One and Two-dimensional Digital Multirate Systems with Applications in Sub-sampling and Bandlimited Signal Reconstruction

Thesis by

Vincent Cheng-Teh Liu

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1990

(submitted June 13, 1989)

Acknowledgement

Foremost, I wish to thank Professor P. P. Vaidyanathan for his continual input and encouragement without which this work would not have been possible. I also wish to thank my colleagues, Phuong-Quan Hoang, Zinnur Doganata, David Koilpillai, Truong Nguyen and Vinay Sathe, for their assistance and interaction with me which contributed significantly to the work in this thesis. A speical word of thanks goes to members of my examination committee, Professor Abu-Mostafa, Professor McEliece and Professor Psaltis, for their interest and criticism of this work. The National Science Foundation provided most of the financial resources for the work undertaken here. I would like to thank my parents for their encouragement and understanding. Most importantly, my deepest gratitude goes to my wife, Carina, whose love and support made everything brighter.

Abstract

This thesis deals with the two-dimensional (2D) multirate quadrature mirror filter (QMF) bank and new applications of 1D and 2D multirate filter bank concepts to the periodic nonuniform sampling and reconstruction of bandlimited signals. The potential use of multirate filter banks in the statistically optimal estimation of signals in the presence of wide-sense cyclostationary noise is also examined. The two-dimensional QMF bank is free from aliasing if and only if a certain polyphase matrix product related to the filter bank possesses the 2D pseudo-circulant property. A 2D FIR filter bank can be designed with the perfect reconstruction property if the polyphase matrix of its analysis filter bank is constrained to be a 2D lossless matrix. A design example is included. The losslessness constraint is satisfied by imposing a cascaded structure of first-degree lossless sections on the polyphase matrix. A limited factroization theorem is derived for 2D FIR lossless systems where the order in one of the two dimensions is limited to unity. In the area of nonuniform sampling of multiband bandlimited signals, the filter bank approach is utilized to derive a computationally efficient method for reconstructing bandlimited signals. The above scheme can also be viewed as a mean of compressing and reconstructing an oversampled bandlimited signal. It is shown that such a scheme has lower computational complexity than traditional methods of sampling rate alteration. The results can be extended to nonuniform sampling in two-dimensions using integer lattices. A further application of the multirate filter bank is in signal estimation in the presence of cyclostationary noise. The necessary and sufficient condition for the filter bank to preserve the wide-sense stationarity of the input is derived. Several applications where cyclostationary noise is present are indicated, and through

the use of simulations the performance of the optimal filter bank can be compared with the conventional scalar optimal filter. The roundoff noise in orthogonal matrix building blocks is analyzed, since these building blocks are commonly present in filter bank implementations.

Table of Contents

List of Figures

List of Tables

Chapter I. Introduction

Multirate signal processing has received the attention of many researchers in the past [CR83]. It basically involves the processing of digitally sampled signals at various sampling rates. The change in sampling-rate is accomplished through devices – that are commonly called decimators and interpolators. The symbols for a decimator and an interpolator are depicted in Fig. 1.1. One widely studied multirate signal processing system is the quadrature mirror filter (QMF) bank. It has applications in sub-band coding of speech and images [CR83] [WO86], spectral analysis [WA86], voice privacy systems [CO87] and transmultiplexers [SC81]. The purpose of this thesis is to study the various engineering issues associated with the two-dimensional QMF bank, and the application of the multirate filter bank as an efficient method of recovering bandlimited signals from their periodically nonuniform samples. We shall also look at the design of a multirate system as a statistically optimal filter for estimating cyclostationary random processes.

1.1 Preliminaries and Mathematical Notations

A discrete-time signal, $x(n)$, is a sequence of real or complex numbers [OP75]. The sequence can be doubly infinitely long. The z-transform of the signal $x(n)$ is defined as $X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$. The Fourier transform of $x(n)$ (if it exists) is given by $X(z)$ evaluated on the unit circle $z = e^{j\omega}$. We will write the Fourier transform of $x(n)$ as $X(e^{j\omega})$. Notice that $X(e^{j\omega})$ has period 2π in the variable ω. Following convention, when a transform $X(e^{j\omega})$ is plotted as a function of frequency, the frequency scale used is that of normalized frequency f which is defined to be $f = \frac{\omega}{2\pi}$.

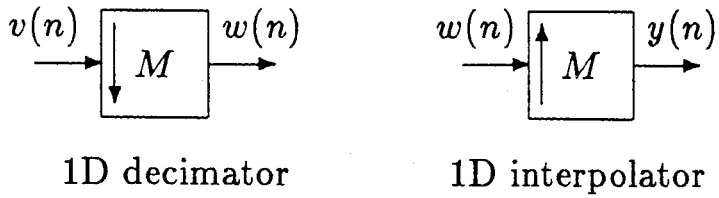

Fig. 1.1. Decimators and interpolators for one-dimensional signals.

a linear time-invariant digital filter, its input $x(n)$ and its output $y(n)$ are related by a convolution $y(n) = \sum_{m=-\infty}^{\infty} h(m)x(n-m)$. The sequence $h(m)$ is the impulse response of the filter. The filter can also be described by its transfer function $H(z)$ which is the z-transform of $h(m)$, and the filter output is given by $Y(z) = H(z)X(z)$. A finite impulse response (FIR) filter is one for which the impulse response is finitely long. A causal filter is one for which the present output depends only upon the present and past inputs, therefore its impulse response $h(n) = 0$ for all $n < 0$. Combining the two properties, we see that the transfer function of a causal FIR filter is a polynomial in z^{-1} (i.e., it contains negative powers of z). For a causal FIR system $H(z)$, the highest power in z^{-1} will be called the order of the system, while the degree of the system refers to the McMillan degree (which is defined as the smallest number of delay elements necessary in realizing a system. Mathematically, a delay element is represented by z^{-1}, so if $Y(z) = z^{-1}X(z)$ then in discrete time $y(n) = x(n-1)$.) In a single-input/single-output (SISO) system, the order and the degree of a system are the same. However, in a multiple-input/multiple-output (MIMO) system (in which case the transfer function is a matrix, $\mathbf{H}(z)$), the order of the system will in general be different from its McMillan degree.

In the thesis, matrices are denoted by upper case bold letters and lower case

bold letters are column vectors. The (i,j)th entry of a matrix $\mathbf{A}$ is denoted by $[\mathbf{A}]_{i,j}$ and the kth entry for the column vector $\mathbf{a}$ is $[\mathbf{a}]_k$. The notation $\mathbf{A}^T$ means the transposition of $\mathbf{A}$, while $\mathbf{A}^\dagger$ is the transposed conjugate. $\mathbf{I}_k$ is the identity matrix of dimension $k \times k$. If there is closed-form expressions for the individual entries of $\mathbf{A}$, then the matrix may be denoted as $(a_{i,k})$ where $a_{i,k}$ is the formula for the (i,k)th entry. A diagonal matrix is written as

$$\text{diag.} \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{M-1} \end{pmatrix} = \begin{pmatrix} d_0 & 0 & \dots & \\ 0 & d_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_{M-1} \end{pmatrix}. \tag{1.1}$$

The letter $\mathcal{W}$ stands for $e^{-j2\pi/M}$. The matrix $\mathbf{W}$ is an $M \times M$ DFT matrix, i.e., $\mathbf{W} = (\mathcal{W}^{ik})$. The columns of $\mathbf{W}$ are denoted as $\mathbf{u}_k$. For any integer k, $((k))$ denotes its residue modulo M.

The Kronecker product of two matrices is defined as followed

$$\begin{pmatrix} a_{0,0} & a_{0,1} & \dots & a_{0,L-1} \\ a_{1,0} & a_{1,1} & \dots & a_{1,L-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K-1,0} & a_{K-1,1} & \dots & a_{K-1,L-1} \end{pmatrix} \otimes \mathbf{B} = \begin{pmatrix} a_{0,0}\mathbf{B} & a_{0,1}\mathbf{B} & \dots & a_{0,L-1}\mathbf{B} \\ a_{1,0}\mathbf{B} & a_{1,1}\mathbf{B} & \dots & a_{1,L-1}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K-1,0}\mathbf{B} & a_{K-1,1}\mathbf{B} & \dots & a_{K-1,L-1}\mathbf{B} \end{pmatrix}. \tag{1.2}$$

For a matrix or a scalar that is a function z, such as $\mathbf{H}(z)$, the notation $\mathbf{H}_*(z)$ means conjugating the coefficients of the function. For example, if $H(z) = a + bz^{-1}$ then $H_*(z) = a^* + b^* z^{-1}$. The tilde notation $\tilde{\mathbf{H}}(z)$ is used to denote $\mathbf{H}_*^T(z^{-1})$. On the unit circle, $z = e^{j\omega}$, one can verify that $\tilde{\mathbf{H}}(e^{j\omega}) = \mathbf{H}^\dagger(e^{j\omega})$.

1.2 Basic Building Blocks in Multirate Signal Processing

The processing of discrete-time signals having different sampling rates is known as multirate signal processing [CR83]. The conversion of a signal at one sampling

rate to another that is at $\frac{1}{M}$ of the original rate (where M is an integer) can be accomplished by the use of a 'decimator', shown in Fig. 1.1 as the box with the down-going arrow. The output $w(n)$ is related to the input $v(n)$ by the formula

$$w(n) = v(Mn). \tag{1.3}$$

Thus, only one out of every M input samples are being kept by the decimator. In this way, the sampling rate is reduced by $\frac{1}{M}$. Such a straight forward method of sampling rate alteration causes aliasing. Namely, the Fourier transform of $v(n)$, (denoted as $V(e^{j\omega})$ which is periodic with period 2π), becomes stretched and the adjacent terms start to overlap with each other. The effect of decimation in the frequency domain can be described as

$$W(e^{j\omega}) = \frac{1}{M}\sum_{m=0}^{M-1} V(e^{j\omega/M}\mathcal{W}^m) \tag{1.4}$$

where $\mathcal{W} = e^{-j\frac{2\pi}{M}}$. For the purpose of increasing the sampling rate of a signal, the device used is an 'interpolator', shown in Fig. 1.1 with the up-going arrow. The output $y(n)$ has a sampling rate that is M times that of $w(n)$, the input signal. The interpolator accomplishes the change in rate by simply filling in $M-1$ zero's in between adjacent samples of $w(n)$. Mathematically, this is described by

$$y(n) = \begin{cases} w(n/M) & \text{when } n \text{ is divisible by } M; \\ 0 & \text{otherwise.} \end{cases} \tag{1.5}$$

The effects on the transforms of the signals is given by

$$Y(e^{j\omega}) = W(e^{j\omega M}). \tag{1.6}$$

1.3 The QMF Bank and Related Issues

One commonly studied multirate system is the quadrature mirror filter (QMF) bank [SM87a,VA87a], shown in Fig. 1.2. The filters, $H_k(z)$, are called the analysis

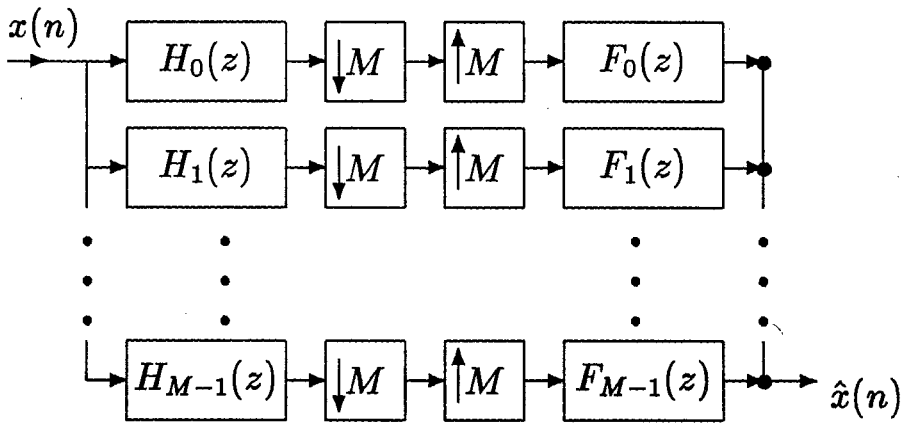

Fig. 1.2. A quadrature mirror filter (QMF) bank.

filters for they split the incoming signal $x(n)$ into M signals for the various purposes of spectral analysis [WA86], sub-band coding [CR83], etc. Traditionally the filters $H_k(z)$ are bandpass filters with consecutive (disjoint) passbands and overlapping transition bands so that the entire frequency axis is covered. In effect, the bank of analysis filters splits the input signal $x(n)$ into M sub-band signals in such a way that each sub-band signal contains a portion of the frequency information of the original signal. The sampling rate for each sub-band signal is reduced by a factor of M therefore the total data rate (in terms of the number of samples per second) for the M sub-bands is the same as the data rate of the original signal. There is an advantage in quantizing and encoding the sub-band signals instead of the original signal. As an example, in speech coding applications, the speech signal often has an uneven energy distribution over its frequency range. A coding gain is realized if some sub-bands can be encoded using fewer number of bit. The bit allocation for each channel may be decided experimentally or it may be chosen so that a certain overall distortion measure is minimized [JAY84] [WE88].

The original signal $x(n)$ is reconstructed using interpolators and the bank of

synthesis filters $F_k(z)$'s. The synthesis filters are chosen to reduce the effects of aliasing (caused by decimators) and amplitude and phase distortions (caused by nonideal nature of the filters $H_k(z)$ and $F_k(z)$). Using (1.4) and (1.6), one can show that the reconstructed signal is given by

$$\hat{X}(e^{j\omega}) = \frac{1}{M}\sum_{k=0}^{M-1} F_k(e^{j\omega}) \sum_{m=0}^{M-1} H_k(e^{j\omega}\mathcal{W}^m) X(e^{j\omega}\mathcal{W}). \tag{1.7}$$

The filter bank should be designed so that $\hat{X}(e^{j\omega})$ approximates $X(e^{j\omega})$ as best as possible. As seen from (1.7), $\hat{X}(e^{j\omega})$ is a linear combination of the original spectrum $X(e^{j\omega})$ and its frequency shifted versions, $X(e^{j\omega}\mathcal{W})$. The frequency shifted terms, $X(e^{j\omega}\mathcal{W}^m)$ with $m \neq 0$, are called the aliasing components or simply the alias terms. So one of the requirements on the filter bank design is that the analysis and synthesis filters should be chosen so that the alias terms are cancelled. This means the two set of filters should satisfy the equations

$$\sum_{k=0}^{M-1} F_k(e^{j\omega}) H_k(e^{j\omega}\mathcal{W}^m) = 0 \quad \text{for } m \neq 0. \tag{1.8}$$

In the absence of aliasing, the output signal is related to the input by $\hat{X}(e^{j\omega}) = \frac{1}{M} X(e^{j\omega}) \sum_{k=0}^{M-1} F_k(e^{j\omega}) H_k(e^{j\omega})$. And we see there is still an overall distortion on the signal. This can be written as $\hat{X}(e^{j\omega}) = T(e^{j\omega}) X(e^{j\omega})$ where $T(e^{j\omega})$ is called the transfer function of the system in the absence of aliasing. One may make a distinction between two types of distortions. One is caused by the amplitude of $T(e^{j\omega})$ which is a function of frequency ω. Thus the spectrum of $\hat{X}(e^{j\omega})$ will in general be different from the spectrum of $X(e^{j\omega})$. The other type of distortion comes from the complex phase of the transfer function. If one can design the filters to satisfy (1.8) and $T(z) = c \cdot z^{-N}$, then the reconstructed signal $\hat{x}(n) = cx(n-N)$ and so the original signal is recovered exactly except for a delay of N units in time and a scale factor. Under these circumstances, the QMF bank is said to have the

perfect reconstruction property.

Following the notations in [VA87a], let us denote the analysis filter bank by

$$\mathbf{h}(z) = \begin{pmatrix} H_0(z) \\ H_1(z) \\ \vdots \\ H_{M-1}(z) \end{pmatrix}. \tag{1.9}$$

The synthesis filter bank can be represented similarly by $\mathbf{f}^T(z) = (\ F_0(z)\ F_1(z)\ \ldots\ F_{M-1}(z)\ )^T$. Any scalar digital transfer function $P(z)$ can be written in terms of its polyphase representations [BE76] either as

$$P(z) = \sum_{k=0}^{M-1} z^{-k} P_k(z^M) \tag{1.10}$$

which is called the Type I polyphase representation, or as

$$P(z) = \sum_{k=0}^{M-1} z^{-M+k} Q_k(z^M) \tag{1.11}$$

which is referred to as Type II polyphase. The sub-filters $P_k(z)$ and $Q_k(z)$ are known respectively as the Type I and Type II polyphase components of $P(z)$. For a filter bank $\mathbf{h}(z)$, it also has a polyphase representation as

$$\mathbf{h}(z) = \mathbf{E}(z^M) \begin{pmatrix} 1 \\ z^{-1} \\ \vdots \\ z^{-M+1} \end{pmatrix}, \tag{1.12}$$

where $\mathbf{E}(z)$ is an $M \times M$ matrix, called the polyphase component matrix [VA87a] of $\mathbf{h}(z)$. From (1.8), we see that $H_k(z) = \sum_{l=0}^{M-1} [\mathbf{E}(z^M)]_{k,l} z^{-l}$. Thus the kth column of $\mathbf{E}(z)$ contains the Type I polyphase components of $H_k(z)$. The synthesis filter bank can be written in terms of a polyphase component matrix as

$$\mathbf{f}^T(z) = (\, z^{-M+1} \quad z^{-M+2} \quad \ldots \quad z^{-1} \quad 1\,)\, \mathbf{R}(z^M). \tag{1.13}$$

Using (1.12) and (1.13), the QMF bank can be re-drawn as in Fig. 1.3. Since $\mathbf{E}(z^M)$ is a function of z^M only, it can be interchanged with the decimators on

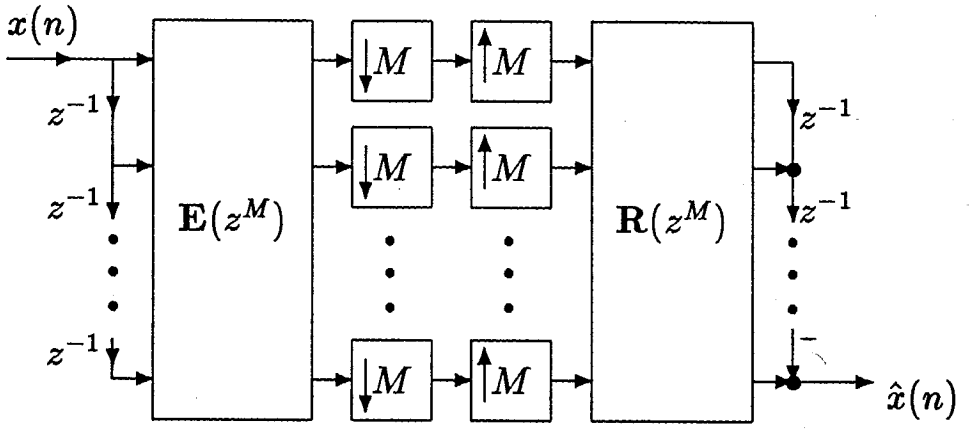

Fig. 1.3. Polyphase representation of the QMF bank.

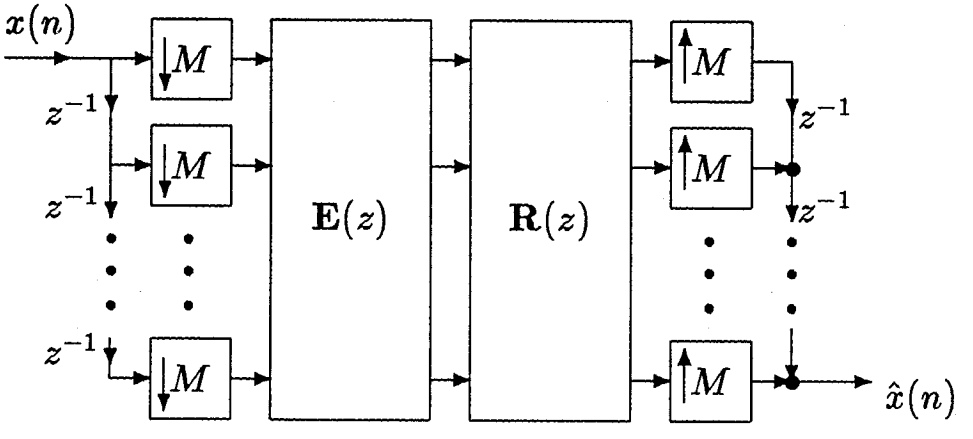

Fig. 1.4. An equivalent representation of the polyphase structure.

its right, provided that $\mathbf{E}(z^M)$ is replaced with $\mathbf{E}(z)$. Similarly, $\mathbf{R}(z^M)$ can be interchanged with the interpolators, and as a result $\mathbf{R}(z^M)$ is replaced by $\mathbf{R}(z)$. This is illustrated in Fig. 1.4. Now by defining a new matrix $\mathbf{P}(z) = \mathbf{R}(z)\mathbf{E}(z)$ (Fig. 1.5), Vaidyanathan et al. have shown that the necessary and sufficient condition for the QMF bank to be free from aliasing is that $\mathbf{P}(z)$ must be a pseudo-circulant matrix [VA87b]. Furthermore, there is no magnitude distortion in the reconstructed signal if and only if the matrix $\mathbf{P}(z)$ is lossless. A matrix $\mathbf{P}(z)$ is defined to be lossless if it is stable and $\mathbf{P}^{-1}(z) = \tilde{\mathbf{P}}(z)$.

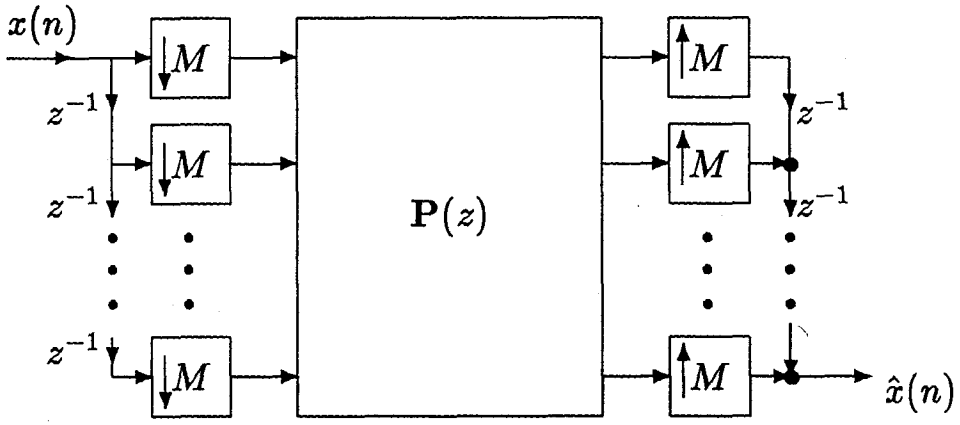

Fig. 1.5. The position of $\mathbf{P}(z)$ within the QMF bank.

In order to obtain perfect reconstruction, (1.8) has to hold exactly, and one way of obtaining this is to make $\mathbf{R}(z) = \mathbf{E}^{-1}(z)$. This is impractical for IIR filters, because if $\mathbf{h}(z)$ represents a stable IIR filter bank then the above choice of $\mathbf{R}(z)$ gives rise to unstable synthesis filters. Even if $\mathbf{h}(z)$ is an FIR filter bank (which means $\mathbf{E}(z)$ is a polynomial matrix in z^{-1}), $\mathbf{E}^{-1}(z)$ will not in general be a polynomial matrix, so for perfect reconstruction $\mathbf{f}(z)$ would be IIR in general and possibly unstable. The only remaining choice is to constrain both $\mathbf{E}(z)$ and $\mathbf{E}^{-1}(z)$ to be FIR, and one way of achieving this is to constrain $\mathbf{E}(z)$ to be an FIR lossless matrix. For in that case, $\mathbf{E}^{-1}(z) = \tilde{\mathbf{E}}(z)$ and the choice of $\mathbf{R}(z) = \tilde{\mathbf{E}}(z)$ guarantees that perfect reconstruction is achieved using FIR filter banks only. The filter bank design problem becomes one of designing $\mathbf{E}(z)$, under the constraint that it must be FIR lossless, so that the analysis filters $H_k(z)$ are good filters. The study of lossless systems in [VA88c] shows that any FIR lossless matrix may be factorized into a cascade of degree-one lossless factors, and a structure for implementing a general FIR lossless system $\mathbf{E}(z)$ is found for which the lossless property is inherent in the structure. This means lossless constraint is automatically satisfied if we implement $\mathbf{E}(z)$ using the

structure in [VA88c], and the design process consists of optimizing the parameters of the structure so that it will give rise to good analysis filters [DO88].

1.4 Outline of Thesis

In the following chapters, we shall look at the multirate filter bank for both one-dimensional (1D) and two-dimensional (2D) signals. First, in Chapter II the theory of [VA87b] is extended to the two-dimensional filter bank. Most of the results in this chapter can also be found in [LI88a]. We will see how the 2D pseudo-circulant matrix plays an important role in the two-dimensional QMF bank. Namely, the necessary and sufficient condition for alias-cancellation is that the polyphase matrix product $\mathbf{P}(z_1, z_2)$ should be 2D pseudo-circulant. Similar to the 1D case, the system is free from magnitude distortion if and only if $\mathbf{P}(z_1, z_2)$ is 2D lossless. In the design of 2D QMF banks having perfect reconstruction property, one desires to impose the lossless constraint on $\mathbf{E}(z_1, z_2)$ for similar reasons as in the 1D case. However, unlike the results in [VA88c], we cannot find a general procedure for factorizing 2D FIR lossless matrices. We will show that a factorization exists for a sub-class of these matrices.

In Chapter III, a new application for the multirate filter bank is presented. A filter bank structure is used to reconstruct a bandlimited signal from its nonuniform samples. (In this thesis, the nonuniform samples shall refer to samples taken on a nonuniform grid which is periodically repeating.) Consider a signal $x(n)$ that is band limited to a certain frequency band, say $-\frac{L\pi}{M} < \omega < \frac{L\pi}{M}$ where $L < M$. This means its Fourier transform $X(e^{j\omega})$ is zero for $-\pi \leq \omega \leq -\frac{L\pi}{M}$ and $\frac{L\pi}{M} \leq \omega \leq \pi$. Here L and M are restricted to be integers. If this is the sampled version of a continuous time signal $x_a(t)$ with sampling frequency F samples/sec., then from the sampling

theorem [SH49] [JE77] we know the sampling rate of $x(n)$ can be reduced to $\frac{L}{M}$ of the original rate and the new samples will still represent the original signal. Thus the data rate can be reduced by $\frac{L}{M}$.

Given the sequence $x(n)$ at the higher sampling rate, there are several ways of reducing its data rate. One method is to convert $x(n)$ back to a continuous-time signal through lowpass filtering and then re-sample it at the lower rate. With the use of multi-rate digital systems, the same can be achieved using interpolators, discrete-time filters and decimators alone [CR83], so the need for an intermediate continuous-time signal is eliminated (Fig. 1.6). In both of these methods, in order to obtain a signal at the lower rate, a non-trivial filtering operation has to be performed. The method which will be presented here has the advantage that the procedure for compressing $x(n)$ is simple. It requires no filtering in order to compress the signal. We imagine that the time axis is divided into intervals of length M each, and the compression scheme keeps only L out of M samples in each interval. This will be called the 'L out of M' compression technique. This method of compression can be viewed as a discrete-time version of the nonuniform sampling theorem [JE77]. Although the compression scheme is very simple, the tradeoff for this simplicity is that the resulting compressed signal is not a uniformly sampled signal, and if we want to convert it back to uniform samples the reconstruction process is somewhat more complicated. The basic ideas of this method have already been treated in [VA88d], and the two-dimensional equivalent can be found in [LI88c].

However, in Chapter III we will show that the above compression scheme can be applied to multiband signals (Sec. 3.2), and the compressed signal can be reconstructed in a manner which is far more efficient than what had been done in

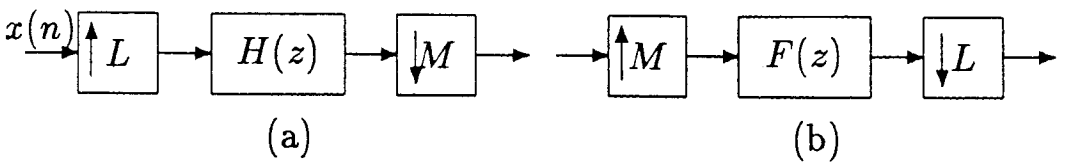

Fig. 1.6. Conventional sampling rate alteration: (a) rate reduction; (b) rate increase.

[VA88d] and [LI88c]. In particular, it will be shown that a polyphase approach can be used to design and implement the reconstruction filters (Sec. 3.3). This approach allows us to derive a structure and design procedure that will work for arbitrary integers L and M. The procedure is based on the design of a single Mth-band filter [MIN82] and its implementation in polyphase form. In contrast, the structures presented in [VA88d] hold only for the speical cases of $M = 3$, $L = 2$ and $M = 4$, $L = 3$. Also the polyphase structure has lower complexity when compared with the previous methods in [VA88d]. This presents a new and efficient way of compressing and reconstructing bandlimited signals. Through the use of design examples (Sec. 3.4), we will show that the signal compression/reconstruction system actually has lower complexity than the conventional means of rational sampling rate alteration [CR83] (see, for example, Fig. 1.6 in this thesis). The algorithm used in Mth-band filter design is presented in Appendix A. The analysis of noise in our compression/reconstruction system is presented in Appendix C. A closed-form expression can be derived for the noise gain. However, the exact value of the noise gain depends upon the subsampling scheme chosen and the frequency bandlimits of the multiband signal.

The above compression scheme is equally applicable to 2D bandlimited signals.

By using general lattice sampling, the bandlimits for the 2D signals are not restricted to rectangular shapes but rather they can be arbitrary parallelograms in the 2D frequency plane (Sec. 3.5). We will also look at the sub-sampling and reconstruction problem in the wider context of bandlimited signal extrapolation. In Sec. 3.6, an energy minimization approach is proposed for extrapolating finite length signals that are close to being bandlimited. This method can be applied to the sub-sampling and reconstruction of bandlimited signals (Sec. 3.7). One may also refer to [LI89a] for the material in Sec. 3.6 and 3.7.

In Chapter IV, random process inputs to the multirate filter bank is considered. A necessary and sufficient condition is derived for which the filter bank will preserve the wide-sense stationarity of the input. Also we will see how a multirate filter bank can be used for the estimation or prediction of signals that are wide-sense cyclostationary. The estimation of cyclostationary processes can be done using block filtering [MA86]. However, due to the effects of blocking the estimator uses data in an asymmetrical manner. By using the filter bank approach in Sec. 4.3, we arrive at a solution that is different from the one obtained through block filtering. In making the estimation, it always uses the data symmetrically. In the limiting case where the filter order is allowed to be infinite, the two methods approach the same theoretical optimal Wiener solution.

Finally, in Chapter V we study the problem of roundoff noise in common filter bank structures. The noise is due to the finite word length computations performed by orthogonal building blocks. We are interested in the cross-correlation of the errors generated at the outputs of orthogonal matrix building blocks. These building blocks are often found in lossless FIR and IIR lattices and in orthogonal digital

filters. The material of this chapter can be found in [LI88b].

Chapter II. Alias-cancellation and Distortion Elimination for Two-dimensional QMF Banks

2.1 The Two-dimensional Filter Bank

The extension of the one-dimensional (1D) quadrature mirror filter (QMF) bank into 2D can be applied to the sub-band coding of images [WO86] and short-space Fourier spectral analysis [WA86]. The 2D QMF bank (Fig. 2.1) has been investigated in [VE84] [WO86] [VA87c], especially with regards to the perfect reconstruction of the input signal. In [VE84] the analysis and synthesis filters are separable filters designed for the 1D QMF bank, and perfect reconstruction is achieved by arranging the filter bank in a tree structure. In [VA87c] a design approach based upon 2D lossless matrices is used which guarantees perfect reconstruction and the resulting filters are not restricted to be separable. However, there has not been a general theory of conditions for alias cancellation and perfect reconstruction in these filter banks. In Sec. 2.2, the necessary and sufficient condition for the filter bank to be free of aliasing is derived. The condition is based upon the pseudo-circulant structure (to be elaborated later) of a polyphase matrix product $\mathbf{P}(z_1, z_2)$. In Sec. 2.3, we will show that for an alias-free QMF bank there is no magnitude distortion if and only if the matrix $\mathbf{P}(z_1, z_2)$ is lossless. For the two-dimensional QMF bank, perfect reconstruction can be achieved if one designs the analysis filter bank such that its polyphase component matrix $\mathbf{E}(z_1, z_2)$ is FIR and lossless, in which case the synthesis filter bank should have its polyphase matrix equal to $\tilde{\mathbf{E}}(z_1, z_2)$ [VA87c]. A factorization theorem is shown to exist for 1D FIR lossless matrices [DO88] and it has been applied to the design of 1D QMF banks. However, we have not been able to find a general factorization theorem that will work for 2D systems.

A factorization exists for a restricted class of 2D lossless systems. In Sec. 2.4, it is shown that a sub-class of 2D FIR lossless matrices can be factorized into degree-one factors. This factorization is presented in the hope that it will motivate one to work out a theorem for factorizing general 2D FIR lossless systems. A 2D QMF bank design example based on lossless FIR polyphase matrices is presented in Sec. 2.5. In Sec. 2.6, block filtering [BU72] [MIT78] [BA80] in 2D is considered and it will be shown that the transfer matrix of a block filter has the 2D pseudo-circulant property. The results presented in these sections can be generalized and extended to multi-dimensional filter banks. For this we need to define what a multi-dimensional pseudo-circulant matrix is. A recursive definition is given in Sec. 2.7.

A 2D QMF bank is shown in Fig. 2.1. The structure contains 2D decimators and interpolators which are rate-varying building blocks. However, the overall output signal $\hat{x}(n_1, n_2)$ has the same sampling rate as the input. Fig. 2.1 can be thought of as a periodic space-variant linear system. For such a system, the transform of the output signal can be written in terms of the input as

$$\hat{X}(z_1, z_2) = \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} X(z_1 W_1^{k_1}, z_2 W_2^{k_2}) A_{k_1,k_2}(z_1, z_2). \tag{2.1}$$

The frequency-shifted versions of $X(z_1, z_2)$ are known as the alias terms. In [VE84] [WO86], the filter banks were designed so as to eliminate the alias terms from $\hat{X}(z_1, z_2)$. However, the design methods presented there is restricted to only separable filters. We will derive a necessary and sufficient condition of alias-cancellation for both separable and non-separable filters. This condition, similar to the one for 1D QMF banks in [VA88b], is based upon the pseudo-circulant property of certain matrices. The elimination of magnitude or phase distortions are also discussed in the context of pseudo-circulant matrices.

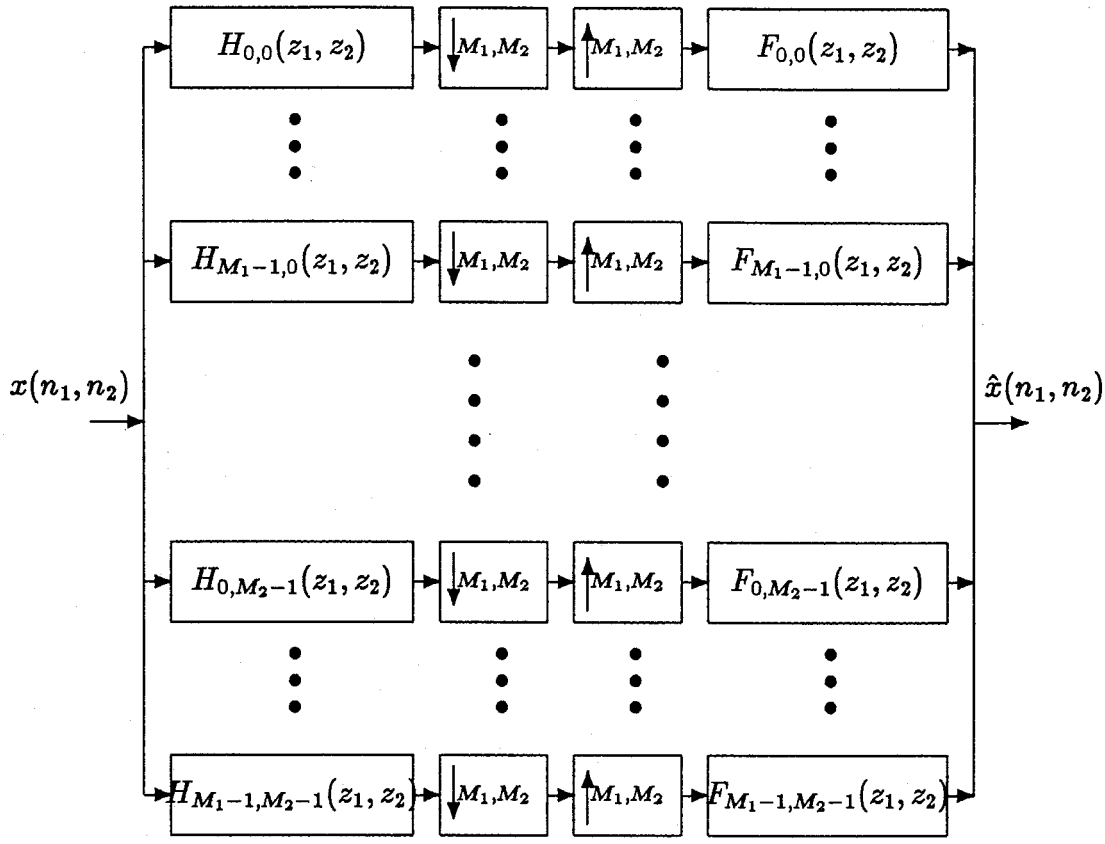

Fig. 2.1. The two-dimensional QMF bank.

A 1D $M \times M$ pseudo-circulant matrix in the variable z_1 has the following format:

$$\begin{pmatrix} c_0 & c_1 & \dots & c_{M-1} \\ z_1^{-1}c_{M-1} & c_0 & \dots & c_{M-2} \\ \vdots & \vdots & \ddots & \\ z_1^{-1}c_1 & z_1^{-1}c_2 & \dots & c_0 \end{pmatrix}. \tag{2.2}$$

The entries c_i can be either scalar constants or scalar functions of several variables z_1, z_2, z_3, etc. The 2D $MN \times MN$ pseudo-circulant matrix takes on the form

$$\begin{pmatrix} \mathbf{P}_0 & \mathbf{P}_1 & \dots & \mathbf{P}_{N-1} \\ z_2^{-1}\mathbf{P}_{N-1} & \mathbf{P}_0 & & \mathbf{P}_{N-2} \\ \vdots & \vdots & \ddots & \\ z_2^{-1}\mathbf{P}_1 & z_2^{-1}\mathbf{P}_2 & \dots & \mathbf{P}_0 \end{pmatrix} \tag{2.3}$$

where each $\mathbf{P}_i$ is a 1D pseudo-circulant matrix as shown in (2.2). Note that for a 1D pseudo-circulant matrix, its entries need not be functions of one variable only.

Similarly, for two dimensions the matrix entries can be functions of two or more variables. Pseudo-circulant matrices of higher dimensions can be defined recursively in terms of the lower dimension ones.

For any 2D digital transfer function $H(z_1, z_2)$, there exists a rectangular polyphase representation for it as follows

$$H(z_1, z_2) = \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} z_1^{-k_1} z_2^{-k_2} H_{k_1,k_2}(z_1^{M_1}, z_2^{M_2}). \tag{2.4}$$

The sub-filters, $H_{k_1,k_2}(z_1, z_2)$, are called the polyphase components of $H(z_1, z_2)$.

2.2 Condition for Alias-cancellation

We shall consider a QMF bank with $M_1 \times M_2$ branches. Fig. 2.2 shows the components of the $(m_1 + m_2 M_1)$th branch, where $0 \leq m_1 < M_1$ and $0 \leq m_2 < M_2$. Let $x(n_1, n_2)$ be the 2D input signal and $\hat{x}(n_1, n_2)$ be the reconstructed signal. $\hat{x}(n_1, n_2)$ is the sum of the output from all the branches. The QMF bank consists of a parallel connection of these $M_1 \times M_2$ branches. The 2D decimator works as follows. Let $u(n_1, n_2)$ be the input and $w(n_1, n_2)$ be the output of the decimator, then $w(n_1, n_2) = u(n_1 M_1, n_2 M_2)$. In the z-transform domain, the effect of this operation is

$$W(z_1, z_2) = \frac{1}{M_1 M_2} \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} U(z_1^{1/M_1} \mathcal{W}_1^{k_1}, z_2^{1/M_2} \mathcal{W}_2^{k_2}). \tag{2.5}$$

Similarly, the 2D interpolation can be described in the z-domain as $W(z_1, z_2)$ $= U(z_1^{M_1}, z_2^{M_2})$.

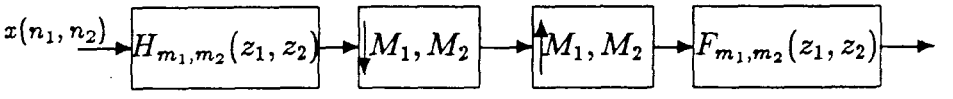

Fig. 2.2. One branch of the 2D QMF bank.

Let the analysis filter bank be denoted by $\mathbf{h}(z_1, z_2)$ and the synthesis filter bank by $\mathbf{f}(z_1, z_2)$, i.e.,

$$\mathbf{h}(z_1, z_2) = \begin{pmatrix} H_{0,0}(z_1, z_2) \\ H_{1,0}(z_1, z_2) \\ \vdots \\ H_{M_1-1,M_2-1}(z_1, z_2) \end{pmatrix} \quad \mathbf{f}(z_1, z_2) = \begin{pmatrix} F_{0,0}(z_1, z_2) \\ F_{1,0}(z_1, z_2) \\ \vdots \\ F_{M_1-1,M_2-1}(z_1, z_2) \end{pmatrix}. \tag{2.6}$$

Define the vectors $\mathbf{e}(z_1, z_2)$ and $\mathbf{r}(z_1, z_2)$ to be

$$\mathbf{e}(z_1, z_2) = \mathbf{e}_2(z_2) \otimes \mathbf{e}_1(z_1) = \begin{pmatrix} 1 \\ z_2^{-1} \\ \vdots \\ z_2^{-M_2+2} \\ z_2^{-M_2+1} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ z_1^{-1} \\ \vdots \\ z_1^{-M_1+2} \\ z_1^{-M_1+1} \end{pmatrix} \quad \text{and} \tag{2.7a}$$

$$\mathbf{r}(z_1, z_2) = \mathbf{r}_2(z_2) \otimes \mathbf{r}_1(z_1) = \begin{pmatrix} z_2^{-M_2+1} \\ z_2^{-M_2+2} \\ \cdots \\ z_2^{-1} \\ 1 \end{pmatrix} \otimes \begin{pmatrix} z_1^{-M_1+1} \\ z_1^{-M_1+2} \\ \cdots \\ z_1^{-1} \\ 1 \end{pmatrix}, \tag{2.7b}$$

then the analysis filter bank can be expressed as $\mathbf{h}(z_1, z_2) = \mathbf{E}(z_1^{M_1}, z_2^{M_2})\mathbf{e}(z_1, z_2)$, and similarly we can write the synthesis bank as $\mathbf{f}^T(z_1, z_2) = \mathbf{r}(z_1, z_2)\mathbf{R}(z_1^{M_1}, z_2^{M_2})$. The matrices $\mathbf{E}(z_1, z_2)$ and $\mathbf{R}(z_1, z_2)$, which were defined in [VA87c], contain the ployphase components of the analysis and synthesis filters respectively. The QMF bank can now be re-drawn as in Fig. 2.3. The building block D represents M_1 decimators and U denotes M_2 interpolators, each having the forms as in Fig. 2.2.

Similar to the 1D case, $\mathbf{E}(z_1^{M_1}, z_2^{M_2})$ and the column of decimators can be inter-

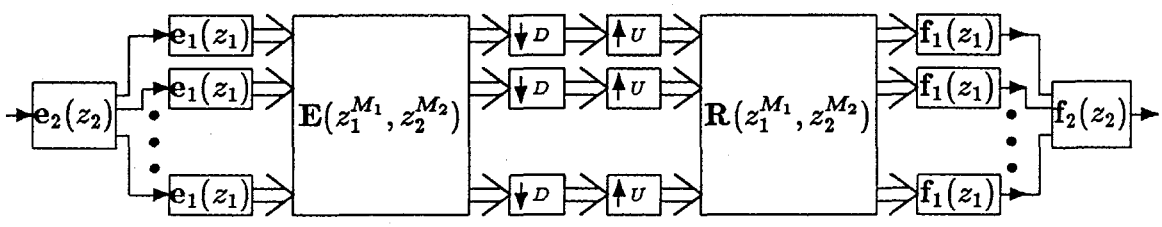

Fig. 2.3. Two-dimensional polyphase representation of the QMF bank

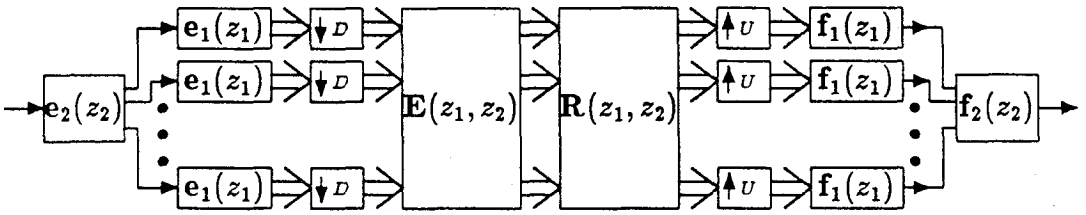

Fig. 2.4. Alternative representation for the polyphase structure.

changed. Performing the same operation on $\mathbf{R}(z_1^{M_1}, z_2^{M_2})$ and the interpolators, we arrive at a representation for the QMF bank shown in Fig. 2.4. Let us define a matrix $\mathbf{P}(z_1, z_2)$ as the product $\mathbf{R}(z_1, z_2)\mathbf{E}(z_1, z_2)$ (Fig. 2.5), then the reconstructed signal $\hat{X}(z_1, z_2)$ can be expressed in terms of $X(z_1, z_2)$ and the entries of $\mathbf{P}(z_1, z_2)$.

Let the signal at the $(m_1 + m_2 M_1)$th input of $\mathbf{P}(z_1, z_2)$ be denoted by $G_{m_1,m_2}(z_1, z_2)$. Similarly, let the $(m_1 + m_2 M_1)$th output of $\mathbf{P}(z_1, z_2)$ be $C_{m_1,m_2}(z_1, z_2)$. Using (2.5), the signal $G_{m_1,m_2}(z_1, z_2)$ for each branch of the analysis stage can be written as

$$G_{m_1,m_2}(z_1, z_2) = \frac{1}{M_1 M_2} \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} z_1^{-m_1/M_1} \mathcal{W}_1^{-m_1 k_1} z_2^{-m_2/M_2} \mathcal{W}_2^{-m_2 k_2} X(z_1^{1/M_1} \mathcal{W}_1^{k_1}, z_2^{1/M_2} \mathcal{W}_2^{k_2}). \tag{2.8}$$

Passing the signals through $\mathbf{P}(z_1, z_2)$, we get

$$C_{i,j}(z_1, z_2) = \sum_{m_1=0}^{M_1-1} \sum_{m_2=0}^{M_2-1} [\mathbf{P}(z_1, z_2)]_{i+jM_1, m_1+m_2M_1} G_{m_1,m_2}(z_1, z_2), \tag{2.9}$$

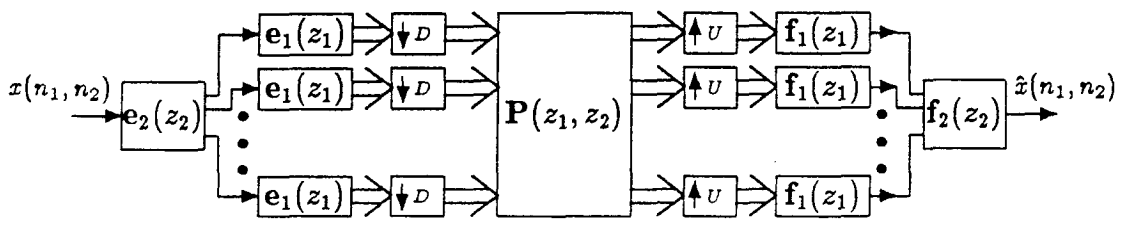

Fig. 2.5. Position of the matrix $\mathbf{P}(z_1, z_2)$ in the filter bank.

and after interpolating and recombining the above signals the reconstructed signal can be expressed as

$$\hat{X}(z_1, z_2) = \sum_{i=0}^{M_1-1} \sum_{j=0}^{M_2-1} z_1^{-(M_1-1-i)} z_2^{-(M_2-1-j)} C_{i,j}(z_1^{M_1}, z_2^{M_2}). \tag{2.10}$$

Combining (2.8- 2.10), the reconstructed signal becomes

$$\hat{X}(z_1, z_2) = \sum_{k_1=0}^{M_1-1} \sum_{k_2=0}^{M_2-1} X(z_1 \mathcal{W}_1^{k_1}, z_2 \mathcal{W}_2^{k_2}) A_{k_1,k_2}(z_1, z_2), \tag{2.11}$$

where $A_{k_1,k_2}(z_1, z_2)$ can be thought of as the weighting function associated with the alias term $X(z_1 W_1^{k_1}, z_2 W_2^{k_2})$. It is given by

$$A_{k_1,k_2}(z_1, z_2) = \frac{1}{M_1 M_2} z_1^{-M_1+1} z_2^{-M_2+1} \sum_{m_1=0}^{M_1-1} \sum_{m_2=0}^{M_2-1} \mathcal{W}_1^{-m_1 k_1} \mathcal{W}_2^{-m_2 k_2} \times$$

$$\sum_{i=0}^{M_1-1} \sum_{j=0}^{M_2-1} z_1^{i-m_1} z_2^{j-m_2} [\mathbf{P}(z_1^{M_1}, z_2^{M_2})]_{i+jM_1, m_1+m_2M_2}. \tag{2.12}$$

In order to eliminate alias terms in (2.11), we must have $A_{0,0}(z_1, z_2) \neq 0$ and $A_{k_1,k_2}(z_1, z_2) = 0$ for all other k_1, k_2. This condition is necessary and sufficient for alias cancellation. (In the absence of aliasing, $A_{0,0}(z_1, z_2)$ becomes the overall tranfer function of the filter bank).

