STATISTICAL OPTIMIZATION OF MULTIRATE SYSTEMS AND ORTHONORMAL FILTER BANKS

Thesis by

Jamal Tuqan

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1998

(Submitted July 30^{th}, 1997)

Acknowledgements

First of all, I would like to thank Professor P. P. Vaidyanathan. It is my greatest fortune to have him as my advisor. It is usually a very rewarding experience to work under the guidance of the best in the field and with P. P., this is no exception. P. P. has always made himself available to help and guide me throughout my PhD studies at Caltech, no matter how busy he is. I admire his superb technical abilities and his depth in analyzing any research problem from which I have personally learned a lot. His exceptional enthusiasm and continuous motivation to pursue his research goals makes him a top notch researcher, a role model I found very difficult to follow. P. P. is also an excellent teacher. His particular approach in teaching Digital Signal Processing (DSP) has greatly enhanced my understanding of the material and gave me a solid preparation to pursue my own research. It is somehow ironic that my biggest struggle during my PhD studies has always been the quest to meet the high standards he has set. My sincere hope is to be up to his expectations in the future.

The generous financial support provided in parts by the Office of Naval research (grant N00014-93-1-0231), Tektronix, Inc., and Rockwell International is greatly appreciated.

I would like also to thank Professors Robert J. McEliece, Yaser S. Abu-Mostafa, Joel N. Franklin, Michelle Effros, Abeer Alwan (UCLA) and Dr. Marvin K. Simon for serving on my candidacy and defense committees. I appreciate their comments and suggestions that enhanced the quality of my thesis. Professor Effros, in particular, was a careful reader and provided some very useful feedback. I am particularly indebted to Professors Robert McEliece and Yaser Abu-Mostafa for their time (they served on both committees), for the classes they have taught me and for their continuous support and encouragements during the course of my studies.

My working and life experience at Caltech has been an unforgettable one due, in great part, to my colleagues: Dr. Igor Djokovic, Dr. See-May Phoong, Dr. Yuan-Pei Lin, Ahmet Kirac, Murat Mese and Sony Akkarakaran. Special thanks goes to my friend, colleague and office mate, Ahmet Kirac, whose helpful criticism and penetrating questions has always improved the quality of my work. Outside the lab, I have also cherished the company of some good friends, namely Mohamed-Slim Alouini and Professor Amir Atiya (Cairo University, Egypt). Finally, I would like to take this opportunity to thank Robert Freeman, the system manager, for all the help he has provided as well as Lilian Porter and Lavonne Martin for their administrative assistance.

Last and certainly not least, I would like to thank my "two" families: my parents for their love, encouragements and all the sacrifices they have gone through to provide me with the best possible education and my wife Fatima and our son Sami, for their patience, love and moral support. There is

no way I can thank them enough. My parents taught me the importance of hard work and perseverance and always encouraged and supported me to pursue my ultimate dreams. Fatima has invariably been there for me, showing infinite patience and understanding during my absorption in this work. I thank her for putting up with me for all these years and for all the love and moral support that only she can offer. Sami has brought an unprecedented joy and happiness in my life. I thank him for being here. Both are my best blessings.

Abstract

The design of multirate systems and/or filter banks adapted to the input signal statistics is a generic problem that arises naturally in variety of communications and signal processing applications. The two main applications we have in mind are the statistical optimization of subband coders for signal compression and the multirate modeling of WSS random processes. These two applications lead naturally to the important concepts of energy compaction filters and principal component filter banks. In this thesis, we study three problems that are directly related to the above mentioned applications. The first problem is motivated by the observation that in the presence of subband quantizers, it is a loss of generality to assume that the synthesis section in a filter bank is the inverse of the analysis section. We therefore consider the statistical optimization of linear time invariant (LTI) pre- and postfilters surrounding a quantization system. Unlike in previous work, the postfilter is not restricted to be the inverse of the prefilter. Closed form expressions for the optimum filters as well as the resulting minimum mean square error (m.m.s.e.) are derived. The importance of the m.m.s.e. expression is that it clearly quantifies the additional gain obtained by relaxing the perfect reconstruction assumption. In the second problem, we study the quantization of a certain class of *non bandlimited* signals, modeled as the output of $L < M$ interpolation filters where M is the interpolation factor. Using the fact that these signals are *oversampled*, we show how to decrease substantially the quantization noise variance using appropriate multirate reconstruction schemes. We also optimize a variety of noise shapers, indicating the corresponding additional reduction in the average mean square error for each case. The results of this chapter extend, using multirate signal processing theory, some well known techniques of efficient A/D converters (e.g. sigma-delta modulators) that usually apply only to bandlimited signals. In the last problem, a novel procedure to design *globally optimal* FIR energy compaction filters is presented. Energy compaction filters are important due to their close connection to *orthonormal* filter banks adapted to the input signal statistics. In fact, for the two channel case, the problems are equivalent. A special case of compaction filters arise also in applications such as echo cancelation, time varying systems identification, standard subband filter design and optimal transmitter and receiver design in digital communications. The new proposed approach guarantees *theoretical* optimality which previous methods could not achieve. Furthermore, the new algorithm is:

i) extremely general in the sense that it can be tailored to cover any of the above applications.

ii) numerically robust.

iii) can be solved efficiently using interior point methods.

The design of a special class of two channel IIR compaction filters is also considered. We show that,

in general, this class of optimum IIR compaction filters, parameterized by a single coefficient, are competitive with very high order optimum FIR filters.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The main theme of this thesis is the optimization of multirate systems and orthonormal filter banks according to the second order statistics of the input signal. In general, the statistical optimization of multirate systems and/or filter banks is an important problem that arises naturally in a wide variety of communication and signal processing applications. The two fundamental applications that motivate our work are the optimization of subband coders for signal compression and the multirate modeling of WSS random processes. To provide the reader with a general overview of the thesis, we start this chapter with a short treatment of each of the above mentioned applications. In particular, we discuss in a qualitative manner the main underlying ideas pertaining to each topic and review some useful background information as well as the major developments in each area. Important connections between the two subjects are then drawn by introducing the concepts of principal component filter banks and energy compaction filters. Finally, we end the chapter with a brief description of each of the three problems we have addressed and the corresponding results.

1.1 The optimal design of Subband Coders

Subband coders were originally introduced for speech coding [15, 16] and since then, have been extensively used in audio, image and video compression (see [73, 86] and the references therein). They play a fundamental role in standard coders such as JPEG (Joint Picture Expert Group) and MPEG (Motion Picture Expert Group) as well as in the state of the art compression schemes that are based on perceptual measures.

1.1.1 Perfect reconstruction filter banks

Central to the implementation of a subband coder is an M-channel maximally decimated uniform filter bank shown in Fig. 1.1.

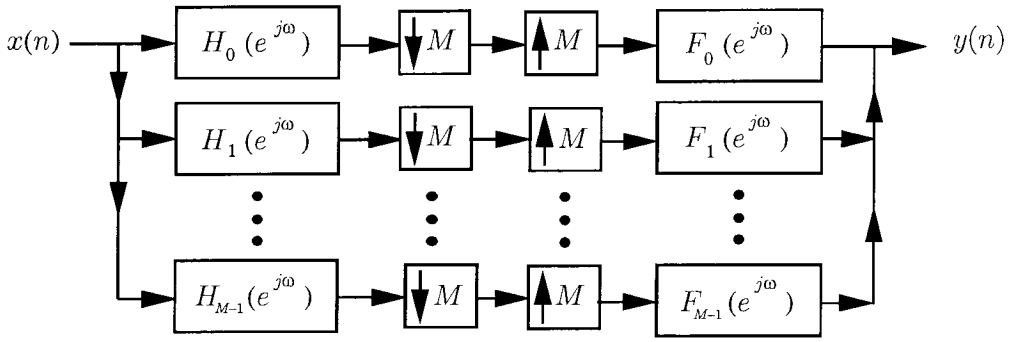

Fig. 1.1: An M-channel maximally decimated uniform filter bank

The input signal $x(n)$ is split into M-subbands in the frequency domain by a bank of linear time invariant (LTI) filters, $H_k(e^{j\omega})$, $k = 0, 1, \ldots, M-1$, termed the analysis filters. Since the filters have typically good frequency responses, their outputs can be considered to be bandlimited, and we can therefore decimate these signals. The decimation operation is denoted by the boxes $\downarrow M$ and consists of retaining the samples of each subband filter output that occur at time $n = iM$ where i is an integer. The signals in each subband are then processed according to the application at hand. At the receiving end, the sampling rates in each of the subbands are increased to their original value by the expanders, denoted by $\uparrow M$, which insert $M-1$ zero valued samples between each two adjacent sample of their input. The signals are then passed through the synthesis filters, $F_k(e^{j\omega})$, $k = 0, 1, \ldots, M-1$, whose magnitude responses typically resemble those of the analysis filters. The output of the synthesis filters are combined to give the reconstructed signal $y(n)$.

The theory and design perfect reconstruction filter banks, i.e., filter banks for which the output $y(n) = c\, x(n - n_0)$, a scaled and delayed version of the input, has been the main focus of research for more than a decade. Since the analysis and synthesis filters are usually not ideal, the filter bank of Fig. 1.1 produces three types of distortion. First, due to the presence of the decimators, aliasing occurs. In 1976, Croisier et al. [16] showed that it is possible to completely eliminate aliasing distortion for the two-channel case by a clever choice of the synthesis filters. It can then be shown that, an alias free filter bank reduces to an LTI system with a certain phase and magnitude response. The frequency response of the alias free filter bank produces therefore phase and magnitude distortion. In the mid-eighties, this remaining form of error was finally removed by Smith and Barnwell [60] and Mintzer [50] for the two-channel case using finite order filters. These papers therefore demonstrated that perfect reconstruction can be achieved at finite cost. Although the extension of these results to the M-channel case was an obvious generalization, the problem was an order of magnitude more difficult. A solution was finally found by Vetterli [85] and independently by Vaidyanathan [70] using the polyphase approach. Indeed, the perfect reconstruction property is perhaps best understood by using the polyphase matrix representation of a maximally decimated uniform filter bank, shown in

Fig. 1.2.

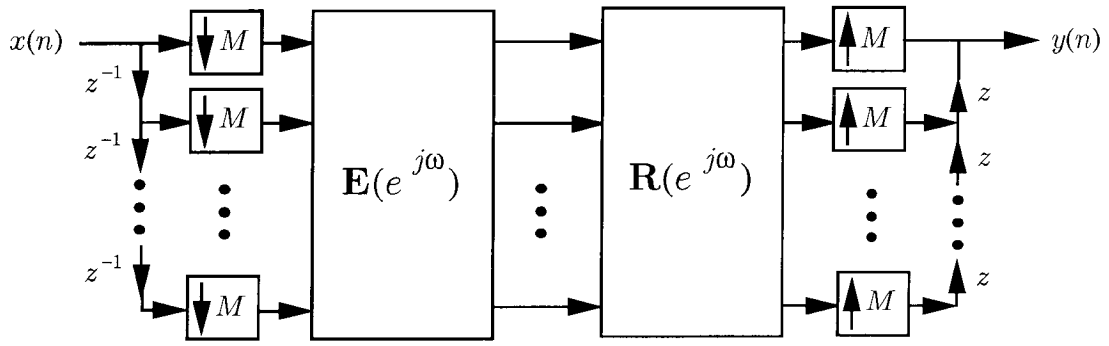

Fig. 1.2: The polyphase representation of an M-channel maximally decimated uniform filter bank

In this figure, $\mathbf{E}(z)$ and $\mathbf{R}(z)$ are the polyphase matrices corresponding respectively to the analysis and synthesis filters. A maximally decimated uniform filter bank can be always redrawn in the polyphase form and exhibits the perfect reconstruction property, if and only if, $\mathbf{R}(z)\mathbf{E}^{-1}(z) = cz^{-n_0}\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. A perfect reconstruction filter bank (PRFB) is also known as a biorthogonal filter bank. An extremely important subclass of perfect reconstruction filter banks is the class of *paraunitary* (PU) or *orthonormal* filter banks. In this case, the analysis polyphase matrix satisfies the lossless property, mathematically expressed as $\mathbf{E}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega}) = \mathbf{I}\ \forall\ \omega$, where the superscript $\dagger$ denotes the conjugate transpose operation. By choosing the synthesis polyphase matrix $\mathbf{R}(e^{j\omega})$ to be equal to $\mathbf{E}^{\dagger}(e^{j\omega})$, perfect reconstruction is guaranteed. The paraunitary property of filter banks offers many advantages: First, the analysis of orthonormal filter banks is usually much more simple than in the biorthogonal case. Second, the design process is greatly simplified due to the complete parameterization of orthonormal filter banks [73]. The synthesis filters $F_k(z)$ are FIR if the analysis filters $H_k(z)$ are chosen to be FIR. In fact, the analysis and synthesis filters will have the same length and $F_k(z)$ can be found by inspection from $H_k(z)$. Finally, the two channel paraunitary filter bank is a basic building block in the generation of orthonormal wavelet basis.

1.1.2 Filter banks adapted to the input signal statistics

Consider now the subband coding scheme shown in Fig. 1.3. The subband coder (SBC) is basically a filter bank structure where the subband quantizers, labeled by $\mathcal{Q}$, are now present. The subband quantizers represent uniform scalar quantizers which are modeled by additive white noise sources $q_k(n)$. The input $x(n)$ is assumed to be a wide-sense stationary signal with a known power spectrum $S_{xx}(e^{j\omega})$. In the presence of quantizers, perfect reconstruction is not possible because quantization is a lossy process. The output $\hat{x}(n)$ is equal to the original signal $x(n)$ plus a filtered version of the quantization noise, labeled by $e(n)$. Given a fixed budget of b bits for the subband quantizers, the subband coding problem is to jointly optimize the analysis and synthesis filters and to choose

a subband bit allocation strategy such that the mean square value of the reconstruction error is minimized.

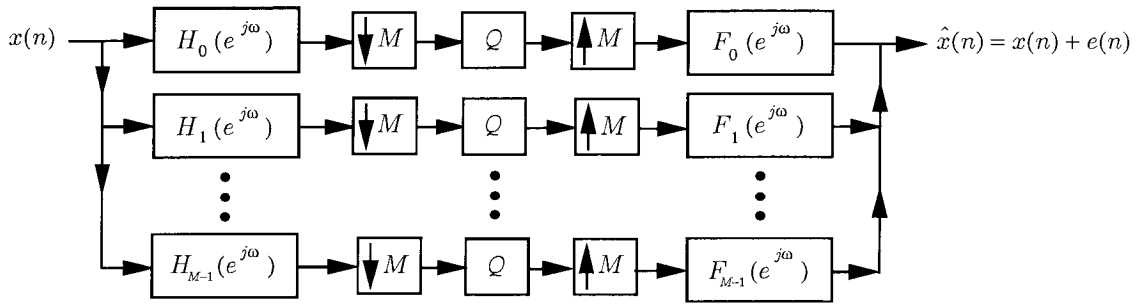

Fig. 1.3: An M-channel uniform maximally decimated subband coder

If the subband quantizers are modeled using the so-called high bit rate assumption (we will discuss this issue in some detail in later chapters of the thesis), it can be shown that the optimum design of subband coders (in the sense described above) involves the optimization of the subband filters according to the second order statistics of the input signal. Additional optimization constraints usually exist and depend on the class of filter banks we are optimizing over (e.g. orthonormal versus biorthogonal). The design of filter banks adapted to the input signal statistics arise therefore in a natural way in the design of optimal subband coders.

The special case where $\mathbf{E}(e^{j\omega})$ is memoryless (i.e. $\mathbf{E}(e^{j\omega}) = \mathbf{T}$ for all ω) is the transform coder problem, and was addressed by Huang and Schultheiss [34], and in greater detail by Segall [59]. It is well known that, the optimal solution, in this case, is the Karhunen Loeve Transform (KLT). Recently, several authors have considered the extension of the KLT result to the case where $\mathbf{E}(e^{j\omega})$ has memory, i.e., $\mathbf{E}(z) = \sum_n \mathbf{E}_n(z)z^{-n}$. Note that the index n can run from $-\infty$ to ∞ because the analysis filters $H_k(e^{j\omega})$ can have arbitrarily large orders. For the case of unconstrained filter orders, theoretical results on the statistical optimization of two channel *orthonormal* filter banks were given by Unser [68]. A general set of necessary and sufficient conditions for the optimality of an M-channel maximally decimated *orthonormal* uniform filter bank were derived in [76]. Some results regarding the optimization of a subband coder over the more general class of biorthogonal filter banks can be found in [2] and [80]. For the case of finite order filters, although theoretical results are not available, design methods have already been reported (see for example [69, 10, 17, 30]). In chapter 4, we introduce a novel procedure for the design of an FIR two channel orthonormal filter bank where, unlike in previous work, *theoretical* optimality is guaranteed.

We would like to end this section by pointing out that, in the presence of quantizers, the synthesis polyphase matrix should not be the the inverse of the analysis polyphase matrix. In other words, optimizing the subband coder over the class of perfect reconstruction filter banks when quantizers are present is a loss of generality [75]. A similar observation is given by Gosse and Duhamel [26]. In their

paper, the most general class of subband coders is called the minimum mean square error (MMSE) filter banks. Kovacevic also reaches the same conclusion for the case where the subband quantizers Q are modeled as Lloyd-Max quantizers [46]. This crucial observation is in fact the main motivation and starting point for the work described in chapter 2.

1.2 Multirate signal modeling

Signal modeling is an important problem that arises in a variety of signal processing applications. Typically, the knowledge of a signal model can be exploited to improve the performance of existing schemes. The main underlying assumption is that the model should approximate the original signal in a fairly accurate manner. In speech processing, for example, the speech waveform is modeled as an all pole LTI filter driven by a weighted sum of impulses if the speech segment is voiced or by white noise if the speech segment is unvoiced. This particular model is then exploited in *linear predictive coding* (LPC) to obtain a more efficient representation of the speech signal. In specific, instead of storing/transmitting the speech waveform (the data), we can simply store/transmit the all pole filter coefficients corresponding to each voiced and unvoiced segment of speech. The original waveform can then be reconstructed using the model described above. This is an example of *parametric* modeling. The tradeoff between the original signal representation and its model representation is usually one of accuracy versus efficiency.

In this thesis, we are interested in the multirate modeling of WSS random processes. The signals of interest are modeled as the output of a single interpolation filter as shown in Fig. 1.4 (the single band model) or more generally, as the sum of the outputs of $L < M$ interpolation filters as shown Fig. 1.5 (the multi-band model). The signal $y(n)$ in Fig. 1.4 is a zero mean WSS process and the signals $y_k(n)$, $k = 0, 1, \ldots, L-1$, in Fig. 1.5 are assumed to be zero mean jointly wide sense stationary random processes. In both cases, the model filter(s) $F_k(e^{j\omega})$ are assumed to be FIR. The reader should note that the output $x(n)$ is in general a zero mean cyclo-widesense stationary random process of period M $[(CWSS)_M]$ [58]. So, unlike in standard stochastic rational modeling (e.g. AR, MA and ARMA modeling), a WSS signal in this case is "approximated" by a $(CWSS)_M$ signal.

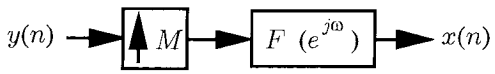

Fig. 1.4: The single band model

1.2.1 Application in quantization

The interest in the previous multirate models is partly motivated by the fact that the signal $x(n)$ in Fig. 1.4 and Fig. 1.5 can be recovered from its decimated version $x(Mn)$, *even though aliasing occurs.*

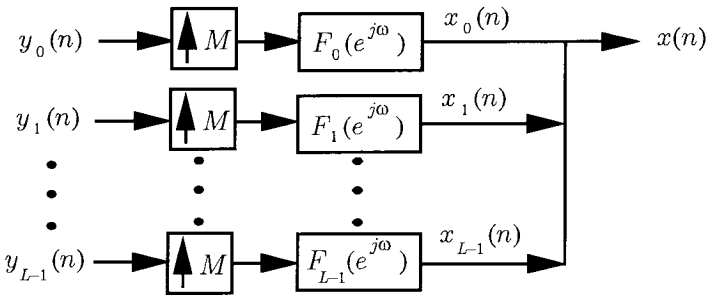

Fig. 1.5: The multiband model

As a quick example, assume that $x(n)$ is modeled as in Fig. 1.4 and consider $x(Mn)$, the M-fold decimated version of $x(n)$. If $F(e^{j\omega})$ is a Nyquist(M) filter (see definition in the last section of this chapter), then, $x(Mn)$ is equal to $y(n)$ and we have the relation $x(n) = \sum_k x(kM)f(n - kM)$. In other terms, $x(n)$ is completely defined by the samples $x(Mn)$ even though the filter $F(e^{j\omega})$ is not ideal. In this sense, the signal $x(n)$ can be considered as *an oversampled signal* although it is actually not bandlimited. In chapter 3 of the thesis, assuming that the signal can be accurately modeled by Fig. 1.5 (The single band case is a special case with $L = 1$), we exploit the oversampled nature of $x(n)$ and show how to reap advantages similar to those obtained by well known efficient A/D conversion schemes. Well known A/D conversion principles such as, oversampling PCM techniques and noise shaping, for example, apply only to bandlimited signals. The results of chapter 3 can be therefore interpreted as the extension of these techniques to a class of non-bandlimited signals using multirate signal processing theory.

1.2.2 Finding a signal model

What kind of signals can be realistically modeled as in Fig. 1.4 or more generally as in Fig. 1.5 ? To answer this, consider again the filter bank system of Fig. 1.1, where a WSS signal $x(n)$ is split into M subbands and reconstructed perfectly from its maximally decimated versions. Suppose now that the signal $x(n)$ has most of its energy concentrated in L subbands, which we number as the first L subbands. Then, the signal model of Fig. 1.5 is a good approximation of the original signal. Thus, given a signal $x(n)$ with energy concentrated mostly in certain subbands, the problem of finding the best signal model reduces to that of finding the filter bank that produces the L most dominant subbands. If the filter bank in Fig. 1.1 is orthonormal (paraunitary), the modeling issue reduces to the design of the so-called principal component filter banks for the multiband case and the design of energy compaction filters for the single band case. These important concepts are discussed next.

1.3 Principal component filter banks and energy compaction filters

Consider Fig. 1.6 where $(M-P)$ channels are dropped in the synthesis part of an M-channel *orthonormal* filter bank. An orthonormal filter bank that minimizes the average mean square reconstruction error for *any* P is called a principal component filter bank (PCFB). Using the orthonormality property [67, 64], it can be shown that, a principal component filter bank produces a decreasing arrangement of the the subband variances $\sigma_{x_1}^2 \geq \sigma_{x_2}^2 \ldots \geq \sigma_{x_P}^2$ such that, for any $1 \leq P < M$, $\sum_{k=1}^{P} \sigma_{x_k}^2$ is maximized. For $P = M$, $\sum_{k=1}^{M} \sigma_{x_k}^2 = M\sigma_x^2$ and is therefore fixed. The set of subband variances $\{\sigma_{x_k}^2\}$ generated by a principal component filter bank is said to "majorize" any other arbitrary set of subband variance $\{\sigma_{y_k}^2\}$. For the case of $P = 1$, the problem becomes one of designing a single analysis filter such that its output variance is maximized under the constraint that its magnitude squared response is Nyquist(M) (See definition at the end of this chapter). The resulting filter is termed an energy compaction filter.

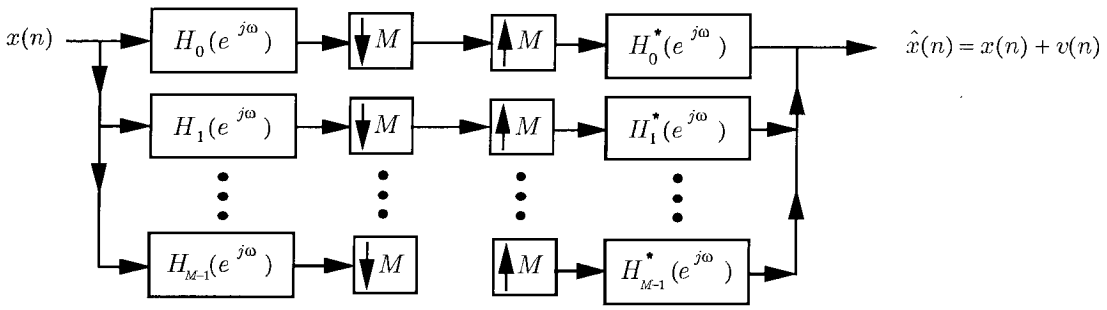

Fig. 1.6: An M-channel FIR principal component filter bank where $P = 2$

1.3.1 Relation to the subband coding problem

Interestingly enough, the design of a principal component filter bank (PCFB) is also closely related to the statistical optimization of an orthonormal filter bank. It can be actually shown (see chapter 4 for details) that, for the two channel case, and with the high bit rate quantizer assumptions, designing one of the subband filters for maximum energy compaction is equivalent to statistically optimizing an orthonormal filter bank. For the M-channel case, we can further show, using a well known majorization theorem (see for example [49, pages 11–12]), that a PCFB is also a statistically optimized subband coder under the orthonormality constraint. This was perhaps first observed by Unser [67] and an independent proof was given by Xuan and Bamberger in [90]. Therefore, by designing a principal component filter bank, we automatically optimize the orthonormal filter bank according to the input signal statistics. We would like to point out that the other direction is not true, i.e., a statistically optimized orthonormal filter bank is not necessarily a principal component filter bank.

There are two extreme examples (depending on the filter order N) worth mentioning: the case where $N < M$ and the case with no order constraint (ideal filters). The first case is the transform coder described previously. The optimum solution is the Karhunen Loeve Transform which is basically the eigen vector matrix corresponding to the positive definite autocorrelation matrix of the WSS input signal $x(n)$. The optimum subband variances are the eigen values and are known to majorize any other possible set of subband variances [33]. Note that the KLT is one example where the principal component solution is *equivalent* to the optimum orthonormal filter bank solution. The ideal filter case has been studied by Unser [68] for the two channel case and more extensively by Vaidyanathan [79] and Tsatsanis and Giannakis [65] for the M-channel case. Vaidyanathan's work is motivated by the subband coding problem whereas Tsatsanis and Giannakis 's work is motivated by the problem of finding the best basis, that is, the best set of orthonormal filters to represent a given signal such that the approximation error is minimized (in the mean square sense). It turns out that, for also this case, the principal component filter bank solution is the same as the optimum orthonormal filter bank solution. The magnitude response of the resulting optimal filters has an on-off behavior (either $\sqrt{M}$ or 0 values) depending on the amplitude of the input power spectral density in certain frequency regions.

The design of *globally optimum* FIR energy compaction filters of arbitrary order N is discussed in chapter 4. The FIR principal component filter bank case, however, remains at this moment in time unresolved. One major obstacle is that the *existence* of a principal component filter bank is currently only established for the above two extremes. More explicitly, given the space of all possible subband variances generated by an FIR orthonormal filter bank (order $N \geq M$ and $M > 2$), the *existence* of a specific set of variances $\{\sigma_{y_k}^2\}$ that will majorize all other possible subband variances is to be proven.

1.4 Thesis outline and overview

Outline. The thesis is organized into five chapters. The three problems described in the abstract are treated in chapters 2, 3 and 4 respectively and represent the main body of the thesis. The last chapter describes some of the remaining open problems and possible extensions of this thesis. The rest of this section provides a brief overview of chapters $2-4$.

1.4.1 Statistically Optimum Pre- and Post Filtering in Quantization

Consider the general scheme shown in Fig. 1.7 where the box labeled $\mathcal{QS}$ represents a quantization system. The goal is to optimize the filters such that the mean square error (m.s.e.) is minimized under the key constraint that the quantization noise variance is directly proportional to the variance of the quantization system input. Unlike some previous work, the postfilter is not restricted to be the inverse of the prefilter.

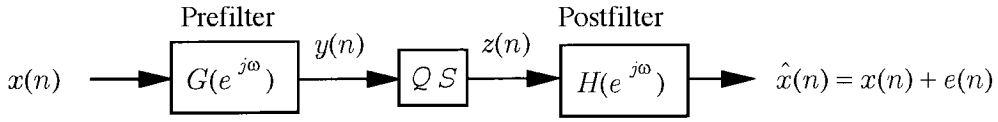

Fig. 1.7: A general pre- and post filtering scheme

With no order constraint on the filters, we present closed form solutions for the optimum pre- and post filters when the quantization system is a uniform quantizer. Using these optimum solutions, we obtain a coding gain expression for the system under study. The coding gain expression clearly indicates that, at high bit rates, there is no loss in generality in restricting the postfilter to be the inverse of the prefilter. We then repeat the same analysis with first order pre- and post filters in the form $1+\alpha z^{-1}$ and $1/(1+\gamma z^{-1})$. In specific, we study two cases : (a) FIR prefilter, IIR postfilter and (b) IIR prefilter, FIR postfilter. For each case, we obtain a mean square error expression, optimize the coefficients α and γ and provide some examples where we compare the coding gain performance with the case of $\alpha = \gamma$. In the last section, we assume that the quantization system is an orthonormal perfect reconstruction filter bank. To apply the optimum pre- and post filters derived earlier, the output of the filter bank must be WSS which, in general, is not true. We provide two theorems, each under a different set of assumptions, that guarantee the wide sense stationarity of the filter bank output. We then propose a suboptimum procedure to increase the coding gain of the orthonormal filter bank.

1.4.2 Oversampling PCM Techniques and Optimum Noise Shapers for Quantizing a Class of Nonbandlimited Signals

We consider the "efficient" quantization of a class of *non bandlimited* signals, namely the class of discrete time signals that can be recovered from its decimated versions. The signal of interest is assumed to be the output of a single interpolation filter (Fig. 1.4) or more generally the sum of the outputs of $L < M$ interpolation filters (Fig. 1.5). By definition, the signal is oversampled and it is reasonable to expect that we can reap the same benefits of well known efficient A/D techniques. In fact, by using appropriate multirate models and reconstruction schemes, we first show that we can obtain a great reduction in the quantization noise variance due to the oversampled nature of the signal. Alternatively, we also show that we can achieve a substantial decrease in bit rate by appropriately decimating the signal and then quantizing it. To further increase the effective quantizer resolution, noise shaping is introduced by optimizing pre- and post filters around the quantizer. We start with a scalar time invariant quantizer and study two important cases of LTI filters, namely the case where the postfilter is the inverse of the prefilter and the more general case where the postfilter is not related to the prefilter. Closed form expressions for the optimum filters and minimum mean squared error

are derived in each case for both the single band and multiband models. Due to the statistical nature of the signal of interest, the class of noise shaping filters and quantizers is then enlarged to include linear periodically time varying $(LPTV)_M$ filters and periodically time varying quantizers of period M. Because the general $(LPTV)_M$ case is difficult to track analytically, we study two special cases in great detail and give complete solutions for both the single band and multiband models. Examples are also provided for performance comparisons between the LTI case and the corresponding $(LPTV)_M$ one.

1.4.3 The design of optimum FIR and IIR energy compaction filters

We propose a new approach to design (in the mean square sense) a weighted FIR filter of arbitrary order N under the constraint that its magnitude squared response is Nyquist(M). The optimization problem can be expressed as follows:

$$\max_{H(e^{j\omega})} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 W(e^{j\omega}) \frac{d\omega}{2\pi} \tag{1.1}$$

subject to

$$\frac{1}{M} \sum_{k=0}^{M-1} |H(e^{j(\omega - 2\pi k/M)})|^2 = |H(e^{j\omega})|^2|_{\downarrow M} = 1 \tag{1.2}$$

where $H(e^{j\omega})$ is a real FIR filter of fixed order N. The constraint (4.2) means in particular that the magnitude squared response $|H(e^{j\omega})|^2$ is Nyquist(M). When $W(e^{j\omega}) = S_{xx}(e^{j\omega})$ where $S_{xx}(e^{j\omega})$ is the power spectral density of the wide-sense stationary input signal $x(n)$, the problem is termed the FIR energy compaction problem.

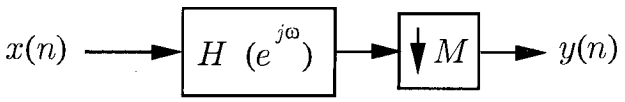

Fig. 1.8: Schematic of the FIR energy compaction problem.

The problem described above arises in many applications such as echo cancelation, time varying systems identification, standard subband filter design and optimal transmitter and receiver design in digital communications to name a few. Although our new formulation is general enough to cover a wide variety of applications depending on the objective function, the focus of the last chapter is on the design of optimum compaction filters. The optimization of such filters has received considerable attention due partly to the fact that they are related to the design of M-channel orthonormal filter banks adapted to the input signal statistics. In particular, for the two channel case, the two problems are equivalent. Similar to some previous work, the new procedure finds the optimum product filter $F_{opt}(e^{j\omega}) = |H_{opt}(e^{j\omega})|^2$ corresponding to the compaction filter $H_{opt}(e^{j\omega})$. Nevertheless, the new design method does not suffer from any of the drawbacks of all previous work:

i) Using well known results in linear system theory, we can automatically ensure the positivity of the resulting optimum product filter $F_{opt}(e^{j\omega})$ *over all frequencies* simultaneously with the Nyquist constraint.

ii) For an input power spectrum, the resulting filter $F_{opt}(z)$ is guaranteed to be *a global* optimum due to the convexity of the new formulation (which other methods could not achieve).

iii) The new design method is expressed as a multi-objective semi definite programming problem which can be solved very efficiently and with great accuracy using recently developed interior point methods.

iv) The new algorithm is extremely general in the sense that it works for any chosen M and any arbitrary given input power spectrum.

v) Unlike previous methods, obtaining $H_{opt}(z)$ from $F_{opt}(z)$ does not require an additional spectral factorization step. The minimum phase spectral factor can be obtained automatically by relating the state space realization of $F_{opt}(z)$ to that of $H_{opt}(z)$.

Several design examples are provided for comparison between the new technique and some previous approaches. In the last section of the chapter, we statistically optimize a well known class of two channel IIR orthonormal filter bank. The whole filter bank is parameterized in this case by *a single* coefficient. The optimization procedure is extremely simple and very fast compared to the FIR case. It is found that for most types of input power spectrums, the compaction gain obtained using this single coefficient IIR filter is very close to the one obtained using very high order FIR filters.

1.5 Standard notations and definitions

1. Lower case letters are used for scalar time domain sequences. Upper case letters are used for transform domain expressions. Bold faced quantities represent vectors and matrices.

2. The superscripts T, $*$ and $\dagger$ denote respectively the transpose, conjugate and conjugate transpose operations for vectors and matrices.

3. $Tr(A)$ denotes the trace of the matrix A.

4. The M-fold downsampler has an input-output relation $y(n) = x(n)\downarrow_M = x(Mn)$. The M-fold expander's input-output relation is $y(n) = x(n)\uparrow_M = x(n/M)$ when $n =$ multiple of M and $y(n) = 0$ otherwise.

5. The M-fold polyphase representation of $X(e^{j\omega})$ is given by $X(e^{j\omega}) = X_0(e^{jM\omega}) + e^{-j\omega}X_1(e^{jM\omega}) + e^{-j2\omega}X_2(e^{jM\omega}) + \ldots + e^{-j(M-1)\omega}X_{M-1}(e^{jM\omega})$. The polyphase components are given by $x_k(n) = x(Mn+k)$ or, in the frequency domain by $X_k(e^{j\omega}) = \big(e^{j\omega k}X(e^{j\omega})\big)\downarrow_M$.

6. The tilde accent on a function $\mathbf{F}(z)$ is defined such that $\tilde{\mathbf{F}}(z)$ is the conjugate transpose of $\mathbf{F}(z)$, i.e., $\tilde{\mathbf{F}}(z) = \mathbf{F}^{\dagger}(1/z^*)$.

7. **Antialias(M) filters.** $F(e^{j\omega})$ is said to be an antialias(M) filter if its output can be decimated M-fold without aliasing, no matter what the input is. Equivalently, there is no overlap between the

plots $F(e^{j(\omega-(2\pi k/M))})$ for distinct k in $0 \leq k \leq M-1$. Since this requires a stopband with infinite attenuation, these are ideal filters.

8. **Orthonormal filter bank.** An M-channel maximally decimated uniform filter bank (FB) is said to have the perfect reconstruction (PR) property when $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{-1}(e^{j\omega})$ where $\mathbf{E}(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$ denote respectively the analysis and synthesis polyphase matrices [73]. In the case of an orthonormal filter bank, the analysis polyphase matrix is paraunitary, i.e., $\mathbf{E}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega}) = \mathbf{I}\ \forall\ \omega$ and we choose $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{\dagger}(e^{j\omega})$ for perfect reconstruction. The analysis and synthesis filters are related by $F_k(e^{j\omega}) = \tilde{H}_k(e^{j\omega})$, that is $f_k(n) = h_k^*(-n)$. It follows that, for an orthonormal filter bank, the energy of each analysis/synthesis filter equals unity, that is $\int_{-\pi}^{\pi} |F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1$.

9. **Nyquist property.** A sequence $x(n)$ is said to be Nyquist if its decimated version $x(Mn) = \delta(n)$ [73]. In the frequency domain, the previous condition becomes

$$\sum_{k=0}^{M-1} X(e^{j(\omega-2\pi k/M)})) = M$$

The above specifically means that $x(n)$ has zero samples at $n = Mi$ where $i \neq 0$ is an integer. Alternatively, $X(e^{j\omega})$ and the $M-1$ shifted copies created by downsampling tile the frequency plane.

Chapter 2

Statistically Optimum Pre- and Post Filtering in Quantization

2.1 Introduction

Consider the general scheme shown in Fig. 2.1 where the box labeled $\mathcal{QS}$ represents a quantization system. The input sequence $x(n)$ is passed through a prefilter $G(e^{j\omega})$ and produces an output $y(n)$. The sequence $y(n)$ is then quantized and filtered with a postfilter $H(e^{j\omega})$ to reproduce an estimate of the input denoted by $\hat{x}(n)$. The quantization system $\mathcal{QS}$ can be a simple uniform quantizer, which we denote by $\mathcal{Q}$, or a more sophisticated quantization system such as the M-channel uniform subband coder (SBC) shown in Fig. 2.2. Assuming that the quantization system is constrained to have a budget of b bits, the main theme in this chapter is to jointly optimize the prefilter $G(e^{j\omega})$ and the postfilter $H(e^{j\omega})$ such that the mean square value $E\{e^2(n)\}$ of the reconstruction error, $e(n) \stackrel{\triangle}{=} \hat{x}(n) - x(n)$, is minimized.

