form implementation, the symmetry of the impulse response can be exploited to reduce the complexity by one half. Furthermore, the phase remains linear even in the presence of quantization. In the previous discussion, the FB convolvers do not take advantage of the symmetry. In the following, we will see how to preserve the advantages of linear phase and at the

same time achieve high coding gain for FB convolvers. Since the method works for PU filter banks only, we will assume that $\{H_k(z)\}$ and $\{F_k(z)\}$ are PU. Assume that the length of the filter $N = jL$, where L is the lcm of n_k shown in Fig. 2.1.1. We will derive the case where $g(n)$ is symmetric with even length (derivations for other cases are very similar).

$$G(z) = \underbrace{\sum_{n=0}^{N/2-1} g(n)z^{-n}}_{\bar{G}(z)} + \underbrace{\sum_{n=N/2}^{N-1} g(n)z^{-n}}_{z^{-N}\bar{G}(z^{-1})}. \tag{2.8.3}$$

By using the GPP representation, we decompose $X(z)$ and $\bar{G}(z)$ respectively as:

$$X(z) = \sum_{k=0}^{M-1} X_k(z^{n_k})F_k(z) = \sum_{k=0}^{M-1} X'_k(z^{n_k})H_k(z), \tag{2.8.4a}$$

$$\bar{G}(z) = \sum_{k=0}^{M-1} \bar{G}_k(z^{n_k})H_k(z). \tag{2.8.4b}$$

If the filter bank formed by $H_k(z)$ and $F_k(z)$ is PU so that $H_k(z^{-1}) = F_k(z)$, then we can write

$$z^{-N}\bar{G}(z^{-1}) = \sum_{k=0}^{M-1} z^{-N}\bar{G}_k(z^{-n_k})F_k(z). \tag{2.8.5}$$

By using the relations in (2.8.3)–(2.8.5), the decimated output of the convolution $x(n) * g(n)$ can be written as:

$$\begin{aligned}\left[G(z)X(z)\right]_{\downarrow L} &= \left[\sum_{k=0}^{M-1}\sum_{i=0}^{M-1} \bar{G}_k(z^{n_k})X_i(z^{n_i})H_k(z)F_i(z)\right]_{\downarrow L} + \left[\sum_{k=0}^{M-1}\sum_{i=0}^{M-1} z^{-N}\bar{G}_k(z^{-n_k})X'_i(z^{n_i})F_k(z)H_i(z)\right]_{\downarrow L} \\ &= \sum_{k=0}^{M-1}\Big(\bar{G}_k(z)X_k(z)\Big)_{\downarrow p_k} + z^{-j}\sum_{k=0}^{M-1}\Big(\bar{G}_k(z^{-1})X'_k(z)\Big)_{\downarrow p_k},\end{aligned} \tag{2.8.6}$$

where $p_k = L/n_k$. Since $\bar{G}_k(z)$ and $\bar{G}_k(z^{-1})$ are time-reversed versions of each other, their multipliers can be shared. We have sucessfully reduced the complexity to one half and preserved the phase linearity of the overall filter even in the presence of quantizers. However the implementation in (2.8.6) may not give high coding gain because in general the filter $\bar{G}(z)$ is not frequency selective due to the artificial discontinuity introduced by truncation. A technique was proposed in [Pho93a] to solve this problem. Instead of partitioning $G(z)$ in a non overlapping manner as in (2.8.3), if we allow some small overlapping in the partition (which would introduce some computational overhead), then it was shown in [Pho93a] that good coding gain can be achieved by using a raised-cosine function to shape the overlapping region. Therefore there is a tradeoff between the complexity and the coding gain. Experiments [Pho93a] showed that an overlap of less than 10 taps can provide high coding gain.

2.8. OTHER CONSIDERATIONS

2.8.1. IFIR Filter as a Special Case of Subband Convolver

IFIR filters were introduced in [Neu84] to design narrowband filters. In lowpass case, if the stopband edge is smaller than π/M, then $G(z)$ can be approximated by a cascade of two filters as:

$$G(z) \approx G^{(0)}(z^M)I(z), \tag{2.8.1}$$

where $I(z)$ is a low cost filter. The number of coefficients in $G^{(0)}(z)$ is roughly equal to $1/M$ of that in $G(z)$. Fig. 2.8.1(a) shows the implementation of an IFIR filter.

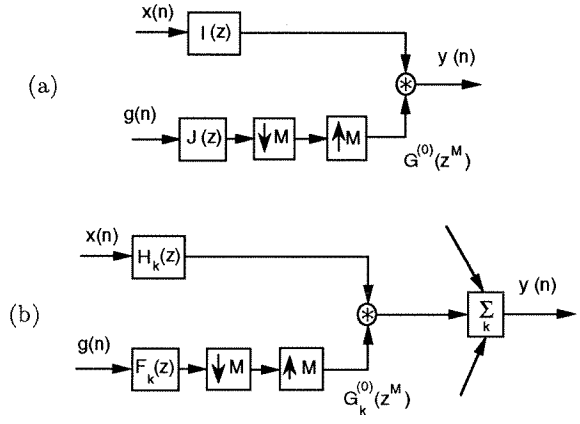

Fig. 2.8.1. Relationship between convolver and IFIR filter: (a) Implementation of IFIR filter; (b) implementation of convolver.

From Fig. 2.1.1(b), $G(z)$ can be decomposed into its GPP components as

$$G(z) = \sum_{k=0}^{M-1} G_k^{(0)}(z^M)H_k(z). \tag{2.8.2}$$

The decomposition is exact. Fig. 2.8.1(b) shows the implementaion. If $G(z)$ has passband smaller than π/M, then only $G_0^{(0)}(z)$ in (2.8.2) has significant energy. By dropping all the other unimportant channels in Fig. 2.8.1(b) corresponding to $G_k^{(0)}(z)$ ($k = 1, 2, \ldots, M-1$), Fig. 2.8.1(b) reduces to Fig. 2.8.1(a) (with $H_0(z)$ and $F_0(z)$ regarded as $I(z)$ and $J(z)$ respectively). Therefore, more generally, if $G(z)$ is a multiband filter, the subband convolver can be used to approximate $G(z)$ by retaining the channels which contain most of the energy.

2.8.2. Low Sensitivity Structures for IIR Filters

The application of the FB convolvers in the implementation of IIR filters is not very useful because of the following reasons:

1. The FB convolver for IIR filters involves the implementation of an LPTV system in the feedback loop, and it is very difficult to ensure the stability in the presence of quantizers.

2. A causal FB convolver will introduce some delay for the output in the feedback loop and this will make the overall system noncausal.
3. IIR filters seldom have an order $N > 10$, so it is not efficient to implement IIR filters with FB convolvers.

2.9. CONCLUSIONS

In this chapter, we have generalized the subband convolution theorem in [Vai93a]. We have derived the coding gain for the generalized convolver, and it was proved that this coding gain is always greater than that of the one-level FB convolver in [Vai93a]. We also unified the subband convolvers, GPP representation, block filtering, LPTV filters, and IFIR filters under one framework. This framework provides us a better understanding of the subband convolvers. As an application of the convolution theorem, a low sensitivity structure for FIR filters is proposed. We have defined the deterministic coding gain of the low sensitivity structure and demonstrated that the coding gain is high. Even when the filter coefficients are quantized to a very low bit rate, we can get filters of small passband ripple and large stopband attenuation.

2.10. APPENDICES

Appendix A. Proof of (2.4.11)

By definition of $g_{k,max}$, we have

$$\begin{aligned} g_{k,max} &= \max_{i,n} |g_k^{(i)}(n)| = \max_{i,n} |\big(g(n-i) * h_k(n)\big)_{\downarrow n_k}| \\ &= \max_{n} |g(n) * h_k(n)| \le g_{max} \sum_{n=0}^{L_{H_k}-1} |h_k(n)| \end{aligned} \tag{2.A.1}$$

The third equality follows from the fact that $\big(g(n-i) * h_k(n)\big)_{\downarrow n_k}$ is one of the polyphase components of $g(n) * h_k(n)$. The last inequality follows directly from triangular inequality. Applying the facts that: (i) $\| \mathbf{v} \|_1 \le \sqrt{N} \| \mathbf{v} \|_2$, where $\| \cdot \|_1$ and $\| \cdot \|_2$ denote 1-norm and 2-norm respectively; (ii) $h_k(n)$ has unit energy (2-norm is unity), (2.4.11) follows immediately.

Appendix B. Proof of Some Facts in Block Filtering

Lemma 2.B.1. *One-Level FB Convolver:* Suppose that the set of filters $\{H_k(z)\}$ is PU. Then the matrix $\mathbf{G}_{\text{one}}(z)$ defined in Fig. 2.6.1(b) is PU if and only if the filter $G(z)$ is an allpass function. ■

Proof: By writing (2.6.7) for all values of $i = 0,\ 1,\ \ldots,\ M-1$, we have the following matrix equation:

$$\begin{pmatrix} G(z) \\ z^{-1}G(z) \\ \vdots \\ z^{-M+1}G(z) \end{pmatrix} = \mathbf{G}_{\text{one}}(z^M) \begin{pmatrix} H_0(z) \\ H_1(z) \\ \vdots \\ H_{M-1}(z) \end{pmatrix} = \mathbf{G}_{\text{one}}(z^M)\mathbf{E}(z^M)\mathbf{e}(z), \tag{2.B.1}$$

where the matrix $\mathbf{E}(z)$ is the polyphase matrix of the filters $\{H_k(z)\}$ and $\mathbf{e}(z) = \begin{bmatrix} 1 & z^{-1} & \dots & z^{-M+1} \end{bmatrix}^T$. Substituting z with zW^{-i} for $i = 0,\ 1,\ \dots,\ M-1$ into the above equation, we get

$$\mathbf{\Lambda}(z)\mathbf{W}\mathbf{\Phi}_G(z) = \mathbf{G}_{\text{one}}(z^M)\mathbf{E}(z^M)\mathbf{\Lambda}(z)\mathbf{W}, \tag{2.B.2}$$

where $\mathbf{\Lambda}(z)$ is the diagonal matrix $diag\begin{bmatrix} 1 & z^{-1} & \dots & z^{-M+1} \end{bmatrix}$, $\mathbf{\Phi}_G(z) = diag\begin{bmatrix} G(z) & G(zW) & \dots & G(zW^{M-1}) \end{bmatrix}$ and $\mathbf{W}$ is the $M \times M$ DFT matrix with $\begin{bmatrix}\mathbf{W}\end{bmatrix}_{ki} = W^{ki}$. Since $\mathbf{\Lambda}(z)$, $\mathbf{W}$ and $\mathbf{E}(z)$ are PU matrices, $\mathbf{G}_{\text{one}}(z)$ is PU if and only if $\mathbf{\Phi}_G(z)$ is. ■

Lemma 2.B.2. *Two-Level FB Convolver:* Suppose that the sets of filters $\{H_k(z)\}$ and $\{H'_k(z)\}$ are PU. Then the matrix $\mathbf{G}_{\text{two}}(z)$ defined in Fig. 2.6.1(c) is PU if and only if the filter $G(z)$ is an allpass function.

■

3

A New Class of Two-Channel Biorthogonal Filter Banks and Wavelet Bases

3.1. INTRODUCTION

Fig. 1.1.3 shows a two-channel maximally decimated filter bank. For the convenience of discussion, we reproduce the two-channel filter bank in Fig. 3.1.1 (where its polyphase form is also shown). A number of PR or nearly PR two-channel systems have been reported before [Cro83, Joh80, Ram84, Smi87, Min85, Vai87b, Ngu89]. In this chapter we develop several new results for two-channel biorthogonal filter banks based on a useful class of polyphase matrices.

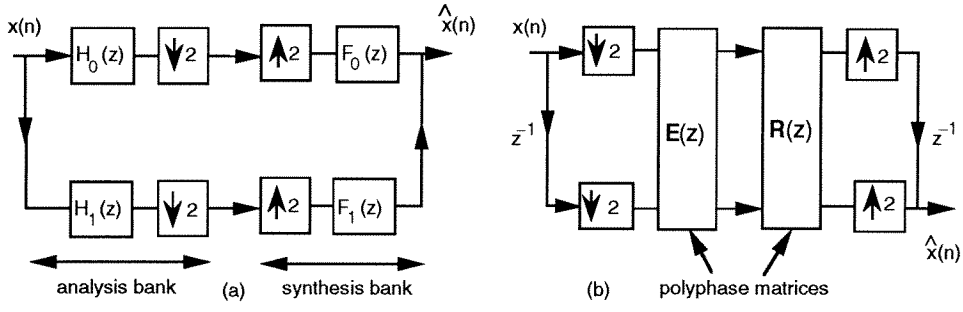

Fig. 3.1.1. (a) Two-channel analysis/synthesis filter bank;
(b) redrawing of (a) by using the polyphase representation.

3.1.1. Previous Work

In FIR filter banks, all the four filters H_0, H_1, F_0, and F_1, are FIR filters while in the case of IIR filter banks, some or all of these filters are IIR filters. The earliest good designs for the IIR case were such that the analysis bank was PU and the polyphase components of $H_0(z)$ and $H_1(z)$ were allpass [Vai87b]. Even though all the IIR filters are causal stable, the reconstructed signal suffers from phase distortion. IIR PR filter banks typically have noncausal stable filters or causal unstable filters [Ram88, Her93, Mit92].

Recently the authors in [Bas95] proposed a IIR PR technique providing causal stable solutions, but no satisfactory design method was given.

In earlier design of 2D filter banks, separable filters have been considered because of their advantage of low complexity. However nonseparable filters offer more freedom in the design and hence in general will give better performance. Recently, some results on the nonseparable filter banks have emerged. In [Lin95a, Lin95b], the authors present some break through works on the theory and design of nonseparable M-channel filter banks. A number of important issues in the design of 2D M-channel filter banks are studied thoroughly. In particular, the authors show that for the class of two-parallelogram filter banks ($M > 2$), it is not possible to obtain PR filter banks with good filters [Lin95a]. The PR conditions for 2D cosine modulated filter banks are derived for the first time. For the class of four-parallelogram filter banks, the authors successfully design PR filter banks with good filters [Lin95b]. In [Vis88], a design method based on space domain approach is given. In [Kar90], a subclass of 2D PU systems (which can be represented as a cascade of 1D PU systems of degree one) is considered. However in both of the polyphase approaches above, the optimization in the designs involves a large number of nonlinear constraints. Thus other approaches, such as 1D to 2D mapping, have been considered [Sha92, Ans87, Che93, Coh93, Tay93]. In [Sha92], even though PR property is preserved by the mapping, the frequency responses of the filters will change. In [Ans87, Che93], a mapping of 1D filter banks to 2D filter banks is given. The authors apply the technique on a 1D two-channel orthogonal IIR system to achieve a 2D IIR filter bank. The resulting systems have either phase distortion or stability problem. In [Coh93], the authors employ McClellan's transformation on the 1D maximally flat FIR halfband filters to obtain a 2D biorthogonal filter bank. However because of the lack of factorization theorems in the 2D case, one of the lowpass filters is constrained to have all its zeros at the aliasing frequency. And there is no simple way to ensure the frequency selectivity of all the filters. In [Tay93], the authors introduce a mapping which can be viewed as the generalization McClellan's transformation. 2D two-channel PR systems with good frequency selectivity can be obtained by judiciously designing the mapping. However, the mapping works for the FIR case only and the resulting filters usually have a large number of coefficients. For an excellent review of various design techniques for two-channel filter banks, the readers are refered to [Lin96].

3.1.2. The New Idea and its Merits

In this chapter, we constrain the polyphase matrix $\mathbf{E}(z)$ such that $\det[\mathbf{E}(z)]$ is a delay. Furthermore we consider $\mathbf{E}(z)$ and $\mathbf{R}(z)$ to be either (i) both causal stable IIR or (ii) both FIR. In each case, the following properties can be simultaneously satisfied.

1. PR is preserved structurally and the structural complexity is very low.
2. All analysis and synthesis filters are designed by controlling a single transfer function $\beta(z)$ [allpass in the IIR case, and Type 2 (i.e., odd order symmetric linear phase FIR) in the FIR case]. So the design procedure is very simple. It is very easy to design $\beta(z)$ so that all filters have good responses (lowpass or highpass as the case may be).

3. In the IIR case, all the analysis and synthesis filters are causal and stable.
4. In some applications such as image coding, the linear phase property of the analysis and synthesis filters is desired. In the FIR case, the filters are exact linear-phase. In the IIR case, we can force the phase response of the filters to be nearly linear in the passband, as we shall explain and demonstrate.
5. The lowpass analysis filter $H_0(z)$ can be forced to have arbitrary number of zeros at $\omega = \pi$. Furthermore the lowpass synthesis filter $F_0(z)$ is guaranteed to have the same number of zeros at π as $H_0(z)$. In both of the IIR and FIR cases, we give closed form expression for the filter coefficients that provide maximum number of zeros at π.

A new class of biorthogonal wavelet bases can be generated from the above filter bank. The regularity property can be directly controlled by imposing multiple zeros at π as desired. In the IIR case, since all filters are causal (in addition to being stable), the basis functions are all causal. In the FIR case, the linear phase property ensures symmetry of the wavelets, while at the same time providing a simple control on regularity (because the number of zeros at π is trivially controlled).

A 1D to 2D Mapping: Furthermore, we also provide a novel mapping of the proposed 1D filter banks into the 2D quincunx case, preserving all the desirable properties. In particular, there is the following:

1. The PR property is preserved.
2. In the IIR case all the analysis and synthesis filters remain causal and stable. In the FIR case the linear phase property is preserved.
3. Even though the filter bank is nonseparable, the complexity is that of a separable filter bank, growing linearly with the filter order.
4. The frequency response supports for the filters are the diamond and diamond-complement as desired for the quincunx case [Ans87, Vai93]. Moreover the filter frequency responses are ensured to be good simply by designing the 1D filter having a good frequency response. Any desired specifications can be met by designing a 1D transfer function $\beta(z)$ appropriately as we shall demonstrate.
5. If the 1D lowpass filter $H_0(z)$ has k zeros at π, then the resulting 2D lowpass filter will have its i-th order total derivative equal to zero at (π, π), for $i = 0, 1, \ldots, k-1$. See Section 3.5 for details.

We also provide a design example to show that the mapping can be easily applied to any dilation matrix (i.e, decimation matrix) with determinant 2.

Relation to Other Results in the Literature: All the designs proposed in this chapter are based on a single class of polyphase matrices, to be described in Section 3.2. However some of the filter banks reported by other researchers are related to our work. In [Kiy92], the authors derive a class of biorthogonal linear phase FIR filter bank which turns out to be a special case of our two-channel framework. In the IIR maximally flat halfband case, our solution is different from the traditional IIR Butterworth design and has approximately linear-phase in the passband. In the FIR maximally flat halfband case, the solution agrees with the classical FIR maximally flat design [Her71]. But our construction is different from those in [Ans87, Dau88] since the analysis filters are factors of maximally flat halfband filters in [Ans87, Dau88]

while our analysis filters are themselves maximally flat halfband. The 2D mapping proposed earlier in [Ans87, Che93] is different from ours because it is known that the earlier mapping will not preserve the PR property in general.

3.1.3. Chapter Outline

Our presentation will go as follows: In the next section, we will derive a framework for the two-channel biorthogonal filter banks. Some properties of such class will be described in detail. In Section 3.3, we will discuss both the IIR and FIR filter banks which are covered in the proposed framework. In Section 3.4, wavelet basis functions generated from the proposed filter banks will be presented and imposition of zeros at aliasing frequency will be considered. Two new classes of IIR maximally flat solution are given in closed form. In Section 3.5, we will first introduce a novel 2D mapping for the quincunx case. Some properties of the mapping are discussed. Then both the IIR and FIR cases are considered. Furthermore numerical examples will be provided throughout the discussion to demonstrate the idea.

Definition. *Halfband Filters:* Consider Fig. 3.1.1. Using the polyphase representation described in Section 1.2, the filters $\{H_k(z), F_k(z)\}$ are related to the polyphase matrices $\mathbf{E}(z)$ and $\mathbf{R}(z)$ as:

$$H_k(z) = E_{k,0}(z^2) + z^{-1}E_{k,1}(z^2), \quad \text{and} \quad F_k(z) = z^{-1}R_{0,k}(z^2) + R_{1,k}(z^2), \tag{3.1.1}$$

where $E_{i,j}(z)$ and $R_{i,j}(z)$ are respectively the ij-th elements of the matrices $\mathbf{E}(z)$ and $\mathbf{R}(z)$. A filter $H_k(z)$ is *halfband* if either one of its polyphase components $E_{k,0}(z)$, $E_{k,1}(z)$ is a delay.

3.2. A FRAMEWORK FOR 1D BIORTHOGONAL FILTER BANKS

Consider Fig. 3.1.1. The system has PR if and only if $\mathbf{R}(z) = \mathbf{E}^{-1}(z)$. It is not easy to constrain $[\det \mathbf{E}(z)]$ to be minimum phase for stability of $\mathbf{R}(z)$; therefore, let us make it a delay. An example is

$$\mathbf{E}(z) = \begin{pmatrix} z^{-N} & \beta(z) \\ 0 & z^{-N'} \end{pmatrix}. \tag{3.2.1}$$

With this we obtain

$$H_0(z) = z^{-2N} + z^{-1}\beta(z^2), \tag{3.2.2}$$

but $H_1(z) = z^{-(2N'+1)}$ which is a delay. Thus even though $H_0(z)$ can be designed to be a good lowpass filter (as we will show), $H_1(z)$ is allpass and this is not useful for subband coding applications. We can modify $H_1(z)$ without affecting $H_0(z)$ by taking the polyphase matrix to be

$$\mathbf{E}(z) = \begin{pmatrix} 0.5 & 0 \\ -0.5\alpha(z) & 1 \end{pmatrix} \begin{pmatrix} z^{-N} & \beta(z) \\ 0 & z^{-N'} \end{pmatrix} = \begin{pmatrix} 0.5z^{-N} & 0.5\beta(z) \\ -0.5z^{-N}\alpha(z) & -0.5\alpha(z)\beta(z) + z^{-N'} \end{pmatrix}. \tag{3.2.3}$$

Then we get the following expressions for the analysis filters:

$$H_0(z) = \frac{(z^{-2N} + z^{-1}\beta(z^2))}{2}, \quad H_1(z) = -\alpha(z^2)H_0(z) + z^{-2N'-1}. \tag{3.2.4}$$

3.2.1. Obtaining Ideal Responses

First notice that the filter $H_0(z)$ can be made an ideal lowpass filter if $\beta(z)$ has the following magnitude and phase responses:

$$|\beta(e^{j2\omega})| = 1, \qquad \forall\ \omega \tag{3.2.5a}$$

$$\angle\beta(e^{j2\omega}) = \begin{cases} (-2N+1)\omega, & \text{for } \omega \in [0, \pi/2]; \\ (-2N+1)\omega \pm \pi, & \text{for } \omega \in (\pi/2, \pi]. \end{cases} \tag{3.2.5b}$$

From (3.2.4), we see that in the high frequency region, $H_1(e^{j\omega})$ has unity gain since $|H_0(e^{j\omega})| = 0$. The function $\alpha(z)$ does not affect $H_0(z)$ and can be freely chosen to shape the response of $H_1(z)$. It should be chosen such that in the low frequency region, $\alpha(z^2)H_0(z)$ cancels with $z^{-2N'-1}$. For exact magnitude cancellation, $|\alpha(e^{j\omega})|$ must be unity. Since $H_0(z)$ is linear phase, it is necessary that $\alpha(z)$ has linear phase in the low frequency region. Comparing these two requirements and the conditions in (3.2.5), we realize that $\beta(z)$ is a suitable candidate for $\alpha(z)$. Indeed, if $N' = 2N-1$, $H_1(z)$ is an ideal highpass filter. In this case, we have an ideal filter bank, and the polyphase matrix $\mathbf{E}(z)$ in Fig. 3.1.1(b) is

$$\mathbf{E}(z) = \begin{pmatrix} 0.5 & 0 \\ -0.5\beta(z) & 1 \end{pmatrix} \begin{pmatrix} z^{-N} & \beta(z) \\ 0 & z^{-2N+1} \end{pmatrix} = \begin{pmatrix} 0.5z^{-N} & 0.5\beta(z) \\ -0.5z^{-N}\beta(z) & -0.5\beta^2(z) + z^{-2N+1} \end{pmatrix}. \tag{3.2.6}$$

With this we get the following expressions for the analysis filters, which we will repeatedly use in this chapter.

$$H_0(z) = \frac{(z^{-2N} + z^{-1}\beta(z^2))}{2}, \quad H_1(z) = -\beta(z^2)H_0(z) + z^{-4N+1}. \tag{3.2.7}$$

PR can be achieved by choosing $\mathbf{R}(z)$ in Fig. 3.1.1(b) to be:

$$\mathbf{R}(z) = \begin{pmatrix} z^{-2N+1} & -\beta(z) \\ 0 & z^{-N} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5\beta(z) & 0.5 \end{pmatrix} = \begin{pmatrix} z^{-2N+1} - 0.5\beta^2(z) & -0.5\beta(z) \\ 0.5z^{-N}\beta(z) & 0.5z^{-N} \end{pmatrix}. \tag{3.2.8}$$

The corresponding synthesis filters can be verified to have the following form:

$$F_0(z) = -H_1(-z), \quad F_1(z) = H_0(-z). \tag{3.2.9}$$

This choice of synthesis filters in (3.2.9) ensures that $\{F_0(z), F_1(z)\}$ will be a lowpass/highpass pair if $\{H_0(z), H_1(z)\}$ is a lowpass/highpass pair. From (3.2.6) and (3.2.8), we have the implementation of the filter bank shown in Fig. 3.2.1. The structure is similar to a ladder network structure [Bru92].

Remark: Of course, the $\alpha(z)$ in (3.2.3) can be taken as functions different from $\beta(z)$, as in the case of [Kim91, Kim91a, Kiy92]. This will provide more freedom in the design. However, by taking them to be the same, the biorthogonal systems can have some additional useful properties. Therefore, we will only consider the case when $\alpha(z) = \beta(z)$.

3.2.2. Two Useful Approximations

The ideal choice of $\beta(z)$ as in (3.2.5) requires infinite complexity. Therefore, we have to design $\beta(z)$ to approximate the conditions in (3.2.5). However the approximation will not change the PR property

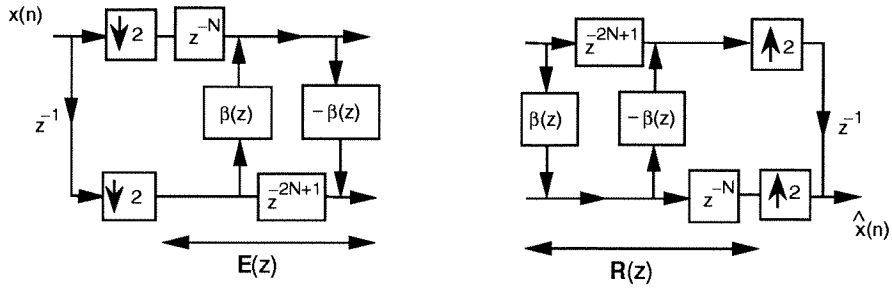

Fig. 3.2.1. Implementation of the proposed biorthogonal filter bank.

because $\mathbf{E}(z)$ in (3.2.6) and $\mathbf{R}(z)$ in (3.2.8) satisfy $\mathbf{R}(z)\mathbf{E}(z) = 0.5z^{-3N+1}\mathbf{I}$, regardless of the choice of $\beta(z)$. Fig. 3.2.1 shows that the frequency responses of all the analysis and synthesis filters depend on one single function $\beta(z)$ only. The frequency selectivity of all four filters depends on how well $\beta(z)$ approximates conditions (3.2.5). This makes the design procedure simple. In the next section, we will provide two simple but useful approximations which correspond to the following two cases:

1. *Stable IIR case:* Here, $\beta(z)$ is chosen to be a causal stable allpass function so that (3.2.5a) is met exactly. We design the phase response of the allpass filter so that (3.2.5b) is approximately satisfied. This leads to a biorthogonal system with *causal stable* IIR analysis and synthesis filters.
2. *Linear phase FIR case:* To satisfy the condition (3.2.5b), $\beta(z)$ can be chosen as a Type 2 linear phase function [Vai93] (filter with a symmetric impulse response of even length). The magnitude response of $\beta(z)$ is optimized to be as close to unity as possible so that (3.2.5a) is well-approximated. This leads to a *linear phase* biorthogonal system.

3.2.3. Additional Properties of the Proposed Filter Banks

In Section 3.1, we have outlined some properties. Properties 1–4 mentioned at the begining of Section 3.1.2 are clear from the above discussion and Property 5 will be discussed later in the Section 3.4. In addition to these five properties, we have:

1. *Double halfband property:* In all previous constructions of two-channel PR filter banks, $H_0(z)F_0(z)$ is a halfband filter, where $H_0(z)$ is not necessary halfband but a factor of a halfband filter. However in our construction above, one can verify that not only the product $H_0(z)F_0(z)$ but also the filter $H_0(z)$ is halfband.
2. *Poles of filters:* In the IIR case, notice from Fig. 3.2.1 that there is no feedback loop in both the analysis and synthesis ends in the ladder network. Therefore the filters have the same poles as those of $\beta(z^2)$ and stability depends solely on the allpass function $\beta(z)$. Moveover in the IIR case if the allpass filter $\beta(z)$ is implemented by using the robust lattice structure [Vai93], the filter bank is stable even when it is realized with finite wordlength.
3. *Robustness to round off noise:* The ladder structure shown in Fig. 3.2.1 is similar to the structure considered in [Bru92]. By using the same reasoning in [Bru92], it can be verified that the round

off noise in the analysis end is compensated by that in the synthesis end. Combining this with the structurally PR property, we conclude that the implementation in Fig. 3.2.1 preserves PR even when all the coefficients are quantized to a finite percision and all the intermediate results are rounded off. However, if the subband signals are quantized (which is usually the case), this property is lost.

4. *Zeros of the filters:* We can verify that $F_0(z)$ and $H_1(z)$ in (3.2.9) and (3.2.7) can respectively be rewritten as:

$$F_0(z) = \big(2z^{-2N+1} - \beta(z^2)\big)H_0(z), \quad H_1(z) = \big(2z^{-2N+1} + \beta(z^2)\big)F_1(z). \tag{3.2.10}$$

These factorizations give the filter bank an interesting structure shown in Fig. 3.2.2. From (3.2.10), it is clear that if $\beta(z)$ is FIR, the zeros of $H_0(z)$ are also zeros of $F_0(z)$. Even when $\beta(z)$ is an irreducible IIR transfer function, this is true since $H_0(z)$ is in the form of (3.2.7) and the zeros of denominator of $\beta(z^2)$ cannot cancel the zeros of $H_0(z)$. Moreover if $|\beta(e^{j\omega})| < 2$, both $F_0(e^{j\omega})$ and $H_0(e^{j\omega})$ have the same set of zeros on the unit circle. The same is true for the pair of $H_1(z)$ and $F_1(z)$. In particular, if $H_0(z)$ has r zeros at $z = -1$, this implies that $F_0(z)$ has no fewer than r zeros at the same point. This property is important in the generation of wavelets since for biorthogonal wavelets, we need both of the analysis and the synthesis wavelets to be regular. By increasing the number of zeros of $H_0(z)$ at $z = -1$, our construction ensures that $F_0(z)$ has at least the same number of zeros at $z = -1$. This is the property which does not appear in the previously existing constructions of biorthogonal filter banks.

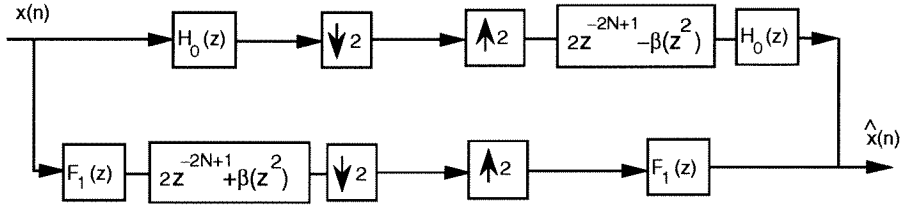

Fig. 3.2.2. Redrawing of Fig. 3.2.1, where $H_0(z) = 0.5(z^{-2N} + z^{-1}\beta(z^2))$, and $F_1(z) = H_0(-z)$.

5. *Ripple sizes of the filters:* Since $H_0(z)$ is a halfband filter and $H_0(z) + F_1(z) = z^{-2N}$, we have the following relationship between the passband ripple δ_p and the stopband ripple δ_s:

$$\delta_p(H_0) = \delta_s(H_0) = \delta_p(F_1) = \delta_s(F_1). \tag{3.2.11}$$

Moreover by using (3.2.10) and the fact that $\beta(z^2) \approx -z^{-2N+1}$ in the high frequency region, we get

$$\delta_s(F_0) \approx 3\delta_s(H_0), \quad \delta_s(H_1) \approx 3\delta_s(F_1), \tag{3.2.12}$$

where $20 \log 3 \approx 9.5$ dB. This property ensures that by designing $H_0(z)$ to have sufficiently high stopband attenuation, we can ensure that all the other three filters will also have good frequency selectivity.

6. *Complexity:* From Fig. 3.2.1, it is very clear that the analysis and synthesis banks have the same complexity. Assume that $\beta(z)$ has order N. For the IIR case, by using the one multiplier lattice structure for allpass function [Vai93], we need approximately $2N$ multiplications, $6N$ additions, and $5N$ delays. For the FIR case, by exploiting the symmetry, we need approximately N multiplications, $2N$ additions and $3.5N$ delays. All the operations are at a lower rate. So the analysis (or synthesis) bank requires N and $0.5N$ multiplications per input sample for the IIR and FIR case respectively.
7. *Near linear phase in the IIR cases:* From (3.2.7), since in the passband the magnitude response of $H_0(z)$ is approximately one, the transfer function $\beta(z^2) \approx z^{-2N+1}$. Therefore $H_0(z)$ has approximately linear phase in the passband. Similar arguement is true for $H_1(z)$.

3.3. DESIGN PROCEDURES FOR THE TWO CLASSES OF PR FILTER BANKS

In this section, we will discuss the two cases of the approximations of (3.2.5) given in the last section. Simple design procedures will be given for both cases.

3.3.1. Causal Stable IIR Biorthogonal Filter Banks

In this section, $\beta(z)$ in (3.2.6)–(3.2.9) is taken to be the causal stable real allpass function:

$$A_{N_1}(z) = \frac{\sum_{k=0}^{N_1} a_{N_1,N_1-k} z^{-k}}{\sum_{k=0}^{N_1} a_{N_1,k} z^{-k}}, \tag{3.3.1}$$

where $a_{N_1,0} = 1$ and $a_{N_1,k}$ are real. In this case, $H_0(z)$ is a sum of a delay and an allpass function. See Eq. (3.2.7). It is an IIR halfband filter and has been studied by some researchers [Ans83, Ren87]. $H_0(z)$ can be made lowpass with large stopband attenuation and small passband ripples by designing the phase response of the allpass function to approximate (3.2.5b) [Ngu94].

Choice of N_1***:*** From the monotone decreasing phase property [Vai93] of a causal stable allpass function, we know that the phase of $A_{N_1}(z^2)$ spans a range of $4N_1\pi$ when ω spans a range of 2π. But from (3.2.5b), $\beta(z^2)$ spans a range of $4N\pi$ or $4(N-1)\pi$. To make the range spanned by both of the functions equal, we set $N_1 = N$ or $N-1$ and this results in two classes of causal stable IIR filter banks. Since the derivation and properties of both of the classes are very similar, in the rest of the chapter, we consider only the case $N_1 = N$ (we will point out at those places where the second class has a different property). With this choice, the analysis filters can be written as

$$H_0(z) = \frac{(z^{-2N} + z^{-1}A_N(z^2))}{2}, \quad H_1(z) = -A_N(z^2)H_0(z) + z^{-4N+1}. \tag{3.3.2}$$

The relationship between the synthesis and analysis filters is the same as (3.2.9).

Additional Properties of the Above IIR Filter Banks:

1. *Preservation of zero at aliasing frequency:* Substituting $z = -1$ into the expression of $H_0(z)$ in (3.3.2), we find that $H_0(z)$ always have a zero at $z = -1$, independent of the coefficients $a_{N,k}$. In

particular, the zero is preserved even when all $a_{N,k}$ are quantized coarsely. This means that one zero at $z = -1$ is structurally imposed. This is important in the generation of wavelet bases since one zero at $z = -1$ is a necessary condition for the existence of the wavelet functions [Dau88]. Note also that $H_1(z)$ will always have a structurally imposed zero at $z = 1$.

2. *Low sensitivity:* Since there exists low sensitivity lattice structure for allpass function [Vai93], the filters have low passband sensitivity. Since the halfband property of $H_0(z)$ is structurally imposed, it has low stopband sensitivity as well.
3. *Bump in the transition band:* Substituting $\omega = \pi/2$ into the expression for $H_1(e^{j\omega})$ and $F_0(e^{j\omega})$ and using the fact that $A_N(-1) = (-1)^N$, we find that $|H_1(e^{j\omega})| = |F_0(e^{j\omega})| = \sqrt{2.5}$ at $\omega = \pi/2$, independent of the allpass function $A_N(z)$. This means that $|H_1(e^{j\omega})|$ and $|F_0(e^{j\omega})|$ always have a bump of approximately 4 dB at $\omega = \pi/2$, no matter how we design $A_N(z)$. The width but not amplitude of the bump can be reduced by increasing the complexity of $A_N(z)$.

Example 3.3.1. *1D Causal Stable IIR Filter Banks:* In this example, $N = 3$. So $A_N(z)$ is a third order allpass function. The filter bank has very low complexity: To implement the analysis (or synthesis) bank, we need only 3 multiplications per input sample! By using the eigenfilter approach for allpass functions [Ngu94] we optimize the coefficients a_k such that maximum attenuation in the stopband of $H_0(z)$ is achieved. The coefficients are obtained as $a_{3,1} = 0.473$, $a_{3,2} = -0.094$, and $a_{3,3} = 0.025$. For the filter $H_0(z)$, the passband edge $\omega_p = 0.4\pi$ and the stopband egde $\omega_s = 0.6\pi$. The stopband attenuation $\delta_s(H_0) = 41.9$ dB. The magnitude responses of the all four filters are shown in Fig. 3.3.1(a). From the plots, relations of ripple sizes in (3.2.11) and (3.2.12) can be verified and it is clear that $H_0(z)$ and $F_0(z)$ have the same set of zeros on the unit circle. The bump of approximately 4 dB around $\pi/2$ is clearly seen. The group delay for $H_0(z)$ and $H_1(z)$ is shown in Fig. 3.3.1(b). The filters are approximately linear phase in the passband and the stopband. ■

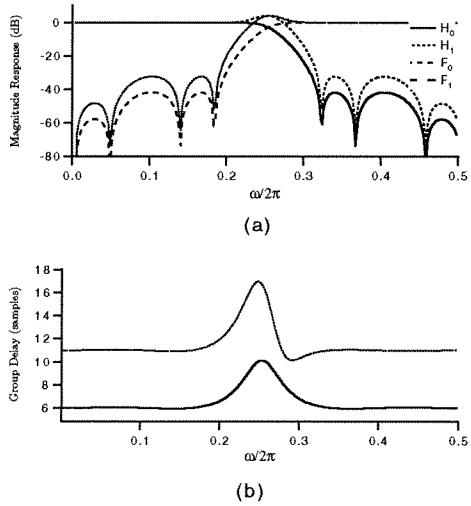

Fig. 3.3.1. Example 3.3.1–Frequency responses of the causal stable IIR filter bank: (a) Magnitude responses of the analysis and synthesis filters; (b) group delays of $H_0(z)$ and $H_1(z)$.

3.3.2. Linear Phase FIR Biorthogonal Filter Banks

In the linear phase FIR case, since $H_0(z)$ is a linear phase halfband filter, it can be designed by employing the trick developed in [Vai87], viz, by taking $\beta(z)$ in (3.2.5)–(3.2.8) to be a Type 2 filter [Vai93] which has a symmetric impulse response of length $2N_1$. In this case, the number of multiplications required to implement $\beta(z)$ is N_1, the same as the N_1-th order allpass function $A_{N_1}(z)$ in (3.3.1). More precisely, let $\beta(z)$ have the following form:

$$V(z) = \sum_{k=1}^{N_1} v_k \times (z^{-N_1+k} + z^{-N_1-k+1}), \tag{3.3.3}$$

where the coefficients v_k satisfy:

$$\sum_{k=1}^{N_1} v_k = 0.5, \tag{3.3.4}$$

so that $V(e^{j0}) = 1$ and $H_0(e^{j0}) = 1$. It is well-known that a Type 2 linear phase filter always has a zero at $z = -1$. In order to satisfy the condition (3.2.5b) exactly, it can be verified that N_1 should be equal to N. By employing the trick in [Vai87], the coefficients v_k can be optimized such that the amplitude response of $V(e^{j\omega})$ is as close to unity as possible. In this case, the analysis filters are:

$$H_0(z) = \frac{(z^{-2N} + z^{-1}V(z^2))}{2}, \quad H_1(z) = -V(z^2)H_0(z) + z^{-4N+1}. \tag{3.3.5}$$

Example 3.3.2. *1D Linear Phase FIR Filter Banks:* $N = 6$. To implement the analysis bank, we need 6 multiplications per input sample, double the number in Example 3.3.1. The Type 2 linear phase function $V(z)$ is designed by using McClellan-Park algorithm. The coefficients are obtained as $v_1 = 0.630$, $v_2 = -0.193$, $v_3 = 0.0972$, $v_4 = -0.0526$, $v_5 = 0.0272$ and $v_6 = -0.0144$. For the filter $H_0(z)$, the passband edge $\omega_p = 0.4\pi$ and the stopband edge $\omega_s = 0.6\pi$, same condition as Example 3.3.1. The stopband attenuation $\delta_s(H_0) = 39.2$ dB and $\delta_s(H_1) = 30$ dB. The magnitude responses of all four filters are shown in Fig. 3.3.2. The relations of ripple sizes in (3.2.11) and (3.2.12) can be verified. ■

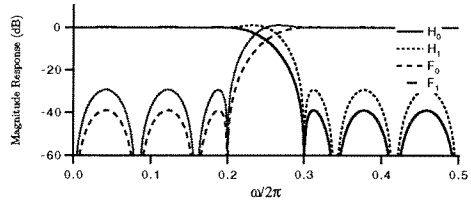

Fig. 3.3.2. Example 3.3.2–Magnitude responses of the linear phase FIR filter bank.

Comparison with Johnston's Filters: For comparison, we will consider Johnston's design [Joh80] with nearly the same specifications. The Johnston's filter 24C in Appendix 7.1 of [Cro83] has $\delta_s = 30$ dB and $\omega_s = 0.586\pi$. For Johnston's filter 32D, $\delta_s = 38$ dB and $\omega_s = 0.586\pi$. To implement the analysis bank,

we need respectively 12 multiplications and 16 multiplications per input sample for the above two cases. Thus as compared to 6 multiplications in our filter bank, the Johnston's design has more complexity than our design. Moreover, there is reconstruction error (0.1 dB for 24C and 0.025 dB for 32D) in Johnston's filter bank.

3.4. IMPOSITION OF MULTIPLE ZEROS AT π

The relation between continuous-time wavelet and discrete-time PR filter bank is well known. A way to construct the scaling and wavelet functions from the filter coefficients was first given by Daubechies in [Dau88]. Starting from the impulse response coefficients $h_0(n)$ and $h_1(n)$, a pair of continuous-time functions $\phi_{H_0}(x)$ and $\psi_{H_1}(x)$ are constructed such that they satisfy:

$$\phi_{H_0}(x) = \sum_{n=0}^{\infty} h_0(n)\phi_{H_0}(2x-n), \tag{3.4.1a}$$

$$\psi_{H_1}(x) = \sum_{n=0}^{\infty} h_1(n)\phi_{H_0}(2x-n). \tag{3.4.1b}$$

Here $\phi_{H_0}(x)$ and $\psi_{H_1}(x)$ are respectively called the analysis scaling and wavelet functions. For the synthesis end, we can write similar expressions for the synthesis scaling and wavelet functions, $\phi_{F_0}(x)$ and $\psi_{F_1}(x)$. The conditions for the existence of such limit functions were given in [Dau88]. It is always desirable to have smooth or "regular" limit functions. It was shown that in order to achieve limit functions of high regularity, we need to have a sufficient number of zeros at the aliasing frequency π. Therefore in the rest of this section, we will show how to impose zeros at π for the proposed filter banks.

3.4.1. Causal Stable IIR Wavelet Bases

For the purpose of achieving regularity, we impose multiple zeros of $H_0(z)$ at π. Since the denominator does not provide any zeros, we consider only the numerator of $H_0(z)$. Except for a delay, the numerator of $H_0(z)$ can be written in terms of $a_{N,k}$ as follows:

$$P_R(\omega) = \sum_{k=0}^{N} a_{N,k}\cos(2k-1/2)\omega. \tag{3.4.2}$$

To obtain r zeros at $z=-1$, we set

$$P_R^{(i)}(\pi) \triangleq \frac{d^{(i)}}{d\omega^{(i)}} P_R(\omega)\Big|_{\omega=\pi} = 0, \quad \text{for } i = 1,\ 2,\ \ldots,\ r-1. \tag{3.4.3}$$

Note that when i is even, $P_R^{(i)}(\pi)$ is always equal to zero. This proves that $P_R(\omega)$ always has an *odd* number of zeros at $\omega=\pi$. Therefore, we can write $r = 2r_0+1$. In this case, we obtain a set of r_0 linear constraints as follows:

$$\sum_{k=0}^{N} a_{N,k}(1-4k)^{2i-1} = 0, \text{ for } i = 1,\ 2,\ \ldots,\ r_0. \tag{3.4.4}$$

The set of linear constraints in (3.4.4) can be satisfied *exactly* in the optimization of the phase response of the allpass function $A_N(z)$ by using the efficient eigenfilter approach [Ngu94], [Che91].

Maximally Flat IIR Wavelets: To obtain a maximally flat solution, i.e., maximum possible number of zeros at π consistent with the constraint that $H_0(z) = 0.5(z^{-2N} + z^{-1}A_N(z^2))$, we set r_0 in (3.4.4) as large as possible. However if $r_0 \geq N+1$, then we can list the first $(N+1)$ linear constraints given by (3.4.4) as follows:

$$\underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0^2 & x_1^2 & \cdots & x_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_0^{2N} & x_1^{2N} & \cdots & x_N^{2N} \end{pmatrix}}_{Vandermonde} \begin{pmatrix} x_0 & & & \\ & x_1 & & \\ & & \ddots & \\ & & & x_N \end{pmatrix} \begin{pmatrix} a_{N,0} \\ a_{N,1} \\ \vdots \\ a_{N,N} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.4.5}$$

where $x_k = 1 - 4k$. Since all the x_k are nonzero and distinct, the two matrices on the left hand side are nonsingular and hence invertible. We get $[a_{N,0}\, a_{N,1} \ \ldots \ a_{N,N}]^T = \mathbf{0}$ which violates the requirement that $a_{N,0} = 1$. This proves under the constraint that $H_0(z) = 0.5[z^{-2N} + z^{-1}A_N(z^2)]$, the filter $H_0(z)$ can have at most $2N+1$ zeros at π. Indeed we can show that the maximally flat IIR filter has *exactly* $2N+1$ zeros at π. To see this, we set $r_0 = N$ and rewrite the set of N linear equations given by (3.4.4) as follows:

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1^2 & x_2^2 & \cdots & x_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{2N-2} & x_2^{2N-2} & \cdots & x_N^{2N-2} \end{pmatrix} \begin{pmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_N \end{pmatrix} \begin{pmatrix} a_{N,1} \\ a_{N,2} \\ \vdots \\ a_{N,N} \end{pmatrix} = -\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{3.4.6}$$

where $x_k = 1 - 4k$ and the fact that $a_{N,0} = 1$ has been imposed. These equations fully determine $A_N(z)$ (hence all the filters) and there is no further parameter to be optimized numerically. As the matrices are invertible, the solution for $a_{N,k}$ always exists and it is *unique*. Furthermore, it is shown in Appendix 3.A that $a_{N,k}$ has the following closed form solution:

$$a_{N,k} = \frac{(-1)^{k-1}}{2k-1}\binom{N}{k}\prod_{i=1}^{N}\frac{(2i-1)}{(2k+2i-1)}, \quad 0 \leq k \leq N, \tag{3.4.7}$$

where $\binom{N}{k} = \frac{N!}{(N-k)!k!}$. The frequency responses of $H_0(z)$ corresponding to $N = 1,\ 2,\ \cdots,\ 10$ are shown in Fig. 3.4.1. Note that although these filters have a numerator of degree $4N-1$ (excluding the trivial delay factor), they have only $2N+1$ zeros at $z = -1$. This implies that some of the zeros are not at $z = -1$ for $N > 1$ and therefore these IIR maximally flat filters are different from the Butterworth halfband filters. Moreover they have nearly linear phase in the passband, as justified at the end of Section 3.2 and demonstrated in Fig. 3.3.1(b). For the case of $N = 1$, one can verify that the solution is a third order Butterworth filter.

Remarks:

1. If the function $\beta(z)$ is taken as $(N-1)$-th order allpass filter (i.e. $N_1 = N-1$), then we will get a second class of causal stable IIR wavelet. In this case, under the constraint that $H_0(z) =$

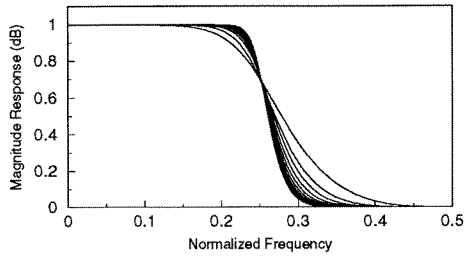

Fig. 3.4.1. Magnitude responses of the IIR maximally flat filters of the form $0.5[z^{-2N} + z^{-1}A_N(z^2)]$ where $A_N(z)$ is a N-th order allpass function, for $N = 1, 2, \ldots, 10$.

$0.5[z^{-2N} + z^{-1}A_{N-1}(z^2)]$, the process of imposition of zeros at π is very similar to the derivation above. The maximally flat IIR filter $H_0(z)$ of this second class will have $2N-1$ zeros at π. The closed form solution for $a_{N-1,k}$ is given as:

$$a_{N-1,k} = \frac{(-1)^{k-1}}{2k+1}\binom{N-1}{k}\prod_{i=1}^{N-1}\frac{(2i+1)}{(2k+2i+1)}, \quad 1 \le k \le N-1, \tag{3.4.8}$$

and $a_{N-1,0} = 1$.

2. Notice that for a PR system, if we interchange the analysis and synthesis filters, the PR property is retained. In many applications such as coding, compression, storage and approximation, the regularity of the synthesis functions is more important [Coh93]. Thus we can choose the wavelet with higher regularity among $\psi_{H_0}(x)$ and $\psi_{F_0}(x)$ as the synthesis wavelet.
3. As the proposed IIR wavelets are generated from rational transfer functions, there is an efficient recursive way to compute the limit functions [Pho95d].

Example 3.4.1. *Causal IIR Wavelets:* We generate the limit functions, ϕ_{H_0}, ψ_{H_1}, ϕ_{F_0}, and ψ_{F_1}, corresponding to the filter bank in Fig. 3.1.1(a). To generate the analysis/synthesis scaling and wavelet functions, we use the cascade algorithm in [Dau88] for eight iterations. We consider the following two cases:

(i) No linear constraint is set, $H_0(z)$ has only one zero at π. The analysis and synthesis filters are the same as those in Example 3.3.1. For the analysis bank, the scaling and wavelet functions, ϕ_{H_0} and ψ_{H_1}, are respectively shown in Fig. 3.4.2(a) and (b). The scaling and wavelet functions corresponding to the synthesis bank, ϕ_{F_0} and ψ_{F_1}, are shown in Fig. 3.4.2(c) and (d).

(ii) As a comparison, we also generate the scaling and wavelet functions corresponding to the IIR maximally flat filters (ϕ_{max} and ψ_{max}) for $N = 3$. In this case, the filter $H_0(z)$ has seven zeros at π. The limit functions are shown in Fig. 3.4.3. For a better comparison on smoothness, in Fig. 3.4.4 we show a zoom-in for Fig. 3.4.2(a) and Fig. 3.4.3(a). We see that the limit functions in Fig. 3.4.3 are more regular than the functions shown in Fig. 3.4.2. ■

3.4.2. Linear Phase FIR Wavelet Bases

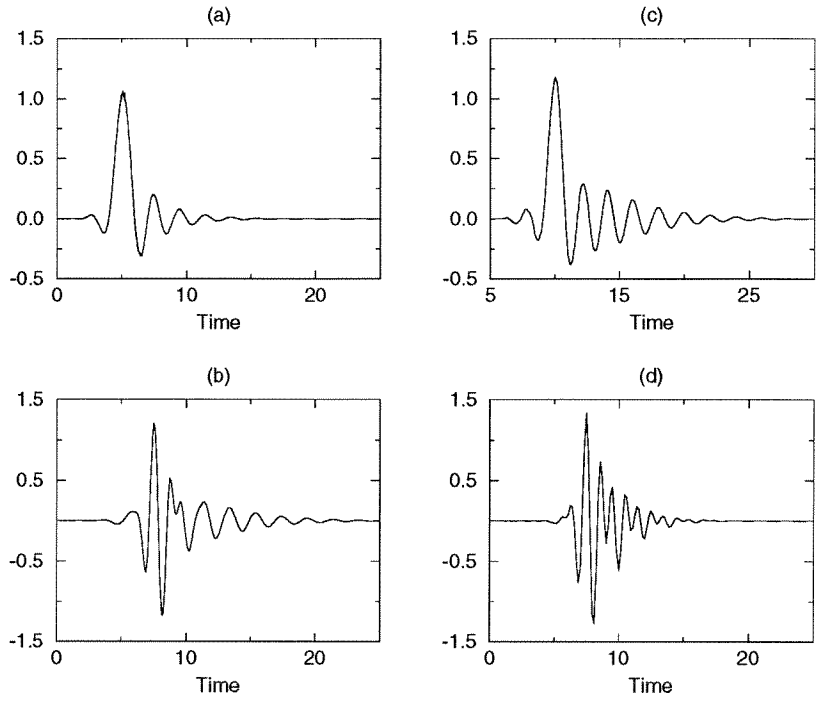

Fig. 3.4.2. Example 3.4.1(i)–Limit functions generated by using the IIR filter bank in Example 3.3.1 ($H_0(z)$ has one zero at π): (a) Analysis scaling function; (b) analysis wavelet function; (c) synthesis scaling function; (d) synthesis wavelet function.

To impose multiple zeros at π for the linear phase FIR case, the procedure is very similar to that given above. Another set of linear constraints can be obtained and incorporated in the procedure of optimization. It can be verified that for this case, $H_0(z)$ always has an *even* number of zeros at π.