To examine the above condition more closely, we can re-write (2.12) as

$$A_{k_1,k_2}(z_1, z_2) = \frac{1}{M_1 M_2} \sum_{m_1=0}^{M_1-1} \sum_{m_2=0}^{M_2-1} \mathcal{W}_1^{-m_1 k_1} \mathcal{W}_2^{-m_2 k_2} \hat{A}_{m_1,m_2}(z_1, z_2) \tag{2.13}$$

where $\hat{A}_{m_1,m_2}(z_1, z_2)$ is defined appropriately according to (2.12). The sequence of functions, $\{A_{k_1,k_2}(z_1, z_2)\}$, is simply a 2D DFT of the sequence $\{\hat{A}_{m-1,m_2}(z_1, z_2)\}$, so aliasing is cancelled if and only if $\hat{A}_{m_1,m_2}(z_1, z_2) = A_{0,0}(z_1, z_2)$ for all m_1 and m_2. This is the same as

$$\sum_{i=0}^{M_1-1} \sum_{j=0}^{M_2-1} z_1^{i-m_1} z_2^{j-m_2} [\mathbf{P}(z_1^{M_1}, z_2^{M_2})]_{i+jM_1, m_1+m_2M_1}$$

$$= z_1^{M_1-1} z_2^{M_2-1} A_{0,0}(z_1, z_2) \quad \forall m_1, m_2. \tag{2.14}$$

Introducing a new function

$$S(z_1, z_2) = z_1^{M_1-1} z_2^{M_2-1} A_{0,0}(z_1, z_2), \tag{2.15}$$

$S(z_1, z_2)$ can be written in terms of its polyphase components as

$$S(z_1, z_2) = \sum_{l_1=0}^{M_1-1} \sum_{l_2=0}^{M_2-1} S_{l_1,l_2}(z_1^{M_1}, z_2^{M-2}) z_1^{-l_1} z_2^{-l_2}. \tag{2.16}$$

By subsituting this into (2.14) and equating like powers of z_1 and z_2, we get

$$[\mathbf{P}(z_1, z_2)]_{i+jM_1, m_1+m_2M_1} =$$

$$\begin{cases} S_{m_1-i,m_2-j}(z_1, z_2) & \text{for } i - m_1 \leq 0 \text{ and } j - m_2 \leq 0; \\ z_1^{-1} S_{m_1-i+M_1,m_2-j}(z_1, z_2) & \text{for } i - m_1 > 0 \text{ and } j - m_2 \leq 0; \\ z_2^{-1} S_{m_1-i,m_2-j+M_2}(z_1, z_2) & \text{for } i - m_1 \leq 0 \text{ and } j - m_2 > 0; \\ z_1^{-1} z_2^{-1} S_{m_1-i+M_1,m_2-j+M_2}(z_1, z_2) & \text{for } i - m_1 > 0 \text{ and } j - m_2 > 0. \end{cases} \tag{2.17}$$

A close examination of (2.17) shows that $\mathbf{P}(z_1, z_2)$ has the pseudo-circulant format as given in (2.3), where each $\mathbf{P}_k$ is 1D pseudo-circulant in z_1 as shown in (2.2). The entries in the top row of each $\mathbf{P}_k(z_1, z_2)$ are given by $S_{l,k}(z_1, z_2)$. In conclusion, the condition that is necessary and sufficient for alias-cancellation is that $\mathbf{P}(z_1, z_2)$ must be a 2D pseudo-circulant matrix. The top row of $\mathbf{P}(z_1, z_2)$ is related to the polyphase components of $S(z_1, z_2)$.

$$[\mathbf{P}(z_1, z_2)]_{0,l+kM_1} = S_{l,k}(z_1, z_2) \tag{2.18}$$

Once aliasing has been cancelled, the QMF bank becomes a linear shift-invariant system, and the overall system tranfer function is $z_1^{-M_1+1}z_2^{-M_2+1}S(z_1,z_2)$. In other words,

$$\hat{X}(z_1,z_2) = z_1^{-M_1+1}z_2^{-M_2+1}S(z_1,z_2)X(z_1,z_2). \tag{2.19}$$

This result can be generalized for higher dimensional QMF banks.

Note that $\mathbf{P}(z_1,z_2)$ in (2.17) is *not* a Kronecker product of two 1D pseudo-circulant matrices. Therefore, the results here cannot be derived directly from [VA88b], except for the special case where $H_{m_1,m_2}(z_1,z_2)$ and $F_{m_1,m_2}(z_1,z_2)$ are separable filters. In the case of separable filters, $P(z_1,z_2)$ becomes a Kronecker product of two one-dimensional matrices which have to be 1D pseudo-circulant. So alias-cancellation is achieved in each dimension separately.

2.3 Distortion Elimination

Even without aliasing, the reconstructed signal suffers from an overall distortion $S(z_1,z_2)$. There are two types of distortion present in $S(z_1,z_2)$, viz., magnitude and phase distortion. From (2.19), we see

$$|\hat{X}(e^{j\omega_1},e^{j\omega_2})| = |S(e^{j\omega_1},e^{j\omega_2})||X(e^{j\omega_1},e^{j\omega_2})|. \tag{2.20}$$

Thus, there is a distortion in the magnitude of $\hat{X}(e^{j\omega_1},e^{j\omega_2})$. This distortion is eliminated if $|S(e^{j\omega_1},e^{j\omega_2})| = 1$ for all ω_1 and ω_2. This is the same as saying $\tilde{S}(z_1,z_2)S(z_1,z_2) = 1$ (which is commonly called the allpass property). Given an alias-free QMF bank, we now show that $S(z_1,z_2)$ is an allpass function if and only if the matrix $\mathbf{P}(z_1,z_2)$ is lossless, i.e., $\mathbf{P}(z_1,z_2)$ is stable and

$$\tilde{\mathbf{P}}(z_1,z_2)\mathbf{P}(z_1,z_2) = \mathbf{I}. \tag{2.21}$$

Let $\mathbf{p}_{k_1+k_2M_1}(z_1,z_2)$ denote the $(k_1+k_2M_1)$th row of the matrix $\mathbf{P}(z_1,z_2)$. Since the zeroth row of $\mathbf{P}(z_1,z_2)$ contains all the polyphase components of $S(z_1,z_2)$, we can express $S(z_1,z_2)$ in vector form as $S(z_1,z_2)=\mathbf{p}_0^T(z_1^{M_1},z_2^{M_2})\cdot\mathbf{e}(z_1,z_2)$, where $\mathbf{e}(z_1,z_2)$ is defined previously in (2.7). Due to the pseudo-circulant nature of $\mathbf{P}(z_1,z_2)$, we can write $z_1^{-k_1}z_2^{-k_2}S(z_1,z_2)$ as

$$z_1^{-k_1}z_2^{-k_2}S(z_1,z_2)=\mathbf{p}_{k_1+k_2M_1}^T(z_1^{M_1},z_2^{M_2})\cdot\mathbf{e}(z_1,z_2) \tag{2.22}$$

With $k_1=0,\ldots,M_1-1$ and $k_2=0,\ldots,M_2-1$, there are altogether M_1M_2 equations in (2.22). All of these equations can be written collectively as

$$\mathbf{e}(z_1,z_2)S(z_1,z_2)=\mathbf{P}(z_1^{M_1},z_2^{M_2})\mathbf{e}(z_1,z_2) \tag{2.23}$$

This leads to

$$\tilde{S}(z_1,z_2)\tilde{\mathbf{e}}(z_1,z_2)\mathbf{e}(z_1,z_2)S(z_1,z_2)=\tilde{\mathbf{e}}(z_1,z_2)\tilde{\mathbf{P}}(z_1^{M_1},z_2^{M_2})\mathbf{P}(z_1^{M_1},z_2^{M_2})\mathbf{e}(z_1,z_2) \tag{2.24}$$

Since $\tilde{\mathbf{e}}(z_1,z_2)\mathbf{e}(z_1,z_2)=M_1M_2$, and if we assume (2.21) holds, then $\tilde{S}(z_1,z_2)S(z_1,z_2)=1$. Thus the losslessness of $\mathbf{P}(z_1,z_2)$ implies that $S(z_1,z_2)$ is allpass.

We now consider the converse. Starting with (2.23), if we replace z_1 with $z_1\mathcal{W}_1^{l_1}$ and z_2 with $z_2\mathcal{W}_2^{l_2}$, the result will be

$$\mathbf{\Lambda}(z_1,z_2)\mathbf{e}(\mathcal{W}_1^{l_1},\mathcal{W}_2^{l_2})S(z_1\mathcal{W}_1^{l_1},z_2\mathcal{W}_2^{l_2})=\mathbf{P}(z_1^{M_1},z_2^{M_2})\mathbf{\Lambda}(z_1,z_2)\mathbf{e}(\mathcal{W}_1^{l_1},\mathcal{W}_2^{l_2}) \tag{2.25}$$

The matrix $\mathbf{\Lambda}(z_1,z_2)$ is diagonal with the form

$$\mathbf{\Lambda}(z_1,z_2)=\begin{pmatrix}1 & & & \\ & z_2^{-1} & & \\ & & \ddots & \\ & & & z_2^{-M_2+1}\end{pmatrix}\otimes\begin{pmatrix}1 & & & \\ & z_1^{-1} & & \\ & & \ddots & \\ & & & z_1^{-M_1+1}\end{pmatrix}. \tag{2.26}$$

For $0 \leq l_1 < M_1$ and $0 \leq l_2 < M_2$, there are altogether $M_1 M_2$ column equations all of the same form as (2.25). Stacking these columns into matrix form, we can write

$$\mathbf{\Lambda}(z_1, z_2)\mathbf{W}\mathbf{S}(z_1, z_2) = \mathbf{P}(z_1^{M_1}, z_2^{M_2})\mathbf{\Lambda}(z_1, z_2)\mathbf{W}, \tag{2.27}$$

where $\mathbf{S}(z_1, z_2)$ is diagonal whose $(l_1 + l_2 M_1)$th entry is $S(z_1 \mathcal{W}_1^{l_1}, z_2 \mathcal{W}_2^{l_2})$. The matrix $\mathbf{W}$ is the result of stacking the columns, $\mathbf{e}(\mathcal{W}_1^{l_1}, \mathcal{W}_2^{l_2})$, and it is the two-dimensional DFT matrix. Now $\mathbf{\Lambda}(z_1, z_2)$ and $\mathbf{W}$ are both lossless. If $S(z_1, z_2)$ is a stable, allpass function, then $\mathbf{S}(z_1, z_2)$ is lossless also. The product of lossless matrices is also lossless. Consequently, $\mathbf{P}(z_1, z_2)$ must be lossless, proving the desired converse.

For an alias-free QMF bank, it is free from phase distortion if and only if $S(z_1, z_2)$ has linear phase. For a causal, stable filter to have linear phase, it must be FIR. Thus let $S(z_1, z_2)$ be an FIR filter with order $N_1 - 1$ and $N_2 - 1$, that is $S(z_1, z_2) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} s(n_1, n_2) z_1^{-n_1} z_2^{-n_2}$. To obtain linear phase, the sequence $\{s(n_1, n_2)\}$ must be either symmetric, $s(n_1, n_2) = s(N_1 - 1 - n_1, N_2 - 1 - n_2)$, or anti-symmetric $s(n_1, n_2) = -s(N_1 - 1 - n_1, N_2 - 1 - n_2)$. For anti-symmetric FIR filters, the frequency response is zero at $(\omega_1, \omega_2) = (0, 0)$. This means there is severe magnitude distortion around that frequency, hence it is undesirable for QMF bank applications. Only the symmetric FIR filter should be considered. For $S(z_1, z_2)$ to have a symmetric

impulse response its polyphase components must satisfy

$$S_{k,l}(z_1,z_2) = \begin{cases} z_1^{-m_1} z_2^{-m_2} S_{(n_1-k),(n_2-l)} & \\ & \text{for } 0 \le k \le n_1,\ 0 \le l \le n_2. \\ z_1^{-m_1+1} z_2^{-m_2} S_{(M_1+n_1-k),(n_2-l)} & \\ & \text{for } n_1 \le k < M_1,\ 0 \le l \le n_2. \\ z_1^{-m_1} z_2^{-m_2+1} S_{(n_1-k),(M_2+n_2-l)} & \\ & \text{for } 0 \le k \le n_1,\ n_2 \le l < M_2. \\ z_1^{-m_1+1} z_2^{-m_2} S_{(M_1+n_1-k),(M_2+n_2-l)} & \\ & \text{for } n_1 \le k < M_1,\ n_2 \le l < M_2. \end{cases} \tag{2.28}$$

where n_1 and m_1 are unique integers that satisfy $n_1 + m_1 M_1 = N_1 - 1$ with $0 \le n_1 \le M_1$, and similarly $n_2 + m_2 M_2 = N_2 - 1$ with $0 \le n_2 \le M_2 - 1$. (2.17) and (2.28) together give us the necessary and sufficient condition for eliminating phase distortion in the QMF bank.

2.4 FIR Perfect Reconstruction Systems

The QMF bank is said to have perfect reconstruction when $\hat{x}(n_1,n_2) = c \cdot x(n_1 - N_1, n_2 - N_2)$. Without loss of generality, the constant c can be assumed to be unity. In the z-transform domain, $\hat{X}(z_1,z_2) = z_1^{-N_1} z_2^{-N_2} X(z_1,z_2)$, so the overall system function is required to be $A_{0,0}(z_1,z_2) = z_1^{-N_1} z_2^{-N_2}$. For simplicity, assume that $N_1 - M_1 + 1$ is divisible by M_1 so $N_1 - M_1 + 1 = M_1 L_1$, and also $N_2 - M_2 + 1$ is divisible by M_2 with $N_2 - M_2 + 1 = M_2 L_2$. From (2.15), we get $S(z_1,z_2) = z_1^{-M_1 L_1} z_2^{-M_2 L_2}$ which in turn means $\mathbf{P}(z_1,z_2) = z_1^{-L_1} z_2^{-L_2} \mathbf{I}$. The only way to achieve this is to make $\mathbf{R}(z_1,z_2) = z_1^{-L_1} z_2^{-L_2} \mathbf{E}^{-1}(z_1,z_2)$. This suggests possible ways of designing QMF banks that have the perfect reconstruction property.

The problem basically involves the design of the analysis filters $H_{m,n}(z_1,z_2)$ so that each of them is a good bandpass filter with passbands disjointed from each

other. At the same time, the polyphase matrix of the analysis filter bank is constrained to be FIR and lossless, for if $\mathbf{E}(z_1, z_2)$ is FIR and lossless then the polyphase matrix for the synthesis filter bank is given simply by $\mathbf{R}(z_1, z_2) = z_1^{-L_1} z_2^{-L_2} \tilde{\mathbf{E}}(z_1, z_2)$. The above choice will guarantee perfect reconstruction. (The integers L_1 and L_2 may be chosen to cancel out the highest positive powers of z_1 and z_2, so that $\mathbf{R}(z_1, z_2)$ contains no positive powers of z_1, z_2. This is important for causality requirements.)

For the 1D QMF bank, the lossless polyphase matrix $\mathbf{E}(z)$ has been successfully used in the design of analysis filters [VA87b]. This is due to the fact that there exists a factorization theorem for one-dimensional lossless matrices. Let $\mathbf{E}(z)$ be any $M \times M$ causal FIR lossless matrix. Since it is causal and FIR (i.e., it is a matrix polynomial in z^{-1}), its determinant must be a polynomial in z^{-1}. Due to the lossless property, the determinant is an allpass function. For a polynomial to be an allpass function, it must have the form of a monomial z^{-K}. The theorem in [DO88] allows us to represent $\mathbf{E}(z)$, as a product of matrix factors such as

$$\mathbf{E}(z) = \mathbf{V}_1(z)\mathbf{V}_2(z)\ldots\mathbf{V}_K(z)\mathbf{R}, \tag{2.29}$$

where $\mathbf{R}$ is an orthogonal matrix and the other factors $\mathbf{V}_i(z_1)$ have the form $(\mathbf{I} - (1 - z_1^{-1})\mathbf{v}_i\mathbf{v}_i^\dagger)$. The vector $\mathbf{v}_i$ is normalized to have unit norm, $\mathbf{v}_i^\dagger\mathbf{v}_i = 1$. One can verify that each $\mathbf{V}_i(z)$ is a lossless matrix and its determinant is z^{-1}. In (2.29), $\mathbf{E}(z)$ is completely characterized by the unit-norm vectors $\mathbf{v}_i$ plus the orthogonal matrix $\mathbf{R}$, which in turn can be decomposed into a number of planar rotations. In all, this gives us the minimum number of parameters needed to characterize any $M \times M$ causal FIR lossless matrix with determinant equal to z^{-K}. By adjusting these parameters by non-linear optimization procedures, one can generate analysis filters having good desired response.

The factorization of 2D FIR lossless matrices $\mathbf{E}(z_1, z_2)$ is much harder. It is not known whether an equivalent factorization for 2D FIR lossless matrices exists. However, under certain restrictive conditions, we can show that a factorization into degree-one factors is possible. This is stated in the following lemma.

Lemma 2.1 Let $\mathbf{E}(z_1, z_2)$ be an $M \times M$ causal FIR lossless matrix with order equal to unity in the z_1 variable and arbitrary order in z_2 (i.e., $\mathbf{E}(z_1, z_2)$ is a matrix polynomial of the form $\mathbf{E}(z_1, z_2) = \sum_{n_1=0}^{1} \sum_{n_2=0}^{J_2} \mathbf{e}_{n_1,n_2} z_1^{-n_1} z_2^{-n_2}$ where $\mathbf{e}_{n_1,n_2}$ are $M \times M$ matrices.) Without loss of generality, one can write the determinant of $\mathbf{E}(z_1, z_2)$ as $z_1^{-K_1} z_2^{-K_2}$. Under these conditions, the lemma states that $\mathbf{E}(z_1, z_2)$ can be factorized as

$$\mathbf{E}(z_1, z_2) = \mathbf{V}_1(z_2)\mathbf{V}_2(z_2) \ldots \mathbf{V}_J(z_2)\mathbf{U}(z_1)\mathbf{V}_{J+1}(z_2) \ldots \mathbf{V}_{K_2}(z_2). \tag{2.30}$$

where the factors $\mathbf{V}_i(z_2)$ have the form $(\mathbf{I} - (1 - z_2^{-1})\mathbf{v}_i\mathbf{v}_i^\dagger)$ with $\mathbf{v}_i$ being an unit-norm vector of length M. The factor $\mathbf{U}(z_1)$ is a one-dimensional FIR lossless system, with determinant equal to $z_1^{-K_1}$. And J is some integer less than or equal to K_2.

To prove the above lemma, we will first show that $\mathbf{E}(z_1, z_2)$ can be factorized in one of two ways,

$$\mathbf{E}(z_1, z_2) = \mathbf{V}(z_2)\hat{\mathbf{E}}(z_1, z_2) \tag{2.31a}$$

$$\text{or} \quad \mathbf{E}(z_1, z_2) = \hat{\mathbf{E}}(z_1, z_2)\mathbf{V}(z_2). \tag{2.31b}$$

The factor $\hat{\mathbf{E}}(z_1, z_2)$ is required to be causal, FIR and lossless. Since $\mathbf{V}(z_2)$ is lossless, the equations in (2.31) can be re-written as

$$\hat{\mathbf{E}}(z_1, z_2) = (\mathbf{I} - (1 - z_2)\mathbf{v}\mathbf{v}^\dagger)\mathbf{E}(z_1, z_2) \tag{2.32a}$$

$$\text{or} \quad \hat{\mathbf{E}}(z_1, z_2) = \mathbf{E}(z_1, z_2)(\mathbf{I} - (1 - z_2)\mathbf{v}\mathbf{v}^\dagger). \tag{2.32b}$$

Thus we see immediately that $\hat{\mathbf{E}}(z_1, z_2)$ is FIR and lossless with determinant $z_1^{-K_1} z_2^{-(K_2-1)}$. The degree of the determinant is reduce by one in the z_2 variable when the factorization of (2.31a) or (2.31b) is carried out. To satisfy the causality requirement, we need to have

$$\mathbf{v}^\dagger \mathbf{e}_{0,0} = \mathbf{v}^\dagger \mathbf{e}_{1,0} = \mathbf{0} \tag{2.33a}$$

(in which case (2.31a) will follow) or

$$\mathbf{e}_{0,0}\mathbf{v} = \mathbf{e}_{1,0}\mathbf{v} = \mathbf{0} \tag{2.33b}$$

which gives us the factorization of (2.31b). The lossless property states that $\tilde{\mathbf{E}}(z_1, z_2)\, \mathbf{E}(z_1, z_2) = \mathbf{I}$, and in particular the $z_1 z_2^{J_2}$-term, $z_2^{J_2}$-term, and $z_1^{-1} z_2^{J_2}$-term of $\tilde{\mathbf{E}}(z_1, z_2)\mathbf{E}(z_1, z_2)$ should be zero:

$$\tilde{\mathbf{e}}_{1,J_2}\mathbf{e}_{0,0} = \mathbf{0}, \tag{2.34a}$$

$$\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0} + \tilde{\mathbf{e}}_{1,J_2}\mathbf{e}_{1,0} = \mathbf{0}, \tag{2.34b}$$

$$\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{1,0} = \mathbf{0}. \tag{2.34c}$$

On the other hand, $\mathbf{E}(z_1, z_2)\tilde{\mathbf{E}}(z_1, z_2) = \mathbf{I}$ and so

$$\mathbf{e}_{0,0}\tilde{\mathbf{e}}_{1,J_2} = \mathbf{0}, \tag{2.34d}$$

$$\mathbf{e}_{0,0}\tilde{\mathbf{e}}_{0,J_2} + \mathbf{e}_{1,0}\tilde{\mathbf{e}}_{1,J_2} = \mathbf{0}, \tag{2.34e}$$

$$\mathbf{e}_{1,0}\tilde{\mathbf{e}}_{0,J_2} = \mathbf{0}. \tag{2.34f}$$

We will assume that $\mathbf{e}_{0,0}$ and $\mathbf{e}_{1,0}$ are not both zero (otherwise one may factorize out a trivial factor $z_2^{-1}\mathbf{I}$ from $\mathbf{E}(z_1, z_2)$), and also $\mathbf{e}_{0,J_2}$ and $\mathbf{e}_{1,J_2}$ are not both zero.

The exsitence of a nonzero vector $\mathbf{v}$ that will satisfy either (2.33a) or (2.33b) is proved as follows.

1.) If $\tilde{\mathbf{e}}_{1,J_2} = \mathbf{0}$, then from (2.34b) and (2.34c) we know $\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0} = \mathbf{0}$ and $\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{1,0} = \mathbf{0}$. So by letting $\mathbf{v}^\dagger$ be a nonzero row of $\tilde{\mathbf{e}}_{0,J_2}$, (2.33a) will hold;

2.) If $\tilde{\mathbf{e}}_{1,J_2} \neq \mathbf{0}$ and $\tilde{\mathbf{e}}_{1,J_2}\mathbf{e}_{1,0} = \mathbf{0}$, then (2.33a) will hold with $\mathbf{v}^\dagger$ being a nonzero row of $\tilde{\mathbf{e}}_{1,J_2}$;

3.) Lastly, if $\tilde{\mathbf{e}}_{1,J_2} \neq \mathbf{0}$ and $\tilde{\mathbf{e}}_{1,J_2}\mathbf{e}_{1,0} \neq \mathbf{0}$, then consider the matrix $\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0}$ which is nonzero due to (2.34b). Post-multiplying both sides of (2.34f) by $\mathbf{e}_{0,0}$ gives us $\mathbf{e}_{1,0}\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0} = \mathbf{0}$. Performing the same multiplication on (2.34e) gives us $\mathbf{e}_{0,0}\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0} + \mathbf{e}_{1,0}\tilde{\mathbf{e}}_{1,J_2}\mathbf{e}_{0,0} = \mathbf{0}$ which is reduced to $\mathbf{e}_{0,0}\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0} = \mathbf{0}$ by (2.34a). By letting $\mathbf{v}$ be a nonzero column of $\tilde{\mathbf{e}}_{0,J_2}\mathbf{e}_{0,0}$ we get (2.33b).

Hence, we have proved that at least one of the factorizations in (2.31) is possible. By applying (2.31) recursively to $\hat{\mathbf{E}}(z_1, z_2)$, the z_2-degree of the determinant is reduced by one each time. After repeating the process K_2 number of times, $\mathbf{E}(z_1, z_2)$ can be decomposed as

$$\mathbf{E}(z_1, z_2) = \mathbf{V}_1(z_2)\ldots\mathbf{V}_J(z_2)\mathbf{W}(z_1, z_2)\mathbf{V}_{J+1}(z_2)\ldots\mathbf{V}_{K_2}(z_2). \tag{2.35}$$

where $\mathbf{W}(z_1, z_2)$ is a causal FIR lossless matrix with determinant equal to $z_1^{-K_1}$. Now we claim that $\mathbf{W}(z_1, z_2)$ does not contain any powers of z_2. It does not contain positive powers of z_2, since in each step of the factorization we have ensured that each factor is causal. Now, if $\mathbf{W}(z_1, z_2)$ contains negative powers of z_2, then the factorization can continue and we get

$$\mathbf{W}(z_1, z_2) = \hat{\mathbf{W}}(z_1, z_2)\mathbf{V}(z_2) \quad \text{or} \quad \mathbf{W}(z_1, z_2) = \mathbf{V}(z_2)\hat{\mathbf{W}}(z_1, z_2) \tag{2.36}$$

with $\hat{\mathbf{W}}(z_1, z_2)$ being a causal FIR matrix with determinant equal to $z_1^{-K_1}z_2$! This is impossible, so $\mathbf{W}(z_1, z_2)$ is a function of z_1 only. We can write $\mathbf{W}(z_1, z_2) = \mathbf{U}(z_1)$ where $\mathbf{U}(z_1, z_2)$ is a causal 1D FIR lossless matrix with determinant equal to $z_1^{-K_1}$. Thus the factorization in (2.30) results.

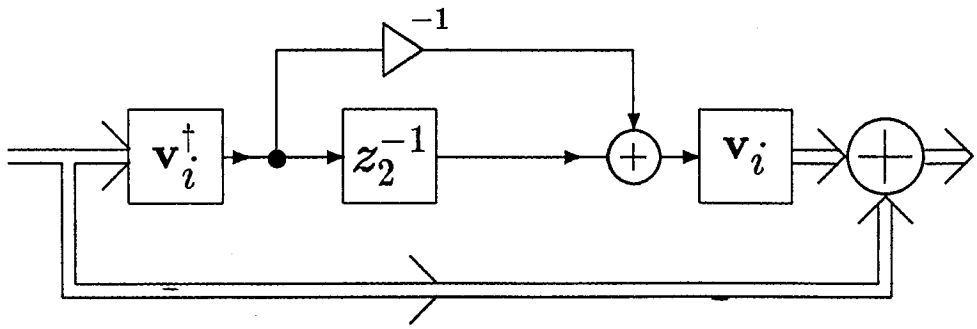

Fig. 2.6. An implementation of the degree-one lossless system, $\mathbf{V}_i(z_2)$.

Each of the $\mathbf{V}_i(z_2)$ factors in (2.30) can be implemented using one delay in z_2. A possible implementation is shown in Fig. 2.6 for $\mathbf{V}_i(z_2)$. The implementation for $\mathbf{U}(z)$ is done using (2.29) where each $\mathbf{V}_i(z_1)$ again requires only one delay in z_1. The total number of delay elements needed for $\mathbf{E}(z_1, z_2)$ is $K_1 + K_2$. This turns out to be the minimum number of delays that is necessary for implementing $\mathbf{E}(z_1, z_2)$. Let us assume that there exists a structure that requires a total of fewer than $K_1 + K_2$ delays, then by changing all the z_2^{-1} elements in that strucuture into z_1^{-1} we can obtain an implementation for $\mathbf{E}(z_1, z_1)$ with fewer delays than K_1+K_2. But $\mathbf{E}(z_1, z_1)$ is a 1D FIR lossless system with determinant $z_1^{-(K_1+K_2)}$, and in [VA88b] it has been shown that the McMillan degree of such a system is $K_1 + K_2$. (For 1D systems, the McMillan degree is defined to be the minimum number of delays necessary for implementing the system.) This contradicts our initial assumption that there exists a structure for $\mathbf{E}(z_1, z_2)$ which utilizes fewer than $K_1 + K_2$ delays. Hence, $K_1 + K_2$ is the minimum number of delay required for implementing $\mathbf{E}(z_1, z_2)$

Similar result holds when $\mathbf{E}(z_1, z_2)$ has its order in z_2 equal to one and arbitrary order in z_1. In other words, $\mathbf{E}(z_1, z_2) = \sum_{n_1=0}^{J_1} \sum_{n_2=0}^{1} \mathbf{e}_{n_1,n_2} z_1^{-n_1} z_2^{-n_2}$.

The above factorization holds for special cases of $\mathbf{E}(z_1, z_2)$. However, it is not clear if a general causal FIR lossless matrix can be factorized in a similar way.

2.5 Design of Analysis Filters Using Lossless Polyphase Matrix

In lieu of a general structure that would represent all causal FIR lossless matrices, an ad hoc method for designing the analysis filters is suggested here. Take the case of $M_1 = M_2 = 2$. The QMF bank has four branches. By requiring $\mathbf{E}(z_1, z_2)$ to be a lossless FIR matrix, the perfect reconstruction property can be obtained as shown in Sec. 2.4. To impose the losslessness constraint on $\mathbf{E}(z_1, z_2)$, it is restricted to have the following form

$$\mathbf{E}(z_1, z_2) = \mathbf{R}\mathbf{V}_L(z_2)\mathbf{V}_{L-1}(z_1)\ldots\mathbf{V}_4(z_2)\mathbf{V}_3(z_1)\mathbf{V}_2(z_2)\mathbf{V}_1(z_1), \tag{2.37}$$

where $\mathbf{R}$ is a 4×4 orthogonal matrix and $\mathbf{V}_i(z)$ is of the form $\left(\mathbf{I}-(1-z^{-1})\mathbf{v}_i\mathbf{v}_i^T\right)$ as discussed previously in Sec. 2.4. This structure is motivated by the one-dimensional structure used in [VA88a] for 1D QMF design. Since each factor in (2.37) is lossless, therefore the polyphase matrix is constrained to be lossless also. Each of the unit-norm vectors $\mathbf{v}_i$ can be characterized by three parameters, for example one can write the vectors as $(v_{i,1}v_{i,2}v_{i,3}v_{i,4})^T$ where $v_{i,1} = \cos(\theta_i)$, $v_{i,2} = \sin(\theta_i)\cos(\phi_i)$, $v_{i,3} = \sin(\theta_i)\sin(\phi_i)\cos(\psi_i)$, $v_{i,4} = \sin(\theta_i)\sin(\phi_i)\sin(\psi_i)$. The orthogonal matrix $\mathbf{R}$ can be characterized by 6 planar rotations. Thus $\mathbf{E}(z_1, z_2)$ is characterized by a total of $3(L+2)$ parameters.

The order of the analysis filters is determined by $L+1$. Since the anaylsis filter bank is given by $\mathbf{h}(z_1, z_2) = \mathbf{E}(z_1, z_2)\mathbf{e}(z_1, z_2)$, each the four filters are functions of these parameter. The passbands of the analysis filters are decided by the particular application in mind. For the purpose of our example here, we will choose the

passband for $H_0(z_1, z_2)$ to be a "fan" shaped region which is the union of two triangle

$$\Big\{(\omega_1, \omega_2)|\omega_1 > 0,\ \omega_2 > 0,\ \omega_1 + \omega_2 < \pi\Big\} \cup$$

$$\Big\{(\omega_1, \omega_2)|\omega_1 < 0,\ \omega_2 < 0,\ \omega_1 + \omega_2 > -\pi\Big\}. \qquad (2.38)$$

The passband for $H_1(z_1, z_2)$ will be the above region shifted along the ω_1 axis by π. The passband for $H_2(z_1, z_2)$ is given by (2.38) shifted by π along the ω_2 direction. As for $H_3(z_1, z_2)$, it passband region will consist of (2.38) shifted by π in both directions. One can verify that the four filters have non-overlapping passbands and the union of the four passbands covers the entire frequency plane.

The filters should be designed so that in the stopband region of each filter its magnitude response is close to zero, $|H_i(e^{j\omega_1}, e^{j\omega_2})| \approx 0$. Let us define the stopband energy of the filter $H_i(z)$ to be the integral

$$E_s^{(i)} = \int\int H_i^*(e^{j\omega_1}, e^{j\omega_2}) H_i(e^{j\omega_1}, e^{j\omega_2}) \frac{d\omega_1 d\omega_2}{4\pi^2} \qquad (2.39)$$

where the region of integration is the stopband region of the filter. With each $H_i(z_1, z_2)$ being an FIR filter of the form $\sum_{n_1=0}^{L-1} \sum_{n_2=0}^{L-1} h_i(n_1, n_2) z_1^{-n_1} z_2^{-n_2}$, the quantity $E_s^{(i)}$ can be written as a function of the impulse response coefficients, namely

$$E_s^{(i)} = g_i(0,0)\alpha_i(0,0) + 2\sum_{n_2=1}^{L-1} g_i(0, n_2)\alpha_i(0, n_2)$$

$$+ 2\sum_{n_1=1}^{L-1} \sum_{n_2=-L+1}^{L-1} g_i(n_1, n_2)\alpha_i(n_1, n_2). \qquad (2.40)$$

where $g_i(n_1, n_2)$ is the auto-correlation of $h_i(n_1, n_2)$ and $\alpha_i(n_1, n_2)$ is given by $\int\int \cos(n_1\omega_1 + n_2\omega_2) d\omega_1 d\omega_2 / 4\pi^2$. The integrals in the expression for $\alpha_i(n_1, n_2)$ can be evaluated either analytically or by numerical integration (if a close-form

expression cannot be readily found). The impulse response coefficients are functions of the parameters of $\mathbf{E}(z_1, z_2)$ in (2.37). Using an optimization algorithm, the total stopband energy, $E_s^{(0)} + E_s^{(1)} + E_s^{(2)} + E_s^{(3)}$, can be minimized with respect to the $3(L+2)$ parameters.

If the stopband energy for each filter is sufficiently small, then the response of the filter is close to zero in the stopband. This in turn guarantees that the passband response is close to unity. The reason for this is as follows. Due to the lossless property of $\mathbf{E}(z_1, z_2)$,

$$\sum_{i=0}^{3} \tilde{H}_i(z_1, z_2) H_i(z_1, z_2) = \tilde{\mathbf{e}}(z_1, z_2)\tilde{\mathbf{E}}(z_1, z_2)\mathbf{E}(z_1, z_2)\mathbf{e}(z_1, z_2) = M_1 M_2 \qquad (2.41)$$

On the unit bi-sphere, this means

$$\sum_{i=0}^{3} H_i^*(e^{j\omega_1}, e^{j\omega_2}) H_i(e^{j\omega_1}, e^{j\omega_2}) = M_1 M_2. \qquad (2.42)$$

Since the four filters have non-overlapping passbands, the passband of any one filter will lie within the stopbands of the other three. If each filter $H_i(e^{j\omega_1}, e^{j\omega_2}) \approx 0$ in its respective stopband, then we know from (2.42) that $H_i(e^{j\omega_1}, e^{j\omega_2}) \approx M_1 M_2$ in its passband. Hence good stopband attenuation for all four of the filters will also mean a good passband performance for the filters.

Design Example 2.1 To demonstrate the above design procedure, we let $\mathbf{E}(z_1, z_2)$ have the form in (2.37) with $L = 30$. So there are 96 parameters in all. A mathematical library subroutine (ZXMIN in the International Mathematical and Statistical Library) is used to minimize the total stopband energy as discussed before. The subroutine is based on the quasi-Newton method for minimization. A local optimum solution is listed in Table 2.1 in terms of the entries for each of the vectors

$\mathbf{v}_1$	0.290675	-.875347	-.028410	0.385316
$\mathbf{v}_2$	0.484210	0.592716	-.047326	0.641863
$\mathbf{v}_3$	0.384293	0.915114	-.103294	-.064924
$\mathbf{v}_4$	-.626686	-.622921	0.022367	0.467690
$\mathbf{v}_5$	0.444359	-.790436	-.186446	-.378147
$\mathbf{v}_6$	0.129461	0.438447	-.680554	-.572582
$\mathbf{v}_7$	-.490077	-.303727	-.795030	0.188418
$\mathbf{v}_8$	0.017694	0.770706	0.212216	-.600553
$\mathbf{v}_9$	0.799441	-.506313	-.135045	0.293777
$\mathbf{v}_{10}$	0.361046	-.754792	-.063723	-.543944
$\mathbf{v}_{11}$	0.918421	0.123016	0.244872	-.285319
$\mathbf{v}_{12}$	-.850465	-.203020	0.485093	-.013286
$\mathbf{v}_{13}$	0.735829	0.186510	-.647566	0.066546
$\mathbf{v}_{14}$	-.641098	-.289323	0.707548	0.068281
$\mathbf{v}_{15}$	-.441061	-.487900	0.753271	0.001003
$\mathbf{v}_{16}$	0.415231	-.128779	0.649627	-.623686
$\mathbf{v}_{17}$	0.026506	-.099967	0.705148	0.701477
$\mathbf{v}_{18}$	0.065981	0.906860	0.371562	-.187596
$\mathbf{v}_{19}$	0.001800	0.927606	0.090004	-.362551
$\mathbf{v}_{20}$	0.846187	-.086868	0.357794	0.385234
$\mathbf{v}_{21}$	-.599662	-.078675	0.656369	0.450994
$\mathbf{v}_{22}$	0.856627	-.026813	-.409761	-.312357
$\mathbf{v}_{23}$	0.736520	0.029068	-.671718	-.074089
$\mathbf{v}_{24}$	-.735058	-.102114	0.475132	-.472770
$\mathbf{v}_{25}$	-.583123	-.154593	0.796474	0.041198
$\mathbf{v}_{26}$	0.065606	-.953211	0.294922	0.010272
$\mathbf{v}_{27}$	0.012881	-.807459	-.010576	0.589689
$\mathbf{v}_{28}$	0.549584	-.185934	0.664653	0.470768
$\mathbf{v}_{29}$	0.276359	0.470955	-.668846	0.504452
$\mathbf{v}_{30}$	0.150582	0.439369	-.795841	0.388481
$\mathbf{R}$	0.179484	-.248734	0.152732	0.939463
	0.494943	0.496166	0.709010	-.078459
	-.776960	0.539794	0.193391	0.259915
	-.345182	-.632909	0.660740	-.209042

Table 2.1 Entries of $\mathbf{v}_i$'s and $\mathbf{R}$ for the optimized 2D filter bank.

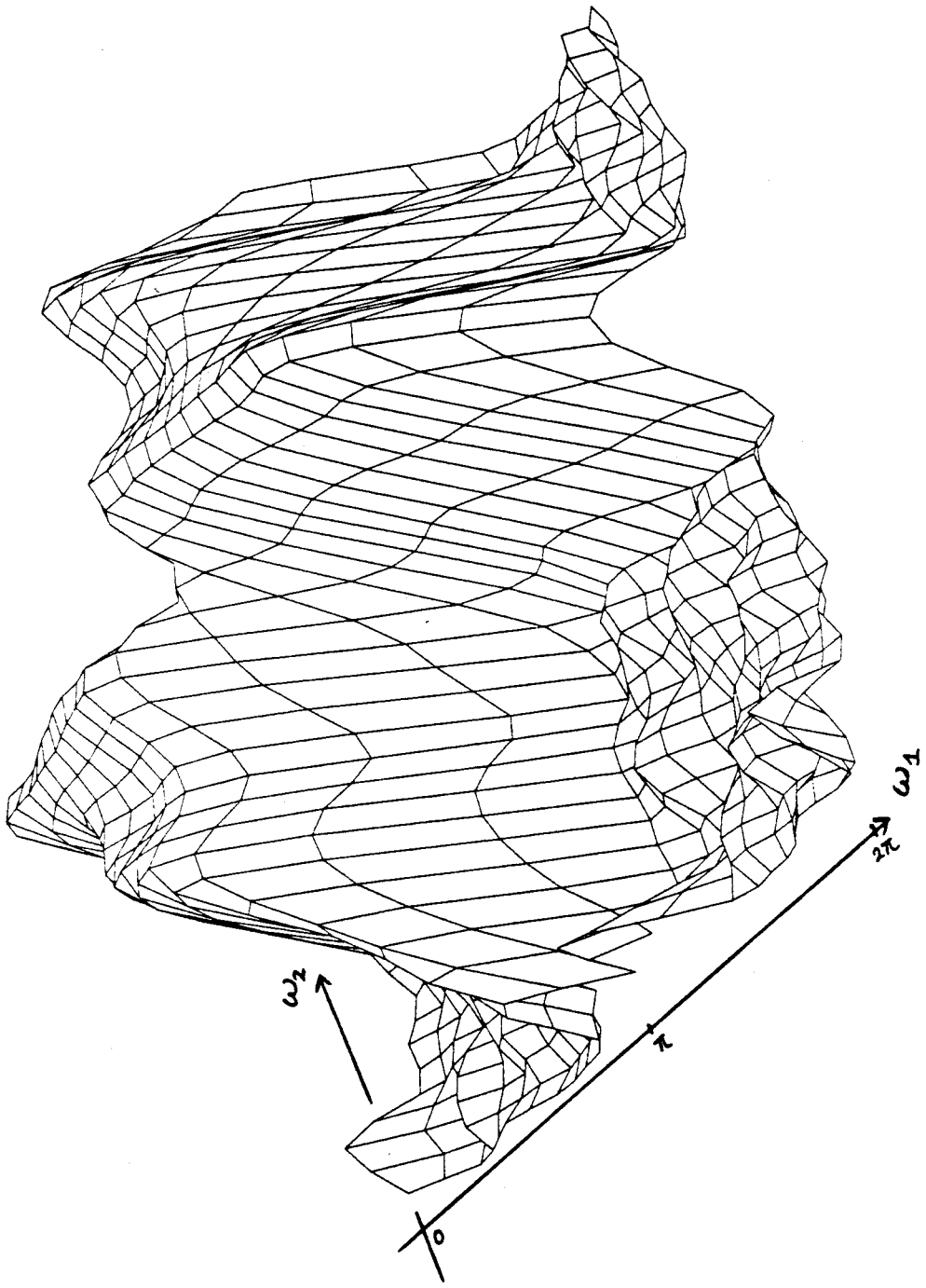

Fig. 2.7. Magnitude response of the filter $H_3(z_1, z_2)$.

$\mathbf{v}_i$'s and the entries of $\mathbf{R}$. The magnitude response $|H_3(e^{j\omega_1}, e^{j\omega_1})|$ is plotted in Fig. 2.7. It shows roughly the fan-shaped region of the passband.

There are several drawbacks to the above design method. One is that the structure in (2.37) might not be able to generate all FIR lossless matrices of a given order. Second, as the filter order $L+1$ increases, the amount of computation needed to evaluate the stopband energy increases rapidly. Using fast algorithms, the number of multiplications required to compute the auto-correlation of $h_i(n_1, n_2)$ is on the order of $L^2 \log(L)$, while the sum in (2.40) requires about $2L^2$ multiplications. Therefore the time for each iteration increases on the order of $L^2 \log(L)$.

2.6 Block Filtering and Pseudo-circulants

Block implementation of digital filters has been explored by various authors in the past [BU72] [MIT78] [BA80]. For a single-input single-output (SISO) filter, its input consists of a sequence of numbers. Block filtering involves the conversion of this serial input into parallel inputs in the form of a vector sequence. And the filtering operation is performed on the sequence of vectors instead of on the original scalar sequence. The aim is to increase computational speed through parallel processing, and also certain effects due to the finite word-length of the filter, e.g., sensitivity and noise gain, are reduced as a result of blocking. In [VA88b], it has been shown how the pseudo-circulant matrix is related to block filtering in one dimension. The two-dimensional pseudo-circulant matrices discussed here also appear in the context of 2D block filtering. Let $y(n_1, n_2)$ be the output of the filter $S(z_1, z_2)$ in response to the input $x(n_1, n_2)$.

The blocked version of $y(n_1, n_2)$ is defined as a sequence of vectors $\{\mathbf{y}_B(n_1, n_2)\}$

where each vector has length MN and

$$\mathbf{y}_B(n_1,n_2) = \begin{pmatrix} \mathbf{y}_{M_2-1}(n_1,n_2) \\ \mathbf{y}_{M_2-2}(n_1,n_2) \\ \vdots \\ \mathbf{y}_1(n_1,n_2) \\ \mathbf{y}_0(n_1,n_2) \end{pmatrix}, \tag{2.43}$$

with each sub-vector defined as

$$\mathbf{y}_i(n_1,n_2) = \begin{pmatrix} y(n_1M_1+M_1-1, n_2M_2+i) \\ y(n_1M_1+M_1-2, n_2M_2+i) \\ \vdots \\ y(n_1M_1+1, n_2M_2+i) \\ y(n_1M_1, n_2M_2+i) \end{pmatrix}. \tag{2.44}$$

The blocked version of $x(n_1,n_2)$ is defined similarly. Taking the z-transforms of $\mathbf{y}_B(n_1,n_2)$ and $\mathbf{x}_B(n_1,n_2)$ respectively, it is possible to find a 2D matrix function $\mathbf{H}(z_1,z_2)$ that would relate the two as $\mathbf{Y}_B(z_1,z_2) = \mathbf{H}(z_1,z_2)\mathbf{X}_B(z_1,z_2)$. The matrix $\mathbf{H}(z_1,z_2)$ describes a multi-input multi-output (MIMO) system. Here we will show that $\mathbf{H}(z_1,z_2)$ is pseudo-circulant, and the entries of the zeroth row of $\mathbf{H}(z_1,z_2)$ are precisely the polyphase components of the original SISO filter $S(z_1,z_2)$. As a result, given $S(z_1,z_2)$ we are able to find $\mathbf{H}(z_1,z_2)$ directly.

Let $Y_{l_1,l_2}(z_1,z_2)$ be the (l_1,l_2)th polyphase component of $Y(z_1,z_2)$, then $[\mathbf{Y}_B(z_1,z_2)]_{m_1+m_2M}$ equals the polyphase term associated with $z_1^{-M_1+1+m_1}$ $z_2^{-M_2+1+m_2}$, i.e.,

$$\left[\mathbf{Y}_B(z_1,z_2)\right]_{m_1+m_2M} = Y_{M_1-1-m_1,M_2-1-m_2}(z_1,z_2). \tag{2.45}$$

The same relationship holds between $\mathbf{X}_B(z_1,z_2)$ and $X(z_1,z_2)$.

Let us apply to the filter $S(z_1,z_2)$ an input $x(n_1,n_2)$ having a z-transform of the form $X(z_1,z_2) = z_1^{-(M_1-1-k_1)}z_2^{-(M_2-1-k_2)}$, then the output transform is given by

$$Y(z_1,z_2) = z_1^{-(M_1-1-k_1)}z_2^{-(M_2-1-k_2)}S(z_1,z_2). \tag{2.46}$$

We will let k_1 and k_2 be restricted to the range, $0 \leq k_1 < M_1$ and $0 \leq k_2 < M_2$. Using the relationship in (2.45) which is also valid for $\mathbf{X}_B(z_1, z_2)$, the blocked version of $X(z_1, z_2)$ is found to be

$$\left[\mathbf{X}_B(z_1, z_2)\right]_{m_1+m_2M_1} = \begin{cases} 1 & m_1 = k_1, m_2 = k_2 \\ 0 & \text{otherwise} \end{cases} \tag{2.47}$$

Hence the blocked output $\mathbf{Y}_B(z_1, z_2)$ equals the $(k_1 + k_2M_1)$th column of $\mathbf{H}(z_1, z_2)$,

$$\left[\mathbf{Y}_B(z_1, z_2)\right]_{m_1+m_2M_1} = \left[\mathbf{H}(z_1, z_2)\right]_{m_1+m_2M_1, k_1+k_2M_1} \tag{2.48}$$

When compared with (2.45), the above entry of $\mathbf{H}(z_1, z_2)$ corresponds to the appropriate polyphase term of $Y(z_1, z_2)$. And from (2.46), we get

$$\begin{aligned} &\left[\mathbf{H}(z_1, z_2)\right]_{m_1+m_2M_1, k_1+k_2M_1} \\ &= Y_{M_1-1-m_1, M_2-1-m_2}(z_1, z_2) \\ &= \begin{cases} S_{k_1-m_1, k_2-m_2}(z_1, z_2) & \text{for } k_1 \geq m_1, \quad k_2 \geq m_2; \\ z_1^{-1} S_{k_1-m_1+M_1, k_2-m_2}(z_1, z_2) & \text{for } k_1 < m_1, \quad k_2 \geq m_2; \\ z_2^{-1} S_{k_1-m_1, k_2-m_2+M_2}(z_1, z_2) & \text{for } k_1 \geq m_1, \quad k_2 < m_2; \\ z_1^{-1} z_2^{-1} S_{k_1-m_1+M_1, k_2-m_2+M_2}(z_1, z_2) & \text{for } k_1 < m_1, \quad k_2 < m_2. \end{cases} \end{aligned} \tag{2.49}$$

This shows that $\mathbf{H}(z_1, z_2)$ has exactly the same form as the matrix $\mathbf{P}(z_1, z_2)$ in (2.17), which is 2D pseudo-circulant. Thus $\mathbf{H}(z_1, z_2)$ is a 2D pseudo-circulant matrix whose first row is given by the polyphase components of $S(z_1, z_2)$.