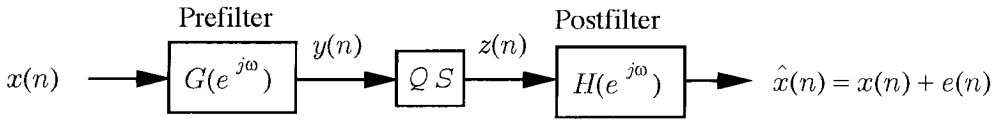

Fig. 2.1: A general pre- and post filtering scheme

The renewed interest in the above classic problem was motivated by its relation to some issues in the area of subband coding. To elaborate, consider the M-channel uniform SBC of Fig. 2.2. The boxes labeled $\mathcal{Q}$ represent subband quantizers, a set of uniform quantizers which are modeled by additive noise sources. An equivalent representation of the uniform SBC is given in Fig. 2.3. It consists of two matrices $\mathbf{E}(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$, known respectively as the analysis and synthesis polyphase matrices. In

the absence of quantizers, the filter bank (FB) is said to have the perfect reconstruction (PR) property if and only if $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{-1}(e^{j\omega})$ [73]. A perfect reconstruction filter bank (PRFB) is also known as a biorthogonal FB. An important subclass of uniform PR filter banks is the class of orthonormal or paraunitary (PU) filter banks. In this case, the analysis polyphase matrix exhibits the lossless property, mathematically expressed as $\mathbf{E}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega}) = \mathbf{I}\ \forall\ \omega$, where the superscript $\dagger$ denotes the conjugate transpose operation. By choosing the synthesis polyphase matrix $\mathbf{R}(e^{j\omega})$ to be equal to $\mathbf{E}^{\dagger}(e^{j\omega})$, perfect reconstruction is guaranteed.

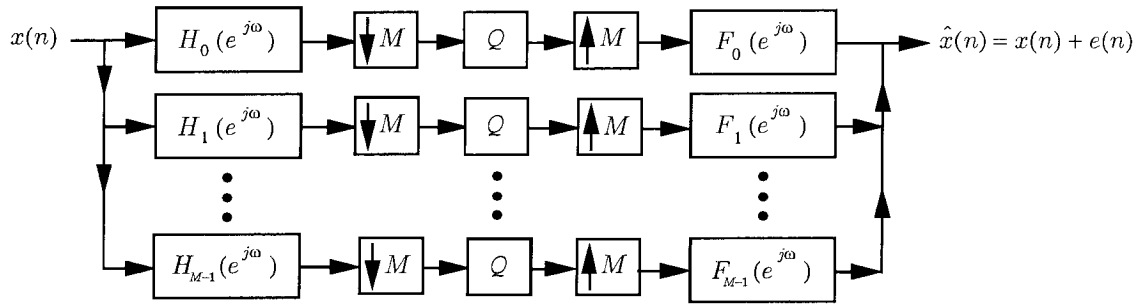

Fig. 2.2: An M-channel uniform maximally decimated subband coder

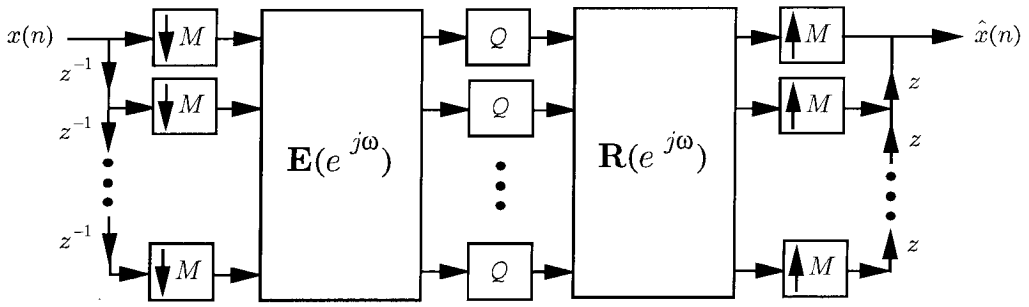

Fig. 2.3: The equivalent polyphase representation for a uniform M-channel subband coder (SBC)

In the presence of quantizers, perfect reconstruction is not possible because quantization is a lossy process. The FB output $\hat{x}(n)$ in this case is the original input $x(n)$ plus a filtered version of the quantization noise denoted by $e(n)$. Recently, several authors have considered the optimization of filter banks when quantizers are present [46, 69, 26, 17]. Given a fixed budget of b bits for the subband quantizers, the aim is to minimize the average variance of $e(n)$. This problem involves optimizing the analysis and synthesis filters and choosing a subband bit allocation strategy. For the sake of further discussions, we will from now on refer to the problem of optimizing a FB in the presence of quantizers as the subband coding problem. In a parallel fashion, interest in the so called energy compaction problem was growing [68, 18, 65]. Although the energy compaction problem might at first seem decoupled from the subband coding problem, Vaidyanathan [76] recently showed that the energy compaction problem and the subband coding problem for the case of an orthonormal SBC are

actually highly connected. In fact, the orthonormal filter bank solution given in [76] for the subband coding problem turns out to be similar to the one given in [65] for the energy compaction problem (ideal filter case). Such filter banks are referred to as ***optimum orthonormal filter banks***. We will use some of the results of optimum orthonormal filter banks later in section 2.5.

Although the subband coding problem was carefully analyzed and solved for the class of orthonormal FB [ideal filter case], the M channel [$M \neq 1$] maximally decimated optimum biorthogonal FB is at this point in time an open problem. Only the solution of the one channel case is well established [37]. Furthermore, it is well known [75] that, in the presence of quantizers, the synthesis polyphase matrix is not necessarily the inverse of the analysis polyphase matrix. Restricting ourselves to the class of biorthogonal FB when quantizers are present is therefore a loss of generality. A similar observation was given by Gosse and Duhamel [26] calling this more general class of filter banks minimum mean square error (MMSE) filter banks. Kovacevic [46] also reaches the same conclusion for the case where the subband quantizer Q is modeled as a Lloyd-Max quantizer. While the synthesis bank was optimized in [75], Vaidyanathan and Chen did not address the issue of optimizing the analysis bank nor the optimum allocation of subband bits.

The joint optimization of the analysis bank and the synthesis bank together with the allocation of subband bits is quite a challenging problem. In this chapter, we will provide a joint optimum solution of the pre- and post filters for the special case of $M = 1$. The system of Fig. 2.1 when the quantization system $\mathcal{QS}$ is a uniform quantizer can indeed be seen as the one channel case of the more general and difficult M channel problem. It is also a generalization of the so-called half-whitening scheme [37] where the postfilter is assumed to be the inverse of the prefilter. A summary of all the chapter's results is given below.

2.1.1 Brief Overview of Past Related Work

The problem of finding optimum pre and postfilters around a noisy processor has been considered by various researchers especially in the field of communication theory. Costas [14] has jointly optimized pre and post filters over an analog communication channel subject to a power constraint on the prefilter. Chan and Donaldson [12] considered the same problem with the input to the postfilter sampled every T seconds. Berger and Tufts [8] optimized transmission and receiving filters in PAM communication systems to minimize the mean square error distortion resulting from channel noise and intersymbol interference. Malvar and Staelin [48] offered an iterative algorithm to design FIR pre- and postfilters in the presence of a downsampler and an upsampler.

The first fundamental difference between the above problems and the quantization problem under study in this chapter is the nature of the noise variance. In specific, we will assume throughout this chapter that the quantization noise variance σ_q^2 is directly proportional to the variance of the input to the quantization system. Such a constraint describes in a fairly accurate manner the interaction

between the quantization system granular noise output and the dynamic range of the quantization system input process. A simple example would be the relation $\sigma_q^2 = c2^{-2b}\sigma_y^2$ used in [37] for the case of a uniform quantizer. In a communication problem setting, the noise source variance is always assumed to be independent of the channel input signal statistics. The second main difference is that, in a communication problem, the prefilter is usually power constrained. This is not the case for the quantizer problem.

Taking a different approach than the one used in communications, Jayant and Noll analyzed the case where the quantization system $\mathcal{QS}$ is a simple uniform quantizer and the postfilter $H(e^{j\omega})$ is simply the inverse of the prefilter, i.e., $H(e^{j\omega}) = 1/G(e^{j\omega})$. Applying the Cauchy-Schwartz inequality, the magnitude response of the optimum filter can be found to be $|G_{opt}(e^{j\omega})| = 1/S_{xx}(e^{j\omega})^{1/4}$. The system was therefore called the half whitening scheme [37] and represents an optimum one channel biorthogonal FB. Recently, Djokovic and Vaidyanathan repeated the analysis for the case where the quantization system $\mathcal{QS}$ is a uniform orthonormal FB [20].

2.1.2 Main Results and Outline of the chapter

1. In the early sections of this chapter, we will assume that the quantization system $\mathcal{QS}$ is a *uniform scalar quantizer*, which we denote by $\mathcal{Q}$. With similar assumptions as the one used by Jayant and Noll in the derivation of the half whitening solution, we derive optimal solutions for the more general scheme of Fig. 2.1. In specific, closed form expressions for the optimum ideal pre- and postfilters are derived in section 2.2.

2. In section 2.3, using the optimum pre- and post filters of section 2.2, we derive an expression for the so called coding gain of the scheme of Fig. 2.1. The beauty of this expression is that it clearly indicates that there is no loss of generality in using the half whitening scheme if we are quantizing at high bit rate, a result that is intuitively very appealing.

3. In section 2.4, we repeat the same type of analysis with first order pre- and post filters with monic polynomials. We derive an expression for the mean square error for the cases of (a) FIR prefilter, IIR postfilter and (b) IIR prefilter, FIR postfilter. We then provide some examples where the coefficients of the filters can be computed numerically. We compare the coding gain of such cases with the one obtained from a first order one channel perfect reconstruction system. Our results indicate again that unless we are quantizing at a very low bit rate, the solution of the more general scheme of Fig. 2.1 tends to the biorthogonal one.

4. In section 2.5, we assume that the quantization system $\mathcal{QS}$ is an orthonormal uniform PRFB. We do not however try to generalize the scheme proposed by Djokovic and Vaidyanathan [20]. Instead, we propose a suboptimum procedure. We first develop two theorems that give sufficient

conditions for wide sense stationarity of the output noise of a non-uniform orthonormal PRFB. We then apply the optimum pre- and post filters of section 2.2 at the input and output of the FB respectively to improve the performance of the original orthonormal PRFB.

2.2 Optimum unconstrained pre- and post filters

The main goal of this section is to jointly optimize the prefilter $G(e^{j\omega})$ and postfilter $H(e^{j\omega})$ of Fig. 2.1 [$\mathcal{QS}$ is a uniform quantizer] to minimize the mean square error $\mathcal{E} \triangleq E\{\hat{x}(n) - x(n)\}^2$ subject to the constraint

$$\sigma_q^2 = c2^{-2b}\sigma_y^2 \tag{2.1}$$

where σ_q^2 is the quantization noise variance, c is a constant that depends on the statistical distribution of $y(n)$ and the overflow probability, and σ_y^2 is the variance of the quantizer input. Our main assumptions for this section are summarized as follows :

1. All random processes are zero mean, real and jointly wide sense stationary.
2. The input $x(n)$ and the quantization noise $q(n)$ are uncorrelated processes, i.e.,

$$E\{x(n)q(m)\} = 0 \ \forall \ n, m.$$

3. The quantization noise $q(n)$ is white with variance σ_q^2 as in (2.1).
4. The filters $H(e^{j\omega})$ and $G(e^{j\omega})$ are not constrained to be rational functions, i.e., the optimum $H(e^{j\omega})$ and $G(e^{j\omega})$ can be ideal filters. Furthermore, no causality constraint is imposed.
5. The power spectral density $S_{xx}(e^{j\omega})$ is positive for all ω. Moreover, when deriving the optimum solution for the prefilter, we will also require $S_{xx}(e^{j\omega})$ and its first derivative to be continuous functions of frequency.

2.2.1 The Optimum Postfilter

To develop optimum closed form solutions for both filters, we first fix the prefilter $G(e^{j\omega})$ and optimize $H(e^{j\omega})$. The optimum postfilter solution is given in the following theorem.

Theorem 1 *For a fixed prefilter $G(e^{j\omega})$, the optimum postfilter $H_{opt}(e^{j\omega})$ is the well-known Wiener filter and is given by :*

$$H_{opt}(e^{j\omega}) = \frac{1}{G(e^{j\omega})} \cdot \frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) + \dfrac{c2^{-2b}}{|G(e^{j\omega})|^2}\displaystyle\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \frac{d\omega}{2\pi}} \tag{2.2}$$

Proof. For a fixed prefilter $G(e^{j\omega})$, the input to the postfilter $H(e^{j\omega})$ is a filtered version of the desired signal embedded in quantization noise. This is a classical Wiener filtering setting and hence, the optimum postfilter is given by [29] $H_{opt}(e^{j\omega}) = S_{xz}(e^{j\omega})/S_{zz}(e^{j\omega})$ where $z(n) = y(n) + q(n)$ is the noisy input to the wiener filter. Since $x(n)$ and $q(n)$ are assumed uncorrelated, it is easy to see that $S_{xz}(e^{j\omega}) = S_{xy}(e^{j\omega}) = G(e^{j\omega})^* S_{xx}(e^{j\omega})$ and $S_{zz}(e^{j\omega}) = S_{yy}(e^{j\omega}) + \sigma_q^2 = |G(e^{j\omega})|^2 S_{xx}(e^{j\omega}) + \sigma_q^2$ where the * denotes complex conjugation. Substituting in the above, we get

$$H_{opt}(e^{j\omega}) = \frac{G(e^{j\omega})^* S_{xx}(e^{j\omega})}{|G(e^{j\omega})|^2 S_{xx}(e^{j\omega}) + \sigma_q^2} = \frac{1}{G(e^{j\omega})} \cdot \frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) + \dfrac{\sigma_q^2}{|G(e^{j\omega})|^2}} \tag{2.3}$$

Substituting the constraint (2.1) in this last equation, we obtain the above solution. ∎

The optimum postfilter can be drawn as in Fig. 2.4. The Wiener filter of (2.2) is therefore expressed as a cascade of two filters : The first filter is the inverse of the prefilter $G(e^{j\omega})$. Its output is simply the original input $x(n)$ embedded in a filtered version of the quantization noise process. The power spectral density of the filtered quantization noise process is $\dfrac{{\sigma_q}^2}{|G(e^{j\omega})|^2}$. The second filter is the optimum Wiener filter for the output of the inverse filter.

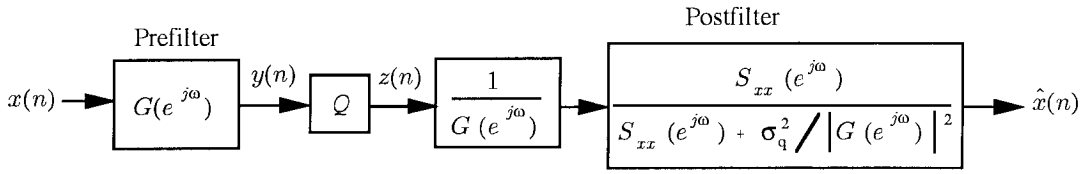

Fig. 2.4: The pre- and post filtering scheme with a uniform quantizer and an optimum postfilter

Using the optimum post filter solution (2.2) and the constraint (2.1), we can now derive an expression for the mean square error only in terms of the prefilter $G(e^{j\omega})$.

$$\begin{aligned}
\mathcal{E} &= E\{e^2(n)\} = E\{e(n) \cdot (\hat{x}(n) - x(n))\} \\
&= E\{e(n) \cdot x(n)\} = E\{(\hat{x}(n) - x(n)) \cdot x(n)\} \\
&= R_{xx}(0) - \sum_{k=-\infty}^{\infty} h(k) \cdot E\{x(n)z(n-k)\} = R_{xx}(0) - \sum_{k=-\infty}^{\infty} h(k) R_{xz}(k)
\end{aligned} \tag{2.4}$$

The second line is obtained from the first using the orthogonality principle [29]. By Parseval's relation, we can then write

$$\begin{aligned}
\mathcal{E} &= R_{xx}(0) - \int_{-\pi}^{\pi} {S_{xz}}^*(e^{j\omega}) H_{opt}(e^{j\omega}) \frac{d\omega}{2\pi} \\
&= \int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) - {S_{xz}}^*(e^{j\omega}) H(e^{j\omega}) \frac{d\omega}{2\pi}
\end{aligned}$$

Substituting with ${S_{xz}}^*(e^{j\omega}) = {S_{xy}}^*(e^{j\omega}) = G(e^{j\omega})S_{xx}(e^{j\omega})$, we obtain

$$\mathcal{E} = \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})(1 - H_{opt}(e^{j\omega})G(e^{j\omega}))\frac{d\omega}{2\pi} \tag{2.5}$$

We note that the previous equation (2.5) holds only for $H_{opt}(e^{j\omega})$. The reason is the use of the orthogonality principle in the derivation of (2.5). To obtain $\mathcal{E}$ only as a function of the prefilter $G(e^{j\omega})$, we substitute $H_{opt}(e^{j\omega})$ into (2.5):

$$\mathcal{E}(|G|, b) = \int_{-\pi}^{\pi} \frac{c2^{-2b}S_{xx}(e^{j\omega})\int_{-\pi}^{\pi} S_{xx}(e^{ju})|G(e^{ju})|^2\frac{du}{2\pi}}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}\int_{-\pi}^{\pi} S_{xx}(e^{ju})|G(e^{ju})|^2\frac{du}{2\pi}}\frac{d\omega}{2\pi}. \tag{2.6}$$

The problem now reduces to finding the prefilter $G(e^{j\omega})$ that minimizes $\mathcal{E}$ as given in (2.6). Two points are in order :

1. Since the mean square error expression (2.6) is a function of $|G(e^{j\omega})|^2$ only, we will be actually seeking an expression for the squared magnitude response of the prefilter rather than $G(e^{j\omega})$.

2. It is clear from (2.6) that trying to derive an optimum analytical expression for $|G(e^{j\omega})|^2$ can be quite tedious. Instead of attacking the problem as it is, the idea is to transform the above unconstrained integral (2.6) into another integral with a power constraint on the prefilter output. The problem then becomes more mathematically tractable and a closed form expression for $|G(e^{j\omega})|^2$ can be obtained. It remains to show that the solution of both problems, the original one and the equivalent one, is the same. This is done in the following claim.

Theorem 2 *The squared magnitude response $|G_{opt}(e^{j\omega})|^2$ that minimizes $\mathcal{E}(|G|, b)$, given as in (2.6), is also the solution of the following constrained optimization problem:*

$$\min_{|G(e^{j\omega})|^2} \int_{-\pi}^{\pi} \frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}}\frac{d\omega}{2\pi} \tag{2.7}$$

subject to:

$$\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G(e^{j\omega})|^2\frac{d\omega}{2\pi} = 1 \tag{2.8}$$

Proof. The role of the magnitude response of the prefilter is basically two fold : It affects the spectral shape of the quantizer input signal $y(n)$ and it changes the quantizer input variance ${\sigma_y}^2$ and therefore the noise variance. The idea is to insert a multiplier α directly before the quantizer. The insertion of this multiplier affect only the variance of the quantizer input. One can then show that the mean square error at the output of this new system is unaffected by this multiplier. This, in turn, indicates that we can always fix the variance of the quantizer input signal $y(n)$ without changing the solution

of our original problem. To prove the argument formally, we proceed as follows: define

$$G^{'}(e^{j\omega}) \triangleq \alpha G(e^{j\omega}) \quad \text{such that} \quad \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G^{'}(e^{j\omega})|^2 \frac{d\omega}{2\pi} = 1 \tag{2.9}$$

Hence,

$$\begin{aligned} \sigma_q^2 &= c2^{-2b} \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \frac{d\omega}{2\pi} \\ &= \frac{1}{\alpha^2} c2^{-2b} \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})|G^{'}(e^{j\omega})|^2 \frac{d\omega}{2\pi} \\ &= \frac{1}{\alpha^2} c2^{-2b} = \frac{1}{\alpha^2}\sigma_{q'}^2 \end{aligned} \tag{2.10}$$

where $\sigma_{q'}^2$ is the quantization noise variance of the system of Fig. 2.5 and is equal to $c2^{-2b}$. The postfilter $H^{'}{}_{opt}(e^{j\omega})$ of the new system is given by (2.3) with $\sigma_{q'}^2$ and $G^{'}(e^{j\omega})$ replacing σ_q^2 and $G(e^{j\omega})$ respectively.

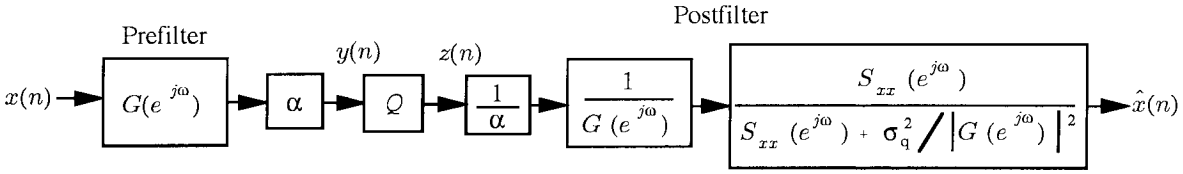

Fig. 2.5: Inserting a multiplier after the prefilter to study the effect of the quantizer input variance

Substituting with $\sigma_{q'}^2$ as in (2.10) and with $G^{'}(e^{j\omega})$ as in (2.9), it is easy to see that

$$H^{'}{}_{opt}(e^{j\omega}) = \frac{1}{\alpha} H_{opt}(e^{j\omega}) \tag{2.11}$$

The filtering scheme of Fig. 2.5 can be redrawn as in Fig. 2.6.

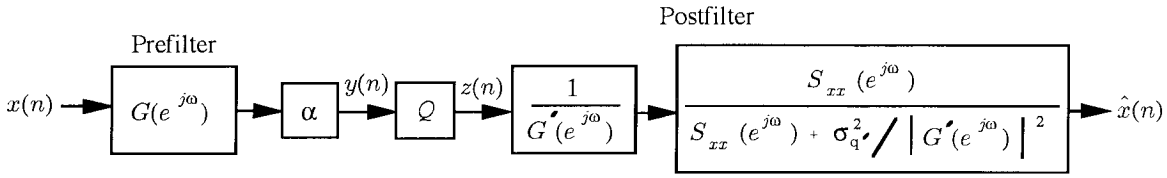

Fig. 2.6: The equivalent pre- and post filtering scheme after the insertion of the multiplier

Following the same type of reasoning as before, the mean square error of the scheme of Fig. 2.6 can be thus expressed as:

$$\mathcal{E}^{'} = \int_{-\pi}^{\pi} S_{xx}(e^{j\omega})(1 - H^{'}{}_{opt}(e^{j\omega})G^{'}(e^{j\omega}))\frac{d\omega}{2\pi} \tag{2.12}$$

By substituting $G^{'}(e^{j\omega})$ and $H^{'}{}_{opt}(e^{j\omega})$ in (2.12) we can immediately see that $\mathcal{E}^{'} = \mathcal{E}$. ■

As a consequence of the above analysis, the mean square error expression reduces to the integral in (2.7).

2.2.2 The Optimum Prefilter

The goal now is to find $|G(e^{j\omega})|^2$ that minimizes the functional (2.7) under the integral constraint (2.8). Since the magnitude squared response is always a non negative function of ω, the optimum minimizing solution we seek must be non negative. This implicit condition is incorporated in the optimization problem as a *pointwise inequality* constraint. The next theorem gives an expression for the optimum magnitude squared response of the prefilter.

Theorem 3 *The prefilter $|G_{opt}(e^{j\omega})|^2$ that minimizes (2.7) under the constraint (2.8) must have a magnitude squared response $|G_{opt}(e^{j\omega})|^2$ in the following form:*

$$|G_{opt}(e^{j\omega})|^2 = \max(0, \frac{1}{\sqrt{S_{xx}(e^{j\omega})}}\left(\frac{1 + c2^{-2b}}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}} - \frac{c2^{-2b}}{\sqrt{S_{xx}(e^{j\omega})}}\right)) \quad \forall\ \omega \in [-\pi, \pi] \tag{2.13}$$

Proof. The minimization of the functional (2.7) under the integral constraint (2.8) and the positivity condition belongs to a class of calculus of variation problems known as isoperimetric problems [24, 63]. An outline of the major steps of the proof with the corresponding equations is given below. For more details, we refer the reader to appendix A.

Step 1. Problem set up. We transform the above constrained problem into an unconstrained one by lumping the integrand of (2.8) to the integrand of (2.7) by a parameter $\lambda(\omega)$. This leads to the following equation:

$$\mathcal{E}_{new} = \int_{-\pi}^{\pi}\left(\frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2\right)\frac{d\omega}{2\pi} \tag{2.14}$$

The parameter $\lambda(\omega)$ takes care of the integral constraint (2.8) which is independent of frequency. We can therefore treat $\lambda(\omega)$ as a constant λ. This last statement can be indeed proved formally [44, page 175]. The optimum magnitude response we seek must obviously be positive over all frequencies. To incorporate this constraint in our problem, we introduce an unspecified parameter $\beta(\omega)$ and consider now the problem of minimizing:

$$\int_{-\pi}^{\pi}\left(\frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + \beta(\omega)|G(e^{j\omega})|^2\right)\frac{d\omega}{2\pi} \tag{2.15}$$

The value of the parameter $\beta(\omega)$ is set in a way that assures that the positivity constraint is never violated. We note that, unlike the parameter λ, $\beta(\omega)$, in this case, takes care of a pointwise constraint. It must therefore be a function of ω.

Step 2. Necessary conditions for an extremum. The key necessary condition for a calculus of variation problem is the Euler-Lagrange equation. For this problem, this is equivalent to requiring $|G(e^{j\omega})|^2$ to satisfy the following equation at all frequencies :

$$\frac{\partial}{\partial |G(e^{j\omega})|^2} \left(\frac{c2^{-2b} S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 \right) = -\beta(\omega) \tag{2.16}$$

Solving the above equation leads to expression (2.13).

Step 3. Sufficient condition for an extremum. The derivation in step 2 indicates that any minimizing curve for (2.7) under the integral constraint (2.8) and the implicit positivity constraint ***must*** have a magnitude response (2.13). Using the convexity of functionals, we finally prove that the solution (2.13) is not only necessary but also sufficient for a minimizing extremum. ■

It follows immediately from this last theorem that the optimum prefilter $G_{opt}(e^{j\omega})$ is not unique since its phase response can be arbitrary set. This is not the case for the optimum postfilter $H_{opt}(e^{j\omega})$. From (2.5), we observe that the mean square error is minimized with respect to the phase response of the filters if the product $G_{opt}(e^{j\omega})H_{opt}(e^{j\omega})$ has zero phase. The phase response of $G_{opt}(e^{j\omega})$ must therefore be the complementary phase of $H_{opt}(e^{j\omega})$. We also note that whenever $|G_{opt}(e^{j\omega})|^2 = 0$, equation (2.2) simplifies to $H_{opt}(e^{j\omega}) = 0$ as well. Finally, for an intuitive interpretation of the above result, we can see, from (2.45) in appendix A, that the magnitude response of the prefilter is set to zero at those frequencies where the noise variance $\sigma_q^2 = c2^{-2b}$ exceeds $\gamma S_{xx}(e^{j\omega})$, γ being a constant defined as in (2.47). *It is therefore better not to prefilter the signal at those frequencies where the noise level is higher (by a certain threshold) than the signal level.*

2.3 Further Analysis of The Optimum One-Channel System

2.3.1 The Coding Gain Expression

Assume that we quantize $x(n)$ directly with b bits. We denote the corresponding mean square error by $\mathcal{E}_{direct}$. We then use the optimum pre and post filters around the quantizer. With the rate of the quantizer fixed to the same value b, we denote the mean square error in this case by $\mathcal{E}_{new}$. The ratio $\mathcal{G}_{opt} \triangleq \mathcal{E}_{direct}/\mathcal{E}_{new}$ is called the coding gain of the new system and, as the name suggests, is a measure of the benefits provided by the pre/post filtering operation. The coding gain expression for the system of Fig. 2.1 with the optimum pre- and postfilters is given in the following theorem.

Theorem 4 *With the optimal choice of pre- and postfilters, the coding gain expression for the scheme of Fig. 2.1 is*

$$\mathcal{G}_{opt} = (1 + c2^{-2b})\mathcal{G}_{hw} \tag{2.17}$$

as long as the right hand side of (2.13) is non-negative $\forall\ \omega$. *Here* $\mathcal{G}_{hw}$ *is the coding gain of the half whitening scheme and is given by*

$$\frac{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2} \tag{2.18}$$

Proof. Following the above definition, the coding gain of the system of Fig. 2.1 can be expressed as:

$$\mathcal{G}_{opt} = \frac{c2^{-2b}\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{c2^{-2b}\int_{-\pi}^{\pi}\frac{S_{xx}(e^{j\omega})}{\left(S_{xx}(e^{j\omega})|G_{opt}(e^{j\omega})|^2 + c2^{-2b}\right)}\frac{d\omega}{2\pi}} \tag{2.19}$$

assuming that the right hand side of (2.13) is always positive. From (2.13), one can then write:

$$|G_{opt}(e^{j\omega})|^2 S_{xx}(e^{j\omega}) + c2^{-2b} = \frac{(1+c2^{-2b})\sqrt{S_{xx}(e^{j\omega})}}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}} \tag{2.20}$$

Substituting (2.20) into (2.19) and simplifying, we obtain (2.17). ■

In the case where the right hand side of (2.13) is set to zero at certain frequencies, we obtain the following coding gain expression:

$$\mathcal{G}_{opt} = (1+c2^{-2b})\frac{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}{\left(\int_{\Omega}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2 + (1+1/c2^{-2b})\int_{\Omega'} S_{xx}(e^{j\omega})\frac{d\omega}{2\pi}}$$

where Ω and Ω' are the set of frequencies over which the right hand side of (2.13) $>\ 0$ and $<\ 0$ respectively. We still expect the filtering scheme under study to outperform the half whitening scheme in this case but it is not clear how one can compare analytically $\mathcal{G}_{opt}$ to $\mathcal{G}_{hw}$.

Example 1. *White input still produces gain.* In this example, we assume that the input $x(n)$ is a white process with variance equal to one. It can be verified for this case that, $|G(e^{j\omega})|^2 = 1 \quad \forall\ \omega$. This is consistent with our earlier observation about the prefilter, namely that it exploits the spectral shape of the input. The postfilter $H(e^{j\omega})$ is a constant, independent of frequency. The coding gain of the half whitening scheme is one since it depends only on the spectral shape of the input. However, the more general system still produces a coding gain $(1+c2^{-2b})$. The gain results from the "Wiener filter part" of the postfilter and, consequently , from the resulting prefilter expression.

Remarks on The Coding Gain Expression

1. ***The coding gain expression for low bit rates.*** It is quite clear from Theorem 4 that the system of Fig. 2.1 will always outperform the half whitening scheme as long as the right hand side in equation (2.13) remains non negative for all frequencies ω. The difference in performance is basically a function of the probabilistic distribution of the quantizer input and more important of the bit rate. Two points are in order: First, the reader should keep in mind that as we quantize at lower bit

rates, the quantizer assumptions made at the beginning of this section become inaccurate questioning therefore the validity of the previous analysis. Second, even if those assumptions hold, the excess gain obtained by using the more general scheme is not worth the extra complexity. For example, for a gaussian input source $x(n)$ and assuming an optimum uniform scalar quantizer, the factor $1 + c2^{-2b}$ provides an extra gain of 0.48 db at $b = 2$, 0.16 db at $b = 3$ and 0.05 db at $b = 4$.

2. ***The coding gain expression for high bit rates.*** By letting b go to infinity, one can easily check that *the right hand side* of equation (2.14) becomes

$$\frac{1}{\sqrt{S_{xx}(e^{j\omega})}}\left(\frac{1}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}}\right) \quad \forall\, \omega \in [-\pi, \pi] \tag{2.21}$$

and is positive $\forall\, \omega$. Therefore, the coding gain expression derived in Theorem 4 can be used and as b goes to infinity, $\mathcal{G}_{opt}$ becomes equal to $\mathcal{G}_{hw}$. At high bit rate ($b \geq 4$), the half whitening scheme is good enough. A similar observation was first mentioned by Goodman and Drouillet [27]. Although the final conclusion is the same, there are main differences between their work and ours. First, Goodman and Drouillet did not derive any coding gain expression. It was quantitatively unclear how much we can benefit from using the more general system of Fig. 2.4. Second, whereas our system is a discrete time system, the system analyzed in [27] was continuous time pre and post filters surrounding a sampler and a quantizer. Moreover, Goodman and Drouillet assumed an additive white noise source model for the quantizer where the noise source is uniform and independent of the quantizer input and its statistics. Although this model is a valid one, we prefer to use the different noise model proposed in [37] by imposing the constraint (2.1) in the beginning of our study. Finally, Goodman and Drouillet replaced the sampler and the quantizer by an additive independent noise source. By doing so, the system becomes identical to the communication system analyzed by Costas [14]. The starting point of Goodman and Drouillet's correspondence is therefore Costas result. This is a different problem as we pointed out in the introduction of this chapter. In our case, we cannot use Costas result directly. The use Theorem 2 is essential in our derivation and it is because of this theorem that the quantization problem under study becomes similar to a communication problem.

2.3.2 Analysis under a Colored Quantization Noise Assumption

The previous analysis can be repeated assuming that the quantization noise is now colored. The noise power spectral density $S_{qq}(e^{j\omega})$ becomes a function of frequency. The remaining assumptions are kept the same. The optimum postfilter in this case can be easily rederived and is given by:

$$H_{opt}(e^{j\omega}) = \frac{1}{G(e^{j\omega})} \cdot \frac{S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) + \dfrac{S_{qq}(e^{j\omega})}{|G(e^{j\omega})|^2}} \tag{2.22}$$

The corresponding mean square error expression can be found to be:

$$\mathcal{E} = \int_{-\pi}^{\pi} \frac{S_{qq}(e^{j\omega})S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + S_{qq}(e^{j\omega})} \frac{d\omega}{2\pi} \tag{2.23}$$

We can again argue that the mean square error at the output of the system does not change by inserting a multiplier before the quantizer. The same type of analysis can therefore be carried out producing the following expression for the magnitude response of the optimum prefilter:

$$|G_{opt}(e^{j\omega})|^2 = \max\left(0, \frac{1}{\sqrt{S_{xx}(e^{j\omega})}}\left(\frac{(1+c2^{-2b})S_{qq}(e^{j\omega})}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}} - \frac{S_{qq}(e^{j\omega})}{\sqrt{S_{xx}(e^{j\omega})}}\right)\right) \quad \forall \quad \omega \in [-\pi, \pi] \tag{2.24}$$

2.4 Optimum pre and post filtering with first order filters

The goal of this section is to try to mimic the same kind of analysis as before with finite order filters. In specific, we will constrain $H(e^{j\omega})$ and $G(e^{j\omega})$ to be first order causal filters with monic polynomials in the form $1 - \alpha z^{-1}$ and $\frac{1}{1-\gamma z^{-1}}$. These first order filters can provide substantial coding gain, are easy to track mathematically and are very economic to implement. The quantization system in Fig. 2.1 is still a uniform quantizer. We will again jointly optimize the first order pre- and post filters to minimize the m.s.e. under the constraint (2.1). All the other assumptions of section 2.2 are the same. We will consider two main cases: a) an FIR prefilter with an IIR postfilter and b) an IIR prefilter with an FIR postfilter. The choice of this combination is not as arbitrary as it may seem. The case where α is not equal to γ can be seen as the first order "version" of the general system of Fig. 2.1 whereas the case of $\alpha = \gamma$ can be interpreted as the first order "version" of the half whitening scheme. In the case of ideal filters, interchanging $G(e^{j\omega})$ and $H(e^{j\omega})$ is merely a change of notation but when dealing with finite order filters, the performance of an FIR prefilter (postfilter) is not in general similar to the performance of an IIR prefilter (postfilter). The two cases must be considered and different results can occur as we will observe through some examples.

2.4.1 The FIR prefilter - IIR postfilter case

In this subsection, the prefilter is in the form $1-\alpha z^{-1}$. The postfilter takes the form $\frac{1}{1-\gamma z^{-1}}$. Under the constraint (2.1), the mean square error expression is derived. It is a function of two variables α and γ and the goal is to jointly optimize these coefficients to minimize the mean square error. The next theorem gives the expression of the mean square error.

Theorem 5 *Assume that the prefilter $G(e^{j\omega})$ is given by $1-\alpha e^{-j\omega}$ and that the postfilter $H(e^{j\omega})$ is $\frac{1}{1-\gamma e^{-j\omega}}$. The mean square error as a function of α and γ, under the constraint (2.1), is given by:*

$$\mathcal{E}(\alpha,\gamma)=\frac{c2^{-2b}\left((1+\alpha^2)R_{xx}(0)-2\alpha R_{xx}(1)\right)}{(1-\gamma^2)}+\frac{(\alpha-\gamma)^2}{(1-\gamma^2)}(R_{xx}(0)+2\sum_{m=1}^{\infty}\gamma^m\ R_{xx}(m)) \tag{2.25}$$

Proof. See appendix B. ∎

By using $\sigma_q^2=c2^{-2b}\sigma_y^2$ where $\sigma_y^2=(1+\alpha^2)R_{xx}(0)-2\alpha R_{xx}(1)$, the mean square error expression of the Theorem 5 can be rewritten as follows:

$$\mathcal{E}(\alpha,\gamma)=\frac{\sigma_q^2}{(1-\gamma^2)}+\frac{(\alpha-\gamma)^2}{(1-\gamma^2)}(R_{xx}(0)+2\sum_{m=1}^{\infty}\gamma^m\ R_{xx}(m)) \tag{2.26}$$

The first term in (2.26) disappears when we do not quantize the signal. In this case, the mean square error can be reduced to zero by setting α equal to γ, i.e., the postfilter is the inverse of the prefilter. However, in the presence of the quantizer, the choice of $\alpha = \gamma$ is not optimal since the choice of γ affects the two terms in (2.26) in different ways. Equation (2.26) also suggests that, at high bit rate, the contribution of the first term in the equation will be almost negligible compared to the contribution of the second term. Hence, as b increases, we should expect the optimum coefficients α_{opt} and γ_{opt} to numerically approach each other.

Even in this very simple case, the problem is highly non-linear in the filter coefficients α and γ. Closed form expressions for the coefficients of the filters in terms of only the second-order statistics of the signal cannot be obtained. However, minimization of the mean square error can be done numerically using for example MATLAB's optimization toolbox.