Maximally Flat Linear Phase FIR Wavelets: The FIR maximally flat filters have been studied by a number of researchers [Her71, Gum78, Dau88, Ans87]. In [Dau88, Ans87], a maximally flat halfband FIR filter is used to construct compactly supported maximally flat wavelets. In our linear phase FIR filter bank, if all the freedom is used to impose zeros at π, we will arrive at the same solution as that in [Dau88, Ans87]. The closed form solution for FIR maximally flat halfband filters was in [Gum78, Ans87] as:

$$v_k = \frac{(-1)^{N+k-1}\prod_{i=0}^{2N}(N+1/2-i)}{2(N-k)!(N-1+k)!(2k-1)}. \tag{3.4.9}$$

Differences Between Our Construction and Those in [Ans87, Dau88]: In [Ans87], $H_0(z)$ is taken to be a factor of a maximally flat halfband filter. In [Dau88], power spectral factorization is considered. However, in our linear phase structure, $H_0(z)$ is taken to be this halfband filter itself, and not a factor. Since the $H_0(z)$ constructed in [Dau88] is a power spectral factor of the $H_0(z)$ in our structure, our linear phase scaling function $\phi_{LP}(x)$ is related to that constructed by Daubechies in [Dau88], $\phi_D(x)$ as:

$$\phi_{LP}(x) = \phi_D(x) * \phi_D^*(-x), \tag{3.4.10}$$

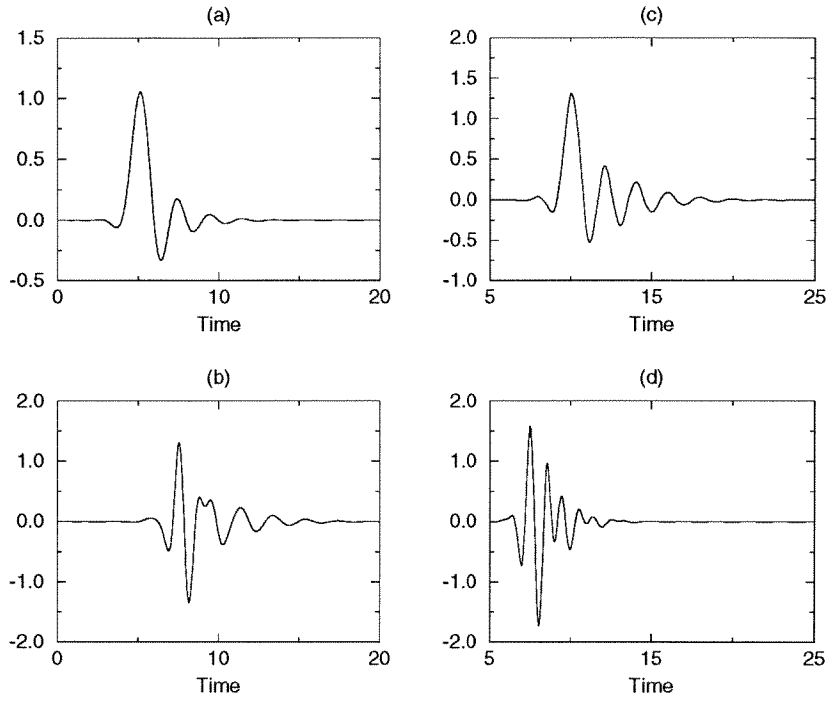

Fig. 3.4.3. Example 3.4.1(ii)–Limit functions generated by using the IIR maximally flat filter bank ($H_0(z)$ has 7 zeros at π): (a) Analysis scaling function; (b) analysis wavelet function; (c) synthesis scaling function; (d) synthesis wavelet function.

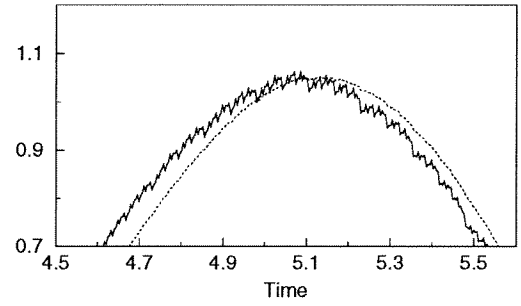

Fig. 3.4.4. Example 3.4.1–Zoom-in for Fig. 3.4.2(a) (solid line) and Fig. 3.4.3(a) (dotted line) demonstrating the improved "regularity" obtained by imposing zeros at π.

where $*$ denotes convolution and ϕ_D^* denotes the complex conjugate of ϕ_D. From (3.4.10), it is clear that the regularity of $\phi_{LP}(x)$ is twice that of $\phi_D(x)$. However the order (and the number of zeros at π) of $H_0(z)$ in our construction is twice that of $H_0(z)$ in the construction in [Dau88]. Comparing the complexity, both of the constructions have approximately the same number of multiplications (because in our construction, linear phase property can be exploited).

Example 3.4.2. *FIR Symmetric Wavelets:* In this example, we construct the limit functions corresponding to the filter bank in Fig. 3.1.1(a) for the linear phase FIR case. The cascade algorithm is used for eight

iterations. We consider two cases:

(i) First, $H_0(z)$ is designed such that no linear constraint other than (3.3.4) is satisfied, therefore it has two zeros at π. The analysis and synthesis filters are the same as those in Example 3.3.2. The limit functions (ϕ_{H_0}, ψ_{H_1}, ϕ_{F_0} and ψ_{F_1}) are respectively shown in Fig. 3.4.5(a), (b), (c) and (d).

(ii) For a comparison, we show the limit functions of the maximum flat case (ϕ_{max} and ψ_{max}) for $N = 6$. In this case, $H_0(z)$ has twelve zeros at π. The plots are shown in Fig. 3.4.6. It can be verified that the limit functions in Fig. 3.4.6 are smoother than those in Fig. 3.4.5. ■

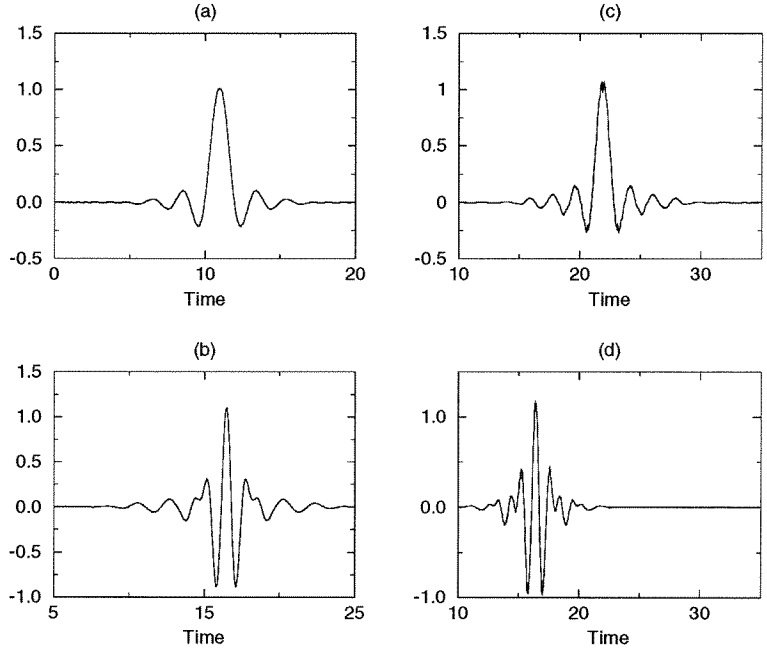

Fig. 3.4.5. Example 3.4.2(i)–Symmetric limit functions generated by using the FIR filter bank in Example 3.3.2 ($H_0(z)$ has 2 zeros at π): (a) Analysis scaling function; (b) analysis wavelet function; (c) synthesis scaling function; (d) synthesis wavelet function.

3.5. MAPPING INTO 2D QUINCUNX PR FILTER BANKS

In this section, we will generalize the 1D framework discussed in Section 3.2 to the 2D case. We will focus on the quincunx subsampling case which has the subsampling lattice shown in Fig. 3.5.1. Notice that the dilation matrix has determinant 2. The corresponding maximally decimated filter bank has only two channels. Furthermore it represents the simplest nonseparable subsampling lattice.

In the 2D case, we know that the desired passband supports of the filters depend not only on the lattice but also on the choice of dilation matrix $\mathbf{M}$ [Vis88]. In the rest of this section, we will consider

$$\mathbf{M} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{3.5.1}$$

The coset vectors are respectively:

$$\mathbf{k}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{k}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{3.5.2}$$

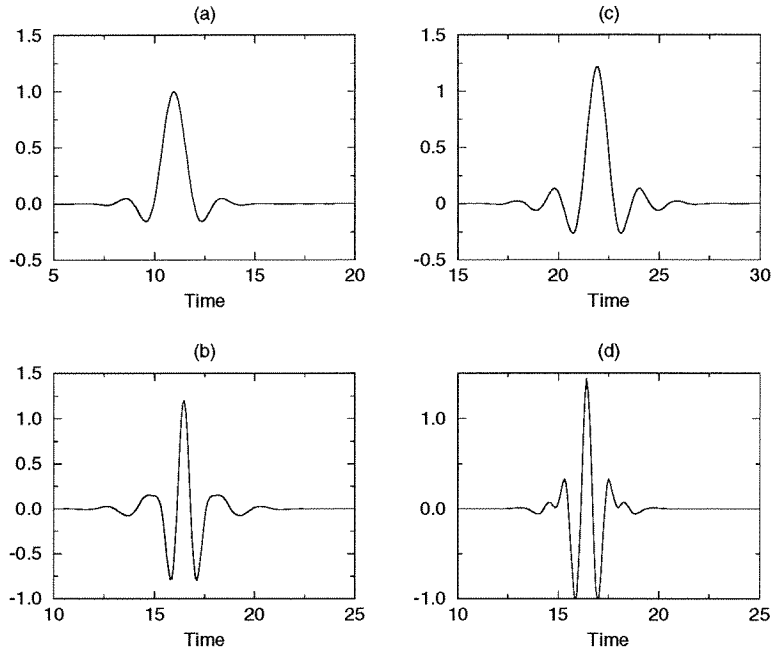

Fig. 3.4.6. Example 3.4.2(ii)–Symmetric limit functions generated by using the FIR maximally flat filter bank ($H_0(z)$ has 12 zeros at π): (a) Analysis scaling function; (b) analysis wavelet function; (c) synthesis scaling function; (d) synthesis wavelet function.

With this $\mathbf{M}$, the ideal supports for alias free decimation, $SPD(\pi\mathbf{M}^{-T})$ [Vai93] is shown in Fig. 3.5.2, where the diamond and diamond-complement, Ω_0 and Ω_1, correspond to the low frequency and high frequency regions respectively. One can verify that $\mathbf{M}$ defined in (3.5.1) has its eigenvalues λ_i equal to $\pm\sqrt{2}$ and $\mathbf{M}^2 = 2\mathbf{I}$. It has a dilation in both the directions. Therefore, $\mathbf{M}$ satisfies the conditions for a *well-behaved* matrix defined in [Kov92]. Given the dilation matrix $\mathbf{M}$ as in (3.5.1) and the coset vectors in (3.5.2), the simple delay chain system and the noble identities are shown in Fig. 3.5.3(a) and (b) respectively. Although the discussion in this chapter is mainly on the quincunx subsampling case with the dilation matrix $\mathbf{M}$ and the coset vectors $\mathbf{k}_i$ defined above, we will provide a design example in the last section to show that the method discussed in this section can be easily generalized to any 2D system with decimation matrix $\mathbf{M}$ having $[\det \mathbf{M}] = 2$.

3.5.1. A 1D to 2D Mapping

In this subsection, we will first give a 2D mapping and then apply the mapping to the framework developed in Section 3.2. Given any 1D biorthogonal systems with the polyphase matrices of the form in (3.2.6) and (3.2.8), we will use the following transformation on the polyphase components:

1. First replace the 1D transfer function $\beta(z)$ with the separable 2D transfer function $\beta(z_0)\beta(z_1)$.
2. Replace all the remaining 1D delay z^{-1} with the 2D delay $z_0^{-1}z_1^{-1}$.

This results in nonseparable analysis and synthesis filters as we will see. Under this transformation,

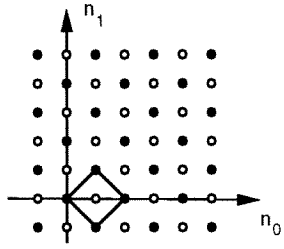

Fig. 3.5.1. Quincunx subsampling lattice.

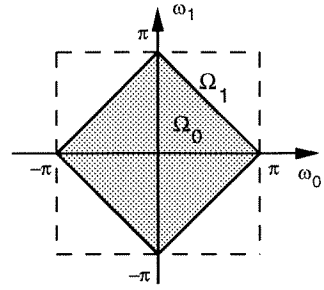

Fig. 3.5.2. Ideal supports for alias-free decimation in quincunx case.

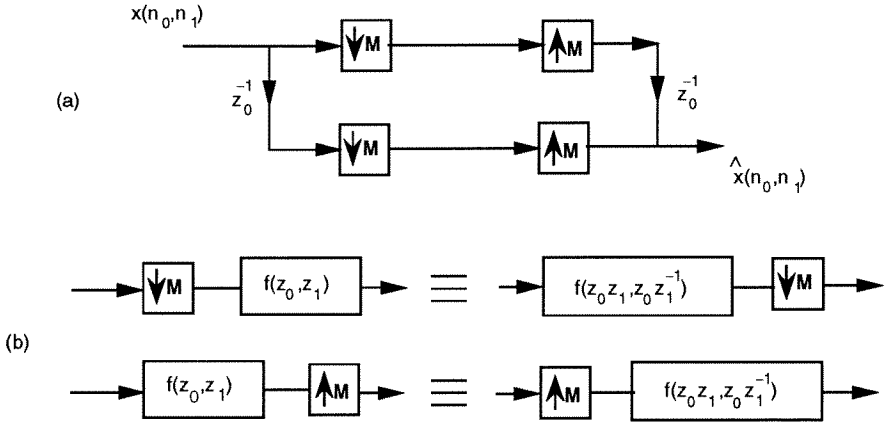

Fig. 3.5.3. Some details for the quincunx decimator: (a) Delay chain; (b) noble identities.

the polyphase matrices $\mathbf{E}^{2D}(z_0, z_1)$ and $\mathbf{R}^{2D}(z_0, z_1)$ of the 2D system can be written respectively as:

$$\begin{aligned}\mathbf{E}^{2D}(z_0,z_1) &= \begin{pmatrix} 0.5 & 0 \\ -0.5\beta(z_0)\beta(z_1) & 1 \end{pmatrix} \begin{pmatrix} (z_0z_1)^{-N} & \beta(z_0)\beta(z_1) \\ 0 & (z_0z_1)^{-2N+1} \end{pmatrix} \\ &= \begin{pmatrix} 0.5(z_0z_1)^{-N} & 0.5\beta(z_0)\beta(z_1) \\ -0.5(z_0z_1)^{-N}\beta(z_0)\beta(z_1) & -0.5\beta^2(z_0)\beta^2(z_1) + (z_0z_1)^{-2N+1} \end{pmatrix}. \end{aligned} \tag{3.5.3}$$

$$\begin{aligned}\mathbf{R}^{2D}(z_0,z_1) &= \begin{pmatrix} (z_0z_1)^{-2N+1} & -\beta(z_0)\beta(z_1) \\ 0 & (z_0z_1)^{-N} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5\beta(z_0)\beta(z_1) & 0.5 \end{pmatrix} \\ &= \begin{pmatrix} (z_0z_1)^{-2N+1} - 0.5\beta^2(z_0)\beta^2(z_1) & -0.5\beta(z_0)\beta(z_1) \\ 0.5(z_0z_1)^{-N}\beta(z_0)\beta(z_1) & 0.5(z_0z_1)^{-N} \end{pmatrix}. \end{aligned} \tag{3.5.4}$$

From the above two equations, we have the implementation of the 2D PR filter bank as Fig. 3.5.4. By using the noble identities in Fig. 3.5.3, we can write the analysis and synthesis filters as:

$$\begin{aligned} H_0(z_0,z_1) &= \frac{z_0^{-2N} + z_0^{-1}\beta(z_0z_1^{-1})\beta(z_0z_1)}{2}, \\ H_1(z_0,z_1) &= -\beta(z_0z_1^{-1})\beta(z_0z_1)H_0(z_0,z_1) + z_0^{-4N+1}, \end{aligned} \tag{3.5.5a}$$

$$F_0(z_0, z_1) = -H_1(-z_0, -z_1), \qquad F_1(z_0, z_1) = H_0(-z_0, -z_1). \tag{3.5.5b}$$

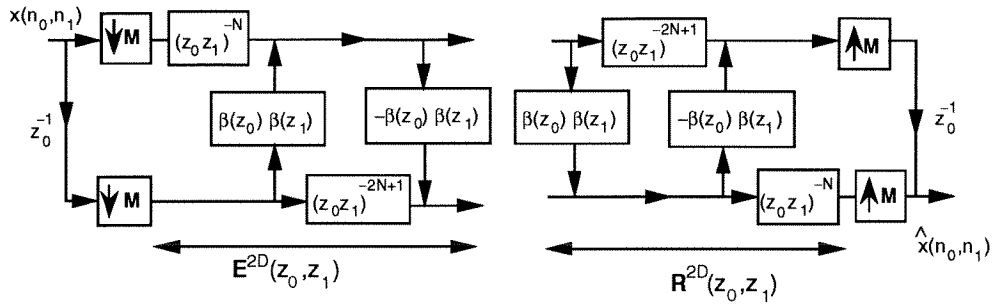

Fig. 3.5.4. 2D biorthogonal filter bank obtained from Fig. 3.2.1 by mapping.

Comparison of our Transformation with Those in [Ans87, Che93, Coh93, Tay93]: McClellan's transformation is used in [Coh93, tay] to obtain a FIR maximally flat halfband filter. The transformation proposed in this chapter differs from McClellan's transformation in the sense that the former operates on the polyphase components while the latter operates directly on the filter. In [Ans87, Che93], the authors obtain a 2D filter bank from 1D by employing the following transformation:

$$\mathbf{E}_{i,j}^{2D}(z_0, z_1) = \mathbf{E}_{i,j}(z_0)\mathbf{E}_{i,j}(z_1), \tag{3.5.6}$$

where $\mathbf{E}_{i,j}$ is the ij-th element of $\mathbf{E}$. We see that in our transformation $\mathbf{E}_{1,1}^{2D}(z_0, z_1) \neq \mathbf{E}_{1,1}(z_0)\mathbf{E}_{1,1}(z_1)$. Therefore our mapping is different from that in (3.5.6).

3.5.2. Properties of the Proposed 2D Filter Banks

Properties 1–5 in Section 3.2 continue to hold after minor modifications to suit the 2D context. In addition, the 2D filter bank satisfies the following properties:

1. *Double halfband property:* It is easy to see that $H_0(z_0, z_1)$ satisfies $H_0(z_0, z_1) + H_0(-z_0, -z_1) = z_0^{-2N}$ and $H_0(z_0, z_1)F_0(z_0, z_1)$ satisfies a similar property. This is the extention of the 1D double halfband property in the 2D quincunx case.
2. *Stability of the 2D analysis and synthesis filters:* If the 1D transfer function $\beta(z)$ is causal then so are the functions $\beta(z_0)\beta(z_1)$ in Fig. 3.5.4. That is $\beta(z_0)\beta(z_1)$ is a first-quadrant filter (the impulse response is zero unless $n_0 \geq 0$ and $n_1 \geq 0$). If $\beta(z)$ is BIBO stable, then so is $\beta(z_0)\beta(z_1)$ so that the polyphase matrix in Fig. 3.5.4 is also BIBO stable. Since the analysis filters are obtained from this stable structure, these filters are guaranteed to be BIBO stable. However we see that the term $\beta(z_0 z_1^{-1})$ has entered the expressions for the analysis filters because of the noble identities, see Fig. 3.5.3(b). It can be shown that this violates the condition for the so-called first-quadrant stability (pp. 166 of [Bos82]). This is explained by the fact that the analysis filters are *not* first-quadrant filters, even though BIBO stable. This is consistent with the observation that the quincunx

decimator $\mathbf{M}$ in (3.5.1) has the negative entry -1. Indeed, the expression $y(\mathbf{n}) = x(\mathbf{Mn})$ means $y(n_0, n_1) = x(n_0 + n_1, n_0 - n_1)$ so that there is a time-reversal operation buried in the decimation process. The same remarks apply for the synthesis filters, that is the 2D synthesis filters $F_0(z_0, z_1)$ and $F_1(z_0, z_1)$ are BIBO stable even though they are not first-quadrant filters.

3. PR is preserved.
4. If the 1D lowpass filter $H_0(z)$ has k zeros at π, then the frequency response of $H_0(e^{j\omega_0}, e^{j\omega_1})$ can be written as

$$H_0(e^{j\omega_0}, e^{j\omega_1}) = \left(1 + e^{-j\frac{\omega_0+\omega_1}{2}}\right)^k P_1(\omega_0, \omega_1) - \left(1 + e^{-j\frac{\omega_0-\omega_1+2\pi}{2}}\right)^k P_2(\omega_0, \omega_1), \tag{3.5.7}$$

where $|P_1(\pi, \pi)|$ and $|P_2(\pi, \pi)|$ are finite quantities. The proof of (3.5.7) is given in Appendix 3.B. Notice that both of the factors $[1 + e^{-0.5j(\omega_0+\omega_1)}]$ and $[1 + e^{-0.5j(\omega_0-\omega_1+2\pi)}]$ are zero at (π, π). Furthermore one can verify that all the mixed partial derivatives satisfy

$$\frac{\partial^{i+l}}{\partial\omega_0^i \partial\omega_1^l} H_0(e^{j\omega_0}, e^{j\omega_1})\Big|_{(\pi,\pi)} = 0, \quad \text{for } i + l < k. \tag{3.5.8}$$

From (3.5.8), we conclude that the total derivatives [Lan87]

$$d^n H_0(\pi, \pi) = \sum_{i=0}^{n} \binom{n}{i} d\omega_0^i d\omega_1^{n-i} \frac{\partial^n}{\partial\omega_0^i \partial\omega_1^{n-i}} H_0(\pi, \pi) = 0, \quad \text{for } n < k. \tag{3.5.9}$$

According to [Dau93], (3.5.9) is a necessary condition for the regularity of 2D wavelet. The necessary and sufficient condition is still unknown.

5. In the FIR case the linear phase property of the analysis and synthesis filters is preserved.
6. In the IIR case, the 2D analysis and synthesis filters have a line of zeros in the frequency plane at $\omega_0 = 0$ or at $\omega_0 = \pi$.

   Proof: Substituting $z_0 = -1$ into the expression for $H_0(z_0, z_1)$ in (3.5.5a) and using the fact that $\beta(z_0 z_1)$ is allpass, one immediately finds that $H_0(-1, z_1) = 0$, $\forall\ z_1$. Since $F_0(z_0, z_1)$ contains $H_0(z_0, z_1)$ as a factor, $F_0(-1, z_1) = 0$. Similarly, we can prove that $F_1(1, z_1) = H_1(1, z_1) = 0$, $\forall\ z_1$.
7. The lowpass/highpass characteristics of the frequency responses of the filters are preserved.

   Proof: Assume that $\beta(z)$ satisfies the ideal conditions in (3.2.5). Then we have

$$\beta(z_0 z_1) = \begin{cases} (z_0 z_1)^{\frac{-2N+1}{2}}, & \text{for } \frac{\omega_0+\omega_1}{2} \in [0, \pi/2]; \\ -(z_0 z_1)^{\frac{-2N+1}{2}}, & \text{for } \frac{\omega_0+\omega_1}{2} \in (\pi/2, \pi]. \end{cases} \tag{3.5.10a}$$

$$\beta(z_0 z_1^{-1}) = \begin{cases} (z_0 z_1^{-1})^{\frac{-2N+1}{2}}, & \text{for } \frac{\omega_0-\omega_1}{2} \in [0, \pi/2]; \\ -(z_0 z_1^{-1})^{\frac{-2N+1}{2}}, & \text{for } \frac{\omega_0-\omega_1}{2} \in (\pi/2, \pi]. \end{cases} \tag{3.5.10b}$$

   By using the above equations, we find that $\beta(z_0 z_1)\beta(z_0 z_1^{-1})$ is equal to z_0^{-2N+1} when $(\omega_0, \omega_1) \in \Omega_0$ and equal to $-z_0^{-2N+1}$ when $(\omega_0, \omega_1) \in \Omega_1$. This proves that $H_0(z_0, z_1)$ has the ideal diamond support Ω_0. Similarly it can be shown that $H_1(z_0, z_1)$ will have the support of ideal diamond-complement.

Thus when the conditions in (3.2.5) are well-approximated by the 1D transfer function $\beta(z)$, the response of the 2D filters will be good.

8. *Low Complexity:* Though the 2D analysis and synthesis filters are nonseparable, the polyphase components are separable. Hence the complexity of the 2D filter bank is comparable to that of a separable filter bank. More precisely, it is equal to twice the complexity of the 1D transfer function $\beta(z)$.

3.5.2. Examples

Example 3.5.1. *2D IIR Filter Banks:* In this example, we transform the 1D filter bank in Example 3.3.1 into the 2D case by using above mapping. Since $N = 3$, the allpass function $A_3(z)$ needs only 3 multiplications. Since the complexity of the 2D analysis (or synthesis) bank is equal to twice that of $A_3(z)$, we need only 6 multiplications per input pixel to implement the analysis (synthesis) bank. The responses of $H_0(z_0, z_1)$ and $H_1(z_0, z_1)$ are shown in Fig. 3.5.5(a) and (b) respectively. The supports of the two filters are diamond and diamond-complement respectively as desired. The stopband attenuation $\delta_s(H_0) \approx 42$ dB and $\delta_s(H_1) \approx 32$ dB. Again, we see that H_1 is about 10 dB worse than H_0 in the stopband. The line of zero of H_1 at $\omega_0 = 0$ is clearly seen in Fig. 3.5.5(b). ■

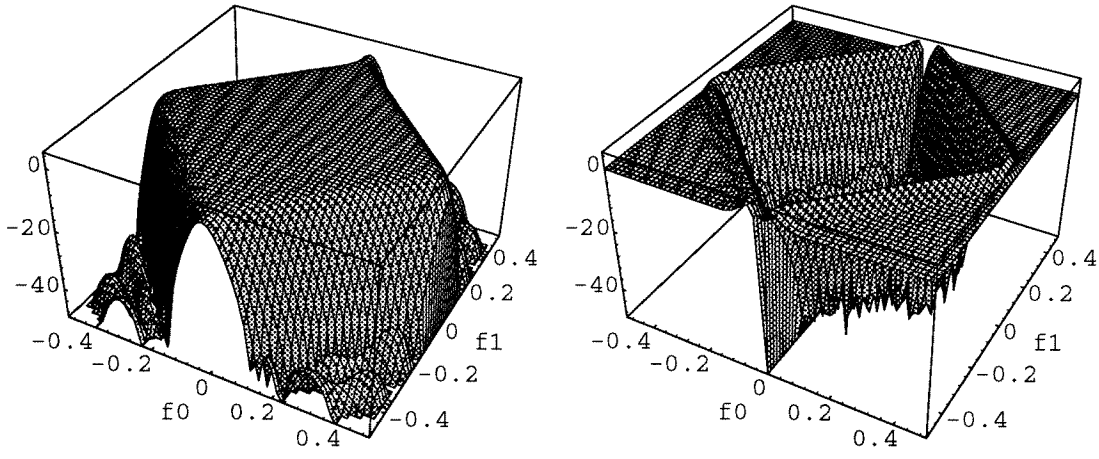

Fig. 3.5.5. Example 3.5.1–Magnitude responses of the PR IIR analysis bank: (a) $H_0(z_0, z_1)$; (b) $H_1(z_0, z_1)$. The normalized frequency $f_i = \omega_i/2\pi$.

Example 3.5.2. *2D FIR Filter Banks:* In this example, the 1D filter bank in Example 3.3.2 is transformed into the 2D case. To implement the 2D analysis (or synthesis) bank, we need 12 multiplications per input pixel. The magnitude responses of $H_0(z_0, z_1)$ and $H_1(z_0, z_1)$ are shown in Fig. 3.5.6(a) and (b) respectively. The stopband attenuation $\delta_s(H_0) \approx 40$ dB and $\delta_s(H_1) \approx 30$ dB. ■

3.6. CONCLUDING REMARKS

In this chapter, we have derived a framework for a new class of two-channel biorthogonal filter banks.

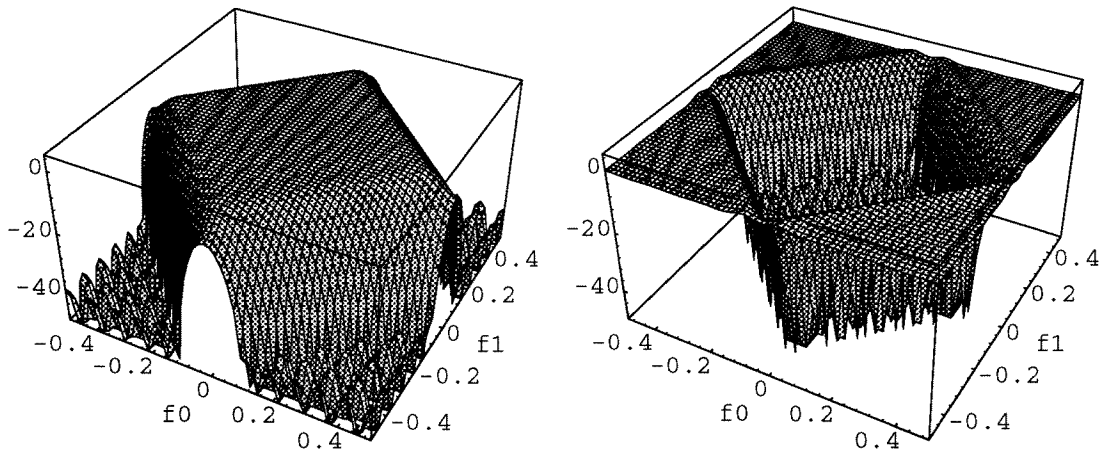

Fig. 3.5.6. Example 3.5.2–Magnitude responses of the PR FIR analysis bank: (a) $H_0(z_0, z_1)$; (b) $H_1(z_0, z_1)$.

The filter banks under the framework allow a structurally PR implementation as in Fig. 3.2.1. It is interesting that we can arrive at precisely the same ladder in Fig. 3.2.1 by using the novel approach in [Bru92] developed for a totally different application, namely cancellation of roundoff error. The proposed systems have very low complexity. Filter banks of high frequency selectivity can be achieved by controlling a single transfer function $\beta(z)$ in Fig. 3.2.1. Two different choices of $\beta(z)$ lead to causal stable IIR and linear phase FIR filter banks respectively. The properties of the proposed filter banks were discussed in detail. We showed that zeros at aliasing frequency can be imposed. Two new types of IIR maximally flat filters were derived and the solutions were given in closed form. In addition to PR property, these IIR filters have nearly linear phase in the passband. Furthermore, we also mapped the 1D filter banks derived in this chapter into 2D cases. The design of a 2D biorthogonal (stable IIR or linear phase FIR) filter bank reduces to the design of a single 1D transfer function. The new transformation preserves many of the properties of the 1D systems. Before we conclude the chapter, we would like to provide an example to demonstrate that the mapping in Section 3.5 can be easily generalized to arbitrary dilation matrix $\mathbf{M}$ with determinant equal to 2.

Generalization of Ladder Structure to the M-Channel Case: In [Pho94a], we generalize the robust ladder structure in Fig. 3.2.1 to the more general M-channel case. Some successful design examples are given for the FIR case. However in the IIR case, the resulting filters will always have bumps in the stopband which is undesirable in most applications.

Example 3.6.1. *2D IIR Filter Banks:* The 1D prototype filter bank is taken to be that in Example 3.3.1. The dilation matrix and the coset vectors are respectively:

$$\mathbf{M} = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{k_0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \mathbf{k_1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{3.6.1}$$

With the above matrix and coset vectors, the ideal passband support for $H_0(z_0, z_1)$ is $SPD(\pi \mathbf{M}^{-T})$, which is shown in Fig. 3.6.1 (shaded area). By using the transformation introduced in Section 3.5, we

find that the polyphase matrices in this example are the same as those in Example 3.5.1. Thus it also has very low complexity. The only differences are the dilation matrix and the coset vectors. With the $\mathbf{M}$ and $\mathbf{k_i}$ chosen as (3.6.1), the responses of $H_0(z_0, z_1)$ and $H_1(z_0, z_1)$ are shown in Fig. 3.6.1(a) and (b). We see that H_0 and H_1 have approximately the desired support. ■

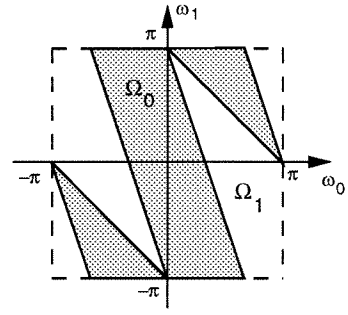

Fig. 3.6.1. Ideal supports for alias-free decimation for $\mathbf{M}$ defined in (3.6.1).

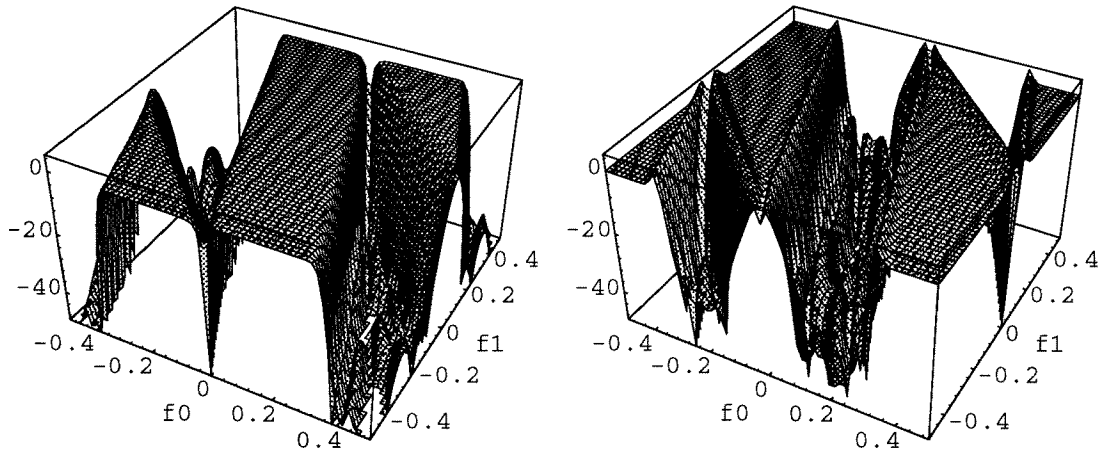

Fig. 3.6.2. Example 3.6.1–Magnitude responses of the PR analysis bank with the decimator $\mathbf{M}$ defined in (3.6.1): (a) $H_0(z_0, z_1)$; (b) $H_1(z_0, z_1)$.

3.7. APPENDICES

Appendix A. Maximally Flat IIR Solutions

In the following, we will prove that the coefficients of the maximally flat IIR solutions are given in (3.4.7). It is shown in [Gum78] that there exists closed form solution for $a_{N,k}$ satisfying system of linear equations:

$$1 + \sum_{k=1}^{N} a_{N,k} x_k^{2r} = 0, \quad \text{for } r = 1, 2, \ldots, N. \tag{3.A.1}$$

With some modification, the solution to (3.4.6) can be written as:

$$a_{N,k} = -\frac{1}{x_k}\prod_{i=1}^{k-1}\frac{1-x_{k-i}^2}{x_k^2-x_{k-i}^2}\prod_{j=k+1}^{N}\frac{x_j^2-1}{x_j^2-x_k^2}, \tag{3.A.2}$$

where $x_k = 1-4k$. Substituting the value for x_k into the equation, we find that

$$\prod_{i=1}^{k-1}\frac{1-x_{k-i}^2}{x_k^2-x_{k-i}^2} = (-1)^{k-1}\prod_{i=1}^{k-1}\frac{(2i-1)}{(4k-2i-1)}, \tag{3.A.3a}$$

$$\prod_{j=k+1}^{N}\frac{x_j^2-1}{x_j^2-x_k^2} = \binom{N}{k}\prod_{j=k+1}^{N}\frac{(2j-1)}{(2k+2j-1)}. \tag{3.A.3b}$$

Combining (3.A.3a) and (3.A.3b), we get (3.4.7).

Appendix B. Preservation of Zeros at Aliasing Frequency

We will prove the Property 4 in Section 3.5.1 in this appendix. Supposing that the 1D filter $H_0(z)$ has k zeros at π, then we have

$$e^{j\omega}H_0(e^{j\omega}) = e^{(-2N+1)j\omega} + \beta(e^{2j\omega}) = \left(1+e^{j\omega}\right)^k p(e^{j\omega}), \tag{3.B.1}$$

where $|p(-1)|$ is a finite nonzero constant. From (3.5.5a), we have

$$e^{j\omega_0}H_0(e^{j\omega_0}, e^{j\omega_1}) = e^{(-2N+1)j\omega_0} + \beta(e^{j(\omega_0+\omega_1)})\beta(e^{j(\omega_0-\omega_1)}). \tag{3.B.2}$$

From (3.B.2), $e^{j\omega_0}H_0(e^{j\omega_0}, e^{j\omega_1})$ can be rewritten as

$$\begin{aligned} e^{j\omega_0}H_0(e^{j\omega_0}, e^{j\omega_1}) =& \beta(e^{j(\omega_0-\omega_1)})\left(e^{(-2N+1)j\frac{\omega_0+\omega_1}{2}} + \beta(e^{j(\omega_0+\omega_1)})\right) \\ & - e^{(-2N+1)j\frac{\omega_0+\omega_1}{2}}\left(\beta(e^{j(\omega_0-\omega_1)}) - e^{(-2N+1)j\frac{\omega_0-\omega_1}{2}}\right). \end{aligned} \tag{3.B.3}$$

By using (3.B.1) and the fact that $\beta(e^{j\omega})$ is of period 2π, we get (3.5.7).

4

Paraunitary Filter Banks over Finite Fields

4.1. INTRODUCTION

Filter banks (FB) have found many successful applications in the subband coding of audio, image and video signals [Woo91, Mal92, Vai93, Fli94, Vet95, Tek95]. Figs. 1.1.1 and 1.2.1 show respectively an M-channel FB and its polyphase representation. For the convenience of discussion, we reproduce these figures in Fig. 4.1.1. Despite the success of real or complex FBs in various applications, little attention has been paid to the case of finite fields. Even though in most of the applications the input is a digital signal which has a finite number of quantization levels, FBs from real or complex field have been used. FBs over finite fields have the advantage that all the roundoff error and the coefficient quantization error can be eliminated completely. In addition, FBs in finite fields have the potential applications in cryptography, in the theory of error-correcting code, and in the coding or analysis of halftone images [Vai90a, Coo94, Swan95]. While these applications still remain to be explored, the immediate purpose of this chapter is to study the theory of PU FBs in finite fields.

4.1.1. Previous Work on Finite Field Filter Banks

The generalization of PU FBs to the case of $GF(2)$ was first done in [Vai90a]. The author showed that even though many properties of PU FBs in complex field continue to hold in the case of $GF(2)$, there were some unexpected properties. Unlike the conventional PU FBs, it was shown that there are PU FBs over $GF(2)$ that cannot be decomposed into degree-one building blocks. In [Coo94], the authors used the alias cancellation (AC) matrix approach to study the theory of FBs over finite fields. In order to obtain PR FBs in finite fields using the AC matrix approach, the authors needed the existence of M-th root of unity in $GF(q)$ for a M-channel FB over $GF(q)$ (which is not always possible). Because of this limitation, the authors in [Coo94] are unable to obtain M-channel PR FBs over $GF(q)$ when $M \geq q$. In [Swan95], the authors proposed a new binary field transform as an alternative to the DFT over $GF(2)$. Using the new transform, the authors were able to define bandwidth, vanishing moments, and spectral content in the filters over $GF(2)$. The application of FBs in $GF(2)$ to the analysis of binary images

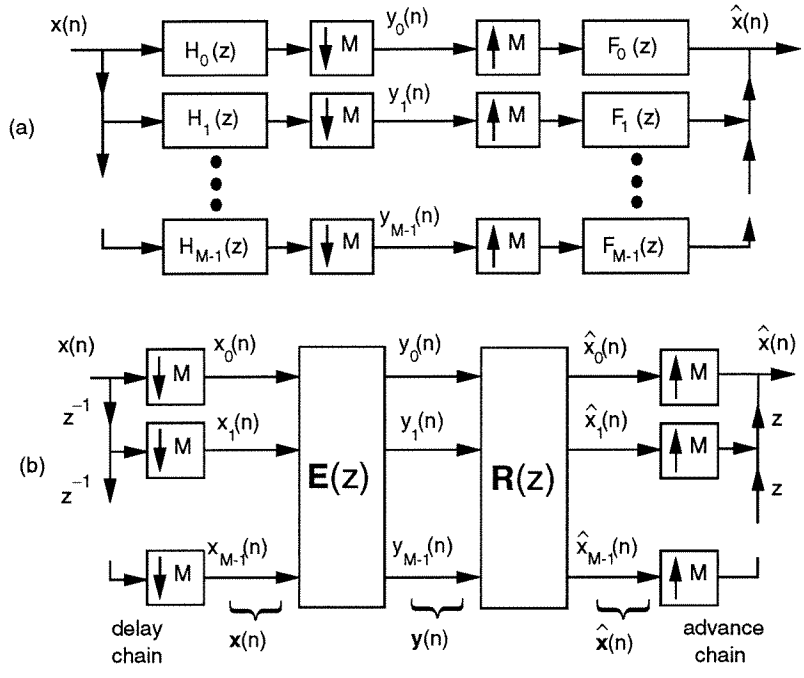

Fig. 4.1.1. (a) M-channel maximally decimated filter bank and (b) its polyphase representation.

was also demonstrated. In [Xia95], the author related the theory of finite field filter bank to the theory of error correcting codes. The application of finite field filter banks to the problem of partial response channel was also studied.

4.1.2. Chapter Outline

Our aim in this chapter is to study theoretical aspects of FBs in finite fields. Our presentation will go as follows: In Section 4.2, we will discuss some basic properties of unitary matrices in $GF(2)$. The $GF(2)$ case are shown to be very different from the complex case. Despite all the unusual properties, we can prove that all unitary matrices can be expressed as a product of permutation matrices and Householder-like matrices. PU matrices in $GF(2)$ are studied in Section 4.3. In Section 4.4, we will derive the most general degree-one building block for PU matrices in $GF(2)$, and derive the conditions under which arbitrary PU matrices in $GF(2)$ can be factorized into these building blocks. A degree-one reduction algorithm will be given. We will also show that there are PU systems in $GF(2)$ that cannot be expressed in terms of these building blocks! We will establish new factorization theorems for PU matrices in Section 4.5. The new theorems involve a building block of degree two. However there are PU systems which cannot be decomposed into any combination of these degree-one and degree-two building blocks. The conventional lapped orthogonal transform (LOT) has been studied in detail [Mal92]. In Section 4.6, we will study the LOT in $GF(2)$. We will show that LOTs in $GF(2)$ can always be factorized in terms of the degree-one and degree-two building blocks. State-space representation of PU systems in $GF(2)$ will be considered in Section 4.7. The implementations based on the factorization are shown to be minimal in terms of delay

elements. Moreover we will show that the well-known LBR lemma [Vai93] cannot be extended to the $GF(2)$ case. In the last section, the theory of PU systems in $GF(2)$ will be extended to the case of $GF(q)$ for prime $q > 2$.

Chapter Related Definitions: In finite fields, since a nonzero vector $\mathbf{v}$ can have $\mathbf{v}^T\mathbf{v} = 0$, the vector space of all M-dimentional vectors is not an inner-product space. Hence orthogonality is not well-defined. However for simplicity, in this chapter we will borrow the jargon from the theory of convolutional codes [For70, McE95]. Two vectors that satisfy $\mathbf{u}^T\mathbf{v} = 0$ are said to be *orthogonal* and matrices that satisfy $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ will be called *unitary* matrices. A rational matrix in finite fields that satisfies $\mathbf{E}^T(z^{-1})\mathbf{E}(z) = \mathbf{I}$ is called a PU matrix. In this chapter, the vector $\mathbf{e}_i$ denotes the i-th column of the identity matrix $\mathbf{I}$.

4.2. UNITARY MATRICES OVER GF(2)

For simplicity, we assume that all the matrices in this section are $M \times M$ square matrices. The result for rectangular matrices can be obtained in a similar manner. In the first part of this section, we will study some basic properties of unitary matrices over $GF(2)$, which we are going to use throughout the chapter. In the second part, we will show that all unitary matrices can be factorized by using some basic building blocks similar to the Householder transformation.

4.2.1. Basic Properties of Unitary Matrices

In $GF(2)$, a matrix $\mathbf{A}$ is said to be unitary if

$$\mathbf{A}^T\mathbf{A} = \mathbf{I}. \tag{4.2.1}$$

One important property of unitary matrices which we are going to use repeatedly later is:

Fact 4.2.1. None of the column (or row) vectors of a unitary matrix in $GF(2)$ can have an even number of 1. ■

It is not difficult to see that if $\mathbf{A}_1$ and $\mathbf{A}_2$ are unitary, so is the product $\mathbf{A}_1\mathbf{A}_2$. Post-multiplying (4.2.1) by $\mathbf{A}^{-1}$, we obtain that $\mathbf{A}^{-1} = \mathbf{A}^T$. Thus if $\mathbf{A}$ is unitary, its inverse is simply its own transpose. Pre-multiplying $\mathbf{A}^{-1} = \mathbf{A}^T$ by $\mathbf{A}$, we get $\mathbf{A}\mathbf{A}^T = \mathbf{I}$. Summarizing the results, we have shown that the following are equivalent: (i) $\mathbf{A}$ is unitary, (ii) $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, (iii) $\mathbf{A}\mathbf{A}^T = \mathbf{I}$, and (iv) $\mathbf{A}^{-1} = \mathbf{A}^T$.

From the above discussion, we see that unitary matrices over $GF(2)$ enjoy many properties similar to unitary matrices over the real or complex field. However there are some differences. For example, it is well-known that in real or complex field a matrix is unitary if and only if it has the property of energy conservation [Vai93]. That means, $\mathbf{A}$ is unitary if and only if $\mathbf{u}^\dagger\mathbf{A}^\dagger\mathbf{A}\mathbf{u} = \mathbf{u}^\dagger\mathbf{u}$ for all $\mathbf{u}$. In $GF(2)$, there are non unitary matrices that satisfy $\mathbf{u}^\dagger\mathbf{A}^\dagger\mathbf{A}\mathbf{u} = \mathbf{u}^\dagger\mathbf{u}$ for all $\mathbf{u}$. To explain this, note that

$$\mathbf{u}^T\mathbf{B}\mathbf{u} = \sum_l u_l u_l b_{ll} + \sum_{i>j} u_i u_j (b_{ij} + b_{ji}). \tag{4.2.2}$$

For any symmetric matrix $\mathbf{B}$ over $GF(2)$, the above equation reduces to $\mathbf{u}^T\mathbf{B}\mathbf{u} = \sum_l u_l b_{ll}$. Thus any symmetric matrix with $b_{ii} = 1$ will satisfy $\mathbf{u}^T\mathbf{B}\mathbf{u} = \mathbf{u}^T\mathbf{u}$. If $\mathbf{A}$ is such that all columns have odd number of nonzero elements, then $\mathbf{A}^T\mathbf{A}$ is symmetric with diagonal elements = 1. Even though $\mathbf{A}$ is not unitary, we have $\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u} = \mathbf{u}^T\mathbf{u}$ for all vectors $\mathbf{u}$. For unitariness of matrices in $GF(2)$, we need a stronger condition as follows:

Fact 4.2.2. If $\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{v} = \mathbf{u}^T\mathbf{v}$ for all possible vectors $\mathbf{u}$ and $\mathbf{v}$, then $\mathbf{A}$ is unitary. ■

Proof: Let $\mathbf{u}$ and $\mathbf{v}$ be respectively the unit vectors $\mathbf{e}_i$ and $\mathbf{e}_j$ defined in Section 4.1. If $\mathbf{e}_i^T\mathbf{A}^T\mathbf{A}\mathbf{e}_j = \mathbf{e}_i^T\mathbf{e}_j$ for all i, j, then we have

$$\begin{bmatrix} \mathbf{e}_0^T \\ \mathbf{e}_1^T \\ \vdots \\ \mathbf{e}_{M-1}^T \end{bmatrix} \mathbf{A}^T\mathbf{A} \begin{bmatrix} \mathbf{e}_0 & \mathbf{e}_1 & \dots & \mathbf{e}_{M-1} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_0^T \\ \mathbf{e}_1^T \\ \vdots \\ \mathbf{e}_{M-1}^T \end{bmatrix} \begin{bmatrix} \mathbf{e}_0 & \mathbf{e}_1 & \dots & \mathbf{e}_{M-1} \end{bmatrix}. \tag{4.2.3}$$

Since $[\mathbf{e}_0 \quad \mathbf{e}_1 \quad \dots \quad \mathbf{e}_{M-1}] = \mathbf{I}$, it immediately follows from the above equation that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. ■

Fact 4.2.3. If $\mathbf{A}$ is a unitary matrix over $GF(2)$, then none of the columns (or rows) can have all elements equal to unity.

Proof: Let $\mathbf{A} = [\mathbf{a}_0 \quad \mathbf{a}_1 \quad \dots \quad \mathbf{a}_{M-1}]$. Suppose $\mathbf{a}_0$ is a column vector with all elements equal to unity. Since $\mathbf{a}_i^T\mathbf{a}_0 = 0$ for $i \neq 0$, we conclude that $\mathbf{a}_i$ must have an even number of unit elements, which is a contradiction to Fact 4.2.1! ■

Combining Facts 4.2.1 and 4.2.3, we conclude that for any $M \times M$ unitary matrix with $M \leq 3$, the column has only one nonzero element. Therefore any $M \times M$ unitary matrix with $M \leq 3$ must be a permutation of the identity matrix. As we will see later in this section, Fact 4.2.3 is very useful in the factorization of unitary matrices. Before we derive the factorization theorem for unitary matrices, we would like to introduce the following building block:

Fact 4.2.4. The matrix $\mathbf{U} = \mathbf{I} + \mathbf{u}\mathbf{u}^T$ with $\mathbf{u}^T\mathbf{u} = 0$ is unitary. ■

The above fact can be proven by direct computation of $\mathbf{U}^T\mathbf{U}$. Moreover it can be verified that $\mathbf{U}$ is its own inverse. As we will see next, the building block in Fact 4.2.4 has a similar function as the Householder transformation.

4.2.2. Factorization of Unitary Matrices over $GF(2)$

In this section, we will show how to parameterize all $M \times M$ unitary matrices. In the real field, all unitary matrices can be written as a product of planar rotations. Since the planar rotations involve sines and cosines, we cannot attempt the same approach in the finite field. Instead, we will use an approach similar to the Householder factorization. In real or complex field, the Householder transformation is a matrix of the form $(\mathbf{I} - \frac{2\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}})$ [Hor85]. In $GF(2)$, since all the computations are performed modulo 2, there is no such Householder transformation in $GF(2)$. However, we will show that we can capture all unitary matrices using the building block $\mathbf{U}$ introduced in Fact 4.2.4. As we have pointed out above, all $M \times M$ unitary matrix with $M \leq 3$ must be a permutation of the identity matrix, so only $M > 3$ is of interest

in the discussion of this section. Before we derive the factorization theorem for $M > 3$, we will show two lemmas which are crucial in this context.

Lemma 4.2.1. Let $\mathbf{v}$ be a vector over $GF(2)$ such that $\mathbf{v}^T\mathbf{v} = 1$ and $v_0 = 0$. Then

$$(\mathbf{I} + \mathbf{w}\mathbf{w}^T)\mathbf{v} = \mathbf{e}_0, \quad \text{and} \quad \mathbf{v}^T(\mathbf{I} + \mathbf{w}\mathbf{w}^T) = \mathbf{e}_0^T, \tag{4.2.4}$$

where $\mathbf{w} = \mathbf{v} + \mathbf{e}_0$, and $\mathbf{e}_0 = [1\ 0\ \ldots\ 0]^T$. ■

The above lemma can be proved by direct substitution. Note that the vector $\mathbf{w}$ has $\mathbf{w}^T\mathbf{w} = 0$ so that $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)$ is unitary (by Fact 4.2.4). The function of $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)$ is similar to the Householder matrix in the real or complex case. The matrix $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)$ will transform the vector $\mathbf{v}$ into the vector $\mathbf{e}_0$. It is not difficult to generalize the result of Lemma 4.2.1 as follows: If $\mathbf{v}$ is a vector such that $\mathbf{v}^T\mathbf{v} = 1$ and $v_i = 0$, then it can be shown that the matrix $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)$ with $\mathbf{w} = \mathbf{v} + \mathbf{e}_i$ transforms the vector $\mathbf{v}$ into $\mathbf{e}_i$. As a consequence of Lemma 4.2.1, we have the following:

Lemma 4.2.2. Let $\mathbf{A}$ be $M \times M$ unitary over $GF(2)$ with $A_{00} = 0$. Define the vector $\mathbf{w} = \mathbf{a}_0 + \mathbf{e}_0$ where $\mathbf{a}_0$ is the 0-th column of $\mathbf{A}$. Then $\mathbf{w}^T\mathbf{w} = 0$, and

$$\mathbf{A} = (\mathbf{I} + \mathbf{w}\mathbf{w}^T)\begin{bmatrix} 1 & 0 \\ 0 & \mathbf{B} \end{bmatrix}, \tag{4.2.5}$$

where $\mathbf{B}$ is $(M-1) \times (M-1)$ unitary. ■

Proof: Since $\mathbf{A}$ is unitary with $A_{00} = 0$, the vector $\mathbf{a}_0$ satisfies the conditions in Lemma 4.2.1. Applying the result of Lemma 4.2.1, we have $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)\mathbf{a}_0 = \mathbf{e}_0$. Thus

$$(\mathbf{I} + \mathbf{w}\mathbf{w}^T)\mathbf{A} = [\mathbf{e}_0 \ \ \mathbf{C}], \tag{4.2.6}$$

where $\mathbf{C}$ is $M \times (M-1)$. Since both $\mathbf{A}$ and $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)$ are unitary, the right hand sight of (4.2.6) is also unitary. Thus the first row of $\mathbf{C}$ contains only zeros. Inverting $(\mathbf{I} + \mathbf{w}\mathbf{w}^T)$ in (4.2.6), we immediately get (4.2.5). ■

With the above two lemmas, we are now ready to prove the main factorization:

Theorem 4.2.1. *Factorization of Unitary Matrices in $GF(2)$:* An $M \times M$ matrix $\mathbf{A}$ over $GF(2)$ (with $M \geq 3$) is unitary if and only if it can be factorized as:

$$\mathbf{A} = \mathbf{P}_M\mathbf{U}_M\mathbf{P}_{M-1}\ldots\mathbf{P}_4\mathbf{U}_4\mathbf{P}_3, \tag{4.2.7}$$

where the unitary matrices $\mathbf{U}_i = \mathbf{I} + \mathbf{u}_i\mathbf{u}_i^T$ with $\mathbf{u}_i^T\mathbf{u}_i = 0$, and $\mathbf{P}_i$ are permutations of identity matrix. ■

Proof: The 'if' part is self-evident. To prove the 'only if' part, assume that $\mathbf{A}$ is unitary. If $A_{00} \neq 0$, we can apply a row permutation such that the $(0,0)$-th element is zero. This is always possible because of Fact 4.2.3. Then the factorization in Lemma 4.2.2 can be applied. Repeat the permutation and factorization operations on the smaller unitary matrix $\mathbf{B}$. Continuing the process, we can successively

generate unitary matrices of smaller and smaller size until we get a 3×3 unitary matrix, which itself is a permutation of the identity matrix. Note that the permutation $\mathbf{P}_i$ in (4.2.7) is $M \times M$ and has the form:

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{I}_{M-i} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{P}}_i \end{bmatrix}, \tag{4.2.8}$$

where $\widehat{\mathbf{P}}_i$ is a $i \times i$ permutation matrix. ■

Remark: In (4.2.5) we have extracted a left factor from $\mathbf{A}$. If we take the 0-th row of $\mathbf{A}$, $\widehat{\mathbf{a}}_0^T$ to form the vector $\widehat{\mathbf{w}} = \widehat{\mathbf{a}}_0 + \mathbf{e}_0$, then we can rewrite (4.2.5) as $\mathbf{A} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{B}} \end{bmatrix} (\mathbf{I} + \widehat{\mathbf{w}}\widehat{\mathbf{w}}^T)$. In this case, we can extract a factor from the right of $\mathbf{A}$. Using this right factor extraction and the column permutation, we can factorize $\mathbf{A}$ as $\mathbf{P}_3\mathbf{U}_4\mathbf{P}_4 \ldots \mathbf{P}_{M-1}\mathbf{U}_M\mathbf{P}_M$, where $\mathbf{P}_i$ are as in (4.2.8).