2.7 Extension to Multi-dimensions

In Sec 2.2, the condition for an alias-free 2D QMF bank is derived. It is found that aliasing is cancelled if and only if the product of polyphase matrices, $\mathbf{R}(z_1, z_2)\mathbf{E}(z_1, z_2)$, is in the form of a 2D pseudo-circulant matrix. This can be generalized to multi-dimensional QMF banks. For example, in K dimensions with decimation factors $(M_1, M_2, \ldots M_K)$, the QMF bank will have $M_0M_1 \ldots M_K$ branches.

Let $M = \prod_{k=0}^{K} M_k$. The necessary and sufficient condition for alias-cancellation is that the matrix $\mathbf{P}(z_1, \ldots, z_K)$ must be a K-dimensional pseudo-circulant matrix. Since there are M branches, $\mathbf{P}(z_1, \ldots, z_K)$ is an $M \times M$ matrix.

The pseudo-circulant property for K-dimensions can be defined recursively by partitioning $\mathbf{P}(z_1, \ldots, z_K)$ into-sub-matrices, each with dimensions $\frac{M}{M_K} \times \frac{M}{M_K}$. Let the (l, m)th sub-matrix be denoted by $\mathbf{Q}_{l,m}(z_1, \ldots, z_K)$ where $0 \leq l, m < M_K$. The sub-matrices must satisfy the following relations:

$$\mathbf{Q}_{l,m}(z_1, \ldots, z_K) = \begin{cases} \mathbf{Q}_{0,m-l}(z_1, \ldots, z_K) & l \leq m; \\ z_K^{-1} \mathbf{Q}_{0,m-l+M_K}(z_1, \ldots, z_K) & l > m, \end{cases} \tag{2.50}$$

and each $\mathbf{Q}_{l,m}(z_1, \ldots, z_K)$ is pseudo-circulant in the first $K - 1$ variables.

Chapter III. Periodic Nonuniform Sampling and the Efficient Reconstruction of Discrete-time Bandlimited Signals

3.1 Introduction

In this chapter we shall study periodic nonuniform sampling as a potential means for data compression of bandlimited signals, and the multirate filter bank concepts discussed previously can be used to derive the compression and reconstruction procedures. It has been well-known that a bandlimited signal can in principle be recovered from its nonuniformly spaced samples provided the 'average sampling rate' exceeds the Nyquist rate. See, for example, [BL53] [JE77] [PAP77a] and references therein. In the chapter we shall address the problem of *efficient reconstruction* of the original signal from such samples.

We shall be concerned only with discrete-time signals here so that nonuniform sampling actually implies *nonuniform decimation.* As a typical example of such an operation, let $x(n)$ be a σ-bandlimited signal, i.e., a signal for which the Fourier transform $X(e^{j\omega})$ is zero for $\sigma \leq |\omega| \leq \pi$. We assume that $\sigma \leq \frac{L}{M}\pi$ where L and M are positive integers with $L < M$. A common procedure [CR83] for compressing the rate of such a signal by M/L would be to use an L-fold interpolator followed by filtering and M-fold decimation (see, for example, Fig. 3.1(a) where $L = 2$ and $M = 3$). The output signal in Fig. 3.1(a) is at $\frac{2}{3}$ times the original sampling rate of $x(n)$. This kind of compression scheme will be referred to as Method I. If the original signal $x(n)$, at its higher sampling rate, is needed, then Fig. 3.1(b) should be used for re-converting to the higher rate.

The advantage of nonuniform decimation (Sec. IV of [VA88d]) in comparison to this is the extreme simplicity of the compression technique. Building upon the

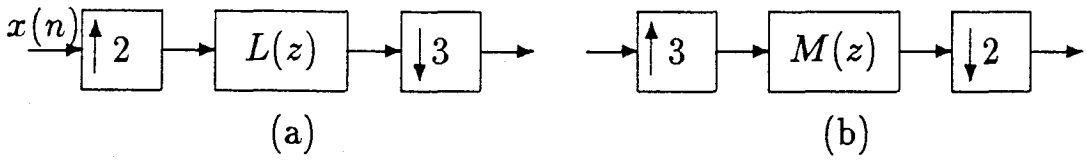

Fig. 3.1. Conventional sampling rate alteration: (a) down-sampling; (b) up-sampling.

previous results in [VA88d], we shall derive a new reconstruction procedure for recovering $x(n)$ from its nonuniformly decimated version, such that the computational complexity is very low. The new reconstruction procedure is applicable to a general class of multi-band bandlimited signals (to be explained later) as well as to the usual lowpass bandlimited signal.

A typical nonuniform decimation of a signal can be performed as follows: divide the time axis n into consecutive intervals of length M and retain L out of M samples in each interval. If this sub-sampling pattern is periodically repeated, the signal $x(n)$ can be recovered from the nonuniformly decimated version, provided $L\pi/M \geq \sigma$. One technique for such reconstruction has been outlined in [VA88d], by formulating the problem as a *multirate filter bank* design problem. To be more specific, it was noted that the nonuniform decimation can be represented by the network of Fig. 3.2, where each box with the down-going arrow represents a uniform M-fold decimator. The problem of reconstructing $x(n)$ was posed as one of designing the filters $F_k(z)$, $0 \leq k \leq L-1$, in Fig. 3.3 so that $\hat{x}(n)$ approximates $x(n)$ as closely as required. It was noted that this problem is mathematically analogous to the problem of designing the M channel QMF bank [SM87a,VA87a], as shown in Fig. 1.2. In traditional QMF applications, the analysis filters are chosen to be good bandpass filters having disjoint passbands and overlapping transition bands so that the whole

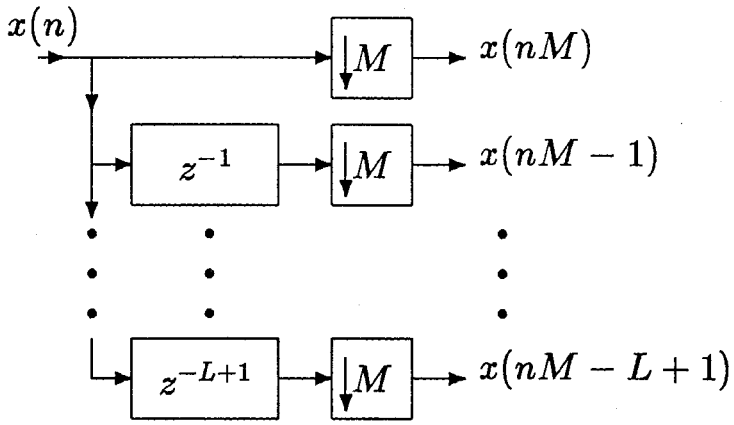

Fig. 3.2. A periodic sub-sampling operation, represented by a multirate filter bank.

frequency range from 0 to 2π are covered.

For the 'nonuniform decimation and reconstruction' application, however, the analysis filters are given by

$$H_k(z) = \begin{cases} z^{-k}, & 0 \le k \le L-1 \\ 0, & L \le k \le M-1 \end{cases} \tag{3.1}$$

and the task is to find a set of synthesis filters $F_k(z)$, (Fig. 3.3), so that aliasing, amplitude distortion and phase distortion are eliminated from $\hat{x}(n)$ under the assumption that $x(n)$ is bandlimited. In [VA88d] it is shown how the filters $F_k(z)$ can be obtained to meet these requirements. The implementation of the synthesis bank in [VA88d] was, however, done in a manner without taking into account the close inter-relationship between the filters $F_k(z)$ in order to minimize cost. For example, a structure for the synthesis banks was presented in Sec. IV of [VA88d] for the special case of $M = 3, L = 2$ (Fig. 17, [VA88d]) and for the $M = 4, L = 3$ case (Fig. 27, [VA88d]) but neither of these structures has the lowest possible com-

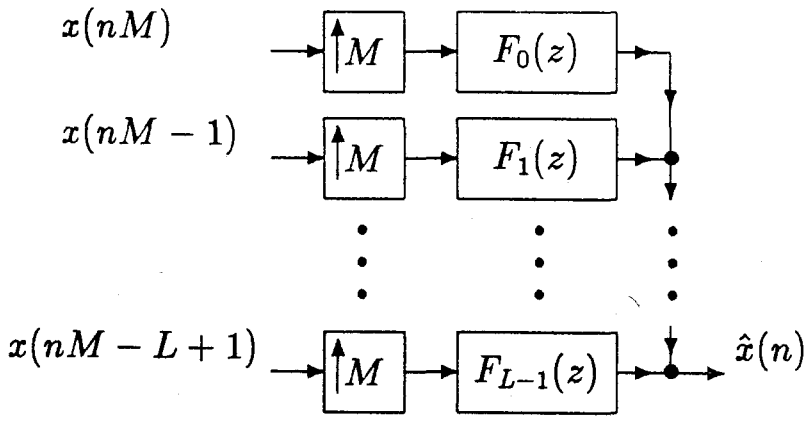

Fig. 3.3. Synthesis filter bank for reconstruction of $x(n)$.

plexity (for a given reconstruction accuracy). Part of the reason for this is that in [VA88d], even though a *filter-bank* approach was used, a polyphase formulation [BE76],[VA87b] was *not* employed. For this same reason, the previous structures presented in Figs. 17 and 24 of [VA88d] (which we shall refer to as Method II) cannot be generalized to arbitrary L and M.

Here, the polyphase approach (to be called Method III) will be introduced in order to obtain a new closed-form solution for the reconstruction filters. There are several advantages to this approach. First, a unified structure is obtained which works for arbitrary L and M. Second, the structure requires significantly fewer computations than the ones in [VA88d], as we shall demonstrate in Sec. 3.4. Lastly, the polyphase structure, when compared to the conventional means of sampling alteration (Fig. 3.1) [CR83], proved to have a lower complexity. This advantage in complexity is achieved only when the polyphase approach is adopted.

The reconstruction of bandlimited signals from nonuniformly spaced samples can be achieved using iterative schemes [SAN63], [WI78]. The filter bank approach

presented here avoids the use of iterations and furthermore, under finite filter length constraints, one has control over the desired amount of alias term attenuation and passband distortion.

In Sec. 3.2, a closed-form filter bank solution is presented for the problem of compressing and reconstructing multi-band bandlimited signals, (i.e., signals which are not necessarily lowpass, but have multiple frequency bands of nonzero energy). A lowpass signal can be thought of as a special case of multi-band signals. The main outcome is the fact that the synthesis filters $F_k(z)$ which result in perfect recovery of $x(n)$ are *multilevel filters,* i.e., filters whose frequency responses are piecewise constants, with each region of constancy having a length of $2\pi/M$. This is an extension of the result in [VA88d].

Since such a piece-constant filter function cannot be implemented in practice, an approximation to the ideal should be made using practical filters such as FIR filters. This gets us into the problem of filter design. In Sec. 3.3, we show how the multilevel specifications reflect into the polyphase components of the filters, thereby a unified design procedure may be developed for these filters. Also the approach of using polyphase components results in an efficient structure for implementing these filters. It will be shown that all the L filters $F_k(z)$, $0 \leq k \leq L-1$, can be obtained by designing a *single lowpass* Mth-band filter [MIN82] and efficiently using its polyphase components. (Even though any Mth-band filter design method may be used for the lowpass filter design [NG88] [SAM88], we have included in Appendix A our own Mth-band filter design method which is used to generate the examples in this thesis.) The cost of the entire synthesis bank is equivalent to the cost of a single Mth-band filter, and this is the key to the improved efficiency of

the new structure. Using the new filter bank structure, we studied in Appendix C the output noise gain due to noise in the subsampled signals. A complete design example is presented in Sec. 3.4 for the $M = 3, L = 2$ case. The complexity of the new structure is compared with a comparable design based on the previous structure in [VA88d]. A comparison with Method I is also included. Finally in Sec. 3.5 we summarize an extension of the 1D results to the case of 2D multi-band signals, where the compression and reconstruction can be achieved by using generalized 2D sampling lattices [ME83] [DU85] [VI88] [AN88].

3.2 Compression and Reconstruction of Multi-band Signals

Let the frequency axis from 0 to 2π be divided into M equal open intervals, each having a length of $2\pi/M$. This division is illustrated in Fig. 3.4. The M intervals are labeled consecutively as: $I_0, I_1, \ldots, I_{M-1}$. Thus we can define I_m as

$$I_m \triangleq \left\{ \omega \left| \frac{2\pi}{M} m < \omega < \frac{2\pi}{M}(m+1) \right. \right\}, \tag{3.2}$$

with $m = 0, 1, \ldots, M-1$. Now consider a band-limited signal $x(n)$ with transform $X(e^{j\omega})$ that occupies only L out of the M frequency intervals. Fig. 3.5 shows an example of such a signal with $L = 4$ and $M = 7$. Next, let us divide the time index

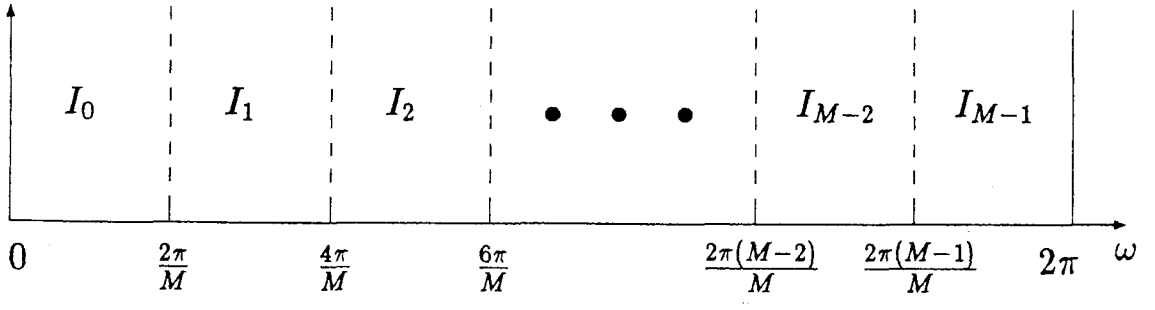

Fig. 3.4. Division of the frequency region $[0, 2\pi]$ into M equal intervals.

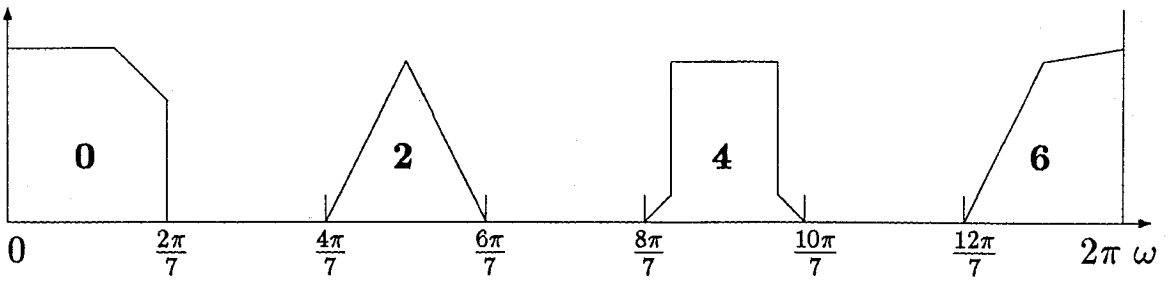

Fig. 3.5. Example of a band-limited spectrum occupying 4 out of 7 intervals.

n into intervals of length M each. Suppose L samples are retained out of M samples in each interval. Let the sub-sampling pattern be the same for each interval. Since $X(e^{j\omega})$ occupies only L out of M intervals, then one might expect that the above sub-sampled signal will still contain enough information to recover $x(n)$. This, as we shall prove, is indeed the case. Such a sub-sampling and reconstruction process can be used for the compression of $x(n)$ and the advantage it has to offer is that the compression scheme is extremely simple.

3.2.1 Retaining consecutive samples

We first consider the case where the samples retained from the original signal are $x(nM), x(nM-1), \ldots x(nM-L+1)$. This is represented in Fig. 3.2. Let the set of integers $\mathcal{L} \triangleq \{l_0, l_1, l_2, \ldots, l_{L-1}\}$ represent all the frequency intervals in which $X(e^{j\omega})$ is nonzero. For example, in Fig. 3.5, the integers $l_0 = 0, l_1 = 2, l_2 = 4$ and $l_3 = 6$. The problem of reconstructing $x(n)$ can be put into the multirate framework as in Fig. 3.3. The reconstructed signal is $\hat{x}(n)$ whose z-transform can be expressed as

$$\hat{X}(z) = \sum_{m=0}^{M-1} X(z\mathcal{W}^m) A_m(z). \tag{3.3}$$

In (3.3), the terms with $m > 0$ represent alias components caused by decimation. The weighting function [VA88d] associated with each alias component is

$$A_m(z) = \frac{1}{M} \sum_{k=0}^{L-1} \mathcal{W}^{-mk} z^{-k} F_k(z). \tag{3.4}$$

To cancel aliasing, one would require that $A_m(z) = 0$ for $m \neq 0$. Once aliasing is cancelled, $T(z) \triangleq A_0(z)$ represents the overall transfer function of the system. For perfect reconstruction, we require $T(z) = 1$ (or more generally $T(z) = cz^{-N}$ with $c \neq 0$). Representing the synthesis filter bank by the vector $\mathbf{f}(z) = (F_0(z)\ F_1(z)\ \ldots\ F_{L-1}(z)\ )^T$, the conditions for alias cancellation can be expressed in matrix form as

$$\begin{pmatrix} A_0(z) \\ A_1(z) \\ \vdots \\ A_{M-1}(z) \end{pmatrix} = \frac{1}{M} \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & \mathcal{W}^{-1} & \ldots & \mathcal{W}^{-(L-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathcal{W}^{-(M-1)} & \ldots & \mathcal{W}^{-(M-1)(L-1)} \end{pmatrix} \mathbf{\Lambda}(z)\mathbf{f}(z) = \begin{pmatrix} T(z) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.5}$$

where $\mathbf{\Lambda}(z)$ is a diagonal matrix of the form

$$\mathbf{\Lambda}(z) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & z^{-1} & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & \ldots & z^{-L+1} \end{pmatrix}. \tag{3.6}$$

In (3.5), the filters $F_k(z)$ are the unknowns to be determined. Since there are more equations than unknowns, so for $T(z) \not\equiv 0$ there might not exist a solution $\mathbf{f}(z)$ that would satisfy (3.5) for all values of z.

Let us examine (3.5) at steady state frequencies $z = e^{j\omega}$. Since $X(e^{j\omega})$ is band-limited to the frequency intervals numbered $\{l_0, l_1, \ldots\ l_{L-1}\}$, not all alias terms are present in each interval. For example in the interval I_p (where p is an integer in the range of 0 to $M-1$), the input spectral terms that are nonzero are $X(e^{j\omega}\mathcal{W}^{p-l_k})$ with $k = 0, \ldots, L-1$. There are L nonzero terms, therefore only L out of M equations in (3.5) need to be satisfied.

Consider the frequency interval I_p where $p \in \mathcal{L}$. For concreteness, let $p = l_q$. As mentioned previously, the nonzero terms are $X(e^{j\omega}\mathcal{W}^{l_q-l_k})$. The weighting function associated with each of these terms is $A_{((l_q-l_k))}(e^{j\omega})$. The notation $((l_q-l_k))$ represents $l_q - l_k$ modulo M. For alias cancellation, we require

$$A_{((l_q-l_k))}(e^{j\omega}) = \begin{cases} 0 & \text{for } 0 \le k < L \text{ and } k \ne q; \\ T(e^{j\omega}) & \text{for } k = q. \end{cases} \tag{3.7}$$

Using (3.5), the L equations in (3.7) can be written as

$$\frac{1}{M}\begin{pmatrix} 1 & \mathcal{W}^{l_0-l_q} & \dots & \mathcal{W}^{(l_0-l_q)(L-1)} \\ 1 & \mathcal{W}^{l_1-l_q} & \dots & \mathcal{W}^{(l_1-l_q)(L-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathcal{W}^{l_{L-1}-l_q} & \dots & \mathcal{W}^{(l_{L-1}-l_q)(L-1)} \end{pmatrix} \mathbf{\Lambda}(e^{j\omega})\mathbf{f}(e^{j\omega}) = \begin{pmatrix} 0 \\ \vdots \\ T(e^{j\omega}) \\ \vdots \\ 0 \end{pmatrix}. \tag{3.8}$$

On the right-hand side of (3.8), the term $T(e^{j\omega})$ appears at the qth entry of the column vector. The left-hand side of (3.8) can be factorized into $\frac{1}{M}\mathbf{U}^{-1}\mathbf{\Lambda}(\mathcal{W}^{l_q})\mathbf{\Lambda}(e^{j\omega})\mathbf{f}(e^{j\omega})$, where $\mathbf{U}^{-1}$ is a constant matrix given by

$$\mathbf{U}^{-1} \triangleq \begin{pmatrix} 1 & \mathcal{W}^{l_0} & \dots & \mathcal{W}^{l_0(L-1)} \\ 1 & \mathcal{W}^{l_1} & \dots & \mathcal{W}^{l_1(L-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathcal{W}^{l_{L-1}} & \dots & \mathcal{W}^{l_{L-1}(L-1)} \end{pmatrix}. \tag{3.9}$$

Notice that $\mathbf{U}^{-1}$ is a Vandermonde matrix. Since $\{l_0, l_1, \dots, l_{L-1}\}$ are all distinct integers within the range 0 to $M-1$, the matrix is nonsingular. By matrix inversion, the solution for (3.8) is found to be

$$\mathbf{\Lambda}(e^{j\omega})\mathbf{f}(e^{j\omega}) = MT(e^{j\omega})\mathbf{\Lambda}(\mathcal{W}^{-l_q})\mathbf{u}_q. \tag{3.10}$$

The vector $\mathbf{u}_q$ is the qth column of $\mathbf{U}$. For perfect reconstruction, set $T(e^{j\omega}) = 1$ in (3.10).

Now for steady state frequencies $\omega \in I_p$ where $p \notin \mathcal{L}$, the transform $X(e^{j\omega})$ is zero within this interval. As a result, we may let $T(e^{j\omega}) = 0$ for $\omega \in I_p$ without

affecting the reconstructed signal, and the alias cancellation condition in (3.5) is easily satisfied by choosing $\mathbf{f}(e^{j\omega}) = \mathbf{0}$.

In summary, the condition for the perfect reconstruction of $x(n)$ is

$$\mathbf{\Lambda}(e^{j\omega})\mathbf{f}(e^{j\omega}) = M\mathbf{b}_p \quad \text{for } \omega \in I_p, \tag{3.11}$$

where $\mathbf{b}_p = \mathbf{0}$ if $p \notin \mathcal{L}$ and $\mathbf{b}_p = \mathbf{\Lambda}(\mathcal{W}^{-p})\mathbf{u}_q$ if $p = l_q \in \mathcal{L}$. In the absence of aliasing, the overall system transfer function $T(e^{j\omega})$ has the following response

$$T(e^{j\omega}) = \begin{cases} 1 & \text{for } \omega \in I_p \text{ where } p \in \mathcal{L}; \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

As seen from (3.11), the frequency response of the filter, $e^{-jk\omega}F_k(e^{j\omega})$, has M bands and within each band the response is a complex constant. Such a piecewise constant function cannot be realized in practice. However, an approximate realization may be achieved using FIR filters. In the next section, we shall derive an approximate realization of (3.11) based on a polyphase approach which yields a structure that is computationally more efficient than the one in [VA88d].

3.2.2 Retaining non-consecutive sub-samples

In Fig. 3.2, the sub-samples that we choose to retain are $x(nM), x(nM-1), \ldots$ $x(nM-L+1)$. Within a period of M samples, these sub-samples are chosen consecutively. Such a choice always allows the missing samples to be reconstructed using a filter bank. However, if one wishes, one can also choose arbitrary sub-samples such as $x(nM-n_0), x(nM-n_1), \ldots x(nM-n_{L-1})$ with n_0, n_1, etc., being distinct integers in the range of 0 to $M-1$. The condition for alias cancellation in (3.5) is replaced by

$$\frac{1}{M}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \mathcal{W}^{-n_0} & \mathcal{W}^{-n_1} & \cdots & \mathcal{W}^{-n_{L-1}} \\ \vdots & \vdots & & \vdots \\ \mathcal{W}^{-(M-1)n_0} & \mathcal{W}^{-(M-1)n_1} & \cdots & \mathcal{W}^{-(M-1)n_{L-1}} \end{pmatrix} \mathbf{\Lambda}'(z)\mathbf{f}(z) = \begin{pmatrix} T(z) \\ \vdots \\ 0 \end{pmatrix}, \tag{3.13}$$

where $\mathbf{\Lambda}'(z)$ is a diagonal matrix with entries: $[\mathbf{\Lambda}'(z)]_{i,i} = z^{-n_i}$. The synthesis filter bank can be derived in a way analogous to the consecutive sub-sampling case, provided that the matrix in (3.9) is replaced by

$$\begin{pmatrix} \mathcal{W}^{l_0 n_0} & \mathcal{W}^{l_0 n_1} & \dots & \mathcal{W}^{l_0 n_{L-1}} \\ \mathcal{W}^{l_1 n_0} & \mathcal{W}^{l_1 n_1} & \dots & \mathcal{W}^{l_1 n_{L-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{W}^{l_{L-1} n_0} & \mathcal{W}^{l_{L-1} n_1} & \dots & \mathcal{W}^{l_{L-1} n_{L-1}} \end{pmatrix}. \tag{3.14}$$

A difficulty is encountered at this point. For an arbitrary set of integers $\{n_0, n_1, \dots n_{L-1}\}$, the matrix in (3.14) is not guaranteed to be nonsingular. If it is singular, then a filter bank type of reconstruction filters may not exist. If we restrict ourselves to consecutive sub-sampling, then a solution always exists.

3.2.3 Error due to the non-bandlimited nature of the signal

In the above analysis, we have assumed that the input signal $x(n)$ is strictly bandlimited. In cases where such an assumption is not satisfied, error is present in the reconstructed signal. We shall analyze the frequency domain error, defined as $E(e^{j\omega}) = \hat{X}(e^{j\omega}) - X(e^{j\omega})$. This error can be expressed as $E(e^{j\omega}) = \Big(A_0(e^{j\omega}) - 1\Big)X(e^{j\omega}) + \sum_{k=1}^{M} A_k(e^{j\omega})X(e^{j\omega}W^k)$. For the case where $\omega \in I_p$ with $p \neq \mathcal{L}$, all the weighting functions $A_k(e^{j\omega})$ are zero, therefore

$$E(e^{j\omega}) = X(e^{j\omega}). \tag{3.15}$$

For the case where $\omega \in I_p$ with $p = l_k$ then the error term is given by

$$E(e^{j\omega}) = \sum_{m=0}^{M-L-1} X(e^{j\omega}W^{l_k - \hat{l}_m})A_{((l_k - \hat{l}_m))}(e^{j\omega}). \tag{3.16}$$

The set $\{\hat{l}_m\}$ with $m = 0, 1, \dots M-L-1$ is defined to be the complement of the set $\{l_k\}$, so $\{\hat{l}_m\} \cup \{l_k\} = \{0, 1, \dots, M-1\}$. Each of the alias term weighting function in (3.16) can be expressed as

$$A_{((l_k - \hat{l}_m))}(e^{j\omega}) = \sum_{i=0}^{L-1} W^{i\hat{l}_m} u_{i,k} \qquad \text{for } \omega \in I_{l_k}. \tag{3.17}$$

For a bandlimited signal, each of the alias terms in (3.16) is strictly zero. Such an equality does not hold if the bandlimited condition is not satisfied. However, if the signal is "almost bandlimited" then one may assume that the alias terms in (3.16) are small in magnitude and so the error term $E(e^{j\omega})$ will remain small in magnitude.

3.3 Multilevel FIR Filters with Adjustable Response Levels

As observed earlier, the filters $e^{-j\omega k}F_k(e^{j\omega})$ are multiple band filters with piecewise constant response. Such filters will be loosely referred to as 'multilevel' filters. We shall first consider the general problem of designing and implementing an FIR multilevel filter. For notational simplicity, $H(z)$ is allowed to be noncausal. The noncausality can be corrected later by adding a sufficient amount of delay. Let $H(z)$ have the following frequency response

$$H(e^{j\omega}) \approx d_p \quad \text{for } \omega \in I_p. \tag{3.18}$$

By decomposing $H(z)$ into its polyphase components, $H(z) = \sum_{k=0}^{M-1} z^{-k}E_k(z^M)$, it can be shown that each of the polyphase terms has the form

$$E_k(e^{j\omega M}) \approx \alpha_k \mathcal{W}^{k/2} e^{j\omega k} \quad \omega \in I_0, \tag{3.19}$$

where α_k's are constants dependent on the response levels d_p. Due to a periodicity of $\frac{2\pi}{M}$, the response of $E_k(e^{j\omega M})$ periodically repeats outside the interval I_0. The approximate equality in (3.19) is proved as follows. Let ω_0 be any frequency point in I_0, then

$$H(e^{j\omega_0}) = \sum_{k=0}^{M-1} e^{-j\omega_0 k}E_k(e^{j\omega_0 M}) \approx d_0, \tag{3.20}$$

and a frequency-shifted version of the above equation can be written as

$$H(e^{j(\omega_0+\frac{2\pi}{M}p)}) = \sum_{k=0}^{M-1} \mathcal{W}^{pk} e^{-j\omega_0 k}E_k(e^{j\omega_0 M}) \approx d_p. \tag{3.21}$$

The equations in (3.21), with $0 \leq p \leq M-1$, give us a system of M equations

$$\mathbf{W} \begin{pmatrix} E_0(e^{j\omega_0}) \\ e^{-j\omega_0} E_1(e^{j\omega_0 M}) \\ \vdots \\ e^{-j\omega_0(M-1)} E_{M-1}(e^{j\omega_0 M}) \end{pmatrix} \approx \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{M-1} \end{pmatrix} \quad \text{for } \omega_0 \in I_0. \tag{3.22}$$

Clearly, d_p's do not depend on the value of ω_0, so on the left-hand side each term $e^{-j\omega_0 k} E_k(e^{j\omega_0 M})$ is approximately equal to a constant. Hence, we can write $E_k(e^{j\omega M})$ as in (3.19). Substituting (3.19) into (3.22), we get an expression for α_k,

$$\mathbf{W} \begin{pmatrix} \alpha_0 \\ \mathcal{W}^{\frac{1}{2}} \alpha_1 \\ \vdots \\ \mathcal{W}^{\frac{(M-1)}{2}} \alpha_{M-1} \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{M-1} \end{pmatrix}. \tag{3.23}$$

By matrix inversion, α_k is found to be

$$\alpha_k = \frac{1}{M} \mathcal{W}^{-k/2} \sum_{p=0}^{M-1} \mathcal{W}^{-kp} d_p. \tag{3.24}$$

Let us designate a new set of filters $G_k(z^M) = \frac{1}{\alpha_k} E_k(z^M)$, then $G_k(z^M)$ satisfies

$$G_k(e^{j\omega M}) \approx \mathcal{W}^{k/2} e^{j\omega k} \quad \text{for } \omega \in I_0. \tag{3.25}$$

If one has a design method for $G_k(z^M)$, then *any* multilevel filter having a desired response of the form (3.18) can be synthesized as

$$H(z) = \sum_{k=0}^{M-1} \alpha_k z^{-k} G_k(z^M) \tag{3.26}$$

with α_k obtained from (3.24). Notice that the ideal solution for $G_0(z^M)$ in (3.25) is $G_0(z^M) = 1$. This means that the impulse response $h(n)$ satisfies $h(Mn) = 0$ for $n \neq 0$. This is commonly called the Nyquist property or the Mth-band property [MIN82]. Thus $H(z)$ is an Mth-band filter. A structure for implementing $H(z)$ is shown in Fig. 3.6(a). This structure can be used in general to implement any Mth-band filter $H(z)$ having a multilevel frequency response as in (3.18). In Fig. 3.6(a), by adjusting the values of α_k's one can change the response levels of the filter at

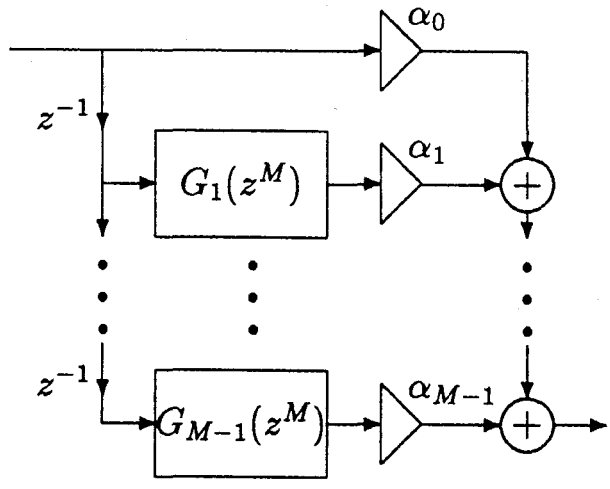

Fig. 3.6(a). Structure for a multiple-band multilevel filter.

will. However, this structure does not allow one to control each band directly. For that purpose, the structure in Fig. 3.6(b) should be used.

The remaining question is how the filters $G_k(z^M)$ can be designed. Since the behavior of $G_k(z^M)$ is independent of the values of d_p, we will first design a very simple prototype filter $P(z)$ which is an Mth-band filter having a lowpass response:

$$P(e^{j\omega}) \approx \begin{cases} 1 & \text{for } \omega \in I_0 \text{ or } \omega \in I_{M-1}; \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

In Appendix A, we included an algorithm for designing such a filter. It is based on a modified version of the algorithm in [PAR72]. It generates an Mth-band FIR filter with symmetric impulse response, and the filter has equiripples in the passband. For simplicity, we will let $P(z)$ be zero-phase (hence noncausal). The Type I polyphase components of $P(z)$ are defined as in (1.8). Due to the Mth-band condition and (3.27), the zeroth polyphase term $P_0(z^M) = \frac{2}{M}$. Since the response of $P(z)$ is a

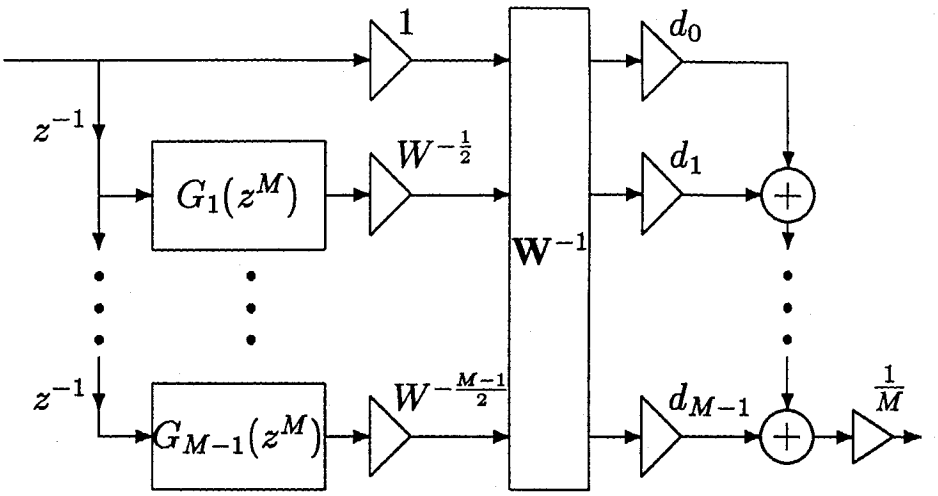

Fig. 3.6(b). Alternative structure for the multilevel filter.

special case of (3.18) with $d_0 = d_{M-1} = 1$ and $d_1 = \ldots = d_{M-2} = 0$, we can write $P(z)$ as

$$P(z) = \beta_0 + \beta_1 z^{-1} G_1(z^M) + \ldots + \beta_{M-1} z^{-M+1} G_{M-1}(z^M), \tag{3.28}$$

where the β_k's are constants. According to the formula in (3.24), β_k is given by $\beta_k = \frac{1}{M}(\mathcal{W}^{-k/2} + \mathcal{W}^{k/2})$. Comparing (3.28) with (1.8), we get $P_k(z^M) = \beta_k G_k(z^M)$. Hence, $G_k(z^M)$ can be obtained as

$$G_k(z^M) = \frac{M}{2\cos(k\pi/M)} P_k(z^M). \tag{3.29}$$

Since the prototype filter has symmetric impulse response, this means the polyphase components come in mirror image pairs $P_k(z^M) = z^M P_{M-k}(z^{-M})$, and so $G_k(z^M)$ $= -z^M G_{M-k}(z^M)$. This relation will be used later.

Design Example 3.1: Since the multilevel filter design is crucial to the reconstruction algorithm, we shall demonstrate it with an example. Suppose we wish

to design and implement a filter $H(z)$ having 5 bands (so $M = 5$) and the filter response is given to be

$$H(e^{j\omega}) \approx \begin{cases} e^{j\pi/2} & \text{for } \omega \in I_0 \\ \frac{1}{4}e^{j\pi/4} & \omega \in I_1 \\ \frac{3}{4} & \omega \in I_2 \\ \frac{1}{4}e^{-j\pi/4} & \omega \in I_3 \\ e^{-j\pi/2} & \text{for } \omega \in I_4 \end{cases} \tag{3.30}$$

In order to make use of Fig. 3.6(a), we need to have a design for $G_1(z^5), \ldots, G_4(z^5)$. We will first design a prototype 5th-band filter $P(z)$ whose response satisfies (3.27)

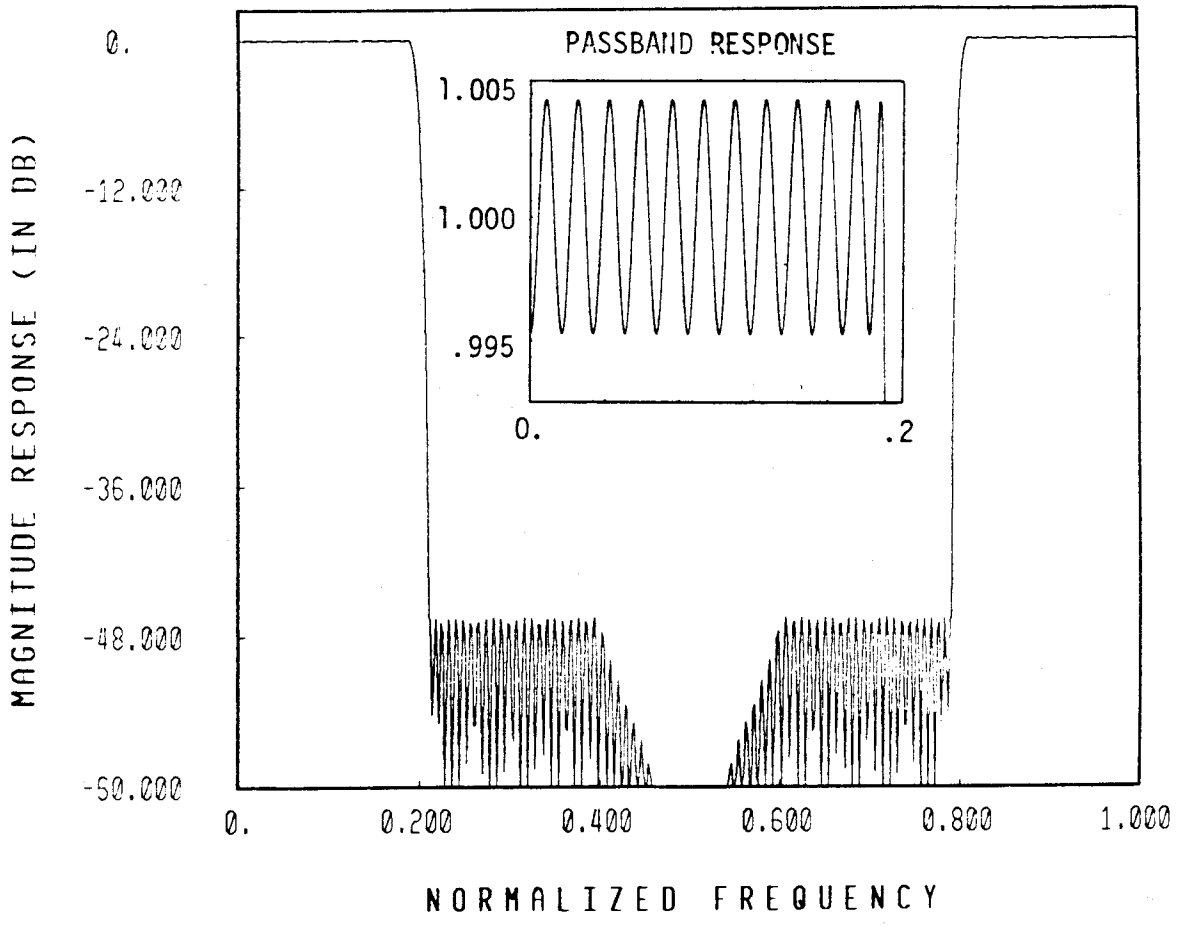

Fig. 3.7. Magnitude response of the prototype filter $P(z)$, used in Design Example 3.1.

with $M = 5$. Using the algorithm in Appendix A, we obtain $P(z)$ whose magnitude response is shown in Fig. 3.7. It has passband edge at 0.38π and stopband edge at 0.42π. The minimum stopband attenuation is $-47dB$, and the peak passband error is 0.005. The order of the filter $N - 1 = 120$. From the polyphase components of $P(z)$, one can obtain $G_k(z^M)$. The phase responses of $G_1(z^M)$ and $G_2(z^M)$ are plotted in Fig. 3.8. This verifies that the filters obtained indeed have the correct phase responses as indicated in (3.25). The filter $H(z)$ is implemented as in Fig. 3.6(a) with α_k computed using (3.24). The magnitude and phase responses of $H(z)$ are shown in Fig. 3.9. As predicted, they match the desired response in (3.30).

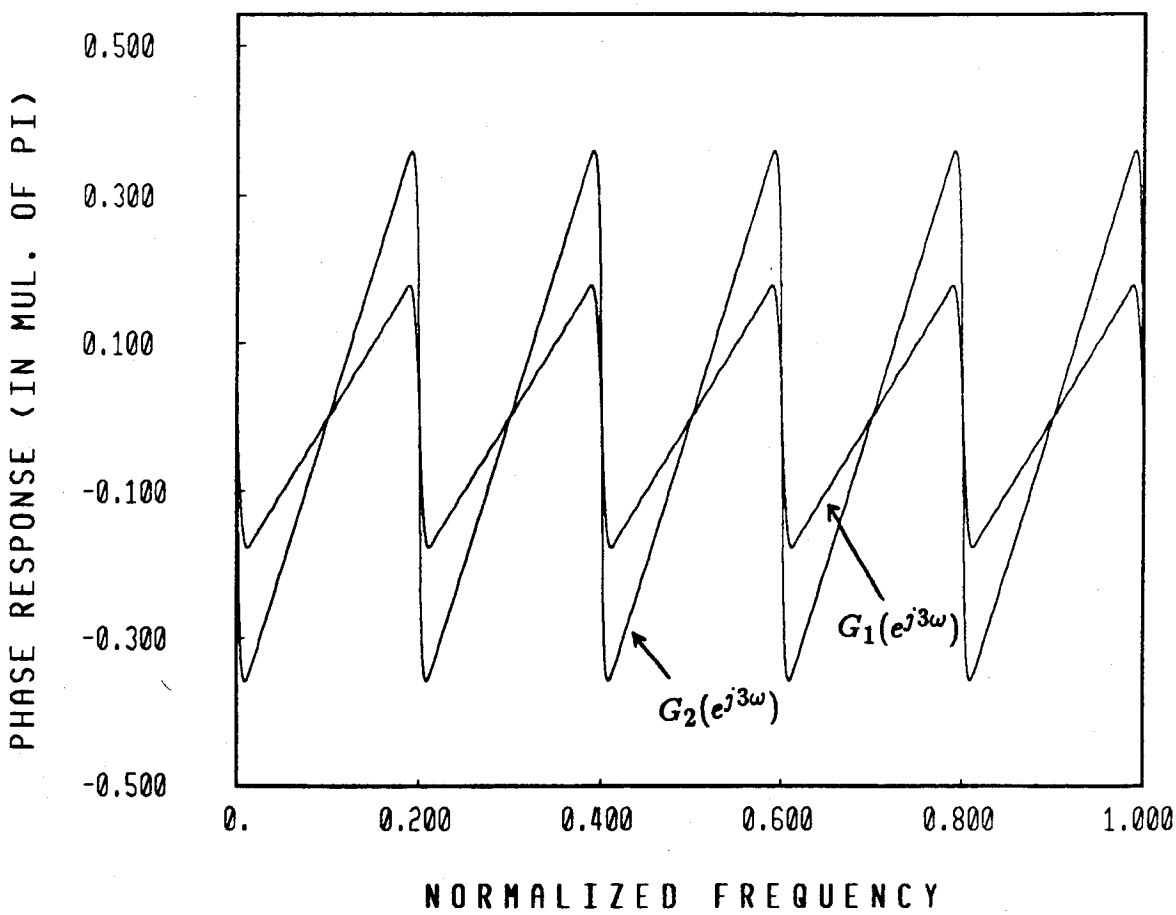

Fig. 3.8. Phase responses of $G_1(e^{j3\omega})$ and $G_2(e^{j3\omega})$.

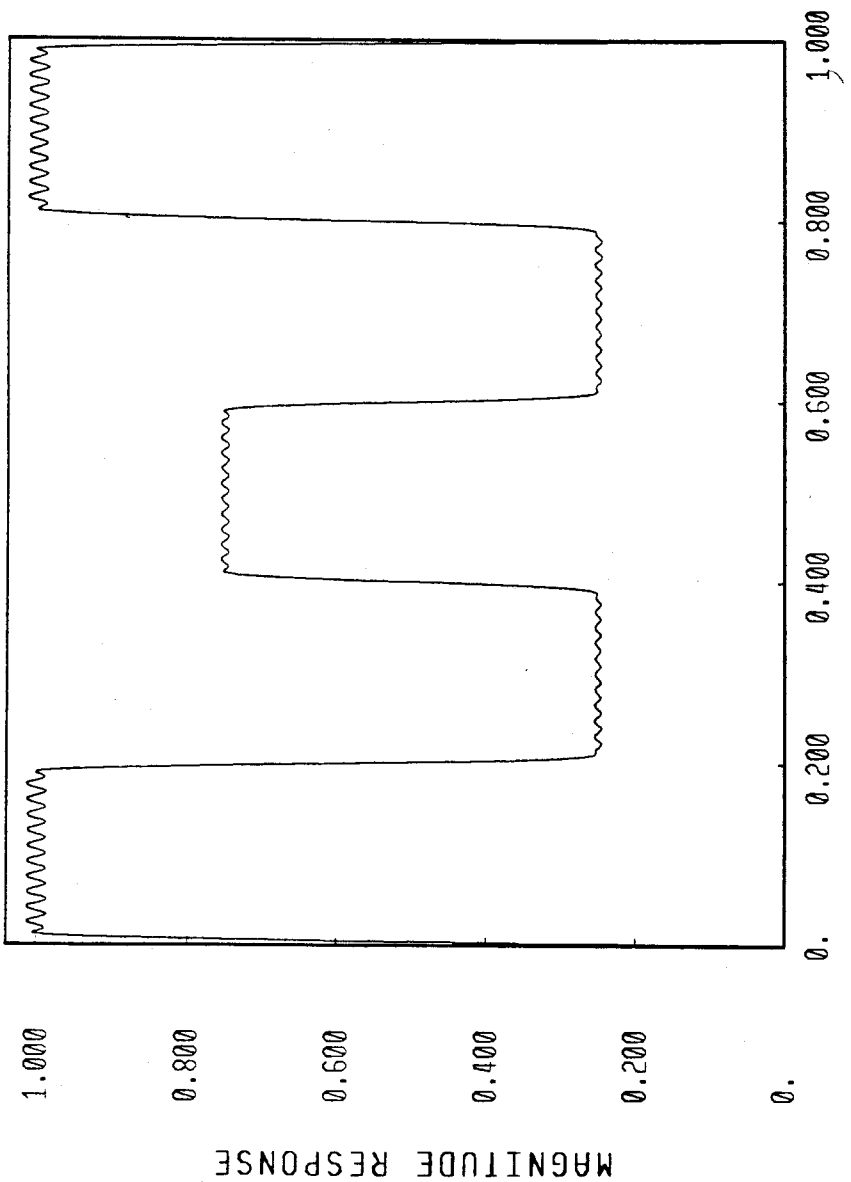

Fig. 3.9(a). Magnitude response of the multilevel filter $H(z)$.