2.4.2 The IIR prefilter - FIR postfilter case

We can easily derive, from equation (2.26), the mean square error for the dual case, namely when the prefilter is $\frac{1}{1-\gamma z^{-1}}$ and the postfilter is $1-\alpha z^{-1}$. To see this, assume first that there is no quantization. It is then clear that the second term in (2.26), the error due to the mismatch of the coefficients, will not change by switching the position of the filters. When quantization is present, the noise term becomes $(1+\alpha^2)\sigma_q^2$ where the noise variance $\sigma_q^2=c2^{-2b}\sigma_y^2=c2^{-2b}\frac{1}{1-\gamma^2}(R_{xx}(0)+2\sum_{m=1}^{\infty}\gamma^m\ R_{xx}(m))$. The mean square error expression is therefore given by :

$$\mathcal{E}(\alpha,\gamma)=c2^{-2b}\frac{(1+\alpha^2)}{(1-\gamma^2)}(R_{xx}(0)+2\sum_{m=1}^{\infty}\gamma^m\ R_{xx}(m))+\frac{(\alpha-\gamma)^2}{(1-\gamma^2)}(R_{xx}(0)+2\sum_{m=1}^{\infty}\gamma^m\ R_{xx}(m)) \tag{2.27}$$

2.4.3 The Special Case of First Order Filters with equal coefficients

A. The FIR prefilter - IIR postfilter case.

When α is equal to γ, the mean square error becomes a function of one parameter α. The coding gain can be then expressed as follows:

$$\mathcal{G}_{opt} = \frac{R_{xx}(0)(1 - \alpha^2{}_{opt})}{(1 + \alpha^2{}_{opt})R_{xx}(0) - 2\alpha_{opt}R_{xx}(1)} \tag{2.28}$$

If $R_{xx}(1) = 0$, then, the above coding gain expression becomes $\frac{(1 - \alpha^2{}_{opt})}{(1 + \alpha^2{}_{opt})}$. It is then quite clear that the optimum coefficient α_{opt} is equal to zero. No pre-and post filtering can enhance the reconstructed output and the coding gain is simply unity. On the other hand, If $R_{xx}(1) = R_{xx}(0)$, then, the coding gain expression becomes $\frac{(1 + \alpha_{opt})}{(1 - \alpha_{opt})}$. As α_{opt} approaches 1, the coding gain grows unbounded. The tradeoff is the stability of the inverse filter.

Having taken care of these two extreme cases, we now assume that $0 < |R_{xx}(1)| < R_{xx}(0)$ and introduce the following notation: $\frac{R_{xx}(1)}{R_{xx}(0)} \triangleq \rho$ where $-1 < \rho < 1$. The problem expressed in this form was considered by Jayant and Noll [37]. We will therefore only give their final results.

1. The optimum coefficient α_{opt} that minimizes the mean square error expression is given by:

$$\alpha_{opt} = \frac{1}{\rho}\left(1 - \sqrt{1 - \rho^2}\right) \qquad \text{if} \quad -1 < \rho < 1 \tag{2.29}$$

2. The coding gain expression as a function of ρ can be found to be

$$\mathcal{G}_{opt} = \frac{1}{\sqrt{1 - \rho^2}} \tag{2.30}$$

We note that the coding gain expression (2.30) is also the coding gain of a 2-channel Karhunen-Loeve transform (KLT) under the assumption of optimum bit allocation. This is then a case of a one channel biorthogonal FB that is as good as a 2 x 2 KLT [an example of a two channel orthonormal filter bank]. This is interesting in view of the fact that the asymptotic coding gain of a KLT is higher than that of a half whitening filter. A natural question then arises: how does the coding gain of a KLT of block length $(M + 1)$ compare to the coding gain of a half whitening like scheme using filters $A(z)$ and $1/A(z)$ of order M? The coding gain of a KLT of block length $(M + 1)$ is well established [37]. For the half whitening like scheme, since the postfilter is assumed to be the inverse of the prefilter $A(z)$, the mean square error expression is due only to the noise component and can be expressed as follows:

$$\mathcal{E} = c2^{-2b}(a^T\mathbf{R}a)\int_{-\pi}^{\pi} \frac{1}{|A(e^{j\omega})|^2}\frac{d\omega}{2\pi} \tag{2.31}$$

where $a^T = (\,1 \quad a_1 \quad \ldots \quad a_{M-1}\,)$, $A(z) = 1+a_1 z^{-1}+\ldots+a_{M-1}z^{-(M-1)}$ and $\mathbf{R}$ is the $(M+1)\times(M+1)$ autocorrelation matrix. The integral in the above expression has a well known closed form expression in terms of the reflection coefficients k_i (See for example [19]). The following closed form expression for the coding gain can be therefore obtained :

$$\mathcal{G}_{opt} = \frac{R_{xx}(0)\prod_{i=1}^{M}(1-k_i^2)}{a^T\mathbf{R}a} \tag{2.32}$$

The reflection coefficients are related in a non linear fashion to the coefficients of the filter $A(z)$ [29]. For the first order case, k_1 is equal to a_1 and (2.32) simplifies to (2.28). The maximization of (2.32) is however beyond the scope of this chapter.

B. The IIR prefilter - FIR postfilter case.

When α is equal to γ, the mean square error is then given by

$$c2^{-2b}\frac{(1+\alpha^2)}{(1-\alpha^2)}(R_{xx}(0)+2\sum_{m=1}^{\infty}\alpha^m\ R_{xx}(m)) \tag{2.33}$$

In this case, the problem is highly non-linear in the filter coefficient α and an analytical solution is difficult to obtain. On the other hand, the minimization of the mean square error can be easily done numerically. Results are illustrated in the next subsection for some specific examples.

2.4.4 Examples of Optimum Filters for Specific inputs

The examples given in this subsection correspond respectively to the cases of a MA(1), an AR(1) and an AR(5) input process $x(n)$. In each case, we compare the coding gain of the general first order system [$\alpha \neq \gamma$] to the coding gain of the first order system with α equal to γ at various bit rates. The optimization of the coefficients is done numerically using MATLAB's optimization toolbox whenever an analytical expression is difficult to obtain. We also include in our comparison the half whitening coding gain $\mathcal{G}_{hw}$ and the coding gain of the system of Fig. 2.1, $\mathcal{G}_{opt}$. $\mathcal{G}_{hw}$ establishes a theoretical bound on the coding gain of the first order system with α equal to γ whereas $\mathcal{G}_{opt}$ represents the theoretical bound for the more general system [$\alpha \neq \gamma$].

Example 2. *MA(1) process.* Assume that the input $x(n)$ is a zero mean gaussian MA(1) process with an autocorrelation sequence in the form

$$R_{xx}(k) = \begin{cases} 1 & k=0 \\ \theta/1+\theta^2 & k=1,-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.34}$$

It is well known that a MA(1) process has to have $\frac{R_{xx}(1)}{R_{xx}(0)} \leq 1/2$ to ensure that the power spectral density is indeed non negative. We therefore restrict θ to be between -1 and 1. For the FIR prefilter - IIR postfilter case, when α is equal to γ, the ratio $R_{xx}(1)/R_{xx}(0)$ is now equal to $\theta/(1 + \theta^2)$. We therefore simply replace ρ in equations (2.29) and (2.30) by $\theta/(1 + \theta^2)$ to obtain expressions for the optimum coefficient α_{opt} and the optimum coding gain $\mathcal{G}_{opt}$. The power spectrum of the MA(1) process is given by:

$$S_{xx}(e^{j\omega}) = 1 - 2\frac{\theta}{(1+\theta^2)}cos(\omega) \tag{2.35}$$

Substituting (2.35) in (2.18), the coding gain expression of the half whitening scheme for a MA(1) process is given by

$$\mathcal{G}_{hw} = \frac{(1+\theta^2)}{\left(\int_{-\pi}^{\pi} \sqrt{(1+\theta^2 - 2\theta cos(\omega))}\, \frac{d\omega}{2\pi}\right)^2} \tag{2.36}$$

The integral in (2.36) is equal to $F(-0.5\,, -0.5\,;\ 1\ ;\ \theta^2)$ where $F(a\ ,\ b\,;\ c\,;\ d)$ is Gauss's hypergeometric function. From [28], $F(-0.5\,,\ -0.5\,;\ 1\,;\theta^2)$ can be rewritten as $(1 + \theta)F(-0.5\,,\ 0.5\,;\ 1\,;4\theta/(1 + \theta)^2)$. This, in turn, can be simplified to $(1 + \theta)\frac{2}{\pi}E(2\sqrt{(|\theta|)}/(1 + \theta))$ where $E(.)$ is the complete elliptic integral of the second kind. Finally, $\mathcal{G}_{opt}$ is given by (2.17).

The optimization of the coefficients for the FIR prefilter-IIR postfilter case and the IIR prefilter-FIR postfilter with $\alpha \neq \gamma$ were all done numerically using MATLAB's optimization toolbox routine "fmins.m". The plots of the coding gain for the FIR/IIR case are illustrated in Fig. 2.7 and Fig. 2.8 for b equal to 2 and 3 respectively. Similarly, the plots of the coding gain for the IIR/FIR case are shown in Fig. 2.9 and Fig. 2.10.

The dotted curve is the coding gain obtained by the first order equal coefficients scheme whereas the dashed curve is the coding gain of the unequal coefficients case. Also included is the half whitening coding gain, $\mathcal{G}_{hw}$, denoted by the dash-dot curve and the coding gain of the system of Fig. 2.1, $\mathcal{G}_{opt}$, denoted by the solid line curve. From these figures, we can observe that as the bit rate increases, there is no loss of generality in assuming α to be equal to γ.

Example 3. *AR(1) process.* Assume that the input $x(n)$ is a zero mean gaussian AR(1) process with an autocorrelation sequence in the form $R_{xx}(k) = \rho^{|k|}$ where $-1 < \rho < 1$.

For the FIR prefilter - IIR postfilter case, when α is equal to γ, the ratio $R_{xx}(1)/R_{xx}(0)$ is equal to ρ. α_{opt} is therefore given by (2.29) and the coding gain $\mathcal{G}_{opt}$ is given by (2.30). The power spectrum of the AR(1) process is

$$S_{xx}(e^{j\omega}) = \frac{1-\rho^2}{1+\rho^2 - 2\rho cos(\omega)} \tag{2.37}$$

Substituting (2.37) in (2.18), the half whitening coding gain expression for the AR(1) process is as

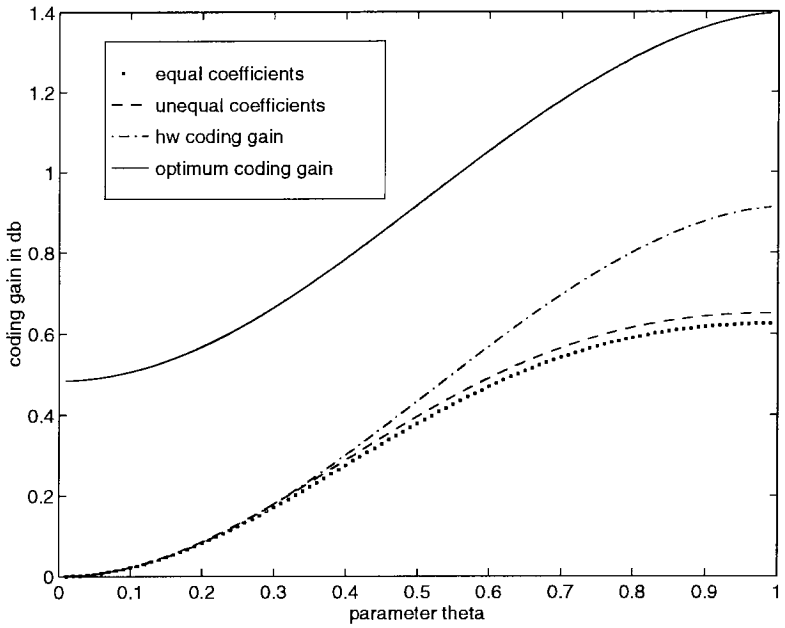

Fig. 2.7: Coding gain curves for the MA(1) case : FIR prefilter, IIR postfilter and $b = 2$

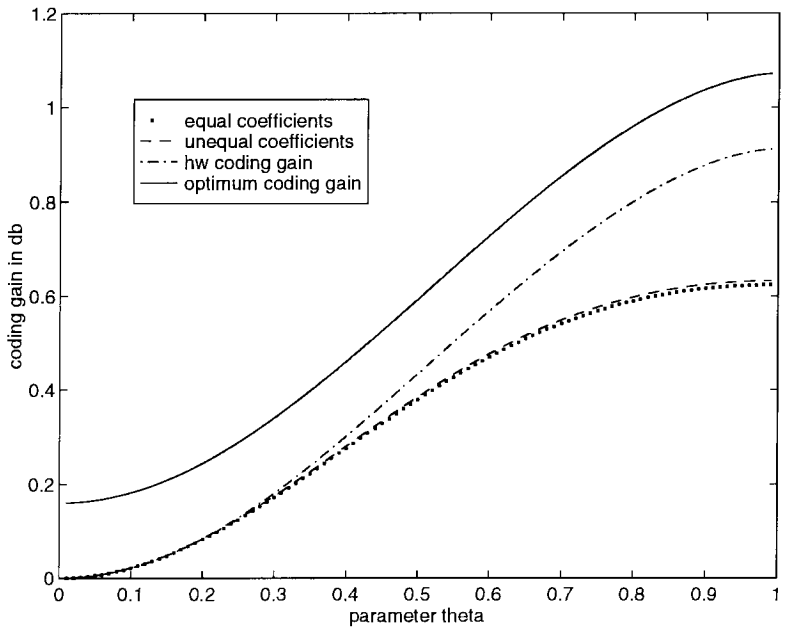

Fig. 2.8: Coding gain curves for the MA(1) case : FIR prefilter, IIR postfilter and $b = 3$

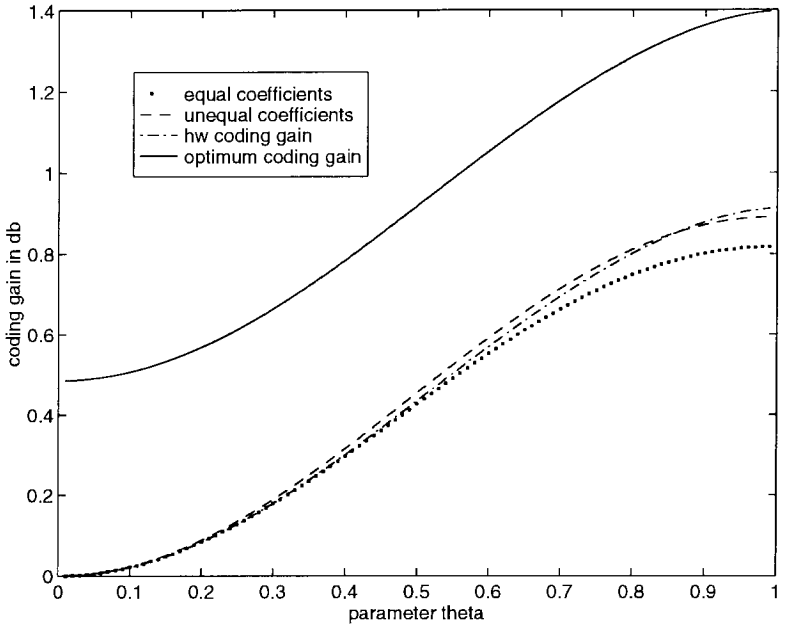

Fig. 2.9: Coding gain curves for the MA(1) case : IIR prefilter, FIR postfilter and $b = 2$

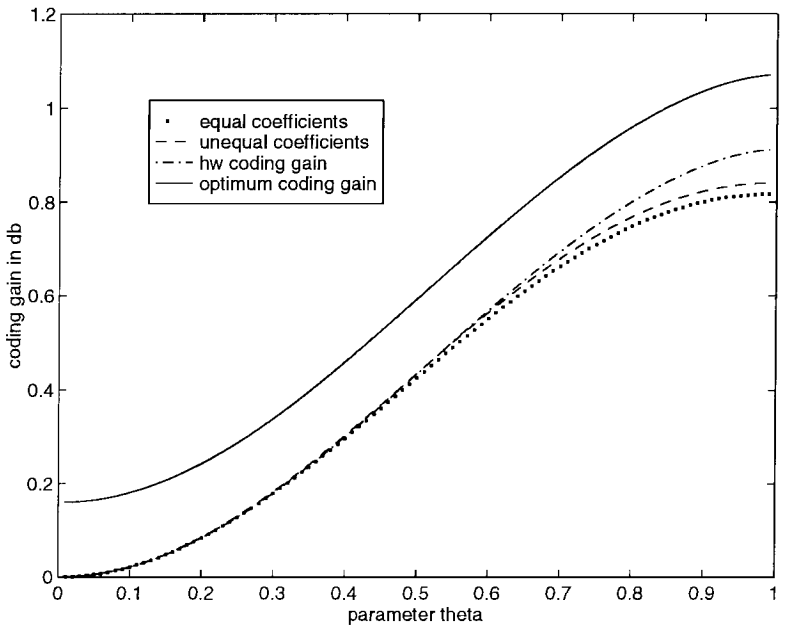

Fig. 2.10: Coding gain curves for the MA(1) case : IIR prefilter, FIR postfilter and $b = 3$

follows:

$$\mathcal{G}_{hw} = \frac{1}{(1-\rho^2)\left(\int_{-\pi}^{\pi} \frac{1}{\sqrt{(1+\rho^2-2\rho cos(\omega))}} \frac{d\omega}{2\pi}\right)^2} \tag{2.38}$$

The integral in (2.38) is equal to $\frac{2}{\pi}K(\rho)$ where $K(\rho)$ is the complete elliptic integral of the first kind [28]. The coding gain of the system of Fig. 2.1, $\mathcal{G}_{opt}$, is again given by (2.17).

The optimization of the coefficients for the FIR prefilter-IIR postfilter case and the IIR prefilter-FIR postfilter with $\alpha \neq \gamma$ were all done numerically using the same MATLAB's optimization toolbox routine "fmins.m". The plots of the coding gain are illustrated in Fig. 2.11 and Fig. 2.12 for the FIR/IIR case and in Fig. 2.13 and Fig. 2.14 for the IIR/FIR case as the bit rate b varies from 2 to 3. The same curve notation as in the previous MA(1) example is used and the same conclusion can be reached.

Example 4. *AR(5) process.* The autocorrelation function of such a process extends to infinity and doesn't have a simple closed form expression. The main problem is the infinite summation in the form $\sum_{m=1}^{\infty} \gamma^m R_{xx}(m)$ found in equations (2.25), (2.27) with $\alpha \neq \gamma$ and (2.33) with $\gamma = \alpha$. Our approach is to truncate this infinite summation with the assumption that after a certain lag m, the correlation coefficients are negligible. For this AR(5) process, we set $R(0) = 1$, $R(1) = 0.86$, $R(2) = 0.64$, $R(3) = 0.4$, $R(4) = 0.26$, $R(5) = 0.2$ and $R(m) = 0 \quad \forall\, m \geq 6$. The values of the correlation coefficients are obtained from [37, page 37]. Table 2.1 summarizes our coding gain results in db for the different cases and bit rates. Again, as b increases, we observe that there is almost no loss in coding gain if we assume that $\alpha = \gamma$. We also observe that, at low bit rate, e.g., $b = 1$, the coding gain of the more general system is very small. This suggests that the gain obtained from searching over a more general class than the biorthogonal class may not be worth the added complexity as we have mentioned previously.

	b = 1	b = 2	b = 3
FIR/IIR $\alpha \neq \gamma$	3.05	2.96	2.94
FIR/IIR $\alpha = \gamma$	2.92	2.92	2.92
IIR/FIR $\alpha \neq \gamma$	3.76	3.52	3.42
IIR/FIR $\alpha = \gamma$	3.37	3.37	3.37

Table 2.1: The coding gain in db obtained from first order filters for the AR(5) process of Example 4.

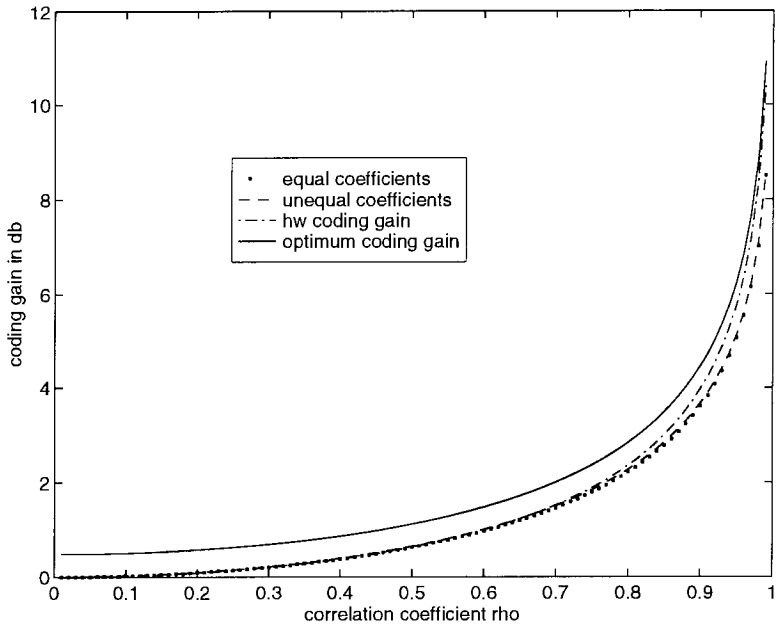

Fig. 2.11: Coding gain curves for the AR(1) case : FIR prefilter, IIR postfilter and $b = 2$

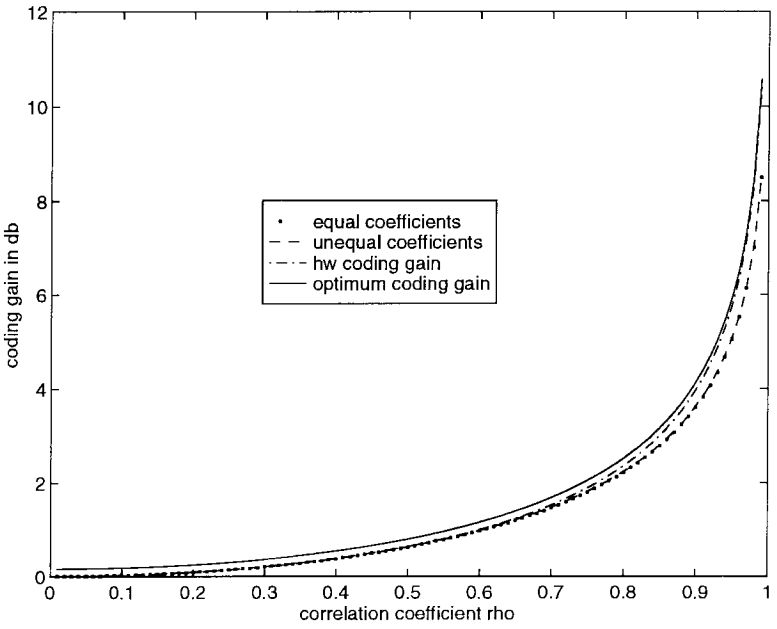

Fig. 2.12: Coding gain curves for the AR(1) case : FIR prefilter, IIR postfilter and $b = 3$

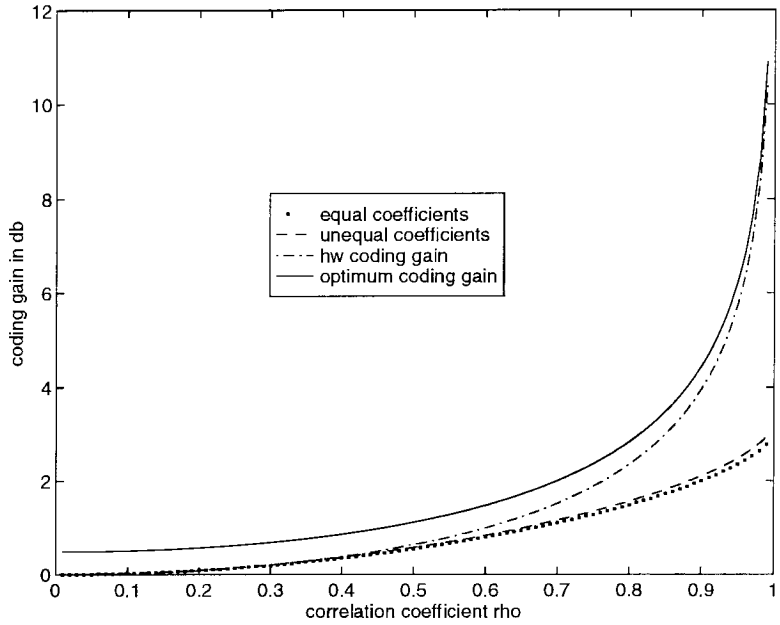

Fig. 2.13: Coding gain curves for the AR(1) case : IIR prefilter, FIR postfilter and $b = 2$

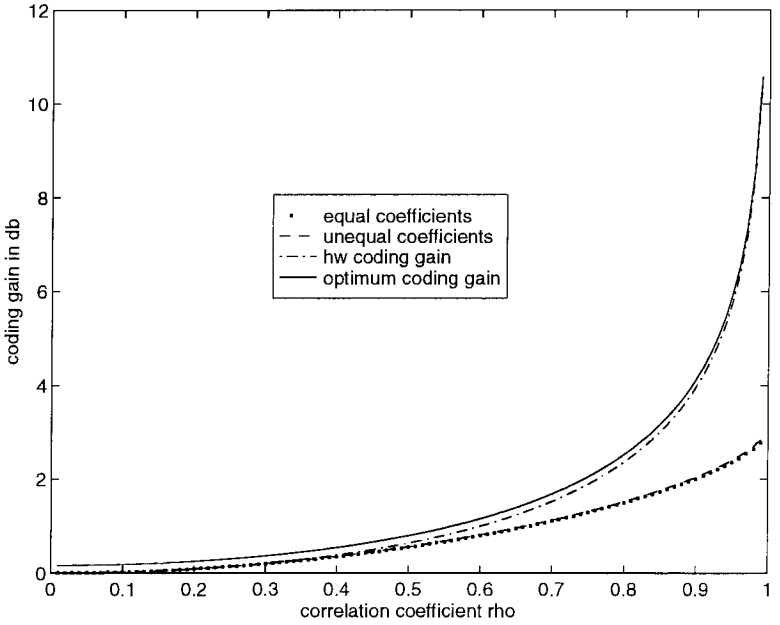

Fig. 2.14: Coding gain curves for the AR(1) case : IIR prefilter, FIR postfilter and $b = 3$

2.5 Replacing the quantizer system by an orthonormal uniform PRFB

Consider the M channel maximally decimated uniform SBC of Fig. 2.2. The boxes labeled $\mathcal{Q}$ are modeled by additive noise sources in the manner described in the introduction. Throughout this section, we will assume that the subband quantization noise sources are white and pairwise uncorrelated. If we interpret this FB as a sophisticated quantizer, the use of pre- and post filters around the FB can increase the coding gain. In a recent paper, Djokovic and Vaidyanathan [20] analyzed the system of Fig. 2.1 where the quantization system $\mathcal{QS}$ is a uniform orthonormal FB and the postfilter is the inverse of the prefilter. The authors gave a formula for the optimum allocation of bits in the subbands. Furthermore, they showed that minimizing the mean square error of the so called prefiltered paraunitary (PPU) FB can be done by separately optimizing the pre/post filtering scheme and the orthonormal filter bank. Their proposed solution was a half whitening scheme surrounding an optimum orthonormal FB. A generalization of the scheme of Djokovic and Vaidyanathan would be again to relax the assumption that the postfilter is the inverse of the prefilter. An analytical *optimum* solution, if it exists, must incorporate the joint optimization of the orthonormal FB and the pre- and post filters. It is not clear that a separate optimization of the pre- and post filters and the orthonormal FB still holds in this case. Furthermore, any developed optimum bit allocation formula must include the pre-and post filtering operation.

In the remainder of this section, we will provide a *suboptimum* procedure that relies on the results derived in section 2.2. We will see that even in this simpler case, two theorems must be first established. The first step in the procedure is to optimize the orthonormal uniform FB for a certain WSS input $x(n)$. Vaidyanathan has recently shown [76] that the optimum orthonormal uniform FB, the one that maximizes the coding gain defined in section 2.3, will consist of antialias filters. Recall from section 1.2 that a discrete time filter is said to be an ***antialias(M)*** filter if its output can be decimated M-fold without aliasing. Since this requires infinite attenuation in the stopbands, anti-alias filters are therefore a class of ideal filters. The second step in the procedure is to perform the optimum bit allocation operation in the usual way [61]. After optimally allocating the bits, we would like to apply the pre- and post filters derived previously in section 2.2. In order to do this, we need first to replace the whole optimum orthonormal FB by an additive noise source, say $v(n)$. This noise source $v(n)$ must be WSS and uncorrelated with the prefilter output $y(n)$. Second, the variance of the input process σ_x^2 must be related to the FB output noise variance $\mathcal{E}_{SBC}$ in a similar fashion as in equation (2.1). A major problem is the following: In the presence of quantizers, it is well known that the output of a uniform/non uniform FB is in general a cyclo-wide sense stationary (CWSS) process. The cyclo-wide sense stationarity is due to the passage of the quantization noise through the interpolators [58]. We provide two results describing important cases that guarantee the wide sense stationarity

of the quantization noise of a uniform ***orthonormal*** FB. Since the results hold for the non uniform decimation case, the proofs will assume a non uniform maximally decimated orthonormal FB case. A non uniform SBC, shown in Fig. 2.15, is a SBC with unequal subband decimation ratios n_k. The boxes labeled $\mathcal{Q}$ represent, as before, uniform quantizers that are modeled by additive noise sources.

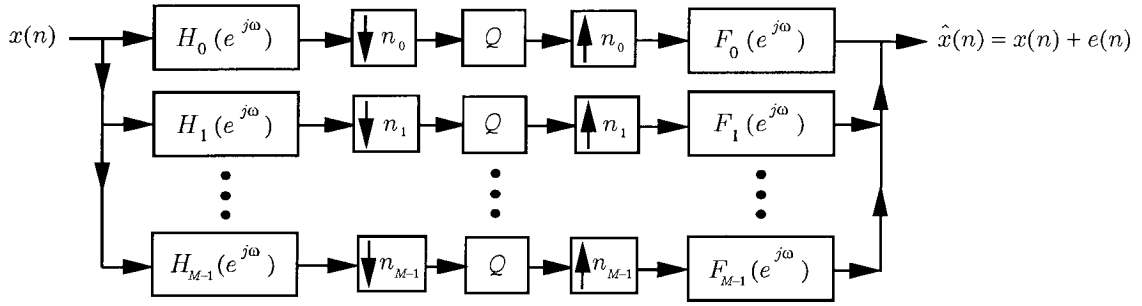

Fig. 2.15: An M-channel non uniform subband coder (SBC)

Theorem 6 *Under optimum bit allocation, the output noise of a [possibly non uniform] orthonormal PRFB is WSS provided the subband quantization noise sources are White, Uncorrelated and Zero Mean (WUZE assumptions).*

Proof. The proof is now established through the following series of steps:

1. Soman and Vaidyanathan [61] showed that for a non uniform orthonormal PRFB, the variances of the subband quantization noises should be equal under optimum bit allocation. Because we are assuming optimum bit allocation in our theorem, we can immediately conclude that the noise variances in the non uniform orthonormal PRFB should be equal to each other.

2. It is well known [45, 32, 74] that an M channel non uniform FB can be redrawn as an L channel maximally decimated uniform FB where $L = n_k p_k$. The set of M analysis and synthesis filters $\{H_k(z), F_k(z)\}$ are replaced by the set of L filters $\{H_k^{'}(z), F_k^{'}(z)\}$ in the uniform system where $L \geq M$. The main goal at this point is to develop the form of the power spectral density matrix of the subband quantization noise $\mathbf{S_{qq}}(e^{j\omega})$ in the equivalent L channel maximally decimated uniform FB. We first observe that the white noise assumption guarantees that, for the k*th* channel, the quantization noises in its corresponding p_k channels are uncorrelated. Furthermore, the variance of the quantization noise is the same in all the p_k channels. Combining this observation with the conclusion of step 1, it is easy to see that $\mathbf{S_{qq}}(e^{j\omega})$ should be equal to $\sigma_q^2 I$ where $\mathbf{S_{qq}}(e^{j\omega})$ is an $L \times L$ matrix.

3. Since the non uniform maximally decimated FB is orthonormal and exhibits the perfect reconstruction (PR) property, then, it follows that the analysis polyphase matrix $\mathbf{E}^{'}(e^{j\omega})$ of the equivalent L channel uniform FB is lossless , i.e., $\mathbf{E}^{'}(e^{j\omega})\mathbf{E}^{'\dagger}(e^{j\omega}) = \mathbf{I}$ (orthonormality) and the synthesis polyphase matrix $\mathbf{R}^{'}(e^{j\omega})$ of the equivalent L channel uniform FB is equal to $\mathbf{E}^{'\dagger}(e^{j\omega})$ (PR property) [74]. The power spectral density matrix of the output quantization noise $\mathbf{S_{vv}}(e^{j\omega})$ is equal to

$\mathbf{R}'(e^{j\omega})\mathbf{S_{qq}}(e^{j\omega})\mathbf{R}'^{\dagger}(e^{j\omega})$ which can be evaluated as $\sigma_q^2 I$ using the above properties. This means that the output noise $v(n)$ is an interleaved version of L uncorrelated white noise sources, each of variance σ_q^2. So, $v(n)$ itself is white with variance σ_q^2. ■

Since the above theorem holds for a non uniform maximally decimated orthonormal PRFB, it includes the uniform decimation case. The output quantization noise $v(n)$ in Theorem 6 is white with variance σ_q^2. Furthermore, $v(n)$ and $y(n)$ are uncorrelated. The problem with the optimum bit allocation is that it yields non integer solution for the bits. If we use a simple rounding procedure or a more sophisticated algorithm [36] to obtain integer solutions, the assumption of equal quantizer noise variances is not valid any more. Nevertheless, in the next theorem, we prove that even with the more practical assumption of different quantization noise variances, the output of a non-uniform orthonormal PRFB with ***antialias*** filters will be wide sense stationary.

Theorem 7 *The output noise of a [possibly nonuniform] orthonormal PRFB consisting of* ***antialias*** *filters is WSS provided the subband quantization noise sources are zero mean and pairwise uncorrelated.*

Proof. Consider the synthesis bank of a non uniform PRFB. The quantization noise sources $q_k(n)$ at the input of the interpolators are assumed to be WSS with power spectrum $S_{q_k}(e^{j\omega})$ and are pairwise uncorrelated. Since the filters $F_k(e^{j\omega})$ are antialias for all k, then, each upsampled and filtered noise sequence $v_k(n)$ is WSS [58]. Furthermore, since the interpolated noise sequences $v_k(n)$ are linear combinations of the input noise sources $q_k(n)$, the uncorrelatedness property is preserved. This can be verified by writing the output vector $\mathbf{v}(n)$ as a time varying linear combination of the vector $\mathbf{q}(n)$ and taking expectations. The interpolated noise sources $v_k(n)$ are therefore jointly wide sense stationary which implies that their sum $v(n)$ is WSS. ■

We emphasize the fact that neither *the whiteness of the noise sources nor the equal variance* assumptions are required for this theorem to hold. We note that the output quantization noise $v(n)$ in Theorem 7 is still uncorrelated with the prefilter output $y(n)$. However, in this case, $v(n)$ is not white. If the subband quantization noise sources are white, it is easy to see that the power spectral density $S_{vv}(e^{j\omega})$ of the PRFB output noise is piecewise constant. The magnitude of each piece of $S_{vv}(e^{j\omega})$ is equal to $\sigma_{q_k}^2$ for some k. The location of the constant piece is determined by the passband of the corresponding synthesis filter $F_k(e^{j\omega})$. The variance of the output noise $v(n)$ is the average of the individual noise variances $\frac{1}{M}\sum_{k=0}^{M-1}\sigma_{q_k}^2$.

The above two theorems permits the continuation of our suboptimum procedure. The optimum bit allocation [without including the pre- and post filters] allows us to relate the variance of the input process σ_x^2 to the FB output noise variance $\mathcal{E}_{SBC}$ by $\mathcal{E}_{SBC} = \frac{c2^{-2b}}{\mathcal{G}_{PU}}\sigma_x^2$. The optimum orthonormal FB is a special case of a non uniform PRFB with antialias filters for which Theorem 7 applies. The FB can be therefore modeled as a WSS noise source that is uncorrelated with the prefilter output sequence $y(n)$ and has a variance proportional to σ_y^2. This is the perfect setting for our previous pre-

and post filtering analysis. The complete system is shown in Fig. 2.16.

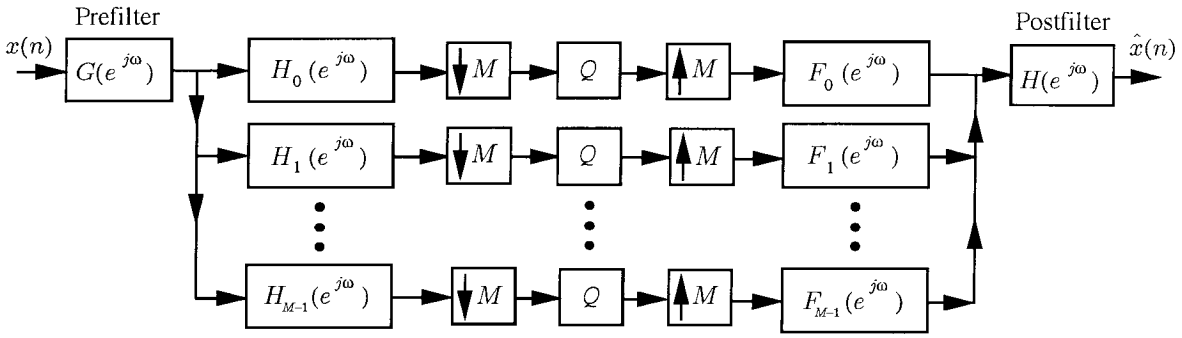

Fig. 2.16: The optimum uniform orthonormal FB with the general pre- and post filtering scheme

The expressions for the optimum postfilter and the magnitude response of the optimum prefilter are given respectively by (2.3) and (2.13) if the noise $v(n)$ is white [case of Theorem 6] or by (2.22) and (2.24) if the noise $v(n)$ is colored [case of Theorem 7]. For either cases, the coding gain of the system of Fig. 2.16 can be easily obtained as $(1 + \frac{c^{-2b}}{\mathcal{G}_{PU}})\, \mathcal{G}_{hw}\mathcal{G}_{PU}$ provided the right hand side of (2.13) or (2.24) is always positive. The next example illustrates the above procedure and provide some numerical results.

Example 5. We assume that the input $x(n)$ is a zero mean, real, WSS random process with a triangular power spectral density $S_{xx}(e^{j\omega})$ as shown in Fig. 2.17.

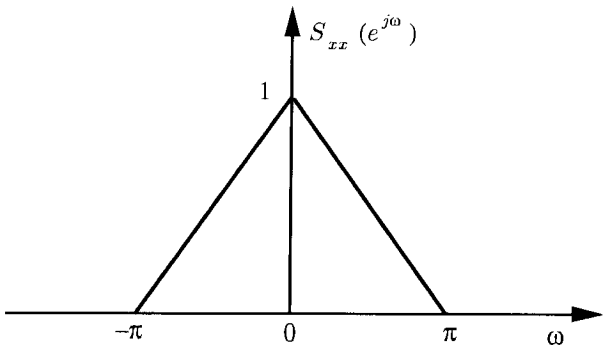

Fig. 2.17: The power spectral density for the input of example 5.