4.3. PU MATRICES AND FILTER BANKS OVER GF(2)

Let $\mathbf{E}(z)$ be a matrix whose entries are rational with coefficients from $GF(2)$. As defined in Section 4.1, the matrix $\mathbf{E}(z)$ is said to be PU if

$$\mathbf{E}^T(z^{-1})\mathbf{E}(z) = \mathbf{I}. \tag{4.3.1}$$

In this section, we will restrict our attention to the FIR case when $\mathbf{E}(z) = \sum_{k=0}^{N} \mathbf{e}(k)z^{-k}$. As we mentioned Section 1.2, the number N is called the order of the system. In the case of real or complex field, the first order PU matrix is called the lapped orthogonal transform (LOT) [Mal92]. The class of LOT in $GF(2)$ can be similarly defined. We will see that this class allows a minimal factorization in terms of smaller PU building blocks. But unlike the complex case, we need both degree-one and degree-two building blocks in the factorization of LOT in $GF(2)$.

4.3.1. Some Basic Properties of PU Matrices

Equation (4.3.1) gives a z-domain characterization of PU matrices. In the time-domain, it can be shown that the impulse response satisfies

$$\sum_n \mathbf{e}^T(n)\mathbf{e}(n+k) = \begin{cases} \mathbf{I}, & k = 0; \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{4.3.2}$$

The conditions in (4.3.2) is very similar to those for PU matrices in real or complex field. Eqn. (4.3.2) gives one time-domain condition for PU matrices. Using the fact that (4.3.1) implies $\mathbf{E}(z)\mathbf{E}^T(z^{-1}) = \mathbf{I}$, we obtain another time-domain condition as:

$$\sum_n \mathbf{e}(n)\mathbf{e}^T(n+k) = \begin{cases} \mathbf{I}, & k = 0; \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{4.3.3}$$

Even though some properties of the real or complex case continue to hold in the case of $GF(2)$, there are some exceptions. For example, in the real or complex field, if the system $\mathbf{E}(z)$ has the input-output

energy preservation property, then it is PU. In general, this is not true for the $GF(2)$ case. A counter example is given by

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{4.3.4}$$

In this case, $\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u} = \mathbf{u}^T\mathbf{u}$ for all $\mathbf{u}$ but $\mathbf{A}^T\mathbf{A} \neq \mathbf{I}$. The precise relation between PU property and input-output mapping is given by the following result:

Lemma 4.3.1. Let $\mathbf{y}_0(n)$ and $\mathbf{y}_1(n)$ be respectively the outputs of $\mathbf{E}(z)$ in response to $\mathbf{u}_0(n)$ and $\mathbf{u}_1(n)$. Then $\mathbf{E}(z)$ (over $GF(2)$) is PU if and only if

$$\sum_n \mathbf{y}_0^T(n)\mathbf{y}_1(n) = \sum_n \mathbf{u}_0^T(n)\mathbf{u}_1(n), \tag{4.3.5}$$

for all possible inputs pairs $\mathbf{u}_0(n)$ and $\mathbf{u}_1(n)$. ■

Proof: The outputs can be written as $\mathbf{y}_i(n) = \sum_k \mathbf{e}(k)\mathbf{u}_i(n-k)$ for $i = 0, 1$. Substituting this into the left hand side of (4.3.5) and rearranging the result, we get

$$\sum_n \mathbf{y}_0^T(n)\mathbf{y}_1(n) = \sum_{k,l} \mathbf{u}_0^T(l) \underbrace{\left(\sum_n \mathbf{e}^T(n)\mathbf{e}(n+k) \right)}_{\mathbf{A}(k)} \mathbf{u}_1(l-k). \tag{4.3.6}$$

If we choose $\mathbf{u}_0(n) = \mathbf{e}_i\delta(n)$ and $\mathbf{u}_1(n) = \mathbf{e}_j\delta(n)$, then the right hand side of (4.3.6) reduces to $a_{ij}(0)$, the (i,j)-th element of $\mathbf{A}(0)$. Using (4.3.5), we conclude that $\mathbf{A}(0) = \mathbf{I}$. Similarly by choosing $\mathbf{u}_0(n) = \mathbf{e}_i\delta(n)$ and $\mathbf{u}_1(n) = \mathbf{e}_j\delta(n+k)$, we can prove that $\mathbf{A}(k) = \mathbf{0}$ for $k \neq 0$. ■

McMillan Degree and Determinant of PU Systems: In the FIR case, the PU property puts a strong constraint on the determinant of $\mathbf{E}(z)$. Taking the determinant of (4.3.1), we get $[\det \mathbf{E}(z)] = z^{-\rho}$ for some integer ρ. In [Vai95c], it is proved for the real and complex fields that the McMillan degree (often called just degree) of causal systems with anti-causal inverses is equal to the degree of the determinant. One can verify that the same proof carries through for systems in the finite fields. In particular, PU system in $GF(2)$ has anti-causal inverse, so degree of the determinant is equal to McMillan degree. The McMillan degree of systems in finite fields has been investigated by researchers in coding theory [For70, McE95]. For a detail study on the topic of McMillan degree, the readers are referred to [For70, Kai80, Vai93, McE95].

4.3.2. PU Filter Banks in $GF(2)$

Consider Figs. 4.1.1. The analysis and synthesis filters are related to the polyphase matrices as:

$$H_k(z) = \sum_{i=0}^{M-1} E_{ki}(z^M)z^{-i}, \quad F_k(z) = \sum_{i=0}^{M-1} R_{ik}(z^M)z^i, \tag{4.3.7}$$

where $E_{ki}(z)$ and $R_{ik}(z)$ are respectively the ki-th and ik-th elements of $\mathbf{E}(z)$ and $\mathbf{R}(z)$. If the analysis polyphase matrix $\mathbf{E}(z)$ is PU, then the polyphase components of the analysis filters satisfy the relation:

$$\sum_i E_{ki}(z)E_{li}(z^{-1}) = \delta(k-l), \tag{4.3.8}$$

which is very similar to the orthogonality condition in the case of real or complex field. Equation (4.3.8) can be rewritten as $[H_l(z^{-1})H_k(z)]_{\downarrow M} = \delta(k-l)$, where $X(z)\big|_{\downarrow M}$ denotes the z-transform of $x(Mn)$. If we take the synthesis polyphase matrix as

$$\mathbf{R}(z) = \mathbf{E}^T(z^{-1}), \tag{4.3.9}$$

then we have a PR FB in $GF(2)$. Using (4.3.7) and (4.3.9), we find that the synthesis filters $F_k(z)$ are time-reversed version of the analysis filters $H_k(z)$:

$$F_k(z) = H_k(z^{-1}). \tag{4.3.10}$$

In the special case of two-channel FIR PU FBs, all the analysis and synthesis are determined by one filter. To be more specific, we have

$$H_1(z) = z^{-N}H_0(z^{-1}), \quad F_0(z) = H_0(z^{-1}), \quad F_1(z) = z^N H_0(z), \tag{4.3.11}$$

where N is the order of the filter $H_0(z)$. The other filters are simply either time-reversed or delayed versions of $H_0(z)$.

4.4. DEGREE-ONE PU SYSTEMS AND FACTORIZATIONS

In this section we introduce the following degree-one causal FIR system over $GF(2)$

$$\mathbf{D}(z) = \mathbf{I} + \mathbf{v}\mathbf{v}^T + z^{-1}\mathbf{v}\mathbf{v}^T, \quad \mathbf{v}^T\mathbf{v} = 1. \tag{4.4.1}$$

By direct computation, we can verify that $\mathbf{D}^T(z^{-1})\mathbf{D}(z) = \mathbf{I}$. So this is a PU system. The system in (4.4.1) has degree one, and Fig. 4.4.1 shows an implementation using one delay. We will study its properties and show that it can be used for the synthesis of more general PU systems.

4.4.1. Basic Properties

1. The inverse system is obtained by replacing z^{-1} with z. That is

$$\mathbf{D}^{-1}(z) = \mathbf{D}(z^{-1}) = \mathbf{I} + \mathbf{v}\mathbf{v}^T + z\mathbf{v}\mathbf{v}^T, \quad \mathbf{v}^T\mathbf{v} = 1. \tag{4.4.2}$$

Fig. 4.4.2 shows an implementation of the inverse $\mathbf{D}^{-1}(z)$.

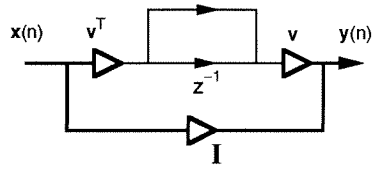

Fig. 4.4.1. Degree-one PU building block. Here $\mathbf{v}^T\mathbf{v} = 1$.

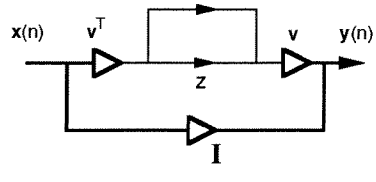

Fig. 4.4.2. The inverse of the degree-one PU system in Fig. 4.4.1.

2. A cascade of k such systems gives $\prod_{k\ times} \mathbf{D}(z) = \mathbf{D}(z^k)$.
3. Let $\{\mathbf{v}_0,\ \mathbf{v}_1,\ \ldots,\ \mathbf{v}_{s-1}\}$ be a set of vectors in $GF(2)$ such that $\mathbf{v}_i^T\mathbf{v}_j = \delta(i-j)$. Then the cascade

$$\prod_{i=0}^{s-1}\left[\mathbf{I} + \mathbf{v}_i\mathbf{v}_i^T + z^{-1}\mathbf{v}_i\mathbf{v}_i^T\right] = \mathbf{I} + \sum_{i=0}^{s-1}\mathbf{v}_i\mathbf{v}_i^T + z^{-1}\sum_{i=0}^{s-1}\mathbf{v}_i\mathbf{v}_i^T. \tag{4.4.3}$$

It is clear from the right-hand side of (4.4.3) that if we interchange any $\mathbf{v}_i$ with any $\mathbf{v}_j$, the system remains the same. Hence the factors $(\mathbf{I}+\mathbf{v}_i\mathbf{v}_i^T+z^{-1}\mathbf{v}_i\mathbf{v}_i^T)$ commute. Moreover it can be shown that a cascade of PU building blocks $\mathbf{D}_i(z)$ with vectors $\mathbf{v}_i$ will have order-one if and only if the vectors satisfy $\mathbf{v}_i^T\mathbf{v}_j = \delta(i-j)$. If we let $\mathbf{V} = [\mathbf{v}_0\ \ldots\ \mathbf{v}_{s-1}]$, then the system in (4.4.3) can be rewritten as $\mathbf{I} + \mathbf{V}\mathbf{V}^T + z^{-1}\mathbf{V}\mathbf{V}^T$.

Lemma 4.4.1. *Most general degree-one PU system:* The most general $M \times M$ causal FIR degree-one PU system over $GF(2)$ can be written as:

$$\mathbf{E}(z) = (\mathbf{I} + \mathbf{v}\mathbf{v}^T + z^{-1}\mathbf{v}\mathbf{v}^T)\mathbf{E}(1), \tag{4.4.4}$$

where $\mathbf{v}$ is a column vector with $\mathbf{v}^T\mathbf{v} = 1$ and $\mathbf{E}(1)$ is a $M \times M$ constant matrix with $\mathbf{E}^T(1)\mathbf{E}(1) = \mathbf{I}$. ■

The proof of Lemma 4.4.1 is very similar to the case of real or complex field [Vai93]. Note that the PU system $\mathbf{E}(z)$ in (4.4.4) can be rewritten as $\mathbf{E}(z) = \mathbf{E}(1)(\mathbf{I}+\mathbf{u}\mathbf{u}^T+z^{-1}\mathbf{u}\mathbf{u}^T)$, with $\mathbf{u} = \mathbf{E}^T(1)\mathbf{v}$ (hence $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 1$).

4.4.2. Degree-one Reduction Using $\mathbf{D}(z)$

To show how we can extract $\mathbf{D}(z)$ from a PU system, we consider the general $M \times M$ PU system of the form $\mathbf{E}(z) = \sum_{k=0}^{N}\mathbf{e}(k)z^{-k}$ with degree ρ. To avoid trivial cases, let $\mathbf{e}(0) \neq \mathbf{0}$ and $\mathbf{e}(N) \neq \mathbf{0}$. So the system $\mathbf{E}(z)$ has order N. From the PU conditions in (4.3.2), we get $\mathbf{e}^T(0)\mathbf{e}(N) = \mathbf{0}$, which implies that both the matrices $\mathbf{e}(0)$ and $\mathbf{e}(N)$ are singular. Let $\mathbf{v}$ be a vector in the null space of $\mathbf{e}(0)$ such that $\mathbf{v}^T\mathbf{v} = 1$. Form the new system

$$\mathbf{E}'(z) = \mathbf{E}(z)(\mathbf{I} + \mathbf{v}\mathbf{v}^T + z\mathbf{v}\mathbf{v}^T). \tag{4.4.5}$$

We say that the degree-one reduction is *successful* if the new system $\mathbf{E}'(z)$ satisfies the following three conditions: (i) it is causal; (ii) it is PU; and (iii) it has degree $\rho - 1$, where ρ is the degree of $\mathbf{E}(z)$. The new system $\mathbf{E}'(z)$ in (4.4.5) is causal because $\mathbf{e}(0)\mathbf{v} = 0$. Since both $\mathbf{E}(z)$ and $(\mathbf{I}+\mathbf{v}\mathbf{v}^T+z\mathbf{v}\mathbf{v}^T)$ are PU, so is $\mathbf{E}'(z)$. Taking the determinant of (4.4.5), we see that the degree of $\mathbf{E}'(z)$ is $\rho' = [\det \mathbf{E}'(z)] = \rho - 1$. Hence $\mathbf{E}'(z)$ satisfies the three conditions mentioned above. We have sucessfully extracted a degree-one building block from $\mathbf{E}(z)$. Inverting $(\mathbf{I} + \mathbf{v}\mathbf{v}^T + z\mathbf{v}\mathbf{v}^T)$, we conclude that $\mathbf{E}(z)$ can be written as $\mathbf{E}(z) = \mathbf{E}'(z)(\mathbf{I} + \mathbf{v}\mathbf{v}^T + z^{-1}\mathbf{v}\mathbf{v}^T)$. If we can successfully repeat the above degree reduction process ρ times, then $\mathbf{E}(z)$ can be written as

$$\mathbf{E}(z) = \mathbf{E}(1)(\mathbf{I}+\mathbf{v}_{\rho-1}\mathbf{v}_{\rho-1}^T + z^{-1}\mathbf{v}_{\rho-1}\mathbf{v}_{\rho-1}^T)\ldots(\mathbf{I}+\mathbf{v}_0\mathbf{v}_0^T + z^{-1}\mathbf{v}_0\mathbf{v}_0^T), \tag{4.4.6}$$

where $\mathbf{v}_i^T\mathbf{v}_i = 1$, and $\mathbf{E}(1)$ is a constant unitary matrix. Similarly one can show that if the null space of $\mathbf{e}^T(0)$ contains a vector $\mathbf{u}$ with $\mathbf{u}^T\mathbf{u} = 1$, then we can write $\mathbf{E}(z)$ as $(\mathbf{I} + \mathbf{u}\mathbf{u}^T + z^{-1}\mathbf{u}\mathbf{u}^T)\mathbf{E}'(z)$ for some causal PU system $\mathbf{E}'(z)$. If $\mathbf{E}(z)$ is completely factorizable into ρ terms, using this degree reduction process from the left, we can write $\mathbf{E}(z)$ as

$$\mathbf{E}(z) = (\mathbf{I} + \mathbf{u}_0\mathbf{u}_0^T + z^{-1}\mathbf{u}_0\mathbf{u}_0^T)\dots(\mathbf{I} + \mathbf{u}_{\rho-1}\mathbf{u}_{\rho-1}^T + z^{-1}\mathbf{u}_{\rho-1}\mathbf{u}_{\rho-1}^T)\mathbf{E}(1). \tag{4.4.7}$$

Equivalence of (4.4.6) and (4.4.7): If $\mathbf{E}(z)$ can be written as the factorized form in (4.4.6), then it can also be expressed as (4.4.7). To prove this, we consider (4.4.6). Starting from the left, we can move the constant matrix $\mathbf{E}(1)$ to the right by letting $\mathbf{u}_i = \mathbf{E}(1)\mathbf{v}_i$.

In the real or complex field, it is well-known that all FIR causal PU matrices of degree ρ can always be factorized into a product of ρ degree-one PU systems of the form $\mathbf{D}(z)$. A similar property is not true in $GF(2)$. To see this, consider the following example:

Example 4.4.1. *A PU system that is unfactorizable in terms of* $\mathbf{D}(z)$*:* Let $\mathbf{G}(z)$ be the following $M \times M$ system with M odd:

$$\mathbf{G}(z) = \mathbf{w}\mathbf{w}^T + z^{-1}(\mathbf{I} + \mathbf{w}\mathbf{w}^T), \tag{4.4.8}$$

where $\mathbf{w} = [1\ 1\ \dots\ 1]^T$ so that $\mathbf{w}^T\mathbf{w} = 1$. It can be verified that $\mathbf{G}^T(z^{-1})\mathbf{G}(z) = \mathbf{I}$. So $\mathbf{G}(z)$ is PU. Let $\{\mathbf{u}_0, \dots, \mathbf{u}_{M-2}\}$ be a set of independent vectors such that $\mathbf{w}^T\mathbf{u}_k = 0$. Then we get $\mathbf{g}(1)\mathbf{w} = \mathbf{0}$ and $\mathbf{g}(1)\mathbf{u}_k = \mathbf{u}_k$ for $0 \le k \le M-2$. So $\mathbf{g}(1)$ has rank $M-1$ and the degree of $\mathbf{G}(z)$ is $M-1$. Suppose that degree reduction from the left is possible. That means, $\mathbf{G}(z) = (\mathbf{I} + \mathbf{v}\mathbf{v}^T + z^{-1}\mathbf{v}\mathbf{v}^T)\mathbf{G}'(z)$, where $\mathbf{v}^T\mathbf{v} = 1$ and $\mathbf{G}'(z)$ is a causal FIR PU system of degree $\rho' = M-2$. Inverting the degree-one system, we have

$$\mathbf{G}'(z) = (\mathbf{I} + \mathbf{v}\mathbf{v}^T + z\mathbf{v}\mathbf{v}^T)\mathbf{G}(z). \tag{4.4.9}$$

Therefore $\mathbf{G}'(z)$ is causal only if $\mathbf{v}^T\mathbf{w} = 0$, which implies that $\mathbf{v}$ has an even number of ones, violating the requirement of $\mathbf{v}^T\mathbf{v} = 1$. Thus degree reduction from the left is impossible. Similarly we can show that degree reduction from the right is also impossible. Therefore we conclude that the system $\mathbf{G}(z)$ in (4.4.8) cannot be factorized in terms of degree-one PU system $\mathbf{D}(z)$! From the above discussion, it is clear that the degree reduction fails because neither the null space of $\mathbf{e}(0)$ nor $\mathbf{e}^T(0)$ contains a vector with an odd number of ones. In the complex field, this can never happen because nonzero vectors always have nonzero norm. ■

Lemma 4.4.2. *Test of Degree-One Factorizability in* $GF(2)$*:* Let $\mathbf{E}(z) = \sum_{k=0}^{N}\mathbf{e}(k)z^{-k}$ be a causal PU system. The degree-one reduction for $\mathbf{E}(z)$ fails if and only if the null spaces of $\mathbf{e}(0)$ and $\mathbf{e}^T(0)$ contain only vectors with an *even* number of ones. ■

The above lemma can be proved in a straightforward manner. Note that it is not necessary to exhaust the whole null space for the test. We need only to look at any basis that spans the null space.

If none of the vectors in this basis has an odd weight, then any linear combination of vectors in the null space has an even weight because in $GF(2)$

$$\Big(\sum_i \mathbf{v}_i\Big)^T\Big(\sum_i \mathbf{v}_i\Big) = \sum_i \mathbf{v}_i^T\mathbf{v}_i. \tag{4.4.10}$$

4.5. DEGREE-TWO PU BUILDING BLOCKS AND FACTORIZATIONS

As we have seen in Example 4.4.1, there are PU systems that cannot be factorized by using the degree-one building blocks. In this section, we will include a degree-two building block in the factorization so that some PU systems that cannot be factorized before can now be factorized. To establish new factorization theorems for PU systems, we introduce the following degree-one system:

$$\mathbf{G}(z) = \mathbf{I} + \mathbf{u}\mathbf{v}^T + z^{-1}\mathbf{u}\mathbf{v}^T, \tag{4.5.1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors over $GF(2)$. The above system is not PU unless $\mathbf{u} = \mathbf{v}$ and $\mathbf{v}^T\mathbf{v} = 1$. To see this, suppose $\mathbf{G}^T(z^{-1})\mathbf{G}(z) = \mathbf{I}$. Computing the coefficient of z^{-1}, we get $\mathbf{u}^T(\mathbf{I}+\mathbf{u}\mathbf{v}^T) = \mathbf{0}$, which implies $\mathbf{u}^T = \mathbf{v}^T$ and $\mathbf{u}^T\mathbf{u} = 1$. The non PU system $\mathbf{G}(z)$ is useful because it can generate degree-two PU building blocks for the new factorization theorem.

Lemma 4.5.1. *Inversion of a Simple Degree-One System:* The system $\mathbf{G}(z)$ over $GF(2)$ in (4.5.1) *always* has a FIR inverse. Its inverse is:

$$\mathbf{G}^{-1}(z) = \begin{cases} \mathbf{G}(z^{-1}), & \text{if } \mathbf{v}^T\mathbf{u} = 1; \\ \mathbf{G}(z), & \text{if } \mathbf{v}^T\mathbf{u} = 0. \end{cases} \tag{4.5.2}$$

■

The above lemma can be proved by direct substitution. It shows that in $GF(2)$ we can have a nontrivial system which is its own inverse, i.e., $\mathbf{G}(z)\mathbf{G}(z) = \mathbf{I}$.

4.5.1. Degree-Two PU Building Blocks

One useful special case of the system $\mathbf{G}(z)$ in (4.5.1) is when the vectors $\mathbf{u}$ and $\mathbf{v}$ satisfy

$$[\mathbf{u}\ \ \mathbf{v}]^T[\mathbf{u}\ \ \mathbf{v}] = \mathbf{J}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{4.5.3}$$

In this case, if we form the following cascade system:

$$\begin{aligned} \mathbf{K}(z) &= (\mathbf{I} + \mathbf{u}\mathbf{v}^T + z^{-1}\mathbf{u}\mathbf{v}^T)(\mathbf{I} + \mathbf{v}\mathbf{u}^T + z^{-1}\mathbf{v}\mathbf{u}^T) \\ &= \underbrace{\mathbf{I} + \mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T}_{\mathbf{k}(0)} + z^{-1}\underbrace{(\mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T)}_{\mathbf{k}(1)}, \end{aligned} \tag{4.5.4}$$

then it can be verified that $\mathbf{K}^T(z^{-1})\mathbf{K}(z) = \mathbf{I}$. So $\mathbf{K}(z)$ is PU even though each individual factor is not PU. From the second equality of (4.5.4), it is clear that $\mathbf{K}(z)$ remains the same if we interchange the

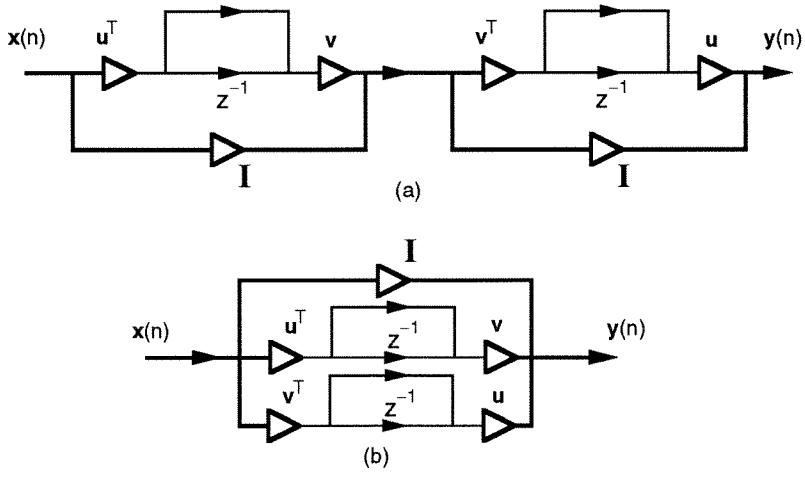

Fig. 4.5.1. (a) Cascade implementation of the degree-two PU system $\mathbf{K}(z)$; (b) parallel implementation of the degree-two PU system $\mathbf{K}(z)$. Here $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 0$, and $\mathbf{v}^T\mathbf{u} = 1$.

vectors $\mathbf{u}$ and $\mathbf{v}$. Therefore we can also write $\mathbf{K}(z)$ as $(\mathbf{I} + \mathbf{v}\mathbf{u}^T + z^{-1}\mathbf{v}\mathbf{u}^T)(\mathbf{I} + \mathbf{u}\mathbf{v}^T + z^{-1}\mathbf{u}\mathbf{v}^T)$. Using (4.5.4), we have the cascade and parallel implementations of $\mathbf{K}(z)$ as shown in Figs. 4.5.1 (a) and (b) respectively.

Basic Properties of $\mathbf{K}(z)$***:***

1. It is symmetric, i.e., $\mathbf{K}^T(z) = \mathbf{K}(z)$. The inverse system is given by $\mathbf{K}^{-1}(z) = \mathbf{K}(z^{-1})$.
2. Note that $\mathbf{k}(1)\mathbf{u} = \mathbf{u}$ and $\mathbf{k}(1)\mathbf{v} = \mathbf{v}$, so the range of $\mathbf{k}(1)$ has rank $= 2$, which implies the system $\mathbf{K}(z)$ has degree two. Hence we have $[\det \mathbf{K}(z)] = z^{-2}$.
3. $\mathbf{K}(z)$ cannot be factorized into building blocks of the form $\mathbf{D}(z)$ in (4.4.1). This can be seen by investigating the null space of the 0-th coefficient $\mathbf{k}(0) = \mathbf{I} + \mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T$. If $\mathbf{w}$ is a vector in the null space of $\mathbf{k}(0)$, then it must satisfy $\mathbf{w} = (\mathbf{v}^T\mathbf{w})\mathbf{u} + (\mathbf{u}^T\mathbf{w})\mathbf{v}$. This implies $\mathbf{w}$ has an even weight since it is a linear combination of two even weight vectors (see (4.4.10)). Using Lemma 4.4.2, we can conclude that $\mathbf{K}(z)$ cannot be written in terms of degree-one PU system $\mathbf{D}(z)$.
4. Let $\mathbf{K}_i(z) = \left[\mathbf{I} + \mathbf{u}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{u}_i^T + z^{-1}(\mathbf{u}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{u}_i^T)\right]$ for $0 \le i \le s-1$ be degree-two PU systems. Then it can be verified that the product $\mathbf{K}(z) = \mathbf{K}_0(z)\ldots\mathbf{K}_{s-1}(z)$ remains order-one if and only if the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are such that the matrix $\mathbf{W} = [\mathbf{u}_0\ \mathbf{v}_0\ \ldots\ \mathbf{u}_{s-1}\ \mathbf{v}_{s-1}]$ satisfies

$$\mathbf{W}^T\mathbf{W} = \mathcal{J}_{2s} = \begin{bmatrix} \mathbf{J}_2 & & & \\ & \mathbf{J}_2 & & \mathbf{0} \\ & & \mathbf{J}_2 & \\ & \mathbf{0} & & \ddots \\ & & & & \mathbf{J}_2 \end{bmatrix}, \tag{4.5.5}$$

where $\mathbf{J}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Furthermore if (4.5.5) is true, then the factors $\mathbf{K}_i(z)$ commute so that we can write the product as

$$\prod_{i=0}^{s-1}\mathbf{K}_i(z) = \mathbf{I} + \sum_{i=0}^{s-1}(\mathbf{u}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{u}_i^T) + z^{-1}\sum_{i=0}^{s-1}(\mathbf{u}_i\mathbf{v}_i^T + \mathbf{v}_i\mathbf{u}_i^T)$$

$$= \mathbf{I} + \mathbf{W}\mathcal{J}_{2s}\mathbf{W}^T + z^{-1}\mathbf{W}\mathcal{J}_{2s}\mathbf{W}^T, \tag{4.5.6}$$

where the $2s \times 2s$ matrix $\mathcal{J}_{2s}$ is as defined in (4.5.5).

4.5.2. Degree-Two Reduction Using $\mathbf{K}(z)$

In Section 4.4.2, we have given a procedure for the extraction of degree-one building block $\mathbf{D}(z)$. Suppose that we have factored out all the extractable degree-one building blocks of the form $\mathbf{D}(z)$ and $\mathbf{E}(z)$ is the remaining system which is unfactorizable in terms of $\mathbf{D}(z)$. Hence the null spaces of $\mathbf{e}(0)$ and $\mathbf{e}^T(0)$ do not contain any vector with an odd weight. Next we will provide an algorithm to extract the degree-two PU building block $\mathbf{K}(z)$ whenever it is possible.

Let $\{\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{s-1}\}$ be a set of independent vectors that span the null space of $\mathbf{e}(0)$. Since there is no degree-one building block, we have $\mathbf{v}_i^T\mathbf{v}_i = 0$ for all i. Suppose there is a pair of vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ such that $\mathbf{v}_i^T\mathbf{v}_j = 1$. Then the following system:

$$\mathbf{E}'(z) = \mathbf{E}(z)\Big(\mathbf{I} + \mathbf{v}_i\mathbf{v}_j^T + \mathbf{v}_j\mathbf{v}_i^T + z(\mathbf{v}_i\mathbf{v}_j^T + \mathbf{v}_j\mathbf{v}_i^T)\Big), \tag{4.5.7}$$

is a causal PU system with degree $\rho' = \rho - 2$, where ρ is the degree of the original system $\mathbf{E}(z)$. The causality of $\mathbf{E}'(z)$ follows from the fact that both $\mathbf{v}_i$ and $\mathbf{v}_j$ are in the null space of $\mathbf{e}(0)$. Since the vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ satisfy the condition (4.5.3), the anticausal system $\Big(\mathbf{I} + \mathbf{v}_i\mathbf{v}_j^T + \mathbf{v}_j\mathbf{v}_i^T + z(\mathbf{v}_i\mathbf{v}_j^T + \mathbf{v}_j\mathbf{v}_i^T)\Big)$ is the inverse of a degree-two PU system $\mathbf{K}(z)$. Therefore $\mathbf{E}'(z)$ is PU. Taking the determinant of (4.5.7), we get $[\det \mathbf{E}'(z)] = z^2 \cdot [\det \mathbf{E}(z)] = z^{-(\rho-2)}$. Since $\mathbf{E}'(z)$ is PU, its degree is equal to $\rho - 2$ (see Section 4.3.1). After rearranging (4.5.7), we get

$$\mathbf{E}(z) = \mathbf{E}'(z)\Big(\mathbf{I} + \mathbf{v}_i\mathbf{v}_j^T + \mathbf{v}_j\mathbf{v}_i^T + z^{-1}(\mathbf{v}_i\mathbf{v}_j^T + \mathbf{v}_j\mathbf{v}_i^T)\Big). \tag{4.5.8}$$

We have sucessfully extracted a degree-two PU building block from the right of $\mathbf{E}(z)$. Note that it is possible that we can extract the degree-one PU building block $\mathbf{D}(z)$ from the right hand side of the reduced PU system $\mathbf{E}'(z)$ (degree-one reduction from the left of $\mathbf{E}'(z)$ is impossible because degree-one reduction from the left of $\mathbf{E}(z)$ fails). Therefore after every degree-two reduction from the right, we must test if there is any degree-one building block. Similarly, if we can find a pair of vectors in the null space of $\mathbf{e}^T(0)$ that satisfies (4.5.3), then we can extract a degree-two factor $\mathbf{K}(z)$ from the left of $\mathbf{E}(z)$.

Example 4.5.1. *A PU system that is factorizable in terms of* $\mathbf{K}(z)$ *but not in terms of* $\mathbf{D}(z)$*:* Consider the PU system $\mathbf{G}(z)$ in (4.4.8) in Example 4.4.1. Let $M = 5$ so that $\mathbf{G}(z)$ can be written as:

$$\mathbf{G}(z) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} + z^{-1} \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}. \tag{4.5.9}$$

The PU system $\mathbf{G}(z)$ has degree equal to 4 as the rank of $\mathbf{g}(1) = 4$. Since the null space of $\mathbf{g}(0)$ consists of vectors with even weight only, degree-one reduction fails (by Lemma 4.4.2). However, one can verify

that $\mathbf{G}(z)$ can be written as a product of two degree-two PU factors $\mathbf{K}(z)$. One of such representations is given as:

$$\mathbf{G}(z) = \left[\mathbf{I} + \mathbf{u}_0\mathbf{v}_0^T + \mathbf{v}_0\mathbf{u}_0^T + z^{-1}(\mathbf{u}_0\mathbf{v}_0^T + \mathbf{v}_0\mathbf{u}_0^T)\right]\left[\mathbf{I} + \mathbf{u}_1\mathbf{v}_1^T + \mathbf{v}_1\mathbf{u}_1^T + z^{-1}(\mathbf{u}_1\mathbf{v}_1^T + \mathbf{v}_1\mathbf{u}_1^T)\right], \quad (4.5.10)$$

where the vectors are

$$\mathbf{u}_0 = \begin{bmatrix}1\\1\\0\\0\\0\end{bmatrix}, \quad \mathbf{v}_0 = \begin{bmatrix}0\\1\\1\\0\\0\end{bmatrix}, \quad \mathbf{u}_1 = \begin{bmatrix}0\\0\\0\\1\\1\end{bmatrix}, \quad \mathbf{v}_1 = \begin{bmatrix}1\\1\\1\\1\\0\end{bmatrix}. \quad (4.5.11)$$

Note that the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ in (4.5.11) satisfy (4.5.3). Moreover one can show that the ordering of $\mathbf{K}(z)$ in (4.5.10) is irrelevant because the two factors commute (see Property 4 of Section 4.5.1). Later we will see that in general it is true that all the factors (degree-one or degree-two) commute for the class of LOTs over $GF(2)$. ■

4.5.3. Non Completeness of $\mathrm{D}(z)$ and $\mathrm{K}(z)$

As we have seen in Example 4.5.1, PU systems that cannot be factorized in terms of $\mathbf{D}(z)$ can sometimes be expressed as a product of $\mathbf{K}(z)$. It is natural to ask if all PU systems can be represented as a product of $\mathbf{D}(z)$ and $\mathbf{K}(z)$. The answer is *no* in general. However we will see in the next section, the class of LOT over $GF(2)$ can always be factorized in terms of $\mathbf{D}(z)$ and $\mathbf{K}(z)$.

Most General Unfactorizable Degree-Two 2×2 ***PU Systems:*** It is shown in Appendix 5.A that the most general 2×2 PU system over $GF(2)$, that cannot be factorized in terms of the degree-one building block $\mathbf{D}(z)$, has the following form:

$$\mathbf{G}(z) = \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + z^{-1}\mathbf{g}(1) + z^{-2}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}, \quad (4.5.12)$$

where $\mathbf{g}(1) = \mathbf{G}(1)$ which equals to either the identity matrix $\mathbf{I}_2$ or the reversal matrix $\mathbf{J}_2$. It can be verified that the coefficients $\mathbf{g}(k)$ satisfy (4.3.2) so that $\mathbf{G}(z)$ is PU. The system $\mathbf{G}(z)$ has degree two because $[\det \mathbf{G}(z)] = z^{-2}$. Fig. 4.5.2 shows a minimal realization of $\mathbf{G}(z)$ when $\mathbf{g}(1) = \mathbf{I}_2$. Using Lemma 4.4.2, we know that $\mathbf{G}(z)$ cannot be factorized in terms of $\mathbf{D}(z)$ because the null spaces of $\mathbf{g}(0)$ and $\mathbf{g}^T(0)$ contain only one vector, namely $\begin{bmatrix}1\\1\end{bmatrix}$, which has an even weight. Moreover the system $\mathbf{G}(z)$ cannot be re-expressed in the form $\mathbf{K}(z)$. If it could, there would exist two vectors $\mathbf{u}$ and $\mathbf{v}$ in the null space of $\mathbf{g}(0)$ or $\mathbf{g}^T(0)$ such that $\mathbf{u}^T\mathbf{v} = 1$ (which is impossible as the null spaces contain only $\begin{bmatrix}1\\1\end{bmatrix}$). Therefore we conclude that $\mathbf{G}(z)$ is a PU system that cannot be written as a product of $\mathbf{D}(z)$ or $\mathbf{K}(z)$.

A Degree-Four Unfactorizable PU System: Consider the following 2×2 PU system:

$$\mathbf{G}'(z) = \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + z^{-2}\mathbf{I} + z^{-4}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}. \quad (4.5.13)$$

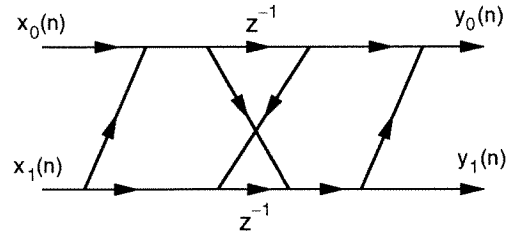

Fig. 4.5.2. Unfactorizable degree-two PU system in $GF(2)$.

The system $\mathbf{G}'(z)$ has degree 4. Both the degree-one reduction by $\mathbf{D}(z)$ and the degree-two reduction by $\mathbf{K}(z)$ are impossible as the null spaces of $\mathbf{g}(0)$ and $\mathbf{g}^T(0)$ contain only the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, which has an even weight. Moreover $\mathbf{G}'(z)$ cannot be written as a product of the system $\mathbf{G}(z)$ in (4.5.12) even though $\mathbf{G}(z)$ is the most general unfactorizable 2×2 PU system. To see this, assume that $\mathbf{G}'(z) = \mathbf{G}_1(z)\mathbf{G}_2(z)$, where $\mathbf{G}_i(z)$ are of the form as in (4.5.12). Comparing the 0-th coefficient, we have

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \mathbf{0}, \tag{4.5.14}$$

a contradiction! Therefore $\mathbf{G}'(z)$ cannot be expressed in terms of $\mathbf{D}(z)$, $\mathbf{K}(z)$ and $\mathbf{G}(z)$.

From the examples given in (4.5.12) and (4.5.13), we know that the PU building blocks $\mathbf{D}(z)$ and $\mathbf{K}(z)$ are *not complete*. For a general PU system $\mathbf{E}(z)$, we have the following test for unfactorizability:

Lemma 4.5.2. *Test of Degree-One and Degree-Two Factorizability in $GF(2)$:* Consider the PU system $\mathbf{E}(z) = \sum_{k=0}^{N} \mathbf{e}(k)z^{-k}$. Let $\{\mathbf{v}_0, \ldots, \mathbf{v}_{s-1}\}$ be *any* basis that spans the null space of $\mathbf{e}(0)$ and $\{\mathbf{u}_0, \ldots, \mathbf{u}_{s-1}\}$ be *any* basis that spans the null space of $\mathbf{e}^T(0)$. Then both the degree-one and degree-two reductions fail if and only if $\mathbf{v}_i^T\mathbf{v}_j = 0$ and $\mathbf{u}_i^T\mathbf{u}_j = 0$ for all i, j. ■

The above lemma can be proved in a straightforward manner. Note that the number of independent vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are the same, as the null spaces of $\mathbf{e}(0)$ and $\mathbf{e}^T(0)$ have the same dimension. Also note that if there is a basis for the null space of $\mathbf{e}(0)$ that satisfies the condition in the above lemma, so are all the other bases because of (4.4.10). Therefore it is sufficient to check one basis. Letting the matrices $\mathbf{V} = [\mathbf{v}_0, \ldots, \mathbf{v}_{s-1}]$ and $\mathbf{U} = [\mathbf{u}_0, \ldots, \mathbf{u}_{s-1}]$, then the condition in Lemma 4.5.2 can be re-stated as $\mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{U} = \mathbf{0}$.

4.6. LAPPED ORTHOGONAL TRANSFORMS OVER GF(2)

In this section, we consider the following $M \times M$ first-order system over $GF(2)$

$$\mathbf{E}(z) = \mathbf{e}(0) + \mathbf{e}(1)z^{-1}. \tag{4.6.1}$$

The rank ρ of the matrix $\mathbf{e}(1)$ is the degree of the system, and $\rho \leq M$. If $\mathbf{E}(z)$ is a PU system, then we call the system $\mathbf{E}(z)$ a lapped orthogonal transform (LOT) over $GF(2)$. The coefficients of the LOT in (4.6.1) should satisfy the PU condition in (4.3.2) which we re-state as follows:

$$\mathbf{e}^T(0)\mathbf{e}(1) = \mathbf{0}, \tag{4.6.2a}$$

$$\mathbf{e}^T(0)\mathbf{e}(0) + \mathbf{e}^T(1)\mathbf{e}(1) = \mathbf{I}. \tag{4.6.2b}$$

In the following, we will first give a minimal parameterization of LOTs over $GF(2)$ and then show the factorization theorem.

4.6.1. Minimal Characterization of LOT

In the case of real or complex field, it is well-known [Vai93, Mal92] that all LOTs of degree ρ can be parameterized by a set of ρ orthonormal vectors and a unitary matrix. We can capture all LOTs by varying the ρ orthonormal vectors and the unitary matrix. Associated with this minimal parameterization, there is an implementation that will structurally guarantee the LOT properties [Vai93, Mal92]. In this section, we will derive a similar result for the $GF(2)$ case.

Theorem 4.6.1. *Minimal Characterization of LOT over $GF(2)$:* In $GF(2)$, the $M \times M$ system $\mathbf{E}(z) = \mathbf{e}(0) + \mathbf{e}(1)z^{-1}$ is a LOT with degree ρ if and only if there is a $M \times \rho$ matrix $\mathbf{U}_\rho = [\mathbf{u}_0\ \mathbf{u}_1\ \ldots\ \mathbf{u}_{\rho-1}]$ such that $\mathbf{U}_\rho^T\mathbf{U}_\rho$ is invertible and

$$\mathbf{E}(z) = \mathbf{E}(1)\big[\mathbf{I} + \mathbf{U}_\rho\mathbf{L}^{-1}\mathbf{U}_\rho^T + z^{-1}\mathbf{U}_\rho\mathbf{L}^{-1}\mathbf{U}_\rho^T\big], \tag{4.6.3}$$

where $\mathbf{L} = \mathbf{U}_\rho^T\mathbf{U}_\rho$ and $\mathbf{E}(1)$ is unitary. ■

Proof: The "if" part can be proved by directly substituting the expression in (4.6.3) into the product $\mathbf{E}^T(z^{-1})\mathbf{E}(z)$. One can verify that $\mathbf{E}^T(z^{-1})\mathbf{E}(z) = \mathbf{I}$. To show the "only if" part, assume $\mathbf{E}(z)$ is LOT with degree ρ. As $\mathbf{E}(z)$ is PU, it can always be rewritten as:

$$\mathbf{E}(z) = \mathbf{E}(1)\Big[\mathbf{I} + \mathbf{W} + z^{-1}\mathbf{W}\Big], \tag{4.6.4}$$

where $\mathbf{E}(1)$ is unitary and $\mathbf{I} + \mathbf{W} + z^{-1}\mathbf{W}$ is PU. Since $\mathbf{E}(z)$ has degree ρ, the matrix $\mathbf{e}(1) = \mathbf{E}(1)\mathbf{W}$ has rank ρ. Thus there are independent vectors $\mathbf{u}_i$ and independent vectors $\mathbf{v}_i$ for $i = 0,\ 1,\ \ldots,\ \rho - 1$ such that $\mathbf{W} = [\mathbf{u}_0\ \mathbf{u}_1\ \ldots\ \mathbf{u}_{\rho-1}][\mathbf{v}_0\ \mathbf{v}_1\ \ldots\ \mathbf{v}_{\rho-1}]^T$. Letting $\mathbf{U}_\rho = [\mathbf{u}_0\ \mathbf{u}_1\ \ldots\ \mathbf{u}_{\rho-1}]$ and $\mathbf{V}_\rho = [\mathbf{v}_0\ \mathbf{v}_1\ \ldots\ \mathbf{v}_{\rho-1}]$, we can rewrite (4.6.4) as

$$\mathbf{E}(z) = \mathbf{E}(1)\big[\mathbf{I} + \mathbf{U}_\rho\mathbf{V}_\rho^T + z^{-1}\mathbf{U}_\rho\mathbf{V}_\rho^T\big]. \tag{4.6.5}$$

Substituting the coefficients into (4.6.2a) and simplifying the result, we get

$$(\mathbf{I} + \mathbf{V}_\rho\mathbf{U}_\rho^T)\mathbf{U}_\rho\mathbf{V}_\rho^T = \mathbf{0}. \tag{4.6.6}$$

As the vectors $\mathbf{v}_i$ are independent, we can conclude from (4.6.6) that

$$\mathbf{U}_\rho = \mathbf{V}_\rho\mathbf{U}_\rho^T\mathbf{U}_\rho. \tag{4.6.7}$$

The above equation has two implications: (i) The vector $\mathbf{u}_i$ is a linear combination of $\mathbf{v}_i$; (ii) The $\rho \times \rho$ matrix $\mathbf{U}_\rho^T\mathbf{U}_\rho$ is invertible as both $\mathbf{U}_\rho$ and $\mathbf{V}_\rho$ have rank equal to ρ. Hence we can write $\mathbf{V}_\rho = \mathbf{U}_\rho\mathbf{L}^{-1}$, where $\mathbf{L} = \mathbf{U}_\rho^T\mathbf{U}_\rho$. Substituting $\mathbf{V}_\rho = \mathbf{U}_\rho\mathbf{L}^{-1}$ into (4.6.5), we immediately get (4.6.3). ■

Note that in the proof of "only if" part, we have not used the second PU condition of (4.6.2b). One can verify that the choice of $\mathbf{V} = \mathbf{U}\mathbf{L}^{-1}$ will automatically satisfy (4.6.2b). From Theorem 4.6.1, we have the implementation of LOT as in Fig. 4.6.1. Note that the matrix $\mathbf{L}$ in Theorem 4.6.1 functions like a "normalization" matrix. In the special case of $\mathbf{L} = \mathbf{I}_\rho$, we can write $\mathbf{E}(z)$ as

$$\mathbf{E}(z) = \mathbf{E}(1)\Big[\mathbf{I} + \mathbf{U}\mathbf{U}^T + z^{-1}\mathbf{U}\mathbf{U}^T\Big], \tag{4.6.8}$$

where the matrix $\mathbf{U}^T\mathbf{U} = \mathbf{I}_\rho$. Using Property 3 of $\mathbf{D}(z)$ in Section 4.4.1, we conclude that in the special case of $\mathbf{L} = \mathbf{I}$, the LOT in (4.6.3) can be written as a product of $\mathbf{D}(z)$:

$$\mathbf{E}(z) = \mathbf{E}(1)\prod_{i=0}^{\rho-1}\Big[\mathbf{I} + \mathbf{u}_i\mathbf{u}_i^T + z^{-1}\mathbf{u}_i\mathbf{u}_i^T\Big]. \tag{4.6.9}$$

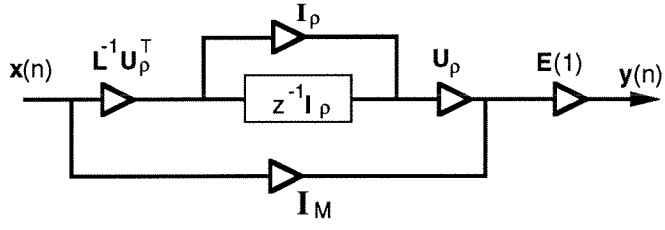

Fig. 4.6.1. Minimal characterization of a LOT with degree ρ. Here $\mathbf{E}(1)$ is a unitary matrix and $\mathbf{L} = \mathbf{U}_\rho^T\mathbf{U}_\rho$.

Remarks:

1. In Theorem 4.6.1, the vectors $\mathbf{u}_i$ cannot be arbitrary independent vectors. They should be chosen such that the matrix $\mathbf{U}_\rho^T\mathbf{U}_\rho$ is invertible. The subtlety is that in finite fields, the independence of $\mathbf{u}_i$ does not always guarantee the invertibility of $\mathbf{U}_\rho^T\mathbf{U}_\rho$. Unlike the real or complex field, the matrices $\mathbf{U}_\rho$ and $\mathbf{U}_\rho^T\mathbf{U}_\rho$ may not have the same rank in finite fields. One such counter example is the matrix $\mathbf{U}_2^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$ in $GF(2)$.
2. In the real or complex field, the matrix $\mathbf{L}$ in (4.6.3) can always be decomposed as $\mathbf{Q}^T\mathbf{Q}$ for some positive definite $\rho \times \rho$ matrix $\mathbf{Q}$. This is the same as saying that the vectors $\mathbf{u}_i$ can always be orthonormalized in the cases of real or complex field.

4.6.2. Complete Factorization of LOT

Consider the first-order system $\mathbf{E}(z)$ in (4.6.1). Assume that $\mathbf{E}(z)$ is a LOT with degree ρ so that the conditions in (4.6.2) are met. To avoid trivial cases, we assume $1 \leq \rho \leq M - 1$. From (4.6.2a), we know that the column vectors in the matrix $\mathbf{e}(1)$ is in the null space of $\mathbf{e}^T(0)$. Suppose that we cannot extract either a degree-one or a degree-two building block from $\mathbf{E}(z)$. By Lemma 4.5.2, it is necessary

that $\mathbf{e}^T(1)\mathbf{e}(1) = \mathbf{0}$, which implies $\mathbf{e}^T(0)\mathbf{e}(0) = \mathbf{I}$ from (4.6.2b). Hence $\mathbf{e}(0)$ is unitary and invertible. Inverting $\mathbf{e}^T(0)$ of (4.6.2a), we have $\mathbf{e}(1) = \mathbf{0}$, which implies $\mathbf{E}(z)$ is a constant unitary matrix. Therefore we conclude that $\mathbf{e}^T(1)\mathbf{e}(1) \neq 0$ if $\rho > 0$. Using Lemma 5.2, we know that we can always extract either the factor $\mathbf{D}(z)$ or the factor $\mathbf{K}(z)$ from $\mathbf{E}(z)$ if its degree $\rho > 0$. After the degree reduction, we will have a new LOT system $\mathbf{E}'(z)$ with degree $\rho' < \rho$. We can further reduce the degree of $\mathbf{E}'(z)$ by extracting a degree-one or degree-two building blocks. Continuing the degree-reduction process, we will finally arrive at a constant unitary matrix. Summarizing the result, we have proved:

Theorem 4.6.2. *Factorization of LOT over GF(2):* All LOTs over $GF(2)$ are factorizable in terms of $\mathbf{D}(z)$ and $\mathbf{K}(z)$. ■

Since the LOTs have order one, the vectors in the factors of $\mathbf{D}(z)$ and $\mathbf{K}(z)$ have to satisfy some constraints so that the product of these first-order building blocks remains a first-order system. Let $\mathbf{w}_i$ be the vectors in $\mathbf{D}_i(z)$ and $(\mathbf{u}_j, \mathbf{v}_j)$ be the vectors in $\mathbf{K}_j(z)$. Then we have

1. The product of $\mathbf{D}_0(z)$ and $\mathbf{D}_1(z)$ has order one if and only if the vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ are such that the matrix $\mathbf{W} = [\mathbf{w}_0\ \mathbf{w}_1]$ satisfies (see Section 4.4.1)

$$\mathbf{W}^T\mathbf{W} = \mathbf{I}_2, \tag{4.6.10}$$

   where $\mathbf{I}_2$ is a 2×2 identity matrix. Moreover $\mathbf{D}_0(z)\mathbf{D}_1(z) = \mathbf{D}_1(z)\mathbf{D}_0(z)$ in this case.

2. The product of $\mathbf{K}_0(z)$ and $\mathbf{K}_1(z)$ has order one if and only if the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are such that the matrix $\mathbf{C} = [\mathbf{u}_0\ \mathbf{v}_0\ \mathbf{u}_1\ \mathbf{v}_1]$ satisfies (see Section 4.5.1)

$$\mathbf{C}^T\mathbf{C} = \begin{bmatrix} \mathbf{J}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{bmatrix}, \tag{4.6.11}$$

   where the matrix $\mathbf{J}_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Moreover $\mathbf{K}_0(z)\mathbf{K}_1(z) = \mathbf{K}_1(z)\mathbf{K}_0(z)$ in this case.

3. The product of $\mathbf{D}_0(z)$ and $\mathbf{K}_0(z)$ has order one if and only if the vector $\mathbf{w}_0$ is such that $\mathbf{w}_0^T\mathbf{u}_0 = 0$ and $\mathbf{w}_0^T\mathbf{v}_0 = 0$. Moreover $\mathbf{D}_0(z)\mathbf{K}_0(z) = \mathbf{K}_0(z)\mathbf{D}_0(z)$ in this case.

Combining the above results with Theorem 4.6.2, we have:

Theorem 4.6.3. *Cascade Form for LOT over GF(2):* The system $\mathbf{E}(z)$ in (4.6.1) is a LOT with degree ρ if and only if it can be written as

$$\mathbf{E}(z) = \mathbf{E}(1) \prod_{i=0}^{\rho_1-1} \left[\mathbf{I} + \mathbf{w}_i\mathbf{w}_i^T + z^{-1}\mathbf{w}_i\mathbf{w}_i^T\right] \prod_{j=0}^{\rho_2-1} \left[\mathbf{I} + \mathbf{u}_j\mathbf{v}_j^T + \mathbf{v}_j\mathbf{u}_j^T + z^{-1}(\mathbf{u}_j\mathbf{v}_j^T + \mathbf{v}_j\mathbf{u}_j^T)\right], \tag{4.6.12}$$

where $\rho = \rho_1 + 2\rho_2$, $\mathbf{E}(1)$ is some unitary matrix, and the vectors are such that the $M \times \rho$ matrix $\mathbf{C} = [\mathbf{w}_0\ \ldots\ \mathbf{w}_{\rho_1-1}\ \mathbf{u}_0\ \mathbf{v}_0\ \ldots\ \mathbf{u}_{\rho_2-1}\ \mathbf{v}_{\rho_2-1}]$ satisfies

$$\mathbf{C}^T\mathbf{C} = \begin{bmatrix} \mathbf{I}_{\rho_1} & & & \\ & \mathbf{J}_2 & & \mathbf{0} \\ & & \mathbf{J}_2 & \\ & \mathbf{0} & & \ddots \\ & & & & \mathbf{J}_2 \end{bmatrix}. \tag{4.6.13}$$

■

Remark: Recall the "normalization" matrix $\mathbf{L}$ in Theorem 4.6.2. Using the result in Theorem 4.6.3, we conclude that there is always a $\mathbf{L}$ of the form (4.6.13).