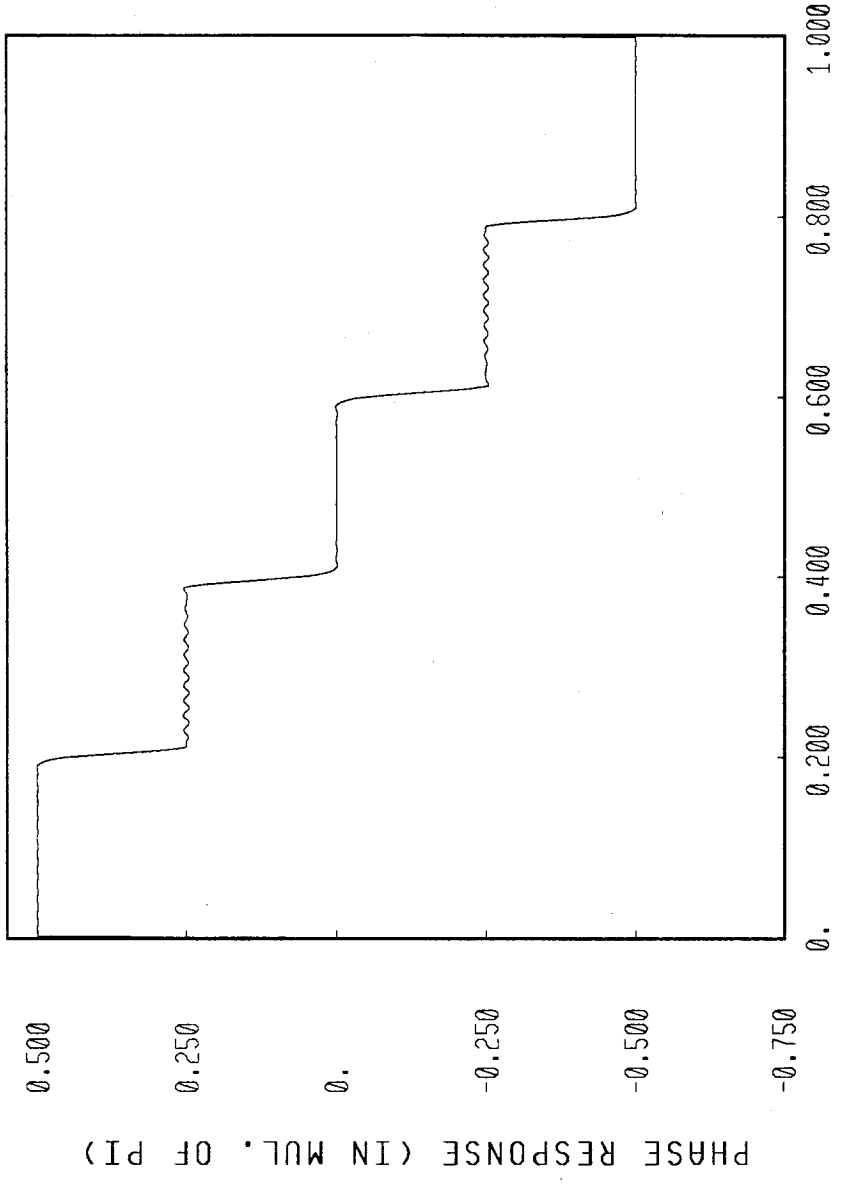

Fig. 3.9(b). Phase response of the multilevel filter $H(z)$.

Now we will return to the synthesis filter bank in (3.11). Each filter $z^{-i}F_i(z)$ has a multilevel response of the form

$$e^{-j\omega i}F_i(e^{j\omega}) = M[\mathbf{b}_p]_i \quad \text{for } \omega \in I_p, \tag{3.31}$$

so it can be implemented as $c_{i,0} + \sum_{k=1}^{M-1} c_{i,k} z^{-k} G_k(z^M)$. The constants $c_{i,k}$ are given by $c_{i,k} = \mathcal{W}^{-k/2} \sum_{p=0}^{M-1} \mathcal{W}^{-kp}[\mathbf{b}_p]_i$. An approximate realization of (3.11) is given by

$$\begin{pmatrix} F_0(z) \\ z^{-1}F_1(z) \\ \vdots \\ z^{-L+1}F_{L-1}(z) \end{pmatrix} = \begin{pmatrix} c_{0,0} & c_{0,1} & \cdots & c_{0,M-1} \\ c_{1,0} & c_{1,1} & \cdots & c_{1,M-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_{L-1,0} & c_{L-1,1} & \cdots & c_{L-1,M-1} \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1}G_1(z^M) \\ \vdots \\ z^{-M+1}G_{M-1}(z^M) \end{pmatrix}. \tag{3.32}$$

Let the matrix in (3.32) be denoted as $\mathbf{C}$ and let $\mathbf{B}$ be a matrix whose pth column is $\mathbf{b}_p$, then the two are related by

$$\mathbf{C} = \mathbf{B}\mathbf{W}^\dagger \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \mathcal{W}^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}^{-\frac{1}{2}(M-1)} \end{pmatrix}. \tag{3.33}$$

The overall synthesis filter bank can now be implemented as

$$\begin{pmatrix} F_0(z) \\ F_1(z) \\ \vdots \\ F_{L-1}(z) \end{pmatrix} = \begin{pmatrix} z^{-L+1} & 0 & \cdots & 0 \\ 0 & z^{-L+2} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{pmatrix} \mathbf{C} \begin{pmatrix} 1 \\ z^{-1}G_1(z^M) \\ \vdots \\ z^{-M+1}G_{M-1}(z^M) \end{pmatrix}. \tag{3.34}$$

This is depicted in Fig. 3.10. The extra delays are added to make the system causal. Notice that we have put the filters $G_k(z^M)$ at the end. Since $G_k(z) = -zG_{M-k}(z^{-1})$, the pair of filters $G_k(z^M)$ and $G_{M-k}(z^M)$ may share the same multipliers.

In Appendix B, we show that $\mathbf{C}$ has L entries that are unity and $L(L-1)$ entries that are zero. Using (3.34) and (A.8) in Appendix B, one can verify that the retained samples of $x(n)$ are not recomputed. In Appendix C, we analyze the problem which occurs when the subsampled signals are corrupted by noise. A

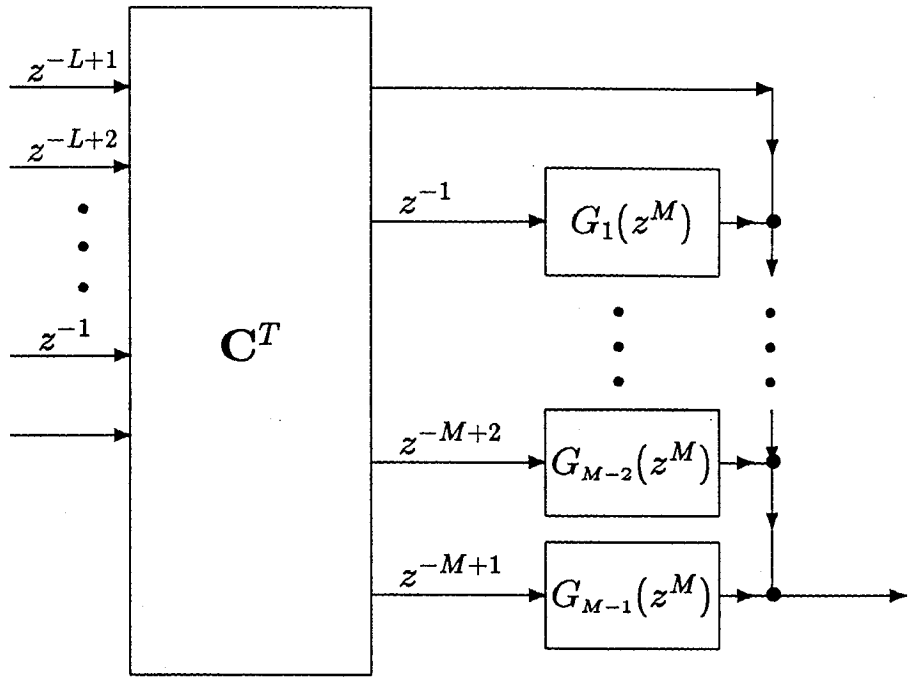

Fig. 3.10. Efficient implementation of the synthesis bank.
($\mathbf{C}$ is a $L \times M$ matrix.)

closed-form expression is derived for the noise gain, but in order to evaluate the noise performance the bandlimits of the signal and the subsampling pattern have to be first given.

3.4 Comparison with Previous Methods for the Case $L = 2$, $M = 3$

We will examine in greater detail the case where the frequency scale is divided into three intervals I_0, I_1 and I_2. Consider a signal $x(n)$ which is bandlimited to $\omega \in I_0$ and $\omega \in I_2$. The band limits of the signal can also be thought of as being $-\frac{2\pi}{3} < \omega < \frac{2\pi}{3}$, so $x(n)$ has the same band limits as in Section IV A of [VA88d]. We should be able to sub-sample $x(n)$ by retaining 2 out of every 3 samples. Using the filter bank in Fig. 3.2 with $M = 3$ and $L = 2$, the sub-samples taken are

$x(3n)$ and $x(3n-1)$. This sub-sampling pattern is the same as in [VA88d]. For the purpose of this section, $\mathcal{W} = e^{-j\frac{2\pi}{3}}$. Substituting the appropriate integers from $\mathcal{L} = \{l_0, l_1\} = \{0, 2\}$ into (3.9), we get

$$\mathbf{U}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & \mathcal{W}^2 \end{pmatrix}. \tag{3.35}$$

This matrix is nonsingular, so the reconstructability of the missing samples is theoretically guaranteed. From $\mathbf{U}$, the matrix $\mathbf{C}$ may be found, and the synthesis filter bank is given by

$$\begin{pmatrix} F_0(z) \\ F_1(z) \end{pmatrix} = \begin{pmatrix} z^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1}G_1(z^3) \\ z^{-2}G_2(z^3) \end{pmatrix}. \tag{3.36}$$

In order to obtain $G_1(z^3)$ and $G_2(z^3)$, we will start with the design of a prototype 3rd-band filter $P(z)$ that satisfies the lowpass response in (3.27) with $M = 3$. After $P(z)$ is obtained, it can be decomposed into its polyphase components

$$P(z) = \frac{2}{3} + z^{-1}P_1(z^3) + z^{-2}P_2(z^3). \tag{3.37}$$

According to (3.29), we let $G_1(z) = 3P_1(z)$ and $G_2(z) = -3P_2(z)$. The structure that implements (3.36) is drawn in Fig. 3.11. The extra delay z^{-K} is added to make the system causal, since originally we took $P(z)$ to be a zero-phase FIR filter.

3.4.1 Design requirements for the filter bank

Because of the FIR approximation in (3.25), two types of errors are introduced into $\hat{x}(n)$. One is due to the imperfection of $T(z)$ which causes a distortion in the original signal. This distortion can be measured in terms of the passband error of $T(e^{j\omega})$. As seen from (3.12), the passband of $T(e^{j\omega})$ is defined to be the frequency intervals where $X(e^{j\omega})$ is nonzero. For the current example, the passband is the

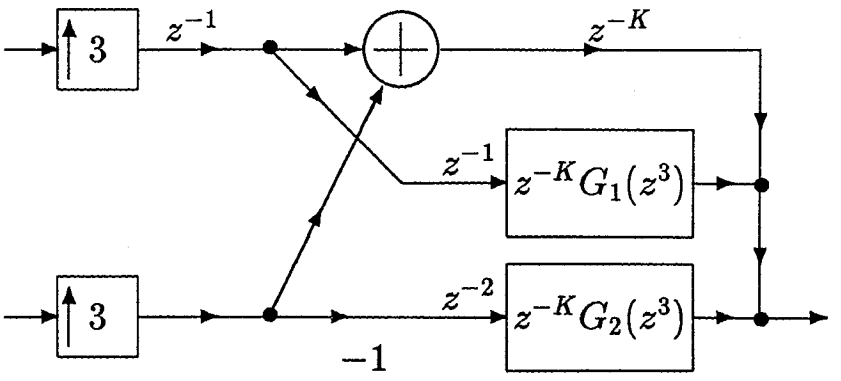

Fig. 3.11. Implementation of the synthesis filter bank using polyphase components of a 3rd-band filter.

union of I_0 and I_2. In practice there will be a transition band in $T(z)$, so its band edges cannot go right upto $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$. Instead the passband edge will be at some frequency $\omega_p < \frac{2\pi}{3}$. This situation is illustrated in Fig. 3.12(a). Let us define a transition bandwidth for $T(z)$ as $\Delta\omega_T = \frac{2\pi}{3} - \omega_p$. What this means in terms of the signal $x(n)$ is that $X(e^{j\omega})$ really should be bandlimited to $-\omega_p < \omega < \omega_p$. So there has to be a guard band between the actual bandlimit of $x(n)$ and $\frac{2\pi}{3}$. The synthesis filter bank is to be designed so that $\Delta\omega_T$ is less than the width of the guard band. Also the passband error for $T(z)$ should be below a prescribed tolerance.

Other sources of error are the alias terms in (3.3) which are not eliminated completely. Each alias term, $X(e^{j\omega}\mathcal{W}^k)$, is attenuated by $A_k(e^{j\omega})$, so a good measure for the error will be the stopband attenuation of $A_k(e^{j\omega})$. The stopband of $A_k(e^{j\omega})$ is defined to be the intervals in which the alias term $X(e^{j\omega}\mathcal{W}^k)$ is nonzero. For the current example, the stopband of $A_1(z)$ is I_0 and I_1. For $A_2(z)$, it is located at I_1 and I_2. Again in practical designs, $A_1(z)$ will have a nonzero transition bandwidth.

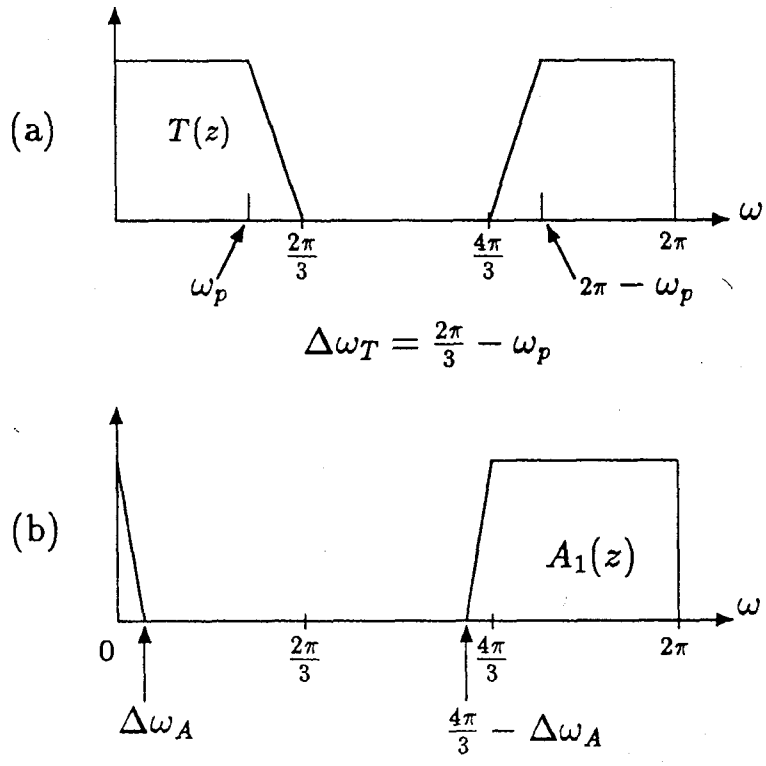

Fig. 3.12. Band edge definitions for the functions $T(z)$ and $A_1(z)$.

If the transition bandwidth of $A_1(z)$ is defined as in Fig. 3.12(b), then for effective alias cancellation $\Delta\omega_A$ is required to be less than the width of the guard band. The minimum stopband attenuation in $A_1(z)$ is also important, for that determines how much of the aliasing error is allowed. Using (3.5) with $M = 3$, we see $|A_2(e^{j\omega})| = |A_1(e^{-j\omega})|$, therefore the stopbands of $A_1(e^{j\omega})$ and $A_2(e^{j\omega})$ have the same attenuation and the transition bandwidth is the same for both functions.

By substituting (3.36) into (3.5) and using the polyphase relations $G_1(z) = 3P_1(z)$ and $G_2(z) = -3P_2(z)$, we get

$$T(z) = z^{-1}P(z) \quad \text{and} \quad A_1(z) = z^{-1}\mathcal{W}\Big(P(z\mathcal{W}) - 1\Big). \tag{3.38}$$

Thus, the passband error of $T(z)$ is the same as the passband error of $P(z)$, and $T(z)$ has the same passband edge as $P(z)$. In (3.38), the function $P(z\mathcal{W}) - 1$ on the unit circle can be derived from a right shifted version of $P(e^{j\omega})$ and then subtracting one from it. Therefore the passband of $P(z)$ becomes the stopband of $A_1(z)$, and the passband error of $P(z)$ equals the stopband error of $A_1(z)$. Also $\Delta\omega_A$ is the same as $\Delta\omega_T$. Suppose the synthesis filter bank is to satisfy the following requirements:

$$\begin{aligned} &\text{peak passband error of } 0.001 \text{ in } T(z), \\ &\text{minimum stopband attenuation of } -60dB \text{ in } A_1(z), \text{ and} \\ &\Delta\omega_T = \Delta\omega_A = 0.034\pi. \end{aligned} \tag{3.39}$$

We will start with the design of a 3rd-band prototype filter $P(z)$ using the algorithm in Appendix A. Its passband edge should be set at $\frac{2\pi}{3} - \Delta\omega_T$. The passband error of $P(z)$ can at most be 0.001. This implies that the stopband attenuation of $A_1(z)$ is at least $-60dB$, so both passband and stopband requirements in (3.39) are satisfied. The lowest order needed for $P(z)$ turns out to be $N - 1 = 94$. From the polyphase components of $P(z)$, the filters $G_1(z)$ and $G_2(z)$ are obtained. The synthesis filter bank is implemented in Fig. 3.11 $z^{-K} = z^{-48}$. The magnitude responses of $T(z)$ and $A_1(z)$ are plotted in Figs. 3.13 and 3.14, respectively.

As a demonstration, the signal $x(n)$ in Fig. 3.15 (a) is bandlimited to $\frac{2\pi}{3}$, this is seen from its Fourier transform $X(e^{j\omega})$ as plotted in Fig. 3.15 (b). We can sub-sample $x(n)$ and perform the reconstruction using the filter bank structure of Fig. 3.11. The reconstructed signal $\hat{x}(n)$ and its transform are plotted in Fig. 3.15 (c) and (d), respectively.

3.4.2 Improvement in efficiency compared to previous methods

Let us label the reconstruction scheme described by (3.36) as Method III. Since

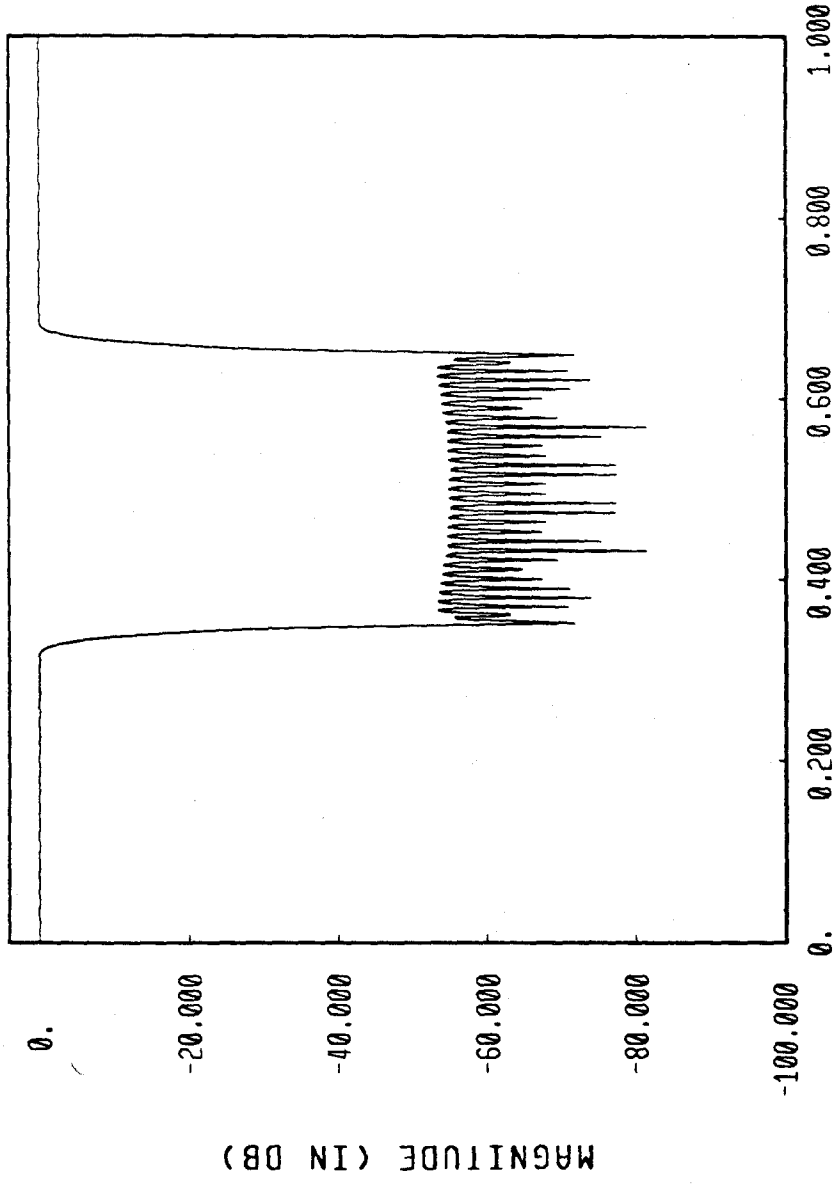

Fig. 3.13. Magnitude of the system transfer function $T(e^{j\omega})$.

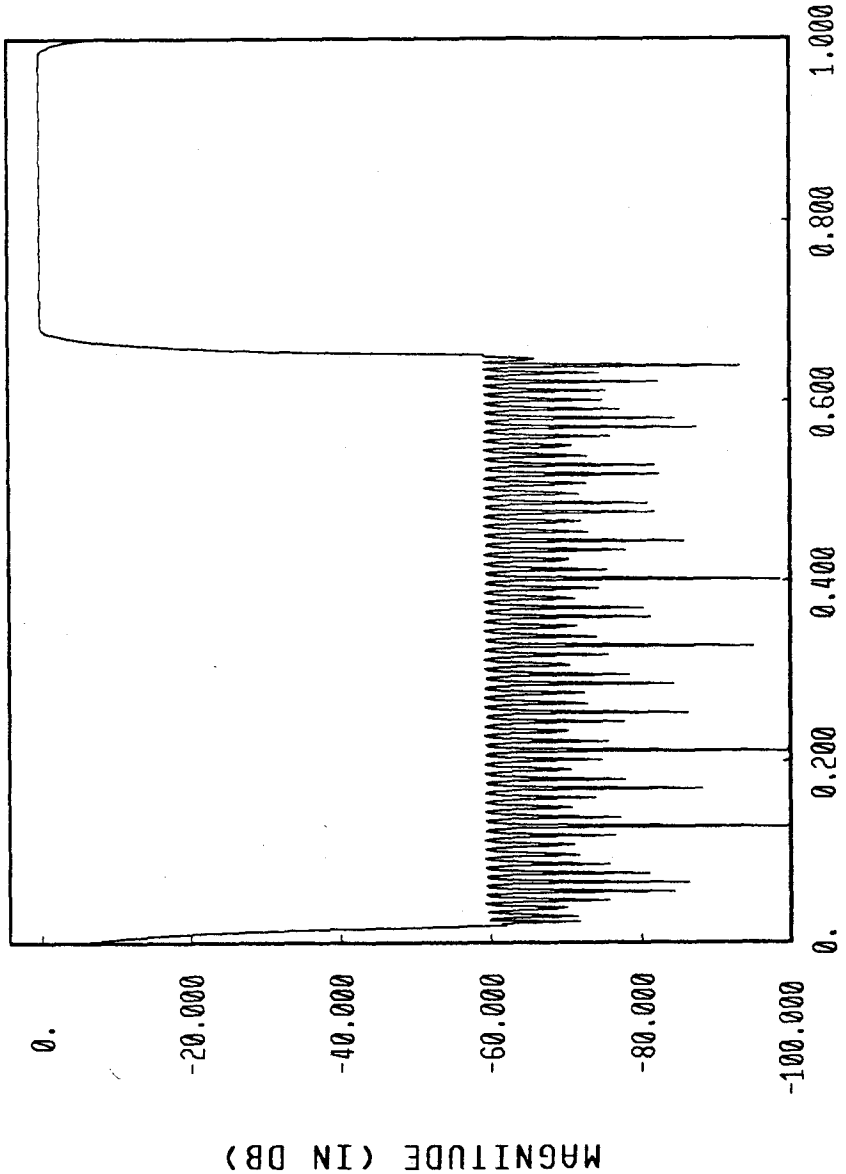

Fig. 3.14. Magnitude of the alias-component weighting function $A_1(e^{j\omega})$.

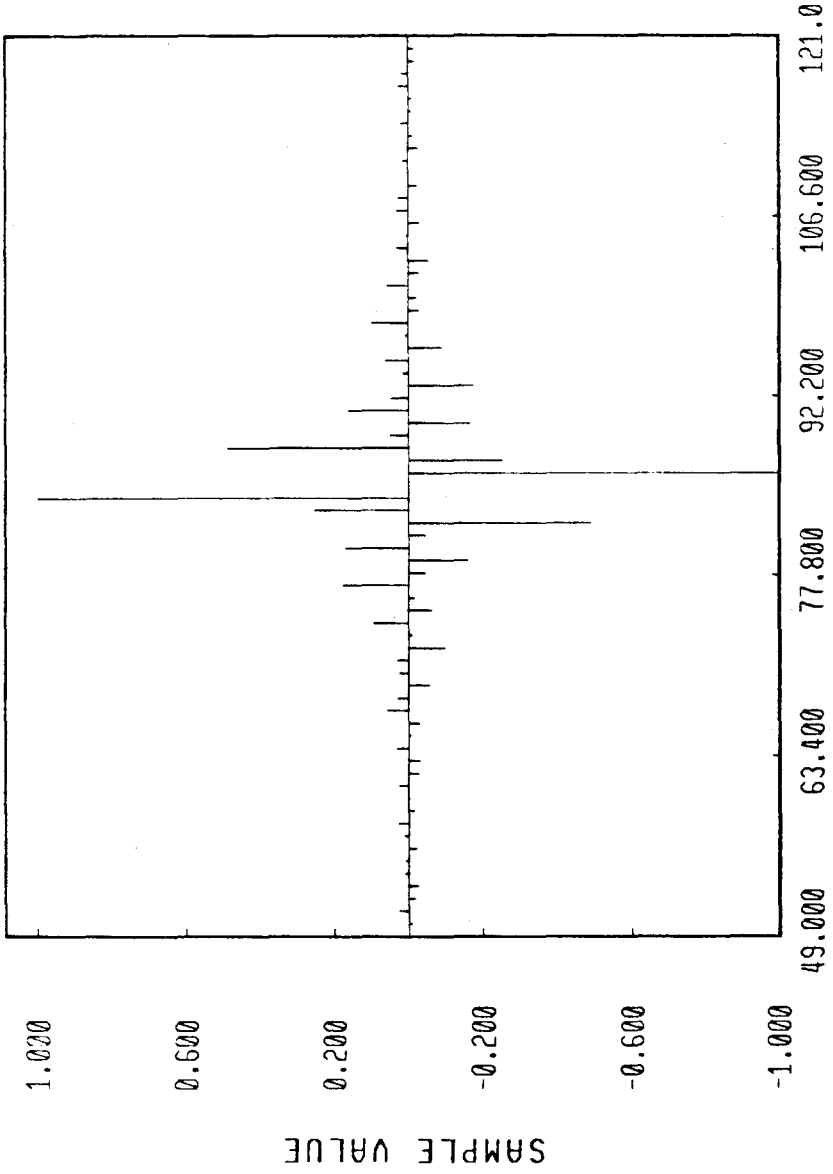

Fig. 3.15(a). The original bandlimited signal $x(n)$.

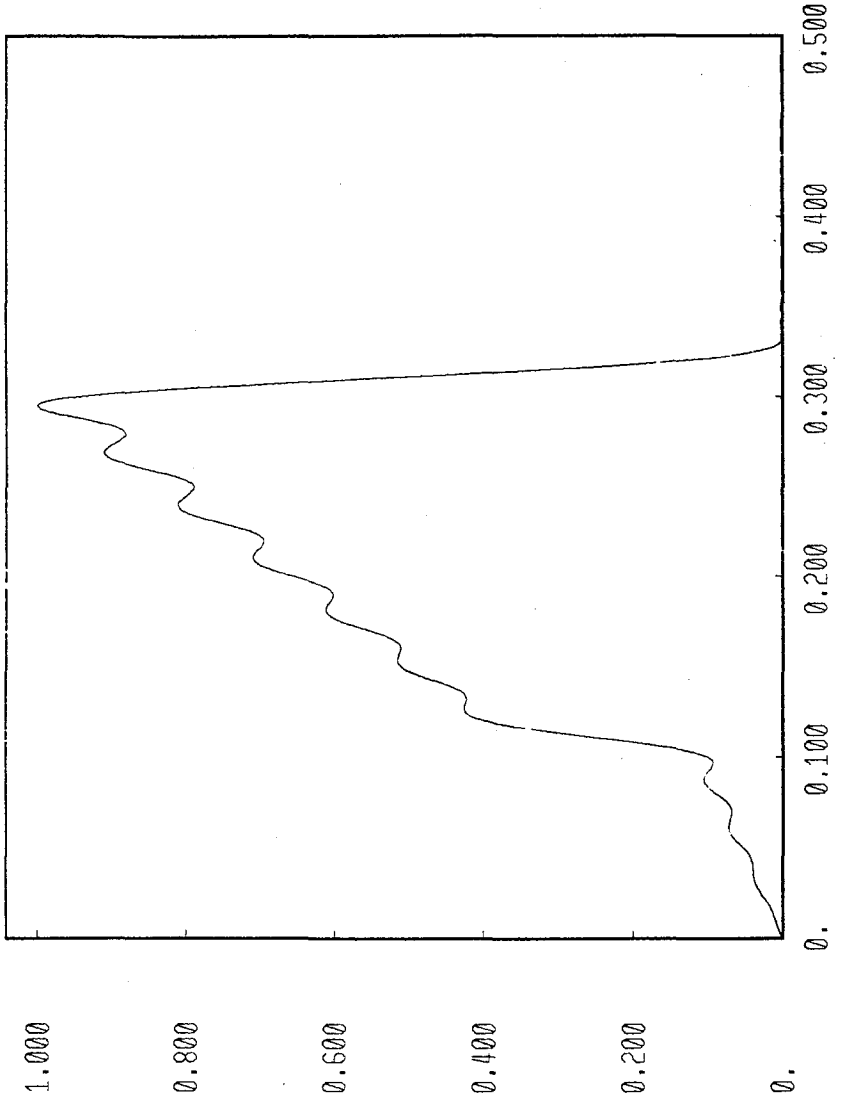

Fig. 3.15(b). Magnitude of the Fourier transform $X(e^{j\omega})$.

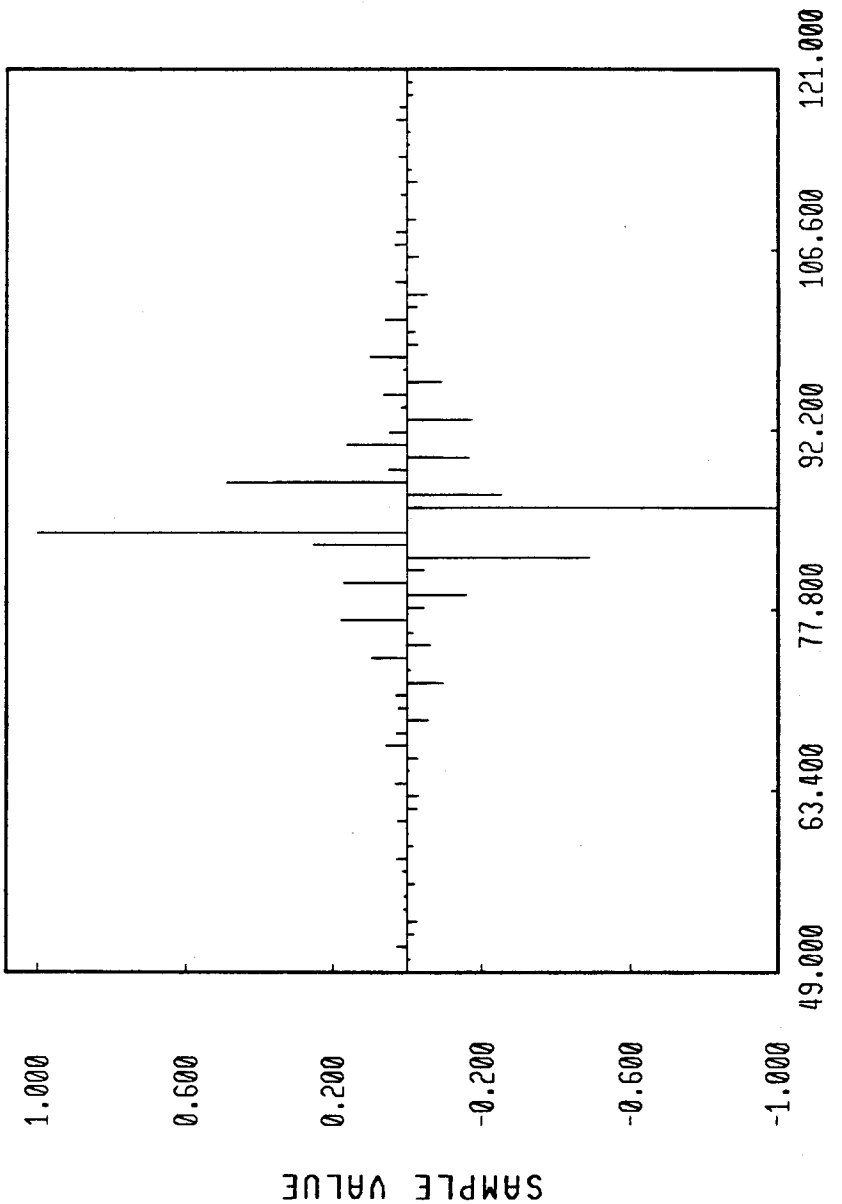

Fig. 3.15(c). The reconstructed signal $\hat{x}(n)$.

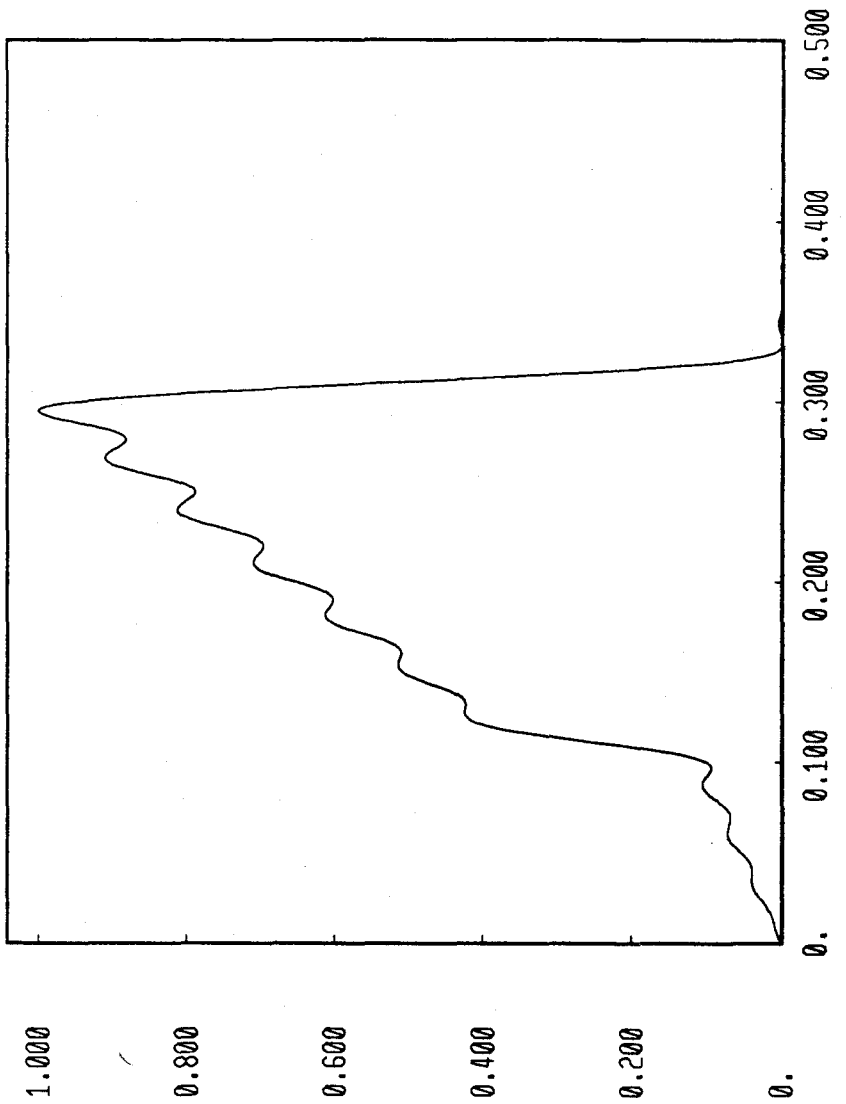

Fig. 3.15(d). The transform of $\hat{x}(n)$.

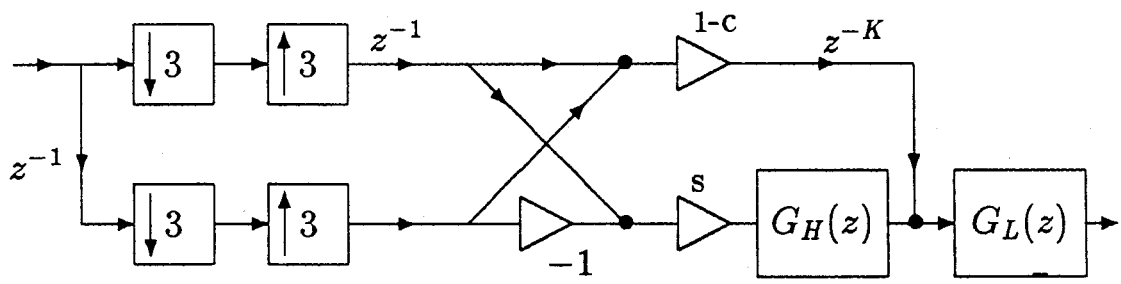

Fig. 3.16. The filter bank structure from Ref. [VA88d].

the order of $P(z)$ is 94, each of the filters, $G_1(z^3)$ and $G_2(z^3)$, requires 32 multipliers for its implementation. But since $P(z)$ has symmetric impulse response, $G_2(z^3) = -z^3 G_1(z^{-3})$. Therefore the two filters may share the same multipliers, and the the total number of multipliers is only 32.

We shall now compare Method III with our previous method derived in Section IV A of [VA88d]. It will be called Method II here. The filter bank structure used for Method II is reproduced in Fig. 3.16. Since the filters given in [VA88d] is designed using a different set of specifications, they cannot be used for comparison here. One can verify that the transfer function and the alias term weighting functions associated with the structure in Fig. 3.16 are given by $T(z) = z^{-1}G_L(z)$, $A_1(z) = \frac{1}{4}z^{-1}G_L(z)\Big[1 - j\sqrt{3} + (\sqrt{3}+j)G_H(z)\Big]$, and $A_2(z) = \frac{1}{4}z^{-1}G_L(z)\Big[1 - j\sqrt{3} - (\sqrt{3}+j)G_H(z)\Big]$. In order for the structure to satisfy the same specifications as in (3.39), the lowpass filter $G_L(z)$ needs to have: a minimum stopband attenuation of $-60dB$; peak passband error of 0.001; passband edge at $\frac{2\pi}{3} - \Delta\omega_T$ and stopband edge at $\frac{2\pi}{3} + \Delta\omega_A$. The requirements for the Hilbert transformer $G_H(z)$ are: peak passband error of 0.002; band edges at $\Delta\omega_A$ and $\pi - \Delta\omega_A$ (chosen for reasons of symmetry

as described in [VA88d]). By designing linear phase FIR filters to satisfy these requirements, the filter order for $G_L(z)$ turns out to be 96 and the filter order of $G_H(z)$ is 98. Due to the linear phase of $G_L(z)$, it requires only 49 multipliers. The Hilbert transformer has anti-symmetric impulse response. Since $|G_H(e^{j\omega})|$ is symmetric with respect to $\pi/2$, the filter can be designed such that about half of its impulse response coefficients are zero. Therefore $G_H(z)$ would require a total of 25 multipliers. Including the two multipliers $1-c$ and s in Fig. 3.16, the total number of multipliers for Method II is 76. In summary, for the same specifications in (3.39) Method III requires 32 multipliers, while Method II needs 78. The saving is by more than a factor of two.

At this point, it will be interesting to compare the complexity of Method III with that of the conventional method of sampling rate alteration [CR83], depicted here as Fig. 3.1(a). The conventional method shall be referred to as Method I. The interpolation filter $L(z)$ is a half-band filter. The specifications in (3.39) was based upon the assumption that the bandlimit of $x(n)$ is at most $\frac{2\pi}{3} - \Delta\omega_T$. After interpolation, the signal at the output of the interpolator will be $X(z^2)$. This means the baseband signal has a bandlimit of $\frac{\pi}{3} - \frac{\Delta\omega_T}{2}$ (this becomes the passband edge for $L(z)$). The image term starts at $\frac{2\pi}{3} + \frac{\Delta\omega_T}{2}$ which defines the stopband edge for $L(z)$. To meet similar requirements as in (3.39), $L(z)$ should have a peak passband error of 0.001 and a minimum stopband attenuation of $-60dB$. A 22nd order linear phase half-band filter is sufficient, and the complexity is only 6 multipliers. The center impulse coefficient is exactly one half, so it is not counted.

Method I has a complexity of 6 multiplications, in order to compress the signal $x(n)$. Method II and III on the other hand, do not involve any compression cost.

It should however be noted that the compressed output of Method I (Fig. 3.1(a)) is a useable signal (because it is a uniformly compressed version) whereas the compressed version in Method II or III is not directly useable without reconverting to the uniform sampling format. A common ground for comparing Method I with Methods II and III can be established as follows: suppose we wish to reconstruct the uniformly sampled signal $x(n)$ at the original oversampled rate. From the above discussion we know that Method II [VA88d] requires 78 multipliers and Method III requires 32 multipliers. For Method I, in order to obtain an approximation to the original $x(n)$, we have to use the non-integer interpolation scheme [CR83] shown in Fig. 3.1(b), where $M(z)$ is a 3rd-band filter [MIN82].

At the input to $M(z)$, the baseband signal has a bandlimit of $\frac{\pi}{3} - \frac{\Delta\omega\pi}{2}$ (which defines the passband for $M(z)$). The filter must also be able to eliminate image terms starting at $\frac{\pi}{3} + \frac{\Delta\omega\pi}{2}$. For the same pass and stopband error specifications as in $L(z)$, the order for $M(z)$ turns out to be 190. Again taking into account the linear phase and 3rd-band property, $M(z)$ has only 65 distinct multipliers. So the cost of implementing Fig. 3.1(a) and (b) combined is 72 multipliers. This is less than Method II. However, Method III is more efficient than both. In conclusion, for applications where the signal needs to be compressed for transmission and then restored to the original rate, the nonuniform sub-sampling approach with polyphase synthesis filter bank gives us a system with lower complexity than the conventional method of Fig. 3.1.

3.4.3 MPU comparisons

Comparisons can also be made in terms of the number of multiplications per unit time (MPU). The time unit is taken to be the time interval between adjacent

samples of $x(n)$. The sampling rate of $x(n)$ is denoted as f_s, so the time unit is $1/f_s$. In Fig. 3.1, one can derive the smallest possible MPU by using the polyphase rearrangement in [HS87]. For Fig. 3.1(a), let the order of the half-band filter $L(z)$ be $N_L - 1$. The filter $L(z)$ can be decomposed into its 6-fold polyphase components as shown in [HS87], and these components are made to operate at the lower sampling rate of $f_s/3$. Using [HS87] and the fact that $L(z)$ has linear phase, the number of MPU's is reduced to about $\frac{N_L+1}{12}$. For our current example, that will be 2 MPU. Suppose that the third-band filter $M(z)$ has order $N_M - 1$, then each of the two non-trivial polyphase components of $M(z)$ has an order of either $\frac{N_M-2}{3}$ or $\frac{N_M-3}{3}$, depending on whichever one is an integer. Due to linear phase, the two polyphase components are mirror images of each other. Again using [HS87] the polyphase components can be made to operate at the lower rate of $f_s/3$. The linear phase property of $M(z)$ can also be exploited in this case. Together the number of MPU's for Fig. 3.1(b) is reduced to about $\frac{1}{3}(J+2)$ where J is the order of the polyphase components. For the current example, $J = 63$ and the number of MPU's is about 22.

In order to achieve the smallest MPU for Method III, the structure in (3.36) has to be re-written as

$$\begin{pmatrix} F_0(z) \\ F_1(z) \end{pmatrix} = \begin{pmatrix} 0 & 1 & G_1(z^3) \\ 1 & 0 & -G_2(z^3) \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1} \\ z^{-2} \end{pmatrix} = \mathbf{R}^T(z^3) \begin{pmatrix} 1 \\ z^{-1} \\ z^{-2} \end{pmatrix}. \tag{3.40}$$

The matrix $\mathbf{R}^T(z^3)$ is the polyphase component matrix of $\mathbf{f}(z)$. It is a function of z^3 only, so it can be interchanged with the interpolators provided z^3 is replaced with z. This is illustrated in Fig. 3.17. The exchange is necessary so that the filters in $\mathbf{R}(z)$ are operating at the lower rate. As discussed previously, $G_1(z)$ and $G_2(z)$ are mirror image polynomials so they may share the same multipliers. Hence, we only

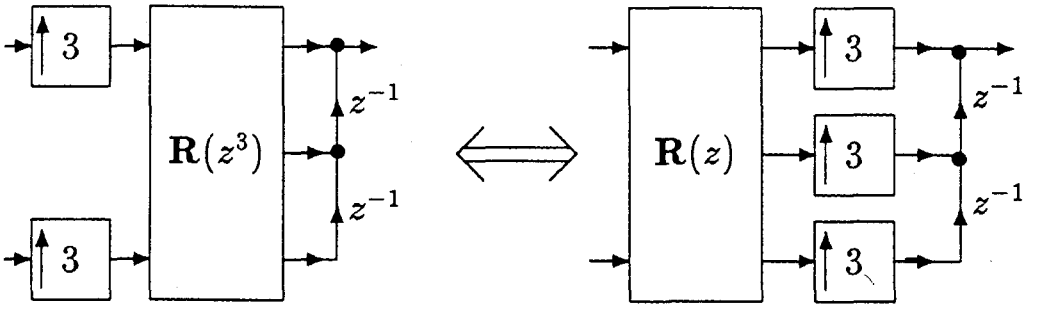

Fig. 3.17. Polyphase representation of the synthesis filter bank.

need 32 multipliers in implementing $\mathbf{R}(z)$. The MPU for Fig. 3.17 is $\frac{32}{3}$ which is about 11. In terms of MPU, Method III is still more efficient than the combination of Fig. 3.1(a) and (b). Here Method II is omitted, since it is worse than the other two methods.

For several other combinations of L and M, we have compared Methods I and III in terms of the numbers of MPU's. The reconstruction accuracy is specified to be: a minimum attenuation of $50dB$ for the alias terms and a peak passband error of 0.003. The transition bandwidth, $\Delta\omega_T$, is taken to be 0.03π. The results of the comparison are listed in Table 3.1. We would like to point out that, except for the case of $L = 2$, Method III requires a smaller number of MPU's than Method I. Furthermore, for a fixed M the number of MPU's decreases as L increases. This can be explained by the fact that for larger L we are keeping a larger proportion of the total number of samples, as a result the average amount of computation that needs to be performed is smaller.

	$L=2$		$L=4$		$L=6$		$L=8$	
Method	I	III	I	III	I	III	I	III
$M=5$	13.6	14.4	40.8	6.4				
$M=7$	10.3	15.4	30.6	16.0	51.4	5.1		
$M=9$	8.3	14.2	25.0	20.0	41.6	14.6	58.3	4.4

Table 3.1. MPU comparisons for various L and M.

3.5 Extension to Two-dimensions

The above reconstruction scheme can be extended to two-dimensional multi-band signals. We will consider the generalized decimation using integer lattice [ME83] [DU85] [VI88] [AN88]. For the 2D decimator in Fig. 3.18, the matrix $\mathbf{D}$ is a non-singular integer matrix whose entries are denoted by [VI88]

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}. \tag{3.41}$$

Let $\mathcal{Z}$ be the set of all integers. The matrix $\mathbf{D}$ in effect defines a sublattice Λ_D which consists of all the points $(n,\hat{n})$ in $\mathcal{Z}^2$ such that $(n,\hat{n})^T = \mathbf{D}(m,\hat{m})^T$ for some $(m,\hat{m}) \in \mathcal{Z}^2$. A coset of Λ_D is an integer shifted version of the lattice. Over the entire plane $\mathcal{Z}^2$, there are M distinct cosets where M is the magnitude of the determinant of $\mathbf{D}$. A set of polyphase shift vectors is any set of M integer

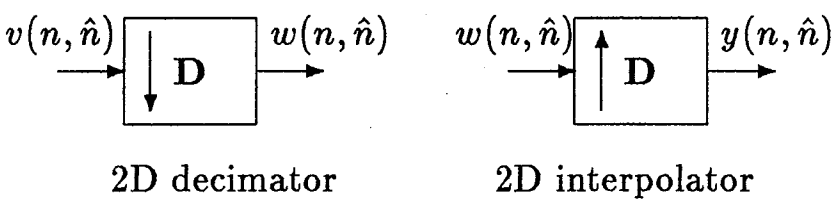

Fig. 3.18. Lattice decimator and interpolator for 2D signals.

vectors which will generate all the distinct cosets of Λ_D. For a given $\mathbf{D}$, the set of polyphase shift vectors is not unique. A polyphase representation for 2D systems may be defined as

$$P(z_1, z_2) = \sum_{i=0}^{M-1} z_1^{-n_i} z_2^{-\hat{n}_i} P_i(z_1^{d_{11}} z_2^{d_{21}}, z_1^{d_{12}} z_2^{d_{22}}), \tag{3.42}$$

where $(n_i, \hat{n}_i)$ is a set of polyphase shift vectors for $\mathbf{D}$. For rectangular sub-sampling, one can simply set $d_{12} = d_{21} = 0$.