The optimum orthonormal FB in this case is the well-known contiguous ideal brick wall FB [68, 18, 65, 76]. The coding gain of an orthonormal FB after the optimum allocation of subband bits is in general given by [61]:

$$\mathcal{G}_{PU} = \frac{\sigma_x^2}{(\prod_{k=0}^{M-1} \sigma_{x_k}^2)^{1/M}}$$

where σ_x^2 is the variance of $x(n)$ and $\sigma_{x_k}^2$ is the variance of the kth subband signal. For the ideal brick

wall FB and a triangular power spectral density, the above coding gain expression can be simplified to the following expression:

$$\mathcal{G}_{PU} \;=\; \frac{0.5}{\left(\prod_{k=0}^{M-1}(1 \;-\; (2k+1)/2M)\right)^{1/M}}$$

We then apply the optimum pre- and post filters at the input and output of the FB respectively. For an average bit rate $b = 3$, the constant $c = 0.75$ and the number of channels $M = 2$, it can be verified that the optimum prefilter [in both cases of white and colored noise] is never set to zero at any frequency and, therefore, we can use the formula $\mathcal{G}_{opt} = (1 \;+\; \frac{c2^{-2b}}{\mathcal{G}_{PU}})\; \mathcal{G}_{hw}\mathcal{G}_{PU}$. Using the above data, we obtain $\mathcal{G}_{PU} = \frac{2}{\sqrt{3}}$ and $\mathcal{G}_{hw} = \frac{9}{8}$. Finally, the theoretical bound on the coding gain, namely the prediction gain, is given by [73]:

$$\mathcal{G}_{th} \;= \frac{\sigma_x^2}{exp\{\int_{-\pi}^{\pi} ln(S_{xx}(e^{j\omega}))\frac{d\omega}{2\pi}\}} \tag{2.39}$$

For this case, $\mathcal{G}_{th}$ is equal to $\frac{e}{2}$. Expressing the above quantities in db, we get : $\mathcal{G}_{PU} = 0.625$ db, $\mathcal{G}_{hw} = 0.51$ db, $\mathcal{G}_{opt} = 1.19$ db and $\mathcal{G}_{th} = 1.33$ db. It is important to observe the relative gain obtained using the pre- and post filtering operation rather than the absolute value of the coding gain. Clearly, we get a substantial increase by using the pre- and post filters as $\mathcal{G}_{opt}$ approaches $\mathcal{G}_{th}$.

Appendix A.

Step 1. We have argued in the proof of Theorem 3 that the parameter $\lambda(\omega)$ is independent of frequency. We proceed to prove that it is a positive constant. Assume for the moment that $\beta(\omega)$ is equal to zero in (2.16) and denote the integrand of (2.16) by $F \;+\; \lambda W$ where

$$F \;= \frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} \quad \text{and} \quad W \;= S_{xx}(e^{j\omega})|G(e^{j\omega})|^2$$

From the theorem in [24, page 43], we see that if $|G(e^{j\omega})|^2$ is an extremum of (2.16) with $\beta(\omega) \;=\; 0$ [but is not in the same time an extremum of W], then, there exists a constant parameter λ such that $\partial F/\partial|G(e^{j\omega})|^2 \;+\; \partial W/\partial|G(e^{j\omega})|^2 \;=\; 0$ for all ω. Since $|G(e^{j\omega})|^2$ is not an extremal for W, then, there is a ω_o such that $\partial W/\partial|G(e^{j\omega})|^2 \neq\; 0$ at $\omega \;=\; \omega_o$. This yields

$$\lambda \;=\; -\left.\frac{\partial F/\partial|G(e^{j\omega})|^2}{\partial W/\partial|G(e^{j\omega})|^2}\right|_{\omega=\omega_o} \tag{2.40}$$

The numerator and denominator of (2.40) are found to be:

$$\frac{\partial F}{\partial |G(e^{j\omega})|^2} = - \frac{c2^{-2b} S_{xx}(e^{j\omega})^2}{\left(S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 + c2^{-2b} \right)^2} \qquad \& \qquad \frac{\partial W}{\partial |G(e^{j\omega})|^2} = S_{xx}(e^{j\omega}) \tag{2.41}$$

Substituting (2.41) into (2.40), we obtain the following:

$$\lambda = \frac{c2^{-2b} S_{xx}(e^{j\omega})}{\left(S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 + c2^{-2b} \right)^2} \tag{2.42}$$

which in particular shows that $\lambda > 0$.

Step 2. The necessary conditions for $|G(e^{j\omega})|^2$ to be a minimum of (2.15) are summarized next.

1. Define G' to be the derivative of $|G(e^{j\omega})|^2$ with respect to ω. The Legendre condition [24],

$$\frac{\partial^2}{\partial G'^2}(F + \lambda W) \geq 0 \quad \forall\, \omega \tag{2.43}$$

must be satisfied. In our case, this condition is satisfied trivially because neither the functional (2.7) nor the constraint (2.8) are functions of the derivative of $|G(e^{j\omega})|^2$.

2. $|G(e^{j\omega})|^2$ must satisfy the Euler-Lagrange equation for the functional (2.15) i.e $|G(e^{j\omega})|^2$ must satisfy (2.16).

The Euler-Lagrange equation (2.16) is a pointwise relation that must be satisfied at all frequencies. The value of the unknown parameter $\beta(\omega)$ in the right hand side is therefore set according to two criterions: First, the choice of $\beta(\omega)$ should not violate the Euler-Lagrange equation at any frequency. Second, the choice of $\beta(\omega)$ should insure the positivity of the solution at all frequencies. There are therefore two possible values for $\beta(\omega)$.

Case of $\beta(\omega) = 0$. Assume first that $\beta(\omega) = 0$. The left hand side of (2.16) is now equal to zero and, in this case, equation (2.16) can be interpreted as the Euler-Lagrange equation for an exactly similar problem *without a positivity constraint* on the solution. Therefore, for those frequencies where $\beta(\omega) = 0$, the positivity constraint is actually ineffective and the solution we obtain must be ≥ 0 at those frequencies. The optimum magnitude squared response $|G_{opt}(e^{j\omega})|^2$ in this case is determined from (2.16) with the right hand side set to zero. Perform the partial differentiation in (2.16) and equating the result to zero, the following equation can be obtained:

$$\left(S_{xx}(e^{j\omega}) |G(e^{j\omega})|^2 + c2^{-2b} \right)^2 = \frac{c2^{-2b}}{\lambda} S_{xx}(e^{j\omega}) \tag{2.44}$$

Taking the square root of (2.44) and simplifying, we get:

$$|G(e^{j\omega})|^2 = \frac{1}{\gamma}\sqrt{\frac{c2^{-2b}}{S_{xx}(e^{j\omega})}} - \frac{c2^{-2b}}{S_{xx}(e^{j\omega})} \tag{2.45}$$

where $\gamma = \sqrt{\lambda}$. Substituting $|G(e^{j\omega})|^2$ as in (2.45) into the constraint (2.8), we obtain:

$$\frac{1}{\gamma}\int_{-\pi}^{\pi}\sqrt{\frac{c2^{-2b}}{S_{xx}(e^{j\omega})}}S_{xx}(e^{j\omega})\,\frac{d\omega}{2\pi} = 1 + \int_{-\pi}^{\pi} c2^{-2b}\,\frac{d\omega}{2\pi} \tag{2.46}$$

Hence, the constant γ is given by:

$$\gamma = \frac{\sqrt{c2^{-2b}}}{1 + c2^{-2b}}\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\,\frac{d\omega}{2\pi} \tag{2.47}$$

Substituting γ as in (2.47) in (2.45), we therefore obtain part of equation (2.13), namely that:

$$|G(e^{j\omega})|^2 = \frac{1}{\sqrt{S_{xx}(e^{j\omega})}}\left(\frac{1 + c2^{-2b}}{\int_{-\pi}^{\pi}\sqrt{S_{xx}(e^{j\omega})}\frac{d\omega}{2\pi}} - \frac{c2^{-2b}}{\sqrt{S_{xx}(e^{j\omega})}}\right) \tag{2.48}$$

for all frequencies for which the right hand side of (2.48) is non-negative.

Case of $\beta(\omega) \neq 0$. At some particular frequency, the solution obtained in case 1 might turn out to be negative. The positivity constraint is obviously violated. At such a frequency, $\beta(\omega)$ should not bet set to zero anymore. Since the Euler-Lagrange equation must be satisfied at all times, we must set $\beta(\omega)$ to be equal to:

$$-\frac{\partial}{\partial|G(e^{j\omega})|^2}\left(\frac{c2^{-2b}S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega})|G(e^{j\omega})|^2 + c2^{-2b}} + \lambda S_{xx}(e^{j\omega})|G(e^{j\omega})|^2\right)$$

The sign of $\beta(\omega)$ in this case is important to make sure that the positivity constraint is not violated. For our problem, $\beta(\omega)$ should be ***non positive***. Finally, it remains to find the value of $|G_{opt}(e^{j\omega})|^2$ when $\beta(\omega) \neq 0$. The Euler-Lagrange equation cannot be used anymore because it determines the unknown parameter $\beta(\omega)$. However, we can simply observe that $|G_{opt}(e^{j\omega})|^2$ cannot be greater than zero. According to the first case, if $|G_{opt}(e^{j\omega})|^2$ is set to any value greater than zero, $\beta(\omega)$ should be zero. The only possible remaining value for $|G_{opt}(e^{j\omega})|^2$ is therefore zero. This argument establishes the complete form of equation (2.13).

From the above construction, we see that $|G_{opt}(e^{j\omega})|^2$ is a smooth function of ω (i.e it is continuous with continuous first order derivative) everywhere except at the frequencies where it has to be forced to zero (so it does not turn negative). The frequencies ω_k at which $|G_{opt}(e^{j\omega})|^2$ is set to zero are called corner points. To be an acceptable piecewise smooth solution, $|G_{opt}(e^{j\omega})|^2$ must satisfy the so called

Weistrass-Erdmann conditions at those frequencies. In our case, the Weistrass-Erdmann conditions reduce to the requirement that the integrand in (2.16) be a continuous function of ω at the corner frequencies ω_k. This requirement is indeed satisfied because the integrand is a continuous function of $|G(e^{j\omega})|^2$ which in turn is continuous in ω even at the corner points ω_k

Step 3. We would like now to prove that the magnitude response expression (2.13) is not only necessary but also sufficient for the optimality of the prefilter. We introduce the following notation by rewriting (2.15) as follows:

$$\int_{-\pi}^{\pi} \Big(f_1(\omega, y(\omega)) + f_2(\omega, y(\omega)) + f_3(\omega, y(\omega)) \Big) \frac{d\omega}{2\pi} \tag{2.49}$$

where

$$\begin{aligned} f_1(\omega, y(\omega)) &= \frac{\sigma_q^2 S_{xx}(e^{j\omega})}{S_{xx}(e^{j\omega}) y(\omega) + \sigma_q^2} \\ f_2(\omega, y(\omega)) &= \lambda S_{xx}(e^{j\omega}) y(\omega) \\ f_3(\omega, y(\omega)) &= \beta(\omega) y(\omega) \\ y(\omega) &= |G(e^{j\omega})|^2 \\ \sigma_q^2 &= c 2^{-2b} \end{aligned} \tag{2.50}$$

Before proceeding further, we can now summarize the following useful facts from [63]:

Fact1. The sum of a convex function with one or more convex functions is again convex.

Fact2. If $f(\omega, y(\omega))$ is convex on $[-\pi, \pi] \times \mathcal{D}$, then, $J[y(\omega)] = \int_{-\pi}^{\pi} f(\omega, y(\omega)) \frac{d\omega}{2\pi}$ is convex on $\mathcal{D}$. Hence, each $y(\omega) \in \mathcal{D}$ that satisfies the necessary conditions of step 2 minimizes $J[y(\omega)]$ on $\mathcal{D}$.

From the above two facts, it is then clear that to prove that the solution (2.13) is a minimizing curve, we simply need to prove the convexity of $f_i(\omega, y(\omega)) \ \forall \ i$. The convexity of $f_i(\omega, y(\omega))$ on $[-\pi, \pi] \times \mathcal{D}$ can be established by using ***anyone*** of the following two conditions:

1. The following inequality must be satisfied $\forall \ (\omega, y(\omega))$ and $\forall \ (\omega, y(\omega) + v(\omega)) \in [-\pi, \pi] \times \mathcal{D}$:

$$f_i(\omega, y(\omega) + v(\omega)) - f_i(\omega, y(\omega)) \geq \left(\frac{\partial}{\partial y(\omega)} f_i(\omega, y(\omega)) \right) \cdot v(\omega) \tag{2.51}$$

2. The matrix of second partial derivatives

$$\begin{bmatrix} f_{yy} & f_{yy'} \\ f_{yy'} & f_{y'y'} \end{bmatrix} \tag{2.52}$$

must be positive semidefinite on $[-\pi, \pi] \times \mathcal{D}$.

In the above two conditions, all the partial derivatives are assumed to be continuous on $[-\pi, \pi] \times \mathcal{D}$. The notation y' is used for the derivative of $y(\omega)$ with respect to ω. We use condition (2.51) to prove

the convexity of $f_2(\omega, y(\omega))$ and $\mathrm{f_3}(\omega, \mathrm{y}(\omega))$ and condition (2.52) to prove the convexity of $f_1(\omega, y(\omega))$.

Convexity of $\mathbf{f_2}(\omega, \mathbf{y}(\omega))$ and $\mathbf{f_3}(\omega, \mathbf{y}(\omega))$. Assume first that $f(\omega, y(\omega)) = f_2(\omega, y(\omega))$ in (2.51). It is then easy to check, in this case, that the right hand side of the equation is equal to the left hand side. In fact, both sides will be equal to $\lambda S_{xx}(e^{j\omega})v(\omega)$. Similarly, when $f(\omega, y(\omega)) = f_3(\omega, y(\omega))$, the right hand side of (2.51) is equal to the left hand side of the same equation. The two sides are, in turn, equal to $\beta(\omega)v(\omega)$. This establish the convexity of both $f_2(\omega, y(\omega))$ and $f_3(\omega, y(\omega))$.

Convexity of $\mathbf{f_1}(\omega, \mathbf{y}(\omega))$. When $f(\omega, y(\omega)) = f_1(\omega, y(\omega))$, then, we first observe that the matrix in (2.52) can be simplified to the following form:

$$\begin{bmatrix} f_{yy} & 0 \\ 0 & 0 \end{bmatrix} \tag{2.53}$$

For this matrix to be positive semidefinite, the principal minors should be non-negative. From (2.53), this is equivalent to proving that $f_{yy} \geq 0 \; \forall \; \omega$. Differentiating $f_1(\omega, y(\omega))$ twice with respect to $y(\omega)$, we obtain the following equation:

$$f_{yy} = \frac{\sigma_q^2 S_{xx}^3(e^{j\omega})}{\left(S_{xx}(e^{j\omega})y(\omega) + \sigma_q^2\right)^3} \tag{2.54}$$

Since all quantities in (2.54) are positive, then, the condition (2.51) is indeed satisfied and $f_1(\omega, y(\omega))$ is convex. Using the convexity of the above functions and facts one and two, we conclude that the solution (2.13) is a minimizing extremum

Appendix B.

Using the following set of equations:

$$\begin{aligned} e(n) &= x(n) - \hat{x}(n) && (2.55) \\ \hat{x}(n) &= z(n) \otimes h(n) = \sum_{k=0}^{\infty} \gamma^k z(n-k) && (2.56) \\ z(n) &= y(n) + q(n) && (2.57) \\ y(n) &= x(n) - \alpha x(n-1) \end{aligned}$$

we can easily verify by direct substitution that the error process at the output of the postfilter is given by:

$$e(n) = x(n) - \sum_{k=0}^{\infty} \gamma^k x(n-k)$$

$$+ \quad \alpha \sum_{k=0}^{\infty} \gamma^k x(n-k-1) - \sum_{k=0}^{\infty} \gamma^k q(n-k)$$

$$= \quad (\frac{\alpha}{\gamma} - 1) \sum_{k=1}^{\infty} \gamma^k x(n-k) - \sum_{k=0}^{\infty} \gamma^k q(n-k) \tag{2.58}$$

The mean square error expression is defined to be $\mathcal{E} \triangleq E\{e^2(n)\}$. This, in turn can be written as:

$$\mathcal{E} = \quad E\Bigg\{ \Big((\tfrac{\alpha}{\gamma} - 1) \textstyle\sum_{k=1}^{\infty} \gamma^k x(n-k) - \sum_{k=0}^{\infty} \gamma^k q(n-k) \Big) \\ \cdot \Big((\tfrac{\alpha}{\gamma} - 1) \textstyle\sum_{l=1}^{\infty} \gamma^l x(n-l) - \sum_{l=0}^{\infty} \gamma^l q(n-l) \Big) \Bigg\} \tag{2.59}$$

This last equation can be simplified using the following assumptions about the noise process $q(n)$:

1. *White noise assumption.* $E\{q(n) \cdot q(n-k)\} = \sigma_q^2 \delta(n-k)$.
2. *Variance constraint assumption.* The noise variance σ_q^2 is equal to $c2^{-2b}\sigma_y^2$ where σ_y^2 is the variance of the quantizer input. Hence, $\sigma_q^2 = c2^{-2b}(R_{xx}(0)(1 + \alpha^2) - 2\alpha R_{xx}(1))$.
3. *Uncorrelatedness with $x(n)$.* The sequence $x(n)$ and $q(n)$ are assumed to be uncorrelated. Hence,

$$E\{x(n) \cdot q(n-k)\} = E\{q(n) \cdot x(n-k)\} = 0 \quad \forall\, k$$

Based on the above assumptions, equation (2.59) can be therefore simplified.The result gives the following expression for the mean square error:

$$\mathcal{E} \quad = \quad c2^{-2b} \Big((1+\alpha^2)\, R_{xx}(0) - 2\alpha R_{xx}(1) \Big) \sum_{k=0}^{\infty} \gamma^{2k}$$

$$+ \quad (\frac{\alpha}{\gamma} - 1)^2 \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \gamma^{l+k} R_{xx}(k-l) \tag{2.60}$$

This last expression consists of two terms and can be further simplified. The first term of (2.60) can be rewritten as follows:

$$c2^{-2b} \Big((1+\alpha^2)\, R_{xx}(0) - 2\alpha R_{xx}(1) \Big) \frac{1}{(1-\gamma^2)} \tag{2.61}$$

The second term can be divided into two subterms, one for $k = l$ and the other for $k \neq l$ to obtain:

$$(\frac{\alpha}{\gamma} - 1)^2 \sum_{\substack{k=1 \\ k=l}}^{\infty} \gamma^{2l} R_{xx}(0) + (\frac{\alpha}{\gamma} - 1)^2 \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \gamma^{l+k} R_{xx}(k-l) \tag{2.62}$$

which in turn can be rewritten as :

$$\Big(\frac{\alpha - \gamma}{\gamma} \Big)^2 \frac{\gamma^2}{1-\gamma^2} R_{xx}(0) + 2 \Big(\frac{\alpha - \gamma}{\gamma} \Big)^2 \frac{\gamma^2}{1-\gamma^2} \sum_{m=1}^{\infty} \gamma^m R_{xx}(m) \tag{2.63}$$

Adding (2.61) and (2.63) we obtain the mean square error expression of Theorem 5.

Chapter 3

Oversampling PCM Techniques and Optimum Noise Shapers for Quantizing a Class of Nonbandlimited Signals

3.1 Introduction

It is well known that if a continuous time signal $x(t)$ is σ-bandlimited, then, it can be recovered uniquely from its samples $x(nT)$ as long as $T \leq \pi/\sigma$. Extensions of the lowpass sampling theorem such as the bandpass, non uniform and derivative sampling theorems can be found in [38]. Recently, Walter [87] showed that, under some conditions, a class of non bandlimited continuous-time signals can be reconstructed from uniformly spaced samples even though aliasing occurs. Vaidyanathan and Phoong [77, 78] developed the discrete time version of Walter's result from a multirate digital filtering perspective. In specific, they considered the class of non bandlimited signals that can be modeled as the output of a single interpolation filter (single band model) as in Fig. 3.1 or as the output of the more general multiband model of Fig. 3.2.

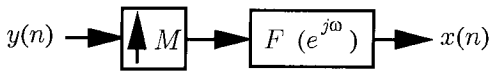

Fig. 3.1: The single band model

The filter $F(e^{j\omega})$ in Fig. 3.1 and the filters $F_k(e^{j\omega})$, $k = 0, 1, \ldots, L-1$, in Fig. 3.2 are usually a set of $L < M$ synthesis filters in an M-channel maximally decimated perfect reconstruction filter bank, although this is not a necessary condition. To give the reader a flavor of the major ideas, consider for the moment the single band model of Fig. 3.1. The discrete time signal $x(n)$ is the output of an

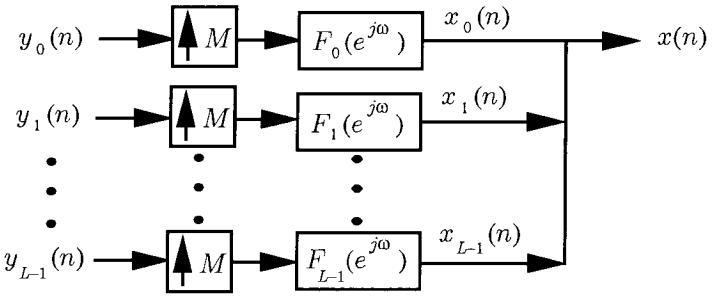

Fig. 3.2: The multiband model

interpolation filter $F(e^{j\omega})$. Even though this signal is not in general bandlimited, it is natural to expect that it can be recovered from its decimated version $x(Mn)$. To see this, assume that $x(n)$ is modeled as in Fig. 3.1 and consider $x(Mn)$, the M-fold decimated version of $x(n)$. If $F(e^{j\omega})$ is a Nyquist(M) filter [73], then, $x(Mn)$ is equal to $y(n)$ and we have the relation $x(n) = \sum_k x(kM)f(n-kM)$. In other words, $x(n)$ is completely defined by the samples $x(Mn)$ even though the filter $F(e^{j\omega})$ is not necessarily ideal. In [78], the authors consider the case where $F(e^{j\omega})$ is not necessarily a Nyquist(M) filter and show how similar reconstruction can be done. They also consider the stability of the reconstruction process. It turns out that if one of the polyphase components of $F(e^{j\omega})$ is free from unit circle zeros, then, stability of reconstruction is guaranteed. Furthermore, even if all the polyphase components of $F(e^{j\omega})$ have unit circle zeros, stable reconstruction can still be achieved by using non uniform decimation. In this case, a sufficient condition for stable reconstruction is that $F(e^{j\omega})$ (assumed FIR) has two polyphase components with no multiple zeros, i.e., each polyphase component has distinct zeros and they do not share any common zero.

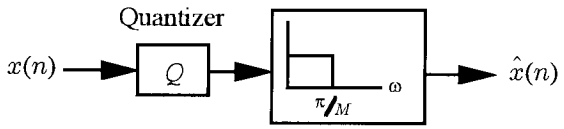

Fig. 3.3: Schematic of the oversampling PCM technique

In this chapter, we consider the efficient quantization of this class of non band-limited signals that can be modeled as in Fig. 3.1 or more generally as in Fig. 3.2. To motivate such a study, consider the schematic shown in Fig. 3.3 where the box labeled Q is a simple uniform roundoff (PCM) quantizer. After going through the quantizer, the signal $x(n)$ is now contaminated by an additive noise component $e(n)$. Assuming that the signal $x(n)$ is bandlimited or equivalently oversampled (since a bandlimited signal can be further downsampled), we can low pass filter the quantized signal $x(n) + e(n)$. The ideal low pass filter on the right removes the noise in the stopband but does not change the signal component. In terms of signal and noise power, the signal power remains unchanged whereas the noise power decreases proportionally to the oversampling ratio, usually expressed in the form 2^r. It can be

shown that for every doubling of the oversampling ratio, i.e., for every unit increment in r, the signal to noise ratio (SNR) improves by about 3 db or equivalently, the quantizer resolution improves by one half bit (see for example [6]). After low-pass filtering, the quantized signal can be downsampled to the Nyquist rate without affecting the signal to noise ratio. The idea is therefore to exploit the oversampled nature of the signal $x(n)$ to trade off quantizer complexity for higher resolution. This technique is usually called oversampled PCM conversion.

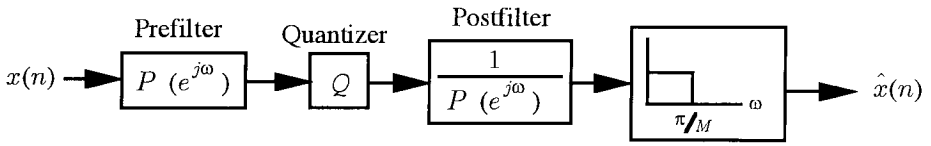

Fig. 3.4: The quantization scheme of Fig. 3.3 with noise shapers

Consider now the system of Fig. 3.4 where $P(e^{j\omega})$ is a linear time-invariant (LTI) filter. The input signal $x(n)$ is still assumed to be oversampled (bandlimited). In addition to the benefits described above, it can be shown that a clever choice of the filter $P(e^{j\omega})$ in Fig. 3.4 produces a further decrease in the noise power. The filter pair $P(e^{j\omega})$ and $1/P(e^{j\omega})$ affects the noise component $e(n)$ but not the input signal $x(n)$. Like a sigma-delta quantizer, the system of Fig. 3.4 introduces *noise shaping* in the signal band to allow higher resolution quantization of bandlimited signals.

With these ideas in mind, observe now the output $x(n)$ of Fig. 3.1. Even though $x(n)$ is not bandlimited, it can be reconstructed from its downsampled version as explained above. In this sense, *it can be considered as an oversampled signal.* The question then arises : Can we obtain advantages similar to those of the above schemes for a non bandlimited signal satisfying the model of Fig. 3.1 and more generally of Fig. 3.2 ? Furthermore, for a fixed filter $F(e^{j\omega})$ (or a set of filters $F_k(e^{j\omega})$, $k = 0, 1, \ldots, L-1,$), what is the best filter prefilter $P(e^{j\omega})$ for minimizing the noise power at the output ? Do we obtain any substantial gain by using a more general postfilter $V(e^{j\omega})$ instead of $\frac{1}{P(e^{j\omega})}$? This is a sample of the type of questions we answer in this chapter. Indeed, we will show that, by replacing the ideal low pass filter with the correct multirate reconstruction system, we can reap the same quantization advantages as in the bandlimited case. As a simple example, consider the scheme of Fig. 3.5 where the finite order filter $F(e^{j\omega})$ is such that its magnitude squared response $|F(e^{j\omega})|^2$ is Nyquist(M), that is, $(|F(e^{j\omega})|^2) \downarrow_M = 1$ (we will motivate such an assumption later in the chapter). With this assumption, it can be shown that the signal $\hat{x}(n)$ in Fig. 3.5 is equal to $x(n)$ in the absence of the quantizer and that the entire scheme of Fig. 3.5 behaves similarly to Fig. 3.3, except that the low pass filtering is now *multirate* and *non ideal.* Thus, generally speaking, if a non bandlimited signal can be reconstructed from its samples $x(Mn)$ because it satisfies a model like Fig. 3.1, then, a low precision quantizer should allow us to produce a high precision version $\hat{x}(n)$.

To bring the analogy closer to the scheme of Fig. 3.4, we should introduce noise shaping. This can

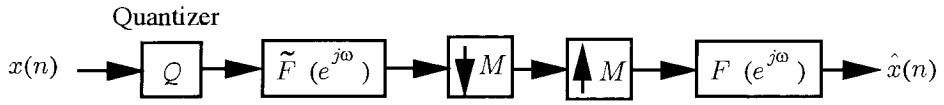

Fig. 3.5: Multirate quantization scheme for the single band case

be done by using a pre- and post filter before and after the quantizer respectively as shown in Fig. 3.6.

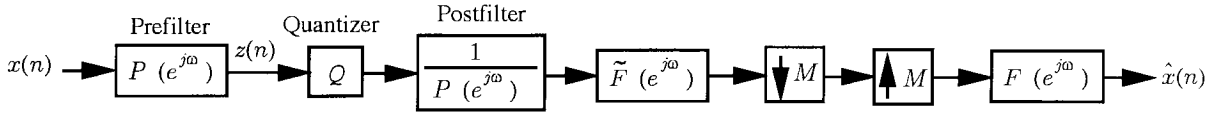

Fig. 3.6: Noise shaping by LTI pre- and post filters for the single band case where the postfilter is assumed to be the inverse of the prefilter

The prefilter $P(e^{j\omega})$ is traditionally an integrating low pass filter. The post filter $1/P(e^{j\omega})$ shapes the noise spectrum in order to further decrease the noise variance. In this chapter, we will derive closed form expressions for the optimal choice of $P(e^{j\omega})$ and the minimum average mean square error obtained from such a scheme. Several extensions to the above noise shaping idea are then introduced. For example, we relax the requirement that the postfilter is the inverse of the prefilter and assume a more general postfilter $V(e^{j\omega})$. Closed form expressions for the optimum filters in this case and the average minimum mean square error are also derived. We would like to warn the reader at this point that no optimization of finite order filters is performed in this chapter. The emphasis is on finding an expression for the theoretically optimum filters (without order constraint) to get an upper bound on the achievable gain with practical inexpensive filters.

The quantization advantage offered by Fig. 3.5 and Fig. 3.6 can be useful, for example, in the following realistic engineering scenario. Suppose $x(n)$ is generated at a point where we cannot afford very complex signal processing (e.g., in deep space) and needs to be transmitted to a distant place (e.g., earth station). If we have the knowledge that $x(n)$ admits a satisfactory model like Fig. 3.1, we can compress it using a very simple low pass filter $P(e^{j\omega})$ with one or two multipliers and then quantize the output before transmission. The post filter $1/P(e^{j\omega})$ and the expensive multirate filter are at the receiver end, where the complexity is acceptable.

Assume now that the main aim is to obtain a reduction in the bit rate (number of bits per second) rather than accuracy (number of bits per sample). If we are allowed to perform discrete time filtering (of arbitrary complexity), we will see that the best approach would be as in Fig. 3.7.

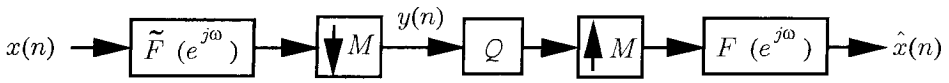

Fig. 3.7: Quantizing the lower rate signal $y(n)$ (single band case)

In this set up, we first generate the driver signal $y(n)$ and then quantize it. The signal $\hat{x}(n)$, which

is equal to $x(n)$ in absence of quantization, is then generated. The lower rate signal $y(n)$ in Fig. 3.7 can be regarded as the principal component signal in an orthonormal subband coder. We will see throughout this chapter that, by choosing this type of quantization system, we can obtain a large reduction in the bit rate and/or the quantization accuracy depending on the particular signal model.

Summarizing, the main issue in this chapter is how to take advantage of the signal model (Fig. 3.1 or Fig. 3.2) in preparing a quantized or compressed version of $x(n)$. Our study is motivated by similar concepts that arises in A/D conversion applications. We find that the choice of a particular scheme depends on how much processing we are allowed to do before quantization. If processing is allowed, we first generate $y(n)$ by filtering and decimation and then quantize it. Otherwise, we quantize $x(n)$ directly and then filter the quantized signal with the appropriate multirate scheme. Noise shaping can be also introduced to obtain better resolution. In any case, an improvement in accuracy and/or bit rate due to the signal model is always achieved.

3.1.1 Main results and outline of the chapter

1. In section 3.2, definitions and well established facts of various multirate and statistical signal processing concepts used in this chapter are reviewed.

2. In section 3.3, new results that describe the statistical behavior of signals as they pass through multirate interconnections are presented. These results will then be used to derive the theorems of interest of the chapter. The filter and quantizer assumptions are stated in section 3.4.

3. In section 3.5, we give several results on the quantization of the non bandlimited signal $x(n)$ modeled as in Fig. 3.1. The signal $x(n)$ is first quantized to an average of b bits per sample and then filtered by the multirate interconnection in Fig. 3.5. We show that the multirate system does not affect the signal component but reduces the noise variance by a factor of M. This amounts to the same quantitative advantage obtained from the oversampling PCM technique (0.5 bit reduction per doubling of the oversampling ratio).

4. In section 3.6, the lower rate signal $y(n)$ is quantized instead of $x(n)$. By quantizing $y(n)$ to b bits per sample, the quantization bit rate (number of bits per second) is decreased by a factor of M but noise reduction due to multirate filtering is now not possible.

5. In section 3.7, noise shaping is introduced in order to obtain better accuracy. First, we consider the use of pre- and post linear time invariant filters $P(e^{j\omega})$ and $\frac{1}{P(e^{j\omega})}$ as in Fig. 3.6 together with a fixed time invariant quantizer $\mathcal{Q}$. For this case, the optimum filter $P_{opt}(e^{j\omega})$ that minimizes the quantization noise variance in the reconstructed output $\hat{x}(n)$ is derived and a closed form expression for the average minimum mean square error is obtained. We then consider the more general pre- and postfilters $P(e^{j\omega})$ and $V(e^{j\omega})$ as in Fig. 3.15. Closed form expressions for the optimum filters and the average minimum mean square error are also found for this case.

6. In section 3.8, we replace the linear time invariant filter $P(e^{j\omega})$ with a more general linear period-

ically time varying filter of period M. This is motivated by the Cyclo-widesense stationarity of $x(n)$. Since the problem of finding the optimum general $(LPTV)_M$ filter (equivalently biorthogonal filter bank) is analytically difficult to track, optimal solutions are given for two special cases of $(LPTV)_M$ filters. The first solution is for the set of M filters $V_k(e^{j\omega})$ shown in Fig. 3.19. The filters $V_k(e^{j\omega})$ and $\frac{1}{V_k(e^{j\omega})}$ act as pre- and post filters for the kth subband quantizer. The second solution is for the case of an orthonormal filter bank or equivalently for a lossless $(LPTV)_M$ filter. The scheme is shown in Fig. 3.23 for the single band case.

7. All the results mentioned above are also generalized for the multiband case. Furthermore, examples are provided whenever necessary for illustrative purposes.

3.2 Chapter specific definitions

1. Blocking a signal. Given a scalar signal $x(n)$, we define its M-fold blocked version $\mathbf{x}(n)$ by

$$\mathbf{x}(n) = \left(x(nM) \quad x(nM-1) \quad \ldots \quad x(nM-M+1) \right)^T \tag{3.1}$$

Equivalently, the scalar sequence $x(n)$ is called the unblocked version of the vector process $\mathbf{x}(n)$. The blocking and unblocking operations are shown in Fig. 3.8. The elements of the blocked version $\mathbf{x}(n)$ are the polyphase components of $x(n)$.

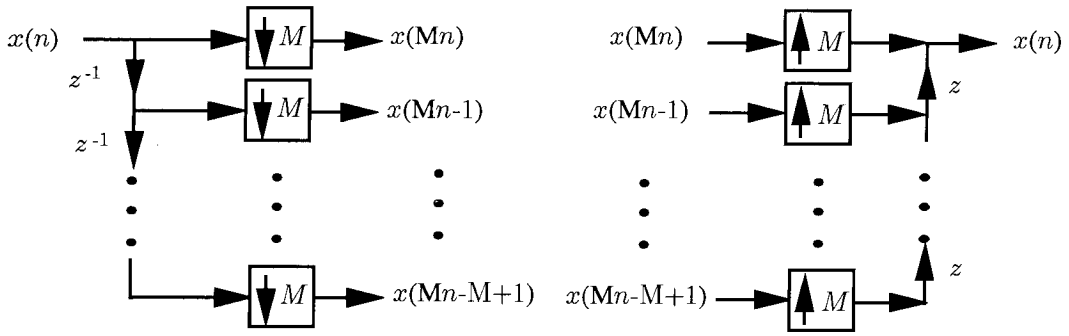

Fig. 3.8: M-fold blocking of a signal and unblocking of an $M \times 1$ vector signal

2. Cyclo-widesense stationary process. A stochastic process $x(n)$ is said to be cyclo-widesense stationary with period M, abbreviated as $(CWSS)_M$, if the M-fold blocked version $\mathbf{x}(n)$ is WSS. Alternatively [25, 58], a process $x(n)$ is $(CWSS)_M$ if the mean and autocorrelation functions of $x(n)$ are periodic with period M, i.e.,

$$E[x(n)] = E[x(n+kM)] \ \forall \ n,k \text{ and } R_{xx}(n,k) = R_{xx}(n+M,k) \ \forall \ n,k. \tag{3.2}$$

where $R_{xx}(n,k) \triangleq E[x(n)x^*(n-k)]$ is the autocorrelation function of $x(n)$.

3. The coding gain of a system. Assume that we quantize $x(n)$ directly with b bits as shown in Fig. 3.9. We denote the corresponding mean square error (m.s.e.) by $\mathcal{E}_{direct}$. We then use the optimum pre and post filters (in the mean square sense) around the quantizer. With the rate of the quantizer fixed to the same value b, we denote the minimum m.s.e. in this case by $\mathcal{E}_{min}$. The ratio $\mathcal{E}_{direct}/\mathcal{E}_{min}$ is called the coding gain of the new system and, as the name suggests, is a measure of the benefits provided by the pre/post filtering operation.

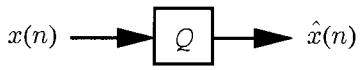

Fig. 3.9: Direct quantization of $x(n)$

3.3 Preliminary Results

Result 1 *Consider any L synthesis filters ($L < M$) of an M-channel orthonormal filter bank as shown in Fig. 3.2. Assume that the L inputs $y_k(n)$ to the synthesis filters $F_k(e^{j\omega})$ are zero mean jointly WSS processes, not necessarily uncorrelated. Then, the statistical correlation (averaged over M samples) between the interpolated subband signal $x_i(n)$ and the M-sample shifted process $x_j(n - Mm)$ is zero, for all values of $i \neq j$ and m, that is:*

$$\frac{1}{M}\sum_{k=0}^{M-1} E[x_i(n-k)x_j{}^*(n-k-Mm)] = 0, \ \forall\ n,m \text{ and } \forall\ i,j \in [0, L-1] \tag{3.3}$$

Proof. The proof can be found in appendix A. ∎

As a consequence, the average variance of the $(CWSS)_M$ output process $x(n)$ of Fig. 3.1, where the filters $F_k(e^{j\omega})$ are any L synthesis filters of an M-channel orthonormal filter bank, is:

$$\sigma_x^2 = \frac{1}{M}\sum_{k=0}^{L-1} \sigma_{y_k}^2 \tag{3.4}$$

This can be seen by substituting $x(n)$ in the formula $\sigma_x^2 = \frac{1}{M}\sum_{n=0}^{M-1} E[|x(n)|^2]$ and using Result 1 for the special case of $m = 0$ and $n = M - 1$. If the L inputs to the synthesis filters $F_k(e^{j\omega})$ are zero mean uncorrelated WSS processes, the previous result holds without the orthonormality requirement on the filters $F_k(e^{j\omega})$, $k = 0, 1, \ldots, L-1$.

Result 2 *Consider the multirate interconnection of Fig. 3.1 where the input $y(n)$ is zero mean WSS random process. If $F(e^{j\omega})$ is a filter (not necessarily ideal) with a Nyquist(M) magnitude squared response, then*

$$\sigma_x^2 = \frac{1}{M}\sigma_y^2 \tag{3.5}$$

where σ_x^2 is the average variance of the $(CWSS)_M$ output $x(n)$.