4.7. STATE-SPACE MANIFESTATION OF PU SYSTEMS

Consider the $M \times M$ causal FIR system $\mathbf{E}(z) = \sum_{i=0}^{N} \mathbf{e}(i) z^{-i}$ in $GF(2)$. Let $\mathbf{x}(n)$ and $\mathbf{y}(n)$ be the input and the output of $\mathbf{E}(z)$ respectively. Then given any structure for $\mathbf{E}(z)$, we can write down two equations of the form:

$$\begin{aligned} \mathbf{s}(n+1) &= \mathbf{A}\mathbf{s}(n) + \mathbf{B}\mathbf{x}(n), \quad \text{(state eqn.)} \\ \mathbf{y}(n) &= \mathbf{C}\mathbf{s}(n) + \mathbf{D}\mathbf{x}(n), \quad \text{(output eqn.)} \end{aligned} \tag{4.7.1}$$

where $\mathbf{A}$ is $\rho \times \rho$, $\mathbf{B}$ is $\rho \times M$, $\mathbf{C}$ is $M \times \rho$ and $\mathbf{D}$ is $M \times M$. The vector $\mathbf{s}(n)$ is called the state vector which consists of the output of delay elements. If the dimension of the matrix $\mathbf{A}$ is the smallest possible, then the structure is said to be *minimal* and ρ is called the McMillan degree of the system. As shown in [Vai93, Vai95c], the McMillan degree of a PU system is equal to the degree of its determinant. Given $(\mathbf{A},\ \mathbf{B},\ \mathbf{C},\ \mathbf{D})$ for a structure, the $(M+\rho) \times (M+\rho)$ matrix

$$\mathcal{R} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \tag{4.7.2}$$

is called the *realization matrix* of the structure. The state-space description (4.7.1) of MIMO systems in the real or complex case has been studied extensively in the past [Vai93, Kai80]. In finite fields, the concept of minimality has been introduced and studied in the area of coding theory [For70, McE95]. Analogous to the complex case, we can define the concepts of complete reachability (cr), complete observability (co) and minimality in finite fields. It can be verified that the following properties continue to hold.

1. A structure is cr if and only if the following matrix $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ has rank ρ in $GF(q)$.

$$\mathcal{S}_{\mathbf{A},\mathbf{B}} = [\mathbf{B}\ \mathbf{A}\mathbf{B}\ \ldots\ \mathbf{A}^{\rho-1}\mathbf{B}]. \tag{4.7.3}$$

2. A structure is co if and only if the following matrix $\mathcal{O}_{\mathbf{A},\mathbf{C}}$ has rank ρ in $GF(q)$.

$$\mathcal{O}_{\mathbf{A},\mathbf{C}} = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^{\rho-1} \end{bmatrix}. \tag{4.7.4}$$

3. A structure is minimal if and only if it is both cr and co.
4. The impulse responses $\mathbf{e}(i)$ are related to $(\mathbf{A},\ \mathbf{B},\ \mathbf{C},\ \mathbf{D})$ as:

$$\mathbf{e}(0) = \mathbf{D}, \quad \mathbf{e}(i) = \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}, \quad for \ \ 1 \le i \le N. \tag{4.7.5}$$

Note that in $GF(q)$, the Cayley-Hamilton theorem continues to hold [Hor85]. For any $\rho \times \rho$ matrix $\mathbf{A}$, its power $\mathbf{A}^\rho$ is a linear combination of $\mathbf{A}^i$ for $0 \leq i \leq \rho - 1$. That means, if the matrix $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ in (4.7.3) does not have rank ρ, then adding more columns of the form $\mathbf{A}^j\mathbf{B}$ for $j \geq \rho$ will not increase the rank. Therefore providing more inputs will not help the controllability of the state. Similarly for the observability.

Example 4.7.1. *Realization Matrices of* $\mathbf{D}(z)$ *and* $\mathbf{K}(z)$*:*

(i) Consider Fig. 4.4.1. The realization matrix of the structure for $\mathbf{D}(z)$ in Fig. 4.4.1 is

$$\mathcal{R} = \begin{bmatrix} 0 & \mathbf{v}^T \\ \mathbf{v} & \mathbf{I} + \mathbf{v}\mathbf{v}^T \end{bmatrix}. \tag{4.7.6}$$

One can verify that $\mathcal{R}^T\mathcal{R} = \mathbf{I}$ so that $\mathcal{R}$ is unitary. It can be shown that the realization matrix for a cascade of $\mathbf{D}(z)$ is also unitary.

(ii) Consider Fig. 4.5.1(b). The realization matrix of the structure for $\mathbf{K}(z)$ in Fig. 4.5.1(b) is

$$\mathcal{R} = \begin{bmatrix} \mathbf{0} & \begin{bmatrix} \mathbf{v}^T \\ \mathbf{u}^T \end{bmatrix} \\ [\mathbf{u} \quad \mathbf{v}] & \mathbf{I} + \mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T \end{bmatrix}. \tag{4.7.7}$$

One can verify that $\mathcal{R}^T\mathcal{R} \neq \mathbf{I}$ so the realization matrix is not unitary. In this case, $\mathcal{S}_{\mathbf{A},\mathbf{B}} = [\mathbf{B}\ \mathbf{B}\mathbf{A}] = \begin{bmatrix} \mathbf{v}^T & 0^T \\ \mathbf{u}^T & 0^T \end{bmatrix}$ and $\mathcal{O}_{\mathbf{A},\mathbf{C}} = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \end{bmatrix} = \begin{bmatrix} \mathbf{u} & \mathbf{v} \\ 0 & 0 \end{bmatrix}$. Since $\mathbf{u}$ and $\mathbf{v}$ are independent, both $\mathcal{S}_{\mathbf{A},\mathbf{B}}$ and $\mathcal{O}_{\mathbf{A},\mathbf{C}}$ have rank two. Thus the structure in Fig. 4.5.1(b) is minimal. ■

Since a cascade of minimal structures is also minimal [Kai80], we conclude that the implementation based on cascade of $\mathbf{D}(z)$ and $\mathbf{K}(z)$ is minimal. In particular, the factorization of LOT given in Theorem 4.6.3 is minimal. Moreover, the realization matrix $\mathcal{R}$ of $\mathbf{D}(z)$ given in (4.7.6) is unitary. Therefore a cascade of $\mathbf{D}(z)$ also has a unitary realization matrix. On the other hand, the realization matrix for $\mathbf{K}(z)$ given in (4.7.7) is not unitary. In fact, later we will show that there does not exist any unitary realization matrix for $\mathbf{K}(z)$. Even though PU systems in $GF(2)$ may not have a unitary realization matrix, the following is true:

Lemma 4.7.1. Consider the causal FIR system $\mathbf{E}(z) = \sum_{i=0}^{N} \mathbf{e}(i)z^{-i}$ in $GF(2)$. If there is a minimal implementation with a unitary realization matrix $\mathcal{R}$, then $\mathbf{E}(z)$ is PU. ■

Proof: Assume that the initial state $\mathbf{s}(n_0) = \mathbf{0}$. Let $\mathbf{x}_0(n)$ and $\mathbf{x}_1(n)$ be two arbitrary finite-length inputs such that the corresponding outputs $\mathbf{y}_0(n)$, $\mathbf{y}_1(n)$ and the state vector $\mathbf{s}(n)$ are zero for $n > K$, for some finite K. Using the unitariness of $\mathcal{R}$, one can show that

$$\sum_{n=n_0}^{K} \mathbf{y}_0^T(n)\mathbf{y}_1(n) = \sum_{n=n_0}^{K} \mathbf{x}_0^T(n)\mathbf{x}_1(n). \tag{4.7.8}$$

Since (4.7.8) holds for any choice of $\mathbf{x}_0(n)$ and $\mathbf{x}_1(n)$, we conclude from Lemma 4.3.1 that $\mathbf{E}(z)$ is PU. ■

One natural question is to ask if the converse of Lemma 4.7.1 is true. The answer is yes when we are dealing with real or complex case [Vai93]. It is shown that in the real or complex case, a system is PU if

and only if there is an implementation with unitary realization matrix. However in $GF(2)$ the converse of Lemma 4.7.1 is not necessarily true as we will see in the following example:

Realization Matrices of $\mathbf{K}(z)$***:*** One minimal realization matrix $\mathcal{R}$ of $\mathbf{K}(z)$ is given in (4.7.7). Since the realization ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$) is minimal, any other minimal realization ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, $\mathbf{D}$) is related to ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$) as follows [Vai93, Kai80]:

$$\mathbf{a} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \quad \mathbf{b} = \mathbf{T}^{-1}\mathbf{B}, \quad \mathbf{c} = \mathbf{C}\mathbf{T}, \tag{4.7.9}$$

for some nonsingular matrix $\mathbf{T}$ in $GF(2)$. If there is a unitary realization for $\mathbf{K}(z)$, there will exist a 2×2 nonsingular matrix $\mathbf{T}$ such that

$$\mathcal{R}' = \begin{bmatrix} \mathbf{0} & \mathbf{T}^{-1}\begin{bmatrix}\mathbf{v}^T \\ \mathbf{u}^T\end{bmatrix} \\ [\mathbf{u} \quad \mathbf{v}]\,\mathbf{T} & \mathbf{I} + \mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T \end{bmatrix} \tag{4.7.10}$$

is unitary. Computing the product $\mathcal{R}'^T\mathcal{R}'$ and equating to the identity, we get

$$\mathbf{T}^T\mathbf{J}_2\mathbf{T} = \mathbf{I}_2, \tag{4.7.11}$$

where $\mathbf{J}_2 = \begin{bmatrix}0 & 1 \\ 1 & 0\end{bmatrix}$. Eqn. (4.7.11) implies $\mathbf{J}_2 = (\mathbf{T}^{-1})^T\mathbf{T}^{-1}$, which is not possible (see Appendix 4.B). Therefore we conclude that the PU system $\mathbf{K}(z)$ does not have a unitary realization matrix.

4.8. UNITARY MATRICES AND PU SYSTEMS OVER GF(q)

In this section, we will generalize the theory developed earlier to the case of $GF(q)$ for any prime number $q > 2$. While many results in $GF(2)$ case can be easily extended to the case of $GF(q)$, there are some exceptions, which we first point out.

4.8.1. Unitary Matrices over $GF(q)$

Let $\mathbf{A}$ be a matrix with elements in $GF(q)$ for some prime $q > 2$. In $GF(q)$, there are a number of properties different from those in $GF(2)$. In particular, the condition that $\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u} = \mathbf{u}^T\mathbf{u}$ for all $\mathbf{u}$ is sufficient to ensure the unitariness of $\mathbf{A}$ in $GF(q)$ for $q > 2$. To be more precise, we have:

Fact 4.8.1. In $GF(q)$ for some prime $q > 2$, $\mathbf{A}$ is unitary if and only if $\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u} = \mathbf{u}^T\mathbf{u}$ for all possible vectors $\mathbf{u}$ in $GF(q)$. ■

Proof: The "only if" part is clear. To show the "if" part, assume that $\mathbf{u}^T\mathbf{u} = \mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u}$. Substituting $\mathbf{B} = \mathbf{A}^T\mathbf{A}$ into (4.2.2), we get

$$\sum_l u_l^2 = \sum_l u_l^2 b_{ll} + 2\sum_{i>j} u_i u_j b_{ij}, \tag{4.8.1}$$

where we have used the fact that $\mathbf{B} = \mathbf{A}^T\mathbf{A}$ is symmetric. Letting $\mathbf{u}$ to be the unit vector $\mathbf{e}_i$, we get $b_{ll} = 1$ from (4.8.1). Using $b_{ll} = 1$, (4.8.1) can be rewritten as $2\sum_{i>j} u_i u_j b_{ij} = 0$. Now if we choose $\mathbf{u} = \mathbf{e}_{i_0} + \mathbf{e}_{j_0}$ for some $i_0 > j_0$, we get $2b_{i_0 j_0} = 0$ which implies $b_{i_0 j_0} = 0$ (as 2 is coprime to q). So $\mathbf{B} = \mathbf{A}^T\mathbf{A} = \mathbf{I}$. ■

Recall from Fact 4.2.3 that in $GF(2)$ none of the columns or rows of a unitary matrix can have all elements equal to 1. The same is not true for unitary matrices in $GF(q)$ for $q > 2$. For example, the following matrix is unitary in $GF(5)$.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 & 2 & 4 & 0 \\ 1 & 1 & 3 & 3 & 4 & 0 \\ 1 & 1 & 0 & 0 & 2 & 0 \\ 1 & 4 & 2 & 3 & 0 & 4 \\ 1 & 4 & 3 & 2 & 0 & 4 \\ 1 & 4 & 0 & 0 & 0 & 2 \end{bmatrix}. \tag{4.8.2}$$

In Section 4.2.2 we have seen that the factorizability of unitary matrices in $GF(2)$ depends on Fact 4.2.3. In $GF(q)$, even though a result similar to Fact 4.2.3 is no longer true, we will see later that all unitary matrices in $GF(q)$ are still factorizable.

Householder-Like Transformation in $GF(q)$***:*** Recall from Section 4.2.2 that for the factorization of unitary matrices in $GF(2)$, we have used the building blocks of the form $[\mathbf{I}+\mathbf{u}\mathbf{u}^T]$, where $\mathbf{u}^T\mathbf{u} = 0$. In the $GF(q)$ case, we will make use of the following building block:

$$\mathbf{U} = \mathbf{I} - 2l_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^T, \tag{4.8.3}$$

where $\mathbf{u}$ is any vector with $l_{\mathbf{u}} = \mathbf{u}^T\mathbf{u} \neq 0$ so that l^{-1} exists. One can verify that $\mathbf{U}$ is unitary and it is its own inverse. Note that unlike the complex field, $l_{\mathbf{u}}$ may not be the square of some number in $GF(q)$. Hence it is not always possible to "normalize" a nonzero vector $\mathbf{u}$ in $GF(q)$ such that $l_{\mathbf{u}} = 1$. One such example is the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ in $GF(3)$. The Householder matrix in (4.8.3) has a very useful property. Given any two vectors $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x}^T\mathbf{x} = \mathbf{y}^T\mathbf{y}$ and $(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) \neq 0$, the Householder matrix in (4.8.3) with $\mathbf{u} = \mathbf{x}-\mathbf{y}$ transforms the vector $\mathbf{x}$ into the vector $\mathbf{y}$. More precisely, we have $\mathbf{U}\mathbf{x} = \mathbf{y}$, where $\mathbf{u} = \mathbf{x}-\mathbf{y}$. Using this transformation property of Householder matrix, we can prove the following lemma:

Lemma 4.8.1. Let $\mathbf{A}$ be $M \times M$ unitary over $GF(q)$ for some prime $q > 2$ and let $A_{00} \neq 1$. Define the vector $\mathbf{u} = \mathbf{e}_0 - \mathbf{a}_0$ where $\mathbf{a}_0$ is the 0-th column of $\mathbf{A}$. Then $l_{\mathbf{u}} = \mathbf{u}^T\mathbf{u} \neq 0$, and

$$\mathbf{A} = (\mathbf{I} - 2l_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^T)\begin{bmatrix} 1 & 0 \\ 0 & \mathbf{B} \end{bmatrix}, \tag{4.8.4}$$

where $\mathbf{B}$ is $(M-1) \times (M-1)$ unitary. ■

Proof: As $A_{00} \neq 1$, we have $l_{\mathbf{u}} = (\mathbf{e}_0 - \mathbf{a}_0)^T(\mathbf{e}_0 - \mathbf{a}_0) = 2 - 2A_{00} \neq 0$. So we can form the unitary matrix $\mathbf{U}$ given in (4.8.3). As we mentioned before, the matrix $\mathbf{U}$ has the property that $\mathbf{U}\mathbf{a}_0 = \mathbf{e}_0$. Therefore we have

$$\mathbf{U}\mathbf{A} = \begin{bmatrix} 1 & \mathbf{v}^T \\ 0 & \mathbf{B} \end{bmatrix}. \tag{4.8.5}$$

The matrix $\mathbf{U}\mathbf{A}$ is unitary as both $\mathbf{A}$ and $\mathbf{U}$ are unitary. Thus $\mathbf{v} = \mathbf{0}$ and $\mathbf{B}$ is unitary. ■

With Lemma 4.8.1, we are ready to prove the factorization theorem for the unitary matrix $\mathbf{A}$ in $GF(q)$. The problem to be solved is, given any unitary matrix $\mathbf{A}$, how to avoid the case where $A_{00} = 1$.

This can be avoided by using both column permutation $\mathbf{P}_{col}$ and row permutation $\mathbf{P}_{row}$. Given any $M \times M$ unitary matrix $\mathbf{A}$ with $M > 1$, there is always an element $A_{ij} \neq 1$. So we can find $\mathbf{P}_{col}$ and $\mathbf{P}_{row}$ such that $\mathbf{A}' = \mathbf{P}_{row}\mathbf{A}\mathbf{P}_{col}$ with $A'_{00} = A_{ij} \neq 1$. Then Lemma 4.8.1 can be applied to $\mathbf{A}'$ and we can write $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{P}_{row}(\mathbf{I} - 2l_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^T)\begin{bmatrix} 1 & 0 \\ 0 & \mathbf{B} \end{bmatrix}\mathbf{P}_{col}, \tag{4.8.6}$$

for some vector $\mathbf{u}$ with $l_{\mathbf{u}} \neq 0$. We can continue the above process and arrive at the following result:

Theorem 4.8.1. *Factorization Unitary Matrices over $GF(q)$:* An $M \times M$ matrix $\mathbf{A}$ over $GF(q)$ (where q is a prime > 2) is unitary if and only if it can be factorized as:

$$\mathbf{A} = \mathbf{P}_M\mathbf{U}_M\mathbf{P}_{M-1}\cdots\mathbf{P}_3\mathbf{U}_3\mathbf{P}_2\mathbf{U}_2\mathbf{P}_1, \tag{4.8.7}$$

where $\mathbf{U}_k$ are as in (4.8.3) and $\mathbf{P}_k$ are permutations of the identity matrix. ■

4.8.2. PU Matrices over $GF(q)$

As we have seen in the previous discussion, many properties of unitary matrices in $GF(q)$ are different from those in $GF(2)$. In this section, we will extend the results of PU matrices in $GF(2)$ derived in Sections 4.3–4.7 to the $GF(q)$ case and point out the differences between these two cases.

Given any vector $\mathbf{v}$ in $GF(q)$ with $l_{\mathbf{v}} = \mathbf{v}^T\mathbf{v} \neq 0$, we form the following degree-one system:

$$\mathbf{D}_q(z) = \mathbf{I} - l_{\mathbf{v}}^{-1}\mathbf{v}\mathbf{v}^T + z^{-1}l_{\mathbf{v}}^{-1}\mathbf{v}\mathbf{v}^T. \tag{4.8.8}$$

Note that $\mathbf{D}_q(z)$ in (4.8.8) is slightly different from the $GF(2)$ degree-one building block $\mathbf{D}(z)$ in (4.4.1). It can be verified that $\mathbf{D}_q^T(z^{-1})\mathbf{D}_q(z) = \mathbf{I}$. So $\mathbf{D}_q(z)$ is PU. For the building block $\mathbf{D}_q(z)$ in (4.8.8), it can be shown that all the properties mentioned in Section 4.4.1 continue to hold. In particular, any $M \times M$ causal FIR degree-one PU system over $GF(q)$ can be written as

$$\mathbf{E}(z) = (\mathbf{I} - l_{\mathbf{v}}^{-1}\mathbf{v}\mathbf{v}^T + z^{-1}l_{\mathbf{v}}^{-1}\mathbf{v}\mathbf{v}^T)\mathbf{E}(1), \tag{4.8.9}$$

for some vector $\mathbf{v}$ such that $l_{\mathbf{v}} = \mathbf{v}^T\mathbf{v} \neq 0$ and unitary matrix $\mathbf{E}(1)$ in $GF(q)$. Given any causal FIR PU system $\mathbf{G}(z)$ in $GF(q)$, the algorithm for degree-one reduction is very similar to that given in Section 4.4.2 for $GF(2)$. The first step is to identify any vector $\mathbf{v}$ with $\mathbf{v}^T\mathbf{v} \neq 0$ in the null space of $\mathbf{g}(0)$ or $\mathbf{g}^T(0)$. Then form the building block $\mathbf{D}_q(z)$ as in (4.8.8). It can be verified that such $\mathbf{D}_q(z)$ can be used to reduce the degree of $\mathbf{G}(z)$. However the degree-one reduction is not always possible. As in the $GF(2)$ case, the degree-one building block $\mathbf{D}_q(z)$ in (4.8.8) is not complete for the class of PU systems over $GF(q)$. There are PU systems in $GF(q)$ that cannot be written as a product of $\mathbf{D}_q(z)$ (see Example 4.8.1). In $GF(q)$ we can use the following to test if we can extract a degree-one PU building block.

Lemma 4.8.2. *Test of Factorizability of PU Matrices over $GF(q)$:* Let $\mathbf{E}(z) = \sum_{k=0}^{N}\mathbf{e}(k)z^{-k}$ be a causal PU system over $GF(q)$ for some prime $q > 2$. Let $\mathbf{U} = [\mathbf{u}_0, \ldots, \mathbf{u}_s]$ and $\mathbf{V} = [\mathbf{v}_0, \ldots, \mathbf{v}_s]$ be *any*

bases that span the null spaces of $\mathbf{e}(0)$ and $\mathbf{e}^T(0)$ respectively. Then we cannot extract a degree-one PU building block from $\mathbf{E}(z)$ if and only if both $\mathbf{U}^T\mathbf{U} = \mathbf{0}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{0}$. ■

Proof: It is not difficult to see that the degree-one reduction fails if and only if neither the null space of $\mathbf{e}(0)$ nor $\mathbf{e}^T(0)$ contains any vector $\mathbf{v}$ with $l_{\mathbf{v}} = \mathbf{v}^T\mathbf{v} \neq 0$. What remains to be shown is that the above condition is equivalent to $\mathbf{U}^T\mathbf{U} = \mathbf{0}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{0}$. To show the "if" part, assume $\mathbf{U}^T\mathbf{U} = \mathbf{0}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{0}$. Then any vector $\mathbf{u}$ in the null space of $\mathbf{e}(0)$ is a linear combination of $\mathbf{u}_i$, i.e., $\mathbf{u} = c_0\mathbf{u}_0 + \ldots + c_s\mathbf{u}_s$ for some constants $c_i \in GF(q)$. Since $\mathbf{U}^T\mathbf{U} = \mathbf{0}$, we have $\mathbf{u}^T\mathbf{u} = 0$. Similarly we can show that $\mathbf{V}^T\mathbf{V} = \mathbf{0}$ implies all the vectors in the null space of $\mathbf{e}^T(0)$ have $\mathbf{v}^T\mathbf{v} = 0$. To prove the "only if" part, assume that $\mathbf{U}^T\mathbf{U} \neq \mathbf{0}$ (the proof is similar if $\mathbf{V}^T\mathbf{V} \neq \mathbf{0}$). If there is any $\mathbf{u}_i$ with $\mathbf{u}_i^T\mathbf{u}_i \neq 0$, then we can form $\mathbf{D}_q(z)$ with $\mathbf{u}_i$, and we are done. Therefore assume that all $\mathbf{u}_i$ have $\mathbf{u}_i^T\mathbf{u}_i = 0$. As $\mathbf{U}^T\mathbf{U} \neq \mathbf{0}$, there are $\mathbf{u}_i$ and $\mathbf{u}_j$ such that $\mathbf{u}_i^T\mathbf{u}_j \neq 0$. With these $\mathbf{u}_i$ and $\mathbf{u}_j$, we form the new vector $\mathbf{u} = \mathbf{u}_i + \mathbf{u}_j$ so that $\mathbf{u}^T\mathbf{u} = 2\mathbf{u}_i^T\mathbf{u}_j \neq 0$ (because 2 is coprime with q). Thus we can form $\mathbf{D}_q(z)$ with the new vector $\mathbf{u}$ and degree-one reduction with $\mathbf{D}_q(z)$ will succeed. Therefore we conclude that if either $\mathbf{U}^T\mathbf{U} \neq \mathbf{0}$ or $\mathbf{V}^T\mathbf{V} \neq \mathbf{0}$, the degree-one reduction will work. The proof is complete. ■

One consequence of Lemma 4.8.2 is that in $GF(q)$ the degree-two PU system $\mathbf{K}(z)$ in (4.5.4) is factorizable in terms of the degree-one PU building block $\mathbf{D}_q(z)$. To see this, recall that

$$\mathbf{K}_q(z) = \mathbf{I} - \mathbf{u}\mathbf{v}^T - \mathbf{v}\mathbf{u}^T + z^{-1}(\mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T), \tag{4.8.10}$$

where the vectors $\mathbf{u}$ and $\mathbf{v}$ are such that $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 0$ and $\mathbf{u}^T\mathbf{v} = 1$ (note that in $GF(2)$, $\mathbf{I} - \mathbf{u}\mathbf{v}^T - \mathbf{v}\mathbf{u}^T = \mathbf{I} + \mathbf{u}\mathbf{v}^T + \mathbf{v}\mathbf{u}^T$)). Form $\mathbf{v}_+ = \mathbf{u} + \mathbf{v}$ and $\mathbf{v}_- = \mathbf{u} - \mathbf{v}$ such that $l_{\mathbf{v}_+} = \mathbf{v}_+^T\mathbf{v}_+ = 2 \neq 0$ and $l_{\mathbf{v}_-} = \mathbf{v}_-^T\mathbf{v}_- = q - 2 \neq 0$ (note that $-2 = q - 2$ in $GF(q)$). With $\mathbf{v}_+$ and $\mathbf{v}_-$, we can factorize the $\mathbf{K}_q(z)$ in (4.8.10) as

$$\mathbf{K}_q(z) = \Big[\underbrace{\mathbf{I} - 2^{-1}\mathbf{v}_+\mathbf{v}_+^T + z^{-1}2^{-1}\mathbf{v}_+\mathbf{v}_+^T}_{\mathbf{D}_{q0}(z)}\Big]\Big[\underbrace{\mathbf{I} + 2^{-1}\mathbf{v}_-\mathbf{v}_-^T - z^{-1}2^{-1}\mathbf{v}_-\mathbf{v}_-^T}_{\mathbf{D}_{q1}(z)}\Big], \tag{4.8.11}$$

where both $\mathbf{D}_{q0}(z)$ and $\mathbf{D}_{q1}(z)$ are degree-one PU systems. In fact, in $GF(q)$ all first order PU systems (i.e., LOT) are factorizable in terms of the degree-one PU system $\mathbf{D}_q(z)$ in (4.8.8).

Theorem 4.8.2. *Complete Factorization of LOT in $GF(q)$:* Consider the first order system $\mathbf{E}(z) = \mathbf{e}(0) + \mathbf{e}(1)z^{-1}$ in $GF(q)$ for some prime $q > 2$. Then $\mathbf{E}(z)$ is a LOT of degree ρ if and only if it can be written as:

$$\mathbf{E}(z) = \mathbf{E}(1)\prod_{i=0}^{\rho-1}\Big[\mathbf{I} - l_{\mathbf{v}_i}^{-1}\mathbf{v}_i\mathbf{v}_i^T + z^{-1}l_{\mathbf{v}_i}^{-1}\mathbf{v}_i\mathbf{v}_i^T\Big], \tag{4.8.12}$$

where $l_{\mathbf{v}_i} = \mathbf{v}_i^T\mathbf{v}_i \neq 0$, the matrix $\mathbf{E}(1)$ is unitary, and the vectors $\mathbf{v}_i$ satisfy $\mathbf{v}_i^T\mathbf{v}_j = l_{\mathbf{v}_i}\delta(i-j)$. ■

The proof of the above theorem is very similar to that of Theorem 4.6.3. The LOT in $GF(q)$ also allows a minimal characterization similar to that given in Theorem 4.6.2. Even though in $GF(q)$ all LOTs are factorizable, there are unfactorizable higher order PU systems.

Example 4.8.1. *A 2×2 Unfactorizable PU System in $GF(5)$:* Consider the following second-order system:

$$\mathbf{G}(z) = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} + z^{-1}\mathbf{I} + z^{-2}\begin{bmatrix} 4 & 3 \\ 3 & 1 \end{bmatrix}. \tag{4.8.13}$$

The system $\mathbf{G}(z)$ is not a LOT because its order > 1. One can verify that the impulse response $\mathbf{g}(i)$ satisfies the condition in (4.3.2) so that $\mathbf{G}(z)$ is PU. Moreover $\mathbf{G}(z)$ has degree two because $[\det \mathbf{G}(z)] = z^{-2}$. Since $\mathbf{g}(0)$ is symmetric, the null spaces of $\mathbf{g}(0)$ and $\mathbf{g}^T(0)$ are identical. The null space of $\mathbf{g}(0)$ consists of vectors of the form $c[3 \;\; 1]^T$ where $c \in GF(5)$. As $[3 \;\; 1][3 \;\; 1]^T = 0$, using Lemma 4.8.2 we conclude that the PU system $\mathbf{G}(z)$ cannot be factorized in terms of $\mathbf{D}_q(z)$. ■

4.8.3. Non Trivial 2×2 Building Block $\mathbf{D}_q(z)$ in $GF(q)$

In the theory of FBs and wavelets in the complex field, one important class is the two-channel PU FBs. The corresponding polyphase matrix is a 2×2 PU matrix which can always be decomposed into degree-one building blocks in the complex case. In the case of finite fields, we know from previous dicussions that the building block $\mathbf{D}_q(z)$ is not complete. In fact, there are not many nontrivial 2×2 PU systems that are factorizable because there are few nontrivial 2×2 degree-one PU systems. (A system $\mathbf{E}(z)$ is said to be *trivial* if it is a diagonal matrix). In particular, all 2×2 degree-one PU systems in $GF(2)$ are diagonal because there is no 2×1 vector $\mathbf{v}$ with $v_0 \neq 0$, $v_1 \neq 0$ and $\mathbf{v}^T\mathbf{v} \neq 0$. Therefore all factorizable 2×2 PU systems in $GF(2)$ are diagonal systems. In the following, we will derive a formula for the number of nontrivial 2×2 degree-one building block $\mathbf{D}_q(z)$ in $GF(q)$ for $q > 2$.

From (4.8.8), we see that $\mathbf{D}_q(z)$ is trivial if and only if the vector $\mathbf{v}$ is either $[v_0 \;\; 0]^T$ or $[0 \;\; v_1]^T$. Therefore it is sufficient to consider vectors with $v_0 \neq 0$ and $v_1 \neq 0$. However the number of nontrivial $\mathbf{D}_q(z)$ is less than the number of distinct vectors (with $v_0 \neq 0$ and $v_1 \neq 0$) because two distinct vectors could generate the same $\mathbf{D}_q(z)$. To be more precise, one can show that two building blocks $\mathbf{D}_q(z)$ generated from two different vectors $\mathbf{u}$ and $\mathbf{v}$ are equivalent if and only if the vectors are related as $\mathbf{u} = k\mathbf{v}$ for some $k \in GF(q)$. Define the set

$$\mathcal{U} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \ldots, \begin{bmatrix} 1 \\ q-1 \end{bmatrix} \right\}. \tag{4.8.14}$$

Then it can be shown that if $\mathbf{u}_0, \mathbf{u}_1 \in \mathcal{U}$, $\mathbf{u}_0 = c\mathbf{u}_1$ if and only if $\mathbf{u}_0 = \mathbf{u}_1$. Therefore if $\mathbf{u}_0$ and $\mathbf{u}_1$ (with $\mathbf{u}_i^T\mathbf{u}_i \neq 0$) are vectors in $\mathcal{U}$, then they generate two distinct nontrivial $\mathbf{D}_q(z)$. Moreover, it is not difficult to show the set $\mathcal{U}$ has the following property: For any $\mathbf{v}$ with nonzero elements, there is an $\mathbf{u} \in \mathcal{U}$ and a $k \in GF(q)$ such that $\mathbf{v} = k\mathbf{u}$. Combining the above results, we can conclude that given any nontrivial degree-one PU building block $\mathbf{D}_q(z)$, there is a *unique* vector $\mathbf{u} \in \mathcal{U}$ such that $\mathbf{D}_q(z) = \mathbf{I} - l_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^T + z^{-1}l_{\mathbf{u}}^{-1}\mathbf{u}\mathbf{u}^T$. Therefore the number of nontrivial 2×2 degree-one PU systems is exactly the number of elements in the following set:

$$\mathcal{U}_1 = \{\mathbf{u} \in \mathcal{U} \big| \mathbf{u}^T\mathbf{u} \neq 0\}. \tag{4.8.15}$$

Note that in $\mathcal{U}$, the number of vectors with $\mathbf{u}^T\mathbf{u} = 0$ is equal to the number of solutions to the equation

$$u^2 = -1 \ mod \ q, \quad \text{for} \ \ u \in GF(q). \tag{4.8.16}$$

Except for $GF(2)$ case (because in $GF(2)$, $-1 = 1$), one can show that u is a solution to (4.8.16) if and only if the *order* of u is 4, i.e., $u^4 = 1 \ mod \ q$ but $u^i \neq 1 \ mod \ q$ for $i < 4$. From number theory [McC79], we know that there is an element of order 4 in $GF(q)$ if and only if $q - 1$ is divisible by 4. Using Euler function [McC79], there are exactly two elements of order 4 if they exist. Therefore we conclude that (4.8.16) can have either no solution or two solutions depending on whether $q - 1$ is divisible by 4. More precisely, we have

$$(\text{number of vectors in } \mathcal{U} \text{ with } \mathbf{u}^T\mathbf{u} = 0) = \begin{cases} 2, & \text{if } q - 1 = 0 \ mod \ 4; \\ 0, & \text{otherwise.} \end{cases} \tag{4.8.17}$$

Combining all the results, we have shown that the number of nontrivial 2×2 degree-one PU systems in $GF(q)$ for $q > 2$ is

$$(q - 1) - 2 \cdot \delta([q - 1]_4), \tag{4.8.18}$$

where $[q - 1]_4$ denotes $(q - 1) \ mod \ 4$. From (4.8.18), we conclude that for $q > 2$, there are at most $(q - 1)$ nontrivial 2×2 building blocks $\mathbf{D}_q(z)$ in $GF(q)$.

4.9. CONCLUSIONS

In this chapter, we gave a detail study on the theory of unitary and PU systems in finite fields. Explicit degree-one and degree-two reduction algorithms for the $GF(2)$ case are given (Sections 4.4.2 and 4.5.2). Several tests for factorizability of PU systems are also given (Lemmas 4.4.2, 4.5.2, and 4.8.2). We have proved a number of factorization theorems for both unitary matrices (Theorems 4.2.1 and 4.8.1) and PU systems (Theorems 4.6.2, 4.6.3, and 4.8.2). In particular, we have shown that all LOTs in $GF(q)$, for any prime number q, are factorizable in terms of smaller (degree-one or degree-two) PU building blocks (Theorems 4.6.3 and 4.8.2). Even though these degree-one or degree-two building blocks are the most general, there are PU systems that cannot be factorized (Examples in (4.5.12) and (4.8.13)).

All the theories in this chapter are developed for finite fields of the form $GF(q)$ with prime q. It would be interesting to extend the results to the fields of the form $GF(q^m)$. In particular, PU systems that cannot be factorized may be factorizable if we use building blocks from extension fields. This is still an open problem. Also we have studied the theory of systems with PU porperty only (except the example in (4.5.1)). It is important to look at other classes such as the unimodular matrices (which are useful in the coding theory [For70, McE95]) and the class of causal matrices with anti-causal inverses [Vai95c] (which cover the PU systems as a special case).

4.10. APPENDICES

Appendix A. Most General 2×2 Degree-two Unfactorizable PU Systems in $GF(2)$

Consider the 2×2 degree-two PU system $\mathbf{G}(z) = \mathbf{g}(0) + \mathbf{g}(1)z^{-1} + \mathbf{g}(2)z^{-2}$. Since $\mathbf{G}(z)$ has degree two, the rank of $\mathbf{g}(2) \leq 1$. If $\mathbf{g}(2) = \mathbf{0}$, then $\mathbf{g}(1)$ has full rank so that the system reduces to the trivial factorizable system $\mathbf{G}(z) = \mathbf{g}(1)z^{-1}$, where $\mathbf{g}(1) = \mathbf{I}_2$ or $\mathbf{J}_2$. So assume rank $\mathbf{g}(2) = 1$. As $\mathbf{G}(z)$ is unfactorizable, the null spaces of $\mathbf{g}(0)$ and $\mathbf{g}^T(0)$ should not contain any vector with an odd weight. This implies $\mathbf{g}(0) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Using the PU conditions $\mathbf{g}^T(0)\mathbf{g}(2) = \mathbf{0}$ and $\mathbf{g}(0)\mathbf{g}^T(2) = \mathbf{0}$, we conclude that $\mathbf{g}(2) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. To find $\mathbf{g}(1)$, we use the condition

$$\mathbf{g}^T(0)\mathbf{g}(0) + \mathbf{g}^T(1)\mathbf{g}(1) + \mathbf{g}^T(2)\mathbf{g}(2) = \mathbf{I}. \tag{4.A.1}$$

Substituting $\mathbf{g}(0)$ and $\mathbf{g}(2)$ into the above equation, we get $\mathbf{g}^T(1)\mathbf{g}(1) = \mathbf{I}$, which implies $\mathbf{g}(1)$ is unitary. The only 2×2 unitary matrices are $\mathbf{I}_2$ and $\mathbf{J}_2$. One can verify that both the choices of $\mathbf{g}(1) = \mathbf{I}_2$ and $\mathbf{g}(1) = \mathbf{J}_2$ give a PU system. Thus we conclude that the most general 2×2 degree-two unfactorizable PU system has the form:

$$\mathbf{G}(z) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \mathbf{g}(1)z^{-1} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} z^{-2}, \tag{4.A.2}$$

where $\mathbf{g}(1) = \mathbf{I}_2$ or $\mathbf{J}_2$.

Appendix B. A Fact for Matrices in $GF(2)$

Fact 4.B.1. Let $\mathbf{A}$ be an $M \times M$ matrix in $GF(2)$ with M even. Then $\mathbf{A}^T\mathbf{A} \neq \mathbf{J}_M$. ■

Note that Fact 4.B.1 is always true for all $M \geq 2$ in the real or complex field as $\mathbf{A}^T\mathbf{A}$ is always semi positive definite while $\mathbf{J}_M$ is not. In $GF(2)$, the lemma does not hold for odd M. To see this, consider $M = 3$. Then it can be verified that the following matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{4.B.1}$$

satisfies $\mathbf{A}^T\mathbf{A} = \mathbf{J}_3$.

Proof of Fact 4.B.1: Suppose that there is a matrix $\mathbf{A}$ such that $\mathbf{A}^T\mathbf{A} = \mathbf{J}_M$. Let $\mathbf{a}_i$ be the i-th column vector of $\mathbf{A}$. Then all $\mathbf{a}_i$ has an even weight because $\mathbf{a}_i^T\mathbf{a}_i = 0$ for all i. Therefore we have $[1\ 1\ \ldots\ 1]\mathbf{A} = \mathbf{0}$, which implies that $\mathbf{A}$ is singular. This contradicts the fact that $\mathbf{A}$ is nonsingular (because $\mathbf{A}^T\mathbf{A} = \mathbf{J}_M$). Thus we conclude that there does not exist such $\mathbf{A}$. ■

5

Basic Principles of Time-Varying Filters and Filter Banks

5.1. INTRODUCTION

Recently there has been considerable interest in both the theory and design of time-varying filter bank (TVFB) [Nay92, Arr93, Her93a, deQ93, Sod94, Gop95, Smi95]. In the applications of subband coding of speech and image signals, the advantages of TVFB are demonstrated in [Her93a, Arr93, Smi95]. It was shown that by using FBs with different responses on different regions (such as smooth areas and edge areas), low bit rate compression with little ringing effect can be achieved. In this chapter, we will focus mainly on the system-theorectic fundamentals of TVFBs whose analysis and synthesis filters are LTV.

TVFBs were first introduced in [Nay92]. The authors used a time-domain approach to formulate the problem of switching between two PR FBs. To preserve the PR property during the transition, a number of synthesis banks is designed. The number of synthesis banks involved is usually quite large and these transition synthesis banks are related to each other. In [Her93a], arbitrary orthonormal tilings of the time-frequency plane were considered. The authors used an approach based on a matrix formulation of the two-channel filter banks. Boundary and transition filters were constructed so that the orthogonality and PR were preserved when one switched from one PU filter bank to another. Orthonormal TV wavelet bases were generated. In [Her95], the results in [Her93a] were extended to the M-channel case. In [Gop95], the authors considered the problem of switching between two LTI PU filter banks, for which the LTI PU FBs were factorized into LTI degree one buiding blocks [Vai93]. The authors combined the result of factorization of LTI PU FBs and the matrix approach in [Her95] to construct transition filters, so that PR was obtained. In all the approaches proposed in [Her93, Nay92, Her95, Gop95] there is a need for the construction of transition filters to preserve PR. The transition period puts a limit on how fast we can vary the filters. A different approach to TVFBs was given in [Arr93, deQ93]. The method was based on a generalization of the FIR lattice [Vai93] to the time-varying case. In [Arr93], the authors considered two-channel TVFBs. By cascading sections of FIR lattice with time-varying parameters, the

authors showed that useful PR TVFBs can be obtained. The method was generalized to the M-channel case in [deQ93], where the planar rotations [Vai93] were made time-varying. A cascade of these TV planar rotations and delays resulted in a lossless PR TVFB. The authors also provided a method to obtain a smooth transition when one switched from one PR filter bank to the other. In the next chapter, we will show that there are infinitely many lossless TVFBs which cannot be realized as a cascade of the time-varying FIR lattice. In [Sod94], instead of redesigning filters during transition period, the authors designed a post filter such that the overall cascade system achieved nearly PR.

In this chapter, we will introduce an approach similar to the conventional polyphase method to study general properties of TVFBs. By using the proposed TV polyphase approach, we will address some basic theory for the general LTV filters and TVFBs. In particular, we are able to show the following differences between a conventional LTI FB and a TVFB:

1. In the LTI case, if a FB has PR, then the FB with analysis and synthesis filters interchanged has also the PR property (Problem 5.17 of [Vai93]). In the LTV case, a similar statement is not true.
2. In the LTI PU case, both the analysis and synthesis banks are lossless [Vai93]. In the LTV case, the losslessness of analysis bank does not imply that of the synthesis bank which yields PR.
3. If we replace the delay z^{-1} in an implementation of a LTI PU system with z^{-L} for some integer L, the system remains PU [Vai93]. This is usually not true for a LTV lossless system.
4. In the LTI case, it is shown [Djo94, Che94] that a PR FB will generate a *discrete-time Riesz basis* for l_2 space. In the LTV case, a PR TVFB will only generate a *frame* for l_2. It becomes a basis only if the time-varying synthesis polyphase matrix is invertible.

There are many other properties of LTV filters and TVFBs which cannot be expected from the LTI cases. In order to show the differences between the LTV filters and the LTI filters, we discuss a single-input single-output (SISO) LTV system (which corresponds to an one-channel FB) in the following subsection.

5.1.1. A Simple Time-Varying System

Example 5.1.1. Consider the LTV system $\mathcal{H}$ given by $y(n) = h_0(n)x(n) + h_1(n)x(n-1)$, where the coefficients are:

$$h_0(n) = \begin{cases} 1, & \text{for } n < 0; \\ 0, & \text{for } n \geq 0, \end{cases} \qquad h_1(n) = \begin{cases} 0, & \text{for } n \leq 0; \\ 1, & \text{for } n > 0. \end{cases} \tag{5.1.1}$$

One can verify that the output of the system is:

$$y(n) = \begin{cases} x(n), & \text{for } n < 0; \\ 0, & \text{for } n = 0; \\ x(n-1), & \text{for } n > 0. \end{cases} \tag{5.1.2}$$

The above system has an inverse $\mathcal{G}$ given as $\widehat{x}(n) = g_0(n)y(n) + g_{-1}(n+1)y(n+1)$, where the coefficients $g_0(n) = h_0(n)$ and $g_{-1}(n) = h_1(n)$. The output of the inverse system is:

$$\widehat{x}(n) = \begin{cases} y(n), & \text{for } n < 0; \\ y(n+1), & \text{for } n \geq 0. \end{cases} \tag{5.1.3}$$

We see that $\widehat{x}(n) = x(n)$ for all n. From this example, we can observe the following:

1. *Losslessness:* From (5.1.2), it is clear that $\sum_n |y(n)|^2 = \sum_n |x(n)|^2$, so the system $\mathcal{H}$ is lossless. In general, given a LTV system, it is difficult to test the losslessness by computing the output energy. In Section 5.5, we will show how to characterize lossless systems in terms of their impulse responses.
2. *Non Trivial SISO Lossless Systems:* The system $\mathcal{H}$ has all the coefficients $h_i(n)$ equal to zero at $n = 0$ and yet the system is lossless. This is not possible in LTI case. In the LTI case, the only scalar lossless system is the trivial system of the form $e^{j\theta}z^{-L}$. For the LTV case, there exist quite non trivial scalar lossless FIR systems as we will explain in Section 5.3 and demonstrate in Section 5.6.
3. *Existence and Uniqueness of Inverse:* In the above example, it can be shown that for any choice of the constant c, the system described by $\widehat{x}(n) = \big(h_0(n) + c\delta(n)\big)y(n) + h_1(n+1)y(n+1)$ is an inverse for the lossless system of (5.1.2). Therefore the inverse of a LTV lossless system is in general not unique. In Section 5.5, we will show that a lossless system is *always* invertible. The conditions for unique invertibility of lossless systems are studied in detail.
4. *Non Invertibility and Non Losslessness of the Inverse*: It is clear from (5.1.3) that the inverse system $\mathcal{G}$ is not invertible because the sample $y(0)$ is lost and can never be recovered. In fact, the inverse system $\mathcal{G}$ is not lossless! This situation is different from the LTI case where the inverse of a lossless system is also invertible and lossless. In Section 5.5, we will study the conditions under which the inverse of a LTV lossless system is invertible and lossless.

5.1.2. Chapter Related Notations and Definitions

Through out this chapter and Chapter 6, we will use the following notations and definitions: The coefficients of a LTV multi-input multi-output (MIMO) filter are denoted by matrices $\mathbf{e}_k(n)$, where n is the time index and k is the coefficient index. All of the MIMO systems considered in this chapter have M inputs and M outputs, therefore $\mathbf{e}_k(n)$ are $M \times M$ matrices. The calligraphic symbols such as $\mathcal{H}$, $\mathcal{G}$ in Example 5.1.1 are used to denote LTV systems. Given two systems $\mathcal{H}_1$ and $\mathcal{H}_2$, a cascade of $\mathcal{H}_1$ followed by $\mathcal{H}_2$ is denoted by $\mathcal{H}_2\mathcal{H}_1$. The output of a system $\mathcal{H}$ corresponds to the input $\mathbf{x}(n)$ is expressed as $\mathbf{y}(n) = \mathcal{H}\mathbf{x}(n)$, where both $\mathbf{y}(n)$ and $\mathbf{x}(n)$ are $M \times 1$ column vectors.

1. *Inner Product:* The inner product of $x_1(n)$ and $x_2(n)$ is defined as $\langle x_1(n), x_2(n)\rangle = \sum_n x_1(n)x_2^*(n)$. For two vector sequences, their inner product is defined as $\langle \mathbf{x}_1(n), \mathbf{x}_2(n)\rangle = \sum_n \mathbf{x}_2^\dagger(n)\mathbf{x}_1(n)$.
2. *l_2-norm and $l_2(M)$ Space:* For a vector sequence $\mathbf{x}(n)$, its l_2-norm is $\| \mathbf{x}(n) \| = \big(\sum_n \mathbf{x}^\dagger(n)\mathbf{x}(n)\big)^{1/2}$. The space of all finite norm M-dimensional vector sequences is denoted by $l_2(M)$. The space of all finite norm scalar sequences is simply represented by l_2.
3. *Lossless System:* A system is lossless if it is stable and preserves input/output energy.
4. *Passive System:* A system is passive if it is stable and its output energy can never be greater than input energy, i.e., $\| \mathbf{y}(n) \|^2 \leq \| \mathbf{x}(n) \|^2$. Note that a lossless system is also passive.

5. *Inverse System:* A system $\mathcal{G}$ is said to be the inverse of a system $\mathcal{H}$ if the cascade of $\mathcal{H}$ followed by $\mathcal{G}$ is the identity system $\mathcal{I}$. The fact that $\mathcal{G}$ is an inverse of $\mathcal{H}$ is denoted by $\mathcal{GH} = \mathcal{I}$. In general, $\mathcal{GH} = \mathcal{I}$ does not imply $\mathcal{HG} = \mathcal{I}$, even for the case of scalar LTV systems (see Example 5.1.1).
6. *Invertible/Non Invertible Inverse Lossless System:* Lossless systems with invertible inverses are called invertible inverse lossless (IIL) systems. Lossless systems with non invertible inverses are called non invertible inverse lossless (NIL) systems. Note that there are no LTI NIL systems.

5.1.3. Chapter Outline

In Section 5.2, we will first review some basics for LTV filters and introduce two direct form structures which will be used throughout the chapter. Then a transform domain description for LTV filters will be defined and studied. By using the transform domain description, we define the polyphase representations of LTV filters in Section 5.3. Noble identities similar to the LTI case hold for LTV systems. Efficient implementations for LTV decimation filters and interpolation filters will be derived. In Section 5.4, we will utilize the proposed polyphase representation to study some basic properties of TVFB, such as PR condition, interchangability of the analysis and synthesis filters, application to PR transmultiplexers, etc. In Section 5.5, lossless LTV filters and filter banks will be considered. All lossless FIR LTV systems are shown to be invertible and their inverses are also FIR. Explicit formula for the inverse will be given. We will also discuss in detail some subtle properties of the inverse of lossless systems, such as uniqueness, invertibility, losslessness, etc. In Section 5.6, some lossless LTV filter and filter bank examples will be provided to demonstrate the theory. In Section 5.7, a time-varying vector space approach to PR TVFBs is given. This can be viewed as a generalization of the approach proposed in [Che94]. By using this LTV vector space approach, we will show that a NIL TVFB only gives rise to a *discrete-time tight frame* with unity frame bound for l_2 space. In the case IIL TVFBs, we can obtain an *orthonormal basis* for l_2.

5.2. DIRECT FORM STRUCTURES AND NEW DESCRIPTIONS FOR LTV FILTERS

A review of notations and different possible representations of LTV filters is given in [Pra92, Cro83]. Her we are going to use only two of the representations, which correspond to two different direct form implementations.

5.2.1. Direct Form A and B Implementations of LTV Filters

Consider Figs. 5.2.1 and 5.2.2 where two different structures to implement a N-th order MIMO causal LTV filter are shown. In the LTI case, these two structures are the same. In the LTV case, there is a simple one to one correspondence between these two structures. We will call the structures in Figs. 5.2.1 and 5.2.2 respectively the direct form A and B implementations. Their system equations can be respectively expressed as

$$\mathbf{y}(n) = \mathbf{e}_0(n)\mathbf{x}(n) + \mathbf{e}_1(n)\mathbf{x}(n-1) + \ldots + \mathbf{e}_N(n)\mathbf{x}(n-N), \tag{5.2.1}$$

$$\mathbf{y}(n) = \mathbf{r}_0(n)\mathbf{x}(n) + \mathbf{r}_1(n-1)\mathbf{x}(n-1) + \ldots + \mathbf{r}_N(n-N)\mathbf{x}(n-N), \tag{5.2.2}$$

where the coefficients $\mathbf{e}_k(n)$ and $\mathbf{r}_k(n)$ are $M \times M$ matrices. From the above two equations, it is clear that if $\mathbf{r}_k(n-k) = \mathbf{e}_k(n)$, the two filters in Figs. 5.2.1 and 5.2.2 are the same. We define the z-transform of the frozen systems in Figs. 5.2.1 and 5.2.2 at time n respectively as

$$\mathbf{E}_n(z) = \sum_k \mathbf{e}_k(n) z^{-k}, \qquad \mathbf{R}_n(z) = \sum_k \mathbf{r}_k(n) z^{-k}, \tag{5.2.3}$$

where $\mathbf{E}_n(z)$ and $\mathbf{R}_n(z)$ are both LTI filters. It is shown in [Pra92] that we can implement the LTV filters by using the LTI filters in (5.2.3) and commutators as shown in Figs. 5.2.3 and 5.2.4. In the case of LPTV filters with period M, we have $\mathbf{E}_{n+M}(z) = \mathbf{E}_n(z)$ and $\mathbf{R}_{n+M}(z) = \mathbf{R}_n(z)$. Figs. 5.2.3 and 5.2.4 respectively reduce to the LPTV systems given in Figs. 3(b) and (a) of [Pra92]. From Fig. 5.2.3, it is clear that it is the most convenient to implement a LTV decimation filter by using direct form A structure. In a decimation filter, the output of the LTV filter is decimated by a factor of M, in this case, we need only to implement the subset $\mathbf{E}_{Mn}(z)$. Similarly, in the case of interpolation filters, the direct form B is the most convenient because the input $\mathbf{x}(n) = \mathbf{0}$ for $n \neq$ multiple of M, and we need only to implement $\mathbf{R}_{Mn}(z)$. Even though both of the direct forms in Figs. 5.2.1 and 5.2.2 are equivalent, we will choose direct form A for decimation filters and direct form B for interpolation filters throughout the chapter. The advantages of making such a choice will become clear as we go through the presentation.

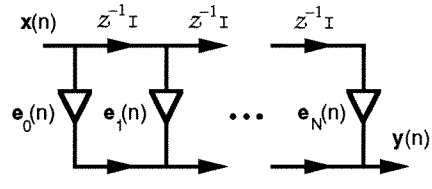

Fig. 5.2.1. Direct form A implementation of a N-th order LTV filter.

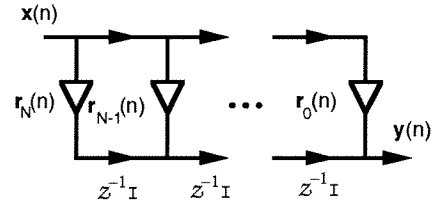

Fig. 5.2.2. Direct form B implementation of a N-th order LTV filter.

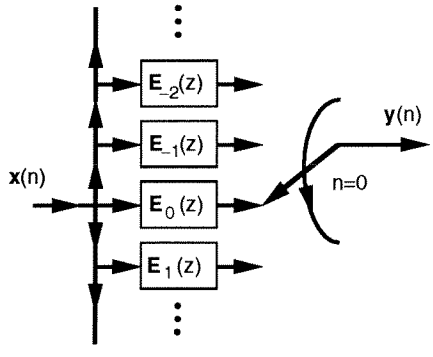

Fig. 5.2.3. Interpretation of direct form A in Fig. 5.2.1 by using a commutator switch and the LTI filters $E_i(z)$ defined in (5.2.3).

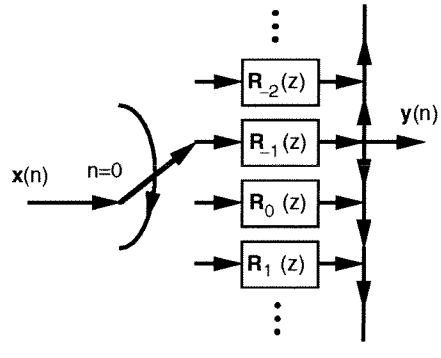

Fig. 5.2.4. Interpretation of direct form B in Fig. 5.2.2 by using a commutator switch and the LTI filters $R_i(z)$ defined in (5.2.3).