In Fig. 3.18, 2D decimation is described by the input/output relation $w(n, \hat{n}) = v(d_{11}n + d_{12}\hat{n}, d_{21}n + d_{22}\hat{n})$. In terms of z-transforms,

$$W(z_1, z_2) \tag{3.43}$$
$$= \frac{1}{M} \sum_{m=0}^{M-1} V(z_1^{d_{22}/M} z_2^{-d_{21}/M} \mathcal{W}^{d_{22}t_m - d_{21}\hat{t}_m}, z_1^{-d_{12}/M} z_2^{d_{11}/M} \mathcal{W}^{-d_{12}t_m + d_{11}\hat{t}_m}).$$

The set of integer vectors, $(t_m, \hat{t}_m)$, is a complete set of polyphase shift vectors for $\mathbf{D}^T$. The interpolation process is given by

$$y(n, \hat{n}) = \begin{cases} w(\frac{d_{22}n - d_{12}\hat{n}}{M}, \frac{-d_{21}n + d_{11}\hat{n}}{M}) & \text{if } (n, \hat{n}) \in \Lambda_D; \\ 0 & \text{otherwise.} \end{cases} \tag{3.44}$$

And in the transform domain,

$$Y(z_1, z_2) = W(z_1^{d_{11}} z_2^{d_{21}}, z_1^{d_{12}} z_2^{d_{22}}). \tag{3.45}$$

Combining (3.43) and (3.45), the overall effect of decimation followed by interpolation is

$$Y(z_1, z_2) = \frac{1}{M} \sum_{m=0}^{M-1} V(z_1 \mathcal{W}^{d_{22}t_m - d_{21}\hat{t}_m}, z_2 \mathcal{W}^{-d_{12}t_m + d_{11}\hat{t}_m}). \tag{3.46}$$

If we pick the shift vector $(t_0, \hat{t}_0)$ to be $(0, 0)$, then in the above summation the $m = 0$ term is the original signal $V(z_1, z_2)$, while the rest are alias terms.

As an example, for the following decimation matrix

$$\mathbf{D} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \tag{3.47}$$

the samples on the 2D plane that are retained are indicated by circles in Fig. 3.19. This forms the sublattice Λ_D. Besides the sublattice, there are 3 distinct cosets which are generated by shifting Λ_D by the following polyphase shift vectors: $(1,0)$, $(1,1)$ and $(2,1)$. Together with the null vector, (0,0), they constitute a complete set of polyphase shift vectors for $\mathbf{D}$. In order to find the locations of the alias terms in (3.46), we need to have a complete set of polyphase shift vectors for $\mathbf{D}^T$. A particular set is given by $(t_m, \hat{t}_m) = \{(0,0)\ ,\ (0,1)\ ,\ (1,1)\ ,\ (1,2)\}$. Substituting $(t_m, \hat{t}_m)$ into (3.46), we see that the first alias term $(m = 1)$ is the original spectrum shifted by $(0, +\pi)$ in the two-dimensional frequency plane. This is illustrated in Fig. 3.20(a) where the diamond shapes represent the original spectrum and its alias terms. The second alias term $(m = 2)$ comes from the original sprectrum shifted by $(\pi\ ,\ \frac{\pi}{2})$, and the third term has a frequency shift of $(\pi\ ,\ \frac{3\pi}{2})$.

Consider again the general form of $\mathbf{D}$ in (3.41). A closer examination of (3.45) reveals that $W(e^{j\omega_1}, e^{j\omega_2})$ is map unto $Y(e^{j(d_{22}\omega_1 - d_{21}\omega_2)/M}, e^{j(-d_{12}\omega_1 + d_{11}\omega_2)/M})$. So the interpolation process in effect maps the content of the frequency plane $0 \leq \omega_1 \leq 2\pi$, $0 \leq \omega_2 \leq 2\pi$ unto a smaller parallelogram whose corners are given by the four points [AN88]:

$$(0,0);\quad \frac{2\pi}{M}(d_{22}, -d_{12});\quad \frac{2\pi}{M}(-d_{21}, d_{11});\quad \frac{2\pi}{M}(d_{22} - d_{21}, -d_{12} + d_{11}). \tag{3.48}$$

We will let the parallelogram defined above be our baseband region I_0, then the other bands are shifted versions of I_0. As seen from (3.46), the band I_m (with $m = 1, \ldots, M-1$) can be thought of as the baseband I_0 frequency-shifted by the

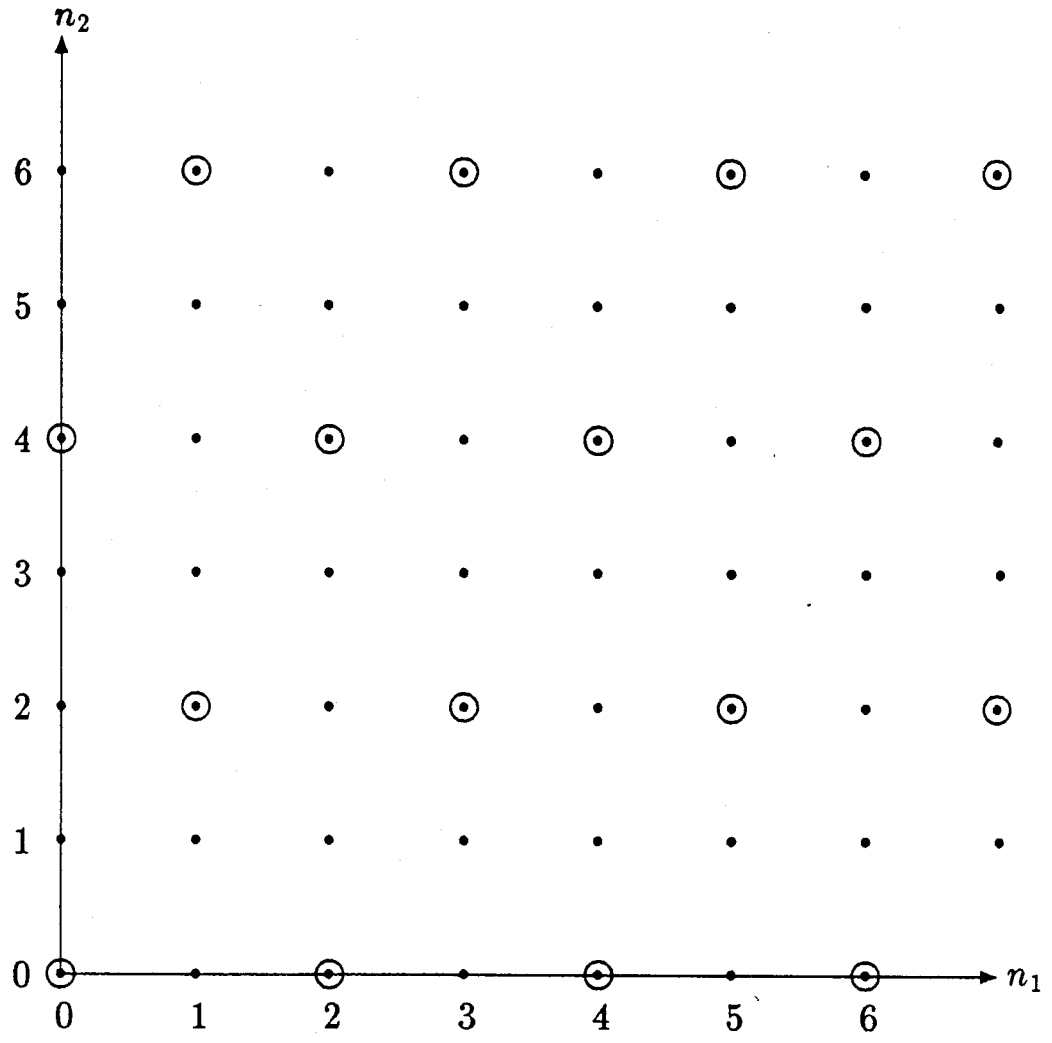

Fig. 3.19. The samples retained by the decimation matrix $\mathbf{D} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$.

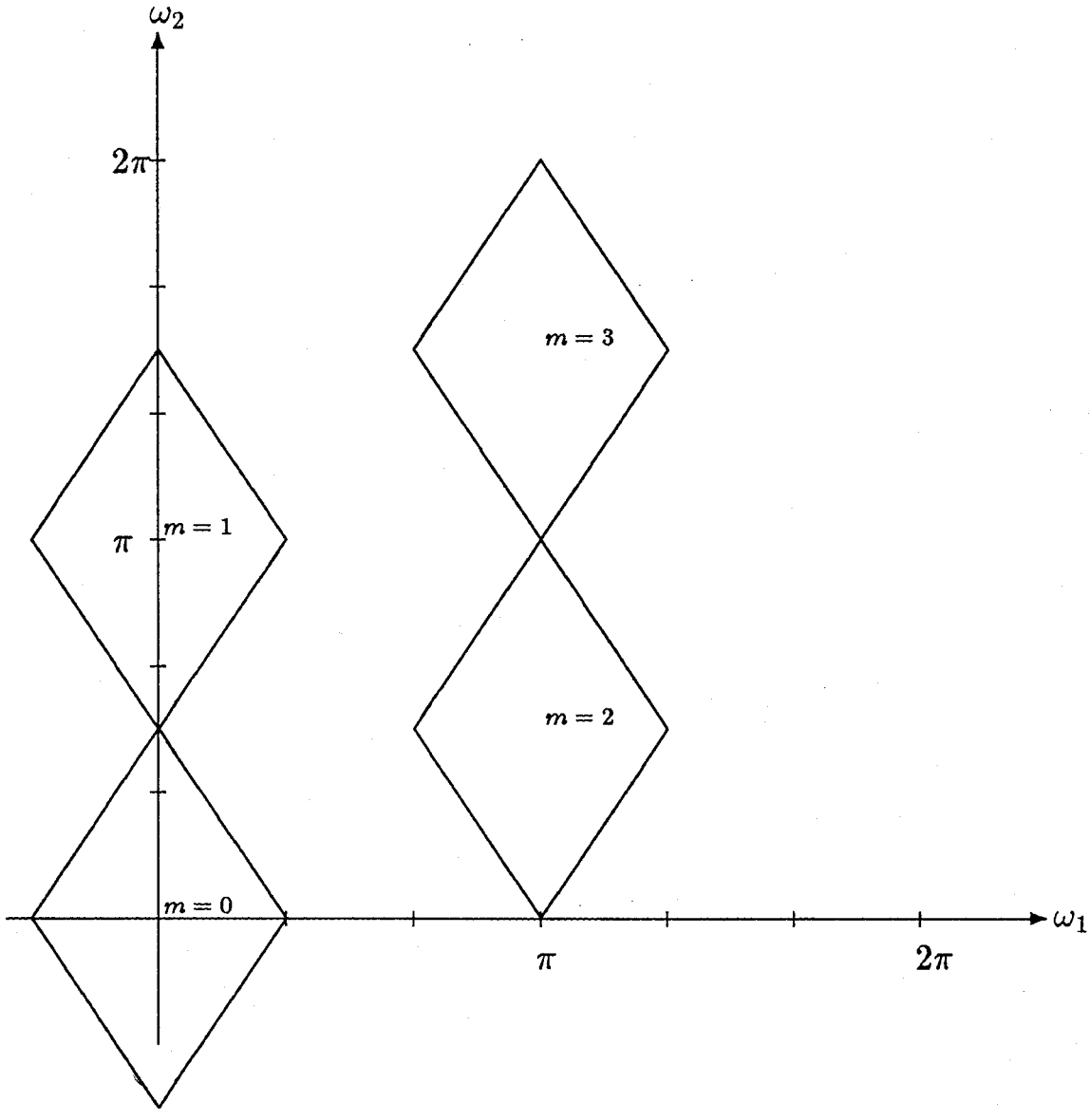

Fig. 3.20(a). Locations of the original signal spectrum and its alias versions.

vector $\frac{2\pi}{M}(a_m, \hat{a}_m)$ where $a_m = d_{22}t_m - d_{21}\hat{t}_m$ and $\hat{a}_m = -d_{12}t_m + d_{11}\hat{t}_m$. (Notice that for a given decimation lattice, there are many possible ways of partitioning the 2D frequency plane and for each partitioning the baseband region is defined differently. For example, using the decimation matrix in (3.47) two ways of partitioning the frequency plane are shown in Fig. 3.20(b). The following derivation of the reconstruction filter bank holds irrespective of the way in which the frequency plane is partitioned.)

For a multi-band signal $x(n, \hat{n})$, suppose its transform $X(e^{j\omega_1}, e^{j\omega_2})$ is nonzero in only L out of a total of M bands, then we can sub-sample the signal by taking L samples out of M without any loss of information. Let the L bands for which $X(e^{j\omega_1}, e^{j\omega_2})$ is nonzero be represented by the set $\mathcal{L} = \{l_0, \ldots l_{L-1}\}$, and let the vectors $(n_i, \hat{n}_i)$ with $i = 0, 1, \ldots, M-1$ be a complete set of polyphase shift vectors for the sampling matrix $\mathbf{D}$, then the L sub-samples can be represented by $x(d_{11}n + d_{12}\hat{n} - n_i, d_{21}n + d_{22}\hat{n} - \hat{n}_i)$ with $i = 0, 1, \ldots\ L-1$. The sub-sampling operation can be implemented in terms of 2D decimators and delay elements, $z_1^{-n_i} z_2^{-\hat{n}_i}$. The reconstruction procedure is analogous to the 1D case (Fig. 3.3) except 2D interpolators and synthesis filters $\hat{F}_k(z_1, z_2)$ are used. The conditions for alias cancellation can be summarized as

$$\frac{1}{M}\hat{\mathbf{V}}\hat{\mathbf{\Lambda}}(z_1, z_2)\hat{\mathbf{f}}(z_1, z_2) = \begin{pmatrix} \hat{T}(z_1, z_2) \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.49}$$

The vector $\hat{\mathbf{f}}(z_1, z_2) = (\ \hat{F}_0(z_1, z_2)\ \hat{F}_1(z_1, z_2)\ \ldots\ \hat{F}_{L-1}(z_1, z_2)\ )^T$ is the 2D synthesis filter bank. Similar to (3.13), $\hat{\mathbf{\Lambda}}(z_1, z_2)$ is a diagonal matrix and $\hat{\mathbf{V}}$ is a $M \times L$ constant matrix. They are defined as follows:

$$[\hat{\mathbf{\Lambda}}(z_1, z_2)]_{i,i} = z_1^{-n_i} z_2^{-\hat{n}_i}, \qquad [\hat{\mathbf{V}}]_{p,i} = \mathcal{W}^{-a_p n_i} \mathcal{W}^{-\hat{a}_p \hat{n}_i}. \tag{3.50}$$

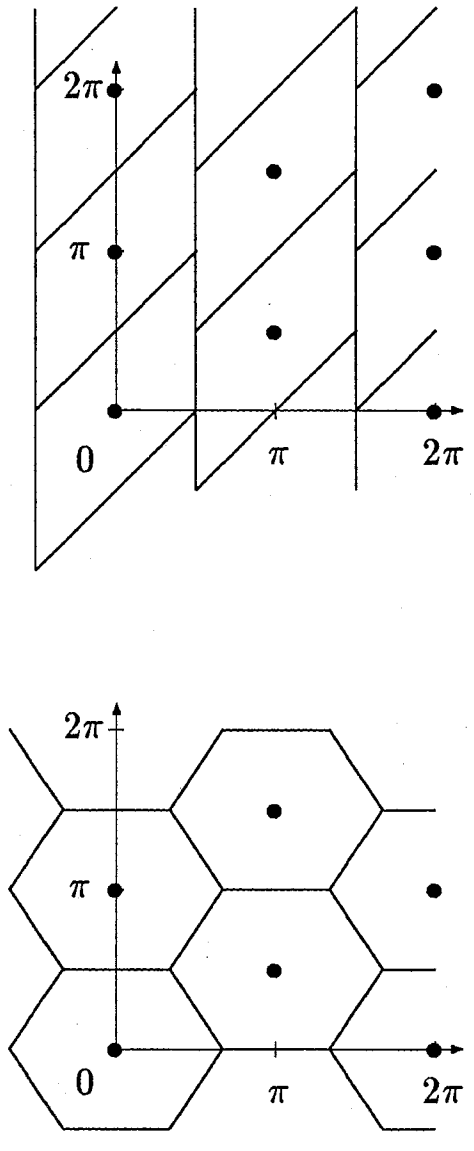

Fig. 3.20(b). Two ways of partitioning the frequency plane.

As in the 1D case, a solution for (3.49) may not exist for all z_1, z_2. However, we can take advantage of the fact that $X(e^{j\omega_1}, e^{j\omega_2})$ is zero in some frequency regions. For any frequency region I_p, there will be only L nonzero spectral terms. They are $X(z_1 \mathcal{W}^{(a_p - a_{l_k})}, z_2 \mathcal{W}^{(\hat{a}_p - \hat{a}_{l_k})})$ where $k = 0, 1, \ldots, L-1$. Thus only L equations in (3.49) need to be satisfied. The existence of a solution depends upon the non-singularity of a $L \times L$ matrix $\hat{\mathbf{Y}}$ whose (k,i)th entry is given by

$$[\hat{\mathbf{Y}}]_{k,i} = \mathcal{W}^{a_{l_k} n_i} \mathcal{W}^{\hat{a}_{l_k} \hat{n}_i}. \tag{3.51}$$

This is the 2D extension of the matrix in (3.14). One has to be careful here with regards to which set of sub-samples to pick. For an arbitrary set of sub-samples, the matrix $\hat{\mathbf{Y}}$ is not guaranteed to be nonsingular. Assuming that it is nonsingular and denoting the qth column of $\hat{\mathbf{Y}}^{-1}$ by $\hat{\mathbf{u}}_q$, the solution for the synthesis filter bank can be expressed as

$$\hat{\mathbf{\Lambda}}(e^{j\omega_1}, e^{j\omega_2}) \hat{\mathbf{f}}(e^{j\omega_1}, e^{j\omega_2}) = M \mathbf{b}_p \quad \text{for} \quad (\omega_1, \omega_2) \in I_p, \tag{3.52}$$

with $\mathbf{b}_p = \mathbf{0}$ when $p \notin \mathcal{L}$, and $\mathbf{b}_p = \hat{\mathbf{\Lambda}}(\mathcal{W}^{-a_p}, \mathcal{W}^{-\hat{a}_p}) \hat{\mathbf{u}}_q$ when $p = l_q \in \mathcal{L}$. Using the polyphase approach, the 2D filter bank can be implemented as in a fashion that is analogous to (3.34)

$$\hat{\mathbf{f}}(z_1, z_2) = \hat{\mathbf{\Lambda}}(z_1^{-1}, z_2^{-1}) \hat{\mathbf{C}} \hat{\mathbf{g}}(z_1, z_2). \tag{3.53}$$

The vector $\hat{\mathbf{g}}(z_1, z_2)$ has M entries. Each entry is a polyphase component filter of the form $z_1^{-n_i} z_2^{-\hat{n}_i} \hat{G}_i(z_1^{d_{11}} z_2^{d_{21}}, z_1^{d_{12}} z_2^{d_{22}})$. The filter $\hat{G}_i(z_1, z_2)$ has the following frequency response:

$$\hat{G}_i(e^{j\omega_1}, e^{j\omega_2}) \approx e^{j\omega_1 (n_i d_{22} - \hat{n}_i d_{12})/M} e^{j\omega_2 (-n_i d_{21} + \hat{n}_i d_{11})/M}. \tag{3.54}$$

The filter $\hat{G}_i(z_1, z_2)$ can be designed as a separable filter, and the design procedure for 1D is applicable here. However, the overall filter bank is not separable.

3.6 Finite Length Extrapolation of Bandlimited Signals

The problem of sub-sampling and reconstruction of bandlimited signals can be thought of as being related to the problem of bandlimited signal extrapolation. In the sub-sampling and reconstruction case, the number of samples that we kept (or that we are given at the receiver end) is a fixed fraction of the total missing samples. While, for signal extrapolation the number of samples given is only an infinitesimal fraction of the missing samples. As the length of the extrapolated signal becomes infinitely long, that fraction goes to zero. In the sub-sampling case, the missing samples can always in theory be reconstructed exactly, so the solution is unique. For discrete-time signal extrapolation, there does not exist an unique bandlimited extrapolation. In fact for a given set of samples and a fixed band limit, there exists infinitely many possible extrapolations that would satisfy the bandlimited condition. However, if we consider only finite length signals, then the two problems become very similar for in both cases the number of missing samples and the number of known samples are finite.

For continuous-time bandlimited signals, it is possible to extrapolate the signal from a given finite segment of it. Various methods exists for such an extrapolation [GE74] [PAP75] [PAP77a] [HO81]. The attempt to do the same for discrete-time bandlimited signals shows that the extrapolation in such a case is not unique. An unique solution exists if one imposes the constraint that the extrapolated signal should have minimum norm [JAI81] [SU84].

In any practical algorithm, the extrapolated signal can only be finite in length. If we consider only finite length signals, then the problem of sub-sampling and reconstruction becomes very similar to the one of extrapolation. A general problem

may be stated as follows: for a given band limit and a finite time interval, several samples are given and one is to find the missing samples so that the overall signal is "as bandlimited as possible."

One method of finding the missing samples will be presented. The overall signal generated will have the minimum out-of-band energy among all permissible extrapolations. This is done by directly minimizing the out-of-band energy of the signal. With a bandlimit of $\frac{L}{M}\pi$ and by taking the number of the observation samples to be L/M of the total number of samples, this method can be applied to the subsampling and reconstruction of lowpass bandlimited signals. Such an application is explored in the next section.

Let $y(n)$ be an unknown bandlimited sequence with only K samples available to us. For notational simplicity, we shall assume $y(n)$ to be a real sequence. Its Fourier transform $Y(e^{j\omega})$ is zero for ω outside the frequency range $|\omega| < \omega_s$. Let the locations of the given samples be denoted by an integer function $f(k)$ with $k = 0, 1, \ldots, K-1$. In other words, the observed samples of $y(n)$ are

$$y(f(0)), \quad y(f(1)), \quad \ldots, \quad y(f(K-1)). \tag{3.55}$$

Here, we assume $f(0) < f(1) < \ldots < f(K-1)$. The problem is to extrapolate these samples into a longer yet finite sequence $x(n)$ of length $(2N+1)$. The length of $x(n)$ should be long enough so that all the observed samples of $y(n)$ lie within it. The extrapolated sequence $x(n)$ matches the observed data, so

$$x(f(k)) = y(f(k)) \qquad \text{for} \quad k = 0, 1, \ldots, K-1. \tag{3.56}$$

A further requirement on $x(n)$ is that it should be almost bandlimited to $|\omega| \leq \omega_s$,

just as $\bar{y}(n)$ is. One way of obtaining a nearly bandlimited extrapolation is to minimize the out-of-band energy of $x(n)$ subject to the constraint in (3.56).

For convenience, we define a complementary function $f'(l)$ with $l = 0, 1, \ldots 2N - K$ so as to indicate the positions of the missing samples. For example, if $N = 5$, $K = 3$ and $f(k) = \{-1, 0, 1\}$, then the function $f'(l)$ will be $f'(l) = \{ -5, -4, -3, -2, 2, 3, 4, 5\}$.

The out-of-band (or stopband) energy of the extrapolated sequence is given by

$$E_s = \frac{1}{2\pi} \int_{\omega_s}^{2\pi-\omega_s} \left| \sum_{n=-N}^{N} x(n) e^{-j\omega n} \right|^2 d\omega \tag{3.57}$$

Denoting the sequence $x(n)$ by a vector $\mathbf{x} = (\ x(-N)\ x(-N+1)\ \ldots\ x(N)\ )^T$, one can rewrite (3.57) as

$$E_s = \mathbf{x}^T \mathbf{Q} \mathbf{x} \tag{3.58}$$

where the $(2N+1) \times (2N+1)$ matrix $\mathbf{Q}$ is defined to be

$$[\mathbf{Q}]_{k,l} = \frac{1}{2\pi} \int_{\omega_s}^{2\pi-\omega_s} e^{j(k-l)\omega} d\omega. \tag{3.59}$$

The above integral can be carried out analytically, and a closed-form expression for $\mathbf{Q}$ is

$$[\mathbf{Q}]_{k,l} = \begin{cases} 1 - \frac{\omega_s}{\pi} & \text{if } k = l; \\ -\frac{\sin(k-l)\omega_s}{(k-l)\pi} & \text{otherwise.} \end{cases} \tag{3.60}$$

In the extrapolation problem, certain samples of $x(n)$ are fixed, namely the observed samples $y(f(k))$. Now by separating the observed samples and the samples to be extrapolated, we can write (3.58) as

$$E_s = \mathbf{y}_0^T \mathbf{Q}_0 \mathbf{y}_0 + 2\mathbf{y}_0^T \mathbf{Q}_1 \mathbf{x}_1 + \mathbf{x}_1^T \mathbf{Q}_2 \mathbf{x}_1. \tag{3.61}$$

where $\mathbf{y}_0$ is a K-dimensional vector containing the observed samples, i.e., $[\mathbf{y}_0]_k = x(f(k)) = y(f(k))$. Similarly, $\mathbf{x}_1$ is of dimension $2N+1-K$ containing the extrapolated samples with $[\mathbf{x}_1]_l = x(f'(l))$. The matrices $\mathbf{Q}_i$ are given by

$$[\mathbf{Q}_0]_{k,l} = [\mathbf{Q}]_{f(k)+N,f(l)+N}, \tag{3.62a}$$

$$[\mathbf{Q}_1]_{k,l} = [\mathbf{Q}]_{f(k)+N,f'(l)+N}, \tag{3.62b}$$

$$[\mathbf{Q}_2]_{k,l} = [\mathbf{Q}]_{f'(k)+N,f'(l)+N}. \tag{3.62c}$$

The dimension of each matrix is defined appropriately according to (3.61). Notice that $\mathbf{Q}_2$ is a symmetric positive definite matrix. The symmetry is due to the symmetry of $\mathbf{Q}$ itself. And since $\mathbf{x}_1^T\mathbf{Q}_2\mathbf{x}_1$ represents the stopband energy of a finite length sequence, i.e.,

$$\frac{1}{2\pi}\int_{\omega_s}^{2\pi-\omega_s}\left|\sum_{l=0}^{2N-K} x(f'(l))e^{-j\omega f'(l)}\right|^2 d\omega, \tag{3.63}$$

this means $\mathbf{x}_1^T\mathbf{Q}_2\mathbf{x}_1 > 0$ for any non-zero $\mathbf{x}_1$. Hence the positive definite nature of $\mathbf{Q}_2$ is established. The inequality is strict because a finite length nonzero sequence cannot be bandlimited. Due to Parseval's relation and (3.63), we also have $\mathbf{x}_1^T\mathbf{Q}_2\mathbf{x}_1 < \mathbf{x}_1^T\mathbf{x}_1$ for any non-zero $\mathbf{x}_1$.

In (3.61) the vector $\mathbf{y}_0$ is fixed, and the only variables are contained in $\mathbf{x}_1$. Thus, we seek to minimize the stopband energy E_s over the $(2N+1-K)$-dimensional vecotr space of $\mathbf{x}_1$. Differentiating E_s with respect to $\mathbf{x}_1$

$$\text{grad}\ \ E_s = 2\mathbf{Q}_1^T\mathbf{y}_0 + 2\mathbf{Q}_2\mathbf{x}_1, \tag{3.64}$$

shows that the minimum energy is obtained by

$$\mathbf{x}_{1_{\min}} = -\mathbf{Q}_2^{-1}\mathbf{Q}_1^T\mathbf{y}_0. \tag{3.65}$$

The above unconstraint minimization of E_s sometimes leads to extrapolations whose norm is much larger than the norm of the observed data. This is specially true when $\mathbf{Q}_2$ becomes ill-conditioned. In conventional bandlimited extrapolation [PAP75] (where the given data are in the middle and one tries to extrapolate on the two ends), the problem has been noted to be ill-conditioned. We shall demonstrate this numerically by calculating the condition number of the $\mathbf{Q}_2$ matrix, $\lambda_{\max}/\lambda_{\min}$, as a function of the number of missing samples. Let us denote the number of missing samples as L and the number of observed samples as K. Figure. 3.21 shows on a logarithmic scale the reciprocal of the condtion number as a function of L, and curves for the following values of K are plotted: $K = 10, 20, 30, \ldots, 60$. The bandlimit ω_s is chosen to be $\pi/2$. We see that the condition number of $\mathbf{Q}_2$ grows very fast as the number of missing samples increases. For a fixed number of missing samples L, increasing the number of observations does not help very much in lowering the condition number. For $K \geq 40$, the condition number is almost an exponential function of L. The same observation can be made for the case when the bandlimit is reduced to $\pi/3$ (Fig. 3.22) except that the growth rate is slower than before. The pattern remains as ω_s is reduced to $\pi/5$ (Fig. 3.23).

As the matrix becomes ill-conditioned, the extrapolated portion of the sequence can have large amplitudes. A common way to combat this phenomenon is to constrain the norm of the extrapolated portion, this was done for the ideal infinite length case [JAI81]. For a finite length sequence, to constrain the extrapolated portion of the sequence one can include the norm of $\mathbf{x}_1$ in the objective function to be minimized, e.g.,

$$\min_{\mathbf{x}_1} \left\{ E_s + \alpha\, \mathbf{x}_1^T \mathbf{x}_1 \right\}. \tag{3.66}$$

The weighting factor α is chosen to be greater or equal to zero. The optimum solution becomes

$$\mathbf{x}_{1_\alpha} = -(\mathbf{Q}_2 + \alpha\mathbf{I})^{-1}\mathbf{Q}_1^T\mathbf{y}_0. \tag{3.67}$$

This solution has the property that it yields the lowest stopband energy among all $\mathbf{x}_1$ vectors with the norm $\mathbf{x}_1^T\mathbf{x}_1 = c$, where the quantity $c = \mathbf{x}_{1_\alpha}^T\mathbf{x}_{1_\alpha}$. The matrix $(\mathbf{Q}_2 + \alpha\mathbf{I})$ is guaranteed to be non-singular for $\alpha > -\lambda_{\min}$ where $\lambda_{\min}$ is the smallest eigenvalue of $\mathbf{Q}_2$.

As a demonstration, let only 15 data points be given from an unknown sequence $y(n)$ as shown in Fig. 3.24. In this case, $K = 15$ and $f(k) = k - 7$ with $0 \leq k \leq 14$. We wish to obtain an extrapolation $x(n)$ of length 55 which is bandlimited to $|\omega| < \pi/3$. The sequence should be centered around $x(0)$. Therefore $N = 27$ and

$$f'(l) = \begin{cases} l - 27 & \text{for } 0 \leq l \leq 19 \\ l - 12 & \text{for } 20 \leq l \leq 39. \end{cases} \tag{3.68}$$

Applying (3.67) with $\alpha = 0.01$, the resulting extrapolated sequence $x(n)$ is computed and it is shown in Fig. 3.25. By examining the Fourier transform of $x(n)$ (Fig. 3.26), one can verify that $x(n)$ is almost bandlimited to $|\omega| < \pi/3$. The peak amplitude of $X(e^{j\omega})$ in the stopband is less than $-30dB$.

3.7 Signal Reconstruction from Sub-samples

The solution in (3.65) can be applied to the reconstruction of finite length sequences which are almost bandlimited. For example, suppose $y(n)$ is a known sequence (see Fig. 3.28) extending from $y(-40)$ to $y(40)$ and it is nearly band-limited to $|\omega| < 2\pi/5$. Then one might expect that the sequence can be compressed by a factor of $5/2$. One compression technique would be to divide the time axis into

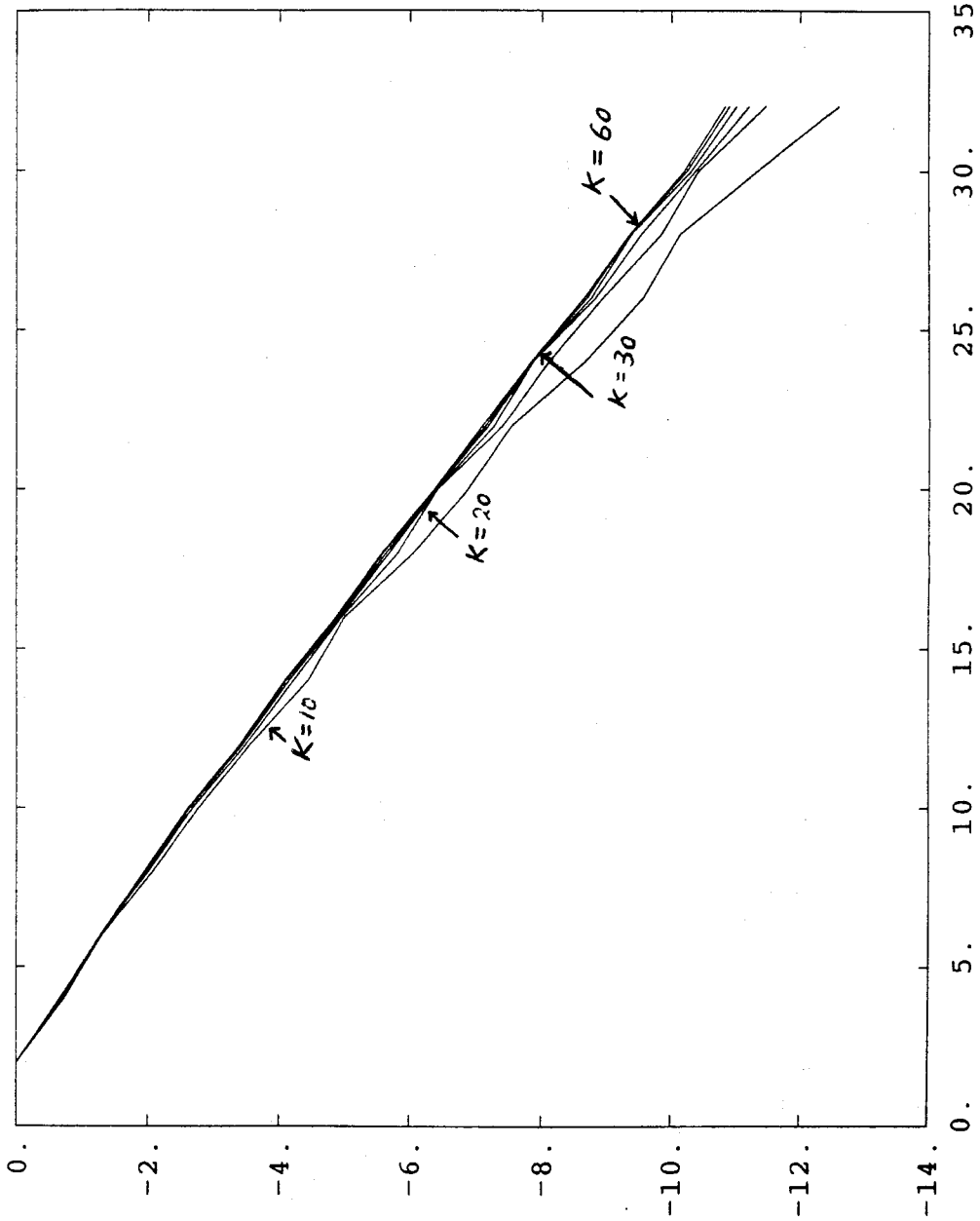

Fig. 3.21. $\log_{10}(\frac{\lambda_{\min}}{\lambda_{\max}})$ as a function of the number of missing samples, L.
$(\omega_s = \frac{\pi}{2})$.

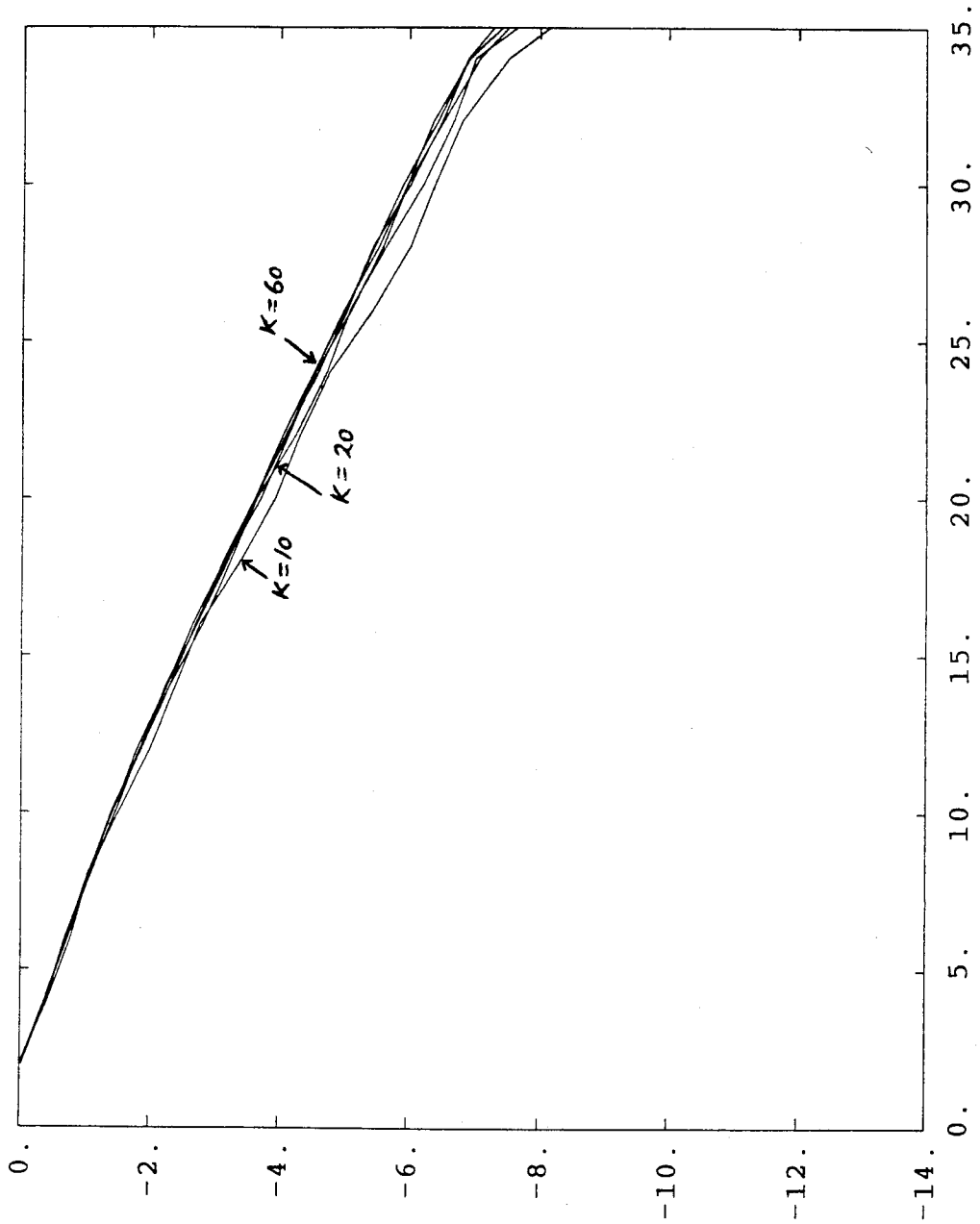

Fig. 3.22. $\log_{10}(\frac{\lambda_{\min}}{\lambda_{\max}})$ as a function of the number of missing samples, L.
$(\omega_s = \frac{\pi}{3})$.

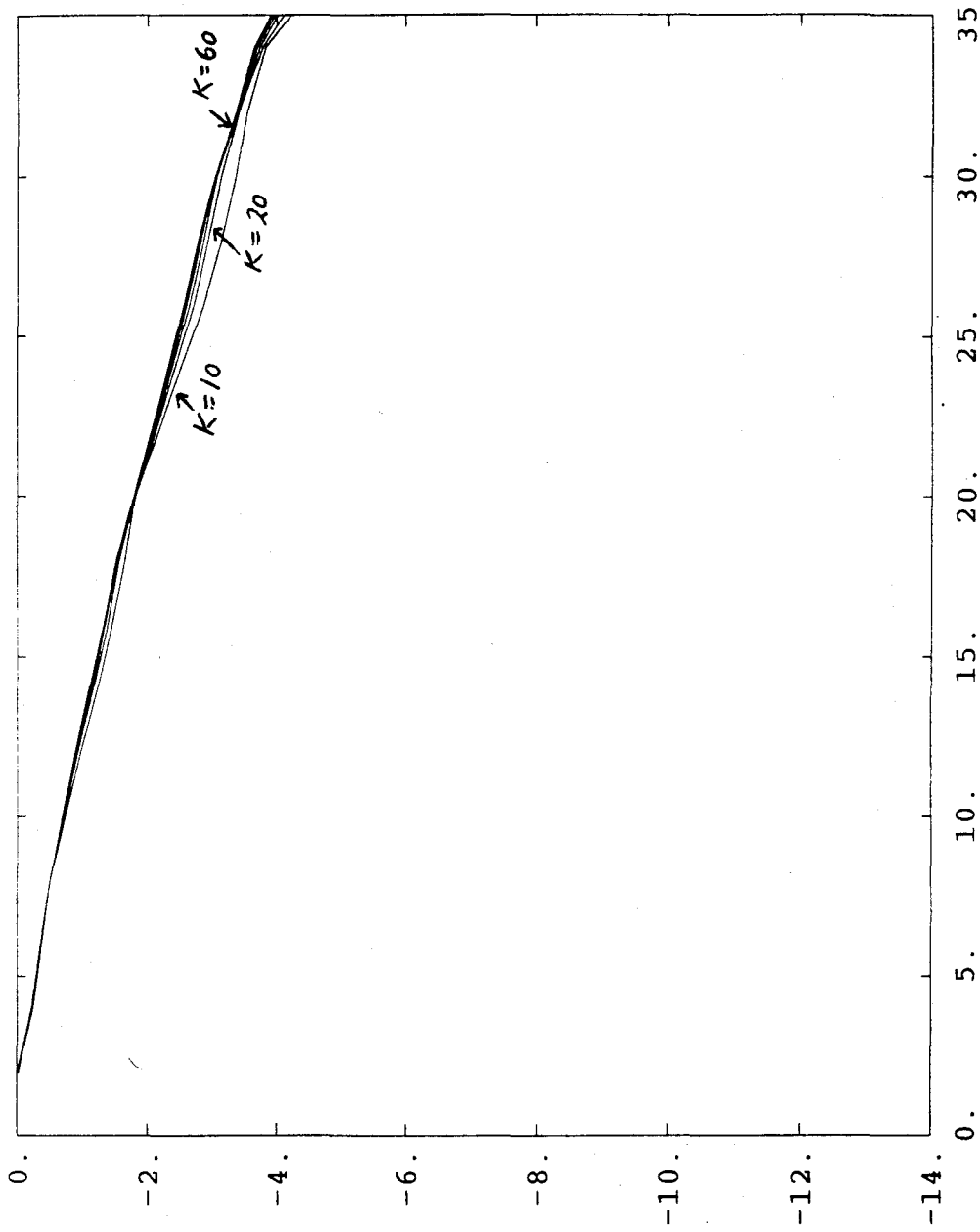

Fig. 3.23. $\log_{10}(\frac{\lambda_{\min}}{\lambda_{\max}})$ as a function of the number of missing samples, L.
$(\omega_s = \frac{\pi}{5})$.

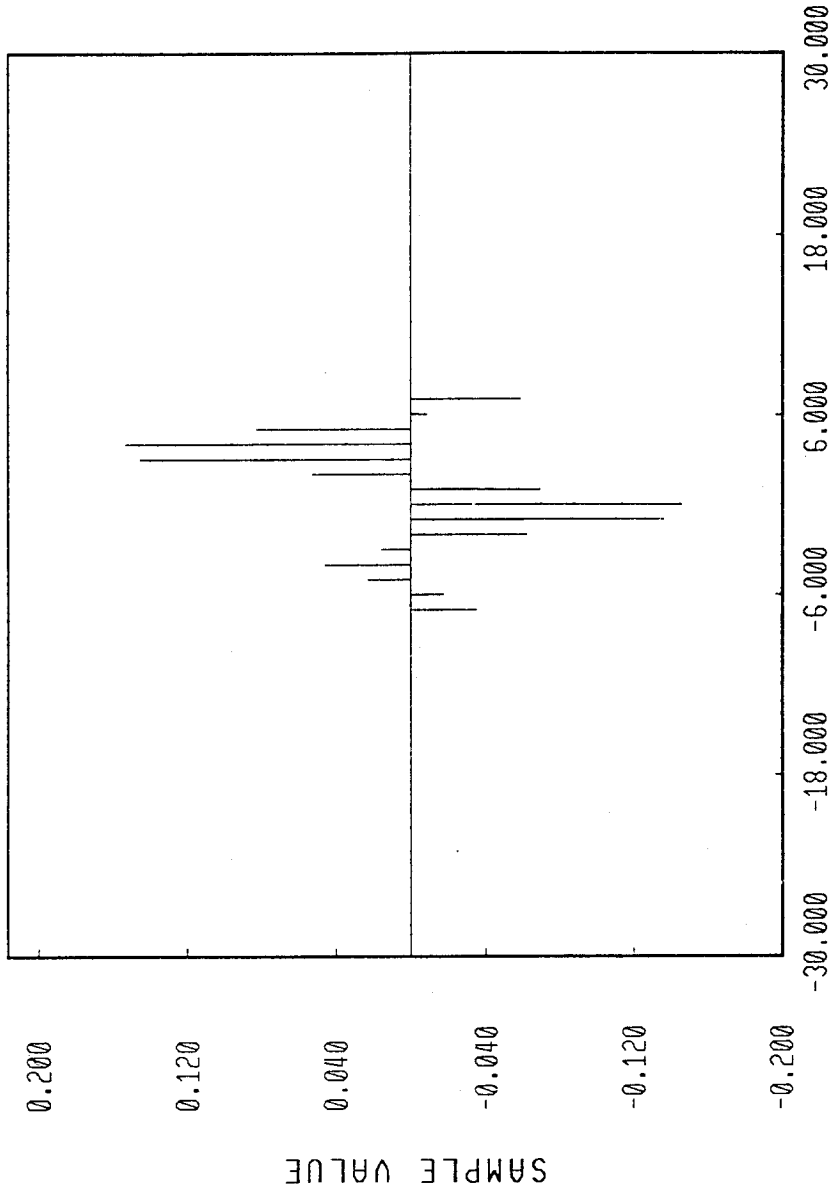

Fig. 3.24. The observed samples of $y(n)$.

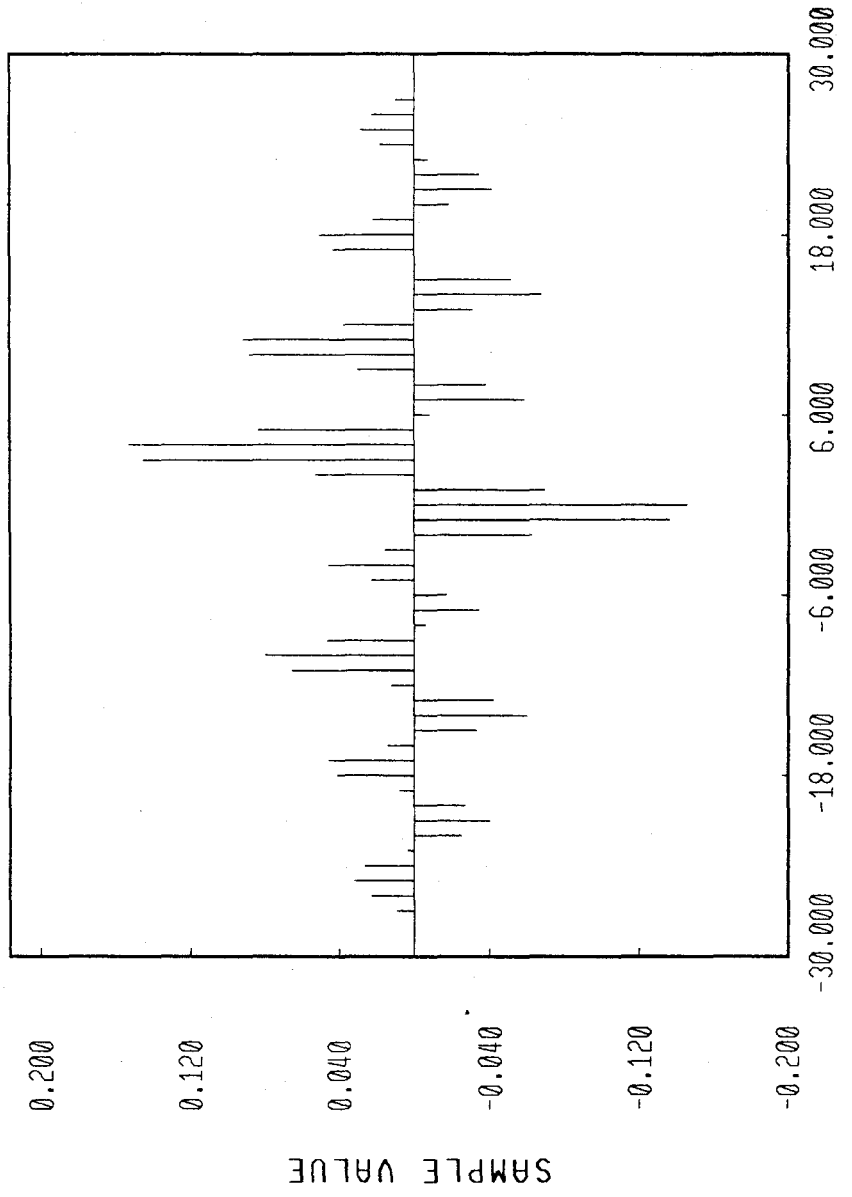

Fig. 3.25. The extrapolated sequence $x(n)$.