Proof. While this is a special case of the above with $L = 1$, the following proof is direct and more instructive. With $F(e^{j\omega})$ expressed in terms of its polyphase components $R_k(e^{j\omega})$, Fig. 3.1 can be redrawn as in Fig. 3.10. The signal $x(n)$ is the interleaved version of the WSS outputs of $R_k(e^{j\omega})$. So, it has zero mean and a variance which is periodic with period M. The average variance is given by :

$$\sigma_x^2 = \frac{1}{M}\sum_{k=0}^{M-1}\sigma^2{}_{x_k} = \frac{1}{M}\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})\sum_{k=0}^{M-1}|R_k(e^{j\omega})|^2\frac{d\omega}{2\pi} \tag{3.6}$$

The Nyquist property of $|F(e^{j\omega})|^2$ implies in particular that $\sum_{k=0}^{M-1}|R_k(e^{j\omega})|^2 = 1$ [73, page 159]. The preceding equation therefore simplifies to $\sigma_x^2 = \frac{1}{M}\int_{-\pi}^{\pi} S_{yy}(e^{j\omega})\frac{d\omega}{2\pi} = \frac{1}{M}\sigma_y^2$ ∎

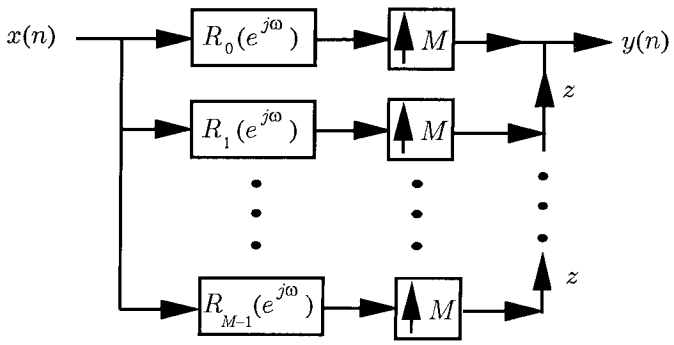

Fig. 3.10: The equivalent polyphase representation of Fig. 3.1

3.4 Filter and Quantizer Assumptions

Filter assumptions. The filters $F(e^{j\omega})$ of and $F_k(e^{j\omega})$, $k = 0, 1, \ldots, L-1$, of Fig. 3.2 are assumed to be the synthesis filters of any L channels of an M-channel maximally decimated *orthonormal* filter bank. Although not necessary for developing the results of this chapter, we will additionally choose the L channels of the M-channel maximally decimated *orthonormal* filter bank to be the most dominant ones in terms of subband energy. The model filters are therefore the so-called optimum energy compaction filters. This last constraint is motivated by the fairly recent result that this particular choice of filters minimizes the average mean square reconstruction error between the original signal $x(n)$ and its approximation $\hat{x}(n)$ [65, 68]. We would like however to emphasize that, unlike in previous work, the filters in this chapter are assumed to be of finite order. Working with ideal brick wall filters will obviously contradict the non-bandlimited assumption.

Quantizer assumption. As is usually the convention in this thesis, the box labeled $\mathcal{Q}$ represents

a scalar uniform (PCM) quantizer and is modeled as an additive zero mean white noise source $q(n)$. Because the model filters are not ideal, the input $x(n)$ is a zero mean $(CWSS)_M$ process. Since the input to the quantizer $x(n)$ is a $(CWSS)_M$ process, its variance $\sigma_x^2(n)$ is a periodic function of n with period M. Define σ_x^2 to be the average variance of $x(n)$, i.e., $\sigma_x^2 = \frac{1}{M}\sum_{n=0}^{M-1}\sigma_x^2(n)$. Then, choose the fixed step size Δ in the uniform quantizer such that the quantization noise variance σ_q^2 is directly proportional to the average variance of the quantizer input $x(n)$, that is

$$\sigma_q^2 = c2^{-2b}\sigma_x^2 \tag{3.7}$$

where σ_q^2 is the quantization noise variance, c is a constant that depends on the statistical distribution of $x(n)$ and the overflow probability, and σ_x^2 is the average variance of the quantizer input. The above relation is justified for a PCM quantizer using 3 (or more) bits per sample (see chapter 4 in [37]). If the input to $\mathcal{Q}$ is wide-sense stationary, the above relation holds with σ_x^2 now denoting the actual variance of the WSS process.

3.5 Increasing the quantizer resolution by multirate filtering

Consider the set up shown in Fig. 3.5 for the single band model and in Fig. 3.11 for the multiband case. In the absence of quantization, the two schemes are perfect reconstruction systems. In the presence of the quantizer, the output $\hat{x}(n)$ in Fig. 3.5 and Fig. 3.11 is equal to the original sequence $x(n)$ plus an error signal $e(n)$ due to quantization. The following result shows that, by using the above schemes, a significant reduction in the average mean square error $\mathcal{E} \triangleq \frac{1}{M}\sum_{n=0}^{M-1} E\{e(n)\}^2$ can be obtained in comparison with the direct quantization of $x(n)$ shown in Fig. 3.9.

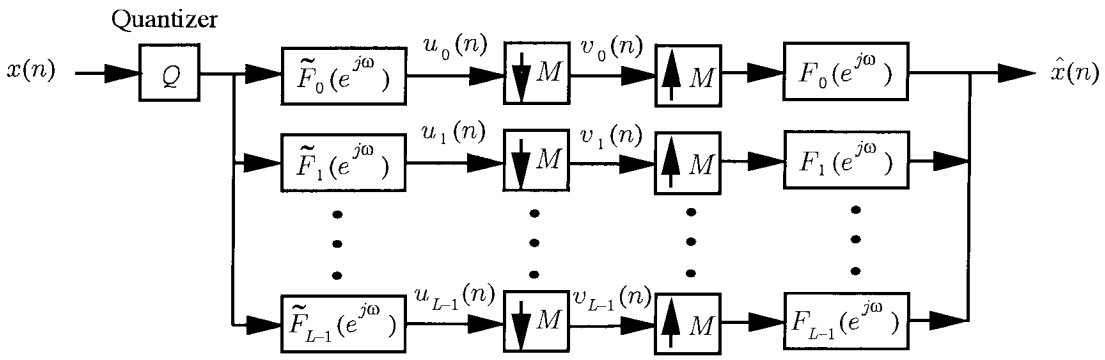

Fig. 3.11: Multirate quantization scheme for the multiband model

Theorem 8 *Consider the scheme of Fig. 3.11 where the L filters $F_k(e^{j\omega})$ are assumed to be any L channels of an M-channel critically sampled orthonormal filter bank. Under the above quantization*

noise assumption, the average mean square error (m.s.e.) $\mathcal{E}$ *is equal to* $\frac{L}{M}\sigma_q^2$.

Proof. Because the system is a perfect reconstruction one, the average error at the output is due only to the quantization noise. The quantization noise $q(n)$ is white and propagates through the L channels of Fig. 3.11. For the *kth* channel, the variance of $u_k(n)$ due to the noise passage through $F_k(e^{j\omega})$ is given by:

$$\sigma_{u_k}^2 = \sigma_q^2 \int_{-\pi}^{\pi} |F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} = \sigma_q^2 \tag{3.8}$$

The second equality follows because the filters have unit energy. The downsampling operation does not alter the variance of a signal. We therefore obtain $\sigma_{v_k}^2 = \sigma_{u_k}^2 = \sigma_q^2$ for all k. Using Result 3.1 of section 3.3, we can write

$$\mathcal{E} = \frac{1}{M} \sum_{k=0}^{L-1} \sigma_{v_k}^2 = \frac{L}{M}\sigma_q^2 \tag{3.9}$$

■

For the scheme of Fig. 3.5, the average m.s.e. $\mathcal{E}$ can be obtained directly by setting $L = 1$ and is therefore equal to $\frac{1}{M}\sigma_q^2$. The quantization noise variance σ_q^2 obtained by directly quantizing $x(n)$ as shown in Fig. 3.9 is now reduced by the oversampling factor M. The signal variance σ_x^2 on the other hand did not change. By expressing the interpolator M in the form 2^r, we can immediately see that we can get the same quantitative advantage of the oversampling PCM technique, namely, an increase in SNR by 3 db for every doubling of the oversampling factor. For example, for the single band case of Fig. 3.5, if $M = 2$, then, we get an SNR increase of 3 db whereas if $M = 4$, the SNR increment is by 6 db. Some important remarks are in order at this point :

1. In the oversampling PCM technique, the quantized bandlimited signal is typically downsampled after the low pass filter [6]. The SNR before and after the downsampler is the same and the increase in SNR is only due to a reduction in noise power. Similarly, the SNR before and after the interpolation filter in Fig. 3.5 does not change. However, the reason for the SNR increase before the interpolation filter is different from the one after the interpolation filter. At the input of the interpolation filter, the signal variance increases proportionally to M since $\sigma_y^2 = M\sigma_x^2$ and the noise power remains fixed. At the output of the interpolation filter, the signal variance doesn't change but the noise power decreases in proportion to M. In both cases, this amounts to the same SNR improvement. This last technical difference arises because our study assumes a statistical framework rather than a deterministic one (typical in A/D conversion applications) and because of our quantizer assumptions.

2. *Intuitive explanation of Theorem* 8. The signal $x(n)$, modeled either as in Fig. 3.1 or Fig. 3.2, is oversampled and therefore, contains redundant information in the form of an excess of samples. It is by quantizing these extra samples that we obtain the reduction in the quantization noise variance (equivalently in the mean square error). We are therefore effectively quantizing with a higher number of bits per sample. This trade off, between the quantization noise variance (effective quantizer resolution)

and the sampling rate is the underlying principle of oversampled A/D converters.

3. *The role of the factor L.* The parameter L, defined to be the number of channels in the multiband case, alternates between two extremes : $L = 1$ and $L = M$. When $L = 1$, we get the best SNR improvement at the expense of a more narrow class of inputs $x(n)$. When $L = M$, it is clear from (3.9) that no noise variance reduction is achieved since the class of signals is now unrestricted. We can also see this by noticing that the multirate interconnection in Fig. 3.11 becomes a perfect reconstruction filter bank that is signal independent. The parameter L therefore determines the tradeoff between the generality of the class of signals $x(n)$ and the reduction in quantization noise variance.

4. *A cascade of the scheme of Fig. 3.5 does not provide any further gain.* Using the scheme of Fig. 3.5, we obtained a reduction in noise by a factor M. If we use a cascade of the same filtering scheme as in Fig. 3.12, no further noise reduction is obtainable. Using the polyphase identity [73] and keeping in mind that $|F(e^{j\omega})|^2$ is Nyquist(M), the product filter $F(e^{j\omega})\tilde{F}(e^{j\omega})$ together with the expander and decimator reduces to an identity system. Fig. 3.6 therefore simplifies to Fig. 3.5 and the average m.s.e. is the same.

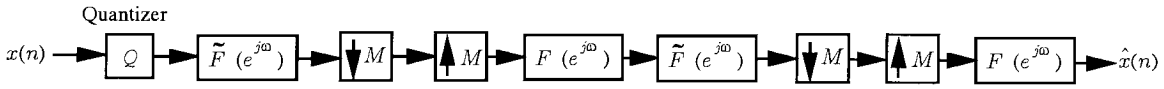

Fig. 3.12: A cascade of two multirate interconnections for the single band case

3.6 Quantizing at lower rate

A consequence of the previous results and discussion is then the natural question: what if the discrete time filtering of the oversampled signal is not a major burden ? If we know that $x(n)$ can be modeled quite accurately by the filter $F(e^{j\omega})$ of Fig. 3.1 or the filters $F_k(e^{j\omega})$, $k = 0, 1, \ldots, L-1$, of Fig. 3.2, we filter and downsample $x(n)$ accordingly to obtain either $y(n)$ or $y_k(n)$, $k = 0, 1, \ldots, L-1$. The quantization systems for the two models are shown in Fig. 3.7 and Fig. 3.13 respectively. In principle, we can then quantize the decimated signal $y(n)$ in Fig. 3.7 with $\hat{b} = Mb$ bits per sample or the signals $y_k(n)$, $k = 0, 1, \ldots, L-1$, of Fig. 3.13 with an average number of bits per sample $\hat{b} = \frac{M}{L}b$ bits. This situation is equivalent to fixing the bit rate (number of bits per second) to be equal to b in order to trade quantization resolution with sampling rate. Moreover, for the multiband case, we can allocate b_k bits to the driving signals $y_k(n)$ in an "appropriate" manner. At this point however, we will assume that the goal is to actually obtain a reduction in the bit rate. To achieve this, we let $\hat{b} = b$ for both cases and analyze the quantization systems of Fig. 3.7 and Fig. 3.13 under this condition. By fixing the number of bits per sample and decreasing the signal rate, the bit rate will automatically decrease by M/L. However, since the quantizer resolution did not increase, the quantization noise variance should not differ from the direct quantization case of Fig. 3.9. This last statement is verified formally

in the next theorems.

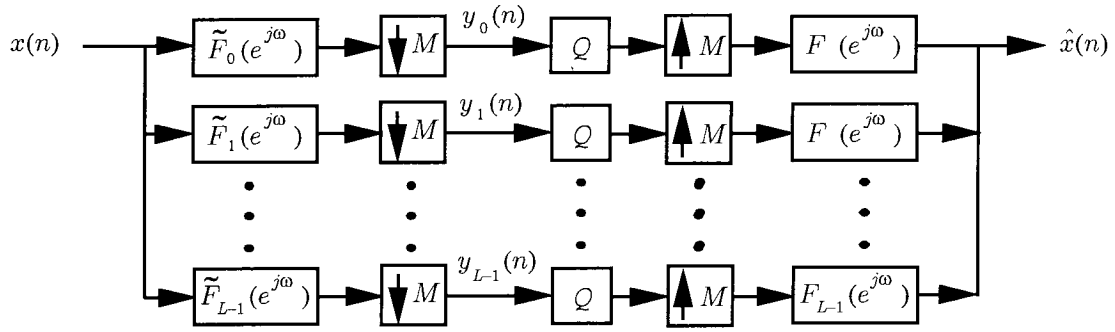

Fig. 3.13: Quantizing the lower rate signals $y_k(n)$ (multiband case)

Theorem 9 *Consider the scheme of Fig. 3.7. Using a fixed number of bits per sample b to quantize $y(n)$, the average mean square error $\mathcal{E}$ is equal to σ_q^2, where σ_q^2 is the noise variance obtained from directly quantizing $x(n)$ using b bits per sample.*

Proof. Let σ_q^2 be the noise variance of Fig. 3.9 and $\mathcal{E}$ be the average mean square error of Fig. 3.7. Using (3.7), we can write $\sigma_q^2 = c2^{-2b}\sigma_x^2$. But, by Result 2 of section 3.3,

$$\mathcal{E} = \frac{1}{M}c2^{-2b}\sigma_y^2 = \frac{1}{M}c2^{-2b}M\sigma_x^2 = \sigma_q^2$$

where σ_x^2 is the average variance of $x(n)$. ■

The theorem indicates that, for the single band model and under a fixed number of quantizer bits b, quantizing the lower rate signal $y(n)$ is as accurate as directly quantizing $x(n)$. This is expected and is in fact consistent with the observation of section 3.5 regarding the tradeoff between the average m.s.e. due to quantization and the rate of the signal. The next theorem for the multiband case gives a similar conclusion.

Theorem 10 *Consider the scheme of Fig. 3.13. Assume that we quantize $y_k(n)$ at b bits per sample for all k. Then, the average mean square error $\mathcal{E}$ is equal to σ_q^2, where σ_q^2 is the noise variance obtained from directly quantizing $x(n)$ using b bits per sample.*

Proof. The average mean square error at the output of Fig. 3.13 is equal to

$$\mathcal{E} = \frac{1}{M}\sum_{k=0}^{L-1}\sigma_{q_k}^2 = \frac{1}{M}c2^{-2b}\sum_{k=0}^{L-1}\sigma_{y_k}^2 \tag{3.10}$$

where b denotes the *fixed* number of bits allocated to the kth channel quantizer. The noise variance σ_q^2 in Fig. 3.9 is equal to $c2^{-2b}\sigma_x^2$, which in turn is equal to (3.10). ■

3.7 Noise shaping by Time-Invariant pre- and post filters

Following the philosophy of sigma-delta modulators, we now perform noise shaping to achieve a further reduction in the average mean square error. To accomplish this, we propose using LTI pre- and post filters around the PCM quantizer as shown in Fig. 3.6 for the single band model and in Fig. 3.14 for the multiband model. We first use a prefilter $P(e^{j\omega})$ and assume that the postfilter is its inverse. We then relax this condition and assume a more general postfilter $V(e^{j\omega})$. The goal is to optimize these filters such that the average m.s.e. at the output of either quantization system is minimized. The noise shaping filters to be optimized are not constrained to be rational functions (i.e., of finite order) and non causal solutions, for example, are accepted.

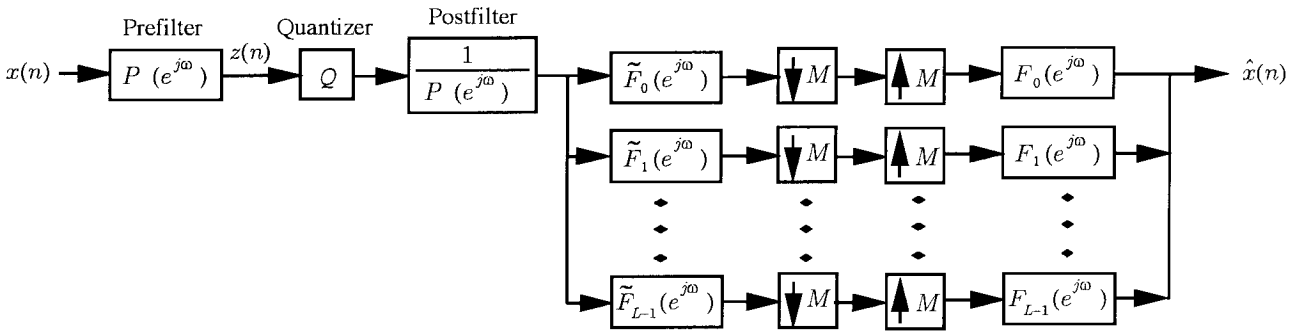

Fig. 3.14: Noise shaping by LTI pre- and post filters for the multiband case where the postfilter is assumed to be the inverse of the prefilter

Although our quantizer design assumptions are the same as before, the quantizer input is not anymore the $(CWSS)_M$ process $x(n)$, but a filtered version of it, which we denote by $z(n)$. Following (3.7), the noise variance in this case is given by $\sigma_q^2 = c2^{-2b}\sigma_z^2$ where σ_z^2 is the average variance of the process $z(n)$. We emphasize that $z(n)$ is a $(CWSS)_M$ process since the output of a linear time invariant filter driven by a $(CWSS)_M$ process is also $(CWSS)_M$ [58]. It is then possible to express σ_z^2 in terms of the prefilter $P(e^{j\omega})$ and the so called average power spectral density (see below) of the process $x(n)$, denoted by $\hat{S}_{xx}(e^{j\omega})$, as follows:

$$\sigma_z^2 = \frac{1}{M}\int_{-\pi}^{\pi} |P(e^{j\omega})|^2 \hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi} \tag{3.11}$$

The proof of (3.11) can be found in Appendix C. The average power spectral density is a familiar concept that arises when "stationarizing" a $(CWSS)_M$ process [23, 35, 22] and satisfies the well known properties of the power spectrum of a WSS process. It is defined to be the discrete time fourier transform of the time averaged autocorrelation function $\hat{R}_{xx}(k)$ given by $\frac{1}{M}\sum_{n=0}^{M-1} E[x(n)x^*(n-k)]$. Another interpretation of the average power spectral density which can be physically more appealing is based on the concept of phase randomization and is reviewed in Appendix B. Finally, if $x(n)$ is

modeled as in Fig. 3.1, it can be shown that :

$$\hat{S}_{xx}(e^{j\omega}) = \frac{1}{M} S_{yy}(e^{j\omega M}) |F(e^{j\omega})|^2 \tag{3.12}$$

whereas if the signal satisfies the multiband model of Fig. 3.2, the average power spectral density takes the following form :

$$\hat{S}_{xx}(e^{j\omega}) = \frac{1}{M} \mathbf{F}^{\dagger}(e^{j\omega}) \mathbf{S_y}(e^{j\omega M}) \mathbf{F}(e^{j\omega}) \tag{3.13}$$

where $\mathbf{F}(e^{j\omega}) = (\, F_0(e^{j\omega}) \quad F_1(e^{j\omega}) \quad \ldots \quad F_{L-1}(e^{j\omega}) \,)^T$ and $\mathbf{S_y}(e^{j\omega})$ is the $L \times L$ power spectral density matrix of the L WSS inputs $y_k(n)$. Note that, when the signals $y_k(n)$ are uncorrelated, equation Fig. 3.9 simplifies to $\frac{1}{M} \sum_{k=0}^{L-1} S_{y_k}(e^{j\omega M}) |F_k(e^{j\omega})|^2$. The proofs of (3.12) and (3.13) are given in appendix D. The expression Fig. 3.8 was derived previously in [58] for the special case where $F(e^{j\omega})$ is an anti-alias(M) filter. Furthermore, the authors prove that the output process $x(n)$ is WSS if and only if $F(e^{j\omega})$ is an anti-alias(M) filter. In summary, the statistical properties of the output $x(n)$ of Fig. 3.1 depend on $F(e^{j\omega})$. If the filter is an anti-alias(M) filter, then, $x(n)$ is WSS with a power spectral density $S_{xx}(e^{j\omega})$ in the same form as (3.12). Otherwise, $x(n)$ is a $(CWSS)_M$ process and in this case, the average power spectral density $\hat{S}_{xx}(e^{j\omega})$ is given by (3.12).

3.7.1 Case where the postfilter is the inverse of the prefilter

Theorem 11 *Consider the scheme of Fig. 3.14 under the same assumptions of section* 3.4. *The optimum prefilter $P(e^{j\omega})$ that minimizes the average mean square reconstruction error has the following magnitude squared response:*

$$|P_{opt}(e^{j\omega})|^2 = \frac{\sqrt{(\sum_{i=0}^{L-1} |F_i(e^{j\omega})|^2)}}{\sqrt{\hat{S}_{xx}(e^{j\omega})}} \tag{3.14}$$

Proof. We first observe that in the absence of quantization, the system of Fig. 3.14 is a perfect reconstruction system. Therefore, the average mean square reconstruction error σ_e^2 at the output is due only to the noise signal. Let $v_k(n)$ be the filtered noise component in the kth channel of the L-channel filter bank of Fig. 3.14. The variance of this signal $\sigma_{v_k}^2$ is equal to

$$\sigma_{v_k}^2 = \int_{-\pi}^{\pi} \sigma_q^2 \frac{|F_k(e^{j\omega})|^2}{|P(e^{j\omega})|^2} \frac{d\omega}{2\pi} \tag{3.15}$$

Since the downsampling operation does not change the variance of a process, we can write

$$\sigma_e^2 = \frac{1}{M} \sum_{k=0}^{L-1} \sigma_{v_k}^2 = \sigma_q^2 \frac{1}{M} \int_{-\pi}^{\pi} \frac{\sum_{k=0}^{L-1} |F_k(e^{j\omega})|^2}{|P(e^{j\omega})|^2} \frac{d\omega}{2\pi} \tag{3.16}$$

Using (3.7) and (3.11), we get

$$\sigma_e^2 = \frac{c2^{-2b}}{M} \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega}) |P(e^{j\omega})|^2 \frac{d\omega}{2\pi} \int_{-\pi}^{\pi} \frac{\sum_{k=0}^{L-1} |F_k(e^{j\omega})|^2}{|P(e^{j\omega})|^2} \frac{d\omega}{2\pi} \tag{3.17}$$

To find the optimum prefilter $P(e^{j\omega})$, we apply Cauchy-Schwartz inequality to (3.17) to obtain:

$$\sigma_e^2 \geq \frac{c2^{-2b}}{M} \left(\int_{-\pi}^{\pi} \sqrt{\hat{S}_{xx}(e^{j\omega}) \left(\sum_{i=0}^{L-1} |F_i(e^{j\omega})|^2 \right)} \frac{d\omega}{2\pi} \right)^2 \tag{3.18}$$

Since this lower bound is independent of $P(e^{j\omega})$, it is indeed the required minimum and is achieved iff

$$\sqrt{\hat{S}_{xx}(e^{j\omega})} |P(e^{j\omega})| = \frac{\sqrt{\left(\sum_{i=0}^{L-1} |F_i(e^{j\omega})|^2 \right)}}{|P(e^{j\omega})|} \tag{3.19}$$

which gives (3.14). ■

A number of observations should be made at this point. First, the optimum filter is not unique since the phase response is not specified. Second, the above derivation assumes that the input average spectrum $\hat{S}_{xx}(e^{j\omega}) \neq 0$ for all ω. The assumption is a reasonable one because $x(n)$ is assumed to be non bandlimited and therefore $\hat{S}_{xx}(e^{j\omega})$ cannot be identically zero on a segment of $[0, 2\pi)$. If $\hat{S}_{xx}(e^{j\omega})$ has an isolated zero for some ω, then, the resulting prefilter will have a zero on the unit circle and is therefore unstable. In any case, a practical system would use only a stable rational approximation of the ideal solution. Finally, we note that the optimum filter for the scheme of Fig. 3.6 can be obtained again as a special case by setting $L = 1$ in (3.14). The optimum prefilter will then have the following magnitude squared response:

$$|P_{opt}(e^{j\omega})|^2 = \frac{|F(e^{j\omega})|}{\sqrt{\hat{S}_{xx}(e^{j\omega})}} \tag{3.20}$$

and can be regarded as a multirate extension of the half whitening filter [37]. Using (3.20), we can derive an interesting expression for the coding gain of the scheme of Fig. 3.6.

Theorem 12 *With the optimum choice of the pre- and post filter, the coding gain expression for the scheme of Fig. 3.6 is*

$$\mathcal{G}_{opt} = \frac{M \int_{-\pi}^{\pi} S_{yy}(e^{j\omega}) \frac{d\omega}{2\pi}}{\left(\int_{-\pi}^{\pi} \sqrt{S_{yy}(e^{j\omega})} \frac{d\omega}{2\pi} \right)^2} = M \mathcal{G}_{hw} \tag{3.21}$$

where $\mathcal{G}_{hw}$ is the half whitening coding gain of the WSS process $y(n)$ [37].

Proof. By definition, the coding gain of the system is given by

$$\mathcal{G}_{opt} = \frac{\sigma_q^2}{\mathcal{E}_{opt}} = \frac{\sigma_x^2}{(\frac{1}{M} \int_{-\pi}^{\pi} \sqrt{\hat{S}_{xx}(e^{j\omega})} |F(e^{j\omega})| \frac{d\omega}{2\pi})^2} = \frac{M \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega}) \frac{d\omega}{2\pi}}{(\int_{-\pi}^{\pi} \sqrt{\hat{S}_{xx}(e^{j\omega})} |F(e^{j\omega})| \frac{d\omega}{2\pi})^2} \tag{3.22}$$

Substituting (3.12) in (3.22) and simplifying, we get

$$\mathcal{G}_{opt} = \frac{M\int_{-\pi}^{\pi} S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2 \frac{d\omega}{2\pi}}{(\int_{-\pi}^{\pi} \sqrt{S_{yy}(e^{jM\omega})}|F(e^{j\omega})|^2 \frac{d\omega}{2\pi})^2} \tag{3.23}$$

The integrals in both the numerator and the denominator can be interpreted as the variance of a WSS random process with a power spectrum density equal to $S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2$ and $\sqrt{S_{yy}(e^{jM\omega})}|F(e^{j\omega})|^2$ respectively. But we know that downsampling a WSS process produces another WSS process with the same variance. Therefore, we can write

$$\mathcal{G}_{opt} = \frac{M\int_{-\pi}^{\pi} (S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2) \downarrow_M \frac{d\omega}{2\pi}}{(\int_{-\pi}^{\pi} (\sqrt{S_{yy}(e^{jM\omega})}|F(e^{j\omega})|^2) \downarrow_M \frac{d\omega}{2\pi})^2} \tag{3.24}$$

Using the fact that $(S_{yy}(e^{jM\omega})|F(e^{j\omega})|^2) \downarrow_M = S_{yy}(e^{j\omega})(|F(e^{j\omega})|^2) \downarrow_M$ and that $(|F(e^{j\omega})|^2) \downarrow_M = 1$, we get (3.21). ■

The factor M in (3.21) is again due to the oversampled nature of the signal $x(n)$. It is interesting to note that the noise shaping contribution to $\mathcal{G}_{opt}$ in (3.21), which we denote by $\mathcal{G}_{hw}$, is exactly the *coding gain we would obtain by half whitening the WSS process $y(n)$ in the usual way* [37]. By appealing to the Cauchy Schwartz inequality again, we can show that $\mathcal{G}_{hw} \geq 1$ with equality iff the power spectral density $S_{yy}(e^{j\omega})$ is a constant, i.e., $y(n)$ is white noise. Therefore, for the particular system of Fig. 3.6, we will not get additional coding gain by noise shaping if the driving WSS process $y(n)$ in Fig. 3.1 is white noise. For completeness, we would like to mention that the following expression for the coding gain of Fig. 3.14 (the multiband case) can be derived under the assumption that the JWSS processes $y_k(n)$, $k = 0, 1, \ldots, L-1$, are uncorrelated :

$$\mathcal{G}_{opt} = \frac{M\int_{-\pi}^{\pi} \sum_{k=0}^{L-1} S_{y_k}(e^{jM\omega})|F_k(e^{j\omega})|^2 \frac{d\omega}{2\pi}}{(\int_{-\pi}^{\pi} \sqrt{\sum_{i=0}^{L-1} S_{y_i}(e^{jM\omega})|F_i(e^{j\omega})|^2} \sqrt{\sum_{n=0}^{L-1} |F_n(e^{j\omega})|^2} \frac{d\omega}{2\pi})^2} \tag{3.25}$$

3.7.2 Using a more general postfilter

Consider now the more general system of Fig. 3.15 where the postfilter is not assumed to be the inverse of the prefilter. The multiband case is shown in Fig. 3.16.

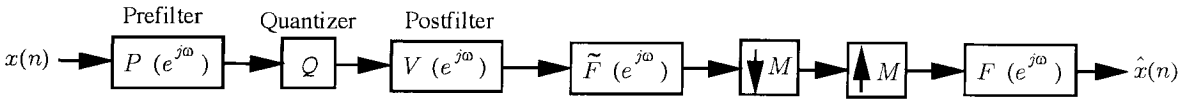

Fig. 3.15: General LTI pre- and post filters for noise shaping for the single band case

The goal is to jointly optimize the prefilter $P(e^{j\omega})$ and the postfilter $V(e^{j\omega})$ to again minimize the average m.s.e. $\triangleq 1/M \sum_{n=0}^{M-1} E\{\hat{x}(n) - x(n)\}^2$ under the following assumptions:

1. The input $x(n)$ is assumed to be a zero mean real wide sense stationary process.

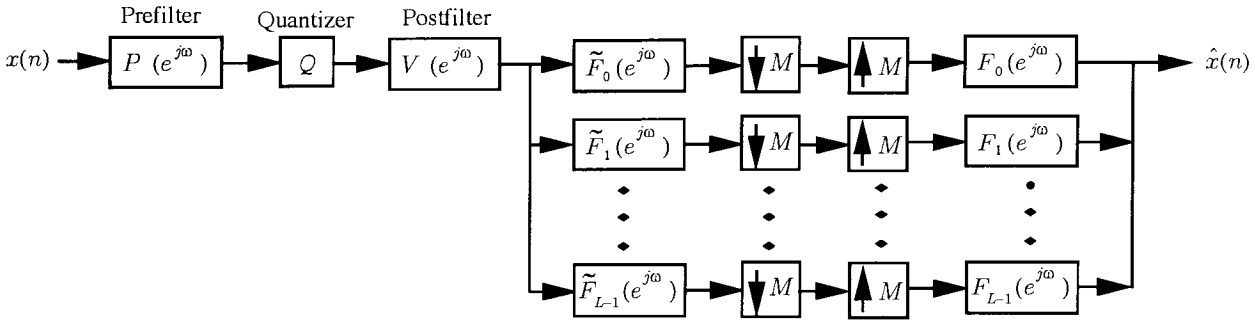

Fig. 3.16: General LTI pre- and post filters for noise shaping for the multiband case

2. The input $x(n)$ and the quantization noise $q(n)$ are uncorrelated processes, i.e.,

$$E\{x(n)q(m)\} \ = \ 0 \ \forall \ n,m$$

3. The quantization noise $q(n)$ is white with variance σ_q^2 as in (3.7).

4. The filters $P(e^{j\omega})$ and $V(e^{j\omega})$ are not constrained to be rational functions and can be non causal.

5. The power spectral density $S_{xx}(e^{j\omega})$ is positive for all ω. Furthermore, for the derivation of the optimum prefilter, we will also require $S_{xx}(e^{j\omega})$ and its first derivative to be continuous functions of frequency.

To solve the above problem, our approach will be the following : First, consider the single band case of Fig. 3.15. Unlike previous quantization schemes, we observe that in the absence of the quantizer, the scheme of Fig. 3.15 is *not* a perfect reconstruction system. The error sequence $e(n) = \hat{x}(n) - x(n)$ has in fact two components: one due to the mismatch between the pre- and post filters and the other due to the filtered quantization noise. We cannot therefore simply minimize the mean square reconstruction error before the downsampler as in the previous sections. Using the m.s.e. definition given above, we derive an expression for the average mean square reconstruction error $1/M \sum_{n=0}^{M-1} E\{e^2(n)\}$ in terms of the filters and the average power spectrum of the signal $x(n)$ and noise $q(n)$. The use of the average power spectral density of the $(CWSS)_M$ input $x(n)$ in this case is not theoretically correct, even under the same quantizer assumptions as before. Nevertheless, it is necessary to work with this quantity to obtain any meaningful comparison between this more general set up and the one of the previous subsection. The calculus of variation is used as a tool to derive closed form expressions for both the optimum pre- and post filters which are then used to obtain the coding gain expression of Fig. 3.15. Finally, we will show how to the generalize the results for the multiband case of Fig. 3.16.

Theorem 13 *For a fixed prefilter $P(e^{j\omega})$ and a given filter $F(e^{j\omega})$, the optimum postfilter $V_{opt}(e^{j\omega})$ is:*

$$V_{opt}(e^{j\omega}) = \frac{1}{P(e^{j\omega})} \frac{\hat{S}_{xx}(e^{j\omega})}{\hat{S}_{xx}(e^{j\omega}) + \frac{c2^{-2b}}{|P(e^{j\omega})|^2} \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega}) |P(e^{j\omega})|^2 \frac{d\omega}{2\pi}} \tag{3.26}$$

Proof. The average mean square reconstruction error can be expressed as follows:

$$\begin{aligned}
\mathcal{E} &= \frac{1}{M}\sum_{n=0}^{M-1} E\{e^2(n)\} \\
&= E\{x^2(n)\} + \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}^2(n)\} - \frac{1}{M}\sum_{n=0}^{M-1} E\{\hat{x}(n)x(n)\} - \frac{1}{M}\sum_{n=0}^{M-1} E\{x(n)\hat{x}(n)\} \\
&= \int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi} + \frac{1}{M}\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2|V(e^{j\omega})|^2|F(e^{j\omega})|^2\frac{d\omega}{2\pi} \\
&+ \frac{1}{M}\int_{-\pi}^{\pi} \sigma_q^2|V(e^{j\omega})|^2|F(e^{j\omega})|^2\frac{d\omega}{2\pi} \\
&- \frac{1}{M}2\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|F(e^{j\omega})|^2\Re\{P(e^{j\omega})V(e^{j\omega})\}\frac{d\omega}{2\pi} \qquad (3.27)
\end{aligned}$$

where $\Re$ stands for the real part. First, observe that the average m.s.e. dependency on the phase of the filters appears only in the last term. To minimize (3.27) with respect to the phase of the filters, the product $P(e^{j\omega})V(e^{j\omega})$ must be zero phase. To see this, simply set $P(e^{j\omega}) = |P(e^{j\omega})|e^{j\phi(\omega)}$ and $V(e^{j\omega}) = |V(e^{j\omega})|e^{j\Phi(\omega)}$. The real part of $P(e^{j\omega})V(e^{j\omega})$ is equal to $|V(e^{j\omega})||P(e^{j\omega})|\cos(\phi(\omega)+\Phi(\omega))$. To minimize (3.27), $\cos(\phi(\omega)+\Phi(\omega))$ must be equal to one. Dropping the real notation $\Re$ in (3.27), we now turn to the magnitude squared response of the filters. We first fix the prefilter $P(e^{j\omega})$ and optimize $|V(e^{j\omega})|$. This can be done by applying the Euler-Lagrange equation from the calculus of variation theory [24] to (3.27). The resulting expression is (3.26). ∎

It is interesting to note that the post filter is independent of $F(e^{j\omega})$. Substituting (3.26) into (3.27), we obtain the following average m.s.e. expression:

$$\mathcal{E}(|P|^2, b) = \int_{-\pi}^{\pi} \frac{\hat{S}_{xx}(e^{j\omega})(\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2(M - |F(e^{jw})|^2) + c2^{-2b}M\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{ju})|P(e^{ju})|^2\frac{du}{2\pi})}{\hat{S}_{xx}(e^{j\omega})|P(e^{jw})|^2 + c2^{-2b}\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{ju})|P(e^{ju})|^2\frac{du}{2\pi}}\frac{d\omega}{2\pi} \qquad (3.28)$$

The above equation is only a function of the magnitude squared response of the prefilter. From this point on, the problem under study is very similar to the one analyzed recently in the previous chapter and in fact, becomes exactly the same by setting M and $F(e^{j\omega})$ to unity in equation (3.28). We will therefore omit the proofs of the upcoming theorems referring the reader to Chapter 2 of the thesis.

Theorem 14 *The squared magnitude response $|P_{opt}(e^{jw})|^2$ that minimizes $\mathcal{E}(|P|^2, b)$, given in (3.28), is also the solution of the following constrained optimization problem:*

$$\min_{|P(e^{jw})|^2} \int_{-\pi}^{\pi} \frac{\hat{S}_{xx}(e^{j\omega})(\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2(1 - |F(e^{jw})|^2) + c2^{-2b})}{\hat{S}_{xx}(e^{j\omega})|P(e^{jw})|^2 + c2^{-2b}}\frac{d\omega}{2\pi} \qquad (3.29)$$

subject to:

$$\int_{-\pi}^{\pi} \hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2\frac{d\omega}{2\pi} = 1 \qquad (3.30)$$

Theorem 15 *The prefilter $|P_{opt}(e^{j\omega})|^2$ that minimizes (3.29) under the constraint (3.30) must have*

a magnitude response $|P_{opt}(e^{j\omega})|^2$ *in the following form:*

$$|P_{opt}(e^{j\omega})|^2 = \max\,(0, \frac{|F(e^{j\omega})|}{\sqrt{\hat{S}_{xx}(e^{j\omega})}}\left(\frac{1+c2^{-2b}}{\int_{-\pi}^{\pi}\sqrt{\hat{S}_{xx}(e^{j\omega})}|F(e^{j\omega})|\frac{d\omega}{2\pi}} - \frac{c2^{-2b}}{\sqrt{\hat{S}_{xx}(e^{j\omega})}}\right) \forall \quad \omega \in [-\pi,\pi] \quad (3.31)$$

Theorem 16 *With the optimal choice of pre- and postfilters, the coding gain expression for the scheme of Fig. 3.15 is*

$$\mathcal{G}_{opt} = (1+c2^{-2b})M\mathcal{G}_{hw} \quad (3.32)$$

as long as $|P_{opt}(e^{j\omega})|^2$ *in (3.32) is never set to zero* $\forall\ \omega$. *Here,* $\mathcal{G}_{hw}$ *is again the half whitening coding gain of the WSS process* $y(n)$.