5.2.2. Transform Domain Representations of LTV Filters

In the LTI case, polyphase representations are very useful tools in both the theory and design of multirate filter banks [Vai93]. In the LTV case, since the conventional z-transform is undefined, we cannot apply the traditional polyphase definitions. We need to define a transform domain description for LTV filters similar to the conventional z-transform for LTI filters. First let us define the delay operator $\mathcal{Z}^{-1}$ as the following: (i) $\mathcal{Z}^{-i}\mathbf{x}(n) = \mathbf{x}(n-i)$ for signal $\mathbf{x}(n)$; (ii) $\mathcal{Z}^{-i_0}\mathcal{Z}^{-i_1} = \mathcal{Z}^{-i_0-i_1}$. The rule for interchanging a time-dependent multiplier and the delay operator is described as: $\mathcal{Z}^{-i}\mathbf{r}_k(n) = \mathbf{r}_k(n-i)\mathcal{Z}^{-i}$, as shown in Fig. 5.2.5. Note that the delay operator $\mathcal{Z}^{-1}$ is different from the z-transform, it does not commute with a multiplier unless the multiplier is time-independent. Therefore the calligraphic symbol $\mathcal{Z}^{-1}$ is used for the delay operator to remind the readers of their difference. With the above delay operator, we can define the TV transform domain descriptions as

$$\mathbf{E}(n,\mathcal{Z}) = \sum_k \mathbf{e}_k(n)\mathcal{Z}^{-k}, \quad \text{(for direct form A; Fig. 5.2.1)}, \tag{5.2.4}$$

$$\mathbf{R}(\mathcal{Z},n) = \sum_k \mathcal{Z}^{-k}\mathbf{r}_k(n), \quad \text{(for direct form B; Fig. 5.2.2)}. \tag{5.2.5}$$

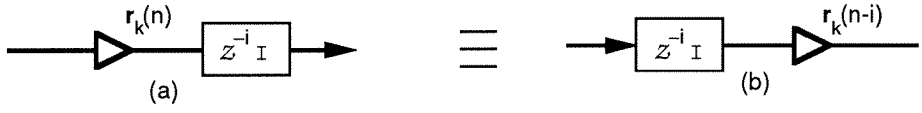

Fig. 5.2.5. Rule for interchanging a delay and a time-varying multiplier.

The readers should note the use of $(n,\mathcal{Z})$ in (5.2.4) corresponding to direct form A in Fig. 5.2.1, and $(\mathcal{Z},n)$ in (5.2.5) corresponding to direct form B in Fig. 5.2.2. By using the above TV transform domain descriptions, the output of the LTV filters in Figs. 5.2.1 and 5.2.2 can be written respectively as

$$\mathbf{y}(n) = \mathbf{E}(n,\mathcal{Z})\mathbf{x}(n) = \sum_k \mathbf{e}_k(n)\mathcal{Z}^{-k}\mathbf{x}(n) = \sum_k \mathbf{e}_k(n)\mathbf{x}(n-k), \tag{5.2.6}$$

$$\mathbf{y}(n) = \mathbf{R}(\mathcal{Z},n)\mathbf{x}(n) = \sum_k \mathcal{Z}^{-k}\mathbf{r}_k(n)\mathbf{x}(n) = \sum_k \mathbf{r}_k(n-k)\mathbf{x}(n-k). \tag{5.2.7}$$

In the transform domain, the rule of interchanging delays and LTV systems in Fig. 5.2.5 reduces to the following: (i) $\mathcal{Z}^{-i}\mathbf{E}(n,\mathcal{Z}) = \mathbf{E}(n-i,\mathcal{Z})\mathcal{Z}^{-i}$, and (ii) $\mathcal{Z}^{-i}\mathbf{R}(\mathcal{Z},n) = \mathbf{R}(\mathcal{Z},n-i)\mathcal{Z}^{-i}$. The cascade of two LTV filters, $\mathbf{E}^{(0)}(n,\mathcal{Z})$ followed by $\mathbf{E}^{(1)}(n,\mathcal{Z})$, can be described as

$$\mathbf{E}^{(1)}(n,\mathcal{Z})\mathbf{E}^{(0)}(n,\mathcal{Z}) = \sum_{k,l} \mathbf{e}_k^{(1)}(n)\mathbf{e}_l^{(0)}(n-k)\mathcal{Z}^{-k-l}, \tag{5.2.8}$$

where $\mathbf{e}_k^{(i)}(n)$ is the k-th coefficients of the LTV filter $\mathbf{E}^{(i)}(n,\mathcal{Z})$. Notice that in general LTV systems do not commute, i.e., $\mathbf{E}^{(1)}(n,\mathcal{Z})\mathbf{E}^{(0)}(n,\mathcal{Z}) \neq \mathbf{E}^{(0)}(n,\mathcal{Z})\mathbf{E}^{(1)}(n,\mathcal{Z})$, even for the simplest case of scalar LTV filters.

5.3. POLYPHASE REPRESENTATIONS, MULTIRATE IDENTITIES AND BLOCK IMPLEMENTATIONS

5.3.1. Polyphase Representations

With the transform domain description defined in the previous section, we are now ready to define the polyphase representations of LTV filters. Similar to the LTI case, we have two types of polyphase representations for both of the LTV filters in Figs. 5.2.1 and 5.2.2. First consider Fig. 5.2.1 (the corresponding transform domain description is given in (5.2.4)). With respect to any positive integer M, the Type 1 polyphase representation of the systems $\mathbf{E}(n, \mathcal{Z})$ can be defined as:

$$\mathbf{E}(n, \mathcal{Z}) = \sum_{i=0}^{M-1} \left(\sum_k \mathbf{e}_{kM+i}(n) \mathcal{Z}^{-kM} \right) \mathcal{Z}^{-i} = \sum_{i=0}^{M-1} \mathbf{E}_i(n, \mathcal{Z}^M) \mathcal{Z}^{-i}, \tag{5.3.1}$$

where $\mathbf{E}_i(n, \mathcal{Z})$ is called the i-th polyphase component of the Type 1 representation at time n. Similarly, for the direct form B in Fig. 5.2.2 (the corresponding transform domain description is given in (5.2.5)), we can define the Type 2 polyphase representation as:

$$\mathbf{R}(\mathcal{Z}, n) = \sum_{i=0}^{M-1} \mathcal{Z}^i \left(\sum_k \mathcal{Z}^{-kM} \mathbf{r}_{kM-i}(n) \right) = \sum_{i=0}^{M-1} \mathcal{Z}^i \mathbf{R}_i(\mathcal{Z}^M, n), \tag{5.3.2}$$

where $\mathbf{R}_i(\mathcal{Z}, n)$ is called the ith polyphase component of the Type 2 representation. Note that for convenience we have used the advance operators $\mathcal{Z}^i$ to define Type 2 polyphase representation. In the LTI case, the definitions in (5.3.1) and (5.3.2) reduce to the conventional definitions of polyphase representations [Vai93].

Similarly we can define the Type 2 and Type 1 polyphase representations for the direct form A and B implementations respectively. However as we explained in Section 5.2.1, we will only use direct form A for decimation filters (for which Type 1 polyphase is useful) and direct form B for interpolation filters (for which Type 2 polyphase is useful). Therefore in the chapter, we need only (5.3.1) and (5.3.2).

5.3.2. Noble Identities and Efficient Structures for Decimation and Interpolation Filters

In the LTI case, the noble identities [Vai93] are very useful in the implementions of decimation and interpolation filters. In the LTV case, similar identities continue to hold. Consider the TV multirate systems given in Figs. 5.3.1(a) and (c), where

$$\mathbf{E}(n, \mathcal{Z}^M) = \sum_k \mathbf{e}_k(n) \mathcal{Z}^{-kM}, \quad \mathbf{R}(\mathcal{Z}^M, n) = \sum_k \mathcal{Z}^{-kM} \mathbf{r}_k(n). \tag{5.3.3}$$

The noble identities say that we can redraw Fig. 5.3.1(a) as in Fig. 5.3.1(b) and Fig. 5.3.1(c) as in Fig. 5.3.1(d), where

$$\mathbf{E}(Mn, \mathcal{Z}) = \sum_k \mathbf{e}_k(Mn) \mathcal{Z}^{-k}, \quad \mathbf{R}(\mathcal{Z}, Mn) = \sum_k \mathcal{Z}^{-k} \mathbf{r}_k(Mn). \tag{5.3.4}$$

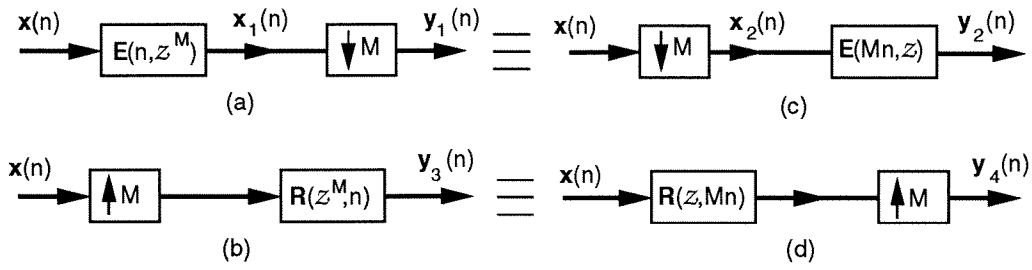

Fig. 5.3.1. Time-varying noble identities for decimators and interpolators.

Notice that even though the coefficients $\mathbf{e}_k(n)$ and $\mathbf{r}_k(n)$ in (5.3.3) are defined for all n, only those coefficients at time $n =$ multiple of M are relevant to the outputs (due to the decimator and interpolator). This is consistent with the fact that in the case of decimation and interpolation filters, only one out of M sets of coefficients are needed (as explained in Section 5.2.1). In the following, we will prove the noble identity for the decimator only, the proof for the interpolator being similar. From Figs. 5.3.1, we have

$$\mathbf{y}_1(n) = \Big[\mathbf{x}_1(m)\Big]_{m=Mn} = \Big[\sum_k \mathbf{e}_k(m)\mathbf{x}(m-Mk)\Big]_{m=Mn} = \sum_k \mathbf{e}_k(Mn)\mathbf{x}(Mn-Mk), \tag{5.3.5}$$

$$\mathbf{y}_3(n) = \sum_k \mathbf{e}_k(Mn)\mathbf{x}_3(n-k) = \sum_k \mathbf{e}_k(Mn)\mathbf{x}(Mn-Mk). \tag{5.3.6}$$

Comparing (5.3.5) and (5.3.6), we have proved the noble identity for decimator.

By using the noble identities, we find another interpretation for the polyphase components of a LTV filter. Consider the cascade system shown in Fig. 5.3.2(a), where $\mathbf{E}(n,\mathcal{Z})\mathcal{Z}^i$ is sandwiched between an interpolator and a decimator. By using the Type 1 polyphase representation in (5.3.1), Fig. 5.3.2(a) can be redrawn as Fig. 5.3.2(b), where $\mathbf{E}_i(n,\mathcal{Z}^M)$ are as defined in (5.3.1). Invoking the noble identity for decimators, Fig. 5.3.2(b) can be redrawn as Fig. 5.3.2(c). Clearly, Fig. 5.3.2(c) is equivalent to Fig. 5.3.2(d). Therefore the circuit in Fig. 5.3.2(a) is equivalent to the i-th polyphase component $\mathbf{E}_i(Mn,\mathcal{Z})$ of the filter $\mathbf{E}(n,\mathcal{Z})$ and it is denoted by the following notation:

$$\Big[\mathbf{E}(n,\mathcal{Z})\mathcal{Z}^i\Big]_{\downarrow M} = \mathbf{E}_i(Mn,\mathcal{Z}). \tag{5.3.7}$$

Similarly if a direct form B filter $\mathcal{Z}^{-i}\mathbf{R}(\mathcal{Z},n)$ is sandwiched between an interpolator and a decimator, one can show that the equivalent system is $\mathbf{R}_i(\mathcal{Z},Mn)$. Using the notation introduced in (5.3.7), we get

$$\Big[\mathcal{Z}^{-i}\mathbf{R}(\mathcal{Z},n)\Big]_{\downarrow M} = \mathbf{R}_i(\mathcal{Z},Mn). \tag{5.3.8}$$

Efficient Structures for Decimation and Interpolation Filters: Consider the decimation filter shown in Fig. 5.3.3(a), where the direct form A implementation of $\mathbf{E}(n,\mathcal{Z})$ is given in Fig. 5.2.1. By using the Type 1 polyphase representation in (5.3.1) and invoking the noble identity, Fig. 5.3.3(a) can be redrawn as in Fig. 5.3.3(b), $\mathbf{E}_i(Mn,\mathcal{Z})$ is the i-th Type 1 polyphase component of $\mathbf{E}(Mn,\mathcal{Z})$. Similarly, the direct

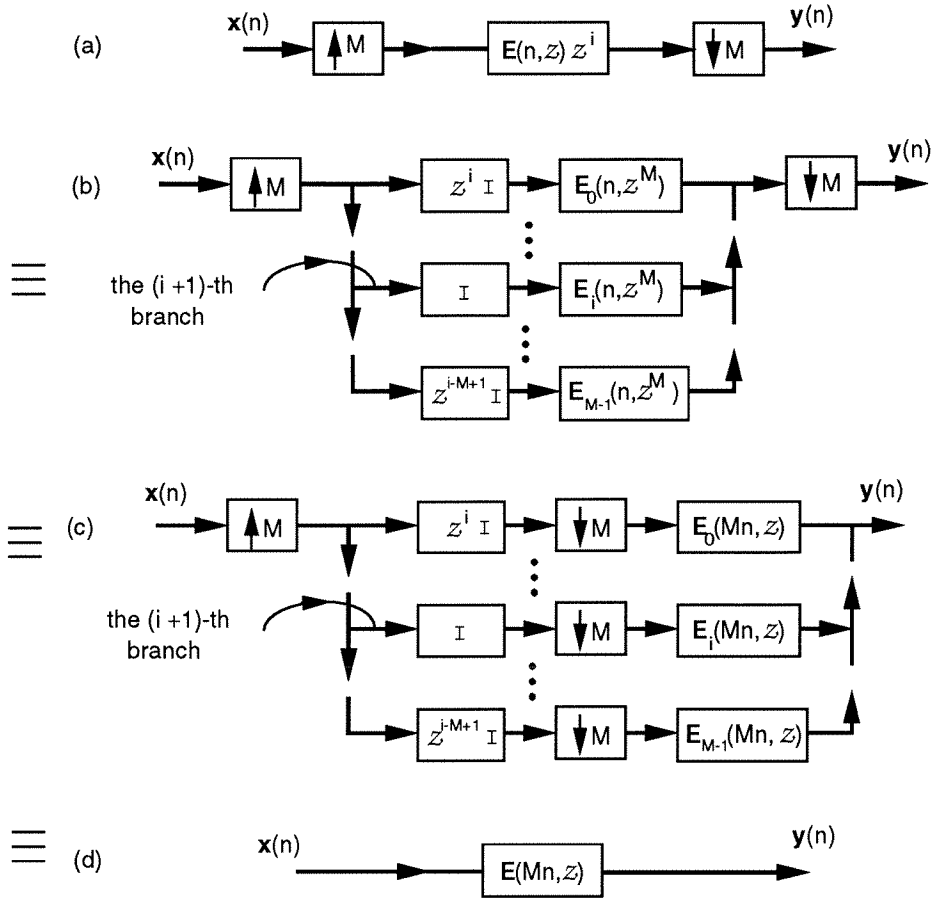

Fig. 5.3.2. Another interpretation of the i-th polyphase component of the filter $E(n, \mathcal{Z})$.

form B implementation of the interpolation filter in Fig. 5.3.4(a) can be implemented as in Fig. 5.3.4(b), where $\mathbf{R}_i(\mathcal{Z}, Mn)$ is the i-th Type 2 polyphase components of $\mathbf{R}(\mathcal{Z}, Mn)$. Figs. 5.3.3(b) and 5.3.4(b) can be regarded as efficient implementations of the decimation and interpolation filters respectively. Consider the case of fractional decimation discussed in (Chapter 4.3.3 of [Vai93]). If the filter is time-varying, one can show that we can use both Type 1 and 2 polyphase representations to simplify the circuit and arrive at an efficient structure similar to Figs. 4.3–8 in [Vai93].

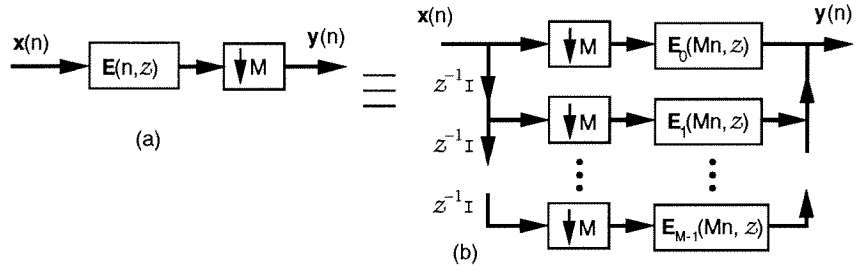

Fig. 5.3.3. Decimation filter and its efficient implementation using polyphase representation.

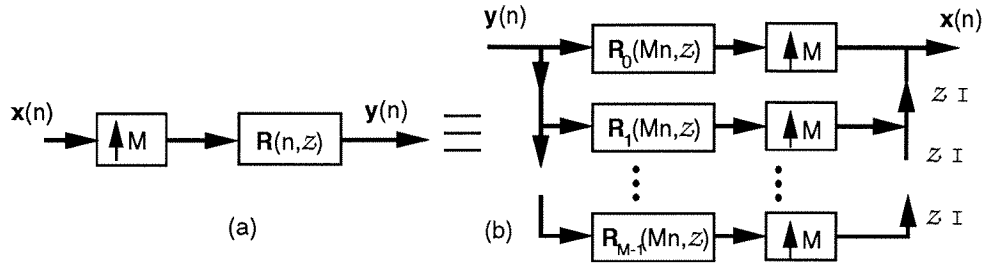

Fig. 5.3.4. Interpolation filter and its efficient implementation using polyphase representation.

5.3.3. Block Implementation of Scalar LTV Filters

Block implementation of LTI filters has been considered in the past [Bur71]. In this section, we will first study the blocking and unblocking of LTV filters by using the transform domain representations. Consider the scalar LTV filter $H(n, \mathcal{Z})$. At time n, its Type 1 polyphase components $E_i(n, \mathcal{Z})$ are defined in (5.3.1). Note that in this case, these $E_i(n, \mathcal{Z})$ are scalar systems. To obtain a block implementation of this scalar filter, we cascade a trivial PR FB (which contains only a delay chain and an advance chain) after the filter as shown in Fig. 5.3.5(a). Using the rule of interchanging the delay $\mathcal{Z}^{-1}$ and LTV filters given in Section 5.2.2, we have $\mathcal{Z}^{-i}H(n, \mathcal{Z}) = H(n-i, \mathcal{Z})\mathcal{Z}^{-i}$, which yields the equivalent structure in Fig. 5.3.5(b). Using the Type 1 polyphase representation for each of the filters $H(n-i, \mathcal{Z})\mathcal{Z}^{-i}$ and applying the noble identity in Fig. 5.3.1, the block implementation can be drawn as Fig. 5.3.5(c), where the polyphase matrix $\mathbf{E}(n, \mathcal{Z})$ is:

$$\mathbf{E}(n, \mathcal{Z}) = \begin{bmatrix} E_0(Mn, \mathcal{Z}) & E_1(Mn, \mathcal{Z}) & \dots & E_{M-1}(Mn, \mathcal{Z}) \\ E_{M-1}(Mn-1, \mathcal{Z})\mathcal{Z}^{-1} & E_0(Mn-1, \mathcal{Z}) & \dots & E_{M-2}(Mn-1, \mathcal{Z}) \\ \vdots & \vdots & \ddots & \vdots \\ E_1(Mn-M+1, \mathcal{Z})\mathcal{Z}^{-1} & E_2(Mn-M+1, \mathcal{Z})\mathcal{Z}^{-1} & \dots & E_0(Mn-M+1, \mathcal{Z}) \end{bmatrix}, \quad (5.3.9)$$

where the scalar system $E_i(j, \mathcal{Z})$ is the i-th polyphase component of $H(j, \mathcal{Z})$. From (5.3.9), the following statements can be verified by directly evaluating the impulse response $h_k(n)$:

1. The scalar system $H(n, \mathcal{Z})$ is FIR if and only if the MIMO system $\mathbf{E}(n, \mathcal{Z})$ is FIR.
2. If the polyphase matrix $\mathbf{E}(n, \mathcal{Z})$ does not depend on n, then the scalar filter $H(n, \mathcal{Z})$ is a linear periodically time-varying filter of period M. This implies the filter coefficients satisfy $h_k(Ml+n) = h_k(n)$.
3. If the polyphase components satisfy $E_i(Mn-j, \mathcal{Z}) = E_i(Mn, \mathcal{Z})$ or $E_i(Mn+j, \mathcal{Z}) = E_i(Mn, \mathcal{Z})$ for $0 \le j \le M-1$, the polyphase matrix is said to be *time-varying pseudocirculant*(M). In this case, the filter coefficients satisfy $h_k(Ml-i) = h_k(Ml)$ or $h_k(Ml+i) = h_k(Ml)$ for $0 \le i \le M-1$.
4. $E_i(Mn-j, \mathcal{Z})$ are independent of both j and n (which implies both 2 and 3), if and only if the polyphase matrix $\mathbf{E}(n, \mathcal{Z})$ is time-invariant pseudocirculant. Therefore the scalar filter $H(n, \mathcal{Z})$ is LTI if and only if both 2 and 3 are true.

Referring to Fig. 5.3.5(c), since the delay chain and advance chain are simply the mechanisms of blocking and unblocking, we have $\sum_n |x(n)|^2 = \sum_n \mathbf{x}^\dagger(n)\mathbf{x}(n)$ and $\sum_n |y(n)|^2 = \sum_n \mathbf{y}^\dagger(n)\mathbf{y}(n)$.

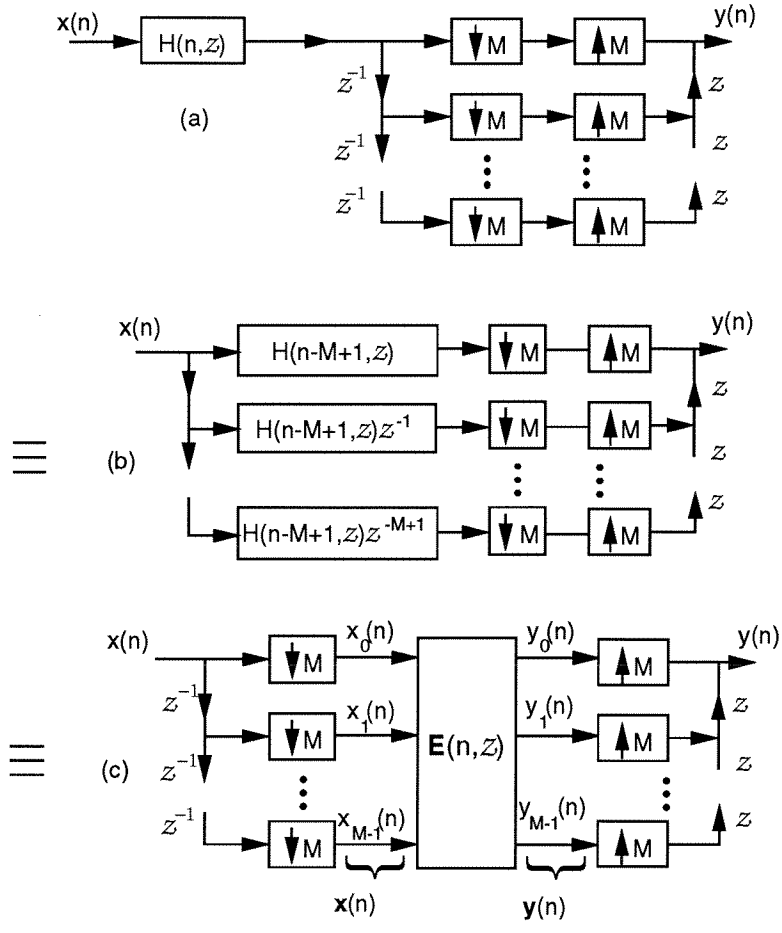

Fig. 5.3.5. Block implementation of the scalar LTV filter $H(n, \mathcal{Z})$, where the matrix $\mathbf{E}(n, \mathcal{Z})$ is defined in (5.3.9).

Therefore scalar system from $x(n)$ to $y(n)$ (i.e., the system $H(n, \mathcal{Z})$ in Fig. 5.3.5(a)) is lossless (passive) if and only if the MIMO system $\mathbf{E}(n, \mathcal{Z})$ in Fig. 5.3.5(c) is lossless (correspondingly passive). In particular, all MIMO LTI PU systems are lossless. If the MIMO system $\mathbf{E}(n, \mathcal{Z})$ in (5.3.9) is chosen as a general LTI PU system, then the resulting scalar system $H(n, \mathcal{Z})$ obtained by unblocking mechanism is lossless. Thus we conclude that *there are non trivial scalar LTV lossless systems.* The invertibility of the scalar systems obtained by unblocking is clearly equivalent to the invertibility of the original MIMO system. Let $\mathbf{R}(n, \mathcal{Z})$ be the blocked version of another filter $F(n, \mathcal{Z})$. It is clear that $\mathbf{R}(n, \mathcal{Z})$ is the inverse of $\mathbf{E}(n, \mathcal{Z})$ if and only if $F(n, \mathcal{Z})$ is the inverse of $H(n, \mathcal{Z})$. The idea of block implementation is useful since it can generate some examples to illuminate the theory of TVFB, such as the existence of non trivial scalar lossless system in Example 5.6.1 and the following 2×2 system which is shown to be *unfactorizable* in Section 4 of Chapter 6.

Generating MIMO Lossless LTV Systems by Partially Unblocking: Consider an $M \times M$ PU LTI system $\mathbf{E}(z)$. Let M_1 be a factor M. Then we can obtain a $M_1 \times M_1$ lossless LTV system by partially unblocking $\mathbf{E}(z)$.

To show how this can be done, we provide a simple example in the following:

Example 5.3.1. *MIMO Lossless LTV Systems from Unblocking:* Let $\mathbf{E}(z)$ be chosen as the following scaled 4×4 (permuted) Hadamard matrix:

$$\mathbf{T} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}. \tag{5.3.10}$$

Clearly $\mathbf{T}$ is unitary. In order to obtain a causal system, we use a delay chain at the synthesis end instead of an advance chain. After some rearrangements, we can get Fig. 5.3.6. Defining $\mathbf{x}(n) = [x_0(n)\ x_1(n)]^T$ and $\mathbf{y}(n) = [y_0(n)\ y_1(n)]^T$, we obtain the following 2×2 system $\mathcal{H}$: $\mathbf{y}(n) = \mathbf{e}_0(n)\mathbf{x}(n) + \mathbf{e}_1(n)\mathbf{x}(n-1) + \mathbf{e}_2(n)\mathbf{x}(n-2)$, where the coefficients are given by

$$\mathbf{e}_0(2n) = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{e}_1(2n) = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{e}_2(2n) = \mathbf{0},$$

$$\mathbf{e}_0(2n+1) = \mathbf{0}, \quad \mathbf{e}_1(2n+1) = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{e}_2(2n+1) = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}. \tag{5.3.11}$$

The above 2×2 system $\mathcal{H}$ is lossless since $\mathbf{T}$ is unitary. Its inverse which can be obtained by unblocking $\mathbf{T}^\dagger$, is also FIR. Thus $\mathcal{H}$ is a FIR lossless system with FIR inverse. ■

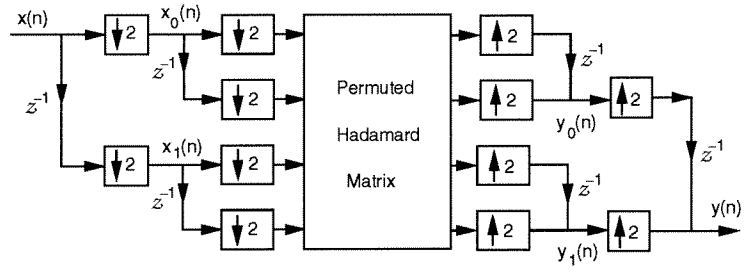

Fig. 5.3.6. Example 5.3.1–MIMO lossless system obtained by partial unblocking.

5.4. POLYPHASE APPROACH TO TVFB AND TRANSMULTIPLEXERS

Consider the TVFB given in Fig. 5.4.1(a). As we explained in Section 5.2.1, for convenience we will choose direct form A for the analysis filters $H^k(n, \mathcal{Z})$ and direct form B for the synthesis filters $F^k(\mathcal{Z}, n)$. Using the transform domain descriptions, the filters can be expressed as follows:

$$H^k(n, \mathcal{Z}) = \sum_i h_i^k(n)\mathcal{Z}^{-i}, \quad \text{and} \quad F^k(\mathcal{Z}, n) = \sum_i \mathcal{Z}^{-i} f_i^k(n), \quad \forall n, \tag{5.4.1}$$

where the superscript k is used to denote the filter number. The subband signals $y_k(n)$ and the output signal $\hat{x}(n)$ (as shown in Fig. 5.4.1(a)) can respectively be expressed as

$$y_k(n) = \sum_l h_l^k(Mn)x(Mn-l), \quad \hat{x}(n) = \sum_{k=0}^{M-1}\sum_m y_k(m) f_{n-Mm}^k(Mm). \tag{5.4.2}$$

From the above equation, it is clear that both the output and decimated subband signals are independent of those filter coefficients $h_i^k(n)$ that occur at $n \neq$ integer multiple of M. Therefore we need to consider only the coefficients $h_i^k(Mn)$ and $f_i^k(Mn)$. Applying the proposed polyphase representations, Fig. 5.4.1(a) can be redrawn as Fig. 5.4.1(b), where the noble identities have been invoked to move the polyphase matrices. The polyphase matrices $\mathbf{E}(Mn, \mathcal{Z})$ and $\mathbf{R}(\mathcal{Z}, Mn)$ are respectively defined as follows:

$$E_{kj}(Mn, \mathcal{Z}) = \left[H^k(n, \mathcal{Z})\mathcal{Z}^j\right]_{\downarrow M}, \quad R_{jk}(\mathcal{Z}, Mn) = \left[\mathcal{Z}^{-j}F^k(\mathcal{Z}, n)\right]_{\downarrow M}, \tag{5.4.3}$$

where the notation $[\bullet]_{\downarrow M}$ is defined in (5.3.7) and (5.3.8). In other words, the kj-th element of $\mathbf{E}(Mn, \mathcal{Z})$ and jk-th element of $\mathbf{R}(\mathcal{Z}, Mn)$ are respectively the j-th polyphase component of $H^k(n, \mathcal{Z})$ and $F^k(\mathcal{Z}, n)$. The relation between the analysis/synthesis filters and the polyphase matrices can be described as

$$[H^0(Mn, \mathcal{Z}) \ \ldots \ H^{M-1}(Mn, \mathcal{Z})]^T = \mathbf{E}(Mn, \mathcal{Z}^M)[1 \ \mathcal{Z}^{-1} \ \ldots \ \mathcal{Z}^{-M+1}]^T, \tag{5.4.4a}$$

$$[F^0(\mathcal{Z}, Mn) \ \ldots \ F^{M-1}(\mathcal{Z}, Mn)] = [1 \ \mathcal{Z} \ \ldots \ \mathcal{Z}^{M-1}]\mathbf{R}(\mathcal{Z}^M, Mn). \tag{5.4.4b}$$

Only $H^k(Mn, \mathcal{Z})$ and $F^k(\mathcal{Z}, Mn)$ are related to the polyphase matrices. For any $i \neq$ multiple of M, $H^k(Mn - i, \mathcal{Z})$ and $F^k(\mathcal{Z}, Mn - i)$ are irrelevant to the output of the FB. Because of the polyphase representation, we can characterize all TVFBs by characterizing the MIMO systems $\mathbf{E}(Mn, \mathcal{Z})$ and $\mathbf{R}(\mathcal{Z}, Mn)$.

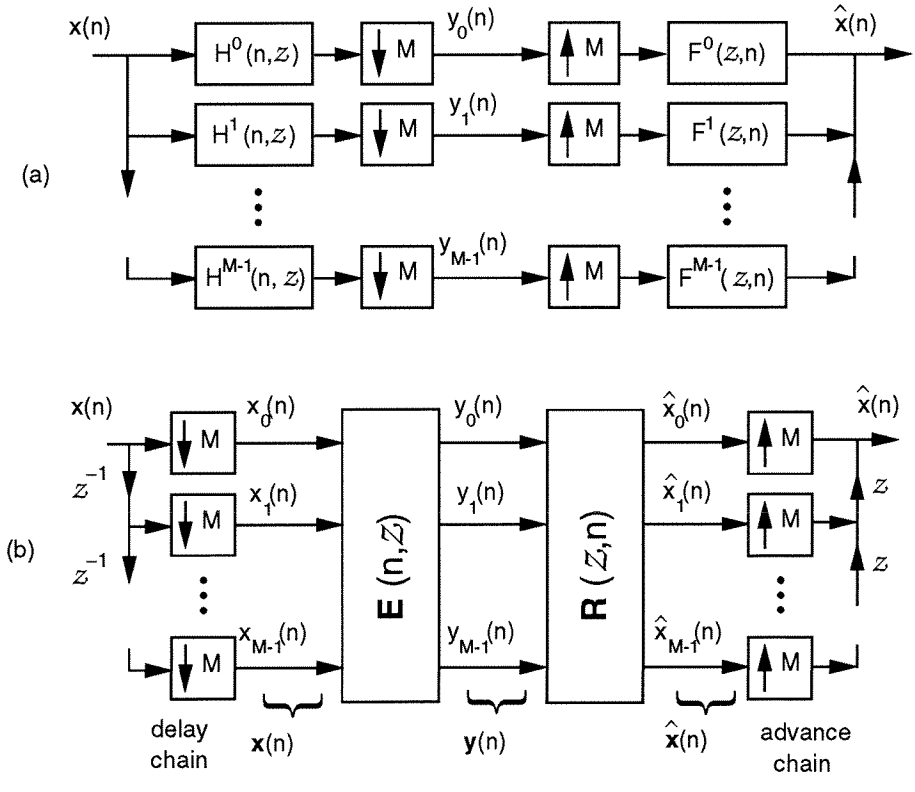

Fig. 5.4.1. Time-varying filter bank and its polyphase implementation.

5.4.1. Time-Varying PR Filter Bank

From the polyphase implementation of the TV filter bank shown in Fig. 5.4.1(b), it is clear that the filter bank achieves PR if and only if

$$\mathbf{R}(\mathcal{Z}, Mn)\mathbf{E}(Mn, \mathcal{Z}) = \mathbf{I}, \quad \forall n. \tag{5.4.5}$$

In other words, we obtain PR if $\mathbf{R}(\mathcal{Z}, Mn)$ is the inverse of MIMO filters $\mathbf{E}(Mn, \mathcal{Z})$. In the general LTV case, eq. (5.4.5) does not imply $\mathbf{E}(Mn, \mathcal{Z})\mathbf{R}(\mathcal{Z}, Mn) = \mathbf{I}$. To see this, consider Example 5.1.1, where an one-channel TVFB is given. In this example, $M = 1$, so $\mathbf{E}(Mn, \mathcal{Z})$ and $\mathbf{R}(\mathcal{Z}, Mn)$ are respectively the scalar filters $\mathcal{H}$ and $\mathcal{G}$. Clearly $\mathbf{E}(Mn, \mathcal{Z})\mathbf{R}(\mathcal{Z}, Mn) \neq \mathbf{I}$ because $\mathbf{R}(\mathcal{Z}, Mn)$ (in this case $\mathcal{G}$) is not invertible! Therefore in general if we interchange the analysis and synthesis polyphase matrix of a PR TVFB, the PR property will be destroyed. This is very different from the conventional PR LTI FB.

However if the matrix $\mathbf{R}(\mathcal{Z}, Mn)$ is also invertible, then its inverse is unique and it is $\mathbf{E}(Mn, \mathcal{Z})$. To prove this, let $\mathbf{R}^{-1}(\mathcal{Z}, Mn)$ be *any* inverse of $\mathbf{R}(\mathcal{Z}, Mn)$. Premultiplying both sides of (5.4.5) by $\mathbf{R}^{-1}(\mathcal{Z}, Mn)$, we get $\mathbf{R}^{-1}(\mathcal{Z}, Mn) = \mathbf{E}(Mn, \mathcal{Z})$. In this case, the cascade $\mathbf{E}(Mn, \mathcal{Z})\mathbf{R}(\mathcal{Z}, Mn) = \mathbf{I}$. This implies the following relation between the analysis and synthesis filters:

$$\Big[H^k(n, \mathcal{Z})F^l(\mathcal{Z}, n)\Big]_{\downarrow M} = \sum_{i=0}^{M-1} E_{ki}(Mn, \mathcal{Z})R_{il}(\mathcal{Z}, Mn) = \delta(k-l), \quad \forall n. \tag{5.4.6}$$

The above equation can be viewed as a generalization of the biorthogonality condition of the analysis/synthesis filters defined for the LTI case [Djo94, Dau92]. To illustrate the above theory, we provide an example in the following:

Example 5.4.1. *PR FIR TVFB:* Consider the M-channel TVFB whose polyphase matrices are

$$\mathbf{E}(Mn, \mathcal{Z}) = \Big[\mathbf{I} - \mathbf{u}(n)\mathbf{v}^\dagger(n)\Big] + \mathbf{u}(n)\mathbf{v}^\dagger(n-1)\mathcal{Z}^{-1}, \tag{5.4.7a}$$

$$\mathbf{R}(\mathcal{Z}, Mn) = \Big[\mathbf{I} + \mathbf{u}(n)\mathbf{v}^\dagger(n)\Big] - \mathcal{Z}^{-1}\mathbf{u}(n+1)\mathbf{v}^\dagger(n), \tag{5.4.7b}$$

where the vectors satisfy $\mathbf{v}^\dagger(n)\mathbf{u}(n) = 0$ for all n. One can verify by direct substitution that the polyphase matrices in (5.4.7) satisfy $\mathbf{R}(\mathcal{Z}, Mn)\mathbf{E}(Mn, \mathcal{Z}) = \mathbf{E}(Mn, \mathcal{Z})\mathbf{R}(\mathcal{Z}, Mn) = \mathbf{I}$. Therefore the TVFB achieves PR and the synthesis polyphase matrix is invertible. The analysis and synthesis filters are respectively

$$H^k(Mn, \mathcal{Z}) = \big(1 - u_k(n)v_k(n)\big)\mathcal{Z}^{-k} - \sum_{i \neq k} u_k(n)v_i(n)\mathcal{Z}^{-i} + \sum_{i=0}^{M-1} u_k(n)v_i(n-1)\mathcal{Z}^{-M-i}, \tag{5.4.8a}$$

$$F^k(\mathcal{Z}, Mn) = \mathcal{Z}^k\big(1 + u_k(n)v_k(n)\big) + \sum_{i \neq k} \mathcal{Z}^i u_i(n)v_k(n) - \sum_{i=0}^{M-1} \mathcal{Z}^{-M+i} u_i(n)v_k(n-1), \tag{5.4.8b}$$

where $u_i(n)$ and $v_i(n)$ are respectively the i-th element of $\mathbf{u}(n)$ and $\mathbf{v}(n)$. We have a PR TVFB where both its analysis and synthesis filters are causal FIR. One can verify that the filters satisfy the biorthogonal condition in (5.4.6). ■

Remark: Applying the result in Section 5.3.3, we know that the TVFB reduces to a scalar LTI system if and only if the cascade $\mathbf{R}(\mathcal{Z},n)\mathbf{E}(n,\mathcal{Z})$ is a time-invariant pseudocirculant matrix. Compare this with the result that an LTI FB reduces to a scalar LTI system if and only if it is alias free [Vai93].

Time-Varying Tilde Operation: In the LTI case, the tilde operation was used in [Vai93] to the study of conventional FBs. Given a LTI system $\mathbf{E}(z)$, the tilde operation is defined in Section 1.2 as $\widetilde{\mathbf{E}}(z) = \mathbf{E}^\dagger(1/z^*)$. This tilde operation was proved to be very useful in the analysis of conventional FB, especially for the LTI PU FBs. In the LTV case, we will define a similar operation. The time-varying tilde operation consists of three steps:

1. Replace the multipliers $\mathbf{e}_k(n)$ and $\mathbf{r}_k(n)$ with their transpose complex-conjugates, $\mathbf{e}_k^\dagger(n)$ and $\mathbf{r}_k^\dagger(n)$ respectively.
2. Interchange the multiplier and the delay, i.e., $\mathbf{e}_k(n)\mathcal{Z}^{-k}$ is replaced with $\mathcal{Z}^{-k}\mathbf{e}_k(n)$ and $\mathcal{Z}^{-k}\mathbf{r}_k(n)$ is replaced with $\mathbf{r}_k(n)\mathcal{Z}^{-k}$. Note that this operation will change the direct form A to the direct form B structure, and vice verse.
3. Replace the delay $\mathcal{Z}^{-1}$ with the advance element $\mathcal{Z}$.

   By using the above definition, the tilde operation on LTV filters can be described as

$$\widetilde{\mathbf{E}}(n,\mathcal{Z}) = \sum_k \mathcal{Z}^k \mathbf{e}^\dagger(n), \quad \widetilde{\mathbf{R}}(\mathcal{Z},n) = \sum_k \mathbf{r}_k^\dagger(n)\mathcal{Z}^k. \tag{5.4.9}$$

Note that $\mathbf{E}(n,\mathcal{Z})$ and $\mathbf{R}(\mathcal{Z},n)$ are respectively in direct form A and B while $\widetilde{\mathbf{E}}(n,\mathcal{Z})$ and $\widetilde{\mathbf{R}}(\mathcal{Z},n)$ are respectively in direct form B and A. The tilde operation will be shown to be very useful as we shall repeatedly see later. Note that the tilde operation is its own inverse, i.e., $\widetilde{\widetilde{\mathbf{E}}}(n,\mathcal{Z}) = \mathbf{I}$. Moreover, the tilde of a cascade of two LTV filters satisfies the following property:

Lemma 5.4.1. Consider two LTV systems $\mathcal{G}_0$ followed by $\mathcal{G}_1$. The tilde of the cascade system is given by $\widetilde{\mathcal{G}_1\mathcal{G}_0} = \widetilde{\mathcal{G}}_0\widetilde{\mathcal{G}}_1$. ■

Proof: We will only prove the fact for the case when both $\mathcal{G}_0$ and $\mathcal{G}_1$ are in direct form A implementation, namely $\mathbf{E}^0(n,\mathcal{Z})$ and $\mathbf{E}^1(n,\mathcal{Z})$ respectively. The proof for other cases being similar. Using (5.2.8) and (5.4.9), we get

$$\begin{aligned}\text{tilde}\left[\mathbf{E}^1(n,\mathcal{Z})\mathbf{E}^0(n,\mathcal{Z})\right] &= \text{tilde}\left[\sum_{k,l}\mathbf{e}_k^1(n)\mathbf{e}_l^0(n-k)\mathcal{Z}^{-k-l}\right] = \sum_{k,l}\mathcal{Z}^{k+l}\mathbf{e}_l^{0\dagger}(n-k)\mathbf{e}_k^{1\dagger}(n)\\ &= \sum_{k,l}\mathcal{Z}^l\mathbf{e}_l^{0\dagger}(n)\mathcal{Z}^k\mathbf{e}_k^{1\dagger}(n) = \widetilde{\mathbf{E}}^0(n,\mathcal{Z})\widetilde{\mathbf{E}}^1(n,\mathcal{Z}).\end{aligned} \tag{5.4.10}$$

The proof is complete. ■

Interchangability of the Analysis/Synthesis Filters of a PR TVFB: In the LTI case, if $\{H^k(z), F^k(z)\}$ form a PR FB, then $\{F^k(z), H^k(z)\}$ will also form a PR analysis/synthesis system (Problem 5.17 of [Vai93]). In the LTV case, if we directly interchange $H^k(n,\mathcal{Z})$ with $F^k(\mathcal{Z},n)$ (and vice verse) without any modification, the PR property in general does not continue to hold. To see this, consider Example 5.1.1, where we show

a one channel PR TVFB with $\mathcal{H}$ as the only analysis filter and $\mathcal{G}$ as the only synthesis filter. Clearly interchanging $\mathcal{G}$ with $\mathcal{H}$ does not preserve the PR property because $\mathcal{G}$ is not invertible! The proper way of interchanging the analysis and synthesis filters is described in the following theorem, which follows directly from (5.4.4) and Lemma 5.4.1.

Theorem 5.4.1. *Interchanging the Analysis and Synthesis Filters:* Let $\{H^k(n,\mathcal{Z}), F^k(\mathcal{Z},n)\}$ be respectively the analysis and synthesis filters of a PR TVFB, then the FB with $\widetilde{F}^k(\mathcal{Z},n)$ as the analysis filters and $\widetilde{H}^k(n,\mathcal{Z})$ as the synthesis filters also achieves PR. ■

5.4.2. Time-Varying PR Transmultiplexers

Transmultiplexers have been used in communication to convert between two formats called *time-division multiplexed* (TDM) format and *frequency-division multiplexed* (FDM) format. The application of conventional FB theory to transmultiplexers was studied in [Vet86, Vai93]. In this subsection, we will generalize the results in [Vet86, Vai93] to the more general LTV case. Consider Fig. 5.4.2 where a TV transmultiplexer is shown. In the traditional theory of transmultiplexers, we have the following two special cases: (i) When $F^k(\mathcal{Z},n) = \mathcal{Z}^k$ and $H^k(n,\mathcal{Z}) = \mathcal{Z}^{-k}$, $y(n)$ is a TDM signal; (ii) When $F^k(\mathcal{Z},n)$ and $H^k(n,\mathcal{Z})$ are LTI ideal bandpass filters, $y(n)$ is a FDM signal. In the above two cases, it is clear that $\hat{\mathbf{x}}(n) = \mathbf{x}(n)$, i.e., the transmultiplexer achieves PR. By using the theory for PR LTI FBs, the author in [Vet86] showed that PR LTI transmultiplexer is possible by using non ideal LTI filters. More precisely, it was shown that if the LTI filters $\{H^k(z), F^k(z)\}$ form a PR analysis/synthesis system, then the corresponding LTI transmultiplexer achieves PR. In the more general LTV case, a similar statement does not hold as shown in the following: By using the polyphase representation, Fig. 5.4.2 can be redrawn as Fig. 5.4.3, where the elements of the matrices $\mathbf{E}(Mn,\mathcal{Z})$ and $\mathbf{R}(\mathcal{Z},Mn)$ are defined in (5.4.3). It is clear from Fig. 5.4.3 that the TV transmultiplexer achieves PR if and only if $\mathbf{E}(Mn,\mathcal{Z})\mathbf{R}(\mathcal{Z},Mn) = \mathbf{I}$. However from the previous subsection, we know that the PR property of $\{H^k(n,\mathcal{Z}), F^k(\mathcal{Z},n)\}$ does not implies $\mathbf{E}(Mn,\mathcal{Z})\mathbf{R}(\mathcal{Z},Mn) = \mathbf{I}$. Therefore the transmultiplexer in Fig. 5.4.2 may not be PR even if $\{H^k(n,\mathcal{Z}), F^k(\mathcal{Z},n)\}$ forms a PR TVFB unless the synthesis system is also invertible (which is not always true). In order to achieve PR TV transmultiplexer, we can use the result of Lemma 5.4.1 to show that PR is attained if $\widetilde{H}^k(n,\mathcal{Z})$ are used as the filters in the multiplexer and $\widetilde{F}^k(\mathcal{Z},n)$ are used as the filters in the demultiplexer.

5.5. LOSSLESS LTV FILTERS AND FILTER BANKS

In the previous section, using the polyphase representation, we have shown that we can characterize all TVFBs by characterizing MIMO LTV filters. Moreover the analysis bank preserves the energy from the input to the subband signals $y_k(m)$ (as shown in Fig. 5.4.1) if and only if the corresponding analysis polyphase matrix is lossless. In this section, we are going to study in detail lossless LTV filters and their inverses. Lossless TVFBs are important in the applications of subband coding since the losslessness implies that the coding gain ≥ 1.

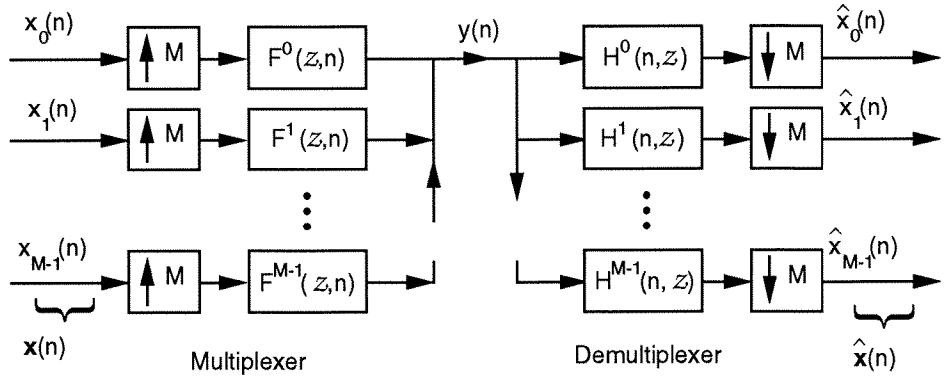

Fig. 5.4.2. Time-varying transmultiplexer.

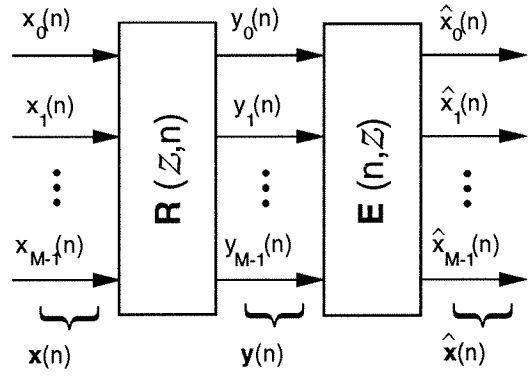

Fig. 5.4.3. Redrawing of the time-varying transmultiplexer by using the polyphase representation.

5.5.1. Characterizations of Lossless LTV Filters and Their Inverses

Impulse Response Characterization: Consider the following MIMO LTV system:

$$\mathbf{y}(n) = \sum_k \mathbf{e}_k(n)\mathbf{x}(n-k). \tag{5.5.1}$$

By definition (Section 5.1.2), the above system is lossless if $\sum_n \mathbf{y}^\dagger(n)\mathbf{y}(n) = \sum_n \mathbf{x}^\dagger(n)\mathbf{x}(n)$ for any $\mathbf{x}(n) \in l_2(M)$. Using (5.5.1), we have

$$\sum_n \mathbf{y}^\dagger(n)\mathbf{y}(n) = \sum_{k,l,n} \mathbf{x}^\dagger(n-k)\mathbf{e}_k^\dagger(n)\mathbf{e}_l(n)\mathbf{x}(n-l). \tag{5.5.2}$$

Making a change of variable, we get the following:

$$\sum_n \mathbf{y}^\dagger(n)\mathbf{y}(n) = \sum_{i,j} \mathbf{x}^\dagger(i) \underbrace{\left[\sum_k \mathbf{e}_k^\dagger(k+n)\mathbf{e}_{k+i-j}(k+n)\right]}_{\mathbf{T}_{ij}} \mathbf{x}(j). \tag{5.5.3}$$

Since the input $\mathbf{x}(n)$ is arbitrary, the right-hand side of (5.5.3) equals $\sum_n \mathbf{x}^\dagger(n)\mathbf{x}(n)$ if and only if the matrices $\mathbf{T}_{ij}$ satisfy $\mathbf{T}_{ij} = \mathbf{I}\delta(i-j)$. Therefore we conclude that the system in (5.5.1) is lossless if and

only if

$$\sum_k \mathbf{e}_k^\dagger(k+n)\mathbf{e}_{k+l}(k+n) = \mathbf{I}\delta(l), \quad \forall n. \tag{5.5.4}$$

Note that the left-hand side of (5.5.4) is different from $\sum_k \mathbf{e}_k^\dagger(n)\mathbf{e}_{k+l}(n)$. In the LTI case, eq. (5.5.4) reduces to $\sum_k \mathbf{e}_k^\dagger \mathbf{e}_{k+l} = \mathbf{I}\delta(l)$. This is consistent with the PU condition [Vai93] $\mathbf{E}^\dagger(1/z^*)\mathbf{E}(z) = \mathbf{I}$, where $\mathbf{E}(z) = \sum_k \mathbf{e}_k z^{-k}$.

For the direct form B implementation of LTV filter which can be described as: $\mathbf{y}(n) = \sum_k \mathbf{r}_k(n-k)\mathbf{x}(n-k)$, the lossless condition for the direct form B can be obtained by simply replacing the coefficients $\mathbf{e}_k(n)$ in (5.5.4) with $\mathbf{r}_k(n-k)$. Therefore a direct form B system is lossless if and only if $\sum_k \mathbf{r}_k^\dagger(n)\mathbf{r}_{k+l}(n-l) = \mathbf{I}\delta(l)$ for all n.

Lossless LTV FIR Systems: Assume that the LTV filter described in (5.5.1) is a N-th order system, i.e., $\mathbf{y}(n) = \sum_{k=0}^{N} \mathbf{e}_k(n)\mathbf{x}(n-k)$. Then by substituting $l = N$ into the lossless condition in (5.5.4), we obtain $\mathbf{e}_0^\dagger(n)\mathbf{e}_N(n) = \mathbf{0}$ for all n. That is, the lowest and highest order coefficients of a lossless FIR system are both singular for each n (if neither of them equals to the null matrix). The sum of their rank cannot exceed M. In the next chapter, we will see that this property is very useful in the factorization of lossless LTV systems in terms of smaller building blocks.

Non Losslessness of the Frozen System: Consider the following system: $\mathbf{y}(n) = \sum_k \mathbf{e}_k(L)\mathbf{x}(n-k)$, where L is some fixed integer. The above system can be thought as the system in (5.5.1) with the coefficients frozen at time L. Note that in general the frozen system, which is a LTI system, does not satisfy the PU condition. Therefore the frozen system might not be lossless!

Replacing Delay z^{-1} with z^L: We know that if we replace the delay z^{-1} in an implementation of a lossless LTI system with z^L for some integer L, the system remains lossless. This is not true in general for the LTV case! To see this, consider Example 5.1.1. If we replace the delay $\mathcal{Z}^{-1}$ in the structure of Fig. 5.2.1 with an advance operator $\mathcal{Z}$, the new system will be $y(n) = h_0(n)x(n) + h_1(n)x(n+L)$ with $h_0(n)$ and $h_1(n)$ given in (5.1.1). The new system is no longer lossless because the samples $x(0)$, ..., $x(L)$ are missing. In the following, we will show how to modify the coefficients such that losslessness is preserved under such transformation.