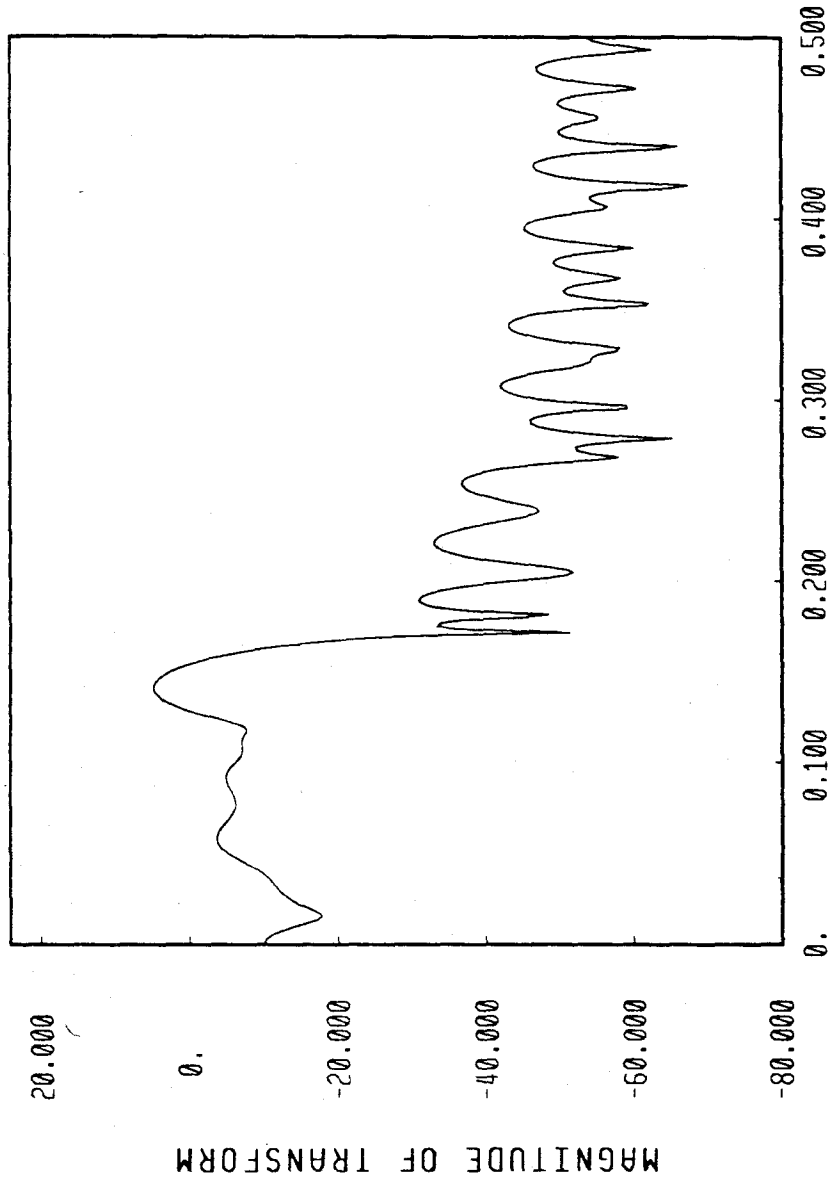

Fig. 3.26. Magnitude (in dB) of the Fourier transform $X(e^{j\omega})$.

intervals of length 5 and retain 2 samples in each interval. As shown previously successful reconstruction of the original signal from its compressed version is shown to be possible with the use of filter banks.

The use of (3.65) presents another method for reconstruction. A periodic sub-sampling scheme is chosen here so that we can ensure the retained samples are evenly spread out between -40 to 40. A scheme which retains samples that are evenly spread out will work better than one that retains all the samples in the middle. In the case where all the observed samples are in the middle, we have seen that the matrix $\mathbf{Q}_2$ is ill-conditioned. For the periodic sub-sampling scheme, the condition number on $\mathbf{Q}_2$ is small, provided that the ratio of the number of missing samples to the total number of samples, $\frac{L}{K+L}$, does not exceed ω_s/π. This is demonstrated in Fig. 3.27 where the reciprocal of the condition number is plotted against the number of missing samples (L). The total number of samples is held constant at $K + L = 81$. With the given bandlimit of $\omega_s = 2/5\pi$, we see that for K exceeding $\frac{2}{5} \times 81$ the condition number grows rapidly. One wants to keep the condition number small, so that if noise exists in the observation data it will not be amplified significantly. Having K restricted to $K < \frac{2}{5}(K + L)$, so with 81 samples in $y(n)$ the largest number for K is 32. If 32 samples are kept using a periodic sub-sampling scheme, then we can be sure that $\mathbf{Q}_2$ is well conditioned.

Let the samples that are retained be

$$y(-39),\ y(-37),\ y(-34),\ y(-32),\ y(-29),\ \ldots y(38). \tag{3.69}$$

Thus, starting from $y(-40)$, in every period of 5 samples two samples are retained. This constitutes a periodic sub-sampling of the original sequence $y(n)$. With the

help of (3.65), the missing samples are interpolated so that the total sequence $x(n)$ has minimum stopband energy subject to the constraint in (3.56). Fig. 3.28 shows an example of $y(n)$ and the reconstructed signal $x(n)$. For comparison, $y(n)-0.15$ is plotted. The displacement of 0.15 is needed to distinguish between the two signals, since the two are nearly identical.

The reconstruction error is given by $e(n) = y(n) - x(n)$. Since $e(f(k)) = 0$ due to (3.56), we shall consider the $f'(l)$ samples only. Let the vectors $\mathbf{e}_1$, $\mathbf{y}_1$ and $\mathbf{x}_1$ be defined as $[\mathbf{e}_1]_l = e(f'(l))$, $[\mathbf{y}_1]_l = y(f'(l))$ and $[\mathbf{x}_1]_l = x(f'(l))$. The stopband energies of $y(n)$ and $x(n)$ are given respectively by

$$E_y = \mathbf{y}_0^T\mathbf{Q}_0\mathbf{y}_0 + 2\mathbf{y}_0^T\mathbf{Q}_1\mathbf{y}_1 + \mathbf{y}_1^T\mathbf{Q}_2\mathbf{y}_1,$$

$$E_x = \mathbf{y}_0^T\mathbf{Q}_0\mathbf{y}_0 + 2\mathbf{y}_0^T\mathbf{Q}_1\mathbf{x}_1 + \mathbf{x}_1^T\mathbf{Q}_2\mathbf{x}_1. \tag{3.70}$$

From (3.65) and (3.70) we get

$$E_y - E_x = (\mathbf{y}_1 - \mathbf{x}_1)^T\mathbf{Q}_2(\mathbf{y}_1 - \mathbf{x}_1)$$

$$= \mathbf{e}_1^T\mathbf{Q}_2\mathbf{e}_1. \tag{3.71}$$

Since $\mathbf{Q}_2$ is symmetric positive definite, let $\{\mathbf{q}_i\}$ be its orthonormal set of eigenvectors and let λ_i be its eigenvalues. Expressing the error vector $\mathbf{e}_1$ as $\mathbf{e}_1 = \sum_{i=0}^{2N-K}\beta_i\mathbf{q}_i$, the energy difference in (3.71) becomes

$$E_y - E_x = \sum_{i=0}^{2N-K}\lambda_i|\beta_i|^2, \tag{3.72}$$

while $\mathbf{e}_1^T\mathbf{e}_1 = \sum_{i=0}^{2N-K}|\beta_i|^2$. Hence, the norm of the error vector is related to the energy difference $E_y - E_x$. In particular, upper and lower bounds for $\mathbf{e}_1^T\mathbf{e}_1$ can be stated as

$$\frac{E_y - E_x}{\lambda_{\max}} \le \mathbf{e}^T\mathbf{e} \le \frac{E_y - E_x}{\lambda_{\min}}. \tag{3.73}$$

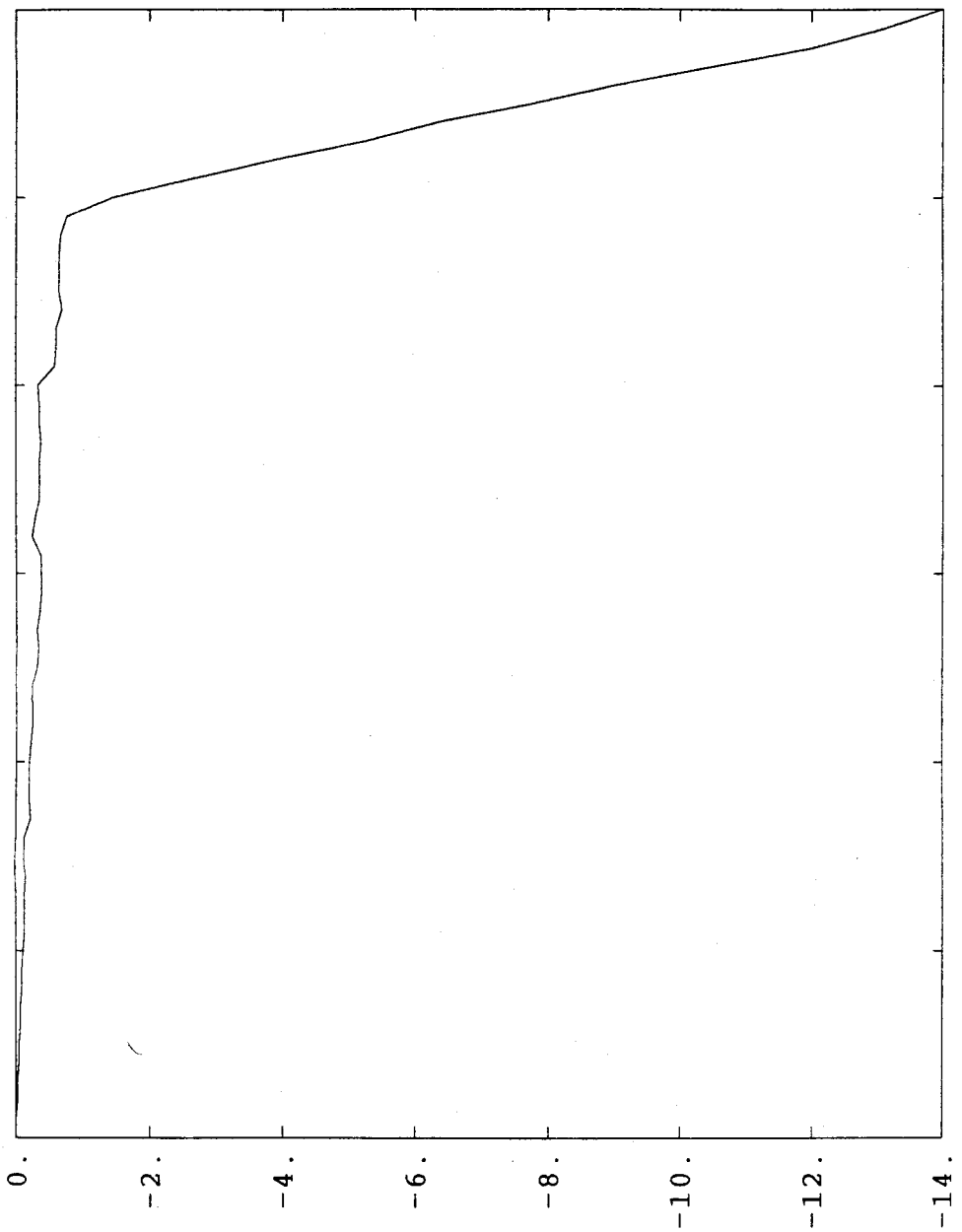

Fig. 3.27. For periodic sub-sampling, $\log_{10}(\frac{\lambda_{min}}{\lambda_{max}})$ as a function of the number of missing samples, L. Total length of sequence is constant $K + L = 81$. $(\omega_s = \frac{2\pi}{5})$

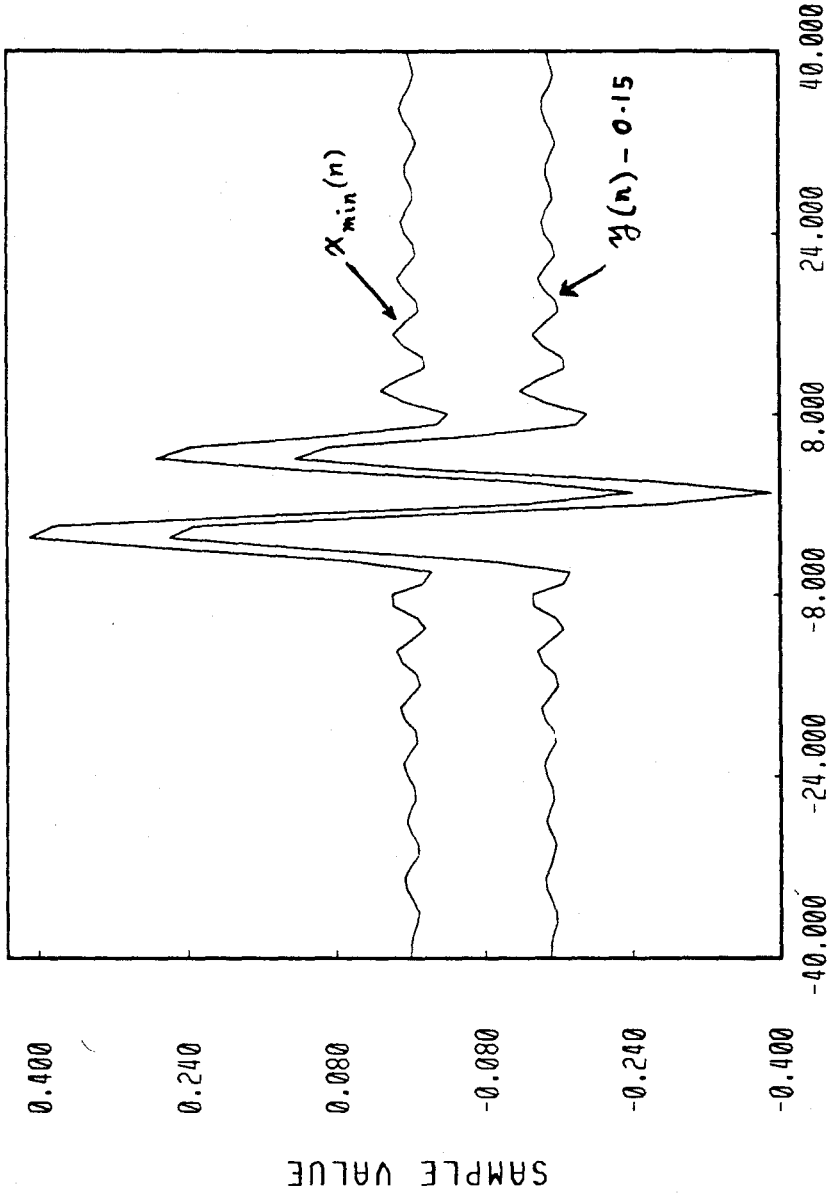

Fig. 3.28. Comparison between the original signal $y(n)$ and the extrapolated signal $x_{\min}(n)$.

In (3.73), since E_x represents the minimum stopband energy possible $E_y \geq E_x$ always and as $y(n)$ becomes more band-limited (in the sense that E_y decreases), the upper and lower bounds decrease with it. Also the particular sub-sampling pattern chosen (which determines $\mathbf{Q}_2$) will have an effect on the error.

3.8 Conclusion

The nonuniform sub-sampling and reconstruction of a multi-band bandlimited signal can be achieved using multi-rate filter banks. The synthesis filters required for perfect reconstruction have multilevel frequency responses. They are not realizable, however, they can be approximated by FIR filters. We considered in general the problem of designing and implementing multilevel filters, and showed that they can be obtained using a lowpass Mth-band filter as a prototype. A new structure was presented for implementing the synthesis filter bank. For the case of signals bandlimited to $\frac{2\pi}{3}$, this structure was demonstrated to be computationally much more efficient than the method in [VA88d] and the traditional non-integer decimation/interpolation scheme.

The 1D results in Sec. 3.2 - 3.4 can be generalized to two-dimensional bandlimited signals. By using an integer lattice decimation scheme, the sub-bands take on the shape of parallelograms. The reconstruction filter bank for 2D can be designed starting from 1D prototype filters, even though the resulting filter bank is not separable.

In Sec. 3.6, we consider the finite-length extrapolation problem. For any given observation data, the problem of finding a finite-length extrapolation which is almost band-limited can be formulated as an energy minimization problem. The

energy quantity to be minimized is the out-of-band energy of the extrapolated sequence $x(n)$. By differentiating this energy with respect to the undetermined values of $x(n)$ and then setting the derivatives to zero, the optimal $x(n)$ can be found. The solution involves the inversion of a positive definite Toeplitz matrix. In the case when the extrapolated samples are too large, an extra term may be added to the objective function to be minimized. This constrains the norm of the extrapolated sequence.

As shown for the example in Sec. 3.7, the above method can be used to reconstruct almost band-limited signals from their sub-samples. The sub-sampling can be periodic as in Sec. 3.2, but the present method allows for any scheme of sub-sampling.

Chapter IV. Random Process Inputs to QMF Bank and Optimal Filtering of Cyclostationary Processes

4.1. Introduction

In this chapter, we shall consider the effects of random process [PAP65] inputs to the multirate filter bank. Since the filter bank is not a time-invariant system, for an input random process $x(n)$ that is wide-sense stationary (WSS) the output $\hat{x}(n)$ will not necessarily be WSS. We will derive in Sec 4.2 the necessary and sufficient condition under which the wide sense stationarity of the input is preserved. In order to obtain the above mentioned condition, we stated and proved two facts concerning random processes and their blocked version. Fact 4.1 deals with a wide sense cyclostationary (WSCS) process [GA75] and its 'blocked' version. It states that a random process is WSCS with period M if and only if its M-fold blocked version is WSS. Fact 4.2 deals with a WSS process and its blocked version. It asserts that a process is WSS if and only if its blocked version is WSS and the blocked vector random process has a power spectral density matrix that is pseudo-circulant. Using these two facts, the necessary and sufficient condition for the filter bank to preserve wide-sense stationarity is derived as Theorem 4.1. It turns out that the filter bank has to satisfy a weaker set of alias cancellation conditions, namely that all the terms $X(z\mathcal{W}^k)$ $(0 \leq k \leq M_1)$ should be cancelled except for one. The resulting output $\hat{x}(n)$ is simply the filtered version of a modulated input. So far the significance of Theorem 4.1 has been purely academic.

In Sec. 4.3, we will look at the problem of estimating a WSCS random process from a noise corrupted version of it. The noise is allowed to be WSCS also. There are several ways of approaching the problem. One is to design a single filter that would

minimize the mean square error averaged over a period of M [GA75]. The optimal filter can be found through the use of the 'averaged' correlation function. Another approach is to block the random process into a vector random process [MIT78] [BA80] and apply multi-channel estimation [MA86] to the blocked version. In this way, the means square error of estimation may be minimized for each time instance. However, due to blocking effects the filters are asymmetrical with respect to the observed samples used for estimation. We shall show that this can be corrected by going to a filter bank formalism. The two methods yield the same theoretical solution in the limit where the filter order is allowed to approach infinity.

In Sec. 4.4, several applications where cyclostationary noise arises are presented. They are in the areas of multirate QMF bank [CR83] [VA87a], digital transmultiplexers [SC81], and the sampling of analog waveforms by multiple A/D converters operating in parallel where the timing jitter in the sampling process introduces a cyclostationary error. The effects of such cyclostationary errors could potentially be reduced by using the filtering scheme discussed above.

For a random process $x(n)$, its statistical auto-correlation function is denoted as $R_{XX}(n_1, n_2)$. This notation will be used for both stationary and non-stationary processes. The auto-correlation function is defined to be $R_{XX}(n_1, n_2) \triangleq Ex(n_1)x^*(n_2)$. A wide-sense stationary random process is one for which $Ex(n) = Ex(n+k)$ for all integers n and k, and

$$R_{XX}(n_1, n_2) = R_{XX}(n_1 + k, n_2 + k) \quad \forall k \in Z. \tag{4.1}$$

The statistical cross-correlation function between $x(n)$ and $y(n)$ is denoted by

$R_{XY}(n_1, n_2)$. The correlation functions are defined as

$$R_{XY}(n_1, n_2) \triangleq E x(n_1) y^*(n_2). \tag{4.2}$$

Two random processes are jointly wide-sense stationary (WSS) if they are individually WSS and their cross-correlation function has the property $R_{XY}(n_1, n_2) = R_{XY}(n_1 + k, n_2 + k)$ for all $k \in Z$. For a vector random process $\mathbf{w}(n)$, its statistical correlation function is a matrix which is defined as

$$\mathbf{R}_{WW}(n_1, n_2) \triangleq E \mathbf{w}(n_1) \mathbf{w}^\dagger(n_2). \tag{4.3}$$

The functions in (4.1-4.3) have two-dimensional z-transforms which will be denoted by $S_{XX}(z_1, z_2)$, $S_{XY}(z_1, z_2)$ and $\mathbf{S}_{WW}(z_1, z_2)$ respectively. Notice that for wide-sense stationary processes, $S_{XX}(e^{j\omega_1}, e^{j\omega_2})$ reduces to the form $\hat{S}(\omega)\delta(\omega_1 + \omega_2)$ where $\hat{S}(\omega)$ is what is commonly called the power spectral density function of $x(n)$.

A random process is called wide-sense cyclostationary (WSCS) with periodicity M if its statistical mean satsifies $Ex(n) = Ex(n + kM)$ for all n and k and

$$R_{XX}(n_1, n_2) = R_{XX}(n_1 + kM, n_2 + kM) \quad \forall n_1, n_2, k \in Z. \tag{4.4}$$

A WSS process can be thought of as being WSCS with an arbitrary period M.

4.2. Wide-sense Stationary Inputs to the QMF Bank

Let $x(n)$ be the input to an M-fold maximally decimated filter bank. Assuming that $x(n)$ is wide-sense stationary, one might ask the question what conditions the analysis and synthesis banks should satisfy in order for the output $\hat{x}(n)$ to remain WSS.

A sufficient conditon is that the filter banks should cancel aliasing for any arbitrary input $x(n)$. In that case, we can write $\hat{x}(n) = \sum_{m=-\infty}^{\infty} t(n - m)x(m)$, since

the whole system acts like a linear time-invariant system and $t(n)$ are the impulse response coefficients of the system transfer function. For a linear time-invariant system, the wide-sense stationary property is preserved [PAP65]. The condition for alias-cancellation can be stated in terms of the analysis and synthesis filters $H_k(z)$ and $F_k(z)$ [VA87a]. The filter bank output can always be written as

$$\hat{X}(z) = \frac{1}{M} \sum_{i=0}^{M-1} X(z\mathcal{W}^i) \sum_{k=0}^{M-1} H_k(z\mathcal{W}^i) F_k(z). \tag{4.5}$$

$X(z\mathcal{W}^i)$ are the alias terms of $X(z)$, and we shall call $\sum_{k=0}^{M-1} H_k(z\mathcal{W}^i)F_k(z)$ the alias-term weighting functions. Denoting these functions as $A_i(z)$, they can be expressed in matrix form

$$\begin{pmatrix} A_0(z) \\ A_1(z) \\ \vdots \\ A_{M-1}(z) \end{pmatrix} =$$

$$\begin{pmatrix} H_0(z) & H_1(z) & \dots & H_{M-1}(z) \\ H_0(z\mathcal{W}) & H_1(z\mathcal{W}) & \dots & H_{M-1}(z\mathcal{W}) \\ \vdots & \vdots & & \vdots \\ H_0(z\mathcal{W}^{M-1}) & H_1(z\mathcal{W}^{M-1}) & \dots & H_{M-1}(z\mathcal{W}^{M-1}) \end{pmatrix} \begin{pmatrix} F_0(z) \\ F_1(z) \\ \vdots \\ F_{M-1}(z) \end{pmatrix}. \tag{4.6}$$

Aliasing is cancelled for all input $x(n)$ if and only if $A_i(z) = 0$ for $i = 1, 2, \dots M-1$.

In deriving a condition that is both necessary and sufficient, we shall make use of the blocked version of $x(n)$. In Fig. 4.1(a), let the output from the kth decimator be denoted as $w_k(n)$, then $w_k(n) = x(nM-k)$. The blocked version of $x(n)$ is defined as

$$\mathbf{w}(n) = (\ w_0(n)\ w_1(n)\ \dots\ w_{M-1}(n)\ )^T. \tag{4.7}$$

A relation between $x(n)$ and $\mathbf{w}(n)$ may be stated as follows.

Fact 4.1 The random process $x(n)$ is WSCS if and only if the vector process $\mathbf{w}(n)$ is WSS.

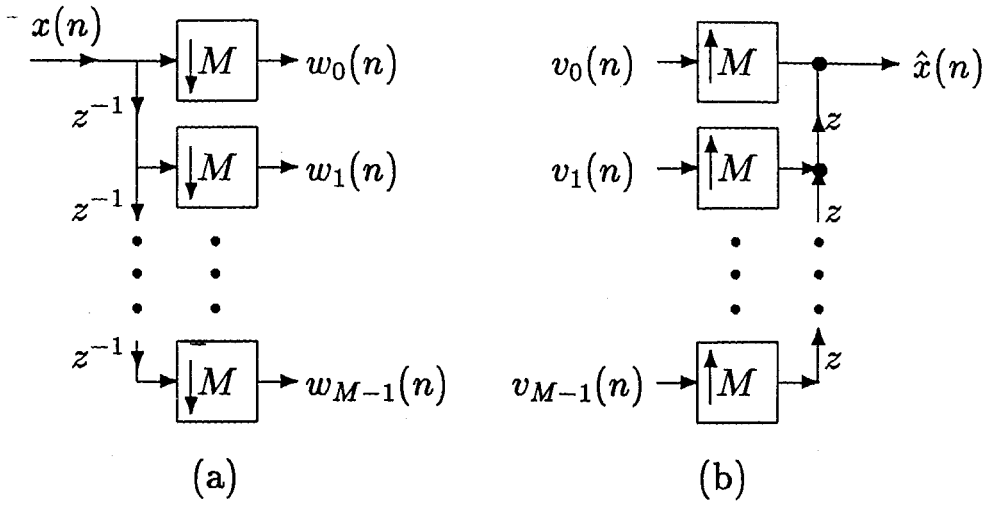

Fig. 4.1. The blocked versions of $x(n)$ and of $\hat{x}(n)$.

Proof of Fact 4.1: The cross-correlation function between $w_i(n)$ and $w_k(n)$ can be written as

$$R_{W_iW_k}(n_1, n_2) = R_{XX}(n_1M - i, n_2M - k). \tag{4.8}$$

If $x(n)$ is wide-sense cyclostationary, then for any integers k and l $(0 \leq k, l \leq M-1)$

$$\begin{aligned} R_{W_iW_k}(n_1, n_2) &= R_{XX}(n_1M - i, n_2M - k) \\ &= R_{XX}(n_1M - i - lM, n_2M - k - lM) \qquad \forall l \\ &= R_{W_iW_k}(n_1 - l, n_2 - l) \qquad \forall l. \end{aligned} \tag{4.9}$$

This means $w_i(n)$ and $w_k(n)$ are jointly WSS for all combinations of i and k, therefore $\mathbf{w}(n)$ is WSS.

Conversely, if $\mathbf{w}(n)$ is WSS, then each entry of $\mathbf{R}_{WW}(n_1, n_2)$ satisfies the property $R_{W_i,W_k}(n_1, n_2) = R_{W_i,W_k}(n_1 - l, n_2 - l)$ for every integer l. Therefore

$$R_{XX}(n_1M - i, n_2M - k) = R_{XX}(n_1M - i - lM, n_2M - k - lM) \qquad \forall l \tag{4.10}$$

which proves that $x(n)$ is WSCS. ■

For the case where $x(n)$ is WSS, the following fact pertains to the z-transform matrix $\mathbf{S}_{\mathbf{WW}}(z_1, z_2)$ which is defined in (4.4).

Fact 4.2 The random process $x(n)$ is WSS if and only if $\mathbf{w}(n)$ is WSS and the matrix $\mathbf{S}_{\mathbf{WW}}(z, z^{-1})$ is pseudo-circulant in z.

Proof of Fact 4.2: If $x(n)$ is WSS, then it is WSCS for any M, so by Fact 4.1 $\mathbf{w}(n)$ is WSS. Furthermore, the entries of $\mathbf{R}_{\mathbf{WW}}(n_1, n_2)$ are related to each other as

$$\begin{aligned} R_{W_i W_k}(n_1, n_2) &= R_{XX}(n_1 M - i, n_2 M - k) \\ &= R_{XX}(n_1 M, n_2 M - k + i) \\ &= \begin{cases} R_{W_0 W_{k-i}}(n_1, n_2) & \text{for } k - i \geq 0; \\ R_{W_0 W_{k-i+M}}(n_1 - 1, n_2) & \text{otherwise.} \end{cases} \end{aligned} \tag{4.11}$$

The z-transform of the above relation is

$$S_{W_i W_k}(z_1, z_2) = \begin{cases} S_{W_0 W_{k-i}}(z_1, z_2) & \text{for } k - i \geq 0; \\ z_1^{-1} S_{W_0 W_{k-i+M}}(z_1, z_2) & \text{otherwise.} \end{cases} \tag{4.12}$$

As a result, the matrix $\mathbf{S}_{\mathbf{WW}}(z_1, z_2)$ has the following structure

$$\begin{pmatrix} S_{W_0 W_0}(z_1, z_2) & S_{W_0 W_1}(z_1, z_2) & \dots & S_{W_0 W_{M-1}}(z_1, z_2) \\ z_1^{-1} S_{W_0 W_{M-1}}(z_1, z_2) & S_{W_0 W_0}(z_1, z_2) & \dots & S_{W_0 W_{M-2}}(z_1, z_2) \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{-1} S_{W_0 W_1}(z_1, z_2) & z_1^{-1} S_{W_0 W_2}(z_1, z_2) & \dots & S_{W_0 W_0}(z_1, z_2) \end{pmatrix}. \tag{4.13}$$

This shows $\mathbf{S}_{\mathbf{WW}}(z_1, z_2)$ is pseudo-circulant in z_1. In particular $\mathbf{S}_{\mathbf{WW}}(z, z^{-1})$ is pseudo-circulant in z.

Now if $\mathbf{w}(n)$ is WSS, then (4.10) holds. Furthermore, if $\mathbf{S}_{\mathbf{WW}}(z, z^{-1})$ is pseudo-circulant in z, then

$$R_{W_i W_k}(n, -n) = \begin{cases} R_{W_0 W_{k-i}}(n, -n) & \text{for } k - i \geq 0 \\ R_{W_0 W_{k-i+M}}(n+1, -n+1) & \text{for } k - i < 0. \end{cases} \tag{4.14}$$

This in turn means $R_{XX}(nM-i,-nM-k) = R_{XX}(nM,-nM-k+i)$ for $0 \leq k,i \leq M-1$. Combining the above equation with (4.10) one can then show that $R_{XX}(n_1,n_2)$ satisfies the WSS property. ■

Facts 4.1 and 4.2 are true also for $\hat{x}(n)$ and its blocked version $\mathbf{v}(n)$ (Fig. 4.1(b)) where the block version is defined as $\mathbf{v}(n) = (\, v_0(n)\ v_1(n)\ \ldots\ v_{M-1}(n))^T$. With the help of the above two facts, we shall now prove that the QMF bank preserves the wide-sense stationarity of the input if and only if the filter bank satisfies a modified set of alias-cancellation conditions. This is stated as Theorem 4.1 whose proof will occupy the remainder of this section.

Theorem 4.1 The random process output $\hat{\mathbf{x}}(n)$ is WSS for every WSS input $x(n)$ if and only if the QMF bank satisfies the conditions that, for some integer p in the range $0 \leq p \leq M-1$, the alias-weighting functions $A_i(z) = 0$ for $i \neq p$.

The sufficiency of the above condition is not surprising, since it merely means $\hat{X}(z) = A_p(z)X(z\mathcal{W}^p)$ which is equivalent to saying $\hat{x}(n)$ is the output of a filter $A_p(z)$ in response to the modulated input $x(n)\mathcal{W}^{pn}$. If $x(n)$ is WSS, then both the modulated input and $\hat{x}(n)$ are WSS. Hence the condition in Theorem 4.1 is sufficient.

Now, we shall prove that it is also necessary. With the filter bank represented as in Fig. 4.2, we see

$$\mathbf{S}_{VV}(z_1,z_2) = \mathbf{P}(z_1)\mathbf{S}_{WW}(z_1,z_2)\mathbf{P}_*^T(z_2). \tag{4.15}$$

If $\hat{x}(n)$ is to be WSS for every WSS input $x(n)$, then by Fact 4.2 the system $\mathbf{P}(z)$ must do two things: (a.)$\mathbf{v}(n)$ should be WSS for every WSS $\mathbf{w}(n)$; (b.) $S_{VV}(z,z^{-1})$

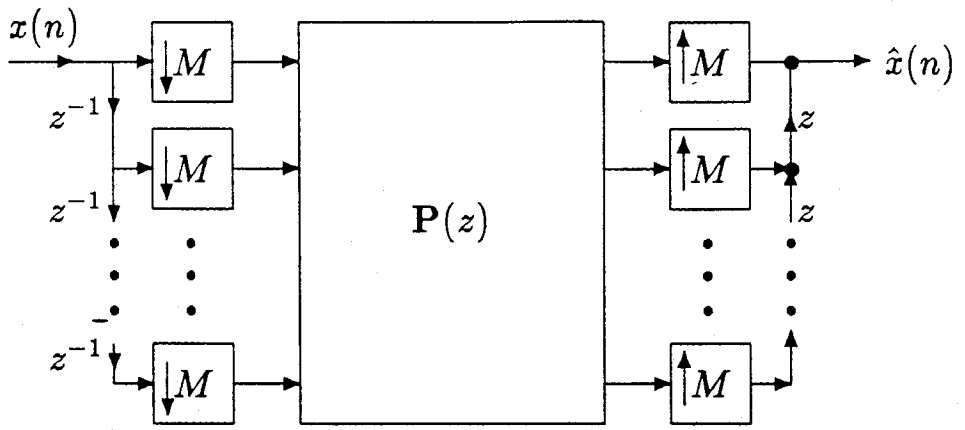

Fig. 4.2. An alternative representation of the QMF bank.

as given in (4.15) must be pseudo-circulant for every matrix $\mathbf{S}_{WW}(z, z^{-1})$ that is pseudo-circulant. The requirement in (a.) is automatically satisfied due to the fact that $\mathbf{P}(z)$ is a linear time-invariant system.

Since $\mathbf{S}_{WW}(z, z^{-1})$ is pseudo-circulant in z, it can be factorized [VA88b] as $\mathbf{D}(z^{\frac{1}{M}})\mathbf{W}\mathbf{\Lambda}_{WW}(z)\mathbf{W}^{\dagger}\mathbf{D}(z^{-\frac{1}{M}})$, where both $\mathbf{\Lambda}_{WW}(z)$ and $\mathbf{D}(z)$ are diagonal matrices with

$$\mathbf{D}(z) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & z^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z^{-M+1} \end{pmatrix}. \tag{4.16}$$

If we define a matrix $\mathbf{\Lambda}_{VV}(z)$ associated with $\mathbf{S}_{VV}(z, z^{-1})$ as in $\mathbf{S}_{VV}(z, z^{-1}) = \mathbf{D}(z^{\frac{1}{M}})$ $\mathbf{W}\ \mathbf{\Lambda}_{VV}(z)\ \mathbf{W}^{\dagger}\ \mathbf{D}(z^{-\frac{1}{M}})$ and also define $\mathbf{L}(z) = \mathbf{W}^{\dagger}\mathbf{D}(z^{-\frac{1}{M}})\mathbf{P}(z)\mathbf{D}(z^{\frac{1}{M}})\mathbf{W}$, then by letting $z_1 = z_2^{-1} = z$ in (4.15) we get

$$\mathbf{\Lambda}_{VV}(z) = \mathbf{L}(z)\mathbf{\Lambda}_{WW}(z)\tilde{\mathbf{L}}(z). \tag{4.17}$$

The matrix $\mathbf{\Lambda}_{VV}(z)$ is diagonal if and only if $\mathbf{S}_{VV}(z, z^{-1})$ is pseudo-circulant. In conclusion, to make $\hat{x}(n)$ WSS for all WSS input we must design $\mathbf{L}(z)$ such that $\mathbf{\Lambda}_{VV}(z)$ is diagonal for any diagonal matrix $\mathbf{\Lambda}_{WW}(z)$. This imposes a certain structure on $\mathbf{L}(z)$ (and hence on $\mathbf{P}(z)$).

We will now derive the necessary structure on $\mathbf{L}(z)$. Since $\mathbf{\Lambda}_{WW}(z)$ is diagonal, the (i,k)th entry of $\mathbf{\Lambda}_{VV}(z)$ will be given by

$$[\mathbf{\Lambda}_{VV}(z)]_{i,k} = \sum_{l=0}^{M-1} [\mathbf{L}(z)]_{i,l}[\mathbf{\Lambda}_{WW}(z)]_{l,l}[\tilde{\mathbf{L}}(z)]_{l,k}. \tag{4.18}$$

If we let $i \neq k$, it is necessary that $[\mathbf{\Lambda}_{VV}(z)]_{i,k} = 0$ for arbitrary diagonal $\mathbf{\Lambda}_{WW}(z)$, and so $[\mathbf{L}(z)]_{i,l}[\tilde{\mathbf{L}}(z)]_{l,k} = 0$ for each l. On the unit circle $z = e^{j\omega}$, this means $[\mathbf{L}(e^{j\omega})]_{i,l}[\mathbf{L}(e^{j\omega})]^*_{k,l} = 0$ which is equivalent to $[\mathbf{L}(e^{j\omega})]_{i,l}[\mathbf{L}(e^{j\omega})]_{k,l} = 0$. This means for every column of $\mathbf{L}(z)$ there can be at most one non-zero entry. $\mathbf{L}(z)$ can be factorized into $\mathbf{P}_L\mathbf{\Lambda}_L(z)$ where $\mathbf{\Lambda}_L(z)$ is diagonal with its kth diagonal entry equal to the non-zero element in the kth column of $\mathbf{L}(z)$. In cases where the kth column of $\mathbf{L}(z)$ is all zero, the corresponding diagonal entry in $\mathbf{\Lambda}_L(z)$ becomes zero. In the matrix $\mathbf{P}_L$, we shall set its entries to unity wherever the corresponding entries in $\mathbf{L}(z)$ is non-zero. If $\mathbf{L}(z)$ has null columns, then we set the topmost entry of the corresponding column in $\mathbf{P}_L$ to be unity. The remaining entries of $\mathbf{P}_L$ are then all set to zero. In this way, it is guaranteed that each column of $\mathbf{P}_L$ has one and only one non-zero element. Defining the product $\mathbf{W}\mathbf{P}_L\mathbf{\Lambda}_L(z)\mathbf{W}^\dagger$ to be $\mathbf{C}(z)$, we get

$$[\mathbf{C}(z)]_{i,k} = \sum_{l=0}^{M-1} W^{ip(l)}[\mathbf{\Lambda}_L(z)]_{l,l}W^{-lk}. \tag{4.19}$$

The sequence $p(l)$ is obtained from $\mathbf{P}_L$ by the rule that $[\mathbf{P}_L]_{p(l),l}$ is non-zero for $l = 0, \ldots, M-1$.

Since $\mathbf{P}(z) = \mathbf{D}(z^{\frac{1}{M}})\mathbf{C}(z)\mathbf{D}(z^{-\frac{1}{M}})$ and for the system $\mathbf{P}(z)$ to be realizable it must not contain fractional powers of z, this imposes certain conditions on $\mathbf{C}(z)$. For simplicity of notations, we shall look instead at the conditions on $\mathbf{C}(z^M)$, which is related by

$$\mathbf{P}(z^M) = \mathbf{D}(z)\mathbf{C}(z^M)\mathbf{D}(z^{-1}) = \mathbf{D}(z)\mathbf{W}\mathbf{P}_L\mathbf{\Lambda}_L(z^M)\mathbf{W}^\dagger\mathbf{D}(z^{-1}). \tag{4.20}$$

In order for $\mathbf{P}(z^M)$ to be a function of z^M, we require that $z^{-i+k}[\mathbf{C}(z^M)]_{i,k}$ must be a function of z^M. The entries of $\mathbf{C}(z^M)$ can also be expressed in polyphase $[\mathbf{C}(z^M)]_{i,k} = \mathbf{b}_{i,k}^T(z^M)\mathbf{e}(z)$, where $\mathbf{b}_{i,k}^T(z^M)$ is a vector containing the polyphase components for the (i,k)th entry and $\mathbf{e}(z) = (1, z^{-1}, \ldots, z^{-M+1})^T$.

On the other hand, from the relation

$$\mathbf{P}_L\mathbf{\Lambda}_L(z^M) = \mathbf{W}^\dagger\mathbf{D}(z^{-1})\mathbf{P}(z^M)\mathbf{D}(z)\mathbf{W} \tag{4.21}$$

where the right-hand side contains only integral powers of z, we can conclude that $\mathbf{\Lambda}_L(z^M)$ contains no fractional powers of z. Thus, the diagonal entries of $\mathbf{\Lambda}_L(z^M)$ can be written in polyphase form

$$[\mathbf{\Lambda}_L(z^M)]_{l,l} = \sum_{n=0}^{M-1} z^{-n}\lambda_{l,n}(z^M). \tag{4.22}$$

Substituting (4.22) into (4.19), the polyphase components of $[\mathbf{C}(z^M)]_{i,k}$ are found to be

$$\mathbf{b}_{i,k}^T(z^M) =$$

$$\mathbf{u}_k^\dagger \text{ diag.} \begin{pmatrix} W^{ip(0)} \\ W^{ip(1)} \\ \vdots \\ W^{ip(M-1)} \end{pmatrix} \begin{pmatrix} \lambda_{0,0}(z^M) & \lambda_{0,1}(z^M) & \cdots & \lambda_{0,M-1}(z^M) \\ \lambda_{1,0}(z^M) & \lambda_{1,1}(z^M) & \cdots & \lambda_{1,M-1}(z^M) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{M-1,0}(z^M) & \lambda_{M-1,1}(z^M) & \cdots & \lambda_{M-1,M-1}(z^M) \end{pmatrix}. \tag{4.23}$$

Recall that $\mathbf{u}_k$ is the kth column of the DFT matrix $\mathbf{W}$. The matrix of $\lambda_{l,n}(z^M)$'s will be denoted by $\mathbf{\Theta}(z^M)$. Since we require $z^{-i+k}[\mathbf{C}(z^M)]_{i,k}$ to be a function of z^M, the entry $[\mathbf{C}(z^M)]_{i,k}$ has only one non-zero polyphase component. That component is the $((M-i+k))$th entry of $\mathbf{b}_{i,k}^T(z^M)$. Denoting that entry by $b_{i,k}(z^M)$ one can write

$$(0 \quad \ldots \quad b_{i,k}(z) \quad \ldots \quad 0) = \mathbf{u}_k^\dagger \text{ diag.} \begin{pmatrix} W^{ip(0)} \\ W^{ip(1)} \\ \vdots \\ W^{ip(M-1)} \end{pmatrix} \mathbf{\Theta}(z), \tag{4.24}$$

where $\bar{b}_{i,k}(z)$ occurs at the $((M-i+k))$th entry on the left-hand side of (4.24). For $k = 0, 1, \ldots, M-1$, there are M row equations in (4.24). By stacking the rows together, they can be written as

$$\begin{pmatrix} \mathbf{0} & \mathbf{I}_i \\ \mathbf{I}_{M-i} & \mathbf{0} \end{pmatrix} \text{diag.} \begin{pmatrix} b_{i,0}(z) \\ b_{i,1}(z) \\ \vdots \\ b_{i,M-1}(z) \end{pmatrix} = \mathbf{W}^\dagger \text{ diag.} \begin{pmatrix} W^{ip(0)} \\ W^{ip(1)} \\ \vdots \\ W^{ip(M-1)} \end{pmatrix} \mathbf{\Theta}(z). \tag{4.25}$$

Setting $i = 0$, one gets an expression for $\mathbf{\Theta}(z)$ as

$$\mathbf{\Theta}(z) = \frac{1}{M} \mathbf{W} \text{diag.} \begin{pmatrix} b_{0,0}(z) \\ b_{0,1}(z) \\ \vdots \\ b_{0,M-1}(z) \end{pmatrix}. \tag{4.26}$$

Substituting it back into (4.25), we get

$$\mathbf{W}^{-1} \begin{pmatrix} W^{ip(0)} & & & \\ & W^{ip(1)} & & \\ & & \ddots & \\ & & & W^{ip(M-1)} \end{pmatrix} \mathbf{W} = c_i \begin{pmatrix} \mathbf{0} & \mathbf{I}_i \\ \mathbf{I}_{M-i} & \mathbf{0} \end{pmatrix}, \tag{4.27}$$

with c_i being a scalar constant dependent on the index i. To solve for $p(l)$, one can use the fact that $p(l) = l$ satisfies (4.27) with $c_i = 1$. Thus the general solution for (4.27) is $p(l+1) = ((p(l) + 1))_M$. The constant $c_i = W^{ip(0)}$, so the choice of $p(0)$ remains arbitrary within the range $0 \leq p(0) < M$. And

$$\mathbf{P}_L = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{p(0)} \\ \mathbf{I}_{M-p(0)} & \mathbf{0} \end{pmatrix}. \tag{4.28}$$

In summary, the necessary and sufficient condition for the output to remain WSS for arbitrary WSS input is that $\mathbf{P}(z^M)$ should be of the form

$$\mathbf{P}(z) = \mathbf{D}(z^{\frac{1}{M}}) \mathbf{W} \begin{pmatrix} \mathbf{0} & \mathbf{I}_{p(0)} \\ \mathbf{I}_{M-p(0)} & \mathbf{0} \end{pmatrix} \mathbf{\Lambda}_L(z) \mathbf{W}^\dagger \mathbf{D}(z^{-\frac{1}{M}}). \tag{4.29}$$

The choice for $p(0)$ remains arbitrary.

The right-hand side of (4.6) can be re-written in terms of $\mathbf{P}(z)$ as $\mathbf{W}^\dagger \mathbf{D}(z) \mathbf{P}^T(z^M) \mathbf{r}(z)$. Substituting (4.29) into (4.6), we find that $A_{p(0)}(z) =$

$Mz^{-M+1}[\mathbf{\Lambda}_L(z^M)]_{M-p(0),M-p(0)}$ and all the remaining $A_i(z)$ terms are zero. This means the filter bank produces the output $\hat{X}(z) = \frac{1}{M}X(zW^{p(0)})A_{p(0)}(z)$ in response to $X(z)$. Thus, the condition in Theorem 4.1 is shown to be necessary. ■

4.3 Estimation of Wide-sense Cyclostationary Processes

A common problem encountered in optimal filtering is the estimation of a random process signal $y(n)$ from a noise corrupted version of the signal $x(n) = y(n) + g(n)$. (We assume here that the signals $y(n)$ and $x(n)$ are real.) It is commonly assumed that $y(n)$ and $g(n)$ are jointly WSS. The best linear estimator is a linear time-invariant filter which takes $x(n)$ as the input and produces an estimate $\hat{y}(n)$ such that the mean square error of estimation $E|y(n) - \hat{y}(n)|^2$ is minimized. If no restriction is placed upon the filter, the optimal solution is given by [PAP65]

$$H(z) = \frac{S_{XY}(z)}{S_{XX}(z)}. \tag{4.30}$$

However, the above solution may turn out to be IIR and unstable. If one restricts $H(z)$ to be an FIR filter, then the stability problem is avoided. For an FIR filter that would estimate $y(n)$ from the $2N+1$ samples $x(-N), x(-N+1), \ldots, x(N)$, the optimal filter coefficients are given by [PAP65]

$$\begin{pmatrix} \hat{R}_{XX}(0) & \hat{R}_{XX}(-1) & \ldots & \hat{R}_{XX}(-2N) \\ \hat{R}_{XX}(1) & \hat{R}_{XX}(0) & \ldots & \hat{R}_{XX}(1-2N) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}_{XX}(2N) & \hat{R}_{XX}(2N-1) & \ldots & \hat{R}_{XX}(0) \end{pmatrix} \begin{pmatrix} h(-N) \\ h(1-N) \\ \vdots \\ h(N) \end{pmatrix}$$

$$= \begin{pmatrix} \hat{R}_{XY}(-N) \\ \hat{R}_{XY}(1-N) \\ \vdots \\ \hat{R}_{XY}(N) \end{pmatrix}, \tag{4.31}$$

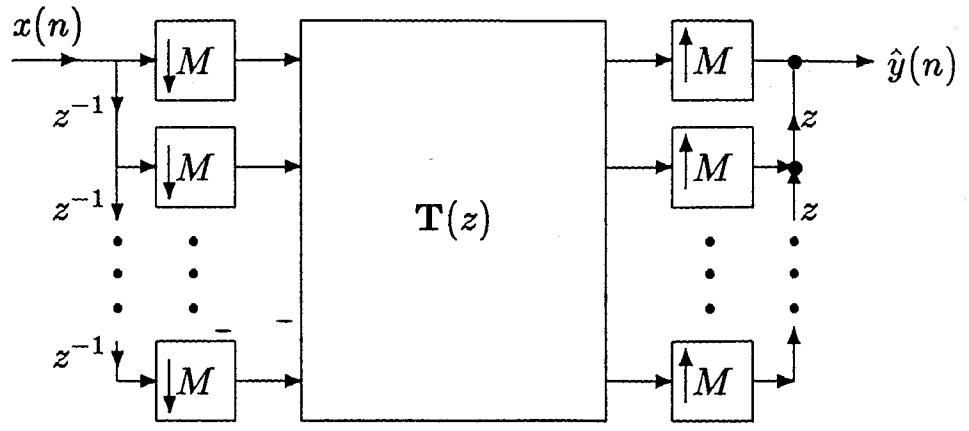

Fig. 4.3. Optimal estimation through block filtering.

where $\hat{R}_{XX}(l) = Ex(n)x(n+l)$ and $\hat{R}_{XY}(l) = Ex(n)y(n+l)$. The filter $H(z) = \sum_{n=-N}^{N} h(n)z^{-n}$.