Note that in this case the coding gain of the more general set up is a concatenation of three factors : $\mathcal{G}_{hw}$ due to the noise shaping, the oversampling factor M due to the signal model and $1+c2^{-2b}$ due to using a more general form of pre- and post filters.

To conclude this section, we would like to repeat the same procedure for the more general scheme of Fig. 3.16. We claim that, for this case, the optimum postfilter is still given by (3.26) and the optimum pre- filter magnitude squared response expression is obtained from (3.32) by simply replacing $|F(e^{j\omega})|$ by $\sqrt{\sum_{k=0}^{L-1}|F_k(e^{j\omega})|^2}$. To prove this, the key is to derive an expression for the average mean square reconstruction error of Fig. 3.16. Clearly, if we can show that $\mathcal{E}$ for the multiband case can be expressed as

$$\begin{aligned}
\mathcal{E} &= E\{x^2(n)\} + \frac{1}{M}\sum_{n=0}^{M-1}E\{\hat{x}^2(n)\} - \frac{1}{M}\sum_{n=0}^{M-1}E\{\hat{x}(n)x(n)\} - \frac{1}{M}\sum_{n=0}^{M-1}E\{x(n)\hat{x}(n)\} \\
&= \int_{-\pi}^{\pi}\hat{S}_{xx}(e^{j\omega})\frac{d\omega}{2\pi} + \frac{1}{M}\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{j\omega})|P(e^{j\omega})|^2|V(e^{j\omega})|^2\sum_{k=0}^{L-1}|F_k(e^{j\omega})|^2\frac{d\omega}{2\pi} \\
&+ \frac{1}{M}\int_{-\pi}^{\pi}\sigma_q^2|V(e^{j\omega})|^2\sum_{k=0}^{L-1}|F(e^{j\omega})|^2\frac{d\omega}{2\pi} \\
&- \frac{1}{M}2\Re\int_{-\pi}^{\pi}\hat{S}_{xx}(e^{j\omega})P(e^{j\omega})V(e^{j\omega})\sum_{k=0}^{L-1}|F(e^{j\omega})|^2\frac{d\omega}{2\pi} \qquad (3.33)
\end{aligned}$$

then, from the previous analysis, the above claim follows immediately. To derive (3.33), we need to only consider the second term and one of the cross terms. The second term $1/M\sum_{n=0}^{M-1}E\{\hat{x}^2(n)\}$ is the variance of the signal estimate at the output of Fig. 3.16. But from Result 3.2 of section 3.3, we know that it is equal to $1/M\sum_{k=0}^{L-1}\sigma_{y_k}^2$ where $\sigma_{y_k}^2$ is the variance of the signal estimate before the kth channel downsampler, $k=0,\ldots,L-1$. Substituting with $\sigma_{y_k}^2$ in this last relation, we obtain the second and third integral in (3.33). Consider now one of the cross terms, say $1/M\sum_{n=0}^{M-1}E\{\hat{x}(n)x(n)\}$. We can rewrite $\hat{x}(n)$ as $\sum_{k=0}^{L-1}\hat{x}_k(n)$ where $\hat{x}_k(n)$ is the signal estimate at the output of the kth channel. By the linearity of the expectation, this gives $1/M\sum_{k=0}^{L-1}\sum_{n=0}^{M-1}E\{\hat{x}_k(n)x(n)\}$. By interpreting the

single band case as the kth channel, the last integral follows easily. Equation (3.33) is therefore established and the claim is proved.

Example 1. *MA(1) process* $y(n)$. Assume that the input $x(n)$ is modeled as in Fig. 3.1 with $M = 2$ and $F(e^{j\omega}) = \frac{1}{\sqrt{2}}(1 + z^{-1})$. Let the driving WSS signal $y(n)$ be a zero mean gaussian MA(1) process with an autocorrelation sequence in the form

$$R_{xx}(k) = \begin{cases} 1 & k = 0 \\ \theta/1 + \theta^2 & k = 1, -1 \\ 0 & \text{otherwise} \end{cases} \tag{3.34}$$

The MA(1) process has to have $\frac{|R_{yy}(1)|}{R_{yy}(0)} \leq 1/2$ to ensure that the power spectral density is indeed non negative. We therefore restrict θ to be between -1 and 1. The power spectrum of the MA(1) process is given by:

$$S_{yy}(e^{j\omega}) = 1 - 2\frac{\theta}{(1+\theta^2)}cos(\omega) \tag{3.35}$$

Substituting (3.35) in (3.21), the coding gain expression of the scheme of Fig. 3.6 becomes

$$\mathcal{G}_{opt} = \frac{2(1+\theta^2)}{\left(\int_{-\pi}^{\pi} \sqrt{(1+\theta^2 - 2\theta cos(\omega))}\, \frac{d\omega}{2\pi}\right)^2} \tag{3.36}$$

The integral in (3.36) is equal to $F(-0.5\,, -0.5\,;\, 1\,;\, \theta^2)$ where $F(a\,,\, b\,;\, c\,;\, d)$ is Gauss's hypergeometric function. From [28], $F(-0.5\,,\, -0.5\,;\, 1\,;\theta^2)$ can be rewritten as $(1+\theta)F(-0.5\,,\, 0.5\,;\, 1\,;4\theta/(1+\theta)^2)$. This, in turn, can be simplified to $(1+\theta)\frac{2}{\pi}E(2\sqrt{(|\theta|)}/(1+\theta))$ where $E(.)$ is the complete elliptic integral of the second kind. The coding gain of the more general system can be obtained by multiplying (3.36) by $(1 + c2^{-2b})$ and obviously depends on the number of bits b. The plots of the coding gain are illustrated in Fig. 3.17 for $b = 3$ and $c = 2.4$.

Example 2. *AR(1) process* $y(n)$. With the same assumptions as in example 1, let the driving signal $y(n)$ be a zero mean gaussian AR(1) process with an autocorrelation sequence in the form $R_{yy}(k) = \rho^{|k|}$ where ρ is between 0 and 1. The power spectrum of the AR(1) process is

$$S_{yy}(e^{j\omega}) = \frac{1-\rho^2}{1+\rho^2 - 2\rho cos(\omega)} \tag{3.37}$$

Substituting (3.37) in (3.21), the coding gain expression for the scheme of Fig. 3.6 is as follows:

$$\mathcal{G}_{opt} = \frac{2}{(1-\rho^2)\left(\int_{-\pi}^{\pi} \frac{1}{\sqrt{(1+\rho^2 - 2\rho cos(\omega))}}\, \frac{d\omega}{2\pi}\right)^2} \tag{3.38}$$

The integral in (3.38) is equal to $\frac{2}{\pi}K(\rho)$ where $K(\rho)$ is the complete elliptic integral of the first kind [28]. Again, the coding gain of the more general system is obtained by multiplying (3.38) by $(1 + c2^{-2b})$. The plots of the coding gain are shown in Fig. 3.18 for $b = 3$ and $c = 2.4$.

3.8 Noise shaping by $(LPTV)_M$ pre- and post filters

In this section, we consider using $(LPTV)_M$ pre- and post filters instead of LTI ones surrounding a periodically time varying ($(PTV)_M$) quantizer. Since the signal model $x(n)$ is $(CWSS)_M$, restricting ourselves to linear time invariant noise shaping filters and quantizers is a loss of generality. Any optimum configuration for such processes should consist of $(LPTV)_M$ filters surrounding a ($(PTV)_M$) quantizer. Using some well known multirate results, it can be shown that this new quantization configuration is equivalent to an M-channel maximally decimated filter bank with M subband quantizers [73]. We will further impose the perfect reconstruction condition in the absence of quantization by confining ourselves to the class of perfect reconstruction filter banks. It follows that $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{-1}(e^{j\omega})$ where $\mathbf{E}(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$ denote respectively the analysis and synthesis polyphase matrices [73]. Equivalently, the analysis and synthesis filters satisfy the biorthogonality condition: $(P_k(e^{j\omega})Q_m(e^{j\omega}))| \downarrow_M = \delta(m-k)$ for all k, m. The goal is then to find the set of M analysis and synthesis filters, $P_k(e^{j\omega})$ and $Q_k(e^{j\omega})$ (equivalently the analysis and synthesis polyphase matrices), that minimize the average mean square error at the output due to the quantization noise. Because the general $(LPTV)_M$ problem is difficult to track analytically, we will only study two special forms of the above set up. The first case assumes that $\mathbf{E}(e^{j\omega})$ is diagonal with diagonal elements equal to $V_k(e^{j\omega})$. It follows that $\mathbf{R}(e^{j\omega})$ is also diagonal with diagonal elements equal to $\frac{1}{V_k(e^{j\omega})}$ for each k. The second case assumes that $\mathbf{E}(e^{j\omega})$ is paraunitary and we choose $\mathbf{R}(e^{j\omega}) = \mathbf{E}^{\dagger}(e^{j\omega})$. Alternatively, the synthesis filters $Q_k(e^{j\omega})$ are equal to $\tilde{P}_k(e^{j\omega})$ for each k and $(P_k(e^{j\omega})\tilde{P}_m(e^{j\omega}))| \downarrow_M = \delta(m-k)$ for all k, m. These two special forms are intermediate between one extreme (the LTI case) and the other (the general $(LPTV)_M$ case).

3.8.1 Letting the synthesis filter be the inverse of the analysis filter

Let $\mathbf{E}(e^{j\omega})$ be a diagonal matrix with diagonal elements equal to $V_k(e^{j\omega})$ and $\mathbf{R}(e^{j\omega})$ be also diagonal with diagonal elements equal to $\frac{1}{V_k(e^{j\omega})}$ for each k. The quantization configuration is shown in Fig. 3.19 for the single band case and Fig. 3.20 for the multiband case. The scalar quantizers labeled $\mathcal{Q}$ are modeled as additive noise sources $q_k(n)$ and individually satisfy relation (3.7).

Throughout this section, we will assume that the subband quantization noise sources $q_k(n)$ are

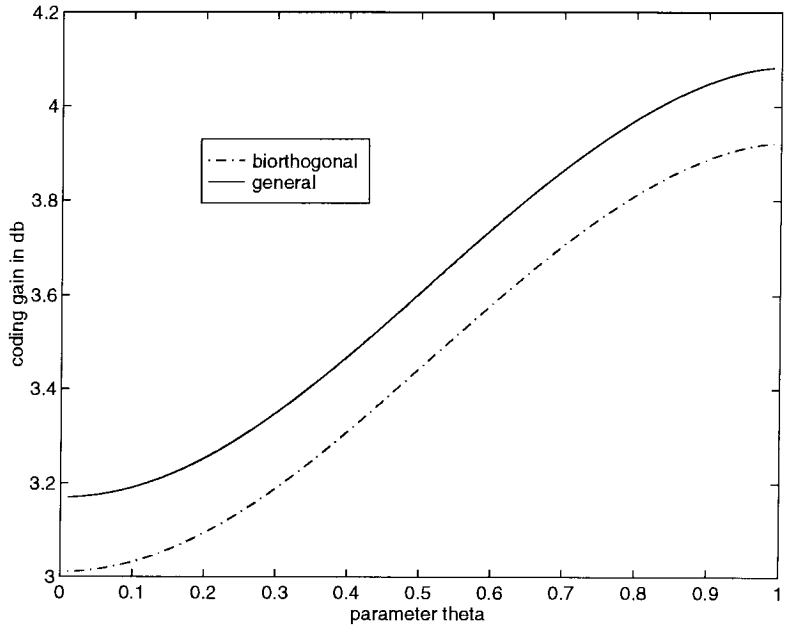

Fig. 3.17: Coding gain curves for the MA(1) case with $b = 3$ and $c = 2.4$

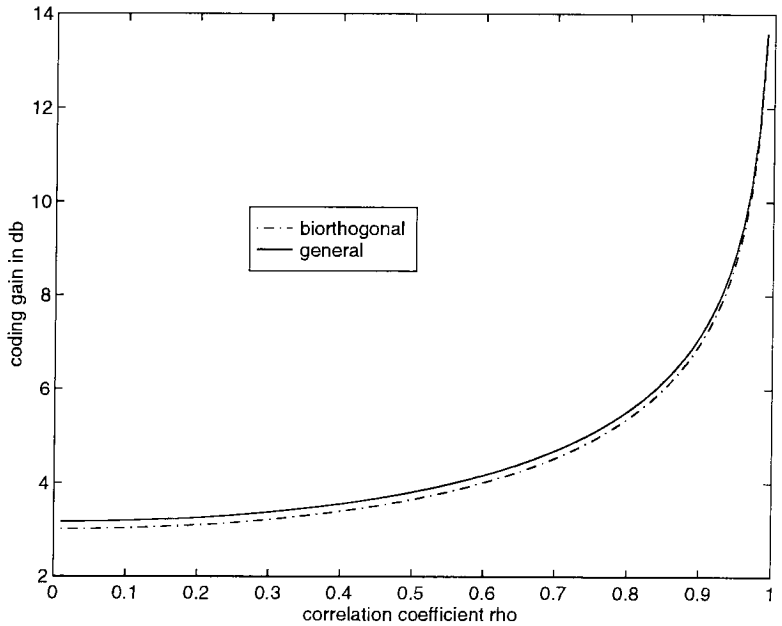

Fig. 3.18: Coding gain curves for the AR(1) case with $b = 3$ and $c = 2.4$

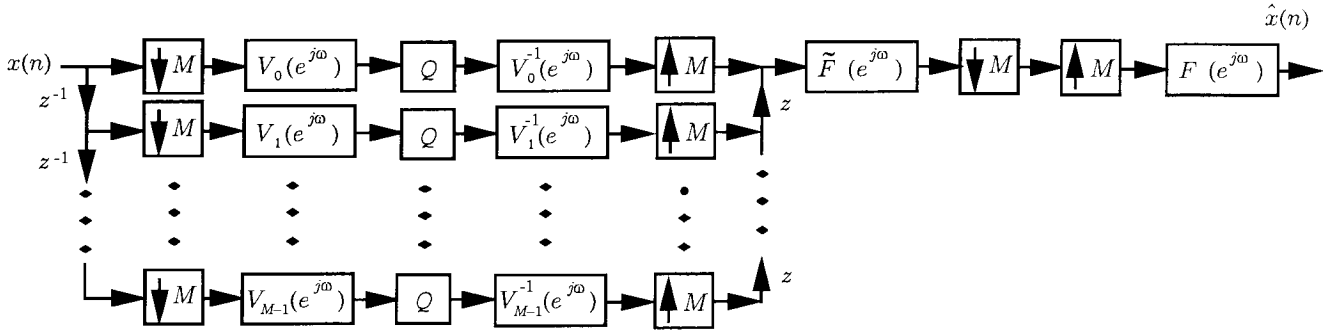

Fig. 3.19: Scheme 1 for noise shaping using $(LPTV)_M$ pre- and post filters (the single band case)

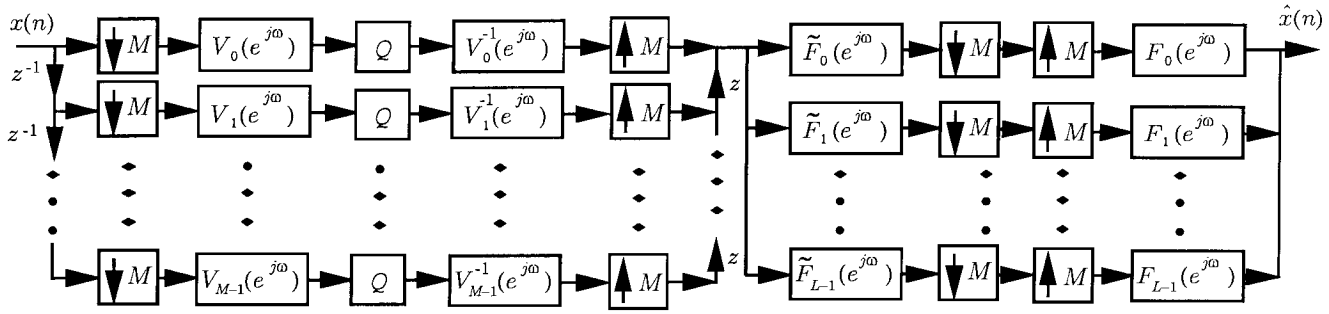

Fig. 3.20: Scheme 1 for noise shaping using $(LPTV)_M$ pre- and post filters (the multiband case)

white and pairwise uncorrelated, i.e., the noise power spectral density matrix is given by

$$\mathbf{S}_{qq}(e^{j\omega}) = \begin{pmatrix} \sigma_{q_0}^2 & 0 & \dots & 0 \\ 0 & \sigma_{q_1}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \sigma_{q_{M-1}}^2 \end{pmatrix} \tag{3.39}$$

The goal is then to jointly allocate the subband bits b_k under a fixed bit rate

$$b = \frac{1}{M} \sum_{k=0}^{M-1} b_k \tag{3.40}$$

and optimize $V_k(e^{j\omega})$ in order to minimize the average m.s.e. at the output of Fig. 3.19 and Fig. 3.20. Our strategy is as follows: we first find the optimum solution for the single band case of Fig. 3.19. Then, by interpreting the single band model as one of the L channels of the more general multiband case, the optimum solution for Fig. 3.20 follows.

Theorem 17 *Consider the scheme of Fig. 3.19 under the above assumptions. The optimum filter* $V_{opt}(e^{j\omega})$ *that minimizes the average mean square reconstruction error at the output is* **independent**

of k *and has the following magnitude squared response:*

$$|V_{opt}(e^{j\omega})|^2 = \frac{1}{\sqrt{S_{yy}(e^{j\omega})}} \tag{3.41}$$

where $S_{yy}(e^{j\omega})$ *is the power spectrum of the WSS process* $y(n)$ *in Fig. 3.1. With the above optimum filter expression, the coding gain of Fig. 3.19 is then given by:*

$$\mathcal{G}_{opt} = \frac{\sigma_y^2}{M(\prod_{k=0}^{M-1} \int_{-\pi}^{\pi} \sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2 \frac{d\omega}{2\pi})^{2/M}} \tag{3.42}$$

where $\tilde{R}_k(e^{j\omega})$ *is the* k*th polyphase component of* $\tilde{F}(e^{j\omega})$.

Proof. Since the system has the perfect reconstruction property in the absence of quantization, the error $e(n)$ at the output is simply the filtered quantization noise signal. After the downsampler, the filtered noise component $w(n)$ is WSS. By Result 2 of section 3.3, $\mathcal{E} = \frac{1}{M}\sigma_w^2$. To compute σ_w^2, we express the filter $\tilde{F}(e^{j\omega})$ in terms of its M polyphase components $\tilde{R}_k(e^{j\omega})$. Because the input signal $x(n)$ is modeled as in Fig. 3.1, we can also invoke the polyphase identity [73, page 133] at the input to simplify Fig. 3.19 to Fig. 3.21 (The interpolation filter was not drawn because we are really interested in evaluating σ_w^2 rather than σ_e^2. Since the quantization noise sources are assumed to be white and uncorrelated, the average mean squared error is therefore given by:

$$\begin{aligned} \mathcal{E} &= \frac{c}{M} \sum_{k=0}^{M-1} 2^{-2b_k} \sigma_{y_k}^2 \int_{-\pi}^{\pi} \frac{|\tilde{R}_k(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2} \frac{d\omega}{2\pi} \\ &= \frac{c}{M} \sum_{k=0}^{M-1} 2^{-2b_k} \int_{-\pi}^{\pi} S_{yy}(e^{j\omega})|R_k(e^{j\omega})|^2|V_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} \int_{-\pi}^{\pi} \frac{|\tilde{R}_k(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2} \frac{d\omega}{2\pi} \end{aligned} \tag{3.43}$$

Using the AM-GM inequality, equation (3.40) and the fact that $|R_k(e^{j\omega})|^2 = |\tilde{R}_k(e^{j\omega})|^2$, equation (3.43) reduces to:

$$\mathcal{E} \geq c2^{-2b}(\prod_{i=0}^{M-1} \int_{-\pi}^{\pi} S_{yy}(e^{j\omega})|V_k(e^{j\omega})|^2|\tilde{R}_k(e^{j\omega})|^2 \frac{d\omega}{2\pi} \int_{-\pi}^{\pi} \frac{|\tilde{R}_k(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2} \frac{d\omega}{2\pi})^{1/M} \tag{3.44}$$

Applying the Cauchy-Schwartz inequality to each term in (3.44), we get:

$$\mathcal{E}_{min} = c2^{-2b}(\prod_{k=0}^{M-1} \int_{-\pi}^{\pi} \sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2 \frac{d\omega}{2\pi})^{2/M} \tag{3.45}$$

This minimum bound is achieved by choosing $|V_{opt}(e^{j\omega})|^2$ as in (3.41). Finally, (3.42) follows immediately from the definition of the coding gain, equation (3.7) and the fact that $\sigma_x^2 = \sigma_y^2/M$. ∎

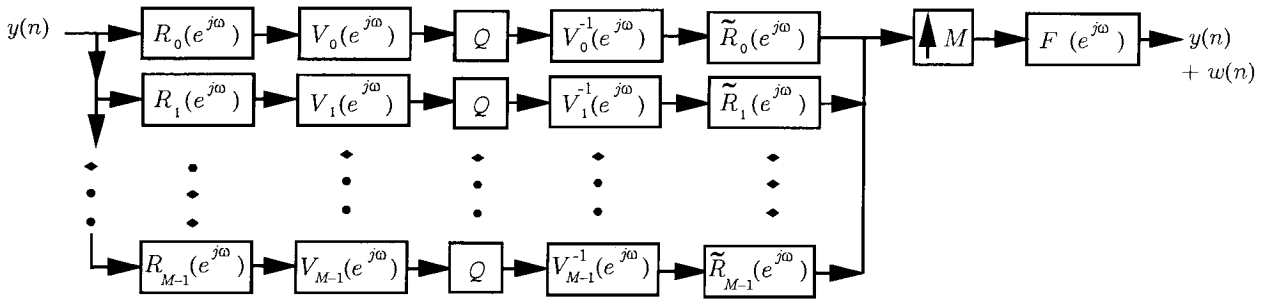

Fig. 3.21: An equivalent representation of Fig. 3.19

The LTI case is indeed a loss of generality. Since the class of $(LPTV)_M$ filters and $(PTV)_M$ quantizers include the LTI case, it is clear that the performance of this more general class of filters and quantizers is at least as good as the LTI one. We have already shown that the optimum $(LPTV)_M$ filter for Fig. 3.19 reduces to a LTI one. The question then becomes : Is the $(PTV)_M$ quantizer providing any excess gain over the LTI case and if so, by how much ? We show next that, even in this restricted form of $(LPTV)_M$ filters, the coding gain of the above scheme is always greater than the LTI one except when the magnitude squared response of the polyphase components $R_k(e^{j\omega})$ of $F(e^{j\omega})$ are equal for all k. Starting from the denominator of (3.22) (the coding gain expression of Fig. 3.6), one can write the following series of steps:

$$\begin{aligned}
\frac{1}{M}\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}\frac{d\omega}{2\pi}\right)^2 &= \frac{1}{M}\left(\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}\sum_{k=0}^{M-1}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^2 \\
&= \frac{1}{M}\left(\sum_{k=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^2 \\
&\geq \frac{1}{M}\times M^2\left(\left(\prod_{i=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^{1/M}\right)^2 \\
&= M\left(\prod_{i=0}^{M-1}\int_{-\pi}^{\pi}\sqrt{S_{yy}(e^{j\omega})}|\tilde{R}_k(e^{j\omega})|^2\frac{d\omega}{2\pi}\right)^{2/M} \qquad (3.46)
\end{aligned}$$

where the last line in (3.46) is the denominator of (3.42). Since the numerator is the same in both cases, the claim is proved. The first equality in (3.46) is obtained by using the power complementary property of the polyphase components of $F(e^{j\omega})$. The second line is a consequence of the linearity of the integral. The third line results from applying the AM-GM inequality. From the AM-GM formula, we know that equality is achieved if and only if all $|\tilde{R}_k(e^{j\omega})|^2$ are equal. From Fig. 3.21 (which was introduced in the proof of Theorem 11), we can see that this makes perfect sense. If all $|\tilde{R}_k(e^{j\omega})|^2$ are equal and since the optimum filters $V_k(e^{j\omega})$ are independent of k, the variance of the subband quantizer inputs will be all equal. There is therefore no variance disparity in the subbands and optimum bit allocation of the subband quantizers (which depends on the AM-GM inequality) can

not produce any gain. Using the single band result, we can now derive closed form expressions for the optimum $V_{opt_k}(e^{j\omega})$ and the average minimum mean squared error for the multiband case.

Theorem 18 *Consider the scheme of Fig. 3.20 under the above assumptions. The optimum filter $V_{opt_k}(e^{j\omega})$ (for each k) that minimizes the average mean square reconstruction error at the output has the following magnitude squared response:*

$$|V_{opt\,k}(e^{j\omega})|^2 = \frac{\sqrt{\sum_{i=0}^{L-1} |\tilde{R}_{ik}(e^{j\omega})|^2}}{\sqrt{S_k(e^{j\omega})}} \tag{3.47}$$

where $S_k(e^{j\omega}) = \sum_{i=0}^{L-1} S_{y_i}(e^{j\omega})|\tilde{R}_{ik}(e^{j\omega})|^2$ is the power spectrum of kth channel and $\tilde{R}_{ik}(e^{j\omega})$ is the kth polyphase component of the ith filter $\tilde{F}_i(e^{j\omega})$. Using the above optimum filters, the coding gain of Fig. 3.20 is then given by:

$$\mathcal{G}_{opt} = \frac{\sigma_y^2}{M(\prod_{k=0}^{M-1} \int_{-\pi}^{\pi} \sqrt{S_k(e^{j\omega})}\sqrt{\sum_{i=0}^{L-1} |\tilde{R}_{ik}(e^{j\omega})|^2}\frac{d\omega}{2\pi})^{2/M}} \tag{3.48}$$

Proof. By interpreting the single band result as one of the L channels of the multiband model and by using Result 2 of section 3.3, the average mean square error can be expressed as follows:

$$\begin{aligned} \mathcal{E} &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_k}\sigma_{y_k}^2 \int_{-\pi}^{\pi} \frac{\sum_{i=0}^{L-1} |\tilde{R}_{ik}(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi} \\ &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_k} \int_{-\pi}^{\pi} S_k(e^{j\omega})|V_k(e^{j\omega})|^2\frac{d\omega}{2\pi} \int_{-\pi}^{\pi} \frac{\sum_{i=0}^{L-1} |\tilde{R}_{ik}(e^{j\omega})|^2}{|V_k(e^{j\omega})|^2}\frac{d\omega}{2\pi} \end{aligned} \tag{3.49}$$

Using the same inequalities as in the proof of Theorem 17, we can immediately derive (3.47) and (3.48). ∎

Following the same type of reasoning as before, we again expect the coding gain of the more general $(LPTV)_M$ case of Fig. 3.20 to be higher than the analogous LTI one of Fig. 3.14. However, the complexity of the expressions (3.25) and (3.48) in this case prevents a formal mathematical proof.

Example 3. *Equal polyphase components.* Assume that the input $x(n)$ is modeled as in Fig. 3.1 where the upsampler $M = 2$ and the driving input $y(n)$ is a zero mean gaussian AR(1) process with correlation coefficient $0 < \rho < 1$. Furthermore, let $F(z)$ be the optimum FIR compaction filter of length two given by $\frac{1}{\sqrt{2}}(1 + z^{-1})$. The filter actually corresponds to one of the channels of a 2×2 KLT which is independent of the input statistics. In this case, the polyphase components of $F(e^{j\omega})$ are $R_0(e^{j\omega}) = R_1(e^{j\omega}) = \frac{1}{\sqrt{2}}$. Substituting in (3.42) and simplifying, we get (3.21), the coding gain expression of Fig. 3.6. In Example 2, a closed form expression was derived for the AR(1) case and a plot of the coding gain is shown in Fig. 3.18.

Example 4. *Unequal polyphase components.* With the same set of assumptions of Example 3, let the filter $F(z)$ be the optimum FIR compaction filter of length four. With $M = 2$ and assuming an AR(1) process, the following closed form expression was derived in [41] for the optimum compaction filter:

$$F(z) = a + cz^{-1} + bz^{-2} + dz^{-3} \tag{3.50}$$

where

$$\begin{aligned} a &= \frac{1}{p\sqrt{2}}\sqrt{\sqrt{p}+\sqrt{q}}\ ,\ b = \frac{1}{p\sqrt{2}}(\sqrt{\sqrt{p}+\sqrt{q}} - \sqrt{p}\sqrt{\sqrt{p}-\sqrt{q}}) \\ c &= \frac{1}{p\sqrt{2}}(\sqrt{p}\sqrt{\sqrt{p}+\sqrt{q}} - \sqrt{\sqrt{p}-\sqrt{q}}\ )\ ,\ d = -\frac{1}{p\sqrt{2}}\sqrt{\sqrt{p}-\sqrt{q}} \end{aligned} \tag{3.51}$$

and $p = 3 + \rho^2$, $q = 2 + \rho^2$. The polyphase components of $F(e^{j\omega})$ are $R_0(e^{j\omega}) = a + be^{-j\omega}$ and $R_1(e^{j\omega}) = c + de^{-j\omega}$. Substituting the power spectrum expression of an AR(1) process given by (3.37) into (3.42) and using some useful integral formulas [28, page 429], we can derive the following coding gain expression for the scheme of Fig. 3.19:

$$\mathcal{G}_{opt} = \frac{1}{2(1-\rho^2)\frac{2}{\pi}((a^2+b^2+\frac{2ab}{\rho})K(\rho) - \frac{2ab}{\rho}E(\rho))\frac{2}{\pi}((c^2+d^2+\frac{2cd}{\rho})K(\rho) - \frac{2cd}{\rho}E(\rho))} \tag{3.52}$$

where K(.) is the complete elliptic integral of the first kind and E(.) is the complete elliptic integral of the second kind. There is a reason for writing the denominator of (3.52) in this form. It can be shown that the factors $\frac{2}{\pi}((a^2+b^2+\frac{2ab}{\rho})\ K(\rho) - \frac{2ab}{\rho}E(\rho))$ and $\frac{2}{\pi}((c^2+d^2+\frac{2cd}{\rho})K(\rho) - \frac{2cd}{\rho}E(\rho))$ represent the variance of the outputs $R_0(e^{j\omega})$ and $R_1(e^{j\omega})$ respectively (with an input with power spectrum $\sqrt{S_{yy}(e^{j\omega})}$). Their product is the geometric mean which produces the extra gain over the LTI case. The further away they are in magnitude, the more gain we will obtain. The plots of the coding gain formulas (3.38) and (3.52) are shown in Fig. 3.22. We notice that the coding gain of the $(LPTV)_M$ case is indeed greater than the LTI one for all values of ρ, although not by a substantial amount for the AR(1) $y(n)$.

3.8.2 Using an orthonormal filter bank

Consider now the M-channel orthonormal filter bank shown in Fig. 3.23 for the single band model and in Fig. 3.25 for the multiband model. As in the previous subsection, we first analyze the single band case in detail and then use the corresponding results to derive analogous expressions for the multiband case. The quantization noise assumptions of the previous subsection are still true here. The goal is again to jointly allocate the subband bits b_k under the constraint (3.40) and optimize the orthonormal filter bank in order to minimize the average m.s.e.

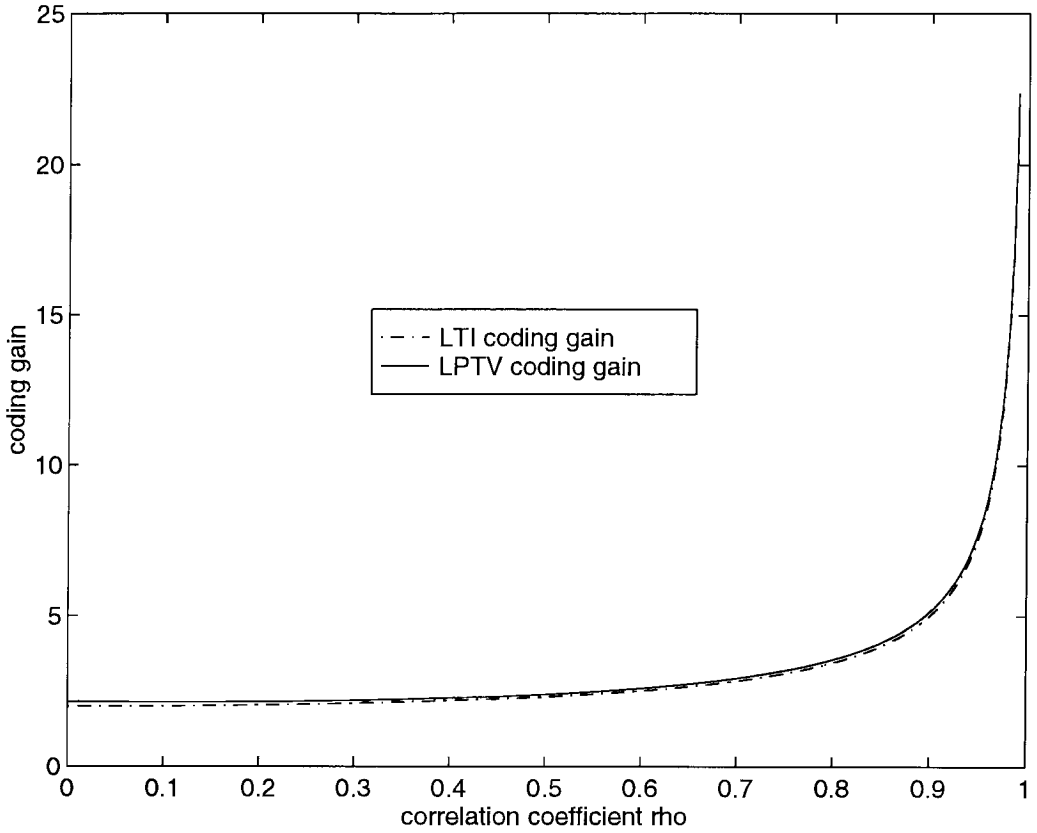

Fig. 3.22: Coding gain curves for the LTI and $(LPTV)_M$ cases under the assumption of a single band model with $M = 2$ and $y(n)$ is an AR(1) process

Theorem 19 *Consider the scheme of Fig. 3.23 under the above assumptions. The synthesis section of the optimum orthonormal filter bank* $\{P_k(e^{j\omega})\}$ *corresponds to choosing one of the filters, say* $\tilde{P}_0(e^{j\omega})$ *to be equal to* $\tilde{F}(e^{j\omega})$ *and the remaining filters* $\tilde{P}_k(e^{j\omega})$, $k = 1, \ldots, M-1$, *to be orthogonal to* $\tilde{P}_0(e^{j\omega})$. *In this case, the optimum orthonormal filter bank reduces to Fig. 3.5 where the quantizer* $\mathcal{Q}$ *is allocated* Mb *bits according to (3.38).*

Proof. By applying the blocking operation and using the polyphase representation [73], the scheme of Fig. 3.23 can be redrawn as in Fig. 3.24, where $\mathbf{E}(e^{j\omega})$ is the polyphase matrix of the analysis bank, $\mathbf{E}^{\dagger}(e^{j\omega})$ is the polyphase matrix of the synthesis bank and $\tilde{R}_k(e^{j\omega})$, $k = 0, \ldots, M-1$, are the M polyphase components of the filter $\tilde{F}(e^{j\omega})$.