Theorem 5.5.1. *Delay Transformation:* Consider a lossless system $\mathbf{e}_k(n)$, $0 \le k \le N$, shown in Fig. 5.2.1. If the delay z^{-1} is replaced with z^{-L}, the direct form A implementations in Fig. 5.2.1 will remain lossless provided that the coefficients for the new system are obtained as:

$$\widehat{\mathbf{e}}_k(Ln+i) = \mathbf{e}_k(n), \quad \text{or} \quad \widehat{\mathbf{e}}_k(Ln-i) = \mathbf{e}_k(n), \quad \text{for} \quad 0 \le i \le L-1, \ \forall n. \tag{5.5.5}$$

Similarly if we replace z^{-1} with z^{-L} in the direct form B shown Fig. 5.2.2, then the system will remain lossless if $\widehat{\mathbf{r}}_k(Ln \pm i) = \mathbf{r}_k(n)$. ■

Theorem 5.5.1 can be proved by direct substitution. Recall from Section 5.3.3 that a LTV system is TV pseudocirculant(M) if $\mathbf{e}_k(Mn-i) = \mathbf{e}_k(Mn)$ or $\mathbf{e}_k(Mn+i) = \mathbf{e}_k(Mn)$ for $0 \le i \le M-1$. Therefore a lossless system remains lossless under the delay transformation if and only if it is TV pseudocirculant(L).

The Inverses for Lossless LTV Filters: For the general LTV filter, it is not easy to determine if the filter is invertible and find its inverse if it is invertible. However if the LTV filter is lossless, it is *always* invertible. Moreover the inverse is FIR if the lossless system is FIR. However unlike the LTI case, the inverse may not be unique, invertible, or lossless as we have demonstrated in Example 5.1.1. In the following, we will construct an anticausal inverse for an LTV lossless system. In the next subsection, we will show that the invertibility and the losslessness of the inverse are closely related to each other. In the LTI case, given a MIMO PU system $\mathbf{E}(z)$, we know [Vai93] that the inverse is simply $\mathbf{E}^\dagger(1/z^*)$. This suggests that the inverse of a LTV lossless system of the form in Fig. 5.2.1 might be in the form in Fig. 5.5.1. This indeed is the case, as verified next. The output of the system in Fig. 5.5.1 can be written as:

$$\hat{\mathbf{x}}(n) = \sum_k \mathbf{e}_k^\dagger(n+k)\mathbf{y}(n+k). \tag{5.5.6}$$

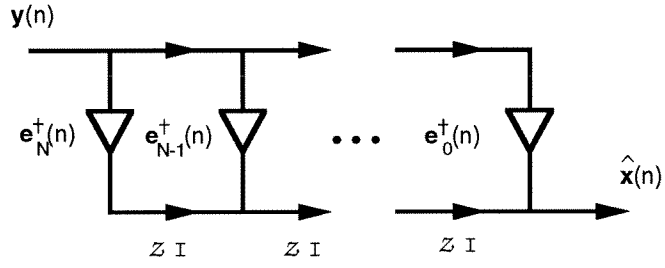

Fig. 5.5.1. Inverse of the lossless system in Fig. 5.2.1.

Substituting (5.5.1) into the above equation and simplifying the result, we have

$$\hat{\mathbf{x}}(n) = \sum_l \left(\sum_k \mathbf{e}_k^\dagger(n+k)\mathbf{e}_{k+l}(n+k) \right) \mathbf{x}(n-l). \tag{5.5.7}$$

Applying the lossless condition (5.5.4) to the above equation, we have $\hat{\mathbf{x}}(n) = \mathbf{x}(n)$ for all n. Therefore we have shown the following:

Theorem 5.5.2. *Inversion of Lossless LTV Systems:* All lossless LTV systems are invertible. Moreover the inverse is FIR as in Fig. 5.1 if the original lossless system is FIR as in Fig. 5.2.1. ■

Unlike the LTI case, we will see later that the inverse of a LTV lossless system may not be unique. Notice that the anticausal inverse given above is implemented in direct form B. Except for this, everything is similar to the LTI case, i.e., the coefficients are mirror-image and transpose-conjugates of the coefficients of the original system. By using a procedure similar to that of Section 5.5.1, one can show that the inverse system in Fig. 5.5.1 is lossless if and only if

$$\sum_k \mathbf{e}_k(n)\mathbf{e}_{k+l}^\dagger(n+l) = \mathbf{I}\delta(l) \quad \forall n. \tag{5.5.8}$$

Therefore if the coefficients of the system in Fig. 5.2.1 satisfy both (5.5.4) and (5.5.8), then it is a lossless system with a lossless inverse. In the LTI case, (5.5.4) and (5.5.8) respectively reduce to the conditions $\sum_k \mathbf{e}_k^\dagger \mathbf{e}_{k+l} = \mathbf{I}\delta(l)$ and $\sum_k \mathbf{e}_k \mathbf{e}_{k+l}^\dagger = \mathbf{I}\delta(l)$. The two conditions in the LTI case are equivalent to each other. Therefore the inverse of a LTI PU system is also PU. However in the LTV case, (5.5.4) does *not* imply (5.5.8). Therefore the inverse of a LTV lossless system might not be lossless (Example 5.6.3).

Transform Domain Characterization of Lossless Filters and Their Inverses: First recall the transform domain description introduced in Section 5.2.2 and the tilde operation defined in Section 5.4.1. Consider the direct form A filter $\mathbf{E}(n, \mathcal{Z})$ in Fig. 5.2.1. Suppose we cascade $\widetilde{\mathbf{E}}(n, \mathcal{Z})$ after the filter $\mathbf{E}(n, \mathcal{Z})$, the resulting system is:

$$\widetilde{\mathbf{E}}(n, \mathcal{Z})\mathbf{E}(n, \mathcal{Z}) = \sum_{i,k} \mathcal{Z}^k \mathbf{e}_k^\dagger(n)\mathbf{e}_i(n)\mathcal{Z}^{-i} = \sum_{i,k} \mathbf{e}_k^\dagger(n+k)\mathbf{e}_i(n+k)\mathcal{Z}^{k-i} = \sum_{k,l} \mathbf{e}_k^\dagger(n+k)\mathbf{e}_{k+l}(n+k)\mathcal{Z}^{-l}. \tag{5.5.9}$$

If the system $\mathbf{E}(n, \mathcal{Z})$ is lossless, i.e., the coefficients $\mathbf{e}_k(n)$ satisfy (5.5.4), then (5.5.9) reduces to the following:

$$\widetilde{\mathbf{E}}(n, \mathcal{Z})\mathbf{E}(n, \mathcal{Z}) = \mathbf{I}. \tag{5.5.10}$$

Therefore a LTV filter $\mathbf{E}(n, \mathcal{Z})$ is lossless if and only if (5.5.10) holds. The beauty of (5.5.10) is that it directly tells us the inverse filter is $\widetilde{\mathbf{E}}(n, \mathcal{Z})$! Note that the inverse $\widetilde{\mathbf{E}}(n, \mathcal{Z})$ is in direct form B while $\mathbf{E}(n, \mathcal{Z})$ is in the direct form A. On the other hand, if a lossless filter is in direct form B, $\mathbf{R}(\mathcal{Z}, n)$, its inverse can be shown to be the direct form A filter $\widetilde{\mathbf{R}}(\mathcal{Z}, n)$. Summarizing the result, we conclude that the inverse of a lossless system $\mathcal{H}$ is given by $\widetilde{\mathcal{H}}$.

5.5.2. Subtle Properties of Inverses of Lossless LTV Systems

From Theorem 5.5.2, we know that lossless LTV systems are always invertible. However the uniqueness, invertibility and losslessness properties of the inverse are not guaranteed (Example 5.1.1). It turns out that these properties are closely related to each other. More precisely, we can prove the following:

Theorem 5.5.3. *Losslessness, Uniqueness and Invertibility of the Inverse:* Given a lossless LTV system $\mathcal{H}$, let $\mathcal{G}$ be one of its inverses. Then the following are equivalent:

(a) $\mathcal{G}$ is *lossless*;

(b) $\mathcal{G}$ is *invertible* and its unique inverse is $\mathcal{H}$;

(c) $\mathcal{H}$ maps $l_2(M)$ **onto** $l_2(M)$. ■

(i) *Proof that (a) implies (b)*: Let the inverse $\mathcal{G}$ be lossless. By Theorem 5.5.2, we know that $\mathcal{G}$ is invertible. We need only to prove that the inverse of $\mathcal{G}$ is unique and equal to $\mathcal{H}$. Let $\mathcal{F}$ be *any* inverse of $\mathcal{G}$, so $\mathcal{F}\mathcal{G} = \mathcal{I}$. By definition, $\mathcal{G}$ is an inverse of $\mathcal{H}$, so we also have $\mathcal{G}\mathcal{H} = \mathcal{I}$. Premultipling both side by $\mathcal{F}$, we get $\mathcal{F}\mathcal{G}\mathcal{H} = \mathcal{F}$. Using the fact $\mathcal{F}\mathcal{G} = \mathcal{I}$ we arrive at $\mathcal{H} = \mathcal{F}$. Thus the *only* inverse of $\mathcal{G}$ is $\mathcal{H}$.

(ii) *Proof that (b) implies (c)*: Suppose the mapping $\mathcal{H}$ from $l_2(M)$ to $l_2(M)$ is not onto. Then there exists $\mathbf{y}_0(n) \in l_2(M)$ such that $\mathbf{y}_0(n) \neq \mathcal{H}\mathbf{x}(n)$ for any $\mathbf{x}(n) \in l_2$. Since $\mathcal{G}$ is invertible, and its unique inverse is $\mathcal{H}$, we have $\mathcal{H}\mathcal{G} = \mathcal{I}$. Applying the previous equation to the signal $\mathbf{y}_0(n)$, we have $\mathcal{H}[\mathcal{G}\mathbf{y}_0(n)] = \mathbf{y}_0(n)$. That means $\mathbf{y}_0(n)$ is in the range of $\mathcal{H}$, a contradiction! Therefore (b) implies (c).

(iii) *Proof that (c) implies (a)*: By the onto property, for all possible sequences $\mathbf{y}(n) \in l_2$, there exists a sequence $\mathbf{x}(n) \in l_2$ such that $\mathcal{H}\mathbf{x}(n) = \mathbf{y}(n)$. Since $\mathcal{H}$ is lossless, we have $\sum_n \mathbf{x}^\dagger(n)\mathbf{x}(n) = \sum_n \mathbf{y}^\dagger(n)\mathbf{y}(n)$. But $\mathcal{G}\mathcal{H}\mathbf{x}(n) = \mathcal{G}\mathbf{y}(n) = \mathbf{x}(n)$ (because $\mathcal{G}\mathcal{H} = \mathcal{I}$). Thus, for all $\mathbf{y}(n) \in l_2$, the signal $\mathcal{G}\mathbf{y}(n)$ has the same energy as $\mathbf{y}(n)$. So $\mathcal{G}$ is lossless. ■

Corollary 5.5.1. *Uniqueness of the Inverse:* If any of the three conditions in Theorem 5.5.3 is true, then the inverse $\mathcal{G}$ given in Theorem 5.5.3 is unique. ■

Proof: We will prove the condition (b) implies that $\mathcal{G}$ is unique. Suppose that there are two different systems $\mathcal{G}_1$ and $\mathcal{G}_2$ such that $\mathcal{G}_1\mathcal{H} = \mathcal{G}_2\mathcal{H} = \mathcal{I}$. From (b), we are given that $\mathcal{G}_2$ has the unique inverse $\mathcal{H}$, i.e. $\mathcal{H}\mathcal{G}_2 = \mathcal{I}$. Premultipling $\mathcal{G}_1$ to the previous equation, we have $\mathcal{G}_1\mathcal{H}\mathcal{G}_2 = \mathcal{G}_1$. Using the identity $\mathcal{G}_1\mathcal{H} = \mathcal{I}$, we have $\mathcal{G}_1 = \mathcal{G}_2$. ■

As a consequence of the above theorem and corollary, the losslessness, invertibility and uniqueness of an inverse of a lossless systems are the same. They are equivalent to the completeness of the range of the original system $\mathcal{H}$. If one of the inverses of a lossless system is non invertible, then all of its inverses are non invertible (hence non lossless). Therefore, an invertible inverse lossless (IIL) system *always* has a unique lossless inverse while a non invertible inverse lossless (NIL) system can *never* have a lossless inverse. Even though the inverses of a NIL system are not lossless, we will show in the following that there will always exist a *unique passive* inverse.

Theorem 5.5.4. *Existence of Unique Passive Inverse:* Given a NIL system $\mathcal{H}$, there is always a *unique* passive inverse. The unique passive inverse is given by $\widetilde{\mathcal{H}}$. ■

Proof: We split the proof into three parts: (i) The existence of passive inverse; (ii) the uniqueness of passive inverse; and (iii) the passiveness of $\widetilde{\mathcal{H}}$. For part (iii), it is sufficient to establish the passiveness of $\widetilde{\mathcal{H}}$ since $\widetilde{\mathcal{H}}$ is always an inverse of $\mathcal{H}$ (Section 5.5.1).

(i) Let $\mathcal{H}$ be a NIL system and $\mathcal{G}$ be *any* inverse of $\mathcal{H}$. Let $\mathcal{R}(\mathcal{H}) \subset l_2(M)$ denote the range of $\mathcal{H}$ and $\mathcal{R}^\perp(\mathcal{H}) \subset l_2(M)$ be the orthogonal complement of $\mathcal{R}(\mathcal{H})$, i.e.,

$$\mathcal{R}(\mathcal{H}) = \{\mathbf{y}(n) \in l_2(M) \mid \mathbf{y}(n) = \mathcal{H}\mathbf{x}(n), \text{ for some } \mathbf{x}(n) \in l_2(M)\}, \tag{5.5.11a}$$

$$\mathcal{R}^\perp(\mathcal{H}) = \{\mathbf{y}(n) \in l_2(M) \mid \langle \mathcal{H}\mathbf{x}(n), \mathbf{y}(n)\rangle = 0, \text{ for all } \mathbf{x}(n) \in l_2(M)\}. \tag{5.5.11b}$$

Since $\mathcal{H}$ is a linear system, both $\mathcal{R}(\mathcal{H})$ and $\mathcal{R}^\perp(\mathcal{H})$ are subspaces of $l_2(M)$ and $\mathcal{R}(\mathcal{H}) \oplus \mathcal{R}^\perp(\mathcal{H}) = l_2(M)$. Since $\mathcal{G}$ is an inverse of $\mathcal{H}$, it maps $\mathcal{R}(\mathcal{H})$ *onto* l_2. Let $\mathcal{G}_0$ be the following linear system:

$$\mathcal{G}_0\mathbf{y}(n) = \begin{cases} \mathcal{G}\mathbf{y}(n), & \text{for } \mathbf{y}(n) \in \mathcal{R}(\mathcal{H}); \\ \mathbf{0}, & \text{for } \mathbf{y}(n) \in \mathcal{R}^\perp(\mathcal{H}). \end{cases} \tag{5.5.12}$$

Clearly $\mathcal{G}_0$ is also an inverse of $\mathcal{H}$. For any $\mathbf{y}(n) \in l_2$, there are unique $\mathbf{y}_0(n) \in \mathcal{R}^{\perp}(\mathcal{H})$ and $\mathbf{y}_1(n) \in \mathcal{R}(\mathcal{H})$ such that $\mathbf{y}(n) = \mathbf{y}_0(n)+\mathbf{y}_1(n)$. The norm of the vectors satisfies $\| \mathbf{y}(n) \|^2 = \| \mathbf{y}_0(n) \|^2 + \| \mathbf{y}_1(n) \|^2$ because $\langle \mathbf{y}_0(n), \mathbf{y}_1(n) \rangle = 0$. Applying $\mathcal{G}_0$ to the input $\mathbf{y}(n)$, we have $\mathcal{G}_0\mathbf{y}(n) = \mathcal{G}\mathbf{y}_1(n)$. Since $\mathcal{H}$ is lossless, we have $\| \mathcal{G}_0\mathbf{y}(n) \|^2 = \| \mathcal{G}\mathbf{y}_1(n) \|^2 = \| \mathbf{y}_1(n) \|^2 \leq \| \mathbf{y}(n) \|^2$. Hence $\mathcal{G}_0$ is a passive inverse of $\mathcal{H}$.

(ii) To prove the uniqueness of $\mathcal{G}_0$, assume that there is another passive inverse $\mathcal{G}_1 \not\equiv \mathcal{G}_0$. Since $\mathcal{G}_1 \not\equiv \mathcal{G}_0$, there is a $\mathbf{y}_0(n) \in \mathcal{R}^{\perp}(\mathcal{H})$ such that $\mathbf{x}_0(n) = \mathcal{G}_1\mathbf{y}_0(n) \neq \mathbf{0}$. Let $\mathbf{y}_1(n) \in \mathcal{R}(\mathcal{H})$ be a signal such that $\mathbf{x}_1(n) = \mathcal{G}_1\mathbf{y}_1(n)$ and $\langle \mathbf{x}_0(n), \mathbf{x}_1(n) \rangle \neq 0$ (this is always possible because $\mathbf{x}_1(n)$ can be an arbitrary $l_2(M)$ signal and $\mathbf{x}_0(n)$ is not identically zero). Consider $\mathbf{x}(n) = \mathcal{G}_1\big(\mathbf{y}_0(n) + c\mathbf{y}_1(n)\big)$. We have

$$\| \mathbf{x}(n) \|^2 = \| \mathbf{x}_0(n) \|^2 + |c|^2 \| \mathbf{y}_1(n) \|^2 + c\sum_n \Big(\mathbf{x}_1^{\dagger}(n)\mathbf{x}_0(n) + \mathbf{x}_0^{\dagger}(n)\mathbf{x}_1(n) \Big), \tag{5.5.13}$$

where we have used $\| \mathbf{x}_1(n) \| = \| \mathbf{y}_1(n) \|$ which follows from the fact that $\mathcal{H}$ is lossless. For given $\mathbf{x}_0(n)$ and $\mathbf{x}_1(n)$, one can always find a constant c such that $c\sum_n \big(\mathbf{x}_1^{\dagger}(n)\mathbf{x}_0(n) + \mathbf{x}_0^{\dagger}(n)\mathbf{x}_1(n) \big) > \| \mathbf{y}_0(n) \|^2$, which implies that $\mathcal{G}_1$ is not passive, a contradiction! Therefore $\mathcal{G}_0$ is the only passive inverse.

(iii) We will prove this for the case where $\mathcal{H}$ is in direct form A. The proof for the direct form B is similar. Assume that the NIL system $\mathcal{H}$ is $\mathbf{E}(n, \mathcal{Z}) = \sum_k \mathbf{e}_k(n)\mathcal{Z}^{-k}$. We only need to establish the passiveness of $\widetilde{\mathbf{E}}(n, \mathcal{Z}) = \sum_k \mathcal{Z}^k \mathbf{e}_k^{\dagger}(n)$, i.e., we need to prove $\widetilde{\mathbf{E}}(n, \mathcal{Z})\mathbf{y}(n) = \mathbf{0}$ for all $\mathbf{y}(n) \in \mathcal{R}^{\perp}(\mathbf{E})$. From (5.5.11b), we have

$$\mathcal{R}^{\perp}(\mathbf{E}) = \{ \mathbf{y}(n) \in l_2(M) \mid \langle \mathbf{E}(n, \mathcal{Z})\mathbf{x}(n), \mathbf{y}(n) \rangle = 0, \quad \text{for all } \mathbf{x}(n) \in l_2(M) \}. \tag{5.5.14}$$

That means, for all $\mathbf{y}(n) \in \mathcal{R}^{\perp}(\mathbf{E})$, we have

$$\sum_n \Big(\sum_k \mathbf{e}_k(n)\mathbf{x}(n-k) \Big)^{\dagger} \mathbf{y}(n) = 0, \quad \text{for all } \mathbf{x}(n) \in l_2(M). \tag{5.5.15}$$

After some simplifications, we get

$$\sum_n \mathbf{x}^{\dagger}(n) \Big(\sum_k \mathbf{e}_k^{\dagger}(n+k)\mathbf{y}(n+k) \Big) = 0, \quad \text{for all } \mathbf{x}(n) \in l_2(M),\ \mathbf{y}(n) \in \mathcal{R}^{\perp}(\mathbf{E}). \tag{5.5.16}$$

Since $\mathbf{x}(n)$ is an arbitrary $l_2(M)$ signal, we conclude that $\widetilde{\mathbf{E}}(n, \mathcal{Z})\mathbf{y}(n) = \sum_k \mathbf{e}_k^{\dagger}(n+k)\mathbf{y}(n+k) = \mathbf{0}$. The proof is complete. ■

5.5.3. Lossless Time-varying Filter Banks

In Section 5.4, we have demonstrated the usefulness of the transform domain description in analyzing the TVFBs. In the following, we will study the lossless TVFBs in the transform domain. To relate all the theory developed so far in this section to the TVFBs, we use the polyphase representation.

TVFB with lossless analysis bank: Consider Fig. 5.4.1. Let $\mathbf{E}(n, \mathcal{Z})$ be lossless. If we take the analysis and synthesis polyphase matrices as $\mathbf{E}(n, \mathcal{Z})$ and $\widetilde{\mathbf{E}}(n, \mathcal{Z})$ respectively, then we have a PR TVFB whose

analysis bank is lossless. Using (5.4.4) and (5.4.9), one can show in this case that the analysis and the synthesis filters are related as

$$F^k(\mathcal{Z}, Mn) = \widetilde{H}^k(Mn, \mathcal{Z}), \qquad f_i^k(Mn) = h_{-i}^{*k}(Mn), \quad \forall n. \tag{5.5.17}$$

Therefore the coefficients of the synthesis filters $f_i^k(Mn)$ are the mirror-image conjugates of the coefficients of the analysis filters $h_i^k(Mn)$. As we explained in Section 5.4, only the coefficients at time Mn are relevant to the TVFB.

IIL TVFB: If the synthesis polyphase matrix $\mathbf{R}(n, \mathcal{Z})$ is invertible (i.e., the TVFB is an IIL system), then we have $\widetilde{\mathbf{R}}(\mathcal{Z}, n)\mathbf{R}(\mathcal{Z}, n) = \mathbf{I}$. This implies that the synthesis filters satisfy the following relation:

$$\left[\widetilde{F}^k(n, \mathcal{Z})F^l(n, \mathcal{Z})\right]_{\downarrow M} = \delta(k - l), \tag{5.5.18}$$

where the notation $[\,\bullet\,]_{\downarrow M}$ is given in (5.3.7). The above equation can be viewed as a time-varying generalization of the orthonormality condition of the synthesis filters defined in [Her93a, Djo94, Vai95, Dau92]. Note that if the analysis bank is an IIL system, the synthesis bank which yields PR is clearly also an IIL system. Therefore we can call a FB an IIL TVFB.

Remarks:

1. Because of Theorem 5.5.4, there is always a passive inverse for any lossless analysis system. If the passive inverse is used in the reconstruction, the noise introduced in the subband will not be amplified. Combining this with the fact that the analysis bank is lossless, we conclude that the coding gain ≥ 1.
2. It should be mentioned that if the inverse of a LTV system is lossless, then the LTV system itself is lossless. Therefore there is no non lossless system with lossless inverse, even in the LTV case.

5.6. EXAMPLES OF LOSSLESS LTV SYSTEMS

Example 5.6.1. *A Non Trivial SISO IIL System:* Consider the following LPTV system $\mathcal{H}_1$: $y(n) = h_{-1}(n)x(n+1) + h_0(n)x(n) + h_1(n)x(n-1)$ where the coefficients are:

$$h_{-1}(n) = \begin{cases} 0, & \text{for } n \text{ even;} \\ \frac{1}{\sqrt{2}}, & \text{for } n \text{ odd,} \end{cases} \quad h_0(n) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{for } n \text{ even;} \\ \frac{-1}{\sqrt{2}}, & \text{for } n \text{ odd,} \end{cases} \quad h_1(n) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{for } n \text{ even;} \\ 0, & \text{for } n \text{ odd.} \end{cases} \tag{5.6.1}$$

The output of the above system is:

$$y(n) = \begin{cases} \frac{x(n)+x(n-1)}{\sqrt{2}}, & \text{for } n \text{ even;} \\ \frac{x(n+1)-x(n)}{\sqrt{2}}, & \text{for } n \text{ odd.} \end{cases} \tag{5.6.2}$$

One can verify that the coefficients defined in (5.6.1) satisfy the lossless condition (5.5.4). As a consistency check, we can verify the ouput energy $\sum_n |y(n)|^2 = \sum_n |x(n)|^2$. This example shows that there exist non trivial scalar lossless systems in the LTV case. The inverse of $\mathcal{H}_1$ is given by $\mathcal{G}_1$: $\widehat{x}(n) = h_{-1}^*(n-1)y(n-$

$1) + h_0^*(n)y(n) + h_1^*(n+1)y(n+1)$, where the coefficients $h_i(n)$ are the same as (5.6.1). The inverse of a scalar lossless FIR system is also FIR! This is impossible in the LTI case except for the trivial system of the form $e^{j\theta}z^{-k}$. Furthermore one can verify that $\mathcal{G}_1 \equiv \mathcal{H}_1$. *We have a system which is its own inverse!* Therefore the inverse $\mathcal{G}_1$ is also lossless which implies that $\mathcal{H}_1$ is an IIL system. Since $M = 1$, the analysis and synthesis filters are themselves the polyphase matrices $\mathcal{H}_1, \mathcal{G}_1$ respectively. ■

Example 5.6.2. *A $M \times M$ IIL System:* Consider the following first-order system $\mathcal{H}_2$:

$$\mathbf{y}(n) = \big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\big]\mathbf{x}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n-1)\mathbf{x}(n-1), \tag{5.6.3}$$

where the $M \times 1$ vector $\mathbf{v}(n)$ satisfies $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$ for all n. The system coefficients are $\mathbf{e}_0(n) = [\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)]$ and $\mathbf{e}_1(n) = \mathbf{v}(n)\mathbf{v}^\dagger(n-1)$. These coefficients have the properties that $\mathbf{e}_0^\dagger(n)\mathbf{e}_1(n) = \mathbf{e}_1^\dagger(n)\mathbf{e}_0(n) = \mathbf{0}$ and $\mathbf{e}_0^\dagger(n)\mathbf{e}_0(n) = \mathbf{e}_0(n)$. By using these properties, one can verify that the coefficients satisfy (5.5.4). Hence the system $\mathcal{H}_2$ in (5.6.3) is lossless. Its inverse is given as $\mathcal{G}_2$:

$$\widehat{\mathbf{x}}(n) = \big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\big]\mathbf{y}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n+1)\mathbf{y}(n+1). \tag{5.6.4}$$

One can show by directly substituting the coefficients into (5.5.8) that the inverse $\mathcal{G}_2$ defined in (5.6.4) is also lossless. Therefore $\mathcal{H}_2$ is an IIL system. The analysis filters are given as:

$$H^k(Mn, \mathcal{Z}) = \big(1 - |v_k(n)|^2\big)\mathcal{Z}^{-k} - \sum_{i \neq k} v_k(n)v_i^*(n)\mathcal{Z}^{-i} + \sum_{i=0}^{M-1} v_k(n)v_i^*(n-1)\mathcal{Z}^{-M-i}, \tag{5.6.5}$$

where $v_k(n)$ is the k-th element of $\mathbf{v}(n)$. The synthesis filters are $F^k(\mathcal{Z}, Mn) = \widetilde{H}^k(Mn, \mathcal{Z})$. ■

Example 5.6.3. *A $M \times M$ NIL System:* Consider the following LTV system $\mathcal{H}_3$: The system equation is the same as (5.6.3) but the vectors are chosen as $\mathbf{v}(n) = \mathbf{0}$ for $n < 0$ and $\mathbf{v}(n) =$ arbitrary unit norm vector for $n \geq 0$. The output of the system can be described as

$$\mathbf{y}(n) = \begin{cases} \mathbf{x}(n), & \text{for } n < 0; \\ [\mathbf{I} - \mathbf{v}(0)\mathbf{v}^\dagger(0)]\mathbf{x}(0), & \text{for } n = 0; \\ [\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)]\mathbf{x}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n-1)\mathbf{x}(n-1), & \text{for } n > 0. \end{cases} \tag{5.6.6}$$

One can verify that the system $\mathcal{H}_3$ described above satisfies (5.5.4). Hence it is lossless. From the above equation, we know that the output $\mathbf{y}(0)$ is always in the range of the matrix $\big[\mathbf{I} - \mathbf{v}(0)\mathbf{v}^\dagger(0)\big]$. Since the matrix $\big[\mathbf{I} - \mathbf{v}(0)\mathbf{v}^\dagger(0)\big]$ with unit norm vector $\mathbf{v}(0)$ is always singular, the output $\mathbf{y}(0)$ therefore cannot be arbitrary. So the lossless system $\mathcal{H}_3$ does not map $l_2(M)$ onto $l_2(M)$. Theorem 5.5.3 implies that its inverse is not lossless (hence not invertible). Thus $\mathcal{H}_3$ is a NIL system. Its unique passive inverse is given as $\mathcal{G}_3$:

$$\widehat{\mathbf{x}}(n) = \begin{cases} \mathbf{y}(n), & \text{for } n < 0; \\ [\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)]\mathbf{y}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n+1)\mathbf{y}(n+1), & \text{for } n \geq 0. \end{cases} \tag{5.6.7}$$

The corresponding analysis filters are given by

$$H^k(Mn, \mathcal{Z}) = \begin{cases} \mathcal{Z}^{-k}, & \text{for } n < 0; \\ \big(1 - |v_k(0)|^2\big)\mathcal{Z}^{-k} - \sum_{i \neq k} v_k(0)v_i^*(0)\mathcal{Z}^{-i}, & \text{for } n = 0; \\ \big(1 - |v_k(n)|^2\big)\mathcal{Z}^{-k} - \sum_{i \neq k} v_k(n)v_i^*(n)\mathcal{Z}^{-i} + \sum_{i=0}^{M-1} v_k(n)v_i^*(n-1)\mathcal{Z}^{-M-i}, & \text{for } n > 0. \end{cases} \tag{5.6.8}$$

The unique passive synthesis bank is given as $F^k(\mathcal{Z}, Mn) = \widetilde{H}^k(Mn, \mathcal{Z})$. ■

Comments:

1. One can verify that the lossless system $\mathcal{H}_1$ in Example 5.6.1 can be obtained by unblocking (see Section 5.3.3) a 2×2 LTI system with transfer matrix $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Since $\mathbf{T}^{-1} = \mathbf{T}$, the inverse of $\mathcal{H}_1$, which the unblocked version of $\mathbf{T}^{-1}$, is identical to $\mathcal{H}_1$ itself!
2. The lossless system $\mathcal{H}_2$ in Example 5.6.2 can be viewed as the generalization of the LTI degree-one building block studied in Chapter 14 of [Vai93]. It will be shown in the Chapter 6 that in some cases, $\mathcal{H}_2$ can be used as a building block to factorize higher order lossless systems.

5.7. DISCRETE-TIME FRAMES AND RIESZ BASES FOR l_2

In [Che94], the authors introduced a vector space framework for LTI FBs. It was shown how the vector space manipulation simplifies the analysis of conventional LTI FB theory. In particular it is true that the synthesis functions of a PR FB form a *discrete-time Riesz basis* for l_2. This result was also shown in [Djo94] by using a different approach. A brief review of these results can be found in Section 1.2. In the case of TVFBs, we will show that a PR TVFB in general will only give rise to a *discrete-time frame*. If in addition being PR, the synthesis polyphase matrix is also invertible, in this case, the TVFB will generate a Riesz basis. In the lossless case, a NIL PR TVFB produces a *tight frame* with a unity frame bound while an IIL PR TVFB will generate an *orthonormal basis*. Excellent tutorials on frames and l_2 bases can be found in [You80, Hei89, Dau92, Vai95].

5.7.1. Frames and Non Invertible Inverse LTV Systems

Recall from Section 5.4 that if a TVFB has PR, then any l_2 signal $x(n)$ can be expressed in terms of the filter coefficients as in (5.4.2). If we define the families of the double index functions $\eta_{km}(n)$ and $\theta_{km}(n)$ respectively as

$$\eta_{km}(n) = f^k_{n-Mm}(Mm), \qquad \theta^*_{km}(n) = h^k_{Mm-n}(Mm), \tag{5.7.1}$$

where $0 \le k \le M-1$ and $-\infty \le m \le \infty$, then (5.4.2) can be rewritten as

$$x(n) = \sum_{k=0}^{M-1} \sum_m y_k(m)\eta_{km}(n), \quad \text{where} \quad y_k(m) = \langle x(n), \theta_{km}(n) \rangle = \sum_n x(n)\theta^*_{km}(n). \tag{5.7.2}$$

The functions $\eta_{km}(n)$ and $\theta_{km}(n)$ are respectively called the synthesis and analysis functions in [Che94]. From (5.7.1), the synthesis functions $\eta_{km}(n)$ are the synthesis filter coefficients at time Mm shifted to the right by Mm. In the LTI case, both $h^k_i(n)$ and $f^k_i(n)$ are independent of n. All the analysis and synthesis functions, $\theta_{km}(n)$ and $\eta_{km}(n)$ for a fixed k, are respectively shifted versions of $\theta_{k0}(n)$ and $\eta_{k0}(n)$. That means for all m, $\theta_{km}(n)$ and $\eta_{km}(n)$ have the same shape as $\theta_{k0}(n)$ and $\eta_{k0}(n)$ respectively. In the LTV case, this property no longer holds as the filters are time-varying. If the analysis bank is lossless, recall that one of the possible set of synthesis filters (in general the synthesis filters are not unique, see Section

5.5.2) for PR is given in (5.5.17). If the synthesis filters are chosen so, the synthesis bank is passive. Using (5.5.17) in (5.7.1), we have

$$\theta_{km}(n) = \eta_{km}(n). \tag{5.7.3}$$

The analysis functions are identical to the synthesis functions. In the following, we will provide an example which shows that the synthesis (or the analysis) functions corresponding to a PR TVFB are in general not independent (hence cannot be a discrete-time basis for l_2).

Example 5.7.1. *Linear Dependency of Analysis/Synthesis Functions:* Consider the PR TVFB given in Example 5.6.3. For simplicity we choose $M = 2$ and $\mathbf{v}(n) = [\frac{1}{\sqrt{2}} \ \frac{1}{\sqrt{2}}]^T$ for $n \geq 0$. In this case, $k = 0$ or 1. The analysis bank is a NIL system, hence the synthesis bank is not invertible. The analysis functions are given as

$$\theta_{0m}(n) = \begin{cases} \delta(n-2m), & \text{for } m < 0; \\ 0.5\delta(n) - 0.5\delta(n+1), & \text{for } m = 0; \\ 0.5\delta(n-2m) - 0.5\delta(n+1-2m) + 0.5\delta(n+2-2m) + 0.5\delta(n+3-2m), & \text{for } m > 0, \end{cases} \tag{5.7.4a}$$

$$\theta_{1m}(n) = \begin{cases} \delta(n+1-2m), & \text{for } m < 0; \\ -0.5\delta(n) + 0.5\delta(n+1), & \text{for } m = 0; \\ -0.5\delta(n-2m) + 0.5\delta(n+1-2m) + 0.5\delta(n+2-2m) + 0.5\delta(n+3-2m), & \text{for } m > 0. \end{cases} \tag{5.7.4b}$$

One set of synthesis functions is given as $\eta_{km}(n) = \theta_{km}(n)$ since the analysis bank is lossless. In this case, one can verify that

$$\theta_{00}(n) - \theta_{01}(n) - \theta_{10}(n) + \theta_{11}(n) = 0, \quad \forall n. \tag{5.7.5}$$

Therefore the analysis (or synthesis) functions are linearly dependent. ■

From the above example, we have seen that the analysis (or synthesis) functions of a PR TVFB does not form a basis. However, we can show the following frame property:

Theorem 5.7.1. *Discrete-Time Frame for l_2:* The synthesis functions $\eta_{km}(n)$ of a stable PR FIR TVFB form a compactly supported *discrete-time frame* for l_2. The dual frame is given by the analysis functions $\theta_{km}(n)$. Moreover if the analysis bank is lossless, then the functions $\eta_{km}(n) = \theta_{km}(n)$ form a *tight frame* with unity frame bound. ■

Proof: If the TVFB is FIR, the functions $\eta_{km}(n)$ and $\theta_{km}(n)$ clearly have compact support. It is shown in Appendix that for any $x(n) \in l_2$, there exist $B_1 < \infty$ and $B_2 < \infty$ such that

$$\sum_{k,m} |\langle x(n), \eta_{km}(n)\rangle|^2 \leq B_1 \parallel x(n) \parallel^2, \tag{5.7.6a}$$

$$\sum_{k,m} |\langle x(n), \theta_{km}(n)\rangle|^2 \leq B_2 \parallel x(n) \parallel^2 . \tag{5.7.6b}$$

Notice that the index k runs from 0 to $M-1$ only. From (5.7.2), we have

$$\begin{aligned} \parallel x(n) \parallel^2 &\leq \sum_{k,m} |\langle x(n), \theta_{km}(n)\rangle| \ |\langle \eta_{km}(n), x(n)\rangle| \\ &\leq \left(\sum_{k,m} |\langle x(n), \theta_{km}(n)\rangle|^2 \right)^{1/2} \left(\sum_{k,m} |\langle x(n), \eta_{km}(n)\rangle|^2 \right)^{1/2}, \end{aligned} \tag{5.7.7}$$

where we have used the Cauchy-Schwartz inequality. Substituting (5.7.6b) into (5.7.7), we have shown that there exist $A_1 = 1/B_2 > 0$ and $B_1 < \infty$ such that

$$A_1 \parallel x(n) \parallel^2 \leq \sum_{k,m} |\langle x(n), \eta_{km}(n)\rangle|^2 \leq B_1 \parallel x(n) \parallel^2, \tag{5.7.8}$$

for all $x(n) \in l_2$. Therefore the synthesis functions $\eta_{km}(n)$ form a frame [You80, Hei89, Dau92, Vai95]. Its dual frame is given by $\theta_{km}(n)$. In the case when the analysis bank is lossless, we have

$$\parallel x(n) \parallel^2 = \sum_{k=0}^{M-1} \sum_{m} |y_k(m)|^2 = \sum_{k,m} |\langle x(n), \theta_{km}(n)\rangle|^2, \tag{5.7.9}$$

which shows that the functions $\theta_{km}(n)$ form a tight frame with unity frame bound. ■

Remarks:

1. Given a frame $\eta_{km}(n)$ for l_2, where $-\infty < m < \infty$ and $0 \leq k \leq M-1$, there is always a dual frame $\theta_{km}(n)$. One can always construct an underlying M-channel PR TVFB by choosing the analysis and synthesis filters as in (5.7.1). Then we can get a stable PR TVFB. In addition, if both the frame and the dual frame have compact support, it is clear that the filters have only finitely many nonzero coefficients.
2. Eq. (5.7.9) can be viewed as the Paserval relation which states the energy preservation.
3. The reason why $\eta_{km}(n)$ might fail to be a Riesz basis is because the functions may no longer be independent as demonstrated in Example 5.7.1. The proof of independence in [Che94] does not go through as $\eta_{km}(n)$ are no longer shifted versions of $\eta_{k0}(n)$.

5.7.2. Bases and Invertible Inverse LTV Systems

Recall that if the synthesis bank of a PR TVFB is also invertible, the filters satisfy the condition in (5.4.6). By using the definition of the analysis and synthesis functions in (5.7.1), we can show by directly expanding (5.4.6) that

$$\langle \eta_{k_0 m_0}(n), \theta_{k_1 m_1}(n)\rangle = \delta(k_0 - k_1)\delta(m_0 - m_1). \tag{5.7.10}$$

Therefore the analysis and the synthesis functions satisfy the biorthogonality condition. In the case of IIL TVFBs, we have $\eta_{km}(n) = \theta_{km}(n)$. Eq. (5.7.10) reduces to the following orthonormality condition:

$$\langle \eta_{k_0 m_0}(n), \eta_{k_1 m_1}(n)\rangle = \delta(k_0 - k_1)\delta(m_0 - m_1). \tag{5.7.11}$$

In this case, we can show the following:

Theorem 5.7.2. *Discrete-Time Basis for l_2:* Given a stable PR FIR TVFB with an invertible synthesis bank, the corresponding synthesis functions $\eta_{km}(n)$ form a compactly supported *discrete-time Riesz basis* for l_2. The dual basis is given by the analysis functions $\theta_{km}(n)$. Moreover if the TVFB is an IIL system, then the functions $\theta_{km}(n) = \eta_{km}(n)$ form an *orthonormal basis* for l_2. ■

Proof: Because of Theorem 5.7.1, we need only to establish the independence of the synthesis functions $\eta_{km}(n)$. The independence of $\eta_{km}(n)$ follows directly from the biorthogonality condition (5.7.10). In the case of IIL TVFB, the synthesis functions $\eta_{km}(n)$ satisfy $\| \eta_{km}(n) \|^2 = 1$ (from (5.7.11)). By using the fact that a normalized tight frame is an orthonormal basis [Dau92, Vai95], the proof is complete. ■

Remark: A PR TVFB in general does not form a biorthogonal system. Therefore a PR TVFB *cannot* be called a biorthogonal TVFB unless its synthesis bank is invertible. The same is true for a lossless TVFB.

5.8. CONCLUSIONS

In this chapter, we have introduced a time-varying polyphase approach to study some basic properties of TVFBs. By using the proposed method, the theory of TVFBs can be developed in analogy to the conventional LTI FBs (Sections 5.3 and 5.4). Even though TVFBs share many properties with the LTI FBs, there are some major differences. We have studied in detail the class of the lossless TVFBs (Section 5.5). Both the time-domain and frequency-domain characterizations of lossless LTV systems are given ((5.5.4) and (5.5.10) respectively). We showed that all lossless LTV systems are invertible and its inverse is FIR provided that the original lossless system is FIR (Theorem 5.5.2). We also showed that the inverse of a lossless LTV system may not be lossless (Examples 5.1.1 and 5.6.3). The losslessness, invertibility, uniqueness and passitivity of the inverse are related as in Theorems 5.5.3 and 5.5.4. We have demonstrated that the synthesis (or analysis) functions of a PR TVFB may not generate a basis for l_2 (Example 5.7.1). These functions however form a frame for l_2 (Theorem 5.7.1) and become a basis for l_2 only if the synthesis polyphase matrix is invertible (Theorem 5.7.2). Moreover, the basis is *orthonormal* if the TVFB is an IIL system. In Chapter 6, we will show that the time-domain characterization of lossless systems in (5.5.4) is useful for the parameterization and factorization of lossless LTV systems.

5.9. APPENDIX

In the following, we will only provide a proof for the inequality in (5.7.6a). The proof for (5.7.6b) is very similar. Since the filters are stable, we have $\sum_i |f_i^k(n)| < \infty$ and $\sum_i |h_i^k(n)| < \infty$, for all k and n. This in particular implies that there is a $B < \infty$ such that $\max_n |f_i^k(n)| \leq B$ for all i and k. Since the filters are FIR, there is a $N < \infty$ such that $f_i^k(n) = 0$ and $h_i^k(n) = 0$ for all $i > N$. By using (5.7.1) and making a change of variable, we can rewrite the left hand side of (5.7.6a) as:

$$\sum_{k=0}^{M-1} \sum_{i=0}^{N-1} \sum_{m} |f_i^k(Mm)x(Mm+i)|^2 \leq \| x(n) \|^2 \sum_{k=0}^{M-1} \sum_{i=0}^{N-1} \left(\max_m |f_i^k(Mm)| \right) \leq N \cdot B \| x(n) \|^2, \quad (5.A.1)$$

By taking $B_1 = NB$, we have proved (5.7.6a).

6

Factorizability of Lossless Time-Varying Filters and Filter Banks

6.1. INTRODUCTION

Fig. 6.1.1 shows a M-channel maximally decimated time-varying filter banks (TVFB). In Chapter 5, we have studied some basic properties of TVFB and showed that there are several unusual properties which are not exhibited by the conventional FBs.

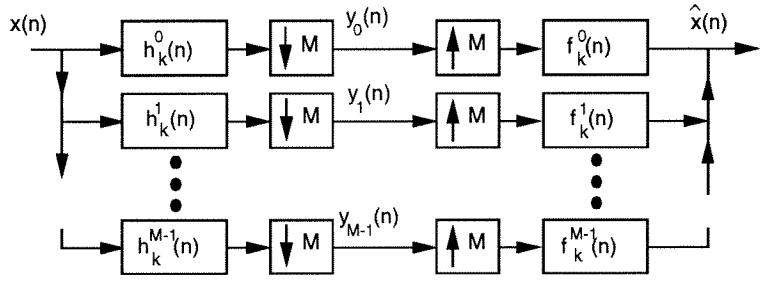

Fig. 6.1.1. M-channel maximally decimated time-varying filter bank.

Consider Fig. 6.1.1, where $h_k^i(n)$ and $f_k^i(n)$ represent the k-th coefficients of the i-th analysis and synthesis filters at time n respectively. Using the LTV polyphase representation introduced in Chapter 5, the M-channel TVFB in Fig. 6.1.1 can be redrawn as in Fig. 6.1.2. We can capture all M-channel TVFBs by characterizing the following M-input M-output LTV filters:

$$\mathbf{y}(n) = \sum_k \mathbf{e}_k(n)\mathbf{x}(n-k), \tag{6.1.1a}$$

$$\hat{\mathbf{x}}(n) = \sum_k \mathbf{r}_k(n-k)\mathbf{y}(n-k). \tag{6.1.1b}$$

In particular, if the $M \times M$ system in (6.1.1b) is the inverse system of (6.1.1a), then we have a PR system.

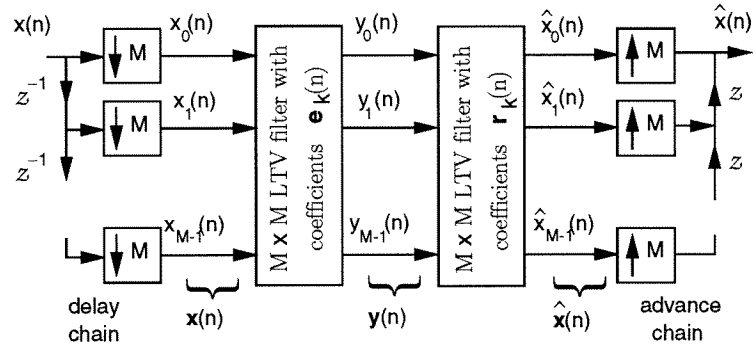

Fig. 6.1.2. Polyphase representation of time-varying filter bank.

In the LTI case, it is well-known [Vai88, Dog88] that all PU FBs can be factorized into degree-one building blocks (see Section 1.2). The factorization is minimal in terms of delay elements. In this chapter, we will study a similar factorizaton for the LTV case. The analysis bank is said to be *lossless* if

$$\sum_n |x(n)|^2 = \sum_{i=0}^{M-1} \sum_n |y_i(n)|^2, \tag{6.1.2}$$

where $y_i(n)$ are the decimated subband signal as shown in Fig. 6.1.1. The class of TVFBs with lossless analysis bank is addressed in detail in Chapter 5. In this chapter, we are going to use the results in Chapter 5 to study the factorizability of this class of TVFBs in terms of lossless LTV building blocks.

Related Works in Literatures: In [Arr93], the authors generalized the two-channel lattice structure derived in [Vai87a] to the LTV case. By cascading these LTV lattice sections, the authors obtained a two-channel PR TVFB in factorized form. The result is generalized to M-channel case in [deQ93] by using the TV planar rotations [Dog88]. In [Gop95], the problem of switching between two LTI PU lattice structures [Vai93] was studied. The above references gave methods for construction of lossless PR TVFBs using TV lattice structures. In this chapter, we are going to address a number of issues, such as the factorizability of general lossless LTV systems, completeness and minimality of the cascade of TV lattice structures, and so forth.

6.1.1. Chapter Outline

In Section 6.2, we will show how to capture all degree-one lossless LTV systems by two time-dependent memoryless unitary matrices. All degree-one lossless systems can be realized as a cascade of a TV memoryless unitary matrix followed by a lossless dyadic-based LTV structure. A number of useful properties (e.g., preservation of losslessness under delay transformation, simple inversion rules, commutivity in cascade, etc.) will be discussed. In Section 6.3, the lapped orthogonal transform (LOT) [Mal92] is extended to the LTV case. We will show that all IIL TVLOTs can be factorized *uniquely* as a cascade of the lossless degree-one building blocks followed by a unitary matrix. The factorization is *minimal* in terms of delay elements. For the NIL TVLOT, we will show factorizable as well as unfactorizable examples. In Section

6.4, we will show how to construct higher degree NIL and IIL systems by using the dyadic-based building blocks. We will give several necessary conditions for the factorizability of a general lossless LTV system and prove that there are unfactorizable IIL systems. A sufficient condition for factorizability, which leads to a order reduction procedure, will also be derived. State-space representation of LTV systems will be discussed in Section 6.5. We introduce the concept of *strong eternal reachability* (SER) and *strong eternal observability* (SEO), which are used to prove that the cascade implementation of factorizable IIL systems is *minimal* in terms of delay elements as well as the number of building blocks. In Section 6.6, we will show that the LTV normalized IIR lattice structure introduced in [Gra75] is bounded input bounded output (BIBO) stable if the TV lattice coefficients $|\alpha_k(n)| \leq \gamma < 1$. We will extend the lossless dyadic-based LTV systems to the non lossless case in Section 6.7. In the LTI case, these lossless systems reduce to the useful degree-one biorthogonal LTI building blocks introduced in [Vai95c, Vai95d].

6.2. THE MOST GENERAL DEGREE-ONE LOSSLESS LTV SYSTEM

Consider the following $M \times M$ first-order system:

$$\mathbf{y}(n) = \mathbf{e}_0(n)\mathbf{x}(n) + \mathbf{e}_1(n)\mathbf{x}(n-1), \tag{6.2.1}$$

where $\mathbf{e}_k(n)$ are $M \times M$ matrices. Then we can prove the following:

Theorem 6.2.1. *Complete Characterization of Degree-One Lossless System:* The first order LTV system defined in (6.2.1) is a degree-one lossless LTV system if and only if for all n, there exist unitary matrices $\mathbf{U}(n) = [\mathbf{u}_0(n)\ \mathbf{u}_1(n)\ \ldots\ \mathbf{u}_{M-1}(n)]$ and $\mathbf{V}(n) = [\mathbf{v}_0(n)\ \mathbf{v}_1(n)\ \ldots\ \mathbf{v}_{M-1}(n)]$ such that

$$\mathbf{e}_0(n) = \mathbf{U}(n)\begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{M-1} \end{bmatrix}\mathbf{V}^\dagger(n) \quad \text{and} \quad \mathbf{e}_1(n) = \mathbf{u}_0(n)\mathbf{v}_0^\dagger(n-1). \tag{6.2.2}$$

■

Proof: The first-order system in (6.2.1) has degree=1 if and only if the rank of $\mathbf{e}_1(n)$ is one for all n. So we can express $\mathbf{e}_1(n) = \mathbf{w}_0(n)\mathbf{v}_0^\dagger(n-1)$, where $\mathbf{w}_0(n)$ and $\mathbf{v}_0(n-1)$ are nonzero vectors. Without lost of generality, we can assume $\mathbf{w}_0^\dagger(n)\mathbf{w}_0(n) = 1$, for all n. Applying the necessary and sufficient condition for losslessness in (5.5.4) to (6.2.1), we obtain

$$\mathbf{e}_0^\dagger(n)\mathbf{e}_0(n) + \mathbf{e}_1^\dagger(n+1)\mathbf{e}_1(n+1) = \mathbf{I}, \tag{6.2.3a}$$

$$\mathbf{e}_0^\dagger(n)\mathbf{e}_1(n) = \mathbf{0}. \tag{6.2.3b}$$

Substituting $\mathbf{e}_1(n) = \mathbf{w}_0(n)\mathbf{v}_0^\dagger(n-1)$ into the above equation, we get

$$\mathbf{e}_0^\dagger(n)\mathbf{e}_0(n) = \mathbf{I} - \mathbf{v}_0(n)\mathbf{v}_0^\dagger(n), \tag{6.2.4a}$$

$$\mathbf{e}_0^\dagger(n)\mathbf{w}_0(n) = \mathbf{0}, \tag{6.2.4b}$$

where the fact that $\mathbf{v}_0(n-1) \neq \mathbf{0}$ has been applied to obtain (6.2.4b). Let $\mathbf{v}_1(n)$, ..., $\mathbf{v}_{M-1}(n)$ be unit norm vectors perpendicular to $\mathbf{v}_0(n)$, i.e., $\mathbf{v}_i^\dagger(n)\mathbf{v}_0(n) = 0$ for $i \neq 0$. We have $[\mathbf{I} - \mathbf{v}_0(n)\mathbf{v}_0^\dagger(n)]\mathbf{v}_i(n) = \lambda_i \mathbf{v}_i(n)$, where the eigenvalues $\lambda_0 = 1 - \mathbf{v}_0^\dagger(n)\mathbf{v}_0(n)$ and $\lambda_i = 1$ for $i \neq 0$. Therefore the matrix $[\mathbf{I} - \mathbf{v}_0(n)\mathbf{v}_0^\dagger(n)]$ is nonsingular unless $\mathbf{v}_0^\dagger(n)\mathbf{v}_0(n) = 1$. But we know from (6.2.4b) that $\mathbf{e}_0^\dagger(n)\mathbf{e}_0(n)$ is singular. Thus it is necessary that $\mathbf{v}_0^\dagger(n)\mathbf{v}_0(n) = 1$ and $\lambda_0 = 0$. Applying the singular value decomposition to the matrix $[\mathbf{I} - \mathbf{v}_0(n)\mathbf{v}_0^\dagger(n)]$, we conclude that there is a unitary matrix $\mathbf{U}'(n)$ such that

$$\mathbf{e}_0(n) = \mathbf{U}'(n) \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I}_{M-1} \end{bmatrix} \mathbf{V}^\dagger(n). \tag{6.2.5}$$

Substituting (6.2.5) into (6.2.4b), we have

$$\mathbf{U}'^\dagger(n)\mathbf{w}_0(n) = [\,\mathbf{u}_0'(n) \quad \mathbf{u}_1'(n) \quad \ldots \quad \mathbf{u}_{M-1}'(n)\,]^\dagger \mathbf{w}_0(n) = \begin{bmatrix} \times \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \tag{6.2.6}$$

where "×" indicates a don't-care term. Eq. (6.2.6) implies $\mathbf{w}_0(n)$ are orthogonal to $\mathbf{u}_1'(n)$, ..., $\mathbf{u}_{M-1}'(n)$. Therefore $\mathbf{w}_0(n) = e^{j\theta(n)}\mathbf{u}_0'(n)$ for some real $\theta(n)$. By letting $\mathbf{u}_0(n) = e^{j\theta(n)}\mathbf{u}_0'(n)$ and $\mathbf{u}_i(n) = \mathbf{u}_i'(n)$ for $i \neq 0$, we have proved the theorem. ■

Implementation Using Planar Rotations and Degree of Freedom: Theorem 6.2.1 tells us how to characterize all the degree-one lossless LTV systems. We can implement (6.2.2) by using the structure shown in Fig. 6.2.1. All lossless degree-one LTV systems can be parameterized by the two time-dependent unitary matrices, $\mathbf{U}(n)$ and $\mathbf{V}(n)$. For the real coefficient case, the unitary matrices $\mathbf{V}(n)$ and $\mathbf{U}(n)$ are real and can be implemented by using planar rotations [Dog88, Vai93]. If the redundant planar rotations of $\mathbf{U}(n)$ are moved into $\mathbf{V}^\dagger(n)$, we can obtain the implementation shown in Fig. 6.2.2. Counting the number of rotations, we know that a degree-one lossless LTV system has only $(M-1)(\frac{M}{2}+1)$ degrees of freedom, instead of $2M^2$, the number of elements in the coefficients $\mathbf{e}_0(n)$ and $\mathbf{e}_1(n)$. The implementation based on planar rotations is *minimal* in terms of free parameters, and it remains lossless even when we change the angles in the rotations.