Now if we consider the case where $y(n)$ and $x(n)$ are not jointly WSS, but rather are jointly WSCS, then the output of $H(z)$ will not be WSS. Also the error $y(n) - \hat{y}(n)$ is WSCS, thus it is not possible to minimize the mean square error $E|y(n) - \hat{y}(n)|^2$ for every n. However, one can still find a filter $H(z)$ that would minimize the error averaged over one period [GA75], $\frac{1}{M}\sum_{n=0}^{M-1} E|y(n) - \hat{y}(n)|^2$. The optimal FIR solution satisfies a similar equation as (4.31) with $\hat{R}_{XX}(l)$ replaced by $\frac{1}{M}\sum_{n=0}^{M-1} R_{XX}(n, n+l)$ and $\hat{R}_{XY}(l)$ replaced by $\frac{1}{M}\sum_{n=0}^{M-1} R_{XY}(n, n+l)$.

A second approach would be to use block filtering (Fig. 4.3). Using block filtering, one may minimize the mean square error for each n individually. Let us define the blocked version of $x(n)$ to be $\mathbf{w}(n)$ as in Fig. 4.1(a) and the blocked version of $y(n)$ to be $\mathbf{v}(n)$,

$$\mathbf{w}(n) = \begin{pmatrix} x(nM) \\ x(nM-1) \\ \vdots \\ x(nM-M+1) \end{pmatrix} \qquad \mathbf{v}(n) = \begin{pmatrix} y(nM) \\ y(nM-1) \\ \vdots \\ y(nM-M+1) \end{pmatrix}. \tag{4.32}$$

From Lemma 4.1, we know that $\mathbf{w}(n)$ and $\mathbf{v}(n)$ are WSS. If we assume each entry of $\mathbf{T}(z)$ is an FIR filter of length $2N+1$, then one may write $\mathbf{T}(z) = \sum_{n=-N}^{N} \mathbf{U}^T(n) z^{-n}$ and

$$\hat{\mathbf{v}}(n) = \sum_{m=-N}^{N} \mathbf{w}^T(n-m)\mathbf{U}(m). \tag{4.33}$$

This corresponds to the case of multi-channel estimation [MA86] where $\mathbf{v}(n)$ is the signal and $\mathbf{w}(n)$ is signal plus noise. The optimal filter coefficients are given by

$$\begin{pmatrix} \hat{\mathbf{R}}_{WW}(0) & \hat{\mathbf{R}}_{WW}(-1) & \dots & \hat{\mathbf{R}}_{WW}(-N) \\ \hat{\mathbf{R}}_{WW}(1) & \hat{\mathbf{R}}_{WW}(0) & \dots & \hat{\mathbf{R}}_{WW}(1-N) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{R}}_{WW}(N) & \hat{\mathbf{R}}_{WW}(N-1) & \dots & \hat{\mathbf{R}}_{WW}(0) \end{pmatrix} \begin{pmatrix} \mathbf{U}(-N) \\ \mathbf{U}(1-N) \\ \vdots \\ \mathbf{U}(N) \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\mathbf{R}}_{WV}(-N) \\ \hat{\mathbf{R}}_{WV}(1-N) \\ \vdots \\ \hat{\mathbf{R}}_{WV}(N) \end{pmatrix} \tag{4.34}$$

where $\hat{\mathbf{R}}_{WW}(l) = E\mathbf{w}(n)\mathbf{w}^T(n+l)$ is the auto-correlation matrix of $\mathbf{w}(n)$ and $\hat{\mathbf{R}}_{WV}(l) = E\mathbf{w}(n)\mathbf{v}^T(n+l)$ is the cross-correlation matrix between $\mathbf{w}(n)$ and $\mathbf{v}(n)$. By taking N to infinity, we arrive at the ideal Wiener solution for block filtering which in the time domain satisfies

$$\sum_{m=-\infty}^{\infty} \hat{\mathbf{R}}_{WW}(n-m)\mathbf{U}(m) = \hat{\mathbf{R}}_{WV}(n), \tag{4.35}$$

therefore in the frequency domain $\mathbf{T}(z) = \mathbf{S}_{WV}(z)\mathbf{S}_{WW}^{-1}(z)$. Equation (4.34) may also be written in terms of the correlation functions $R_{XX}(n_1, n_2)$ and $R_{XY}(n_1, n_2)$. If we let the kth column of the coefficient matrix in (4.34) be denoted as $\mathbf{s}_k$, then the dimension of $\mathbf{s}_k$ is $1 \times (2L+M)$ where $L = MN$ and the vector satisfies the relationship

$$\mathcal{R}_{2L+M}(0,0)\mathbf{s}_k = \begin{pmatrix} R_{XY}(L,-k) \\ R_{XY}(L-1,-k) \\ \vdots \\ R_{XY}(-L-M+1,-k) \end{pmatrix} \tag{4.36}$$

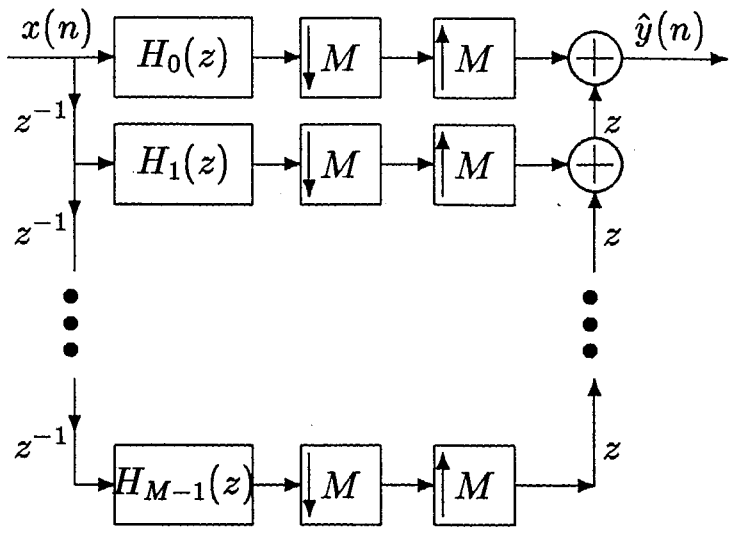

Fig. 4.4. A filter bank used in the optimal estimation of a wide-sense cyclostationary random process.

with $\mathcal{R}_{2L+M}(n_1, n_2)$ being a $(2L+M) \times (2L+M)$ matrix whose (i,k)th entry is defined as $R_{XX}(n_1 - i, n_2 - k)$. By writing n as $Ml - k$ where $0 \leq k \leq M-1$, we see from Fig. 4.3 that the estimate $\hat{y}(n)$ is produced by the kth output of $\mathbf{T}(z)$ at time l. In terms of $\mathbf{s}_k$, this means

$$\begin{aligned} \hat{y}(n) &= \left(\mathbf{w}^T(l+N)\mathbf{w}^T(l+N-1) \quad \ldots \quad \mathbf{w}^T(l-N) \right) \mathbf{s}_k \\ &= \left(x(n+k+NM) \quad \ldots \quad x(n+k-NM-M+1) \right) \mathbf{s}_k. \end{aligned} \tag{4.37}$$

Notice that in making the estimate $\hat{y}(n)$ the $(2N+1)M$ samples of data used are $x(n+k+NM)$ to $x(n+k-NM-M+1)$. These samples are not centered around $x(n)$. If one desires to have an estimator which uses the data symmetrically around the sample to be estimated, then a filter bank of the type shown in Fig. 4.4 may be used.

Let us consider only the case where $H_k(z)$ is an FIR filter of the form $H_k(z) = \sum_{m=-J}^{J} h_k(m) z^{-m}$. At any time instance n, we can write $n = Ml - k$ where $0 \leq$

$k \leq M-1$ and the output $\hat{y}(n)$ is produced by the kth branch of the filter bank at time instance l, i.e.,

$$\hat{y}(n) = \sum_{m=-J}^{J} x(n-m)h_k(m). \tag{4.38}$$

We shall define a vector of impulse response coefficients, $\mathbf{h}_k$, as $(\,h_k(-L) \quad h_k(-L+1) \quad \ldots \quad h_k(L)\,)^T$. Then the estimation error can be written in terms of $\mathbf{h}_k$ as $e(n) = y(n) - (\,x(n+J) \quad x(n+J-1) \quad \ldots \quad x(n-J)\,)\,\mathbf{h}_k$ and the mean square error becomes

$$E|e(n)|^2 = R_{YY}(n,n) - 2\,(\,R_{XY}(n+J,n) \quad \ldots \quad R_{XY}(n-J,n)\,)^T\,\mathbf{h}_k + \mathbf{h}_k^T\,\mathcal{R}_{(2J+1)}(n+J,n+J)\mathbf{h}_k. \tag{4.39}$$

The matrix $\mathcal{R}_{(2J+1)}(n+J,n+J)$ is defined before in the comments following (4.34). Due to the cyclostationary property and $n = Ml - k$, the mean square error can also be written as

$$E|e(n)|^2 = R_{YY}(-k,-k) - 2\,(\,R_{XY}(J-k,-k) \quad \ldots \quad R_{XY}(-J-k,-k)\,)\,\mathbf{h}_k + \mathbf{h}_k^T\,\mathcal{R}_{(2J+1)}(J-k,J-k)\mathbf{h}_k. \tag{4.40}$$

Thus the mean square error is a function of k and not of l. By choosing $\mathbf{h}_k$ properly one can minimize $E|e(Ml-k)|^2$ for all l. As a result $E|e(n)|^2$ is minimized for all time n. The optimal filter coefficients are obtained from the linear equations

$$\mathcal{R}_{2J+1}(J-k,J-k)\mathbf{h}_k = \begin{pmatrix} R_{XY}(J-k,-k) \\ R_{XY}(J-k-1,-k) \\ \vdots \\ R_{XY}(-J-k,-k) \end{pmatrix}. \tag{4.41}$$

From (4.38), one sees that Fig. 4.4 estimates $y(n)$ from the samples $x(n+J)$, $x(n+J-1)\ldots$, $x(n-J)$. These samples are centered around $x(n)$. Thus the estimator in Fig. 4.4 will be different from the one in Fig. 4.3.

In the limiting case where the filter order goes to infinity, the two estimators give the same estimate. This can be shown as follows. By comparing (4.37) with (4.38), we see that the vectors $\mathbf{s}_k$ and $\mathbf{h}_k$ play comparable roles in the estimation of $\hat{y}(n)$. The optimal $\mathbf{s}_k$ is given by (4.36) and $\mathbf{h}_k$ by (4.41). If we increase the filter order for both structures to infinity (i.e., let L in (4.36) and J in (4.41) go to infinity), then $\mathbf{s}_k$ and $\mathbf{h}_k$ become the same except for a linear shift in their components.

The performance of each of the filtering schemes mentioned above can be evaluated through simulations. We shall let the signal $y(n)$ be a WSS auto-regressive Gaussian process. It is generated by passing white Gaussian noise through an all-pole IIR filter. The noise $g(n)$ is chosen to be WSCS with a periodicity of 2. The noise can be generated by interleaving two mutually uncorrelated white Gaussian processes, $g_1(n)$ and $g_2(n)$. The two processes are made to have different variances, σ_1^2 and σ_2^2. (If the two variances are the same, then the combined noise source $g(n)$ will be WSS.)

By fixing σ_1 and σ_2, we shall evaluate each of the filtering schemes by computing the mean square error (MSE) at the output, i.e., $E(y(n)-\hat{y}(n))^2$ where the expected value is approximated by the time averaged value. Let us denote the output MSE due to optimal scalar filtering as σ_{sc}^2, the output MSE due to optimal block filtering (Fig. 4.3) as σ_b^2 and the MSE due to filtering by a multirate filter bank as σ_{fb}^2. With σ_g^2 being the unfiltered noise power (which is taken to be $(\sigma_1^2+\sigma_2^2)/2$), we shall define the normalized MSE for each filtering scheme as, σ_{sc}^2/σ_g^2, σ_b^2/σ_g^2 and σ_{fb}^2/σ_g^2 respectively.

Example 5.1 For a simulation run with $\sigma_1^2 = .5$ and $\sigma_2^2 = .1$, the normalized MSE is plotted in Fig. 4.5 as a function of the filter order. For the scalar filter case,

the filter order is simply the order of the single filter. For the multirate filter bank (Fig. 4.4), the filter order is taken to be the order of any one of the filters, $H_i(z)$. In the block filtering case, we take the filter order to be $MN_B - 1$ where N_B is the length of the block filter $\mathbf{T}(z)$ in Fig. 4.3. As Fig. 4.5 shows, block filtering and the filter bank approach have similar performance, while both methods have a gain of about $3dB$ over the single filter case.

Example 5.2 For cases where the difference between σ_1 and σ_2 is wider, one expects that the interleaved noise process $g(n)$ will be further away from being WSS. Therefore both the filter bank and the block filtering approach will show an increased advantage over the conventional single filter. The simulation results for $\sigma_1^2 = .5$ and $\sigma_2^2 = .005$ is shown in Fig. 4.6. It shows that the filter bank has a noise reduction of $10dB$ over the single filter.

IV. Applications where Cyclostationary Noise Arises

In a multirate QMF bank, noise could arise in each of the sub-band channels. The noise can be due to the encoding of the sub-band signals or due to noise that is commonly present in any communication channel. If one models the noise in each channel as an additive WSS noise source, then the noise at the output of the filter bank is also additive. However, the output noise will in general be WSCS. Let us denote the output noise as $g(n)$. If one makes the assumptions that the synthesis filters are good bandpass filters with non-overlapping passbands and sharp transition bands, and that the noise source in each channel is uncorrelated with the other noise sources, then $g(n)$ would be close to being WSS.

Example 5.3 This is demonstrated by the simulation results in Fig. 4.7, where a two-channel QMF bank is used and the noise at the output of the filter bank

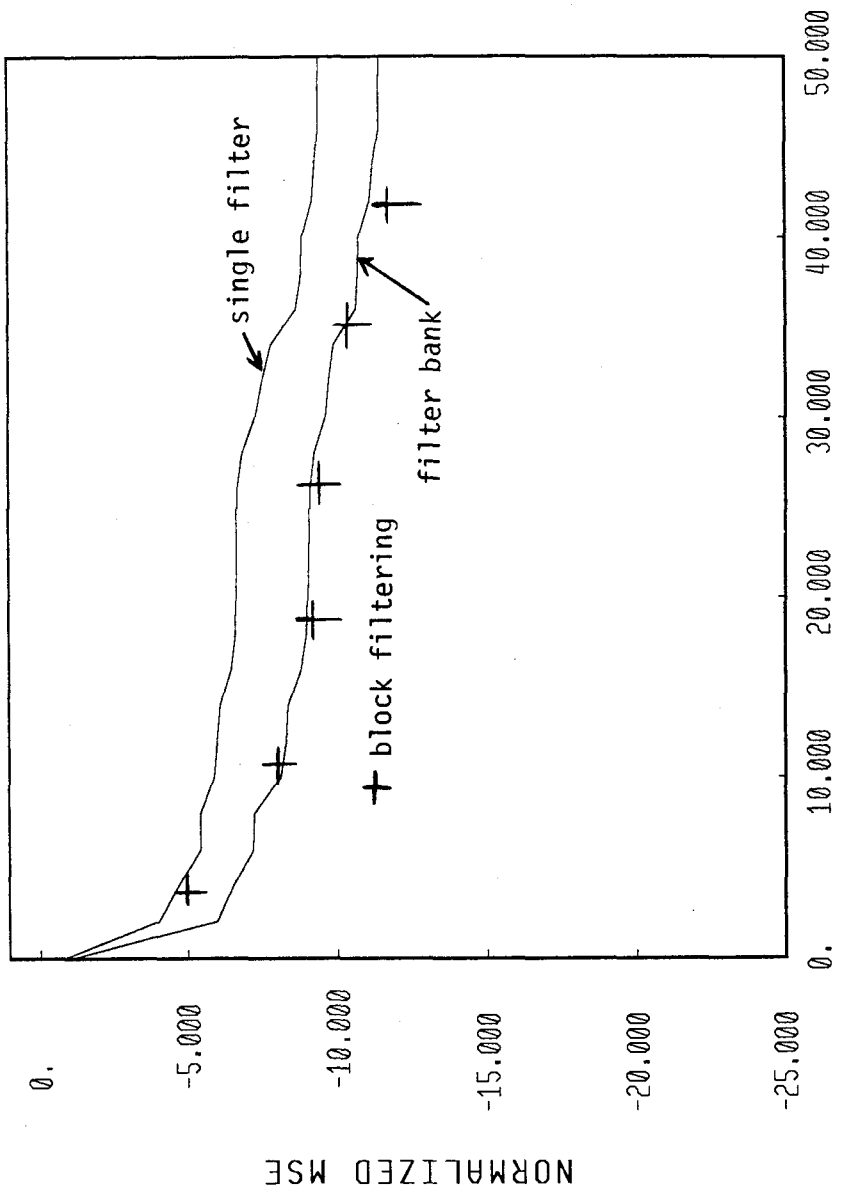

Fig. 4.5. Normalized MSE for various filtering schemes in Example 4.1.

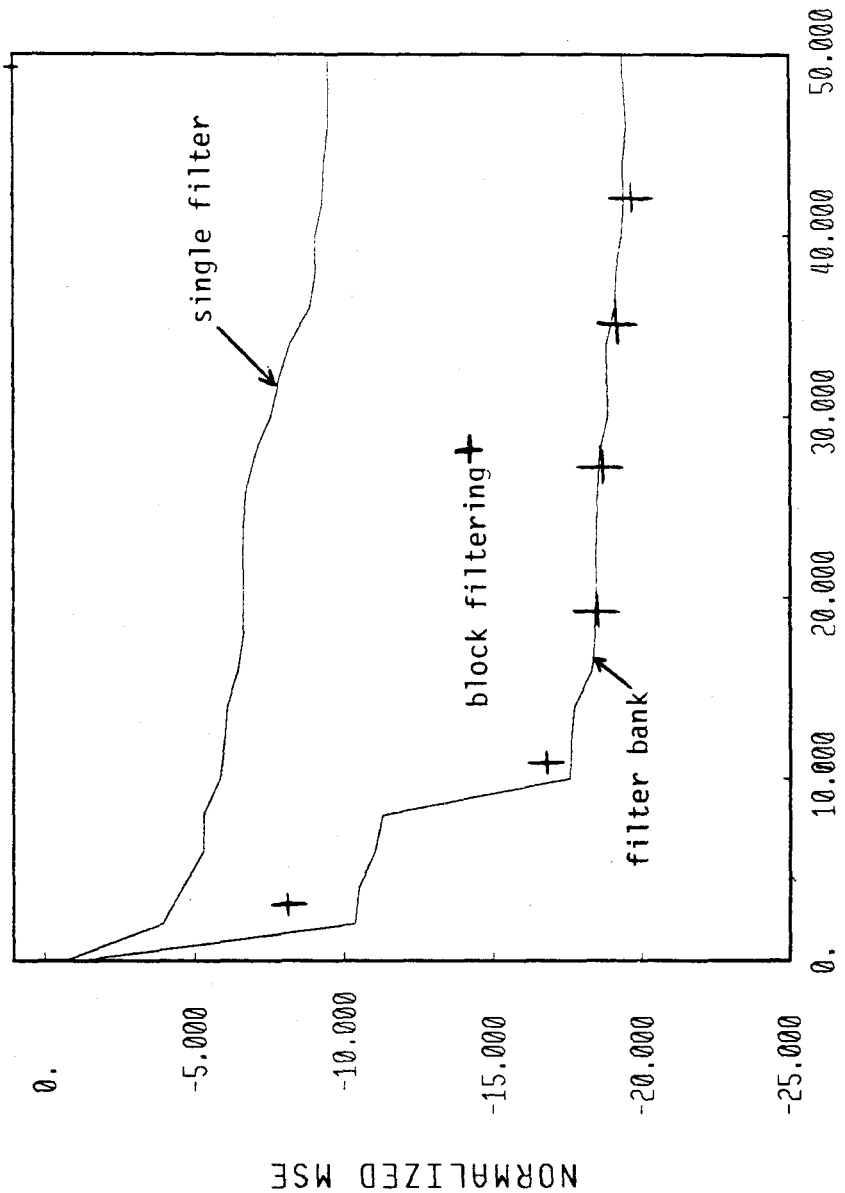

Fig. 4.6. Normalized MSE for various filtering schemes in Example 4.2.

$g(n)$ is generated by adding two uncorrelated white Gaussian noise sources to the two sub-band channels. The signal $y(n)$ is the same as in *Example 5.1* and *5.2*. And one tries to estimate $y(n)$ from $y(n) + g(n)$. The particular two-channel QMF bank used is Filter #48F in [VA88a] where the filters are designed by optimizing the two-channel lossless FIR lattice. The filters have stopband attenuation of $70dB$ and a normalized transition bandwidth of .051. Designating the two noise sources as $g_1(n)$ and $g_2(n)$, their variances are chosen to be: $\sigma_1^2 = .5$ and $\sigma_2^2 = .005$. So just as in *Example 5.2* (see Fig. 4.6), the two noise sources have widely different variances. However, in the QMF case Fig. 4.7 shows that the resulting noise process $g(n)$ is close to being WSS (as measured by the difference between the MSE due to scalar filtering and the MSE due to optimal filter bank).

Example 5.4 If the assumption that the synthesis filters are good bandpass filters is violated, then $g(n)$ is no longer close to being WSS. As an example, we have replaced the QMF bank used in *Example 5.3* by a lower order filter bank, namely Filter #8A in [VA88a]. With the new filter bank, the stopband attenuation is at least $41dB$. The normalized transition bandwidth is .1474. Simulation results are plotted in Fig. 4.8. It shows a wider difference in performance between the optimal scalar filter and the optimal filter bank.

Example 5.5 Going back to the higher order filter bank, we will now show that as the two channel noise sources become correlated the resulting noise $g(n)$ moves away from being WSS. To generate two correlated noise sources, a pair of uncorrelated Gaussian white noise processes is passed through a 2×2 orthogonal matrix building block. If the input noise sources are chosen with different variances, such as $\sigma_1^2 = .5$ and $\sigma_2^2 = .005$, then the outputs from the matrix building block

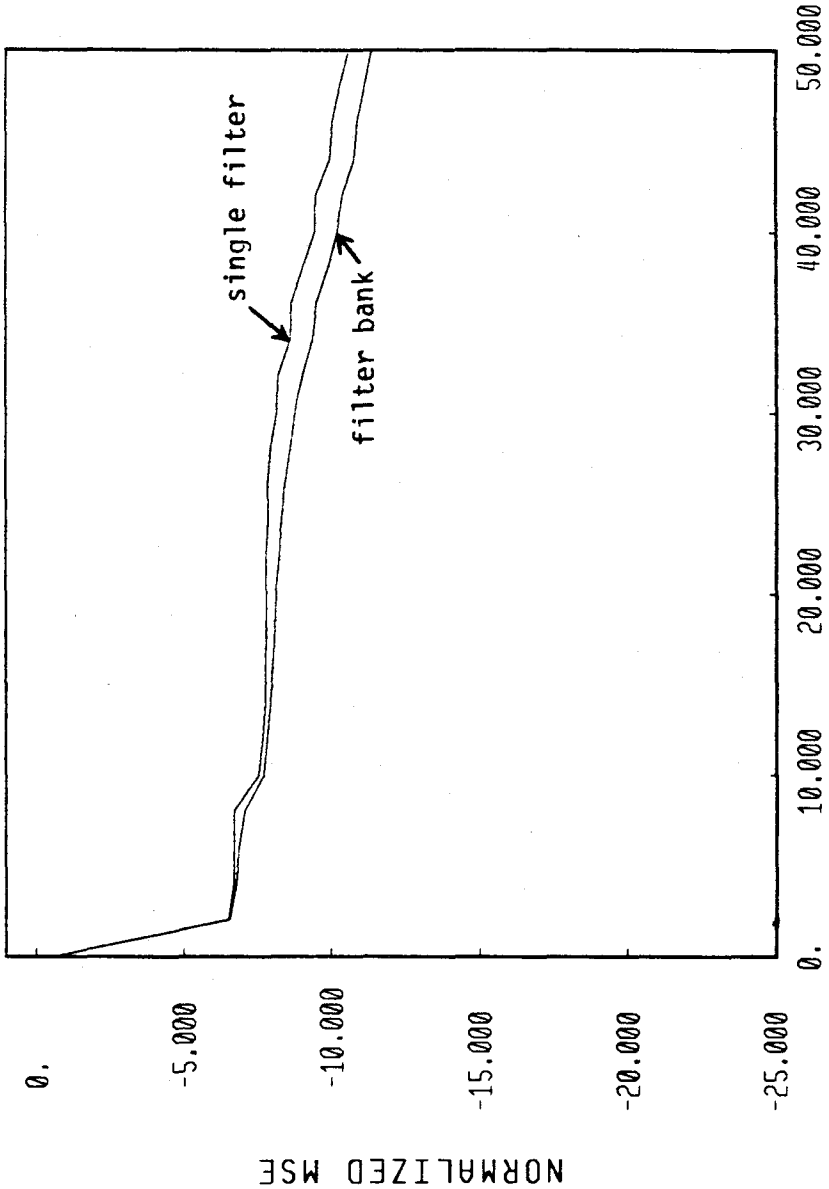

Fig. 4.7. Simulation results for Example 4.3.

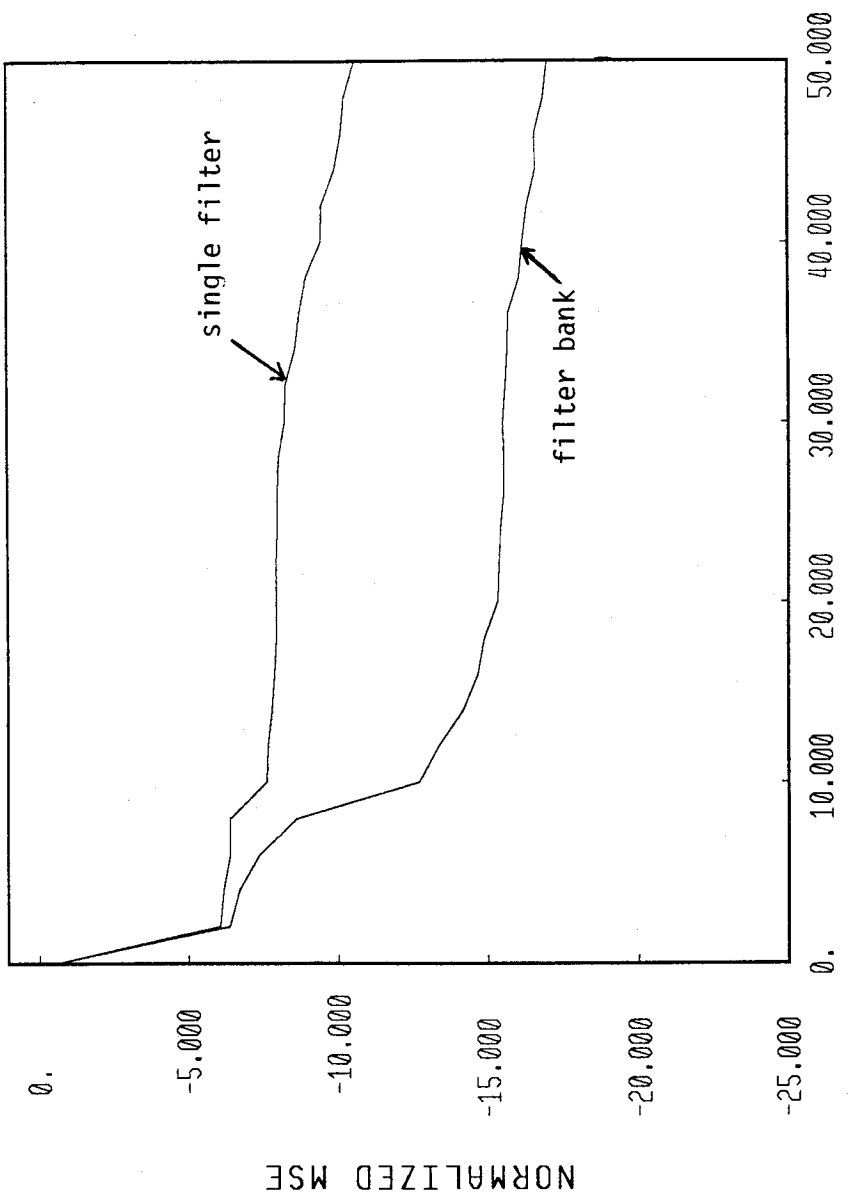

Fig. 4.8. Simulation results for Example 4.4.

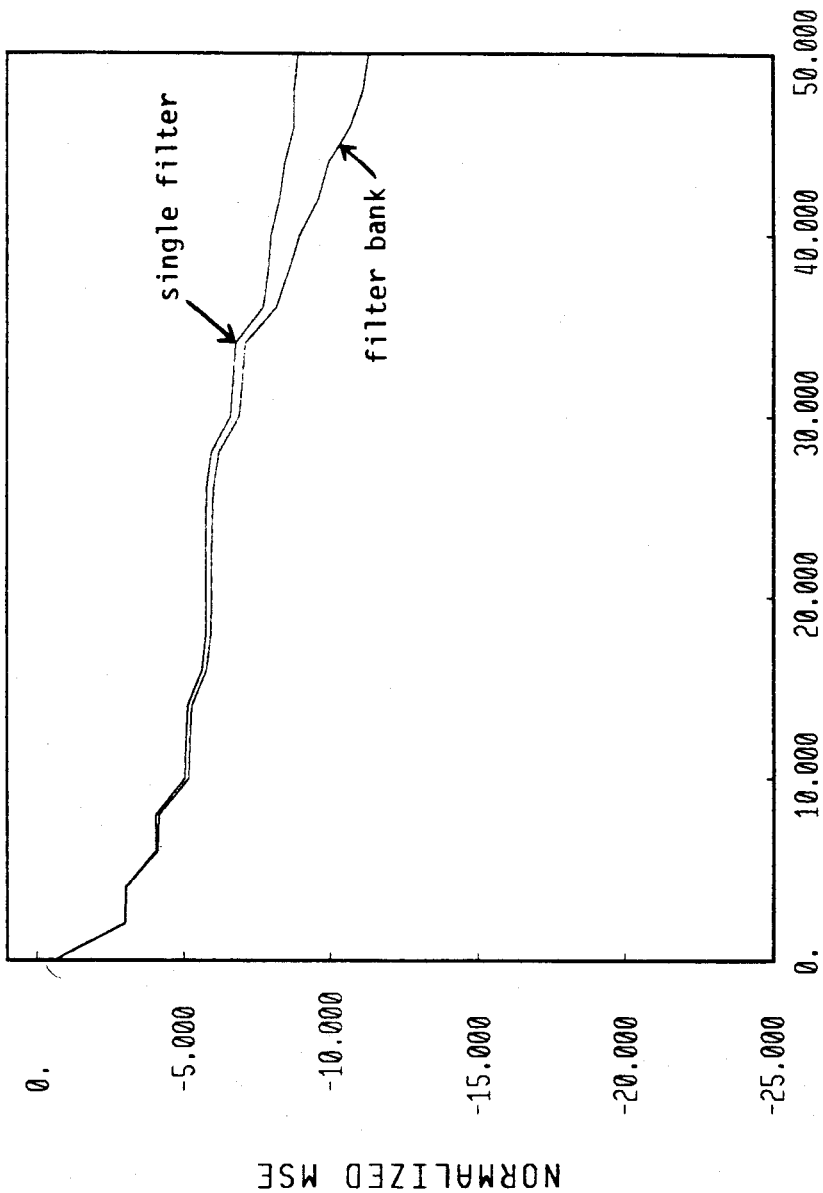

Fig. 4.9. Simulation results for Example 4.5.

will in general be correlated. Simulation shows that if these two correlated noise sources are added to the channels of the QMF bank, then the resulting $g(n)$ is less WSS (Fig. 4.9) as compared to the case with uncorrelated noise sources (Fig. 4.7).

A multirate structure which is closely related to the QMF bank is the digital transmultiplexer [SC81]. In Fig. 4.10 we have drawn in terms of multirate building blocks a structure that converts a time-division-multiplexed (TDM) signal $w(n)$ into a frequency-division-multiplexed (FDM) signal $y(n)$. In Fig. 4.10, the input signal $x(n)$ is the TDM signal whose polyphase components, $x_i(n)$'s, are the individual messages being multiplexed. The signal $x(n)$ is corrupted by additive WSS noise, $a(n)$. The filters $F_i(z)$, similar to the analysis filters in the QMF bank, are good bandpass filters with non-overlapping passbands and sharp transition bandwidths. In general, the signals $x_i(n)$'s could have different statistics. If we assume that each one of them is WSS, then the TDM signal will be WSCS. However, if one further assumes that each message signal $x_i(n)$ is uncorrelated with the other messages (and provided that the filters $F_i(z)$ are good), then the FDM signal, $y(n)$, is close to being a WSS process. Now consider the WSS noise $a(n)$. Except for a few special cases (such as $a(n)$ being white noise), the polyphase components $a_i(n)$ are correlated with each other. Hence, the resulting noise at the output, $b(n)$, is WSCS. The cyclostationarity of $b(n)$ can also be seen from the fact that the filter bank in Fig. 4.10does not satisfy the necessary and sufficient condition in Theorem 4.1 of Sec. 4.2, (otherwise the transmultiplexer can be replaced by a single modulator and a time-invariant filter). Hence, in general a WSS noise source will produce a WSCS output. In conclusion, at the output of the transmultiplexer in Fig. 4.10, the signal is WSS while the noise is WSCS.

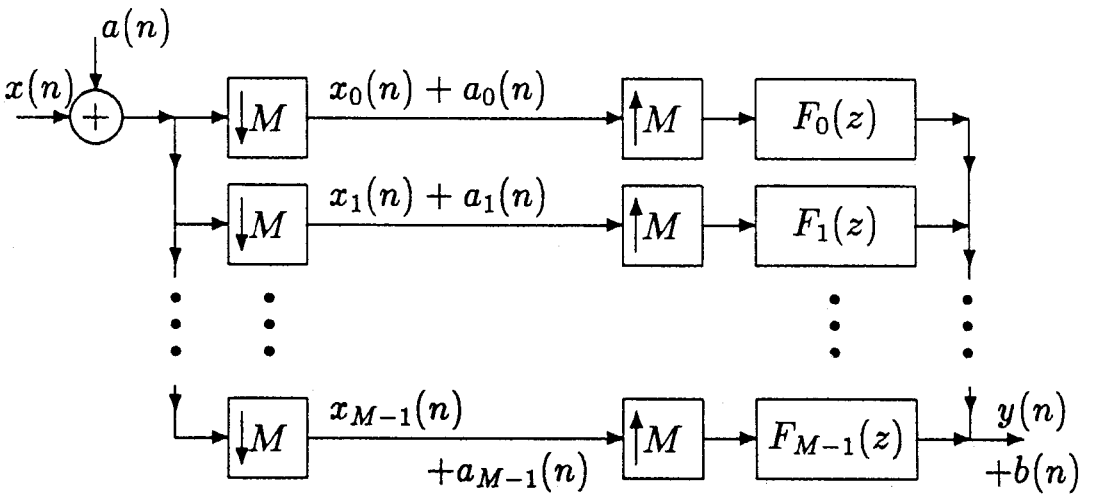

Fig. 4.10. Conversion from TDM signal $x(n)$ to FDM signal $y(n)$.

In Fig. 4.11, the conversion of a FDM signal into a TDM signal is depicted. The FDM signal, $y(n)$, is corrupted by WSS noise due to transmission over a communications channel. The filter bank structure in Fig. 4.11 does not satisfy the condition in Theorem 4.1, therefore the output noise $d(n)$ is in general WSCS.

As a second application, let us consider the sampling of a continuous waveform by an analog-to-digital converter. Let $y_a(t)$ be the continuous-time signal to be sampled at the rate of $\frac{1}{T}$ samples/sec. However, due to timing error in the A/D converter the samples obtained are not $y_a(nT)$ but instead are $y_a(nT+\tau(n))$ where $\tau(n)$ is a random process. This can be modeled as in Fig. 4.12 with the input signal $y_a(t)$ being multiplied by a train of impulses $\sum_{n=-\infty}^{\infty}\delta(nT+\tau(n))$. The output which is taken to be a discrete-time sequence is denoted as $x(n) = y_a(nT+\tau(n))$. We shall assume that the timing error $\tau(n)$ is second order stationary (SOS). That means its second order probability density function $f_T(\tau_1,\tau_2;n_1,n_2)$ is a function of $n_1 - n_2$ only. Assuming further that $x_a(t)$ is WSS and the signal $x_a(nT)$ is independent of the error $\tau(n)$, then the error introduced by $\tau(n)$ can be modeled as a WSS

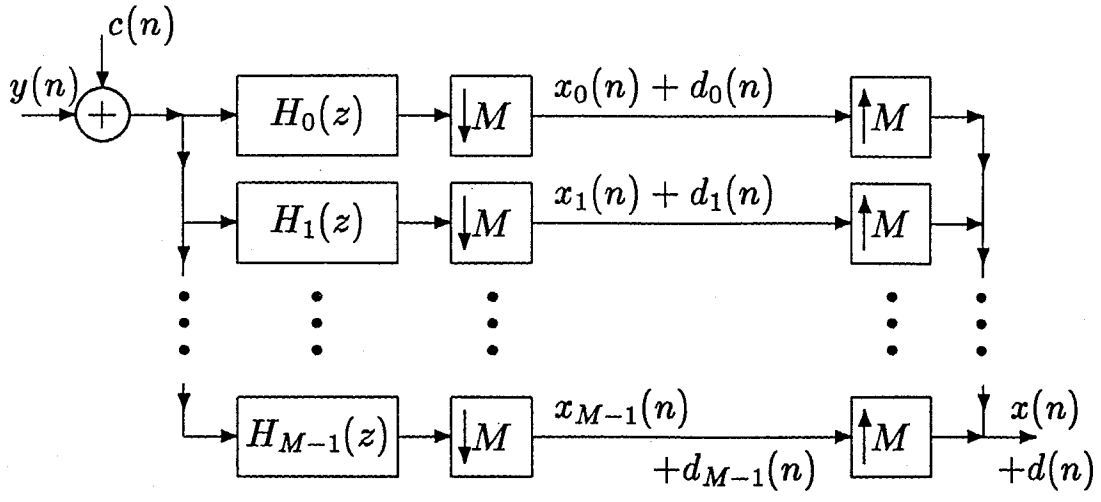

Fig. 4.11. Conversion from FDM signal $y(n)$ to TDM signal $x(n)$.

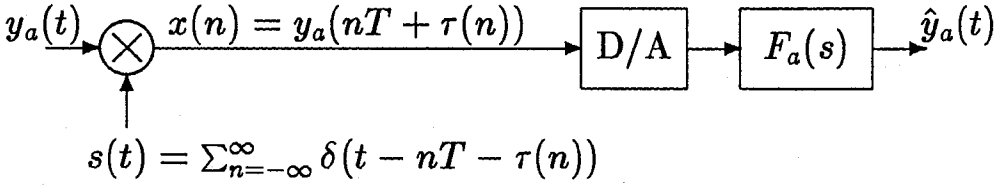

Fig. 4.12. Reconstruction of a randomly sampled analog waveform.

random process. We can write $x(n)$ as $y(n) + g(n)$ where $g(n)$ is the noise term, and $y(n), g(n)$ are jointly WSS.

The interpolation is done by converting $w(n)$ back into an impulse train in continuous-time and then filtered by an interpolation filter $F_a(s)$. The resulting output $\hat{y}_a(t)$ will be jointly WSS with $y_a(t)$. Since $y_a(t)$ is bandlimited (in the sense that $R_{YY}(t_0 + t, t_0)$ as a function of t is bandlimited to $|\Omega| \leq \frac{\pi}{T}$), the interpolation filter should satisfy

$$F_a(j\Omega) = 0 \quad \text{for } |\Omega| > \frac{\pi}{T}. \tag{4.42}$$

The passband of the filter can be chosen so that $E(\hat{y}_a(t) - y_a(t))^2$ is minimized.

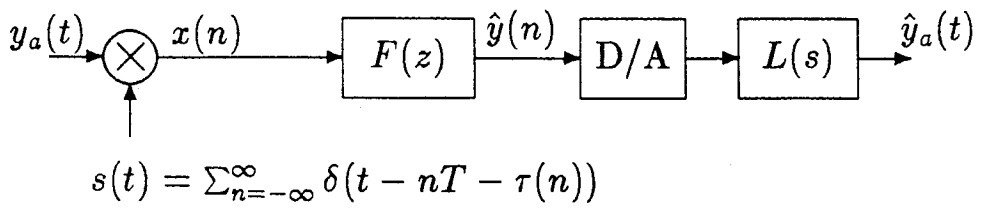

Fig. 4.13. Reconstruction is performed using a digital optimal filter $F(z)$ followed by an ideal lowpass filter $L(s)$.

The filter $F_a(s)$ in Fig. 4.12 can be redrawn as in Fig. 4.13 where $F_a(s)$ is separated into a digital filter $F(z)$ and an ideal lowpass filter $L(s)$. The passband of $L(s)$ is $|\Omega| < \frac{\pi}{T}$. The only function for $L(s)$ is to interpolate the discrete-time signal so that the output is in continuous time. The optimal filtering is provided by $F(z)$ alone. Let the output of $F(z)$ be $\hat{y}(n)$, then minimizing the mean square error $E(\hat{y}_a(t) - y_a(t))^2$ is the same as minimizing $E(\hat{y}(n) - y(n))^2$. So the problem is equivalent to estimating $y(n)$ from $x(n)$ in the WSS case.

In situations where the sampling rate required for A/D conversion is high, it might be more cost effective to use several A/D converters in parallel each operating at a slower rate. With M A/D converters, each of them would only be required to operate at a sampling rate of $\frac{M}{T}$ where $\frac{1}{T}$ is the original sampling rate, and each unit would have a time offset of $\frac{k}{T}$ where $0 \leq k \leq M-1$. Let the samples taken by the kth converter be denoted as $x_k(n)$, we can write $x_k(n) = y_a(nMT - kT + \tau_k(n))$ where $\tau_k(n)$ represents the timing error for the kth converter. By interleaving the samples from the M converters, one obtains a discrete-time signal $x(n)$ that is at the higher sampling rate. The signal $x(n)$ can be written as $x(n) = y_a(nT + \tau(n))$

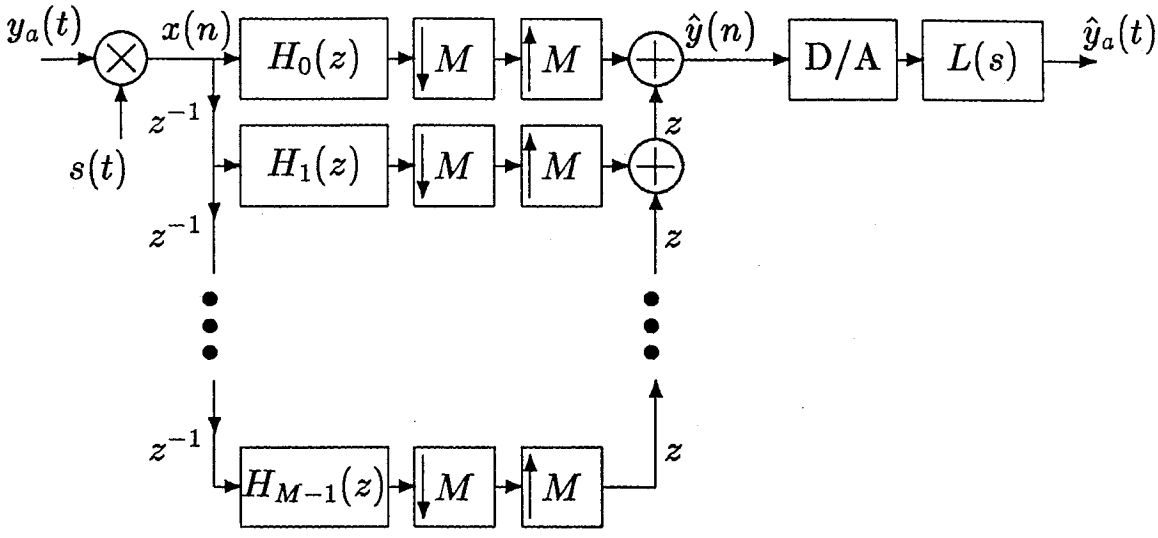

Fig. 4.14. A filter bank used to reconstruct $y_a(t)$.

and $\tau(n)$ is a random process given by

$$\tau(n) = \tau_k(\frac{n+k}{M}) \qquad \text{where } k = ((M-n)). \tag{4.43}$$

If one assumes that the processes $\tau_k(n)$ $(0 \leq k \leq M-1)$ are all jointly second order stationary, then $\tau(n)$ becomes a second order cyclostationary process. As a result, we may model $x(n)$ as $x(n) = y(n) + g(n)$ where $g(n)$ being the noise due to the timing error is WSCS. To recover $y_a(t)$, we can first estimate $y(n)$ using the methods discussed previously for WSCS processes. The estimated signal $\hat{y}(n)$ is then passed through a lowpass interpolation filter to obtain $\hat{y}_a(t)$. This is illustrated in Fig. 4.14 where a filter bank is used for the estimation.

4.5 Conclusion

In Sec. 4.2, we see that the QMF bank will preserve the wide-sense stationarity of its input if and only if the filter bank satisfies a modified set of alias-cancellation conditions. The conditions state that among all the aliasing terms plus the original

term only one term should remain at the output. This is equivalent to modulating the input by $e^{j\frac{2\pi k}{M}n}$ and then passing it through a linear time-invariant filter.

In Sec. 4.3, the problem of estimating a WSCS random process signal from another WSCS random process is addressed. We show that the multirate filter bank can be used as the optimal estimation filter. Such an estimator is different from the one commonly obtained through block filtering approach. Both methods converge to the same theoretical optimal solution when the filter order is taken to infinity. The above estimation problem could arise in the analog-to-digital conversion of continuous-time waveforms when several A/D converters are used in parallel in order to achieve a higher sampling rate.

Chapter V. Roundoff Errors Generated by Orthogonal Matrix Building Blocks

In the study of QMF banks, a recent trend has been towards the use of lossless systems in the design and implementation of the analysis/sythesis filter banks [VA89]. Lossless systems also have applications in the area of low sensitivity filter design, orthogonal filters and lattice filters [DEW80] [HE83] [RA84] [VA85]. In [VA88c], it is shown how orthogonal matrices can be used as the basic building blocks for realizing lossless systems. In the analysis of roundoff noise in conventional digital signal processing structures, a commom assumption made by most researchers [JAC70] [MU76] [BA85] have been that the noise can be modeled as an additive white Gaussian noise which is uncorrelated to the signal being quantized, and each noise source is uncorrelated with the rest. For the case of orthogonal building blocks, some have argued that due to the special structure of the orthogonal matrix, the output signals from an orthogonal matrix building block could be correlated. Hence the roundoff errors produced by the quantization of these output signals could have a nonzero cross-correlation. A nonzero cross-correlation will be of some siginficance, for it can affect the noise power estimate at the output of the overall system. In this chapter, we will investigate numerically this cross-correlation between the roundoff errors.

5.1. Introduction

In this chapter, we are interested in the roundoff errors produced by the multiplication of a vector by an orthogonal matrix. In particular, one would like to know if the errors at the outputs are correlated to each other. The type of quantization being considered here is fixed point rounding. In other words, if B is the number

of bits used for representing numbers in the range of $[0,1]$, then any real number a in that range will have a roundoff value of

$$Q(a) = \begin{cases} 2^{-B}K & \text{if } 2^B a - \lfloor 2^B a \rfloor < \frac{1}{2}; \\ 2^{-B}(K+1) & \text{otherwise,} \end{cases} \tag{5.1}$$

where $K = \lfloor 2^B a \rfloor$. The step size Δ for the above quantization scheme is defined to be 2^{-B}. Fig. 5.1 shows the case of a 2×2 orthogonal matrix, where $(x_1, x_2)^T$ is the input vector quantized to B bits. The result of the matrix-vector multiplication, $(y_1, y_2)^T$, is then quantized to B bits also. The errors generated are defined to be $\epsilon_1 = \hat{y}_1 - y_1$ and $\epsilon_2 = \hat{y}_2 - y_2$. In the analysis of the roundoff errors, we shall make use of the results in [BA85].

We are interested in orthogonal matrices here, because they appear in several low-sensitivity implementations of digital filters, such as the Gray-Markel lattice [GR73], orthogonal filters [DEW80] [HE83] [RA84] [VA85], and the lossless FIR lattice structure [VA86]. Applications of these orthogonal matrices in multirate filter banks have also been reported recently [VA87a], [VA88a]. For example, consider the lossless FIR lattice as shown in Fig. 5.2. The location of the quantizers are exactly at the output of each 2×2 orthogonal matrix (denoted in the figure by $\mathbf{R}_i$). According to common assumptions [JAC70] [MU76], the noise generated at each quantizer is assumed to be uncorrelated to all the other noise sources. However

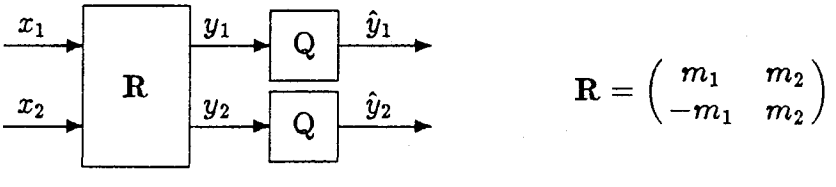

Fig. 5.1. A 2×2 orthorgonal building block with quantized outputs.

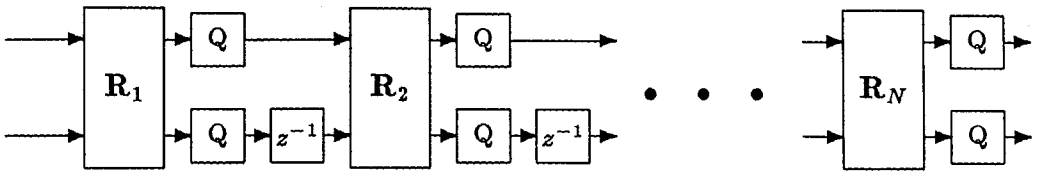

Fig. 5.2. The lossless FIR lattice with quantizers.

since the entries of the 2×2 orthogonal matrix are related to each other as sines and cosines of the same angle θ, one might think that the roundoff errors would be correlated in some ways. In this paper, this issue is addressed in a quantitative way, and the answer is provided by simulation results.