Let $\mathbf{U}(e^{j\omega})$ be the $1 \times M$ vector whose kth element is $\tilde{R}_k(e^{j\omega})$. Then, the average m.s.e. can be expressed as follows:

$$\mathcal{E} = \frac{1}{M} \int_{-\pi}^{\pi} Tr(\mathbf{U}(e^{j\omega})\mathbf{E}(e^{j\omega})\mathbf{S_{qq}}\mathbf{E}^{\dagger}(e^{j\omega})\mathbf{U}^{\dagger}(e^{j\omega})) \frac{d\omega}{2\pi} \tag{3.53}$$

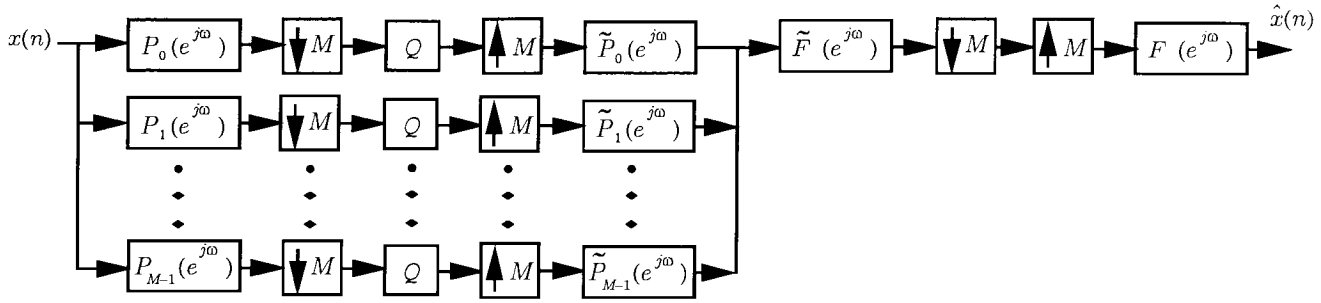

Fig. 3.23: Scheme 2 for noise shaping using $(LPTV)_M$ pre- and post filters (the single band case)

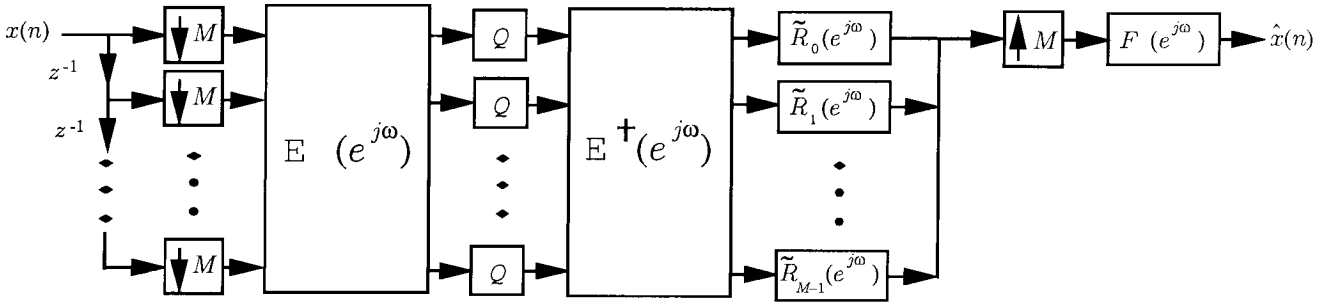

Fig. 3.24: The polyphase representation of Fig. 3.23

Since the integrand is in a quadratic form, the trace operator $Tr(.)$ can be removed. Furthermore, since $\mathbf{E}(e^{j\omega})\mathbf{E}^{\dagger}(e^{j\omega}) = I$ by orthonormality and $\mathbf{U}(e^{j\omega})\mathbf{U}^{\dagger}(e^{j\omega}) = 1$ by the Nyquist property of $F(e^{j\omega})$, we can rewrite (3.53) as follows:

$$\mathcal{E} = \frac{1}{M}\int_{-\pi}^{\pi} \frac{\mathbf{P}(e^{j\omega})\mathbf{S_{qq}}\mathbf{P}^{\dagger}(e^{j\omega})}{\mathbf{P}(e^{j\omega})\mathbf{P}^{\dagger}(e^{j\omega})}\frac{d\omega}{2\pi} \tag{3.54}$$

where $\mathbf{P}(e^{j\omega}) = \mathbf{U}(e^{j\omega})\mathbf{E}(e^{j\omega})$. Since the integrand of (3.54) is positive for all ω, minimizing (3.54) is equivalent to minimizing the integrand at each frequency. But for any fixed frequency ω_0, the ratio $\frac{\mathbf{P}(e^{j\omega_0})\mathbf{S_{qq}}\mathbf{P}^{\dagger}(e^{j\omega_0})}{\mathbf{P}(e^{j\omega_0})\mathbf{P}^{\dagger}(e^{j\omega_0})}$ is a Rayleigh quotient. For each frequency ω, the minimizing vector $\mathbf{P}_{opt}(e^{j\omega})$ has the form $(0 \quad \ldots \quad 1 \quad \ldots \quad 0)$ where the 1 in the ith position corresponds to the minimum noise variance $\sigma_{q_i}^2$. Since $\mathbf{P}(e^{j\omega}) = \mathbf{U}(e^{j\omega})\mathbf{E}(e^{j\omega})$, the minimizing vector $\mathbf{P}_{opt}(e^{j\omega})$ can be obtained by setting the ith column in $\mathbf{E}(e^{j\omega})$ to be equal to $\mathbf{U}^{\dagger}(e^{j\omega})$ and all the remaining columns to be orthogonal to $\mathbf{U}(e^{j\omega})$. This is equivalent to the statement of the theorem. ■

The optimum orthonormal filter bank thus reduces to the scheme of Fig. 3.5 with Mb bits allocated to the quantizer. The result of Theorem 19 is very intuitive and somehow expected: filter and decimate the oversampled signal $x(n)$ according to its model and then quantize $y(n)$ in Fig. 3.24 with $\hat{b} = Mb$ bits per sample. As we mentioned before, this amounts to fixing the bit rate (number of bits per second) in order to trade quantization resolution with sampling rate. It is interesting though to see that this very intuitive scheme is equivalent to using an optimum orthonormal filter bank as a

sophisticated quantizer to the input $x(n)$. With (3.7) in mind, the coding gain expression can be derived following the lines of the proof of Theorem 8 and is equal to $2^{2b(M-1)}$. This is an exponential gain which can be quite large for moderate values of M but unlike all previous schemes, depends on the bit rate b. Finally, to end this section, we would like to derive an analogous result (to Theorem 19) for the multiband case.

Theorem 20 *Consider the scheme of Fig. 3.25 under the same assumptions. The synthesis section of the optimum orthonormal filter bank corresponds to choosing L of the filters to be equal to $\tilde{F}_k(e^{j\omega})$ and the remaining filters $Q_k(e^{j\omega})$, $k = L+1, \ldots, M-1$, to be the $M-L-1$ orthogonal filters to $F_i(e^{j\omega})$, $i = 0, \ldots, L-1$. In this case, the optimum orthonormal filter bank reduces to Fig. 3.13 with an equivalent average number of bits $\hat{b}$ equal to Mb/L bits.*

Proof. By interpreting the single band result as one of the L channels of the multiband model and by using Result 2 and equation (3.40), the result follows immediately. ∎

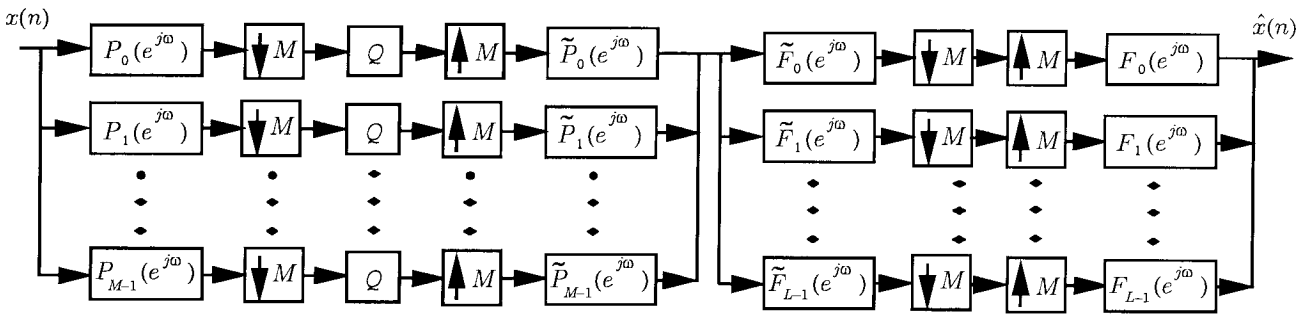

Fig. 3.25: Scheme 2 for noise shaping using $(LPTV)_M$ pre- and post filters (the multiband case)

With the above $\hat{b}$, we can now perform an optimum allocation of subband bits for the scheme of Fig. 3.13. This is a standard allocation problem that arises in subband coding application [37]. By applying the AM-GM inequality to the output error expression $\mathcal{E} = c\frac{1}{M}\sum_{k=0}^{L-1} 2^{-2b_k}\sigma_{y_k}^2$, we get

$$\mathcal{E}_{min} = c2^{-2b\frac{M}{L}}\frac{L}{M}\left(\prod_{i=0}^{L-1}\sigma_{y_i}^2\right)^{1/M} \tag{3.55}$$

which can be achieved by setting $b_k = b + 0.5\log_2\sigma_{y_k}^2 - 0.5\log_2\prod_{i=0}^{L-1}(\sigma_{y_i}^2)^{1/M}$. This optimum bit allocation formula will in almost all cases yield non integer solution for the bits. A quick remedy might be to use a simple rounding procedure or a more sophisticated algorithm [36] to obtain integer solutions. A detailed discussion of the topic of allocating *integer* bits to the channel quantizers is however outside the scope of the chapter. The noise variance in Fig. 3.9 simplifies to $c2^{-2b}\frac{L}{M}\left(\frac{1}{L}\sum_{k=0}^{L-1}\sigma_{y_k}^2\right)$. The coding

gain expression takes therefore the following form:

$$\mathcal{G}_{opt} = 2^{2b(\frac{M}{L}-1)} \frac{AM(\sigma_{y_i}^2)}{GM(\sigma_{y_i}^2)} \tag{3.56}$$

where AM is the arithmetic mean, GM is the geometric mean and $\sigma_{y_i}^2$ is the variance of the ith signal $y_i(n)$ in Fig. 3.2. We observe that when $L = 1$, we get the coding gain of the single band case and when $L = M$, the scheme of Fig. 3.13 reduces to an orthonormal filter bank, the average number of bits is equal to b and (3.56) reduces to the well known expression of the coding gain of an orthonormal filter bank.

Appendix A. Proof of Result 1 in section 3.3

The interpolated subband signals can be expressed as $x_i(n) = \sum_k y_i(k) f_i(n - Mk)$. Hence,

$$E[x_i(n)x_j{}^*(n - Mm)] = E\Big[\sum_{k'} y_i(k') f_i(n - Mk') \sum_k y_j{}^*(k) f_j{}^*(n - Mk - Mm)\Big] \tag{3.57}$$

Let $r(u)$ be the cross correlation between the jointly WSS processes $y_i(n)$ and $y_j(n)$, that is, $r(u) = E[y_i(n)y_j{}^*(n - u)$. Using the change of variable $k' - k = l$, the preceding equation becomes:

$$E[x_i(n)x_j{}^*(n - Mm)] = \sum_l r(l) \sum_{k'} f_i(n - Mk') f_j{}^*(n + M(l - m) - Mk') \tag{3.58}$$

Substituting (3.58) in the left hand side of (3.57), we get:

$$\frac{1}{M} \sum_l r(l) \sum_{k'} \sum_{k=0}^{M-1} f_i(n - (Mk' + k)) f_j{}^*(n + M(l - m) - (Mk' + k)) \tag{3.59}$$

Since M is positive, k' and k are integers and $0 \leq k < M$, we can always replace $Mk' + k$ by an integer u. That is, there always exist an integer u such that k' is the quotient and k is the remainder obtained from dividing u by M. We can therefore rewrite (3.59) as follows:

$$\frac{1}{M} \sum_l r(l) \sum_u f_i(n - u) f_j{}^*(n + M(l - m) - u) = \frac{1}{M} \sum_l r(l) \sum_k f_i(k) f_j{}^*(k + M(l - m)) \tag{3.60}$$

But the orthonormality of the filter bank implies, in particular, that $\sum_k f_i(k) f_j{}^*(k + M(l - m)) = 0$ for all l, m. Thus, the inner sum in (3.60) reduces to zero and the result follows.

Appendix B. Phase randomization of a $(CWSS)_M$ process

A WSS process $\hat{x}(n)$ can be obtained from a $(CWSS)_M$ process $x(n)$ by introducing a random shift θ in the $(CWSS)_M$ signal $x(n)$ [23, 35, 23]. The parameter θ is a discrete random variable that can take any integer value from 0 to $M-1$ with equal probability $1/M$. Furthermore, the random variable θ is assumed to be independent of $x(n)$. The autocorrelation function of $\hat{x}(n)$ is given by:

$$\begin{aligned} R_{\hat{x}\hat{x}}(n,k) &= E\{\hat{x}(n)\hat{x}(n-k)\} \\ &= E_\theta\{E\{x(n-\theta)x(n-k-\theta)|\theta\}\} = E_\theta\{R_{yy}(n-\theta,k)\} \\ &= \sum_{\theta=-\infty}^{\infty} R_{xx}(n-\theta,k)p(\theta) \\ &= \frac{1}{M}\sum_{\theta=0}^{M-1} R_{xx}(n-\theta,k) \\ &= \frac{1}{M}\sum_{m=n}^{M+n-1} R_{xx}(m,k) \end{aligned} \tag{3.61}$$

Now observe that

$$\begin{aligned} \frac{1}{M}\sum_{m=n}^{M+n-1} R_{xx}(m,k) &= \frac{1}{M}\sum_{m=0}^{M-1} R_{xx}(m,k) + \frac{1}{M}\sum_{m=M}^{M+n-1} R_{xx}(m,k) - \frac{1}{M}\sum_{m=0}^{n-1} R_{xx}(m,k) \\ &= \frac{1}{M}\sum_{m=0}^{M-1} R_{xx}(m,k) + \frac{1}{M}\sum_{m=M}^{M+n-1} R_{xx}(m,k) - \frac{1}{M}\sum_{m=M}^{M+n-1} R_{xx}(m,k) \\ &= \frac{1}{M}\sum_{m=0}^{M-1} R_{xx}(m,k) \end{aligned} \tag{3.62}$$

The second line follows because $R_{xx}(m,k) = R_{xx}(m+M,k)$ by cyclostationarity. The last sum is independent of n implying that $R_{\hat{x}\hat{x}}(n,k)$ is a function of k only and that the process $\hat{x}(n)$ is indeed WSS. Furthermore,

$$\hat{R}_{xx}(k) = \frac{1}{M}\sum_{n=0}^{M-1} R_{xx}(n,k) \tag{3.63}$$

Appendix C. Proof of equation (3.11)

Let $x(n)$ be a $(CWSS)_M$ process input to a linear time invariant filter $P(e^{j\omega})$. The output $z(n)$ is a $(CWSS)_M$ process [58] and is related to $x(n)$ by the well known convolution sum $z(n) = \sum_i p(i)x(n-i)$. Our goal is to derive an expression for the average variance of the $(CWSS)_M$ process $z(n)$. So,

$$\sigma_z^2 = \frac{1}{M}\sum_{n=0}^{M-1} E\{|z(n)|^2\}$$

$$
\begin{aligned}
&= \frac{1}{M}\sum_{n=0}^{M-1}\sum_{i}\sum_{j} p(i)p^*(j)E\{x(n-i)x^*(n-j)\} \\
&= \sum_{i}\sum_{j} p(i)p^*(j)\frac{1}{M}\sum_{n=0}^{M-1} R_{xx}(n, j-i) \\
&= \sum_{i}\sum_{j} p(i)p^*(j)\hat{R}_{xx}(j-i) \qquad (3.64)
\end{aligned}
$$

where the last equality follows from equation (3.63). By making the change of variables $j - i = l$, we get :

$$\sigma_z^2 = \sum_{l}\sum_{j} p(j-l)p^*(j)\hat{R}_{xx}(l) = \sum_{l}\sum_{l} r_p(l)\hat{R}_{xx}(l) \qquad (3.65)$$

where $r_p(l) \triangleq \sum_k p^*(k)p(k-n)$ is the deterministic autocorrelation of $p(n)$. Taking the discrete time fourier transform of (3.65), we get (3.11).

Appendix D. Power spectral density of an interpolated random process

Let $y(n)$ be a wide sense stationary (WSS) random process, input to an interpolation filter as shown in Fig. 3.1. The output $x(n)$ is in general a $(CWSS)_M$ process [58]. The average power spectral density of the "stationarized" process has the form

$$\hat{S}_{xx}(e^{j\omega}) = \frac{1}{M}S_{yy}(e^{j\omega M})|F(e^{j\omega})|^2 \qquad (3.66)$$

To derive (3.66), we can use (3.63) to write

$$\hat{R}_{xx}(k) = \frac{1}{M}\sum_{i}\sum_{j} R_{yy}(i-j)\sum_{n=0}^{M-1} f(n-Mi)f(n-k-Mj) \qquad (3.67)$$

Making the consecutive change of variables $i - j = l$ and $n - Mi = u$, equation (3.67) simplifies to:

$$
\begin{aligned}
\hat{R}_{xx}(k) &= \frac{1}{M}\sum_{l} R_{yy}(l)\sum_{i}\sum_{u=-Mi}^{M-1-Mi} f(u)f(u-k+Ml) \\
&= \frac{1}{M}\sum_{l} R_{yy}(l)\sum_{u} f(u)f(u-(k-Ml)) = \frac{1}{M}\sum_{l} R_{yy}(l)r_f(k-Ml) \qquad (3.68)
\end{aligned}
$$

where $r_f(n)$ is the deterministic autocorrelation of $f(n)$ as defined in appendix C. Equation (3.68) can be interpreted as passing the autocorrelation sequence $\frac{1}{M}R_{yy}(n)$ through the interpolation filter $r_f(n)$. Taking the fourier transform of (3.68), we obtain (3.66) or equivalently (3.12). The expression for multiband case, equation (3.15), can be obtained in a similar fashion. Again, from (3.63), one can

write:

$$\begin{aligned}
\hat{R}_{xx}(k) &= \frac{1}{M}\sum_{n=0}^{M-1} E\{x(n)x^*(n-k)\} \\
&= \frac{1}{M}\sum_{n=0}^{M-1}\sum_{i}\sum_{j}\mathbf{f}(n-Mi)E\{\mathbf{y}(n)\mathbf{y}^\dagger(n-k)\}\mathbf{f}^\dagger(n-k-Mj) \\
&= \frac{1}{M}\sum_{n=0}^{M-1}\sum_{i}\sum_{j}\mathbf{f}(n-Mi)\mathbf{R}_{\mathbf{y}}(k)\mathbf{f}^\dagger(n-k-Mj) \qquad (3.69)
\end{aligned}$$

where $\mathbf{f}(n) = (\, f_0(n) \quad f_1(n) \quad \ldots \quad f_{L-1}(n)\,)^T$ and $\mathbf{R}_{\mathbf{y}}(k)$ is the autocorrelation matrix of the L WSS inputs $y_k(n)$. By following the same steps used to derive (3.66), we obtain (3.13).

Chapter 4

The design of optimum FIR and IIR energy compaction filters

4.1 Introduction

Consider the following optimization problem

$$\max_{H(e^{j\omega})} \quad \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 W(e^{j\omega}) \frac{d\omega}{2\pi} \tag{4.1}$$

subject to

$$\frac{1}{M} \sum_{k=0}^{M-1} |H(e^{j(\omega - 2\pi k/M)})|^2 = |H(e^{j\omega})|^2|_{\downarrow M} = 1 \tag{4.2}$$

where $H(e^{j\omega})$ is a real FIR filter of fixed order N. The constraint (4.2) means in particular that the magnitude squared response $|H(e^{j\omega})|^2$ is Nyquist(M). The problem described above has received a lot of attention in the past because of its constant occurrence in different disciplines depending on the choice of the frequency weight function $W(e^{j\omega})$. A first application that comes to our mind is the ***design of FIR orthonormal filter banks***. This is a standard problem and the goal in this case is to obtain filters with good frequency response. For $M = 2$, the FB design problem is a special case of (4.1) and (4.2) with $W(e^{j\omega}) = rect(\omega/\omega_c)$ where ω_c is the passband cut off frequency of an ideal low pass filter. For the M-channel case, the well known factorization result of paraunitary filter bank can be used to design the orthonormal filter bank [73]. In this case, it turns out that a preliminary step in the filter bank design is the optimization problem described in (4.1) and (4.2) where $H(e^{j\omega})$ is the analysis filter in the first subband. The above problem arises also in ***a digital communications context***. In particular, the optimum design of a *matched* FIR transmitter and receiver filters satisfying a zero intersymbol interference condition when cascaded over an ideal channel has been considered in

[11]. The matched property implies that $H_r(e^{j\omega}) = H_t^*(e^{j\omega})$ where $*$ denotes complex conjugation and the zero intersymbol interference constraint is simply (4.2). It is not difficult to see that the problem is identical to the design of the first subband analysis filter in a M-channel orthonormal filter bank where ω_c is now a function of the bandwidth of the channel. Finally, the exact same problem has also recently appeared in the framework of ***echo cancelation*** [39].

Although the new method described in this chapter is general enough to cover the previously mentioned applications, it is the design of FIR orthonormal filter banks adapted to the input signal statistics and their connections to FIR energy compaction filters that provides the main motivation of this work. To give the reader a flavor of the major ideas, consider for the moment the M-channel orthonormal filter bank shown in Fig. 4.1 where the box labeled $\mathcal{Q}$ represents a scalar uniform quantizer.

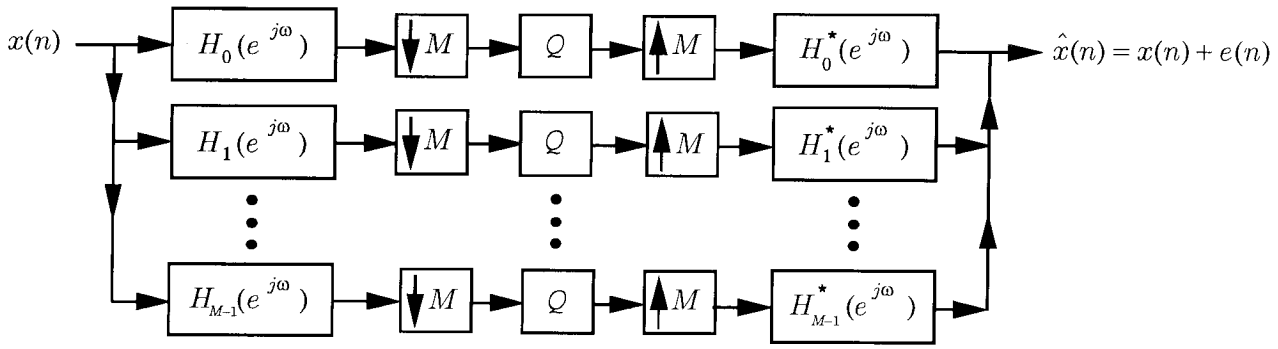

Fig. 4.1: An M-channel FIR orthonormal filter bank with scalar quantizers

Given a fixed budget of b bits for the subband quantizers, the design of an optimum orthonormal filter bank consists of simultaneously optimizing the analysis and synthesis filters as well as choosing a subband bit allocation strategy such that the average variance of the output error $e(n)$ is minimized. This is the classical subband coding problem defined now over the class of orthonormal filter banks. Under optimum bit allocation and with the *high bit rate* quantizer assumptions, the objective function is given by the well known coding gain expression [37]:

$$\mathcal{G}_{SBC}(M) = \sigma_x^2 / (\prod_{k=0}^{M-1} \sigma_{x_k}^2)^{1/M} \tag{4.3}$$

where $\sigma_{x_k}^2$ is the variance of the kth subband signal. Note that this expression holds whether the subband filters are FIR, IIR or ideal. Since σ_x^2 is fixed, the optimization of the analysis filters consists of minimizing the geometric mean of the subband variances under the orthonormality condition. The geometric mean is a concave function making the above problem quite a difficult one to solve both theoretically and practically. An orthonormal filter bank that maximizes (4.3) is termed an optimum orthonormal filter bank.

Consider now Fig. 4.2 where $(M-P)$ channels are dropped in the synthesis part of an M-channel orthonormal filter bank. An orthonormal filter bank that minimizes the average mean square reconstruction error for *any* P is called a principal component filter bank (PCFB). Using the orthonormality property [67, 64], it can be shown that, a principal component filter bank produces a decreasing arrangement of the the subband variances $\sigma_{x_1}^2 \geq \sigma_{x_2}^2 \ldots \geq \sigma_{x_P}^2$ such that, for any $1 \leq P < M$, $\sum_{k=1}^{P} \sigma_{x_k}^2$ is maximized. For $P = M$, $\sum_{k=1}^{M} \sigma_{x_k}^2 = M\sigma_x^2$ and is therefore fixed. The set of subband variances $\{\sigma_{x_k}^2\}$ generated by a principal component filter bank is said to "majorize" any other arbitrary set of subband variance $\{\sigma_{y_k}^2\}$. It is important to note that a *necessary* first step in designing PCFB is the optimization problem described in (4.1) and (4.2) with $W(e^{j\omega}) = S_{xx}(e^{j\omega})$ where $S_{xx}(e^{j\omega})$ is the input power spectrum (ignore the FIR constraint for the moment). The resulting optimum subband filter is called an optimum compaction filter. The interest in principal component filter banks is usually initiated by signal modeling applications, which leads to the standard question of finding the optimum (in the mean square sense) basis (subband filters) for a certain signal representation. The design of such filter banks is important on its own and plays a major role in applications such as the quantization of a class of non bandlimited signals based on multirate interconnections [66], multirate signal modeling [77, 78], optimization of wavelet basis [65] and time varying system identification [64] to name a few.

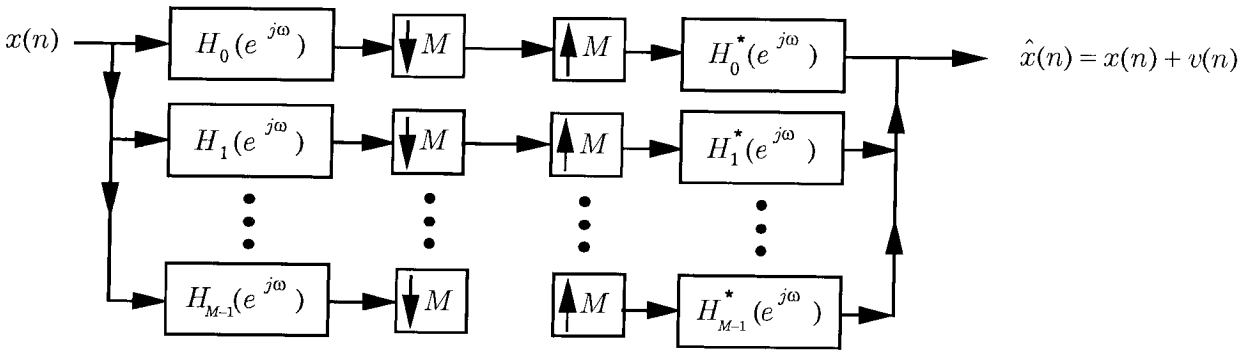

Fig. 4.2: An M-channel FIR principal component filter bank where $P = 2$

The design of a PCFB can also provide a solution to the difficult coding gain maximization problem described above. To see this, we first note that for the *two-channel* case ($M = 2$), designing one of the subband filters for maximum energy compaction is equivalent to obtaining an orthonormal filter bank with maximum coding gain. Using $2\sigma_x^2 = \sigma_{x_0}^2 + \sigma_{x_1}^2$, the coding gain expression can be rewritten as

$$\mathcal{G}_{SBC}(2) = \frac{1}{\sqrt{G_{comp}(2,N)(2 - G_{comp}(2,N))}}$$

where $G_{comp}(2,N)$ is the compaction gain. The compaction gain therefore uniquely determines the

coding gain of a 2-channel orthonormal subband coder. For the M-channel case, it can be shown, using a well known majorization theorem (see for example [49, pages 11–12]), that a PCFB is also an optimum orthonormal filter bank. This was perhaps first observed by Unser [67] and an independent proof was given by Xuan and Bamberger in [90]. Therefore, instead of directly minimizing (4.3), we can in principle optimize an orthonormal filter bank according to the input statistics by designing a PCFB. A recent paper by Moulin et al. [51] provides a procedure to design a PCFB. The proposed method begins by optimizing the first subband filter for maximum energy compaction and the remaining part of the design process reduces to finding a KLT matrix.

Unfortunately, there are still two pending limitations in using the PCFB approach for designing an optimum orthonormal filter bank: First, the *existence* of a principal component filter bank is currently only established for two extremes, namely the KLT case where $N < M$ and the ideal filter case $N = \infty$. More explicitly, given the space of all possible subband variances generated by an FIR orthonormal filter bank (order N is between the two extreme cases and $M > 2$), the *existence* of a specific set of variances $\{\sigma_{y_k}^2\}$ that will majorize all other possible subband variances is to be proven. Second, even if the existence issue is settled, the coding expression (4.3) holds only under the high bit rate assumption and will not be valid for most realistic compression schemes. The first of these two points is of theoretical relevance. In the worse case scenario where theoretical optimality cannot be claimed, the use of compaction filters generates in practice better compression results than conventional subband coders (see for example [62, 18, 90] for image compression applications). The second point however is more crucial. In this case, it remains to see if compression schemes designed under low bit rate assumptions perform significantly better. It is important to keep in mind that the optimization of signal dependent filter banks under low bit rate assumptions is a challenging problem at least in terms of theoretical tractability due to the lack of adequate low bit rate quantizer models (some models actually do exist but are quite difficult to use, see for example [7])

The main contribution of this chapter is the development of an efficient and numerically robust algorithm that finds the *global* optimum for the optimization problem described in (4.1) and (4.2). In fact, besides the analytical solution found by Aas et al. [1] for the special case of $M = 2$ and $W(e^{j\omega}) = rect(\omega/\omega_c)$, all other proposed techniques that we are aware of are *theoretically suboptimum.* We will discuss this issue in more details in the next section. The new approach is extremely general in the sense that it will work for an arbitrary M and any weight function $W(e^{j\omega})$. The problem is expressed in terms of the state space of the causal part of the product filter $F(z) = H(z)H(z^{-1})$ corresponding to the compaction filter $H(z)$. However, unlike previous product filter approaches, we can simultaneously satisfy the Nyquist constraint (4.2) and guarantee the positivity of $F_{opt}(e^{j\omega})$ *over all frequencies* which results in no loss of optimality. Moreover, the usual spectral factorization step required to obtain $H_{opt}(z)$ can be avoided by relating the state space realization of the causal part of $F_{opt}(z)$ to that of $H_{opt}(z)$. The new design method is expressed as a multi-objective semi

definite programming problem which is a convex problem that can be solved with great efficiency using recently developed interior point methods [53].

The chapter is organized as follows : In section 4.2, the FIR energy compaction problem is defined. We then provide the reader with a global overview of the main previous design techniques proposed to solve the general problem described in (4.1) and (4.2). In section 4.3, we formulate the problem in terms of the state space representation of the real part of the product filter $F(z)$ and show how this approach allows us to simultaneously satisfy the positivity and Nyquist constraints. Section 4.4 discusses the minimum phase spectral factor and its properties. A theorem and three corollaries characterizing this particular factor are derived. These results are important in order to avoid an additional spectral factorization step after the optimization procedure. Simplifications of these results for the particular FIR case under study are also presented. In section 4.5, we prove the convexity of the new formulation and briefly discuss the important properties of *semi definite programming* and the particular interior point method chosen to solve the problem. We also show how regularity constraints can be readily added. In section 4.6, numerical examples are provided to illustrate the excellent performance of the proposed algorithm. Finally, in section 4.7, we design a specific class of optimum IIR 2-channel energy compaction filters parameterized by a single coefficient. Some analytical results are derived and numerical examples are provided to illustrate the performance of these filters.

4.1.1 Chapter specific definitions

Definition 1. *Convex cone.* A set C is called a cone if for every $x \in C$ and scalar $\lambda \geq 0$, $\lambda x \in C$. A cone is convex if for $\lambda_1, \lambda_2 \geq 0$ and $x_1, x_2 \in C$, $\lambda_1 x_1 + \lambda_2 x_2 \in C$. The set of symmetric positive semi-definite matrices $\{P | P = P^T, P \succeq 0\}$ is a convex cone.

Definition 2. *Partial order.* The convex cone of symmetric positive semi-definite matrices $K = \{P | P = P^T, P \succeq 0\}$ defines a partial order on the space of symmetric matrices in the following sense: $P_2 \succeq_K P_1$ if and only if $P_2 - P_1$ is positive semi definite.

Definition 3. *Generalized inequalities.* The convex cone of symmetric positive semi definite matrices is also closed and pointed with a non empty interior (**int**K is the set of positive definite matrices). In this case, we refer to the induced partial ordering $\preceq_K$ as a *generalized inequality* associated with K. Furthermore, for this generalized inequality, we define the associated *generalized strict inequality* by $P_1 \prec_K P_2$ if and only if $P_1 - P_2 \in \mathbf{int}K$.

Definition 4. *Minimum element.* We say that $P_{min} \in S$ is a minimum element of S with respect to the (strict) generalized inequality $\preceq_K$ ($\prec_K$) if for every $P \in S$ we have $P_{min} \preceq_K (\prec_K) P$. If a set has a minimum element, this element is unique.

Definition 5. *Congruence.* An $N \times N$ matrix A is said to be congruent to B if there exits a non singular matrix T such that $B = TAT^T$. The following property of congruence with respect to positive semi-definite matrices can be proved :

The partial order induced by the positive semi definite cone is invariant under congruence, i.e.,

$$P_1 \preceq_K P_2 \implies TP_1T^T \preceq_K TP_2T^T \tag{4.4}$$

Assuming that T is *non singular*, a similar relation holds for the positive definite case with $\prec_K$ replacing $\preceq_K$. Note that by taking $P_1 = 0$, it follows that the cone of positive semi definite matrices is invariant under a congruence transformation.

4.2 The FIR energy compaction problem

Consider the scheme of Fig. 4.3 where $H(z)$ is an FIR filter of order N. The input $x(n)$ is assumed to be a zero mean WSS random process with a fixed power spectrum $S_{xx}(e^{j\omega})$. The output of the filter is decimated by M to produce $y(n)$. For a fixed pair (M,N), the FIR energy compaction problem is to maximize the output variance σ_y^2 under the Nyquist(M) constraint on $|H(e^{j\omega})|^2$. Since the decimator does not change the variance of the filter output, we note that σ_y^2 is given by (4.1) with $W(e^{j\omega}) = S_{xx}(e^{j\omega})$.

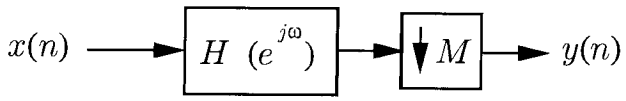

Fig. 4.3: Schematic of the FIR energy compaction problem.

We can therefore define the compaction gain as follows:

$$G_{comp}(M,N) = \frac{\sigma_y^2}{\sigma_x^2} = \frac{\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi}}{\int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi}} \tag{4.5}$$

where σ_x^2 is the variance of $x(n)$. The compaction gain $G_{comp}(M,N)$ is always $\leq M$. To see this, note that the Nyquist constraint (4.2) implies that $|H(e^{j\omega})|^2 \leq M \ \forall \ \omega$. Therefore,

$$\sigma_y^2 = \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} \leq M \int_{-\pi}^{\pi} S_{xx}(e^{j\omega}) \frac{d\omega}{2\pi} = M\sigma_x^2 \tag{4.6}$$

The equality holds if and only if $|H(e^{j\omega})|^2 = M$ for all ω for which $S_{xx}(e^{j\omega}) \neq 0$. Since this requires $|H(e^{j\omega})|^2$ to be identically zero for some region of frequency which is impossible with the finite order case, the FIR compaction gain will always be strictly less than M.

4.2.1 Summary of previous work

The general FIR optimization problem described in (4.1) and (4.2) has been considered by a number of authors in one form or another. The different design approaches can be broadly classified into four

main categories which we describe below.

1. **Optimizing the FIR lattice structure.** It is well known that the class of two-channel FIR orthonormal filter bank is *completely* parameterized by a lattice structure [73]. One can therefore optimize the lattice coefficient, namely a set of N angles θ_k, $1 \leq k \leq N$ to obtain the impulse response $h(n)$. An immediate consequence of this approach is that the problem is now unconstrained, i.e., the Nyquist constraint (4.2) is removed. Furthermore, unlike the product filter approach described below, no spectral factorization is required. There are also several other advantages which can be found in [73, pages 313–314]. The main drawback is that the problem is highly non linear and cannot be expressed as a convex program. The quasi-newton method used in [71] and the ring algorithm used in [18] both converge to a *local* maximum that depends on the starting point of the algorithm. To get a solution close to the global maximum, Taubman and Zakhor [62] propose to use a multi-start algorithm that generates several local optima over a subset of the parameter space. Finally, Xuan and Bamberger [91, 92] discuss similar procedures to optimize $2-D$ principal component filter banks.

2. **Quadratically constrained quadratic programming method.** The problem in this case is formulated in terms of the impulse response of the filter $H(z)$ as follows :

$$\text{maximize} \quad \mathbf{h}^T \mathbf{R_{xx}} \mathbf{h} \tag{4.7}$$

subject to

$$\mathbf{h}^T P_l \mathbf{h} = \delta(l) \quad \text{for } l = 0, 1, \ldots, N/2. \tag{4.8}$$

where $\mathbf{R_{xx}}$ is the Toeplitz hermitian matrix whose first row is $[r(0)\ r(1)\ \ldots\ r(N)]$ where $r(n)$ is the autocorrelation sequence of $x(n)$, $\mathbf{h}^T = [h(0)\ h(1)\ \ldots\ h(M-1)]$ and P_l are permutation matrices with $P_0 = I$. Since the permutation matrices $P(l)$, $l \neq 0$ are indefinite matrices, the above quadratically constrained quadratic optimization problem is non convex and very hard to solve both theoretically and numerically due to the existence of local minima. Several authors have used the classical method of Lagrange multipliers which leads to an iterative augmented Lagrangian algorithm (see for example [10, 84] for $M = 2$ and [11] for arbitrary M.)

3. **Optimizing the product filter.** Instead of optimizing the filter $H(z)$ directly, we first find the optimum product filter $F_{opt}(z) = H(z)H(z^{-1})$ and then, obtain $H_{opt}(z)$ from $F_{opt}(z)$ by spectral factorization. This approach was first introduced by Vaidyanathan et al. [72] in order to design an M-channel orthonormal filter bank with frequency selective (sharp) filters. It turns out that a primary step in the filter bank design is the optimization problem described in the introduction where $H(e^{j\omega})$ is the first subband filter. The goal in this case is to design a low pass filter with good frequency response and their proposed iterative method is based on an eigen

filter approach. Moulin [52] considered the design of energy compaction filters and formulated the problem as a linear semi-infinite programming method. We also mention the work of Pesquet and Combettes [54] who use an alternating projection method rather than linear programming to solve the problem. The main drawback with the algorithms proposed in [52, 54] is that they guarantee *theoretical* optimality only over a finite set of discrete frequencies $\{\omega_i, 0 \leq i \leq L\}$. Recently, Kirac and Vaidyanathan [42] proposed a *suboptimum* algorithm which they call the window method. The idea is to sacrifice optimality for speed by constraining the product filter to be the cascade of a window and a periodic filter. Although suboptimal, the new procedure is extremely efficient and performs very well at high filter order. This is primarily due to the fact that the window method is equivalent to the a linear program with $L = 2N$, where L is the number of discrete frequency points over which the linear program inequality constraints are enforced. Finally, it is important to keep in mind that in the product filter approach, all the above techniques require an additional spectral factorization step to obtain $H_{opt}(z)$.

4. **Analytical methods.** A subset of the product filter technique, the goal in this case is to derive an analytical procedure to obtain $F_{opt}(z)$. The elegance of this approach lies in the fact that no iterative numerical optimization is involved. For $M = 2$ and $W(e^{j\omega}) = rect(\omega/\omega_c)$ (ideal low pass filter with cutoff frequency ω_c), Aas et al. [1] were able to find a way to identify the unit-circle zeros of $F_{opt}(z)$. Once these are known, the other zeros can be found using Gaussian quadrature theory. Based on a characterization of positive definite matrices rather than Gaussian quadrature, Kirac and Vaidyanathan [43] extend the results of [1] for $M = 2$ and $W(e^{j\omega}) = S_{xx}(e^{j\omega})$ where $S_{xx}(e^{j\omega})$ is the power spectrum of $x(n)$. Unfortunately, the method works only for a certain class of random processes. Notice that in the analytical approach, a spectral factorization step is also necessary at the end.

4.3 Formulating the problem in terms of the product filter

From (4.1) and (4.2), we can immediately observe that the optimum solution, if it exists, is only a function of $|H(e^{j\omega})|^2$. By denoting **the product filter** $H(z)H(z^{-1})$ by $F(z)$, the output variance σ_y^2 in (4.1) can be rewritten as

$$\sigma_y^2 = r(0) + 2\sum_{n=1}^{N} f(n)r(n) \tag{4.9}$$

and the constraint (4.2) becomes

$$f(Mn) = \delta(n) \tag{4.10}$$

$$F(e^{j\omega}) \geq 0 \quad \forall\, \omega \tag{4.11}$$

where $r(i)$ denotes the i^{th} autocorrelation coefficient of the input $x(n)$. The objective function is now linear in the optimization variables $f(n)$, $n \geq 1$ at the expense of an additional constraint, namely equation (4.11) which we shall refer to as the positivity constraint. The major difficulty with the product filter approach is to simultaneously satisfy the positivity and Nyquist constraints. A standard way to solve such optimization problems is to consider a finite set of discrete frequencies $\{\omega_i, 0 \leq i \leq L\}$ over the continuous frequency axis and enforce the positivity constraint only at those frequencies (see for example [52, 54]). One major drawback with the "discretization" approach is that, in general, the resulting $G_{opt}(e^{j\omega})$ is negative between the discrete frequencies *no matter how large L is* which, in turn, creates an infeasible spectral factorization step. There are of course several ways to get around this problem (see for example [60]) but the point is no matter which method you choose, there will always be a loss of optimality.