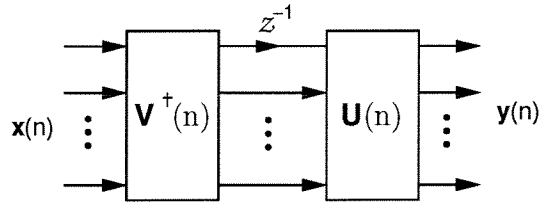

Fig. 6.2.1. Most general degree-one lossless LTV system. Here $\mathbf{U}(n)$ and $\mathbf{V}(n)$ are unitary matrices.

Remark: In the LTI case, it was shown in [Vai93] that the general 2 × 2 LTI PU matrices can be implemented by using the normalized and denormalized FIR lattice structure shown respectively in Figs.

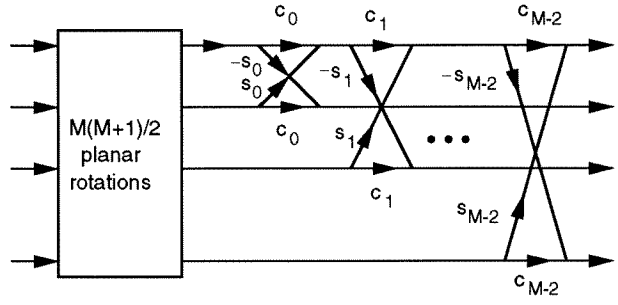

Fig. 6.2.2. Implementation of degree-one lossless real coefficient lossless LTV system based on planar rotation. $c_m = \cos(\theta_m(n))$ and $s_m = \sin(\theta_m(n))$.

6.4-1 and 6.4-2 of [Vai93]. If we make the free parameters (θ_m for the normalized lattice and α_m for the denormalized lattice) time-varying, then one can show that the LTV normalized lattice structure in Fig. 6.4-1 will remain lossless while the denormalized lattice structure in Fig. 6.4-2 will no longer be lossless unless $|\alpha_m(n)|$ are independent of n. In both cases, PR can be achieved by inverting the lattice section by section, as shown in [Arr93].

6.2.1. Dyadic-Based Building Blocks

The implementation based on planar rotations gives a minimal parameterization of degree-one lossless LTV system. However the implementation is not efficient in the sense that it requires more multipliers than necessary. In order to obtain a more efficient implementation, we simplify the coefficients $\mathbf{e}_0(n)$ and $\mathbf{e}_1(n)$ as follows:

$$\mathbf{e}_0(n) = \mathbf{P}(n)\Big[\mathbf{I} - \mathbf{v}_0(n)\mathbf{v}_0^\dagger(n)\Big], \quad \mathbf{e}_1(n) = \mathbf{P}(n)\Big[\mathbf{v}_0(n)\mathbf{v}_0^\dagger(n-1)\Big], \tag{6.2.7}$$

where $\mathbf{P}(n) = \mathbf{U}(n)\mathbf{V}^\dagger(n)$. Since $\mathbf{U}(n)$ can be arbitrary unitary matrix, $\mathbf{P}(n)$ is an arbitrary unitary matrix unrelated to $\mathbf{V}(n)$. Using (6.2.7), we obtain the implementation as in Fig. 6.2.3, where the system $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ is shown in Fig. 6.2.4. We will call the structure in Fig. 6.2.4 a *dyadic-based* structure. Therefore all degree-one lossless LTV systems can be realized as a cascade of a dyadic-based LTV system $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ with $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$, followed by a time-dependent unitary matrix. The dyadic-based structure has only $2M$ multipliers which are fewer than $4M - 2$, the number of multiplications required for the implementation based on planar rotations.

Remarks:

1. Notice that we can also express the coefficients as

$$\mathbf{e}_0(n) = \Big[\mathbf{I} - \mathbf{u}_0(n)\mathbf{u}_0^\dagger(n)\Big]\mathbf{U}(n)\mathbf{V}^\dagger(n), \quad \mathbf{e}_1(n) = \Big[\mathbf{u}_0(n)\mathbf{u}_0^\dagger(n-1)\Big]\mathbf{U}(n-1)\mathbf{V}^\dagger(n-1). \tag{6.2.8}$$

   By using the above equation, we have another implementation the degree-one lossless system as a cascade of a TV unitary matrix followed by the dyadic-based lossless system $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{u}_0(n))$.

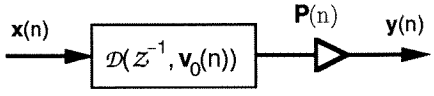

Fig. 6.2.3. Implementation of degree-one lossless LTV system using dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_0(n))$, where $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$ and $\mathbf{P}(n)$ is unitary. $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ is shown in Fig. 6.2.4.

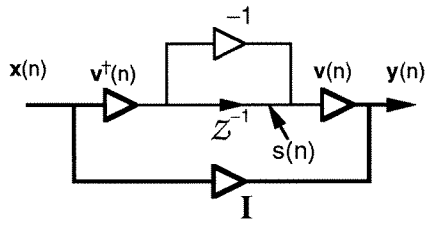

Fig. 6.2.4. Dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$.

2. In the LTI case, the lossless dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ in Fig. 6.2.4 reduces to the degree-one building block given in Fig. 14.5-1 of [Vai93].

Properties of Dyadic-Based Structures $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$**:** The dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ in Fig. 6.2.4 has several nice properties and it can be used as a basic building block to factorize some higher degree lossless LTV systems. The system equation for $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ can be expressed as

$$\mathbf{y}(n) = \Big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\Big]\mathbf{x}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n-1)\mathbf{x}(n-1). \tag{6.2.9}$$

In the following, we list some of its properties:

1. *Identity system:* If $\mathbf{v}(n) = \mathbf{0}$, the dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{0})$ reduces to the identity system.
2. *Losslessness:* In general, it is not easy to satisfy the condition for losslessness in (5.5.4). However for the dyadic based structure in Fig. 6.2.4, if $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$ for all n, then one can show that $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ is lossless. In the presence of quantization, if the vector $\mathbf{v}(n)$ is quantized in such a way that the quantized vector $\mathbf{v}_q(n)$ satisfies $\mathbf{v}_q^\dagger(n)\mathbf{v}_q(n) = 1$, then $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_q(n))$ remains lossless. This implies the implementation in Fig. 6.2.4 is structurally lossless.
3. *Scalar dyadic-based systems:* In the single input single output (scalar) case, the degree-one lossless system degenerates to a delay followed by a unit magnitude multiplier, i.e., $y(n) = c(n)x(n-1)$ for some $|c(n)| = 1$.
4. *Simple inverse system:* It is shown in Chapter 5 that the coefficients of the inverse of a lossless system can be obtained as the mirror image and transpose conjugate of the original system. For a lossless dyadic-based system, the inverse is even simpler. It can be verified that if $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$, the inverse of the degree-one lossless building block $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ in Fig. 6.2.4 can be obtained by simply replacing the delay with an advance operator as shown in Fig. 6.2.5. The inverse system $\mathcal{D}(\mathcal{Z}, \mathbf{v}(n))$ can be expressed as

$$\hat{\mathbf{x}}(n) = \Big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\Big]\mathbf{y}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n+1)\mathbf{y}(n+1). \tag{6.2.10}$$

$\mathcal{D}(\mathcal{Z}, \mathbf{v}(n))$ is anticausal, FIR, and lossless (can be verified by directly substituting the coefficients into (5.5.4)). Therefore $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n)$ is an IIL system if $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$.

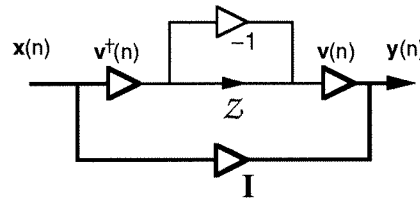

Fig. 6.2.5. Inverse system for the lossless dyadic-based system $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ with $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$.

5. *Commutivity:* Consider $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_0(n))$ and $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_1(n))$, where $\mathbf{v}_0(n)$ and $\mathbf{v}_1(n)$ are unit norm vectors. Then it can be shown that the two building blocks commute with each other if and only if $\mathbf{v}_0(n) = \mathbf{v}_1(n)$ or $\mathbf{v}_0^\dagger(n)\mathbf{v}_1(n) = 0$ (i.e., perpendicular), for all n. If $\mathbf{v}_0^\dagger(n)\mathbf{v}_1(n) = 0$, the building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_0(n))$ and $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_1(n))$ are said to be *perpendicular*. We will see in the next section that precisely this situation arises in the factorization of the TVLOT. The cascade of k perpendicular building blocks can be expressed as $\prod_{i=0}^{k-1} \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$. The ordering of these sections does not matter. The cascade system $\prod_{i=0}^{k-1} \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$ has order one and can be expressed as

$$\mathbf{y}(n) = \Big[\mathbf{I} - \mathbf{V}_k(n)\mathbf{V}_k^\dagger(n)\Big]\mathbf{x}(n) + \mathbf{V}_k(n)\mathbf{V}_k^\dagger(n-1)\mathbf{x}(n-1), \tag{6.2.11}$$

where the $M \times k$ matrix $\mathbf{V}_k(n) = [\mathbf{v}_0(n) \ \ldots \ \mathbf{v}_{k-1}(n)]$. For $M \times M$ systems, if we cascade M such perpendicular lossless building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$, the resulting system reduces to $\mathbf{y}(n) = \mathbf{P}(n)\mathbf{x}(n-1)$ for some unitary $\mathbf{P}(n)$.

6. *Delay transformation:* It is shown in Chapter 5 that if the delay $\mathcal{Z}^{-1}$ in an implementation of a lossless system is replaced by $\mathcal{Z}^{-L}$, the losslessness will usually be destroyed. However the lossless dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ preserves the lossless property under such delay transformation. That is, if the delay in Fig. 6.2.4 is replaced with $\mathcal{Z}^{-L}$ for arbitrary integer L (possibly negative), the new system $\mathcal{D}(\mathcal{Z}^{-L}, \mathbf{v}(n))$ remains lossless. In this case, the system equation is

$$\mathbf{y}(n) = \Big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\Big]\mathbf{x}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n-L)\mathbf{x}(n-L). \tag{6.2.12}$$

Moreover we can show that $\mathcal{D}(\mathcal{Z}^{-L}, \mathbf{v}(n)) = \prod_{L \text{ times}} \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$.

Example 6.2.1. *Lossless and Non Lossless System Obtained from* $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$*:*

(i) If the vector $\mathbf{v}(n)$ in $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ is switched from a zero vector to a unit norm vector at $n = 0$, we know from Example 5.6.3 that the system is lossless. In this case, the system equation is given in (5.6.6), which we reproduce in the following for convenience:

$$\mathbf{y}(n) = \begin{cases} \mathbf{x}(n), & \text{for } n < 0; \\ [\mathbf{I} - \mathbf{v}(0)\mathbf{v}^\dagger(0)]\mathbf{x}(0), & \text{for } n = 0; \\ [\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)]\mathbf{x}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n-1)\mathbf{x}(n-1), & \text{for } n > 0. \end{cases} \tag{6.2.13}$$

The above system is lossless because the coefficients satisfy (5.5.4). And its inverse is not lossless (see Example 5.6.3). Hence it is a NIL system. The fact that the inverse is not invertible also follows from Theorem 6.3.1 which will be proved in the next section.

(ii) Consider another example: If we switch the vector $\mathbf{v}(n)$ in $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ from a unit norm vector to a zero vector at $n = 0$, the resulting system is

$$\mathbf{y}(n) = \begin{cases} \Big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\Big]\mathbf{x}(n) + \mathbf{v}(n)\mathbf{v}^\dagger(n-1)\mathbf{x}(n-1), & \text{for } n < 0; \\ \mathbf{x}(n), & \text{for } n \geq 0. \end{cases} \tag{6.2.14}$$

Notice that $\mathbf{x}(-1)$ appears only in the expression of $\mathbf{y}(-1)$ and it is premultiplied by the singular matrix $[\mathbf{I} - \mathbf{v}(-1)\mathbf{v}^\dagger(-1)]$. Therefore the system in (6.2.14) is not invertible because $\mathbf{x}(-1)$ can never be recovered from $\mathbf{y}(n)$. Hence it cannot be lossless (from Theorem 5.5.2 which says that all lossless systems are invertible).

6.2.2. IIR Lossless LTV Systems Obtained from Dyadic-Based Structures

More generally, if the delay $\mathcal{Z}^{-1}$ in Fig. 6.2.4 is replaced with a BIBO stable scalar lossless LTV system $\mathcal{L}(\mathcal{Z}^{-1}, n)$ (possibly IIR) as shown in Fig. 6.2.6, does the system $\mathcal{D}(\mathcal{L}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$ remain lossless? The answer is in the affirmative. The BIBO stability of $\mathcal{D}(\mathcal{L}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$ can be shown as follows: Since $\mathbf{v}(n)$ has unit norm, the scalar quantity $w_{in}(n)$ is bounded for bounded input $\mathbf{x}(n)$. So the scalar $w_{out}(n)$ is also bounded as $\mathcal{L}(\mathcal{Z}^{-1}, n)$ is a BIBO stable scalar system. Therefore the output vector $\mathbf{y}(n)$ is bounded. To show the losslessness, we write the output as:

$$\mathbf{y}(n) = \Big[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)\Big]\mathbf{x}(n) + w_{out}(n)\mathbf{v}(n), \tag{6.2.15}$$

where $\sum_n |w_{out}(n)|^2 = \sum_n |w_{in}(n)|^2 = \sum_n |\mathbf{v}^\dagger(n)\mathbf{x}(n)|^2$ because the scalar system $\mathcal{L}(\mathcal{Z}^{-1}, n)$ is lossless. Computing the energy of $\mathbf{y}(n)$ from (6.2.15), we get

$$\sum_n \mathbf{y}^\dagger(n)\mathbf{y}(n) = \sum_n \Big(\mathbf{x}^\dagger(n)\mathbf{x}(n) - |\mathbf{v}^\dagger(n)\mathbf{x}(n)|^2 + |w_{out}(n)|^2\Big). \tag{6.2.16}$$

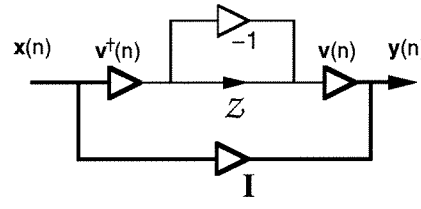

Fig. 6.2.6. Lossless system obtained from the dyadic-based structure. Here $\mathcal{L}(\mathcal{Z}^{-1}, n)$ is an arbitrary lossless scalar system.

Using the fact that $\sum_n |w_{out}(n)|^2 = \sum_n |\mathbf{v}^\dagger(n)\mathbf{x}(n)|^2$, one can show that $\mathcal{D}(\mathcal{L}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$ is lossless. Furthermore it can be verified by direct substitution that the inverse of $\mathcal{D}(\mathcal{L}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$ is given by $\mathcal{D}(\mathcal{L}^{-1}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$, where $\mathcal{L}^{-1}(\mathcal{Z}^{-1}, n)$ is the inverse of $\mathcal{L}(\mathcal{Z}^{-1}, n)$. The existence of inverse $\mathcal{L}^{-1}(\mathcal{Z}^{-1}, n)$ is guaranteed by the losslessness of $\mathcal{L}(\mathcal{Z}^{-1}, n)$.

In particular, the scalar allpass function $A(z)$ is a scalar lossless system. If $\mathcal{L}(\mathcal{Z}^{-1}, n)$ is taken as the stable N-th order allpass function $A_N(z)$, then we can get a subclass of MIMO IIR lossless LTV systems $\mathcal{D}(A_N(z), \mathbf{v}(n))$ with degree N. However in this case, the inverse is anticausal IIR. For the details of implementing IIR anticausal systems when the input is infinite, see [Vai95c].

Remark: More generally, we can obtain a class of invertible non lossless LTV system by replacing $\mathcal{Z}^{-1}$ in Fig. 6.2.4 with an invertible scalar system $\mathcal{T}(\mathcal{Z}^{-1}, n)$ (not necessary lossless). In this case, the BIBO stability of $\mathcal{D}(\mathcal{T}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$ is guaranteed by that of $\mathcal{T}(\mathcal{Z}^{-1}, n)$. Moreover the inverse of $\mathcal{D}(\mathcal{T}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$ can still be obtained simply as $\mathcal{D}(\mathcal{T}^{-1}(\mathcal{Z}^{-1}, n), \mathbf{v}(n))$.

6.3. TIME-VARYING LAPPED ORTHOGONAL TRANSFORMS (TVLOT)

The lapped orthogonal transforms (LOT) have been shown to be very useful in subband coding of image and video signals [Mal92]. They provide satisfactory coding gain and good perceptual quality in these applications, with low complexity. In this section, we will generalize the theory of the conventional LOT system to the time-varying case. Consider the following $M \times M$ first-order system:

$$\mathbf{y}(n) = \mathbf{e}_0(n)\mathbf{x}(n) + \mathbf{e}_1(n)\mathbf{x}(n-1). \tag{6.3.1}$$

If the above system is lossless, then it is called a *time-varying lapped orthogonal transform* (TVLOT). From Theorem 5.5.2, we know that a lossless system is always invertible. Hence a TVLOT is always invertible and its inverse has also order one. However we also know from Theorem 5.5.3 that the inverse system may not be lossless. That means the inverse of a TVLOT may not be a TVLOT system! This is a very different situation from the LTI LOT case. If the inverse of a TVLOT is also lossless, then it is called an *invertible inverse lossless* (IIL) TVLOT. In this case its inverse is also a TVLOT; If the inverse is not invertible, then it is called a *non invertible inverse lossless* (NIL) TVLOT. Note that a dyadic-based structure with unit norm vector in Fig. 6.2.4 is an IIL TVLOT. The existence of NIL TVLOT is shown by Example 6.2.1.

In (6.3.1), since $\mathbf{e}_1(n)$ is time-varying, its rank could also be time-varying. Therefore the degree of a TVLOT is not a constant. We will call the rank of $\mathbf{e}_1(n)$ the *instantaneous degree*, since this is the minimum number of delays required at time n. In the following, we will first show for a TVLOT (either IIL or NIL), the rank of $\mathbf{e}_1(n)$ cannot decrease with n. Moreover, the rank of $\mathbf{e}_1(n)$ is time-invariant if and only if it is an IIL TVLOT. In the second part of this section, we will show that an IIL TVLOT system can *always* be realized as a *unique* cascade of *perpendicular* degree-one building blocks $\mathcal{D}(z^{-1}, \mathbf{v}(n))$ introduced in the Section 6.2. We will also provide an example to show there exist unfactorizable NIL TVLOTs.

6.3.1. The Instantaneous Degree of TVLOT

Theorem 6.3.1. *On the Instantaneous Degree pf TVLOT:* Let $\rho(n)$ be the instantaneous degree of a $M \times M$ TVLOT. Then $\rho(n)$ cannot be decreasing. ■

Proof: If the system is lossless in (6.3.1), then from the lossless condition in (5.5.4) we have:

$$\mathbf{e}_0^\dagger(n)\mathbf{e}_1(n) = \mathbf{0}, \tag{6.3.2a}$$

$$\mathbf{e}_0^\dagger(n)\mathbf{e}_0(n) + \mathbf{e}_1^\dagger(n+1)\mathbf{e}_1(n+1) = \mathbf{I}, \tag{6.3.2b}$$

for all n. By definition, $\rho(n) =$ rank of $\mathbf{e}_1(n)$. From (6.3.2a) we see that rank of $\mathbf{e}_0(n)$ is at most $M-\rho(n)$. Using the facts that the rank of $\mathbf{e}_0^\dagger(n)\mathbf{e}_0(n) =$ the rank of $\mathbf{e}_0(n)$ and the rank of $\left(\mathbf{I}-\mathbf{e}_1^\dagger(n+1)\mathbf{e}_1(n+1)\right)$ is at least $M-\rho(n+1)$, we conclude that

$$\rho(n) \le M - \text{rank of } \mathbf{e}_0(n) \le \rho(n+1). \tag{6.3.3}$$

■

Applying the above theorem to Example 6.2.1, we conclude that the lossless system in (6.2.13) is a NIL system since its instantaneous degree increases from zero to one at time $n = 0$. This result is consistent with that obtained from Theorem 5.5.3.

Theorem 6.3.2. *Degree of IIL TVLOT:* A TVLOT is invertible inverse lossless (IIL) TVLOT if and only if its instantaneous degree $\rho(n)$ is time-invariant. ■

Proof:

1. "If" part: See Section 6.3.2 for a constructive proof.
2. "Only if" part: Assume that the system given in (6.3.1) is an IIL TVLOT. From Section 5.5, we know that the unique inverse is given by $\mathcal{G}$: $\widehat{\mathbf{x}}(n) = \mathbf{e}_0^\dagger(n)\mathbf{y}(n) + \mathbf{e}_1^\dagger(n+1)\mathbf{y}(n+1)$. Consider the system $\mathcal{W}$: $\mathbf{w}(n) = \mathbf{e}_1^\dagger(n)\mathbf{y}(n) + \mathbf{e}_0^\dagger(n)\mathbf{y}(n-1)$. Clearly the system $\mathcal{W}$ is lossless because the system $\mathcal{G}$ is lossless. Therefore we can apply Theorem 6.3.1 to the system $\mathcal{W}$ to obtain the following result:

$$\text{rank of } \mathbf{e}_0(n+1) \le M - \text{rank of } \mathbf{e}_1(n) \le \text{rank of } \mathbf{e}_0(n), \tag{6.3.4}$$

where $\rho(n) =$ rank of $\mathbf{e}_1(n)$. Combining (6.3.3) and (6.3.4), we have proved that $\rho(n)$ is a constant.■

Theorem 6.3.1 gives a simple test of non losslessness of first order LTV systems. If the instantaneous degree of a first order LTV system decreases for some n, then it is guaranteed to be non lossless. Theorem 6.3.2 can be used to verify the losslessness of the inverse of a TVLOT system.

6.3.2. Factorization of TVLOT

In this section, we will show that all IIL TVLOTs (i.e., TVLOT with constant degree $\rho(n) = \rho$) can be factorized *uniquely* as a cascade of ρ perpendicular, degree-one building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ (see Section 6.2.1). Since there is only one delay in each building block, the factorization is minimal in terms of delay. Moreover, the building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ are invertible, and their inverses have the form $\mathcal{D}(\mathcal{Z}, \mathbf{v}(n))$. Therefore the unique inverse of IIL TVLOT is also factorizable. Similar to the case of degree-one lossless system, the coefficients of an TVLOT system satisfy the following:

Theorem 6.3.3. *Complete Characterization of IIL TVLOTs:* The system in (6.3.1) is a TVLOT with a constant degree ρ if and only if the coefficients can be expressed as

$$\mathbf{e}_0(n) = \mathbf{U}(n) \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I}_{M-\rho} \end{bmatrix} \mathbf{V}^\dagger(n), \quad \text{and} \quad \mathbf{e}_1(n) = \mathbf{U}_\rho(n)\mathbf{V}_\rho^\dagger(n-1), \tag{6.3.5}$$

where $\mathbf{U}(n) = [\mathbf{u}_0(n)\ \mathbf{u}_1(n)\ \ldots\ \mathbf{u}_{M-1}(n)]$ and $\mathbf{V}(n) = [\mathbf{v}_0(n)\ \mathbf{v}_1(n)\ \ldots\ \mathbf{v}_{M-1}(n)]$ are arbitrary unitary matrices and $\mathbf{U}_\rho(n)$ and $\mathbf{V}_\rho(n)$ are submatrices defined respectively as

$$\mathbf{U}_\rho(n) = [\mathbf{u}_0(n)\ \mathbf{u}_1(n)\ \ldots\ \mathbf{u}_{\rho-1}(n)] \qquad \mathbf{V}_\rho(n) = [\mathbf{v}_0(n)\ \mathbf{v}_1(n)\ \ldots\ \mathbf{v}_{\rho-1}(n)]. \tag{6.3.6}$$

■

The above theorem can be proved by using a procedure similar to the proof of Theorem 6.2.1. Theorem 6.3.1 tells us how all IIL TVLOTs can be captured by two unitary matrices. For all IIL TVLOTs, the linear span of columns of $\mathbf{e}_0(n)$ is in the orthogonal complement of the columns of $\mathbf{e}_1(n)$; the linear span of rows of $\mathbf{e}_0(n)$ is in the orthogonal complement of the rows of $\mathbf{e}_1(n+1)$.

Implementations and Degree of Freedom: From (6.3.5) and (6.3.6), we have the implementation shown in Fig. 6.3.1. The inverse for Fig. 6.3.1 is given by Fig. 6.3.2. Since the system in Fig. 6.3.2 is a cascade of lossless systems (two unitary matrices and a diagonal system with only advanced elements), the inverse system is also lossless. This is consistence with the fact that the inverse of an IIL system is also lossless (Theorem 5.5.3). In the real coefficient case, the unitary matrices $\mathbf{U}(n)$ and $\mathbf{V}^\dagger(n)$ are real and can be implemented by using TV planar rotations, we can obtain an implementation similar to Fig. 6.2.2. Counting the free parameters, we conclude that for a degree ρ IIL TVLOT, the degree of freedom is $0.5M(M-1) + M\rho - 0.5\rho(\rho+1)$. The implementation based on planar rotations gives a *minimal* characterization of IIL TVLOT.

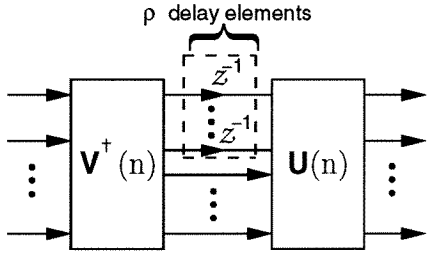

Fig. 6.3.1. Complete characterization of IIL TVLOT of degree ρ. Here $\mathbf{U}(n)$ and $\mathbf{V}(n)$ are unitary matrices.

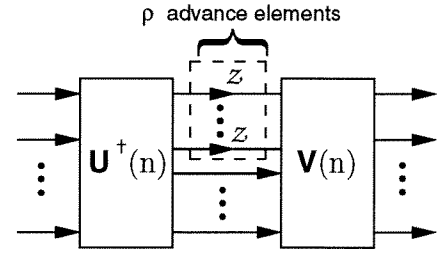

Fig. 6.3.2. Inverse system of the IIL TVLOT in Fig. 6.3.1.

Remarks:

1. We see that for a TVLOT with a constant degree, the inverse shown in Fig. 6.3.2 is lossless. Therefore a TVLOT with a constant degree is an IIL system. The proof for part 1 of Theorem 6.3.2 is complete.

2. When $\rho = 0$, the IIL TVLOT reduces to the special case of LTV transform coding (i.e., a time-dependent unitary matrix); when $\rho = M$, it reduces to a LTV transform coding followed by a pure delay.

Complete Factorization of IIL TVLOTs: Similar to the degree-one lossless case, we can simplify the coefficients as

$$\mathbf{e}_0(n) = \mathbf{P}(n)\big[\mathbf{I} - \mathbf{V}_\rho(n)\mathbf{V}_\rho^\dagger(n)\big], \quad \mathbf{e}_1(n) = \mathbf{P}(n)\big[\mathbf{V}_\rho(n)\mathbf{V}_\rho^\dagger(n-1)\big], \tag{6.3.7}$$

where $\mathbf{P}(n) = \mathbf{U}(n)\mathbf{V}^\dagger(n)$. Since $\mathbf{U}(n)$ can be arbitrary unitary matrix, the unitary matrix $\mathbf{P}(n)$ is arbitrary. Using the above equation, we arrive at the implementation shown in Fig. 6.3.3. Since $\mathbf{V}_\rho(n)$ is a submatrix of a unitary matrix, we have $\mathbf{V}_\rho^\dagger(n)\mathbf{V}_\rho(n) = \mathbf{I}_\rho$. This implies the vectors $\mathbf{v}_k(n)$ for $0 \le k \le \rho - 1$ are perpendicular to each other. Recall from Section 6.2.1 and (6.2.11) that the LTV system from $\mathbf{x}(n)$ to $\mathbf{y}'(n)$ in Fig. 6.3.3 is a cascade of ρ perpendicular lossless dyadic-based building blocks. Using this fact, we arrive at the factorization in Fig. 6.3.4. The ordering of the lossless dyadic-based systems $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$ in Fig. 6.3.3 does not matter because the building blocks are perpendicular. From Fig. 6.3.4, it is clear that the inverse system can be obtained by inverting the building blocks and $\mathbf{P}(n)$. Therefore, the inverse of an IIL TVLOT can be realized as a cascade of $\mathcal{D}(\mathcal{Z}, \mathbf{v}_i(n))$ followed by $\mathbf{P}^\dagger(n)$, as shown in Fig. 6.3.5. Summarizing all the results, we have proved

Theorem 6.3.4. *Complete Factorization of IIL TVLOT:* The first order system in (6.3.1) is an IIL TVLOT with degree ρ if and only if it can be factorized in the perpendicular lossless dyadic-based building blocks as shown in Fig. 6.3.4. Moreover the inverse is given by Fig. 6.3.5. ■.

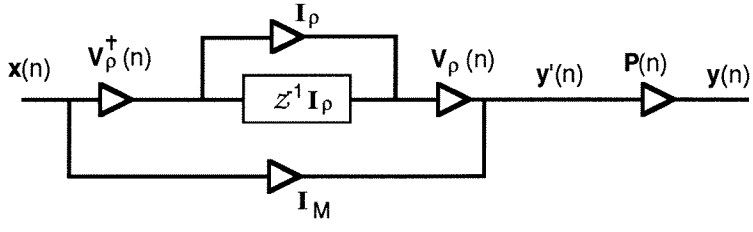

Fig. 6.3.3. Another characterization of IIL TVLOT of degree ρ. The matrix $\mathbf{V}_\rho(n)$ is defined in (6.3.6) and $\mathbf{P}(n)$ is unitary.

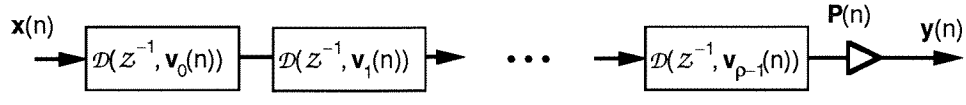

Fig. 6.3.4. Complete factorization of the IIL TVLOT of degree ρ. Here $\mathbf{v}_i^\dagger(n)\mathbf{v}_j(n) = \delta(i-j)$ and $\mathbf{P}(n)$ is unitary.

Remark: We can also simplify the coefficients as in the form similar to (6.2.8). In this case, we can obtain another implementation of the IIL TVLOT as a cascade of $\mathbf{P}(n)$ *followed by* the lossless dyadic-based building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$.

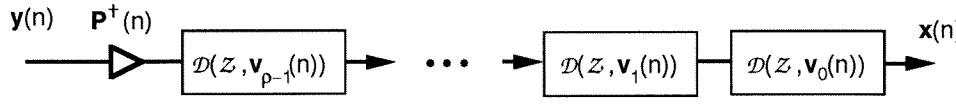

Fig. 6.3.5. Implementation of the inverse of IIL TVLOT in factorized form.

Example 6.3.1. *An Unfactorizable NIL TVLOT:* Consider the following first order system:

$$\mathbf{e}_0(n) = \begin{cases} \mathbf{I}, & \text{for } n < 0; \\ \mathbf{I} - (1 + \sqrt{1-\gamma^2})\mathbf{u}(0)\mathbf{u}^\dagger(0), & \text{for } n = 0; \\ \mathbf{I} - \mathbf{u}(n)\mathbf{u}^\dagger(n), & \text{for } n > 0, \end{cases} \quad \mathbf{e}_1(n) = \begin{cases} \mathbf{0}, & \text{for } n < 1; \\ \gamma\mathbf{u}(1)\mathbf{u}^\dagger(0), & \text{for } n = 1; \\ \mathbf{u}(n)\mathbf{u}^\dagger(n-1), & \text{for } n > 1, \end{cases} \tag{6.3.8}$$

where $|\gamma| \leq 1$ and $\mathbf{u}(n)$ are unit norm vectors. It can be verified by direct substitution into (5.5.4) that the above first order system is lossless, so it is a TVLOT. It is clear that it is a NIL TVLOT since its instantaneous degree increases (Theorem 6.3.2). However, unless $|\gamma| = 1$, the NIL TVLOT in (6.3.8) cannot be factorized into the dyadic-based building block. Because if it could, $\mathbf{e}_0(n)$ should always be of the form $[\mathbf{I} - \mathbf{v}(n)\mathbf{v}^\dagger(n)]\mathbf{P}(n)$, where $\mathbf{v}(n)$ are either zero or unit norm vectors and $\mathbf{P}(n)$ is a unitary matrix. That means, $\mathbf{e}_0(n)$ should be either a singular matrix or a unitary matrix. But from (6.3.8), we see that $\mathbf{e}_0(0)$ is neither singular nor unitary for $|\gamma| \neq 1$.

Recall from Example 6.2.1 that the system in (6.2.13) is a NIL TVLOT. This NIL TVLOT is factorizable because it is already in factorized form. Combining this result and the result in Example 6.3.1, we conclude that there are factorizable and unfactorizable NIL TVLOTs.

6.4. FACTORIZABILITY OF HIGHER ORDER LOSSLESS SYSTEMS

In the previous section, we proved that all IIL TVLOT can be factorized into the degree-one building blocks. We also know that there are factorizable and unfactorizable NIL systems. However we still don't know if all IIL systems are factorizable. More generally, how to determine if a lossless system is factorizable? In this section, we will give several necessary conditions for a factorizable lossless system. These necessary conditions give simple tests for unfactorizable systems. Using these tests, we are able to show some unfactorizable IIL examples. So unlike TVLOT, an IIL system of order > 1 could be unfactorizable. Moreover, we will also give a sufficient condition for factorizability of lossless LTV systems.

6.4.1. Higher Order Lossless Systems and Necessary Conditions for Factorizability

The TVLOTs are first order lossless LTV systems. One way to generate higher order lossless systems is to cascade N sections of the dyadic-based building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$, with $\mathbf{v}_i^\dagger(n)\mathbf{v}_i(n) = 1$. If none of the adjacent building blocks are perpendicular to each other (in the sense defined in Section 6.2.1 Property 5), then the result of the cascade is an N-th order lossless system. If some of the adjacent building blocks are perpendicular, then the order can be smaller than N. In the extreme case of TVLOT, all the building blocks are perpendicular. The lossless systems constructed by this method have the same

number of building blocks for all time n. Since the inverses of the building blocks are lossless, so is their cascade. Therefore we conclude that the above construction will always gives IIL systems.

To construct examples of higher order NIL systems, recall from (6.2.13) of Example 6.2.1 that if the vector $\mathbf{v}(n)$ in a dyadic building block $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ changes from a zero vector to a unit norm vector, then $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ is a lossless system with non lossless inverse, i.e., it is a NIL system. By cascading N sections of such $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$ where $\mathbf{v}_i(n)$ are allowed to switch from zero to unit norm vectors, we can get a NIL system of order N. Also recall from Example 6.2.1 that if the vector $\mathbf{v}(n)$ changes from a zero vector to a unit norm vector, the dyadic building block $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ is no longer lossless. Therefore we conclude that the number of building blocks in a factorizable lossless system cannot be decreasing. Summarizing the results, we have

Theorem 6.4.1. *Number of Building Blocks of Factorizable Systems:* If a lossless LTV system is factorizable in terms of degree-one lossless building blocks, the number of building blocks cannot decrease with time. Moreover the factorizable lossless system is IIL if and only if the number of building blocks is a constant with respect to time. ■

The above theorem can be used to determine if a cascade of building blocks is NIL or IIL. However it is not very useful for testing the factorizability of a lossless system because it assumes that the system is given in factorized form. In the following, we will give some other necessary conditions which lead to simple tests.

Unfactorizability of Non Trivial Scalar Lossless Systems: In the scalar case, we know from Section 6.2 that the degree-one building block $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ reduces to a delay followed by a unit magnitude multiplier. If a scalar lossless system is factorizable using these building blocks, then it should be the cascade of these trivial building blocks. Therefore the output of the factorizable scalar lossless system can always be written as:

$$y(n) = a(n)x(n - k(n)), \tag{6.4.1}$$

where $k(n)$ is a non decreasing (because of Theorem 6.4.1) positive integer and $|a(n)| = 0$ or 1 (because it is a product of either zero or unit norm multipliers). Thus we conclude that *all non trivial* (with at least two non zero coefficients at the same time) *lossless scalar systems are unfactorizable* in terms of degree-one building blocks. Therefore the lossless scalar LTV system given in Example 5.6.1 is an unfactorizable IIL system.

To determine the exact relation between $a(n)$ and $k(n)$, assume that we start the system at $n = n_0$ with the following initial conditions:

$$a(n_0 - 1) = 1, \quad k(n_0 - 1) = 0, \quad x(n_0 - 1) = 0. \tag{6.4.2}$$

Then it can be shown that the coefficient $|a(n)| = 1$ whenever n satisfies the condition $n - k(n) = \hat{n} - k(\hat{n}) + 1$, where $\hat{n}$ is the largest integer $< n$ such that $a(\hat{n}) = 1$.

A Necessary Condition for Factorizability: Consider the $M \times M$ causal lossless system given in (6.1.1a) with the coefficients $\mathbf{e}_k(n)$. Suppose that the system is FIR, that is there is an N such that $\mathbf{e}_k(n) = \mathbf{0}$ for $k > N$ for all n. Let $i(n)$ be the largest integers such that $\mathbf{e}_k(n) = \mathbf{0}$ for $k < i(n)$. Therefore we have $\mathbf{e}_{i(n)}(n) \neq \mathbf{0}$. If the system is factorizable, then it is a cascade of the dyadic-based building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n))$. Since the first coefficient of $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n))$ is of the form $\left[\mathbf{I} - \mathbf{v}_k(n)\mathbf{v}_k^\dagger(n)\right]$, the quantity $\mathbf{e}_{i(n)}(n)$ is a product of matrices $\left[\mathbf{I} - \mathbf{v}_k(n)\mathbf{v}_k^\dagger(n)\right]$ followed by a unitary matrix $\mathbf{P}(n)$, where $\mathbf{v}_k(n)$ is either a zero or unit norm vector. This means that $\mathbf{e}_{i(n)}(n)$ is singular unless $\mathbf{v}_k(n) = \mathbf{0}$ for all k which implies that the system is a trivial system which contains only one non zero coefficient. Therefore we conclude that for a non trivial factorizable lossless system, the first non zero coefficient is singular for each n. This gives a quick test for unfactorizable lossless systems. Applying this result to Example 5.3.1, since the first non zero coefficient is nonsingular for all n, the system is an unfactorizable IIL system.

Summary: Summarizing the results on factorizability of lossless systems, we can make the following conclusions:

1. All IIL TVLOTs are factorizable (Section 6.3).
2. All nontrivial SISO lossless systems are unfactorizable (Section 6.4).
3. There are factorizable and unfactorizable NIL systems (Examples 6.2.1(i) and 3.1 respectively).
4. There are factorizable and unfactorizable IIL systems (IIL TVLOT and Example 5.3.1 respectively).

6.4.2. Sufficient Condition for Factorizability

Consider the following N-th order FIR LTV system $\mathcal{H}$:

$$\mathbf{y}(n) = \sum_{k=0}^{N} \mathbf{e}_k(n)\mathbf{x}(n-k). \tag{6.4.3}$$

Supposing that the system is lossless, i.e., the coefficients satisfy (5.5.4), then we can prove the following

Theorem 6.4.2. *Order Reductibility:* Consider the lossless system $\mathcal{H}$ given in (6.4.3). If the highest order coefficient $\mathbf{e}_N(n)$ has a constant rank ρ for all n, then the system $\mathcal{H}$ can be factorized as a cascade of a causal lossless system $\mathcal{H}'$ whose order is at most $N-1$, followed by an IIL TVLOT block of degree ρ. ■

Before proving the theorem, we notice a few things. If $\mathcal{H}'$ also satisfies the condition in the above theorem, we can apply the above order reduction procedure to $\mathcal{H}'$ to further reduce the order. If the order reduction procedure is applicable at every step, we will finally reduce the lossless system $\mathcal{H}$ to a zeroth-order lossless system, which is a unitary matrix $\mathbf{P}(n)$. In this case, the lossless system can be realized as a cascade of IIL TVLOTs, which implies the lossless system itself is IIL. In the special case of LTI systems, this order reduction procedure is always possible because the coefficient always has a fixed rank. Therefore a LTI PU system is always factorizable. The order reduction for the LTI case is also given in [Gop95]. Before we prove Theorem 6.4.2, it should be mentioned that the constant rank condition on $\mathbf{e}_N(n)$ is not a necessary condition as shown next.

Example 6.4.1. *Factorizable System Which Violates the Constant Rank Condition:* Consider a cascade of two degree one building blocks, $\mathcal{D}(z^{-1}, \mathbf{v}_1(n))\mathcal{D}(z^{-1}, \mathbf{v}_0(n))$. Let $\mathbf{v}_0(n)$ and $\mathbf{v}_1(n)$ be unit norm vectors such that $\mathbf{v}_0^\dagger(2n)\mathbf{v}_1(2n) = 0$ and $\mathbf{v}_0^\dagger(2n-1)\mathbf{v}_1(2n-1) \neq 0$. Then one can verify that the highest order coefficient of $\mathcal{D}(z^{-1}, \mathbf{v}_1(n))\mathcal{D}(z^{-1}, \mathbf{v}_0(n))$, denoted as $\mathbf{e}_2(n)$, has the following form: $\mathbf{e}_2(2n-1) = \mathbf{0}$ and $\mathbf{e}_2(2n) = \Big(\mathbf{v}_1^\dagger(2n-1)\mathbf{v}_0(2n-1)\Big)\mathbf{v}_1(2n)\mathbf{v}_0^\dagger(2n-2)$. It is clear that this IIL system does not satisfy the constant rank condition of Theorem 6.4.2, though it is a cascade of two degree-one factors (hence "factorizable").

Proof of Theorem 6.4.2: If $\rho = M$, then one can verify that the condition for losslessness in (5.5.4) implies that $\mathbf{e}_k(n) = \mathbf{0}$ for $0 \leq k \leq N-1$ and $\mathbf{e}_N^\dagger(n)\mathbf{e}_N(n) = \mathbf{I}$. The system $\mathcal{H}$ reduces to the trivial system $\mathbf{y}(n) = \mathbf{e}_N(n)\mathbf{x}(n-N)$ for unitary $\mathbf{e}_N(n)$. Therefore we can assume $1 \leq \rho \leq M-1$. Since $\mathbf{e}_N(n)$ has rank ρ, it can be written as $\mathbf{e}_N(n) = \mathbf{U}_\rho(n)\mathbf{V}_\rho(n)$, where the $M \times \rho$ matrices $\mathbf{U}_\rho(n)$ and $\mathbf{V}_\rho(n)$ are given in (6.3.6). Since $\mathbf{u}_i(n)$ are independent, we can apply the invertible Gram-Schmidt orthonormalization procedure [Hor85] so that $\mathbf{u}_i(n)$ are orthonormal. Therefore without lost of generality, we can assume that $\mathbf{U}_\rho^\dagger(n)\mathbf{U}_\rho(n) = \mathbf{I}_\rho$. Consider the system $\mathcal{H}'$ which is a cascade of $\mathcal{H}$ followed by the following anticausal system:

$$\mathcal{F}(\mathcal{Z}) = \prod_{k=0}^{\rho-1} \mathcal{D}(\mathcal{Z}, \mathbf{u}_k(n)). \tag{6.4.4}$$

The above LTV system $\mathcal{F}(\mathcal{Z})$ is lossless since it is the inverse of an IIL TVLOT. Note that in this case the ordering of $\mathcal{D}(\mathcal{Z}, \mathbf{u}_k(n))$ does not matter because these building blocks are perpendicular (see Property 5 in Section 6.2.1). The system $\mathcal{H}'$ has order at most equal to N (because the IIL TVLOT is anticausal) and the N-th order coefficient $\mathbf{e}'_N(n)$ can be written as:

$$\mathbf{e}'_N(n) = \Big[\mathbf{I} - \mathbf{U}_\rho(n)\mathbf{U}_\rho^\dagger(n)\Big]\mathbf{U}_\rho(n)\mathbf{V}_\rho^\dagger(n) = \mathbf{0}. \tag{6.4.5}$$

Therefore $\mathcal{H}'$ has order $\leq N-1$. It remains to show that the system $\mathcal{H}'$ is causal and lossless. The losslessness of $\mathcal{H}'$ follows directly from that of $\mathcal{H}$ and $\mathcal{F}(\mathcal{Z})$. To prove the causality, recall from (5.5.4) that we have the condition $\mathbf{e}_N^\dagger(n)\mathbf{e}_0(n) = \mathbf{0}$ for all n. Since the vectors $\mathbf{v}_k(n)$ are independent, the above condition implies that $\mathbf{U}_\rho^\dagger(n)\mathbf{e}_0(n) = \mathbf{0}$. Therefore, we have

$$\mathbf{e}'_{-1}(n) = \mathbf{U}_\rho(n)\mathbf{U}_\rho^\dagger(n)\mathbf{e}_0(n) = \mathbf{0}. \tag{6.4.6}$$

Thus, the causality of $\mathcal{H}'$ follows. Inverting the anticausal system $\mathcal{F}(\mathcal{Z})$, we conclude that $\mathcal{H}$ is a cascade of a causal lossless system $\mathcal{H}'$ whose order is at most $N-1$, followed by the causal IIL TVLOT block $\mathcal{F}(\mathcal{Z}^{-1}) = \prod_{k=0}^{\rho-1} \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{u}_k(n))$. ∎

6.5. STATE-SPACE MANIFESTATION OF FACTORIZABLE IIL SYSTEMS

In this section, we will consider the state-space representation of LTV systems. The theory is well-known in the LTI case [Deca89, Kai80, Vai93]. We will generalize the concept of reachability and observability

to the LTV case in a way most suited for our purpose. We will prove that for the cascade system of an arbitrary number of dyadic building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n))$, the realization matrix is *unitary*. Furthermore the cascade system is *strongly eternally reachable* and *observable*. We will also prove that the strong eternal reachability and observability imply that the *minimality* of the structure. Thus, the implementation based on factorization is *minimal* in terms of delays as well as the number of building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n))$. A brief introduction to the continuous-time reachability and observability of LTV systems is given in [Deca89]. In the following, we will develop the theory for the discrete-time LTV case.

6.5.1. State-Space Representation of LTV Systems

Consider the following state-space realization of an $M \times M$ LTV system:

$$\mathbf{s}(n+1) = \mathbf{A}(n)\mathbf{s}(n) + \mathbf{B}(n)\mathbf{x}(n), \quad \text{(state equation)}, \tag{6.5.1a}$$

$$\mathbf{y}(n) = \mathbf{C}(n)\mathbf{s}(n) + \mathbf{D}(n)\mathbf{x}(n), \quad \text{(output equation)}, \tag{6.5.1b}$$

where $\mathbf{s}(n) = [s_1(n)\ s_2(n)\ \ldots\ s_\rho(n)]^T$ is the state vector, $\mathbf{x}(n)$ and $\mathbf{y}(n)$ are respectively the input and output vectors. The integer ρ is called the dimension of the state space. In (6.5.1) we have assumed that ρ is time invariant. According to Theorem 6.4.1, the instantaneous degree of a factorizable lossless system could be increasing with time n. Thus the constant degree assumption is a loss of generality. However since a factorizable IIL system has a constant number of building blocks (Theorem 6.4.1), we will see that all factorizable IIL lossless systems have constant ρ. From (6.5.1) we have the implementation in Fig. 6.5.1. Note that the system in Fig. 6.5.1 is always causal. The realization matrix $\mathbf{R}(n)$ is given as

$$\mathbf{R}(n) = \begin{bmatrix} \mathbf{A}(n) & \mathbf{B}(n) \\ \mathbf{C}(n) & \mathbf{D}(n) \end{bmatrix}. \tag{6.5.2}$$

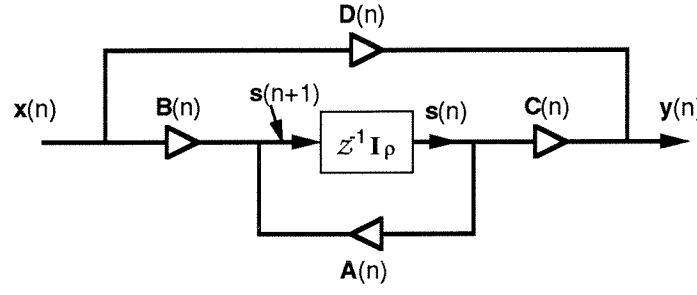

Fig. 6.5.1. State-space implementation of a LTV system.

Time-Varying Impulse Response: Assuming that we start the system at $n = n_0$ with zero initial condition, using (6.5.1) the output of the system can be expressed as:

$$\mathbf{y}(n) = \sum_{k=0}^{n-1-n_0} \mathbf{C}(n)\Phi(n, n-k)\mathbf{B}(n-1-k)\mathbf{x}(n-1-k) + \mathbf{D}(n)\mathbf{x}(n), \tag{6.5.3}$$

where the $\rho \times \rho$ state transition matrix $\Phi(n, m)$ is defined as

$$\Phi(n+1, m) = \mathbf{A}(n)\Phi(n, m), \quad \text{and} \quad \Phi(m, m) = \mathbf{I}. \tag{6.5.4}$$

Comparing (6.5.3) and the direct form implementation $\mathbf{y}(n) = \sum_k \mathbf{e}_k(n)\mathbf{x}(n-k)$, we conclude that the impulse response coefficients are

$$\mathbf{e}_{k+1}(n) = \mathbf{C}(n)\Phi(n, n-k)\mathbf{B}(n-1-k), \; k \geq 0 \;\text{ and }\; \mathbf{e}_0(n) = \mathbf{D}(n). \tag{6.5.5}$$

Reachability of LTV Systems: For the LTI case, there are several equivalent definitions of reachability [Deca89, Kai80, Vai93]. In the following, we generalize the one given in Chapter 13 of [Vai93] to the LTV case. Since the implemenation is time-varying, we have to differentiate between the instantaneous and eternal reachabilities which are defined as follows:

Definition 6.5.1. *Reachability:* An implementation is said to be *reachable* at time n_f if we can reach any specified final state $\mathbf{s}^f$ at time n_f (i.e., $\mathbf{s}(n_f) = \mathbf{s}^f$) starting from any initial state by application of an appropriate *finite length* input. If the implementation is reachable for all n, then we say that it is *eternally reachable* (ER).

Let $(\mathbf{A}(n),\ \mathbf{B}(n),\ \mathbf{C}(n),\ \mathbf{D}(n))$ be the state space representation of an implementation of a LTV system as in (6.5.1). We are going to show how the reachability of an implementation depends only on $\mathbf{A}(n)$ and $\mathbf{B}(n)$. Assuming that we start the system at $n = n_0$ with initial condition $\mathbf{s}(n_0)$, from (6.5.1a) we have

$$\mathbf{s}(n) - \Phi(n, n_0)\mathbf{s}(n_0) = \sum_{k=0}^{n-1-n_0} \Phi(n, n-k)\mathbf{B}(n-1-k)\mathbf{x}(n-1-k), \tag{6.5.6}$$

where $\Phi(n, n-k)$ is defined in (6.5.3). From (6.5.6), we see that an implementation is reachable at n if there is a finite integer L such that the following $\rho \times ML$ matrix $\mathbf{R}_{\mathbf{A},\mathbf{B}}(n-1, n-L)$ has full column rank (i.e. column rank $= \rho$, the dimension of the state space):

$$\mathbf{R}_{\mathbf{A},\mathbf{B}}(n, n-L) = [\,\mathbf{B}(n-1) \quad \Phi(n, n-1)\mathbf{B}(n-2) \quad \ldots \quad \Phi(n, n-L+1)\mathbf{B}(n-L)\,]. \tag{6.5.7}$$

In the LTI case, we know [Deca89, Kai80, Vai93] that if we cannot reach a particular final state by applying an input of length ρ, then the final state cannot be reached by applying more inputs (because of Cayley-Hamilton Theorem). In the LTV case, a similar statement does not hold. The fact that the matrix $\mathbf{R}_{\mathbf{A},\mathbf{B}}(n-1, n-\rho)$ does not have full column rank does not imply that $\mathbf{R}_{\mathbf{A},\mathbf{B}}(n, n-L)$ will not have full column rank for all $L > \rho$. Therefore we cannot determine in finite time if a state is unreachable. Therefore in the LTV case, we have the following definition of strong reachability:

Definition 6.5.2. *Strong Reachability:* An implementation is said to be *strongly reachable* at time n if the $\rho \times M\rho$ matrix $\mathbf{R}_{\mathbf{A},\mathbf{B}}(n, n-\rho)$ defined in (6.5.7) has full column rank, i.e., $\mathbf{R}_{\mathbf{A},\mathbf{B}}(n, n-\rho)\mathbf{R}^{\dagger}_{\mathbf{A},\mathbf{B}}(n, n-\rho)$ is nonsingular. If an implementation is strongly reachable for all n, then we say that it is *strongly eternally reachable* (SER).

Observability of LTV Systems: Similar to the case of reachability, we will generalize the definition of LTI observability given in Chapter 13 of [Vai93] to the LTV case.

Definition 6.5.3. *Observability:* An implementation is said to be *observable* at time n if the state $\mathbf{s}(n)$ can be determined *uniquely* by observing a *finite-length* segment of the output. If the implementation is observable for all n, then we say that it is *eternally observable* (EO).

One can show that a state at time n is observable if and only if there is a finite L such that the following $ML \times \rho$ matrix $\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+L-1,n)$ has full row rank:

$$\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+L-1,n) = \begin{bmatrix} \mathbf{C}(n) \\ \mathbf{C}(n+1)\Phi(n+1,n) \\ \vdots \\ \mathbf{C}(n+L-1)\Phi(n+L-1,n) \end{bmatrix}. \tag{6.5.8}$$

We cannot determine in finite time if a state is not observable. Therefore similar to the case of reachability, we have the following definition:

Definition 6.5.4. *Strong Observability:* An implementation is said to be *strongly observable* at time n if the $M\rho \times \rho$ matrix $\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1,n)$ defined in (6.5.8) has full row rank, i.e., $\mathbf{O}^{\dagger}_{\mathbf{A},\mathbf{C}}(n+\rho-1,n)\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1,n)$ is nonsingular. If the implementation is strongly observable for all n, then we say that it is *strongly eternally observable* (SEO).

Minimality of LTV Systems: The reason we introduce the concepts of SER and SEO as in Definitions 6.5.2 and 6.5.4 is that it leads to the minimality of LTV systems. Let $(\mathbf{A}(n),\ \mathbf{B}(n),\ \mathbf{C}(n),\ \mathbf{D}(n))$ be the state-space representation of the system $\mathbf{y}(n) = \sum_k \mathbf{e}_k(n)\mathbf{x}(n-k)$. By using (6.5.5), (6.5.7) and (6.5.8), one can verify that the $ML \times ML$ product matrix of $\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+L-1,n)\mathbf{R}_{\mathbf{A},\mathbf{B}}(n,n-L)$ is related to the impulse coefficients as

$$\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+L-1,n)\mathbf{R}_{\mathbf{A},\mathbf{B}}(n,n-L) = \begin{bmatrix} \mathbf{e}_1(n) & \mathbf{e}_2(n) & \cdots & \mathbf{e}_L(n) \\ \mathbf{e}_2(n+1) & \mathbf{e}_3(n+1) & \cdots & \mathbf{e}_{L+1}(n+1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_L(n+L-1) & \mathbf{e}_{L+1}(n+L-1) & \cdots & \mathbf{e}_{2L-1}(n+L-1) \end{bmatrix}. \tag{6.5.9}$$

By using the above equation, we can show (see Appendix 6.A) that if an implementation is SER and SEO, then we cannot reduce the number of state variables. That is, the implementation is *minimal.* Thus, we have the following theorem.