5.2. The 2×2 Orthogonal Block

Consider the case of a signal passing through a multiplier followed by a quantizer with quantization level Δ, where $\Delta = 2^{-B}$, as shown in Fig 5.3. The error produced by the quantization process had been analyzed in [BA85]. The input u has a quantization level of Δ also. Let the multiplier value be N/L where N is an odd integer and L is a positive power of two. Thus the product y has a quantization level of $N\Delta/L$, which is then rounded off to B bits, producing $\hat{y}$. The quantization error ϵ (defined to be $\hat{y} - y$) can be expressed in terms of the input u as [BA85]

$$\epsilon = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}u} \tag{5.2}$$

The sequence $\mathcal{E}_{L,N}(k)$ contains the DFT coefficients to the periodic sequence

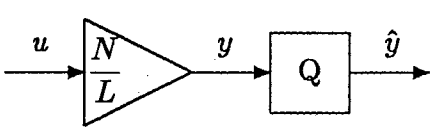

Fig. 5.3. A single multiplier followed by a quantizer.

$\{\frac{\Delta}{L}l(n)\}$ where $l(n)$ is the modulo L solution to the equation $l(n)+nN=0$ Mod L [BA85]. An explicit formula for $\mathcal{E}_{L,N}(k)$ is given by [BA85]

$$\mathcal{E}_{L,N}(k)=\begin{cases}\frac{\Delta}{2L} & \text{when } k\equiv 0 \text{ Mod } L;\\ (-1)^k\frac{\Delta}{2L}\Big(1+j\cot(\pi m(k)/L)\Big) & \text{otherwise,}\end{cases} \tag{5.3}$$

with $m(k)$ being the unique modulo-L solution to the equation $m(k)N+k=0$ Mod L.

In this paper, we shall apply (5.2) to the analysis of roundoff errors generated by orthogonal matrices. The case of the 2×2 orthogonal matrix is shown in Fig. 5.1. The vector $\mathbf{y}=(y_1,y_2)^T$ is related to the input by

$$\mathbf{y}=\mathbf{R}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}m_1 & m_2\\ -m_2 & m_1\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix} \tag{5.4}$$

The entries m_1 and m_2 are the quantized values of $\cos(\theta)$ and $\sin(\theta)$ repectively, where θ represents the angle of rotation for the orthogonal matrix $\mathbf{R}$. Assume the level of quantization for these multipliers is $\Delta_1=2^{-B_1}$, then m_1 can be written as $M_1M\Delta_1$ and m_2 as $M_2M\Delta_1$ with M_1 and M_2 being relatively prime integers. Since x_1 and x_2 are quantized to B bits, let $x_1=n_1\Delta$ and $x_2=n_2\Delta$, then we get

$$\begin{aligned}y_1&=(n_1M_1+n_2M_2)M\Delta_1\Delta\\ y_2&=(-n_1M_2+n_2M_1)M\Delta_1\Delta\end{aligned} \tag{5.5}$$

Due to the fact that M_1 and M_2 are relatively prime, there exist integers n_1 and n_2 to make $n_1M_1+n_2M_2=1$. This means y_1 has a quantization level of $M\Delta_1\Delta$. The same holds for y_2. Both needs to be quantized back to a level of Δ, producing $\hat{y}_1$ and $\hat{y}_2$. Defining two numbers $u_1=y_1/(M\Delta_1)$ and $u_2=y_2/(M\Delta_1)$, it is clear that both u_1 and u_2 have B bits to the right of the binary point. The case of the

single multiplier in Fig. 5.3 becomes applicable here by making the identification of u_1 with u, y_1 with y and $M\Delta_1$ with the multiplier value $\frac{N}{L}$. As a result, (5.2) becomes

$$\epsilon_1 = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}u_1} \tag{5.6}$$

Similarly, the expression for ϵ_2 is

$$\epsilon_2 = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}u_2} \tag{5.7}$$

Assume that ϵ_1 and ϵ_2 have zero mean. From (5.6) and (5.7), expressions for the variances and cross-correlation between ϵ_1 and ϵ_2 can be found.

$$\begin{aligned} E[\epsilon_1^2] &= E\left[\sum_{k_1,k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) e^{j\frac{2\pi}{L\Delta}(k_1+k_2)u_1}\right] \\ &= \sum_{k_1,k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) E\left[e^{j\frac{2\pi}{M\Delta_1 L\Delta}(k_1+k_2)y_1}\right] \\ &= \sum_{k_1,k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) E\left[e^{j\frac{2\pi(k_1+k_2)}{L\Delta}(M_1x_1+M_2x_2)}\right]. \end{aligned} \tag{5.8}$$

Defining the joint characteristic function of x_1 and x_2 as

$$\Phi(\omega_1,\omega_2) = E\left[e^{j(\omega_1x_1+\omega_2x_2)}\right] \tag{5.9}$$

then (5.8) becomes

$$E[\epsilon_1^2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2)\Phi\left(\frac{2\pi M_1(k_1+k_2)}{L\Delta}, \frac{2\pi M_2(k_1+k_2)}{L\Delta}\right). \tag{5.10}$$

Using (5.7), an expression for $E[\epsilon_2^2]$ can also be found

$$E[\epsilon_2^2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2)\Phi\left(-\frac{2\pi M_2(k_1+k_2)}{L\Delta}, \frac{2\pi M_1(k_1+k_2)}{L\Delta}\right). \tag{5.11}$$

Similarly, the cross-correlation between the two errors ϵ_1 and ϵ_2 is

$$E[\epsilon_1\epsilon_2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{N,L}(k_1)\mathcal{E}_{N,L}(k_2) \times$$

$$\Phi\left(\frac{2\pi(M_1k_1 - M_2k_2)}{L\Delta}, \frac{2\pi(M_2k_1 + M_1k_2)}{L\Delta}\right). \tag{5.12}$$

The common assumption being made about roundoff errors in digital filters is that they are uncorrelated with each other and the error is uncorrelated with the signal. For the case of the 2×2 orthogonal matrix, we will see how the cross-correlation of its errors behave in terms of the cross-correlation between the two inputs. Let the cross-correlation coefficient between ϵ_1 and ϵ_2 be defined as

$$\rho_{\epsilon_1,\epsilon_2} = \frac{E[\epsilon_1\epsilon_2]}{E[\epsilon_1^2]^{\frac{1}{2}}E[\epsilon_2^2]^{\frac{1}{2}}}. \tag{5.13}$$

Obviously, $\rho_{\epsilon_1,\epsilon_2}$ will depend on the probability distribution of x_1 and x_2. However, it is not clear how $\rho_{\epsilon_1,\epsilon_2}$ is related to the correlation of the two inputs. If (5.13) is small enough for most inputs commonly encountered, then one may safely assume that ϵ_1 and ϵ_2 are uncorrelated.

In order to calculate (5.13), assumptions need to be made about the joint probability distribution of x_1 and x_2. Let us assume that the joint probability density of x_1 and x_2 follows a Gaussian envelope,

$$f_{X_1,X_2}(x_1,x_2) = B \sum_{m,n=-\infty}^{\infty} \delta(x_1 - m\Delta, x_2 - n\Delta) \times e^{-(x_1^2 - 2\rho_0 x_1 x_2 + x_2^2)/2\sigma_0^2(1-\rho_0^2)}. \tag{5.14}$$

The constant B is chosen such that the total probability is normalized to one. σ_0 is the variance of the Gaussian envelope. It is different from the variance of x_1 and of x_2. Similarly, ρ_0 is not the same as ρ_{x_1,x_2}.

For the input distribution given in (5.14), one can verify that the errors ϵ_1 and ϵ_2 have zero means. Furthermore, if the inputs are uncorrelated, i.e., $\rho_0 = 0$ in (5.14), then the errors are automatically uncorrelated as well.

From (5.14), the characteristic function $\Phi(\omega_1, \omega_2)$ of x_1, x_2 can be found, and substituting it into (5.10) through (5.12) the variances and the cross-correlation of the two errors can be calculated. Note that for a given distribution function, such as the one in (5.13), the quantities $E[\epsilon_1^2]$, $E[\epsilon_2^2]$ and $E[\epsilon_1 \epsilon_2]$ depends on the ratio σ_0/Δ. They do not depend on σ_0 and Δ individually. Therefore, we will use σ_0/Δ as a parameter for the input distribution. This is commonly referred to as the dynamic range of the input.

With a quantization level of $\Delta = \Delta_1 = 2^{-8}$, we shall examine the behavior of $\rho_{\epsilon_1,\epsilon_2}$ as the rotational angle and the dynamic range vary. With a dynamic range of $(\sigma_0/\Delta) = 5$, the correlation coefficient $\rho_{\epsilon_1,\epsilon_2}$ is plotted in Fig. 5.4 as a function of θ for several values of ρ_0. As Fig. 5.4 shows, the cross-correlation between the two errors can be quite high for certain angles (such as $\theta = 36°$ and $38°$). Also, the correlation between the errors is large when the orthogonal matrix has small angles of rotation. Fig. 5.5 plots $\rho_{\epsilon_1,\epsilon_2}$ as a function of ρ_0 for small angles of θ. Here, for small enough angle ($\theta \leq 3°$ the error cross-correlation becomes comparable to the cross-correlation of the inputs.

As we go to inputs with larger dynamic range, the magnitude of $\rho_{\epsilon_1,\epsilon_2}$ tends to decrease, with the exception of a few particular values of θ. Fig. 5.6 represents the case of $\frac{\sigma_0}{\Delta} = 10$ and Fig. 5.7 shows the result for $\frac{\sigma_0}{\Delta} = 25$. Further numerical computation shows that $\rho_{\epsilon_1,\epsilon_2}$ goes to zero, as the dynamic range increases.

5.3. General Orthogonal Blocks

The expression in (5.6) and (5.7) can be generalized to any $K \times K$ orthogonal

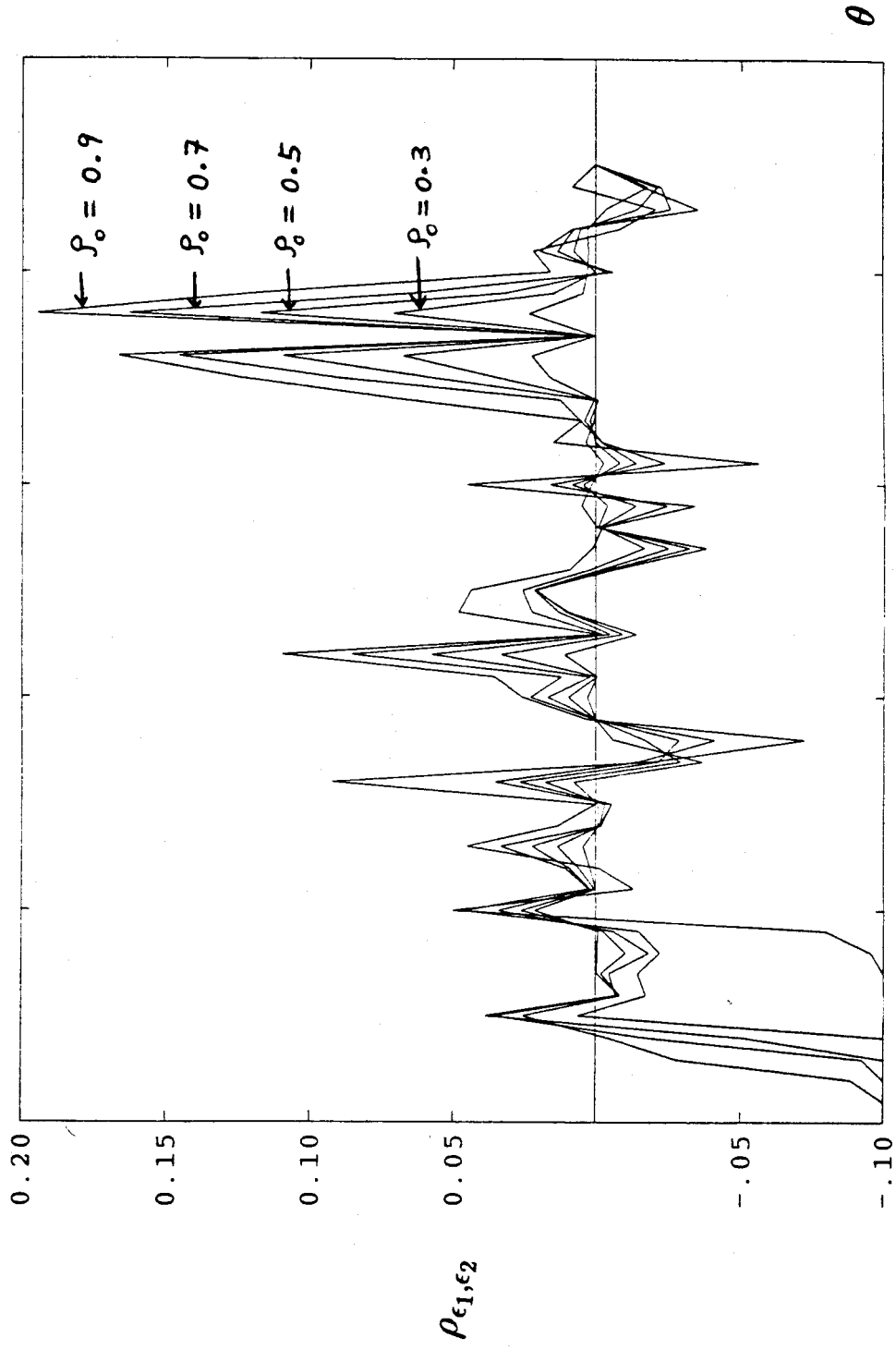

Fig. 5.4. $\rho_{\epsilon_1,\epsilon_2}$ as a function of the rotational angle with dynamic range ($\sigma_0/\Delta = 5$).

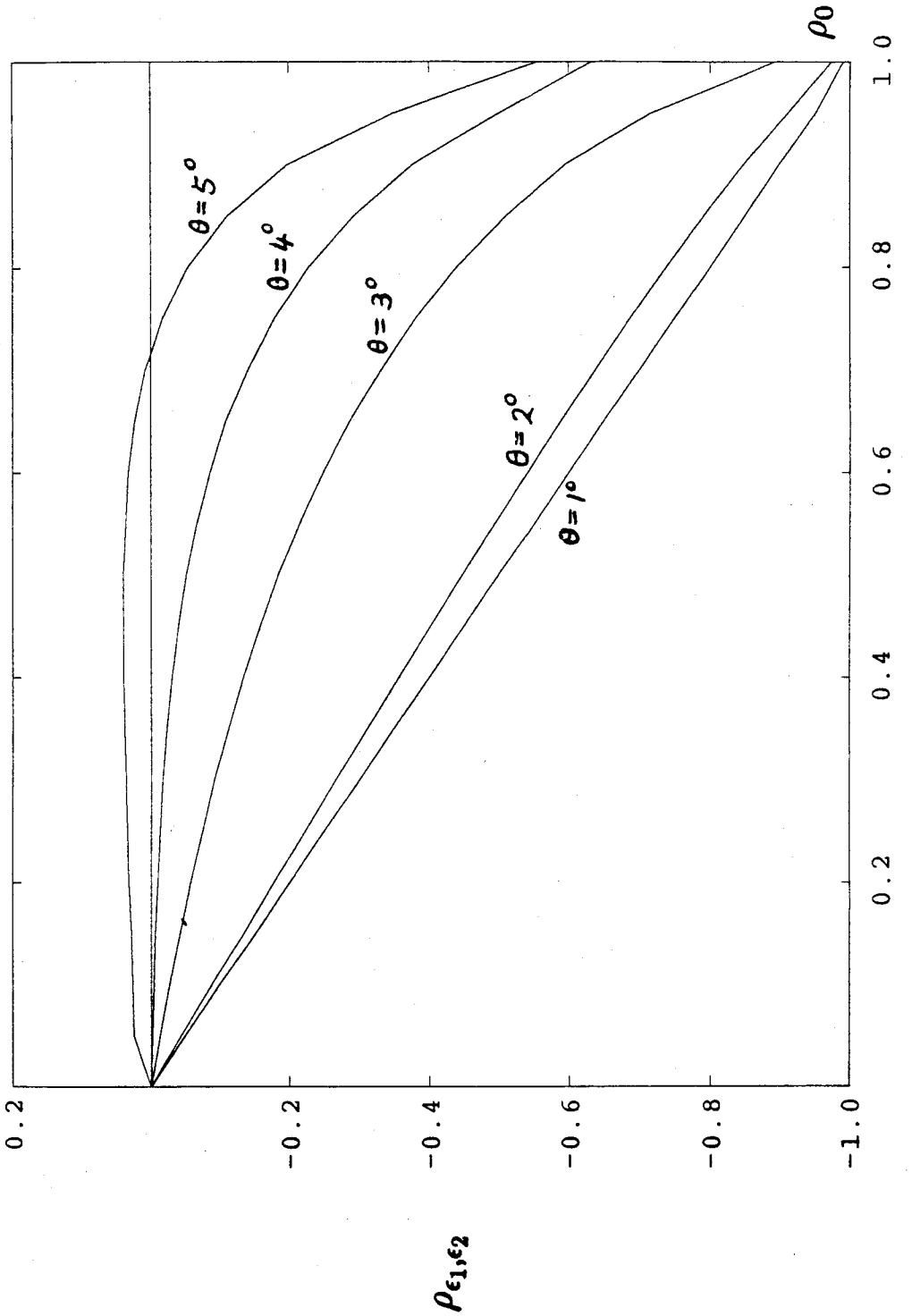

Fig. 5.5. $\rho_{\epsilon_1,\epsilon_2}$ as a function of ρ_0 for small angles of rotation ($\sigma_0/\Delta = 5$).

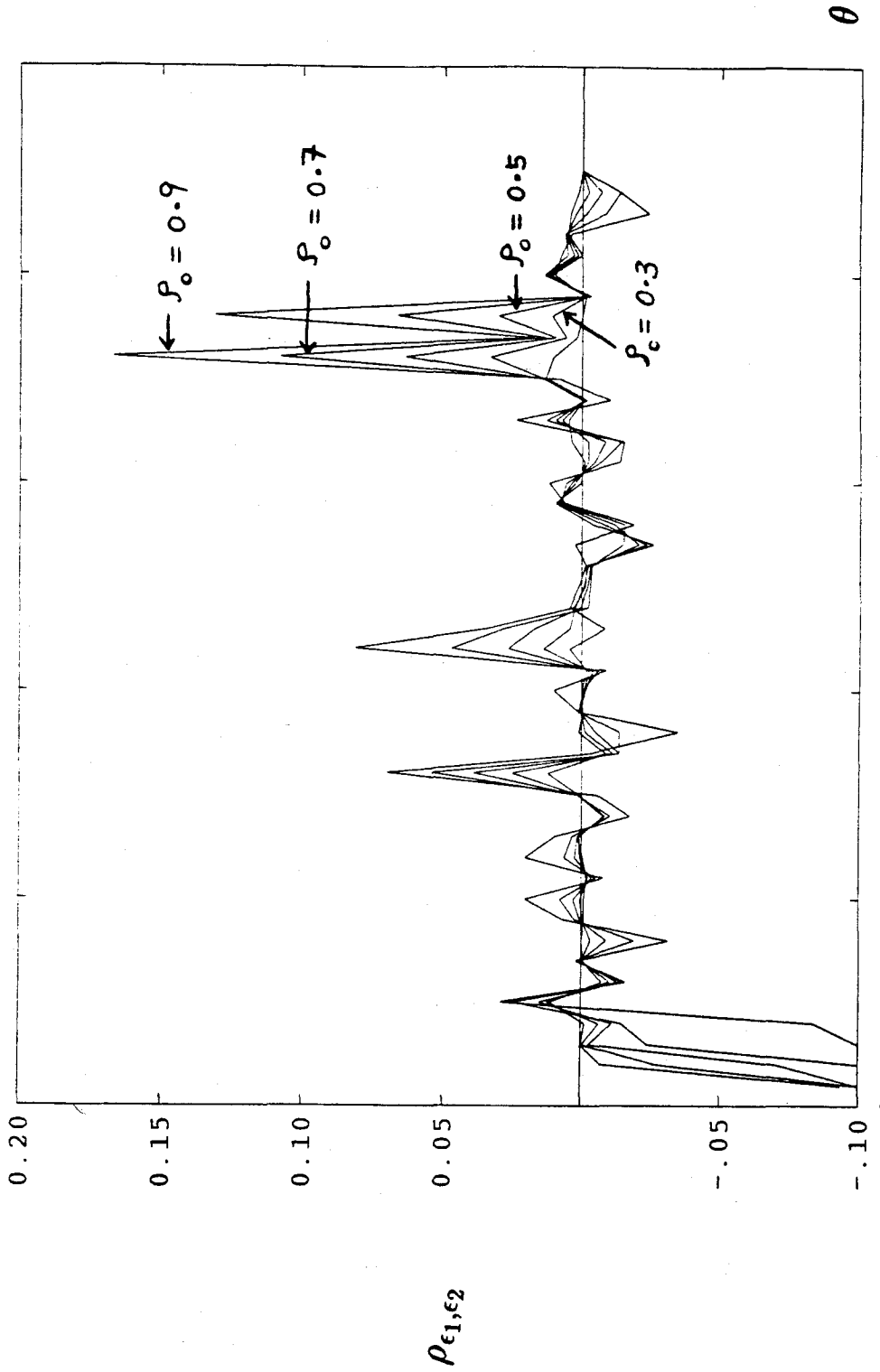

Fig. 5.6. $\rho_{\epsilon_1,\epsilon_2}$ as a function of the rotational angle with dynamic range ($\sigma_0/\Delta = 10$).

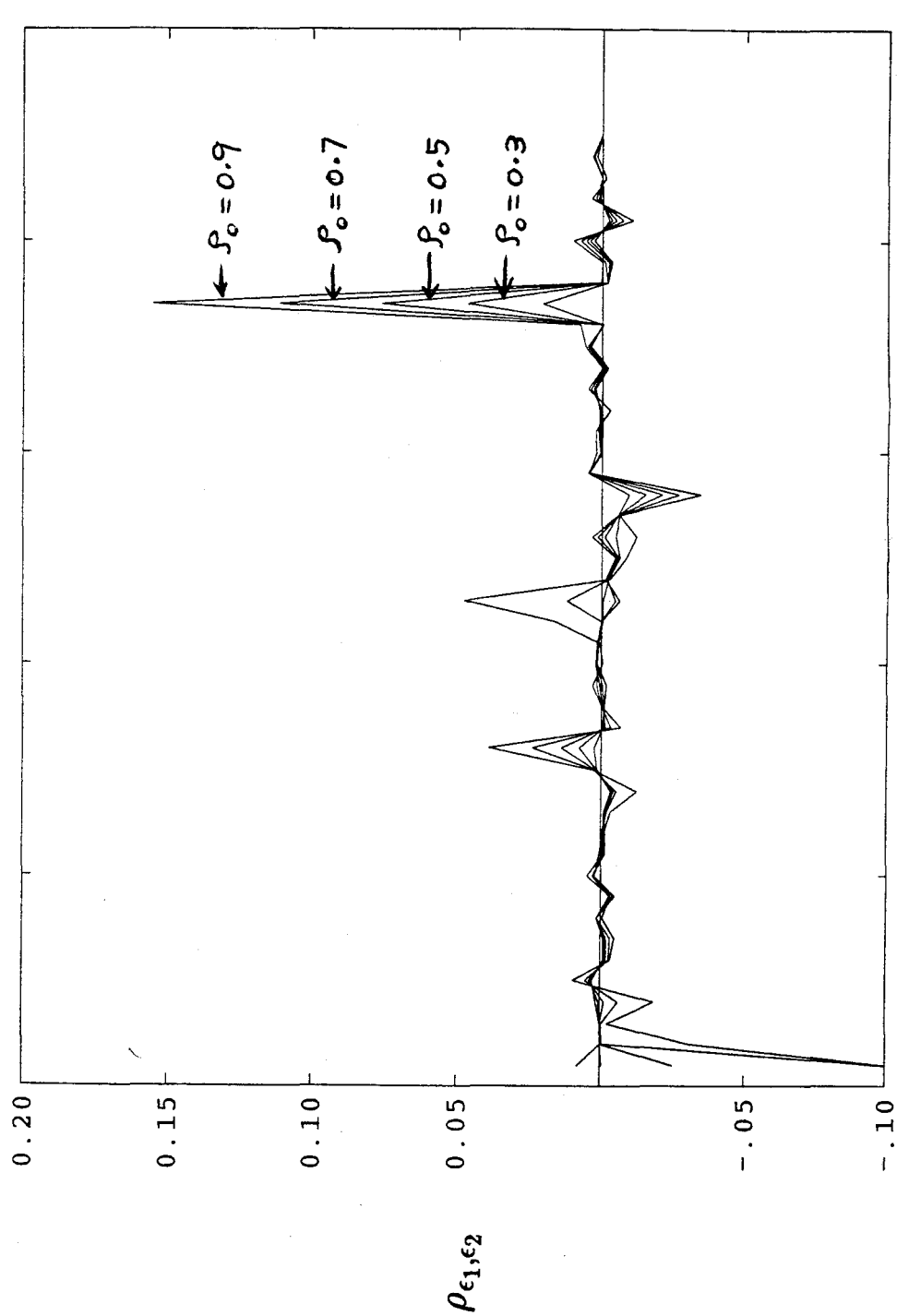

Fig. 5.7. $\rho_{\epsilon_1,\epsilon_2}$ as a function of the rotational angle with dynamic range ($\sigma_0/\Delta = 25$).

matrix. With $\mathbf{x} = (x_1, \ldots, x_K)^T$ as the input, the unquantized output is

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix} = \mathbf{R}\mathbf{x} \tag{5.15}$$

Assuming each entry of $\mathbf{R}$ is quantized to B_1 bits, we can write these entries as $[\mathbf{R}]_{i_1,i_2} = M_{i_1,i_2}M\Delta_1$, where the $\gcd(M_{i_1,i_2}) = 1$. The quantization level of $\mathbf{y}$ is at $M\Delta_1\Delta$. Let $M\Delta_1 = \frac{N}{L}$ as in (5.2), then following argument similar to the 2×2 case, the error ϵ_i can be written as

$$\epsilon_i = \sum_{k=0}^{L-1} \mathcal{E}_{L,N}(k) e^{j\frac{2\pi k}{L\Delta}\mathbf{x}^T\mathbf{m}_i}, \tag{5.16}$$

where $\mathbf{m}_i \triangleq (M_{i,1} \quad \ldots \quad M_{i,K})^T$. Defining the joint characteristic funtion of $\mathbf{x}$ to be $\Phi_K(\mathbf{w}) = E[e^{j\mathbf{w}^T\mathbf{x}}]$ with $\mathbf{w} = (\omega_1, \ldots, \omega_K)^T$, the error variance is given by

$$E[\epsilon_i^2] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{L,N}(k_1)\mathcal{E}_{L,N}(k_2) \times \Phi_K\left(\frac{2\pi(k_1+k_2)}{L\Delta}\mathbf{m}_i\right). \tag{5.17}$$

Similar to (5.12), the cross-correlation between any two errors is

$$E[\epsilon_{i_1}\epsilon_{i_2}] = \sum_{k_1=0}^{L-1}\sum_{k_2=0}^{L-1} \mathcal{E}_{L,N}(k_1)\mathcal{E}_{L,N}(k_2) \times \Phi_K\left(\frac{2\pi}{L\Delta}(k_1\mathbf{m}_{i_1} + k_2\mathbf{m}_{i_2})\right) \tag{5.18}$$

5.4. Conclusion

The roundoff errors generated by orthogonal matrix multiplications were analyzed. The variance for each error can be expressed in terms of the characteristic function of the input, so can the cross-correlation between any pair of errors. For the 2×2 case, by assuming the input distribution to have a jointly Gaussian envelope, the cross-correlation between ϵ_1 and ϵ_2 is computed for various different input

dynamic ranges. It is found that for most angles of rotation the cross-correlation between the two errors decreases rapidly as the dynamic range of the input goes up. However, for a few particular values of θ the cross-correlation between the errors remains high even for large dynamic range.

A special case arises when the angle of rotation is very small. The cross-correlation between the errors becomes comparable to the cross-correlation of the input signals.

Appendix A.

We now describe the design procedure for the Mth-band FIR filter used in Sec. 3.3. Let $P(z)$ be the Mth-band filter with a desired response as in (3.27). We will assume that the allowable maximum passband and stopband errors are the same. If the pass and stopband requirements are different, it can be accomodated by making a minor modification to the procedure. The procedure takes as inputs: the filter order $N-1$ and a passband edge $\omega_p < \frac{2\pi}{M}$. The stopband edge ω_s is fixed to be $\frac{4\pi}{M} - \omega_p$. The output will be the impulse response coefficients of $P(z)$. The desired pass and stopband errors is achieved by adjusting the filter order through trial and error. (As a lower bound, the order estimation formula for linear phase equiripple filters [PAR72] may be used.) We shall write $P(z)$ as $\sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} p(n)z^{-n}$ where $N-1$ is even and $\frac{N+1}{2}$ is restricted to be a multiple of M. Let $\frac{N+1}{2}$ be denoted as J. Assuming that the impulse response is symmetric, we get $p(-n) = p(n)$. Furthermore, due to the Mth-band property $p(nM) = 0$ for $n \neq 0$ and $p(0) = \frac{2}{M}$. On the unit circle, the respose of $P(z)$ can be expressed as

$$P(e^{j\omega}) = p(0) + 2\sum_{n=1}^{J-1} p(n)\cos(n\omega) \qquad (A.1)$$

The function $P(e^{j\omega})$ can be viewed as a polynomial of $\cos(\omega)$ of order $J-1$. Thus, there will be at most $J-2$ extremal frequency points in the range $0 < \omega < \pi$. An extremal frequency point is defined to be a point on the frequency axis where $\frac{\delta P(e^{j\omega})}{\delta\omega} = 0$.

The method used here follows closely the design algorithm of [PAR72]. As in [PAR72], the extremal frequencies of $P(e^{j\omega})$ are first assumed to be known by initially guessing their locations. Since there are only $\frac{J}{M}(M-1)$ coefficients of the impulse response undetermined and the band edges ω_p and ω_2 are already

predetermined, we only have control over $\frac{J}{M}(M-1)-1$ of the extremal frequencies. Let the set of frequency points be: $\{\omega_0, \omega_1, \ldots, \omega_{K-1}\}$ with $K = \frac{J}{M}(M-1)-1$. The first point, ω_0, will always be set to zero. The rest are ordered so that $\omega_k > \omega_{k-1}$. Let us assume that the first $\frac{2J}{M}$ extremal frequencies (including ω_0) lies within the passband, therefore $\omega_{\frac{2J}{M}-1} < \omega_p$, while $\omega_{\frac{2J}{M}} > \omega_s$.

For $k \leq 2\frac{J}{M} - 1$, we impose the following alternation conditions

$$P(e^{j\omega_k}) = 1 + (-1)^{k+1}\delta \qquad (A.2)$$

and at the passband edge $P(e^{j\omega_p}) = 1 - \delta$. Similarly, in the stopband the function $P(e^{j\omega})$ is constrained to satisfy the conditions: $P(e^{j\omega_s}) = \delta$ and

$$P(e^{j\omega_k}) = (-1)^{k+1}\delta \quad \text{for} \quad k \geq 2\frac{J}{M}. \qquad (A.3)$$

The conditions (A.2) and (A.3) together with the conditions at ω_p and ω_s give us $\frac{J}{M}(M-1)+1$ equations in all. Since there are the same number of unknowns (the $\frac{J}{M}(M-1)$ impulse coefficients plus the error δ), we can solve this linear system of equations.

With the unknown $p(n)$'s determined, the actual extrema of $P(e^{j\omega})$ can be located by computing and searching through a dense grid of, say, 1000 points. The first $2\frac{J}{M}$ extrema in the passband and the first $K - 2\frac{J}{M}$ extrema in the stopband will be kept. The values of ω_k are set to the new extremal frequencies, and by substituting these new values into (A.2) and (A.3) we obtain a new system of linear equations. By solving the system of equations, a new set of $p(n)$ can be obtained and the process repeats until there are no more changes to the set $\{\omega_0, \omega_1, \ldots, \omega_{K-1}\}$ from one iteration to the next.

Fig. 3.7 shows us the result $|P(e^{j\omega})|$ at the termination of the above algorithm with $M = 5$, $N - 1 = 120$, $\omega_p = 0.38\pi$ and $\omega_s = 0.42\pi$.

Appendix B.

The matrix $\mathbf{C}$ as defined in (3.33) has at most $ML - L^2$ non-trivial entries, i.e., entries whose values are other than 0 or 1. Consider first the product $\mathbf{BW}^\dagger$. Let us denote the entries of $\mathbf{B}$ as $b_{i,k}$, and the entries of $\mathbf{U}$ as $u_{i,k}$. Since the pth column of $\mathbf{B}$ is zero whenever $p \notin \mathcal{L}$, we can delete those columns from $\mathbf{B}$ and as a result the corresponding row of $\mathbf{W}^\dagger$ are also deleted. The matrix product $\mathbf{BW}^\dagger$ becomes

$$\mathbf{BW}^\dagger = \begin{pmatrix} b_{0,l_0} & b_{0,l_1} & \cdots & b_{0,l_{L-1}} \\ b_{1,l_0} & b_{1,l_1} & \cdots & b_{1,l_{L-1}} \\ \vdots & \vdots & \ddots & \vdots \\ b_{L-1,l_0} & b_{L-1,l_1} & \cdots & b_{L-1,l_{L-1}} \end{pmatrix} \begin{pmatrix} 1 & W^{-l_0} & \cdots & W^{-l_0(M-1)} \\ 1 & W^{-l_1} & \cdots & W^{-l_1(M-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & W^{-l_{L-1}} & \cdots & W^{-l_{L-1}(M-1)} \end{pmatrix}. \tag{A.4}$$

Given that the l_kth column of $\mathbf{B}$ was defined to be $\mathbf{\Lambda}(W^{-l_k})\mathbf{u}_k$, we can write $b_{i,l_k} = u_{i,k}W^{il_k}$. Writing out the matrix multiplication in (A.4) explicitly for each entry, we get

$$[\mathbf{BW}^\dagger]_{i,m} = \sum_{k=0}^{L-1} b_{i,l_k} W^{-l_k m} = \sum_{k=0}^{L-1} u_{i,k} W^{l_k(i-m)}. \tag{A.5}$$

Remember that $u_{i,k}$ are the entries of $\mathbf{U}$ and the matrix $\mathbf{U}^{-1}$ is defined in (3.9). In (A.5), when $i - m = 0$ the above summation is reduced to

$$[\mathbf{BW}^\dagger]_{i,i} = \sum_{k=0}^{L-1} u_{i,k} = \begin{cases} 1 & \text{for } i = 0; \\ 0 & \text{for } i = 1, \ldots, L-1. \end{cases} \tag{A.6}$$

Similarly, for $((i - m)) = 1$

$$[\mathbf{BW}^\dagger]_{i,((i-1))} = \sum_{k=0}^{L-1} u_{i,k} W^{l_k} = \begin{cases} 1 & \text{for } i = 1; \\ 0 & \text{for } i = 0, 2, \ldots, L-1. \end{cases} \tag{A.7}$$

We can write down similar expressions for $((i - m)) = 2, 3, \ldots, L - 1$. As a result,

the matrix $\mathbf{BW}^\dagger$ has the form

$$\mathbf{BW}^\dagger = \begin{pmatrix} 1 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} & 0 & 0 & 0 \\ 1 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} & \mathrm{x} & 0 & 0 \\ 1 & 0 & 0 & \mathrm{x} & \mathrm{x} & \dots & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 \\ 1 & 0 & 0 & 0 & \mathrm{x} & \dots & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}. \qquad (A.8)$$

Here we display $\mathbf{BW}^\dagger$ for the special case of $L = 4$, but the general pattern is readily discernable. In general, there will be L one's and $L(L-1)$ zero's in the matrix. Substituting (A.8) back into (3.33), one can verify that there are only $ML - L^2$ non-trivial entries in $\mathbf{C}$.

Appendix C

Let us assume that each of the sub-sequences $x(nM-k)$ is corrupted by uncorrelated white noise of equal variance σ_i^2. For the purpose of noise analysis, the reconstruction filter bank (shown in Fig. 3.10) can be re-written in terms of its polyphase representation

$$\begin{pmatrix} 0 & \dots & 0 & 1 & r_{0,L}G_1(z^M) & r_{0,L+1}G_2(z^M) & \dots & r_{0,M-1}G_{M-L}(z^M) \\ 0 & \dots & 1 & 0 & r_{1,L}G_2(z^M) & r_{1,L+1}G_3(z^M) & \dots & r_{1,M-1}G_{M-L+1}(z^M) \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & \dots & 0 & 0 & r_{L-1,L}G_L(z^M) & r_{L-1,L+1}G_{L+1}(z^M) & \dots & r_{L-1,M-1}G_{M-1}(z^M) \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1} \\ \vdots \\ z^{-M+1} \end{pmatrix}. \qquad (A.9)$$

The scalar factors, $r_{m,k}$, are related to the entries of the $\mathbf{C}$ matrix as

$$r_{m,k} = c_{m,k+m-L+1}. \qquad (A.10)$$

From the polyphase matrix above, we see that $\hat{x}(nM+k) = x(nM+k-L-1)$ for $k = 0, \dots, L-1$. Thus, the retained samples are produced directly at the output without any computation, and the noise variance for these output samples is σ_0^2.

For the case of $\hat{x}(nM+k)$ where $L \le k \le M-1$, the noise analysis is as follows. Assuming that $G_i(z)$'s are ideal filters, each of the filters has a noise gain of unity.

Therefore the noise gain from the mth input of the polyphase matrix to the kth output is simply $|r_{m,k}|^2$. If one assumes that the L noise sources at the input are uncorrelated, then using (A.5) and (A.10) the noise variance at the kth output is given by

$$\sigma_0^2 \sum_{m=0}^{L-1} |r_{m,k}|^2 = \sigma_0^2 \sum_{m=0}^{L-1} \left| [\mathbf{B}\mathbf{W}^\dagger]_{m,k+m-L+1} \right|^2 = \sigma_0^2 \sum_{m=0}^{L-1} \left| \sum_{n=0}^{L-1} u_{m,n} W^{l_n(L-k-1)} \right|^2. \quad (A.11)$$

References

[AN88] R. Ansari and S. H. Lee, "Two-dimensional multirate processing on non-rectangular grids: theory and filtering procedures," submitted, 1988.

[BA80] C. W. Barnes and S. Shinnaka, "Block-shift invariance and block implementation of discrete-time filters," IEEE Trans. Circuits Syst., vol. CAS-27, no. 8, pp. 667-672, August, 1980.

[BA85] C. W. Barnes, B. N. Tran and S. H. Leung, "On the statistics of fixed-point roundoff error," IEEE Trans. Acoust., Speech, Signal Processing, pp. 595-606, June 1985.

[BE76] M. Bellanger, G. Bonnerot and M. Coudreuse, "Digital filtering by polyphase network: application to sample rate alteration and filter banks," IEEE Trans. Acoust., Speech, Signal Processing, vol. 24, pp. 109-114, Apr., 1976.

[BL53]H. S. Black, *Modulation theory*, New York, van Nostrand, 1953.

[BR81]J. L. Brown, "Multi-channel sampling of lowpass signals," IEEE Trans. Circuits Syst., vol. 28, pp. 101-106, Feb. 1981.

[BU72] C. S. Burrus, "Block realization of digital filters," IEEE Trans. Audio Electroacout., vol. AU-20, pp. 230-235, Oct. 1972.

[CO87]R. V. Cox, D. E. Bock, J. D. Johnston, and J. H. Snyder, "The analog voice privacy system," AT& T Tech. J., vol. 66, pp. 119-131, Jan-Feb 1987.

[CR83]R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[DEL85] Ph. Delsarte, A. J. E. M. Janssen and L. B. Vries, "Discrete prolate spheroidal wave functions and interpolation," SIAM J. Appl. Math., vol. 45, no. 4, pp. 641-650, Aug. 1985.

[DEW80] P. DeWilde and E. Deprettere, "Orthogonal cascade realization of real multiport digital filters," Int. J. Circuit Theory Appl., vol. 8, pp. 245-277, 1980.

[DO88]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. 36, Oct. 1988.

[DU85] E. Dubois, "The sampling and reconstruction of time-varying imagery with application in video systems," Proc. of IEEE, vol. 73, no. 4, pp. 502-522, April 1985.

[GA75]W. A. Gardner and L. E. Franks, "Characterization of cyclostationary random signal processes," IEEE Trans. Inform. Theory, vol. IT-21, pp. 4-14, Jan. 1975.

[GE74]R. W. Gerchberg, "Super-resolution through error energy reduction," Optica Acta, vol. 21, no. 9, pp. 709-720, 1974.

[GR73] A. H. Gray, Jr. and J. D. Markel, "Digital lattice and ladder filter synthesis," IEEE Trans. AU, pp. 491-500, Dec. 1973.

[HE83] D. Henrot and C. T. Mullis, "A modular and orthogonal digital filter structure for parallel processing," in IEEE Int. on Acoust., Speech, Signal Processing, pp. 623-626, Apr. 1983.

[HO81]S. J. Howard, "Method for continuing Fourier spectra given by the fast

Fourier transform," J. Opt. Soc. Am., vol. 71, no. 1, pp. 95-98, 1981.

[HS87] Chia-Chuan Hsiao, "Polyphase filter matrix for rational sampling rate conversions," Proc. IEEE Int. Conf. on ASSP, pp. 2173-2176, Dallas, April 1987.

[JAC70] L. B. Jackson, "On the interaction of roundoff noise and dynamic range in digital filters, Bell Systems Technical Journal, vol. 49, pp. 159-184, 1970.

[JAI81] A. K. Jain and S. Ranganath, "Extrapolation algorithms for discrete signals with application in spectral estimation," IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 830-845, Aug. 1981.

[JAY84]N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

[JE77]A. J. Jerri, "The Shannon sampling theorem - its various extensions and applications: a tutorial review," Proc. of the IEEE, pp. 1565-1596, Nov. 1977.

[LI88a] V. C. Liu and P. P. Vaidyanathan, "Alias cancellation and distortion elimination in multi-dimensional QMF banks," Proc. of IEEE International Symposium on Circuits and Systems, Espoo, Finland, May 1988.

[LI88b] V. C. Liu and P. P. Vaidyanathan, "Roundoff noise generated by orthogonal building blocks in signal processing structures," Proc. of IEEE International Symposium on Circuits and Systems, Espoo, Finland, May 1988.

[LI88c] V. C. Liu and P. P. Vaidyanathan, " Compression of two-dimensional bandlimited signals using subsampling theorems," J. of Institution of Electronics and Telecommunication Engineers, vol. 34, no. 5, pp. 416-422, September-October, 1988.

[LI89a] V. C. Liu and P. P. Vaidyanathan, "Finite length extrapolation of bandlimited signals," Proceedings of the IEEE International Symposium on Circuits and Systems, Portland, Oregon, May 1989.

[LI89b] V. C. Liu and P. P. Vaidyanathan, "Efficient reconstruction of bandlimited signals from nonuniformly decimated versions by use of polyphase filter banks," (submitted for journal publication), Nov. 1988.

[MA86]S. L. Marple, *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[ME83] R. M. Mersereau and T. C. Speake, "The processing of periodically sampled multidimensional signals," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-31, pp. 188-194, Feb. 1983.

[MIN82]F. Mintzer, "On half-band, third-band and Nth-band FIR filters and their design," IEEE Trans. Acoust., Speech, Signal Processing, vol. 30, pp. 734-738, October 1982.

[MIT78] S. K. Mitra and R. Gnanasekaran, "Block implementation of recursive digital filters - new structures and properties," IEEE Trans. Circuits Syst., vol. CAS-25, no. 4, pp. 200-207, April, 1978.

[MU76] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters, IEEE Trans. Circuits Syst., pp. 551-562, Sept. 1976.

[NG88]T. Nguyen, T. Saramaki and P. P. Vaidyanathan, " Eigenfilters for the design of special transfer functions with applications in multirate signal processing," Proc. IEEE Int. Conf. ASSP., pp. 1467-1470, New York, April 1988.

[OP75] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[PAP65] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., New York, 1965.

[PAP75] A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," IEEE Trans. Circuits Syst., vol. 22, pp. 735-742, Sept. 1975.

[PAP77a] A. Papoulis, *Signal Analysis*, McGraw-Hill, Inc., New York, 1977.

[PAP77b]A. Papoulis, "Generalized sampling expansions," IEEE Trans. Circuits Syst., pp. 652-654, Nov. 1977.

[PAR72]T. W. Parks and J. H. McClellan, "A program for the design of linear phase finite impulse response filters," IEEE Trans. Audio Electroacoust., vol. AU-20, no. 3, pp. 195-199, Aug. 1972.

[RA84] S. K. Rao and T. Kailath, "Orthogonal digital filters for VLSI implementation," IEEE Trans. Circuits Syst., vol. CAS-31, pp. 933-945, Nov. 1984.

[SAM88]H. Samueli, "On the design of optimal equiripple FIR digital filters for data transmission applications," IEEE Trans. Circuits Syst., vol. 35, pp. 1542-1546, Dec. 1988.

[SAN63]I. W. Sandberg, "On the properties of some systems that distort signals – I," Bell System Technical Journal, Sept. 1963.

[SC81]H. Scheuermann and H. Gockler, "A comprehensive survey of digital transmultiplexing methods," Proceedings of the IEEE, vol. 69, no. 11, pp. 1419-1450, Nov. 1981.

[SH49]C. E. Shannon, "Communications in the presence of noise," Proc. of the IRE, vol. 37, pp. 10-21, Jan. 1949.

[SM87a] M. J. T. Smith and T. P. Barnwell, III, "A new filter bank theory for time-

frequency representation," IEEE Trans Acoust., Speech, Signal Processing, vol. 35, pp. 314-327, March 1987.

[SM87b] M. J. T. Smith and S. L. Eddins, "Subband coding of images with octave band tree structures," Proc. of the IEEE Int. Conf. ASSP, pp. 1382-1385, Dallas, April 1987.

[SU84] B. J. Sullivan and B. Liu, "On the use of singular value decomposition and decimation in discrete-time band-limited signal extrapolation," IEEE Trans. Acoust., Speech, Signal Processing, vol.32, pp. 1201-1212, Dec. 1984.

[VA85] P. P. Vaidyanathan, "A unified approach to orthogonal digital filters and wave digital filters, based on LBR two-pair extraction," IEEE Trans. Circuits Syst., vol. CAS-32, pp. 673-686, July 1985.

[VA86] P. P. Vaidyanathan, "Passive cascaded-lattice structures for low-sensitivity FIR filter design, with Applications to Filter Banks," IEEE Trans. Circuits Syst., pp. 1045-1064, Nov. 1986.

[VA87a] P. P. Vaidyanathan, "Theory and design of M-Channel maximally decimated quadrature mirror filters with arbitrary M, having the perfect-reconstruction property," IEEE Trans. Acoust., Speech, Signal Processing, pp. 476-492, April, 1987.

[VA87b] P. P. Vaidyanathan, "Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques," IEEE ASSP magazine, vol. 4, pp. 4-20, July 1987.

[VA87c] P. P. Vaidyanathan, "Perfect reconstruction QMF banks for two-dimensional applications," IEEE Trans. Circuits Syst., pp. 976-978, August 1987.

[VA88a] P. P. Vaidyanathan and P.-Q. Hoang, "Lattice structures for optimal de-

sign and robust implementation of two-channel perfect-reconstruction QMF banks" IEEE Trans. Acoust., Speech, Signal Processing, pp. 81-94, Jan. 1988.

[VA88b] 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. Acoust., Speech, Signal Processing, vol. ASSP-36, no. 3, pp. 381-391, March, 1988.

[VA88c] P. P. Vaidyanathan and Z. Doganata, "The role of lossless systems in modern digital signal processing: a tutorial," Caltech, July 1988.

[VA88d] P. P. Vaidyanathan and Vincent C. Liu, "Classical sampling theorems in the context of polyphase digital filter banks," IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, no. 9, pp. 1480-1495, Sept. 1988.

[VA89]P. P. Vaidyanathan, T. Q. Nguyen, Z. Doganata and T. Saramaki, "Improved technique for design of perfect reconstruction FIR QMF banks with lossless polyphase matrices," IEEE Trans. on Acoust., Speech, Signal Processing, vol. 37, July 1989.

[VE84] M. Vetterli, "Multi-dimensional sub-band coding: some theory and algorithms," Signal Processing, vol. 6, pp 97-112, April 1984.

[VI88] E. Viscito and J. Allebach, "Design of perfect reconstruction multi-dimensional filter banks using cascaded Smith form matrices," Proc. of the IEEE Int. Symp. on Circuits and Systems, Espoo, Finland, pp. 831-834, June 1988.

[WA86] G. Wackersreuther, "On two-dimensional polyphase filter banks," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, no. 1, pp. 192-199, Feb. 1986.

[WE88]P. H. Westerink, D. E. Boekee, J. Biemond and J. W. Woods, "Sub-band coding of images using vector quantization," IEEE Trans. Comm., vol. 36, no. 6, pp. 713-719, June 1988.

[WI78]R. G. Wiley, "Recovery of bandlimited signals from unequally spaced samples," IEEE Trans. Comm., vol. COM-26, no. 1, pp. 135-137, Jan. 1978.

[WO86] J. W. Woods and S. D. O'Neil, "Subband coding of Images", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, no. 5, pp. 1278-1288, Oct. 1986.

[YE56]J. L. Yen, "On nonuniform sampling of bandlimited signals," IRE Trans. Circuits Theory, vol. CT-3, pp. 251-257, Dec. 1956.