Theorem 6.5.1. *Minimality:* If an implementation of an LTV is SER and SEO, then it is minimal. ■

6.5.2. State-Space Representation of Factorizable IIL Systems

First let us consider the IIL TVLOT studied in Section 6.3.2. For an IIL TVLOT of degree ρ, we know that the coefficients can be characterized as in (6.3.7). If we take the output of the delay elements in the dyadic-based structure as the state variable, then state vector is $\mathbf{s}_\rho(n) = \mathbf{V}^{\dagger}_{\rho}(n-1)\mathbf{x}(n-1)$. The

state-space representation of the system becomes

$$\begin{bmatrix} \mathbf{s}(n+1) \\ \mathbf{y}(n) \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{V}_\rho^\dagger(n) \\ \mathbf{P}(n)\mathbf{V}_\rho(n) & \mathbf{P}(n)[\mathbf{I} - \mathbf{V}_\rho(n)\mathbf{V}_\rho^\dagger(n)] \end{bmatrix}}_{\mathbf{R}(n)} \begin{bmatrix} \mathbf{s}(n) \\ \mathbf{x}(n) \end{bmatrix}, \tag{6.5.10}$$

where $\mathbf{P}(n)$ is an arbitrary unitary matrix and $\mathbf{V}_\rho^\dagger(n)\mathbf{V}_\rho(n) = \mathbf{I}_\rho$. One can verify by direct substitution that the realization matrix $\mathbf{R}(n)$ is unitary, i.e., $\mathbf{R}^\dagger(n)\mathbf{R}(n) = \mathbf{I}_{M+\rho}$. For the special case of degree one IIL TVLOT, the system reduces to the dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ in Fig. 6.2.4 with $\mathbf{v}^\dagger(n)\mathbf{v}(n) = 1$. In this case, the state vector is the scalar quantity $s(n)$ as indicated in Fig. 6.2.4. For the more general case of the cascade of arbitrary number of building blocks, we have the following:

Theorem 6.5.2. *Unitariness of Realization Matrix:* Consider the cascade implementation of factorizable IIL system $\mathcal{H} = \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n)) \ldots \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_1(n))$ with $\mathbf{v}_i^\dagger(n)\mathbf{v}_i(n) = 1$ for $1 \le i \le k$. Then for any integer k, the realization matrix $\mathbf{R}(n)$ of the cascade implementation $\mathcal{H}$ satisfies the following:

(a) $\mathbf{A}(n)$ in (6.5.2) is a lower (or upper) triangular matrix with zero diagonal elements;

(b) $\mathbf{R}(n)$ is unitary. ■

Proof: Part (a) is clear since the state variable of the i-th section does not depend the state variables of the j-th section for all $j \ge i$. To prove part (b), we denote the output of the i-th section as $\mathbf{y}_i(n)$. Since the realization matrix of the dyadic-based structure is unitary, we have

$$\mathbf{y}_{i-1}^\dagger(n)\mathbf{y}_{i-1}(n) + |s_i(n)|^2 = \mathbf{y}_i^\dagger(n)\mathbf{y}_i(n) + |s_i(n+1)|^2, \quad \text{for } 1 \le i \le k, \tag{6.5.11}$$

where $\mathbf{y}_0(n) = \mathbf{x}(n)$ and $\mathbf{y}_k(n) = \mathbf{y}(n)$. Summing up all the k terms in (6.5.11), we get

$$\sum_{i=1}^{k} |s_i(n+1)|^2 + \mathbf{y}^\dagger(n)\mathbf{y}(n) = \sum_{i=1}^{k} |s_i(n)|^2 + \mathbf{x}^\dagger(n)\mathbf{x}(n), \quad \forall n. \tag{6.5.12}$$

Since the right-hand sight of (6.5.12) is arbitrary, we conclude that the realization matrix of the cascade system $\mathcal{H}$ is unitary. ■

From Theorem 6.4.1, we know that if an IIL system is factorizable, then the number of building blocks $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$ is time-invariant. Therefore combining the results in Theorems 6.4.1 and 6.5.2, we conclude that *there is always a unitary realization matrix for any factorizable IIL system.*

Minimal Factorization of IIL Systems: Consider the dyadic-based structure $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$ with unit norm vector $\mathbf{v}_i(n)$. Taking the state variable to be $s_i(n) = \mathbf{v}_i^\dagger(n)\mathbf{x}(n)$, then it is clear that the dyadic-based structure is SER and SEO. More generally, we can prove the following:

Theorem 6.5.3. *Strong reachability and observability:*
Consider the factorized system $\mathcal{H} = \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n)) \ldots \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_1(n))$, with $\mathbf{v}_i^\dagger(n)\mathbf{v}_i(n) = 1$ for $1 \le i \le k$. This cascade implementation of $\mathcal{H}$ is SER and SEO and hence minimal. ■

Proof: Because of Theorem 6.5.1, we need only to prove the SER and SEO of the structure. Let $s_i(n)$ and $\mathbf{y}_i(n)$ be respectively the state variable and the output of the i-th section $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_i(n))$.

1. *Strong reachability.* The proof is based on induction on k. If $k = 1$, the cascade system $\mathcal{H}$ reduces to the dyadic-based structure which is SCR. Assuming that the theorem is true for $k = L$, we will show that it is true for $k = L + 1$. Let $\mathbf{s}^f = [s_1^f \ \ldots \ s_{L+1}^f]^T$ be the final state vector we want to reach. We will construct $\{\mathbf{x}(n-1), \ldots, \mathbf{x}(n-L-2)\}$ such that $\mathbf{s}(n) = \mathbf{s}^f$. Let $\mathbf{y}_L^f(n-1)$ be such that $s_{L+1}(n) = \mathbf{v}_{L+1}^\dagger(n-1)\mathbf{y}_L^f(n-1) = s_{L+1}^f$ (this is always possible because $\mathbf{v}_{L+1}(n) \neq \mathbf{0}$). Therefore the problem reduces to choosing the input $\mathbf{x}(n)$ such that $\mathbf{y}_L(n-1) = \mathbf{y}_L^f(n-1)$ and the state vector $\hat{\mathbf{s}}(n-1) = [s_1(n-1) \ \ldots \ \mathbf{s}_L(n-1)]^T$ satisfy the following:

$$\mathbf{R}_L(n-1)\begin{bmatrix}\hat{\mathbf{s}}(n-1)\\ \mathbf{x}(n-1)\end{bmatrix} = \begin{bmatrix}\hat{\mathbf{s}}^f(n)\\ \mathbf{y}_L^f(n-1)\end{bmatrix}, \tag{6.5.13}$$

where $\mathbf{R}_L(n)$ is the unitary realization matrix of the cascade of first L sections (see Theorem 6.5.2). By the hypothesis of the induction that $\mathcal{H}$ is strongly reachable for $k = L$, we can choose $\{\mathbf{x}(n-2), \ldots, \mathbf{x}(n-L-2)\}$ so that $\hat{\mathbf{s}}(n-1)$ satisfies (6.5.12). Therefore the cascade of $L+1$ sections is SER.

2. *Strong observability.* We want to determine $\mathbf{s}(n)$ uniquely by observing $\{\mathbf{y}(n), \ldots, \mathbf{y}(n+k-1)\}$. First Note that $\mathbf{y}(n) = \mathbf{v}_k(n)s_k(n) + [\mathbf{I} - \mathbf{v}_k(n)\mathbf{v}_k^\dagger(n)]\mathbf{y}_{k-1}(n)$. Using the fact that $\mathbf{v}_k^\dagger(n)\mathbf{v}_k(n) = 1$, we find that $s_k(n) = \mathbf{v}_k^\dagger(n)\mathbf{y}(n)$. Since $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_k(n))$ is invertible, knowing $\mathbf{y}(n+1), \ldots, \mathbf{y}(n+k-1)$ and $s_k(n)$, we can uniquely determine $\mathbf{y}_{k-1}(n), \ldots, \mathbf{y}_{k-1}(n+k-2)$, from which we can find $s_{k-1}(n)$. Repeat the above process, we can determine all the state variables $s_i(n)$ for $1 \leq i \leq k$. The cascade system $\mathcal{H}$ is therefore SEO. ■

Combining the results in Theorems 6.4.1, 6.5.1 and 6.5.3, we have proved that the factorization of IIL system is *minimal*. We cannot find a structure which has a smaller number of delays.

6.6. IIR LATTICE STRUCTRES FOR LOSSLESS LTV SYSTEMS

In all the previous discussions, we have considered only the FIR case (except Section 6.2.2). In the IIR LTV case, it is not easy to ensure the stability. In the LTV case, there are several types of stability [Deca89,Gra75]. In this section, we will study only two of them, namely the bounded input bounded output (BIBO) stability and l_2 stability, which are defined as follows:

Definition 6.6.1. *BIBO and l_2 Stability:* A system is said to be BIBO stable if bounded input produces bounded output. A system is said to be l_2 stable if a finite energy input generates a finitew energy output.

In general BIBO stability and l_2 stability are different. To see this, consider the idea LTI lowpass filter and the LTV system $y(n) = x(0)$. The former is l_2 stable but not BIBO stable while the latter is BIBO stable but not l_2 stable.

6.6.1. Stability of LTV Normalized IIR Lattice

Consider the LTV normalized IIR lattice structure given in Fig. 6.6.1, where the number of delays ρ is time-invariant. For an introduction to the theory of LTI IIR lattice, see Chapter 7 of [Mit93]. In

the LTV case, it was shown in [Gra75] that the system in Fig. 6.6.1 preserves the energy from input to output. Using this energy balance property, the authors in [Gra75] showed that the normalized IIR lattice structure in Fig. 6.6.1 is l_2 stable if the time-varying lattice coefficients $|\alpha_k(n)| \leq \gamma < 1$. In this section, we will show that the structure in Fig. 6.6.1 is BIBO stable in addition to being l_2 stable. To prove the BIBO stability of the normalized lattice, we need the following lemma and the definition of matrix norm.

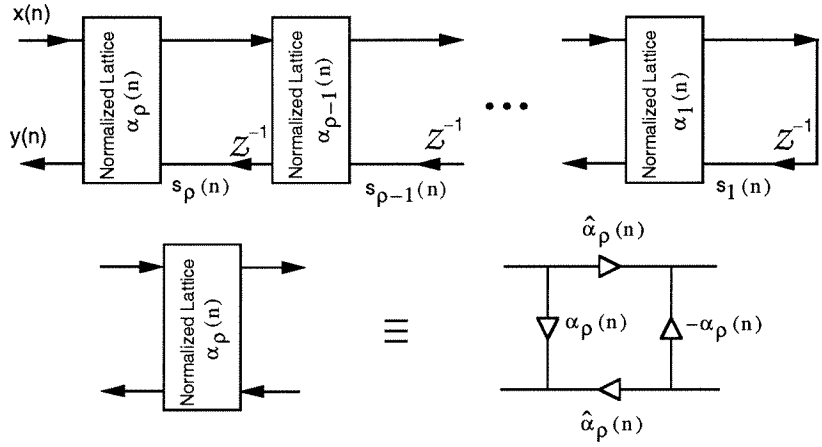

Fig. 6.6.1. LTV normalized IIR lattice structure. Here $\hat{\alpha}_\rho(n) = \sqrt{1-|\alpha_\rho(n)|^2}$.

Definition 6.6.2. *Matrix Norm* [Hor85]: The norm of a matrix $\mathbf{A}$ (denoted as $\| \mathbf{A} \|$) is defined as $\max_{\|\mathbf{v}\|=1} \mathbf{v}^\dagger \mathbf{A}^\dagger \mathbf{A} \mathbf{v}$.

It can be shown [Hor85] that $\| \mathbf{Av} \| \leq \| \mathbf{A} \| \| \mathbf{v} \|$ and $\| \mathbf{A}_1 \mathbf{A}_2 \| \leq \| \mathbf{A}_1 \| \| \mathbf{A}_2 \|$. By using these norm properties, we can prove the following lemma:

Lemma 6.6.1. Let $(\mathbf{A}(n), \mathbf{B}(n), \mathbf{C}(n), D(n))$ be a state-space description of a LTV system such that the realization matrix $\mathbf{R}(n)$ is unitary. Let ρ be the dimension of the state-space. Then for all n, $\| \Phi(n+\rho, n) \| \leq 1$, with equality if and only if the LTV system is not SEO. ■

Proof: From (6.5.4), we have $\| \Phi(n+\rho, n) \| = \| \mathbf{A}(n+\rho-1)\Phi(n+\rho-1, n) \| \leq \prod_{i=0}^{\rho-1} \| \mathbf{A}(n+i) \|$. Since $\mathbf{R}(n)$ is unitary, we have

$$\mathbf{A}^\dagger(n)\mathbf{A}(n) + \mathbf{C}^\dagger(n)\mathbf{C}(n) = \mathbf{I}, \tag{6.6.1}$$

for all n. It immediately follows from (6.6.1) that $\| \Phi(n+\rho, n) \| \leq 1$. Using (6.6.1) and the recursive formula in (6.5.4), we can expand $\Phi^\dagger(n+\rho, n)\Phi(n+\rho, n)$ as follows

$$\begin{aligned}\Phi^\dagger(n+\rho, n)\Phi(n+\rho, n) = \mathbf{I} - \mathbf{C}^\dagger(n)\mathbf{C}(n) - \Phi^\dagger(n+1, n)\mathbf{C}^\dagger(n+1)\mathbf{C}(n+1)\Phi(n+1, n) \\ - \ldots - \Phi^\dagger(n+\rho-1, n)\mathbf{C}^\dagger(n+\rho-1)\mathbf{C}(n+\rho-1)\Phi(n+\rho-1, n).\end{aligned} \tag{6.6.2}$$

By using the definition of $\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1, n)$ in (6.5.8), we see that

$$\Phi^\dagger(n+\rho, n)\Phi(n+\rho, n) = \mathbf{I} - \mathbf{O}_{\mathbf{A},\mathbf{C}}^\dagger(n+\rho-1, n)\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1, n). \tag{6.6.3}$$

Therefore we conclude $\| \Phi(n+\rho, n) \| = 1$ if and only if $\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1, n)$ is singular (which implies that the LTV system is not SEO (Definition 6.5.4)). ■

For the LTI case, it was shown in Chapter 7 of [Mit93] that the realization matrix $\mathbf{R}$ of a normalized IIR lattice structure is unitary. This property continues to hold for the LTV case. Therefore the LTV normalized IIR lattice satisfies the condition given in Lemma 6.6.1. Furthermore it is shown in Appendix 6.B that the system in Fig. 6.6.1 is strongly eternally observable (SEO) if $|\alpha_k(n)| \leq \gamma < 1$. Therefore there is some fixed $\epsilon > 0$ such that $\mathbf{O}^{\dagger}_{\mathbf{A},\mathbf{C}}(n+\rho-1, n)\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1, n) \geq \epsilon\mathbf{I}$ (Appendix 6.B). From (6.6.3), we have $\Phi^{\dagger}(n+\rho, n)\Phi(n+\rho, n) \leq (1-\epsilon)\mathbf{I}$ which implies that $\| \Phi(n+\rho, n) \| \leq (1-\epsilon)$. Using (6.5.3), the output $y(n)$ of the IIR LTV system in Fig. 6.6.1 satisfies the following

$$\begin{aligned} |y(n)| &\leq |Dx(n)| + \sum_{k=0}^{n-1-n_0} \| \mathbf{C}(n) \| \| \Phi(n, n-k) \| \| \mathbf{B}(n-1-k) \| \, |x(n-1-k)| \\ &\leq |x(n)| + \sum_{k=0}^{n-1-n_0} \| \Phi(n, n-k) \| \, |x(n-1-k)|, \end{aligned} \tag{6.6.4}$$

where we have used the fact that $\| \mathbf{C}(n) \| \leq 1$ and $\| \mathbf{B}(n-1-k) \| \leq 1$ (which follow from the unitariness of $\mathbf{R}(n)$) in the second inequality. If there is a $B < \infty$ such that the input $|x(n)| \leq B$, then using the fact that $\| \Phi(n+\rho, n) \| \leq (1-\epsilon)$, we get

$$|y(n)| \leq B + B \sum_{k=0}^{n-1-n_0} (1-\epsilon)^{\lfloor k/\rho \rfloor}, \tag{6.6.5}$$

where $\lfloor k/\rho \rfloor$ denotes the largest integer $\leq k/\rho$. From (6.6.5), we conclude that $|y(n)| \leq (\rho\epsilon^{-1}+1)B$ for all n. Therefore the output is bounded. Summarizing the result, we have shown that:

Theorem 6.6.1. *Stability of LTV Normalized IIR Lattice:* The LTV normalized IIR lattice structure in Fig. 6.6.1 is both BIBO stable and l_2 stable if the lattice coefficients $|\alpha_k(n)| \leq \gamma < 1$. ■

Remarks:

1. In the LTI case, Lemma 6.6.1 reduces to the following [Vai93]: If $\mathbf{A}$ is a $\rho \times \rho$ unit norm stable matrix, then $\| \mathbf{A}^{\rho} \| < \mathbf{I}$. It is shown in [Vai87a, Mit93] that the condition $\| \mathbf{A}^{\rho} \| < \mathbf{I}$ is sufficient for preventing zero-input limit cycles.
2. Since $\| \Phi(n+\rho, n) \| < 1$ for a normalized IIR lattice structure with $|\alpha_k(n)| \leq \gamma < 1$, the energy of the state vector has to decrease after ρ time interval if there is no input. Therefore we conclude that the structure is free from zero-input limit cycles.

6.6.2. Stability of the Two-multiplier IIR Lattice Structures

In the LTI case, we know that the normalized IIR lattice is not efficient in terms of computation though it has a better noise performance. There is a more efficient two-multiplier IIR lattice [Vai93, Gra75, Mit93]. In this subsection, we will generalize the LTI two-multiplier lattice structure to the LTV denormalized IIR lattice as shown in Fig. 6.6.2, where the number of sections ρ is a constant independent of n. After some

simplifications, it can be shown that the LTV system in Fig. 6.6.2 is equivalent to that in Fig. 6.6.3. The structure in Fig. 6.6.3 is very similar to that of the normalized IIR lattice structure in Fig. 6.6.1 except the time-dependent multipliers $\hat{\alpha}_k(n) = \sqrt{1 - |\alpha_k(n)|^2}$ between the sections. Because of these multipliers, it can be shown that the LTV system in Fig. 6.6.3 can *never* be lossless unless $\hat{\alpha}_k(n)$ are time-independent, which is equivalent to saying that the magnitude of the lattice coefficients $|\alpha_k(n)|$ is a constant independent of n. To see this, we consider Fig. 6.6.3. Since the system from $w_2'(n)$ to $s_2'(n+1)$ is lossless, we have $\sum_n |w_2'(n)|^2 = \sum_n |s_2'(n+1)|^2$. This implies that $\sum_n |w_2(n)/\hat{\alpha}_1(n)|^2 = \sum_n |s_2(n)/\hat{\alpha}_1(n)|^2$. Therefore $s_2(n)$ in general does not have the same energy as $w_2(n)$ unless $\hat{\alpha}_1(n)$ is a constant. This prove that the first order denormalized IIR lattice is in general not lossless. Continuing the process, we can show that the output $y(n)$ in general does not have the same energy as the input $x(n)$ unless $|\alpha_k(n)|$ are time-independent.

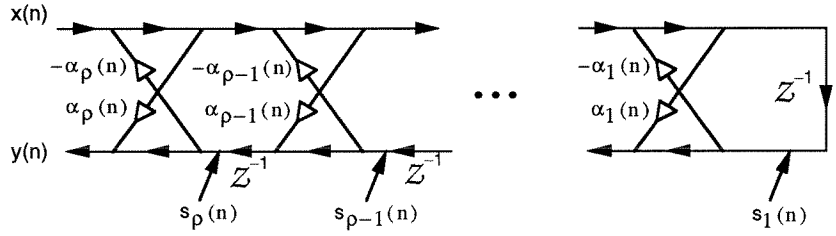

Fig. 6.6.2. LTV denormalized IIR lattice structure.

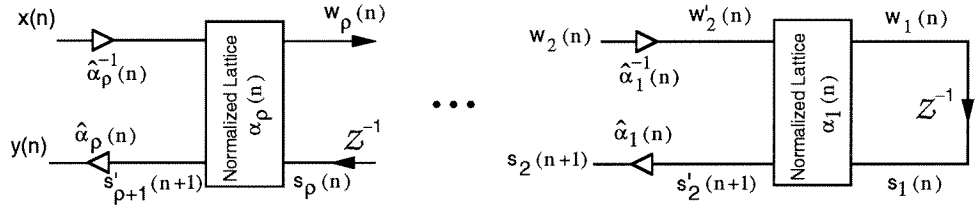

Fig. 6.6.3. Redrawing of Fig. 6.6.2 in terms of normalized building blocks.

In general we cannot prove either the BIBO or l_2 stability of the two-multiplier IIR lattice in Fig. 6.6.2. However in the special case when $|\alpha_k(n)|$ are constant independent of n, both the BIBO and l_2 stability of the structure is guaranteed by the condition $|\alpha_k(n)| \leq \gamma < 1$. The reason is because in this case the time independent multipliers $\hat{\alpha}_k(n)$ can be moved to the left and the resulting structure is very similar to the normalized IIR lattice in Fig. 6.6.1.

6.7. NON LOSSLESS FIR LTV SYSTEMS WITH FIR INVERSES

In this section, we will show how to construct non lossless FIR LTV systems with FIR inverses. The following two classes will be considered: (i) Causal FIR LTV systems with causal FIR inverses, which

are also called the LTV unimodular systems (just by analogy to the LTI case); (ii) causal FIR LTV systems with anticausal FIR inverses (abbreviated as LTV CAFACAFI). For a detail discussion on LTI CAFACAFI systems, see [Vai95c, Vai95d]. First we will construct a degree-one system which can be used to form higher degree systems with FIR inverses.

Theorem 6.7.1. *A Class of Degree-One LTV CAFACAFI:* Consider the following degree-one LTV system $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n), \mathbf{u}(n))$:

$$\mathbf{y}(n) = \Big[\mathbf{I} - \mathbf{u}(n)\mathbf{v}^{\dagger}(n)\Big]\mathbf{x}(n) + \mathbf{u}(n)\mathbf{v}^{\dagger}(n-1)\mathbf{x}(n-1), \tag{6.7.1}$$

where $\mathbf{u}(n)$ and $\mathbf{v}(n-1)$ are non zero vectors. We have the following:

(i) If $\mathbf{u}^{\dagger}(n)\mathbf{v}(n) = 0$ for all n, then $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n), \mathbf{u}(n))$ is a LTV unimodular system. Its unique causal FIR inverse is invertible and can be described as

$$\widehat{\mathbf{x}}(n) = \Big[\mathbf{I} + \mathbf{u}(n)\mathbf{v}^{\dagger}(n)\Big]\mathbf{y}(n) - \mathbf{u}(n)\mathbf{v}^{\dagger}(n-1)\mathbf{y}(n-1). \tag{6.7.2}$$

(ii) If $\mathbf{u}^{\dagger}(n)\mathbf{v}(n) = 1$ for all n, then $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n), \mathbf{u}(n))$ is a LTV CAFACAFI system. Its unique anticausal FIR inverse is invertible and can be described as

$$\widehat{\mathbf{x}}(n) = \Big[\mathbf{I} - \mathbf{u}(n)\mathbf{v}^{\dagger}(n)\Big]\mathbf{y}(n) + \mathbf{u}(n)\mathbf{v}^{\dagger}(n+1)\mathbf{y}(n+1). \tag{6.7.3}$$

Moreover the LTV CAFACAFI system is lossless if and only if $\mathbf{u}(n) = \mathbf{v}(n)$.

The above theorem can be proved by direct substitution. Since the cascade of LTV unimodular systems (or LTV CAFACAFI) is also a LTV unimodular (LTV CAFACAFI) system, we can generate higher degree systems by using the corresponding degree-one system $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n), \mathbf{u}(n))$ given in Theorem 6.7.1. However it should be mentioned that we don't know if the degree-one system in (6.7.1) is a most general LTV unimodular system (or LTV CAFACAFI system). Thus the above construction of higher degree systems might not be complete.

LTV Unimodular Lapped Transform (ULT) and Biorthogonal Lapped Transform (BOLT) [Vai95d]: Consider the cascade system: $\mathcal{H} = \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_{\rho}(n), \mathbf{u}_{\rho}(n)) \ldots \mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}_1(n), \mathbf{u}_1(n))$. Assuming that $\rho < M$. It can be shown that if the vectors $\mathbf{v}_j^{\dagger}(n)\mathbf{u}_i(n) = 0$ for all $i < j$, then the system $\mathcal{H}$ has order one. In this case, we can get either a LTV ULT if $\mathbf{v}_i(n)\mathbf{u}_i(n) = 0$ or a LTV BOLT if $\mathbf{v}_i(n)\mathbf{u}_i(n) = 1$.

6.8. CONCLUDING REMARKS

In this chapter, we showed how to capture all degree-one lossless LTV systems by two unitary matrices (Theorem 6.2.1) and proved that they can be realized as a cascade of a lossless dyadic-based building block and a unitary matrix (Fig. 6.2.3). The dyadic-based building block in Fig. 6.2.4 has many useful properties (Section 6.2.1). The theory of LOT [Mal92] is extended to the LTV case (Section 6.3). We showed that the instantaneous degree of a TVLOT is non decreasing with time n and it is a constant if and only if it is an IIL TVLOT (Theorems 6.3.1 and 6.3.2). All IIL TVLOT can be factorized uniquely

into perpendicular dyadic-based building blocks (Fig. 6.3.4) and the inverse is also factorizable (Fig. 6.3.5). For NIL TVLOT systems, there are factorizable example (Eq. (6.2.13) of Example 6.2.1) and unfactorizable example (Example 6.3.1). Factorizability of higher order lossless LTV systems is also studied (Section 6.4). By using the test for factorizability (Section 6.4.1), we demonstrated that there are unfactorizable IIL systems (Example 5.3.1). A sufficient condition for factorizability which leads to an order reduction procedure is also given (Theorem 6.4.2). We also introduced the concept of SER and SEO (Section 6.5.1). If an implemntation of a LTV system is SER and SEO, then it is minimal (Theorem 6.5.1). In particular, we show that the implementation in terms of building blocks is minimal in terms of delay elements (Theorem 6.5.3). The LTV normalized IIR lattice is proved to be BIBO stable as well as l_2 stable if the lattice coefficients $|\alpha_k(n)| \leq \gamma < 1$ (Theorem 6.6.1).

However there are still many unsolved problems related to the topic of lossless LTV systems. Some of these are stated as follows. From Section 6.3, we know that there exist unfactorizable lossless systems. However in all of our unfactorizable lossless examples, their instantaneous degree is time dependent. This leads us to ask if all lossless systems with time independent degree are factorizable in terms of degree-one lossless building block $\mathcal{D}(\mathcal{Z}^{-1}, \mathbf{v}(n))$ introduced in Section 6.2. In the more general TV biorthogonal case, it is still unknown that if the system given in (6.7.1) is the most general degree one TV unimodular (or TV CAFACAFI) system. A complete characterization of TV ULT or TV BOLT systems is still unknown. In the LTI case, a complete parameterization of the BOLT systems is given in [Vai95d] and it is shown that BOLT systems can always be factorized into degree-one building blocks. In the LTV case, the factorizability of TV BOLT is currently under study.

6.9. APPENDICES

Appendix A. Proof of Theorem 6.5.1

Let $(\mathbf{A}(n),\ \mathbf{B}(n),\ \mathbf{C}(n),\ \mathbf{D}(n))$ be a SER and SEO realization of the LTV system $\mathcal{H}$, with $\mathbf{A}(n)$ being $\rho \times \rho$ matrix. Suppose that there is another SER and SEO realization $(\mathbf{A}'(n),\ \mathbf{B}'(n),\ \mathbf{C}'(n),\ \mathbf{D}'(n))$, with smaller state space dimension. That is, $\mathbf{A}'(n)$ is $\rho' \times \rho'$ with $\rho' < \rho$. By using (6.5.9), we know that

$$\mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho-1,n)\mathbf{R}_{\mathbf{A},\mathbf{B}}(n,n-\rho) = \mathbf{O}_{\mathbf{A}',\mathbf{C}'}(n+\rho-1,n)\mathbf{R}_{\mathbf{A}',\mathbf{B}'}(n,n-\rho). \tag{6.A.1}$$

Premultiplying and postmultiplying (6.A.1) respectively by $\mathbf{O}^{\dagger}_{\mathbf{A},\mathbf{C}}(n+\rho-1,n)$ and $\mathbf{R}^{\dagger}_{\mathbf{A},\mathbf{B}}(n,n-\rho)$, we get

$$\underbrace{\mathbf{O}^{\dagger}_{\mathbf{A},\mathbf{C}}\mathbf{O}_{\mathbf{A},\mathbf{C}}}_{\rho\times\rho}\underbrace{\mathbf{R}_{\mathbf{A},\mathbf{B}}\mathbf{R}^{\dagger}_{\mathbf{A},\mathbf{B}}}_{\rho\times\rho} = \underbrace{\mathbf{O}^{\dagger}_{\mathbf{A},\mathbf{C}}\mathbf{O}_{\mathbf{A}',\mathbf{C}'}}_{\rho\times\rho'}\underbrace{\mathbf{R}_{\mathbf{A}',\mathbf{B}'}\mathbf{R}^{\dagger}_{\mathbf{A},\mathbf{B}}}_{\rho'\times\rho}, \tag{6.A.2}$$

where we have dropped the indices for notational simplicity. Because $(\mathbf{A}(n),\ \mathbf{B}(n),\ \mathbf{C}(n),\ \mathbf{D}(n))$ is SER and SEO, the left-hand side of (6.A.2) is a $\rho \times \rho$ nonsingular matrix. The rank of the matrix on the right-hand side of (6.A.2) is at most $\rho' < \rho$, a contradiction! Therefore we cannot find a realization with fewer than ρ delays.

Appendix B. Proof of SEO of Normalized IIR Lattice

Consider Fig. 6.6.1. Since $\hat{\alpha}_k(n) = \sqrt{1 - |\alpha_k(n)|^2} \geq \beta > 0$ for all n, we have

$$\begin{bmatrix} w_i(n) \\ s_i(n) \end{bmatrix} = \underbrace{\frac{1}{\hat{\alpha}_k(n)} \begin{bmatrix} 1 & -\alpha_k(n) \\ -\alpha_k(n) & 1 \end{bmatrix}}_{\mathbf{T}_k(n)} \begin{bmatrix} w_{i+1}(n) \\ s_{i+1}(n+1) \end{bmatrix}, \tag{6.B.1}$$

where $w_{\rho+1}(n)$ is the input $x(n)$ and $s_{\rho+1}(n+1)$ is the output $y(n)$. Knowing the input $x(i)$ and the output $y(i)$ for $n \leq i \leq n+\rho-1$, we can determine $s_\rho(i)$ and $w_\rho(i)$ for $n \leq i \leq n+\rho-1$ by using (6.B.1). The information of $s_\rho(i)$ and $w_\rho(i)$ for $n \leq i \leq n+\rho-1$ can be used to determine $s_{\rho-1}(i)$ and $w_{\rho-1}(i)$ for $n \leq i \leq n+\rho-2$. Continuing this procedure, we can determine $s_k(n)$ for $1 \leq k \leq \rho$. The structure in Fig. 6.6.1 is therefore SEO. Furthermore, since $\| \mathbf{T}_k(n) \| \leq$ constant $< \infty$, we have $\| \mathbf{O}_{\mathbf{A},\mathbf{C}}(n+\rho, n) \| \geq \epsilon > 0$.

Bibliography

[Ans83] R. Ansari, and B. Liu, "Efficient sampling rate alteration using recursive (IIR) digital filters," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. 31, pp. 1366-73, Dec. 1983.

[Ans87] R. Ansari, and C.-L. Lau, "Two-dimensional IIR filters for exact reconstruction in tree-structured subband decomposition," *Electronics letters,* vol. 23, pp. 633-634, June 1987.

[Ans87] R. Ansari, C. Guillemot, and J. F. Kaiser, "Wavelet construction using Lagrange halfband filters," *IEEE Trans. Circuits and Systems,* vol. 38, no. 9, pp. 1116-1118, Sep. 1991.

[Arr93] J. L. Arrowood Jr. and M. J. T. Smith, "Exact reconstruction analysis/synthesis filter banks with time-varying filters," *in Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing,* pp. 233-236, Minneapolis, MN, April 1993.

[Bar80] C. W.Barnes, and S. Shinnaka, "Block-shift invariance and block implementaion of discrete-time filters," *IEEE Trans. on Circuits and Systems,* vol. CAS-27, pp. 667-672, Aug. 1980.

[Bas95] S. Basu, C.-H. Chiang, and H. M. Choi, "Wavelets and perfect reconstruction subband coding with causal stable IIR filters," *IEEE Trans. Circuits and Systems,* vol. CAS-42, pp. 24-38, Jan. 1995.

[Bel76] M. Bellanger, G. Bonnerot, and M. Coudreuse, "Digital filtering by polyphase network: application to sample rate alteration and filter banks," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. 24, pp. 109-114, April 1976.

[Bla85] R. E. Blahut, *Fast algorithms for digital signal processing,* Addison-Wisley, 1985.

[Bos82] N. K. Bose, *Applied multidimensional systems theory,* New York NY: Van Nostrand Reinhold, 1982.

[Bru92] A. M. Bruekers, and Ad W. M. van den Enden, "New networks for perfect inversion and perfect reconstruction," *IEEE Jour. Selected Areas in Communication,* vol. 10, pp. 130-137, Jan. 1992.

[Bur71] C. S. Burrus, "Block implementation of digital filters," *IEEE Trans. on Circuit Theory,* vol. CT-18, pp. 697-701, Nov. 1971.

[Cha73] S. K. Chan, and L. R. Rabiner,"Analysis of quantization errors in the direct form for finite impulse response digital filters," *IEEE Trans. on Audio and Electroacoustics,* pp. 3354-366, Aug. 1973.

[Che91] T. Chen, and P. P. Vaidyanathan, "Design of IFIR eigenfilters," *in Proc. IEEE Int. Symp. on Circuits and Systems,* pp. 264-267 Singapore 1991.

[Che93] T. Chen, and P. P. Vaidyanathan, "Multidimensional multirate filters and filter banks derived from one-dimensional filters," *IEEE Trans. on Signal Processing,* vol. 41, no. 5, pp. 1749-65, May 1993.

[Che94] T. Chen, and P. P. Vaidyanathan, "Vector space framework for unnification of one- and multidimensional filter bank theory," *IEEE Trans. Signal Processing,* pp. 2006-2022, vol. 42, Aug. 1994.

[Chu93] W. C. Chung, and M. J. T. Smith, "Spatially-varying IIR filter banks for image coding," *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. V 570-573, Minneapolis, April 1993.

[Coh93] A. Cohen, and I. Daubechies, "Non-separable bidimensional wavelet bases," *Preprint,* 1993.

[Coo94] T. Cooklev, A. Nishihara, and M. Sablatash, "Theory of filter banks over finite fields," *in Proc. of Asia Pacific Conference on Circuits and Systems,* pp. 260-265, Taipei, Dec. 1994.

[Cox87] R. V. Cox, et al., "The analog voice privacy system," *AT&T Tech. J.,* 1987.

[Cro83] R. E. Crochiere, and L. R. Rabiner, *Multirate digital signal Processing,* Prentice Hall, 1983.

[Cro77] A. Croisier, D. Esteban, and C. Galand, "Perfect channel splitting by use of interpolation / decimation / tree decomposition techniques," *in Proc. IEEE Int. Symp. on Circuits and Systems,* Patras, Greece, 1977.

[Dau88] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.,* vol. 41, pp. 909-996, Nov. 1988.

[Dau92] I. Daubechies, *Ten lectures on wavelets*, SIAM, CBMS series, April 1992.

[Dau93] I. Daubechies, private communication, 1993.

[Deca89] R. A. Decarlo, *Linear systems: A state variable approach with numerical implementation,* Prentice Hall, Englewood Cliffs, NJ: 1989.

[deQ93] R. L. de Queiroz, and K. R. Rao, "Time-varying lapped transform and wavelet packets," *IEEE Trans. Signal Processing,* pp. 3293-3305, Vol. 41, no. 12, Dec. 1993.

[Djo94] I. Djokovic, and P. P. Vaidyanathan, "Results on biorthogonal filter banks," *Applied and Computational harmonic analysis,* pp. 329-343, vol. 1, 1994.

[Djo94a] I. Djokovic, and P. P. Vaidyanathan, "Generalized sampling theorems in multiresolution subspaces," *Preprint,* Caltech, 1994.

[Dog88] Z. Doganata, P. P. Vaidyanathan and T. Q. Nguyen, "General synthesis procedures for FIR lossless transfer matrices, for perfect-reconstruction multirate filter bank applications," *IEEE Trans. Acoust. Speech, Signal Processing,* Vol. ASSP-36, p. 1561-1574, Oct. 1988.

[Fli94] N. J. Fliege, *Multirate digital signal processing,* Chichester: John Wiley and Sons, 1994.

[For70] G. D. Forney, Jr., "Convolutional codes I: Algebraic structure," *IEEE Trans. Info. Theory,* vol. IT-16, pp. 720-738, Nov. 1970.

[Gil92] A. Gilloire, and M. Vetterli, " Adaptive filtering in subbands with critical sampling: Analysis, experiments and application to acoustic cancellation," *IEEE Trans. on Signal Processing*, vol. 40, pp. 1862-75, Aug. 1992.

[Gop95] R. A. Gopinath, and C. S. Burrus, "Factorization approach to unitary time-varying filter bank trees and wavelets," *IEEE Trans. Signal Processing*, pp. 666-680, vol. 43, No. 3, March 1995.

[Gra75] A. H. Gray, and J. D. Markel, "A normalized digital filter structure," *IEEE Trans. Acoust. Speech, Signal Processing,* Vol. ASSP-23, p. 268-277, June 1975.

[Gum78] C. Gumacos, "Weighting coefficients for certain maximally flat nonrecursive digital filters," *IEEE Trans. Circuits and Systems,* vol. 25, no. 4, pp. 234-235, Apr. 1978.

[Hei89] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms," SIAM review, vol. 31, pp. 628-666, Dec. 1989.

[Her93] C. Herley, and M. Vetterli, "Wavelets and recursive filter banks," *IEEE Trans. on Signal Processing,* vol. 41, no. 8, pp. 2536-56, Aug. 1993.

[Her93a] C. Herley, J. Kovacevic, K. Ramachandran and M. Vetterli, "Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithm," *IEEE Trans. Signal Processing,* pp. 3341-3359, Vol. 41, no. 12, Dec. 1993.

[Her95] C. Herley, "Boundary filters for finite length signals and time-varying filter banks," *IEEE Trans. Circuits and Systems,* vol. 42, pp. 102-114, Feb. 1995.

[Her71] O. Herrmann, "On the approximation problem in nonrecursive digital filter design," *IEEE Trans. Circuit Theory,* vol. 18, pp. 411-413, May 1971.

[Hoa89] P. -H. Hoang, and P. P. Vaidyanathan, "Nonuniform multirate filter banks: theory and design," *in Proc. IEEE Int. Symp. on Circuits and Systems,* pp. 371-374, Portland, Oregon, May 1989.

[Hor85] R. A. Horn, and C. R. Johnson, *Matrix analysis,* Cambridge University Press, 1985.

[Jay84] N. S. Jayant, and P. Noll, *Digital coding of waveforms*, Englewood Cliffs, NJ: Prentice Hall, 1984.

[Joh80] J. D. Johnston, "A filter family designed for the use in quadrature mirror filter banks," *in Proc. Int. Conf. Acoust. Speech, Signal Processing,* pp. 291-294, April 1980.

[Kai80] T. Kailath, *Linear systems,* Englewood Cliffs, NJ: Prentice Hall, 1980.

[Kar90] G. Karlsson, and M. Vetterli, "Theory of two-dimensional multirate filter banks," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. 38, pp. 925-937, June 1990.

[Kim91] C. W. Kim, and R. Ansari, " FIR/IIR exact reconstruction filter banks with applications to subband coding of images," *in Proc. of Midwest CAS Symp.,* Monterey, CA, May 1991.

[Kim91a] C. W. Kim, and R. Ansari, "Subband decomposition procedure for quincunx sampling grids," *in Proc. of SPIE-Visual Comm. and Image Proc.,* Boston, MA, Nov 1991.

[Kiy92] H. Kiya, M. Yae, and M. Iwahashi, "A linear phase two-channel filter bank allowing perfect reconstruction," *in Proc. IEEE Int. Symp. Circuits and Systems,* pp. 951-954, San Diego, 1992.

[Kov92] J. Kovacevic, and M. Vetterli, "Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for R^n," *IEEE Trans. Information Theory,* vol. 38, pp. 533-555, Mar. 1992.

[Lan87] S. Lang, *Calculus of several variables*, New York NY: Springer-Verlag, 1987.

[Lin94] Y. Lin, and P. P. Vaidyanathan, "Application of DFT filter banks and cosine modulated filter banks in filtering," *in Proc. IEEE Asia-Pacific Conference on Circuits and Systems,* Dec. 1994.

[Lin95a] Y. Lin, and P. P. Vaidyanathan, "Theory and design of two-parallelogram filter banks," *Preprint,* Caltech, Nov. 1995.

[Lin95b] Y. Lin, and P. P. Vaidyanathan, "The class of four-parallelogram filter banks," *Preprint*, Caltech, Nov. 1995.

[Lin96] Y. Lin, and P. P. Vaidyanathan, "Theory and design of two-dimensional filter banks: a review," *Multidimensional Systems and Signal Processing*, Academic Press, 1996.

[Mal89] S. Mallet, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Analysis, and Machine Intell.*, vol. 11, pp. 674-693, July 1989.

[Mal92] H. S. Malvar, *Signal processing with lapped orthogonal transforms,* Artech House, 1992.

[McC79] J. H. McClellan, and C. M. Rader, *Number theory in digital signal processing,* Englewood Cliffs, NJ: Prentice Hall, 1979.

[McE95] R. J. McEliece, "The algebraic theory of convolutional codes," in *Handbook of Coding Theory,* 1995.

[Min85] F. Mintzer, "Filters for distortion-free two-band multirate filter banks," *IEEE Trans. on Acoust., Speech and Signal Processing,* vol. ASSP-33, pp. 626-630, June 1985.

[Mit78] S. K. Mitra, and R. Gnanasekaran, "Block implementation of recursive digital filters: new structures and properties," *IEEE Trans. on Circuits and Systems,* vol. CAS-25, pp. 200-207, April 1978.

[Mit92] S. K. Mitra, C. D. Creusere, and H. Babic, "A novel implementation of perfect reconstruction QMF banks using IIR filters for infinite length signals," *in Proc. IEEE Int. Symp. on Circuits and Systems,* pp. 2312-15, San Diego, May 1992.

[Mit93] S. K. Mitra, and J. F. Kaiser, *Handbook for digital signal Processing,* John Wiley and Sons, 1993.

[Nay92] K. Nayebi, M. J. T. Smith and T. P. Barnwell, "Analysis-synthesis systems based on time-varying filter banks," *in Proc. Int. Conf. Acoust. Speech, Signal Processing,* pp. 617-620, San Francisco, March 1992.

[Neu84] Y. Neuvo, C. -Y. Dong, and S. K. Mitra, "Interpolated finite impulse response filters," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. ASSP-32, pp. 563-570, June 1984.

[Ngu89] T. Q. Nguyen, and P. P. Vaidyanathan, "Two-channel perfect reconstruction FIR QMF Structures which yield linear phase analysis and synthesis filters," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. ASSP-37, pp. 676-690, May 1989.

[Ngu94] T. Q. Nguyen, T. I. Laakso, and R. D. Koilpillai, "Eigenfilter approach for the design of allpass filters approximating given phase response," *IEEE Trans. Signal Processing,* pp. 2257-2263, Sep. 94.

[Pho93] S.-M. Phoong, and P. P. Vaidyanathan, "The bi-orthonormal filter-bank convolvers, and applications in low sensitivity FIR filter structures," *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing,* pp. III 165-168, Minneapolis, April 1993.

[Pho93a] S.-M. Phoong, and P. P. Vaidyanathan, "Robust convolution using data-compression schemes," *in Proc. of Asilomar Conf. on Signals, Systems and Comp.,* pp. 1499-1503, Pacific Grove, CA, 1993.

[Pho94] S.-M. Phoong, and P. P. Vaidyanathan, "Two-channel 1D and 2D biorthonormal filter banks with causal stable IIR and linear phase FIR filters," *in Proc. IEEE Int. Symp. on Circuits and Systems,* London, England, May 1994.

[Pho94a] S.-M. Phoong, and P. P. Vaidyanathan, "Robust M-channel biorthogonal filter banks," *in Proc. Sixth IEEE Digital Signal Processing Workshop,* Yosemite, CA, Oct. 1994.

[Pho95] S.-M. Phoong, and P. P. Vaidyanathan, "One- and two-level filter-bank convolvers," *IEEE Trans. Signal Processing,* pp. 116-133, Jan. 1995.

[Pho95a] S.-M. Phoong, C. W. Kim, P. P. Vaidyanathan and R. Ansari, "A new class of two-channel biorthogonal filter banks and wavelet bases," *IEEE Trans. Signal Processing,* pp. 649-665, March 1995.

[Pho95b] S.-M. Phoong, and P. P. Vaidyanathan, "Time-varying filters and filter banks: Some basic principles," *Preprint,* Caltech, March 1995 (Submitted to IEEE Trans. Signal Processing).

[Pho95c] S.-M. Phoong, and P. P. Vaidyanathan, "Factorizability of lossless time-varying filter banks," *Preprint,* Caltech, March 1995 (Submitted to IEEE Trans. Signal Processing).

[Pho95d] S.-M. Phoong and P. P. Vaidyanathan, "Efficient recursive computation of 1D and 2D-quincunx IIR wavelets," *in Proc. IEEE Int. Symp. Circuits and Systems,* Seattle, WA, April 1995.

[Pho95e] S.-M. Phoong, and P. P. Vaidyanathan, "Time-varying lapped orthogonal transforms," *in Proc. Int. Conf. Digital Signal Processing,* Limassol, Cyprus, June 1995.

[Pho95f] S.-M. Phoong, and P. P. Vaidyanathan, "Paraunitary filter banks over finite fields," *Preprint,* Caltech, Sep. 1995 (Submitted to Trans. Signal Processing).

[Pho95g] S.-M. Phoong, and P. P. Vaidyanathan, "On the study of lossless time-varying filter banks," *in Proc. Asilomar Conf. Signals Systems and Comput.,* Pacific Grove, CA, Oct. 1995.

[Pho96] S.-M. Phoong, and P. P. Vaidyanathan, "A polyphase approach to time-varying filter banks," *in Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing,* Atlanta, May 1996.

[Pho96a] S.-M. Phoong, and P. P. Vaidyanathan, "New results on paraunitary filter banks over finite fields," *in Proc. IEEE Int. Symp. Circuits and Systems,* Atlanta, May 1996.

[Pra92] J. S. Prater, and C. M. Loeffler, "Analysis and design of periodically time-varying IIR filters, with applications to transmultiplexing," *IEEE Trans. Signal Processing,* pp. 2715-2725, vol. 40, No. 11, Nov. 1992.

[Ram84] T. A. Ramstad, "Analysis/synthesis filter banks with critical sampling," *in Proc. Int. Conf. on Digital Signal Processing,* Florence, Sep. 1984.

[Ram88] T. A. Ramstad, "IIR filter bank for subband coding of images," *in Proc. IEEE Int. Symp. Circuits and Systems,* pp. 827-830, Espoo, Finland, 1988.

[Ren87] M. Renfors, and T. Saramaki, "Recursive Nth-band digital filters–part I: Design and properties," *IEEE Trans. Circuits and Systems,* vol. 34, no. 1, pp. 24-39, Jan. 1987.

[Sha92] I. A. Shah, and A. A. C. M. Kalker, "Generalized theory of multidimensional M-band filter bank design," *Proc. of the Sixth EUSIPCO,* pp. 969-972, Aug. 1992.

[Smi87] M. J. T. Smith, and T. P. Barnwell, "A new filter-bank theory for time-frequency representation," *IEEE Trans. on Acoustics, Speech and Signal Processing,* pp. 314-327, March 1987.

[Smi95] M. J. T. Smith, and W. C.-L. Chung, "Recursive time-varying filter banks for subband image coding," *IEEE Trans. Image Processing,* vol. 4, pp. 885-895, Jul. 1995.

[Sod94] I. Sodagar, K. Nayebi, and T. P. Barnwell, "Time-varying filter banks and wavelets," *IEEE Trans. Signal Processing,* pp. 3293-3305, Vol. 42, no. 11, Nov. 1994.

[Som93] A. K. Soman, and P. P. Vaidyanathan, "On orthonormal wavelets and paraunitary filter banks," *IEEE Trans. on Signal Processing,* vol. 41, pp. 1170-83, March 1993.

[Som93a] A. Soman, and P. P. Vaidyanathan, "Coding gain in paraunitary analysis/synthesis systems," *IEEE Trans. on Signal Processing,* vol. 41, pp. 1824-35, May 1993.

[Som93b] A. K. Soman, T. Q. Nguyen, and P. P. Vaidyanathan, "Linear phase paraunitary filter banks: Theory, factorizations and designs," *IEEE Trans. Signal Processing,* vol. 41, pp. 3480-3496, Dec. 1993.

[Som94] A. K. Soman, and P. P. Vaidyanathan, "Generalized polyphase representation in multirate signal processing," *IEEE Trans. on Circuits and Systems,* pp. 627-630, Sep. 1994.

[Ste91] A. Steffen, *Digital pulse compression using multirate filter banks,* Hartung-Gorre Verlag, 1991.

[Str96] G. Strang, and T. Q. Troung, *Wavelets and filter banks,* Wellesley-Cambridge Press, 1996.

[Swam86] K. Swaminathan, and P. P. Vaidyanathan, "Theory and design of uniform DFT, parallel, quadrature mirror filter banks," *IEEE Trans. Circuits and Systems,* pp. 1170-1191, Dec. 1986.

[Swan95] M. D. Swanson, and A. H. Tewfik, "A binary wavelet decomposition of binary images," *Preprint,* Feb. 1995.

[Tay93] B. H. Tay, and N. G. Kingsbury, "Flexible design of multidimensional perfect reconstruction FIR 2-band filters using transformations of variables," *IEEE Trans. on Image Processing,* vol. 2, no. 4, pp. 466-480, Oct. 1993.

[Tek95] A. M. Tekalp, *Digital video processing,* Englewood Cliffs, NJ: Prentice Hall, 1995.

[Vai87] P. P. Vaidyanathan, and T. Q. Nguyen, "A "TRICK" for the design of FIR half-band filters," *IEEE Trans. Circuits and Systems,* vol. 34, no. 3, pp. 297-300, March 1987.

[Vai87a] P. P. Vaidyanathan, and V. Liu, "An improved sufficient condition for absence of limit cycles in digital filters," *IEEE Trans Circuits and Systems,* CAS-34, pp. 319-332, March 1987.

[Vai87b] P. P. Vaidyanathan, P. Regalia, and S. K. Mitra, "Design of doubly complementary IIR filters using a single complex allpass filter, with multirate applications," *IEEE Trans. on Circuits and Systems,* vol. 34, pp. 378-389, April 1987.

[Vai88] P. P. Vaidyanathan, and P. -H. Hoang, "Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks," *IEEE Trans. on Acoustics, Speech and Signal Processing* pp. 81-94, Jan. 1988.

[Vai88a] P. P. Vaidyanathan, and S. K. Mitra, "Polyphase networks, block digital filtering, LPTV systems, and alias-free QMF banks: a unified approach based on pseudocirculants," *IEEE Trans. on Acoustics, Speech and Signal Processing,* vol. ASSP-36, pp. 381-391, March 1988.

[Vai90] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proc. of the IEEE,* vol 78, pp. 56-93, Jan. 1990.

[Vai90a] P. P. Vaidyanathan, "Unitary and paraunitary systems in finite fields," *in Proc. IEEE Int. Symp. on Circuits and Systems,* pp. 1189-1192, New Orleans, 1990.

[Vai93] P. P. Vaidyanathan, *Multirate systems and filter banks,* Englewood Cliffs, NJ: Prentice Hall, 1993.

[Vai93a] P. P. Vaidyanathan, "Orthonormal and biorthonormal filter-banks as convolvers, and convolutional coding gain," *IEEE Trans. on Signal Processing,* vol. 41, pp. 2210-2231, June 1993.

[Vai95] P. P. Vaidyanathan, and I. Djokovic, *Wavelet transforms, in Circuits and Filters Handbook,* ed., W. K. Chen, CRC Press, 1995.

[Vai95a] P. P. Vaidyanathan, and S.-M. Phoong, "Reconstruction of sequences from nonuniform sample," *in Proc. IEEE Int. Symp. Circuits and Systems,* Seattle, WA, April 1995.

[Vai95b] P. P. Vaidyanathan, and S.-M. Phoong, "Discrete-time signal which can be recovered from samples," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing,* Detroit, May 1995.

[Vai95c] P. P. Vaidyanathan, and T. Chen, "Role of anticausal inverses in multirate filter-banks, Part I: System-theoretic fundamentals," *IEEE Trans. Signal Processing,* pp. 1090-1103, May 1995.

[Vai95d] P. P. Vaidyanathan, and T. Chen, Role of anticausal inverses in multirate filter-banks, Part II: The FIR case, factorizations, and biorthonormal lapped transforms (BOLT)," *IEEE Trans. on Signal Processing,* vol. 43, no. 5, May 1995.

[Vet86] M. Vetterli, "Perfect transmultiplexers," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing,* pp. 2567-2570, Tokyo, Japan, April 1986.

[Vet86a] M. Vetterli, "Filter banks allowing for perfect reconstruction," *Signal Processing,* pp. 219-244, April 1986.

[Vet88] M. Vetterli,"Running FIR and IIR filtering using multirate filter banks,"*IEEE Trans. on Acoustics, Speech and Signal Processing,* pp. 730-738, Jan. 1988.

[Vet92] M. Vetterli, and C. Herley,"Wavelets and filter banks," *IEEE Trans. on Signal Processing,* Vol. 40, no. 9, pp. 2209-2232, Sep. 1992.

[Vet95] M. Vetterli, and J. Kovacevic, *Wavelets and subband coding,* Prentice Hall, 1995.

[Vis88] E. Viscito, and J. Allebach, " Design of perfect reconstruction multidimensional filter banks using cascaded Smith form matrices," *in Proc. IEEE Int. Symp. on Circuits and Systems,* Espoo, Finland, pp. 831-834, June 1988.

[Wal92] G. G. Walter, "A sampling theorem for wavelet subspaces," *IEEE Trans. Information Theory,* vol. 38, pp. 881-884, 1992.

[Woo91] J. W. Woods, *Subband coding of images,* Kluwer Academic Publishers, Inc., 1991.

[Xia93] X.-G. Xia, and Z. Zhang, "On sampling theorem, wavelets, and wavelet transforms," *IEEE Trans. on Signal Processing,* vol. 41, pp. 3524–3535, Dec. 1993.

[Xia95] X.-G. Xia, "Filterbank approach for error correction codes with applications in partial response channels," *Preprint,* 1995.

[Xia95a] X.-G. Xia, "Nonmaximally decimated multirate filterbanks in partial response channels with perfect reconstruction," *Preprint,* Aug. 1995.

[You80] R. M. Young, *An introduction to nonharmonic Fourier series,* Academic Press, Inc., 1980.