We show next, using a well known result from system theory, that the positivity constraint can be satisfied over all ω at the expense of $N(N+1)/2$ additional optimization variables.

4.3.1 The state space approach

We first observe that since $F(z) = H(z)H(z^{-1})$, the product filter is a two sided symmetric sequence and we can therefore write $F(z)$ as $D(z) + D(z^{-1})$ where $D(z)$ is the causal part of $F(z)$ and $D(z^{-1})$ is the anti-causal part. Due to the symmetry of $F(z)$, it is then clear that $D(z)$ (or $D(z^{-1})$ in that matter) completely characterize $F(z)$. It is therefore natural to wonder whether the positivity condition on $F(e^{j\omega})$ can be reformulated in terms of some other condition(s) on $D(e^{j\omega})$. The answer turns out to be yes and is established by the well known **discrete time positive real lemma** [31]. We first start with a definition.

Definition 6. *Discrete positive real functions.* A square transfer matrix (function) $D(z)$ whose elements are real rational functions analytic in $|z| > 1$ is discrete positive real if, and only if, it satisfies all the following conditions :

$$\text{poles of } D(z) \text{ on } |z| = 1 \text{ are simple} \tag{4.12}$$

$$D(e^{j\omega}) + D(e^{-j\omega}) \geq 0 \quad \forall\, \omega \text{ at which } D(e^{j\omega}) \text{ exists} \tag{4.13}$$

Furthermore, If $z_0 = e^{j\omega_0}$, ω_0 real, is a pole of $D(z)$ and if K is the residue matrix of $D(z)$ at $z = z_0$, the matrix $Q = e^{-j\omega_0}\mathbf{K}$ is hermitian non negative definite.

Assume now that $D(z)$ has the following state space realization :

$$\begin{aligned} x(n+1) &= A_d x(n) + B_d u(n) \\ y(n) &= C_d x(n) + D_d u(n) \end{aligned} \tag{4.14}$$

where A_d is $N \times N$, B_d is $N \times P$, C_d is $M \times N$, and D_d is $M \times P$. For our case, $P = M = 1$. Then, the following lemma can be established.

Fact 1. *The discrete time positive real lemma* [31]. Let $D(z)$ be a transfer matrix (function) with real rational elements that is analytic in $|z| > 1$ with only simple poles on $|z| = 1$. Let (A_d, B_d, C_d, D_d) be a minimal realization of $D(z)$. Then, $D(z)$ is discrete positive real if, and only if, there exist a real symmetric positive definite matrix P_d and real matrices W_d and L_d such that :

$$P_d - A_d^T P_d A_d = L_d L_d^T \tag{4.15}$$

$$C_d^T - A_d^T P_d B_d = L_d W_d \tag{4.16}$$

$$D_d + D_d^T - B_d^T P_d B_d = W_d^T W_d \tag{4.17}$$

It is important to observe that the above equalities (4.15-4.17) can be rewritten as the following matrix inequality :

$$\mathcal{M}_d = \begin{bmatrix} P_d - A_d^T P_d A_d & C_d^T - A_d^T P_d B_d \\ C_d - B_d^T P_d A_d & D_d + D_d^T - B_d^T P_d B_d \end{bmatrix} = \begin{bmatrix} L_d \\ W_d^T \end{bmatrix} [L_d^T \; W_d] \succeq_K 0 \tag{4.18}$$

and therefore represent an equivalent condition for the positivity constraint to be satisfied. We will frequently alternate between the forms (4.15-4.17) and (4.18) to prove some results when we discuss spectral factorization issues.

As usual with the product filter formulation, the major difficulty at this point is to deal simultaneously with the positivity and Nyquist constraints. It turns out that, in this case, the Nyquist constraint can be imposed as an equality constraint in an extremely simple manner. To see this, assume that $D(z)$ is implemented in a direct form structure with the following state space representation:

$$A_d = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}, \quad B_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad C_d = [\, f(N) \; \ldots \; f(1) \,], \quad D_d = \frac{1}{2} \tag{4.19}$$

Clearly, this state space realization is minimal since the number of delay elements is equal to the degree of $D(z)$. Then, the Nyquist constraint can be written as a linear equality constraint:

$$Q \; C_d^T = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.20}$$

where Q is a diagonal matrix with diagonal elements $\{0 \ldots 1 \ldots 1 \ldots 0\}$. The positions of the unity

elements are determined by M. For example, for N = 5 and M = 2, the diagonal elements are $\{0\ 1\ 0\ 1\ 0\}$. In conclusion, we can represent the positivity constraint as a "linear" matrix inequality (LMI) whose entries are affine functions of the variables (P and C_d) and the Nyquist constraint as an equality constraint on C_d. The optimization problem described by (4.9), (4.10) and (4.11) can be rewritten as follows :

$$\max_{C_d} \quad C_d \, \mathbf{R}^T \tag{4.21}$$

where $\mathbf{R} = [r(N) \dots r(1)]$ such that there exists a symmetric positive definite matrix $P_d = P_d^T > 0$ satisfying

$$\begin{bmatrix} P_d - A_d^T P_d A_d & C_d^T - A_d^T P_d B_d \\ C_d - B_d^T P_d A_d & D_d + D_d^T - B_d^T P_d B_d \end{bmatrix} \succeq_K 0 \ , \quad Q \, C_d^T = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.22}$$

This is therefore a maximization problem in the variable vector C_d and a feasibility problem in the matrix P_d. In principle, the problem as stated above can be actually solved (we will discuss this issue in more details in section V). In specific, we can obtain a global optimum $C_{d_{opt}}$ and a feasible matrix P_d that will meet the constraints (4.22) and maximize the objective function (4.21). We can then spectral factorize $F_{opt}(z)$ to obtain $H_{opt}(z)$ using any of the well known algorithms (see for example [73, pages 854–856]). It turns out however that the spectral factorization step can be completely avoided by writing the state space representation of the *minimum phase spectral factor*, $H_{min}(z)$, in terms of the matrices A_d, B_d, C_d, D_d and the minimum element P_{min} of the convex set of positive definite matrices satisfying equation (4.22). The problem is then recast as a multi-objective maximization problem in terms of C_d and $-P_d$.

4.4 The minimum phase spectral factor

We first establish the existence of a spectral factor.

Theorem 21 *Assume that $D(z)$ satisfies the discrete time positive real lemma with a minimal realization (A_d, B_d, C_d, D_d). Then, the transfer function $H(z)$ with the following form :*

$$H(z) = W_d + L_d^T (zI - A_d)^{-1} B_d \tag{4.23}$$

is a spectral factor of $F(z) \triangleq D(z) + D(z^{-1}) = H(z)H(z^{-1})$.

Proof. The proof is given in Appendix A. ∎

The above theorem is analogous to the continuous time result found in [4]. It is important to note the number of rows of W_d and the number columns of L_d are unrestricted while the other dimensions of

P_d, L_d and W_d are automatically fixed. For example, in the SISO case, W_d can be a scalar or a column vector. The remainder of this section is dedicated to the study of the SISO minimum phase spectral factor $H_{min}(z)$. In particular, we will establish that the SISO minimum phase spectral factor $H_{min}(z)$ can be expressed in the form (4.23) with W_d being a scalar and L_d being a column vector. We will then present several alternative characterization of $H_{min}(z)$ which will be useful in the optimization procedure.

The development of these results follows by first applying the bilinear transformation $s = (z-1)/(z+1)$ on the continuous time minimum phase spectral factor and then using some strong results proved for the continuous time case by Willems [88] and Anderson [4, 5]. Unlike however the work in [4]-[5], the discussions and proofs presented here apply only to the scalar case, which is sufficient for the purpose of this chapter. We now introduce some well established facts.

Fact 2. *The continuous time positive real lemma* [4]. Let $D(s)$ be a transfer matrix (function) with real rational elements that is analytic in $Re\ s > 0$ with only simple poles on $Re\ s = 0$. Let (A_c, B_c, C_c, D_c) be a minimal realization of $D(s)$. Then, $D(s)$ is discrete positive real if, and only if, there exist a real symmetric positive definite matrix P_c and real matrices W_c and L_c such that :

$$-A_c^T P_c - P_c A_c = L_c L_c^T \tag{4.24}$$

$$C_c^T - P_c B_c = L_c W_c \tag{4.25}$$

$$D_c + D_c^T = W_c^T W_c \tag{4.26}$$

As in the discrete time case, an equivalent condition for the above equalities is the following matrix inequality:

$$\mathcal{M}_c = \begin{bmatrix} -P_c A_c - A_c^T P_c & C_c^T - P_c B_c \\ C_c - B_c^T P_c & D_c + D_c^T \end{bmatrix} \succeq_K 0 \tag{4.27}$$

The definition of positive real for the continuous time case is analogous to the discrete time case and can be found in [31].

Fact 3. *Spectral factorization* [93]. A transfer matrix $H(s)$ is said to be a spectral factor associated with an $m \times m$ positive real transfer matrix $D(s)$ if $D(s) + D^T(-s) = H^T(-s)H(s)$. $H(s)$ is a minimum spectral factor if it is $r \times m$ where r is the normal rank of $D(s) + D^T(s)$, if it has constant rank r in $Re\ s > 0$ and if its entries have no poles in $Re\ s \geq 0$. Moreover, $H(s)$ is unique to within multiplication on the right by an arbitrary real constant orthogonal matrix.

Fact 4. *The continuous time minimum phase spectral factor* [88],[5] Let $(A_c,\ B_c,\ C_c,\ D_c)$ be a minimal realization of a positive real transfer matrix $D(s)$. Then, the set of symmetric positive definite matrices, $\{P_c = P_c^T \succ_K 0\}$, satisfying the LMI constraint (4.27) has a minimum element $P_{c_{min}}$. This minimum element is associated with a minimum phase continuous time spectral factor $H_{min}(s)$, expressed in the form $H_{min}(s) = W_c + L_c^T(sI - A_c)^{-1}B_c$ where L_c and W_c satisfy equations (4.24-4.26)

with $P_c = P_{c_{min}}$.

Using the above and for the special SISO case, it immediately follows that the minimum phase spectral factor is unique and stable with no zeros in the right half plane $Re\ s > 0$. Furthermore, if $H_{min}(s) = W_c + L_c^T(sI - A_c)^{-1}B_c$, then, W_c is a 1×1 scalar and L is a $M \times 1$ column vector. Finally, all the eigen values of A_c have $Re\ \lambda_i < 0$.

By applying now the bilinear transformation on $H_{min}(s)$, the following results can be established.

Result 1. The bilinear transformation $s = (z-1)/(z+1)$ maps the *unique* minimum phase continuous time spectral factor $H_{min}(s)$ to the *unique* minimum phase discrete time spectral factor $H_{min}(z)$.

Proof. The first statement is a consequence of the s-plane to z-plane mapping property of the bilinear transformation. The uniqueness of $H_{min}(z)$ follows from the uniqueness of $H_{min}(s)$. ■

Theorem 22 *Assume that the continuous time minimum phase spectral factor $H_{min}(s)$ is given in the following form:*

$$H_{min}(s) = W_c + L_c(sI - A_c)^{-1}B_c \tag{4.28}$$

By applying the bilinear transformation, the minimum phase discrete time spectral factor $H_{min}(z)$ can be then expressed in the standard form of Theorem 1, namely,

$$H_{min}(z) = W_d + L_d(sI - A_d)^{-1}B_d \tag{4.29}$$

with

$$A_d = (I - A_c)^{-1}(I + A_c),\ B_d = 2(I - A_c)^{-2}B_c,\ L_d = L_c,\ W_d = W_c + L_c(I - A_c)^{-1}B_c \tag{4.30}$$

Furthermore, if (A_c, B_c, C_c, D_c) is a minimal realization, then, (A_d, B_d, C_d, D_d) is also a minimal realization.

Proof. The proof of the above statements is given in Appendix B and C respectively. ■

Note that $(I - A_c)$ must be non singular. Otherwise, one of the eigen values of A_c is equal to one which contradicts the stability of $H_{min}(s)$. The discrete to continuous relations (also found in [31]) can be easily derived from (4.30) and are as follows:

$$A_c = (A_d + I)^{-1}(A_d - I),\ B_c = 2(A_d + I)^{-2}B_d,\ L_c = L_d,\ W_c = W_d - L_d(A_d + I)^{-1}B_d \tag{4.31}$$

We are now ready to state the main theorem of this section.

Theorem 23 (The discrete time minimum phase spectral factor) . *Let $F(z)=D(z)+D(z^{-1})$ be a real rational function whose elements are analytic in $|z| > 1$. Assume that $D(z)$ satisfies the*

discrete time positive real lemma with a minimal realization $(A_d,\ B_d,\ C_d,\ D_d)$. *Then, the minimum phase spectral factor* $H_{min}(z)$ *is expressed in the standard form:*

$$H_{min}(z) = W_d + L_d^T(zI - A_d)^{-1}B_d \tag{4.32}$$

where

$$W_d = (D_d + D_d^T - B_d^T P_{min} B_d)^{1/2} \tag{4.33}$$

$$L_d^T = (C_d^T - A_d^T P_{min} B_d)(D_d + D_d^T - B_d^T P_{min} B_d)^{-1/2} \tag{4.34}$$

and $P_{d_{min}}$ *is the minimum element in the convex set of symmetric positive definite matrices satisfying the LMI (4.18)*

$$\begin{bmatrix} P_d - A_d^T P_d A_d & C_d^T - A_d^T P_d B_d \\ C_d - B_d^T P_d A_d & D_d + D_d^T - B_d^T P_d B_d \end{bmatrix} \succeq_K 0,$$

as well as the Nyquist constraint (4.20). Alternatively, $P_{d_{min}}$ *is also the unique solution to the following equations:*

$$P_d = A_d^T P_d A_d + (C_d^T - A_d^T P_d B_d)(D_d + D_d^T - B_d^T P_d B_d)^{-1}(C_d^T - A_d^T P_d B_d)^T \tag{4.35}$$

$$P_d = A_1^T P_d A_1 + A_1^T P_d B_d (R - B_d^T P_d B_d)^{-1} B_d^T P_d A_1 + C_d^T R^{-1} C_d$$

$$\text{where} \quad A_1 = A_d - B_d R^{-1} C_d, \quad R = D_d + D_d^T \succ_K 0 \tag{4.36}$$

Proof. The fact that the minimum phase spectral factor has the form (4.32) has been established in Theorem 21. Equations (4.33) and (4.34) are obtained from (4.16) and (4.17) by recalling that, for the SISO minimum phase spectral factor, W_d is a scalar and L_d is a column vector. Now, to prove that for the scalar $H_{min}(z)$, the LMI and the Nyquist constraints are satisfied with $P_d = P_{d_{min}}$, we first observe that the Nyquist constraint (4.20) can be removed by defining $C_{c_{new}} = C_{d_{new}} = (I - Q)C_d$ where I is the $N \times N$ identity matrix. The problem is the same as before but without the equality constraint. In the rest of the proof, the subscript *min* is omitted for convenience. The set $\{P_c = P_c^T \succ_K 0\}$ satisfying (4.27) is the same set of symmetric positive definite matrices satisfying the following matrix inequality:

$$\begin{bmatrix} I & 0 \\ B_d^T(I + A_d^T)^{-1} & I \end{bmatrix} \begin{bmatrix} -P_c A_c - A_c^T P_c & C_c^T - P_c B_c \\ C_c - B_c^T P_c & D_c + D_c^T \end{bmatrix} \begin{bmatrix} I & (I + A_d)^{-1} B_d \\ 0 & I \end{bmatrix} \succeq_K 0 \tag{4.37}$$

The claim is verified by noting that for each P_c, there will correspond a certain matrix $\mathcal{M}_c$, i.e., there exist a one to one relation between the set $\{P_c = P_c^T \succ_K 0\}$ and the set $\{\mathcal{M}_c = \mathcal{M}_c^T \succeq_K 0\}$ generated

by it. Since the left hand side of (4.37) is congruent to $\mathcal{M}_c$ and since the cone of symmetric positive semi- definite matrices is invariant under a congruence transformation, the result follows automatically. The second step is to multiply the above matrices and perform the following substitutions:

$$A_c = (A_d + I)^{-1}(A_d - I),\ B_c = 2(A_d + I)^{-2}B_d,\ C_c = C_d,\ D_c = D_d - C_d(A_d + I)^{-1}B_d \quad (4.38)$$

It can then be shown (see Appendix D) that the above operations produce the linear matrix inequality (4.18) with

$$P_d = 2(A_d^T + I)^{-1}P_c(A_d + I)^{-1}. \quad (4.39)$$

Equation (4.39) describes another non singular congruence transformation applied this time on the set $\{P_c = P_c^T \succ_K 0\}$. The congruence transformation preserves the positive definiteness of the matrices as well as the partial order induced on the set. The above sequence of arguments therefore proves that the set $\{P_d = P_d^T \succ_K 0\}$, satisfying the LMI constraint (4.18) and the Nyquist constraint (4.20), has a minimum element $P_{d_{min}}$ and that minimum is given by $2(A_d^T + I)^{-1}P_{c_{min}}(A_d + I)^{-1}$. It now remains to show that $P_{d_{min}}$ is the positive definite solution associated with the minimum phase spectral factor $H_{min}(z)$. This can be done by starting with equations (4.24-4.26) with $P_c = P_{c_{min}}$, applying the bilinear transformation on $H_{min}(s)$ which produces equation (4.31), making the additional substitutions

$$C_c = C_d,\ D_c = D_d - C_d(A_d + I)^{-1}B_d,\ P_{d_{min}} = 2(A_d^T + I)^{-1}P_{c_{min}}(A_d + I)^{-1}$$

and simplifying to obtain equations (4.15-4.17). The proof is omitted since it is similar to some of the previous derivations. The conclusion that these final equations are associated with the minimum phase spectral factor $H_{min}(z)$ follows from Result 1 and Theorem 22.
Finally, we note that Equation (4.35) follows by substituting (4.34) in (4.15). Equation (4.36) is derived from (4.35) assuming that R is positive definite and the proof can be found in Appendix E. ■

Corollary 1 *Assume that C_d is fixed, i.e., is also given along with the rest of the state space realization. By recognizing that (4.36) is a discrete time algebraic Ricatti equation (ARE), an analytic solution $P_{d_{min}}$ can be obtained by forming the $2N \times 2N$ Hamiltonian matrix [21, pages 430–438]*

$$\mathcal{H}_d = \begin{bmatrix} A_1 - B_d R^{-1} B_d^T A_1^{-T} C_d^T R^{-1} C_d & B_d R^{-1} B_d^T A_1^{-T} \\ -A_1^{-T} C_d^T R^{-1} C_d & A_1^{-T} \end{bmatrix} \quad (4.40)$$

$P_{d_{min}}$ is then equal to UV^{-1} where the $2N \times N$ matrix $[V\ U]^T$ is the matrix of eigen vectors associated with the eigen values that are inside the unit circle.

Corollary 2 *$P_{d_{min}}$ can be obtained from $P_{c_{min}}$ using the congruence relation (4.39) and the fact $P_{c_{min}}$ is the unique solution to the following equations :*

$$\begin{aligned} -A_c^T P_c - P_c A_c &= (C_c^T - P_c B_c)(D_c + D_c)^{-1}(C_c^T - P_c B_c)^T \qquad (4.41) \\ -A_2^T P_c - P_c A_2 &= P_c B_c R^{-1} B_c^T P_c + C_c^T R^{-1} C_c \\ & \text{where} \quad A_2 = A_c - B_c R^{-1} C_c, \quad R = D_c + D_c^T \qquad (4.42) \end{aligned}$$

where (A_c, B_c, C_c, D_c) are given by (4.38).

The proof that $P_{c_{min}}$ is the unique solution to (4.41) and (4.42) can be found in [4].

4.4.1 Simplifications for the FIR case

For the SISO FIR special case under consideration, we can further simplify the previously derived equations as follows: assume that the positive real function $F(z) = D(z) + D(z^{-1})$ where $D(z)$ has the following minimal state space realization:

$$\begin{aligned} A_d = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}, \quad B_d = [0\ 0\ \ldots\ 1]^T \\ C_d = [\ f(N)\ \ldots\ f(1)\ ], \quad D_d = \frac{1}{2} \end{aligned} \qquad (4.43)$$

The minimum phase spectral factor $H_{min}(z)$ of $F(z)$ is then given by:

$$\begin{aligned} H_{min}(z) &= \sqrt{1 - p_{d_{min}}(N,N)} + \frac{(C_d - B_d^T P_{d_{min}} A_d)}{\sqrt{1 - p_{d_{min}}(N,N)}}(zI - A_d)^{-1} B_d \\ &= \sqrt{1 - p_{d_{min}}(N,N)} + \frac{(C_d - B_d^T P_{d_{min}} A_d)}{\sqrt{1 - p_{d_{min}}(N,N)}}[z^{-N} z^{-(N-1)} \ldots z^{-1}]^T \\ &= \frac{1}{\sqrt{1 - p_{d_{min}}(N,N)}}\{1 - p_{d_{min}}(N,N) + (f(1) - p_{d_{min}}(N-1,N))z^{-1} \\ &+ \ \ldots - p_{d_{min}}(N-M,N)z^{-M} + \ \ldots + f(N)z^{-N}\} \end{aligned} \qquad (4.44)$$

The second equality follows by analogy with the transfer function of $D(z)$ and the third equality is obtained by direct substitution of (4.43). Closed form expressions for the continuous time system A_c, B_c, D_c, as well as A_1 and the Hamiltonian matrix $\mathcal{H}_d$ can be also derived for this particular case and

are given by :

$$A_c = \begin{bmatrix} 1 & -1 & \dots & (-1)^{N+1} \\ 0 & 1 & \dots & (-1)^N \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & \vdots \\ \vdots & 0 & \ddots & 1 \\ 0 & 0 & \dots & -1 \end{bmatrix} \tag{4.45}$$

$$B_c = 2 \begin{bmatrix} 1 & -1 & \dots & (-1)^{N+1} \\ 0 & 1 & \dots & (-1)^N \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} (-1)^{N+1} \\ (-1)^N \\ \vdots \\ 1 \end{bmatrix} = 2 \begin{bmatrix} N(-1)^{N+1} \\ (N-1)(-1)^N \\ \vdots \\ 1 \end{bmatrix} \tag{4.46}$$

$$D_c = D(z)|_{z=-1} = 1/2 - f(1) + \ldots + (-1)^N f(N), \quad D_c + D_c^T = F(z)|_{z=-1} \tag{4.47}$$

The above follows by noticing that $A_d + I$ is an $N \times N$ Jordan block. Its inverse is equal to an upper triangular Toeplitz matrix with first row $[1 \ -1 \ 1 \ldots \ (-1)^{N+1}]$. The knowledge of the form of equations (4.45)-(4.47) is useful in order to avoid computing inverses during the optimization process (if the continuous time characterization of corollary 2 is to be used). From (4.36), A_1 is a bottom companion matrix, i.e.,

$$A_1 = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 0 & 0 & 1 \dots & 0 \\ \vdots & \ddots & \ddots & 1 \\ -f(N) & \dots & -f(2) & -f(1) \end{bmatrix} \tag{4.48}$$

and is therefore always invertible if and only if $f(N) \neq 0$. The Hamiltonian matrix $\mathcal{H}_d$ can be also simplified and put in a final form that is only function of the product filter coefficients:

$$\mathcal{H}_d = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \ddots & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & 1 & 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 & -1/f(N) & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 & -f(N-1)/f(N) & 1 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & 0 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -f(1)/f(N) & 0 & \dots & 1 \\ f(N) & \dots & f(2) & f(1) & -1/f(N) & 0 & \dots & 0 \end{bmatrix} \tag{4.49}$$

Equation (4.49) is obtained by noting that the inverse of A_1 is a top companion matrix with first row equal to $-1/f(N)$ $[f(N-1) \ \dots \ f(1) \ 1]$. The transpose of A_1^{-1} gives a left companion matrix. The matrix $B_d R^{-1} B_d^T A_1^{-T}$ can be easily computed and is equal to a lower triangular Toeplitz matrix

with first column $[0\ 0\ \ldots\ -1/f(N)]^T$. The remaining block matrices are obtained by algebraic manipulation.

Using the fact that the inverse of the upper triangular Toeplitz matrix with first row $[\lambda\ -1\ 0\ \ldots\ 0]$ is an upper triangular Toeplitz matrix with first row $[1/\lambda\ 1/\lambda^2\ \ldots\ 1/\lambda^N]$ and using a determinant rule for block matrices, we can show that the characteristic polynomial $\det(\lambda I - \mathcal{H}_d)$ of the Hamiltonian matrix is $\frac{\lambda^N}{f(N)}F(\lambda)$. Therefore, the eigen values of $\mathcal{H}_d$ are simply *the zeros* of $F(z)$. The minimum phase spectral factor $H_{min}(z)$, usually reconstructed by interpolating the zeros that are inside or on the unit circle is now obtained by manipulating the eigen vectors associated with the zeros that are inside or on the unit circle. Before going through an example, we end this section with the following corrolary.

Corollary 3 *For the special SISO FIR case under consideration, the minimum element $P_{d_{min}}$ has the following form:*

$$\begin{aligned} P_{d_{min}} &= \sum_{k=0}^{N-1} {A_d^T}^k L_d^T L_d {A_d}^k \\ &= {\mathcal{S}_{L_d,A_d}}^T \mathcal{S}_{L_d,A_d} \end{aligned}$$

where $\mathcal{S}_{L_d,A_d}$ denotes the observability matrix of the realization $\{A_d, L_d^T\}$. The above result follows from the fact that $P_{d_{min}}$ has to satisfy a discrete time Lyapunov equation (4.15). The solution of a discrete time Lyapunov equation can be found in [73, pages 684–685] and can be further simplified for this case using the fact that $A_d^N = 0$. It follows also that the $Tr(P_{d_{min}}) = \sum_{n=1}^{N} n l_d^2(n)$ and that equation (4.17) is the unit energy constraint enforced on the optimum filter.

Example 1. The 2×2 KLT. Assume that $N = 1$ and $M = 2$. The state space representation for $D(z)$ in this case is: $A_d = 0, B_d = 1, C_d = f(1)$ and $D_d = 1/2$. Using (4.35) and this particular state space realization, the optimization problem can be simplified and recast as follows: maximize $f(1)R(1)$ subject to the equality constraint $\sqrt{P_{d_{min}}(1 - P_{d_{min}})} = f(1)$ where $0 < P_{d_{min}} < 1$. With $A_1 = -f(1)$ and $R = 1$, it is easy to check that the same problem statement is obtained if we used (4.36) instead of (4.35). The above formulation can be rewritten as an "unconstrained" problem in the variable $P_{d_{min}}$, namely: maximize $\sqrt{P_{d_{min}}(1 - P_{d_{min}})}R(1)$ where $0 < P_{d_{min}} < 1$. Using the AM-GM inequality, the convex objective function is upper bounded by $R(1)/2$ which is independent of $f(1)$. The bound is achieved if and only if $1 - P_{d_{min}} = P_{d_{min}}$, i.e., $P_{d_{min}} = 1/2$. From $\sqrt{P_{d_{min}}(1 - P_{d_{min}})} = f(1)$, it then follows that $f(1) = 1/2$. Using (4.32), (4.34) and the above state space representation, the minimum phase spectral factor has the form : $H_{min}(z) = (1 - P_{d_{min}})^{1/2} + f(1)(1 - P_{d_{min}})^{-1/2}z^{-1}$. By substituting $P_{d_{min}} = f(1) = 1/2$, we get: $H_{min}(z) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}z^{-1}$ which corresponds to the first row of the 2×2 universal KLT. We also note that the above result could have been obtained from corrolary 3 which in in this case reduces to the two equations: $f(1) = l^2$ and $1 - l^2 = w^2$. Neither the

product filter nor the spectral factor coefficients depend on the value of $R(1)$. The compaction gain is however equal to $1+|R(1)|/R(0)$. We can also check the alternative characterizations for $P_{d_{min}}$ given in corollaries 1 and 2. Assuming $C_d = f(1) = 1/2$, The Hamiltonian matrix of corollary 1 is then the 2×2 matrix:

$$\mathcal{H}_d = \begin{bmatrix} 0 & -1/f(1) \\ f(1) & -1/f(1) \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ 1/2 & -2 \end{bmatrix}$$

The characteristic polynomial is given by: $\lambda^2 + 2\lambda + 1$ which has a repeated eigen value equal to -1 (case of a double root on the unit circle). If we denote the eigen vector by $[v\ u]^T$, it is easy to see from the eigen value decomposition of $\mathcal{H}_d$ that $u = v/2$. Therefore, $P_{d_{min}} = uv^{-1} = 1/2$ as expected. To check corollary 2, note that $A_c = -1, B_c = 2, C_c = f(1), D_c = 1/2 - f(1)$ and with $A_2 = -1/(1-2f(1))$, $R = 1-2f(1)$, (4.41) and (4.42) reduce to: $\sqrt{2P_{c_{min}}(1-2P_{c_{min}})} = f(1)$ where $0 < P_{c_{min}} < 1$. The problem can be put in the following form : maximize $\sqrt{2P_{c_{min}}(1-2P_{c_{min}})}R(1)$, which can be solved in the same way as the discrete time case. The final result is $C_c = f(1) = C_d = 1/2$ and $P_{c_{min}} = 1/4$. We also note that $P_{d_{min}} = 2(A_d^T + I)^{-1}P_c(A_d + I)^{-1} = 1/2$. Finally, the continuous time spectral factor $H_{min}(s)$ is equal to $\sqrt{2}/(s+1)$. It can be easily verified that this is the result we obtain by applying the bilinear transformation $z^{-1} = (1-s)/(1+s)$ over $H_{min}(z)$.

Although the above example uses primarily conditions (4.35) and (4.36) and/or their continuous time equivalents as the constraining equations to solve the maximization problem, it is actually the linear matrix inequalities $\mathcal{M}_d$ in (4.18) and $\mathcal{M}_c$ in (4.27) that come into play when using *semi-definite programming* to solve the general (N, M) case as we discuss in the next section.

4.5 The optimization procedure

Since the minimum phase spectral factor is determined by $P_{d_{min}}$, we can certainly attempt to include P_d in our objective function (4.21). Minimizing P_d directly will produce a vector valued objective function. It would be nice if we can minimize instead a scalar valued function of P_d and this can be established by the following observation.

Observation 1. Assume that $P_{d_{min}}$ is the minimum element in the convex set of symmetric positive definite matrices satisfying the LMI constraint (4.18). Then, $P_d = P_{d_{min}}$ if and only if $Tr(WP_d)$ is minimum for any diagonal positive semi-definite matrix W.

Proof. The necessary part is obvious because $P_1 \succ_K P_2$ implies that $Tr(WP_1) > Tr(WP_2)$. The sufficiency part follows from the uniqueness of the minimum element $P_{d_{min}}$. ∎

The optimization problem formulated at the end of section III reduces now to the following final form:

$$\max_{C_d, P_d} \quad C_d \, \mathbf{R}^T - Tr(W P_d) \tag{4.50}$$

where $\mathbf{R}^T = [r(N) \dots r(1)]^T$ and W is a diagonal positive semi definite weight matrix such that

$$\mathcal{M}_d = \begin{bmatrix} P_d - A_d^T P_d A_d & C_d^T - A_d^T P_d B_d \\ C_d - B_d^T P_d A_d & D_d + D_d^T - B_d^T P_d B_d \end{bmatrix} \succeq_K 0 \ , \quad Q \, C_d^T = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.51}$$

and is therefore a maximization problem in the variable vector C_d and a minimization problem in the matrix P_d.

Observation 2. The optimization problem described by (4.50) and (4.51) is a convex problem in the variables C_d and P_d.

Proof. The reader who is familiar with semi definite programming will immediately recognize that the above optimization problem can be transformed to a semi definite program which by definition is convex. The following proof however does not require this background and might be more instructive. The objective function is linear in both C_d and P_d and is therefore convex. The matrix $\mathcal{M}_d$ is K-convex with respect to C_d since for all C_{d_1} and $C_{d_2} \in R^N$ and for all $0 \leq \lambda \leq 1$,

$$\mathcal{M}_d(\lambda C_{d_1} + (1-\lambda) C_{d_2}) = \lambda \mathcal{M}_d(C_{d_1}) + (1-\lambda) \mathcal{M}_d(C_{d_2})$$

The same argument holds for P_d. Finally, the equality constraint is linear in C_d (and trivially in P_d) and is therefore convex. In fact, Equations (4.50) and (4.51) are in a standard form convex optimization problem with generalized inequality constraints. ∎

The above formulation is therefore a convex multi-objective optimization problem for which any local solution is also a global one. The solution obviously depends on the choice of W and the vector $\mathbf{R}$. We note that the result of observation 1 will actually hold if, for example, we used the determinant function instead of the trace. The particular choice of the trace function was intentional in order to use semi definite programming. The weight matrix W was included in the problem because, unlike in section 4.3.1 where the objective function was uniquely determined by C_d, we now have two *competing* objectives. In other terms, for each C_d, there will correspond a certain P_d and the idea is to choose the weight so that optimality of C_d is never compromised, i.e., in order to prohibit $Tr(W P_d)$ from becoming the dominant factor in (4.50). Finally, we would like to mention that the continuous time characterization can also be used. In particular, with the continuous time state space realization

described in (4.45), (4.46), (4.47) and $C_c = C_d$, the optimization problem becomes:

$$\max_{C_c, P_c} \quad C_c \, \mathbf{R}^T - Tr(WP_c) \tag{4.52}$$

where $\mathbf{R}^T$ and W are defined as before such that

$$\mathcal{M}_c = \begin{bmatrix} -A_c^T P_c - P_c A_c & C_c^T - P_c B_c \\ C_c - B_c^T P_c & D_c + D_c^T \end{bmatrix} \succeq_K 0 \,, \quad Q \, C_c^T = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.53}$$

The matrix $P_{d_{min}}$ can be then obtained from $P_{c_{min}}$ using (4.39). Unfortunately, from the optimization results, it is not clear which of the two characterization to choose. For the compaction filter application, the two problems perform equally well and we have therefore chosen in the more natural setting of the problem, the discrete time domain. It is important however to keep this dual characterization in mind, first because it might be of use in other applications (one characterization might be more suitable than the other) and second as a prospective tool to develop theoretical results (again, one domain might provide more insight than the other).

Linear phase filters. It is well known that an FIR filter of order N is linear phase if and only if the impulse response is generalized hermitian, i.e., $h(n) = dh^*(N-n)$ for some d with $|d| = 1$ [73]. Since such a constraint will leave $(N+1)/2$ degrees of freedom for odd N and $N/2$ degrees of freedom for even N, we expect a significant decrease in compaction gain. In fact, for $M = 2$, a real FIR linear phase filter with a Nyquist(2) magnitude squared response must take the form $1 + z^{-N}$. To see this, observe that the Nyquist constraint on the magnitude squared response is equivalent to the polyphase components of $H(z)$ being power complementary, i.e., $|E_0(e^{j\omega})|^2 + |E_1(e^{j\omega})|^2 = 2$. Using the fact that an FIR filter with a Nyquist(2) magnitude squared response has to have odd order, we can easily conclude that $E_1(e^{j\omega}) = \tilde{E}_0(e^{j\omega})$. Therefore, $E_0(z)$ and $E_1(z)$ have the same magnitude responses which implies that $|E_0(e^{j\omega})|^2 = |E_1(e^{j\omega})|^2 = 1$. Using the assumption that the filter is real FIR, we obtain the above form. In general, only approximate linear phase solutions are feasible (see [60] for $M = 2$).

Adding regularity constraints. The regularity property is important in wavelet applications and consist of forcing L zeros at the aliasing frequencies $\omega = 2\pi m/M$ for $1 \leq m < M$. For example, for $M = 2$, this amounts to forcing r zeros at $z = -1$. The first of these zeros ($r = 0$) is simply obtained from $F(-1) = 0$ (because $F(e^{j\omega}) \geq 0 \quad \forall \, \omega$, there will actually be a double zero at π). The second zero $r = 1$ is obtained by differentiating $F(e^{j\omega})$ twice with respect to ω, evaluating the result at π and setting it to zero. Repeating this procedure, we can easily derive the following set of equations:

$$D_d - C_d(A_d + I)^{-1} B_d = 0, \quad r = 0 \tag{4.54}$$

$$C_d \begin{bmatrix} (2N+1)^{2r} \\ \vdots \\ (2k+1)^{2r} \\ \vdots \\ 3^{2r} \end{bmatrix} = 0, \quad 1 \leq r < L \tag{4.55}$$

For general M, the regularity condition can be expressed as the following linear constraint on the filter product coefficients [51] :

$$2(-4\pi^2)^{2r} \sum_{n=1}^{N} f(n) n^{2r} \cos(\frac{2\pi nm}{M}) = \delta(r), \quad 0 \leq r < L, \quad 1 \leq m < M \tag{4.56}$$

The current versions of the running programs assume the so called Slater conditions (existence of a strict feasible primal or dual). Unfortunately, adding the regularity constraint seem to violate those conditions indicating either an infeasible solution (which would be an interesting result) or more likely, a non strict primal and/or dual solution. Optimality conditions for such situations have been already derived [56] but have not been incorporated in any software package yet. Because we strongly believe that this is just a matter of time, we have decided to keep this small paragraph on how to add regularity constraints.

4.5.1 Semidefinite programming

Before going through the numerical results and comparisons with previous work, it is important to put into perspective some of the virtues of semi definite programming and more specifically the particular algorithm we are using. We also briefly discuss the effect of certain program parameters as well as the choice of the weight matrix W on the speed of convergence and accuracy of the results. A thorough review of the material is not attempted here and the reader is referred to the appropriate references for more details. Broadly speaking, semi definite programming can be regarded as an extension of linear programming where the component wise inequalities between vectors are replaced by matrix inequalities (generalized inequalities). The problem is however non linear (convex) and has the following general form:

$$\begin{aligned} &\min_{y} \quad \mathbf{e}^{\mathrm{T}} y \\ &\text{subject to} \quad G(y) \succeq_K 0, \quad Sy = q \end{aligned} \tag{4.57}$$

where $G(y) = G_0 + \sum_{i=1}^{M} y_i G_i$. The problem data are the vector $\mathbf{e}^{\mathrm{T}} \in R^M$ and the $M+1$ symmetric matrices $G_0, G_1, \ldots, G_{M+1} \in R^{N \times N}$. The variable $y \in R^M$ in (4.57) is a vectorized version of C_d and the matrix P_d, i.e., the stacked vector $y = [p_{d_{11}} \ldots p_{d_{NN}} f(N) \ldots f(1)]^T$. The vector $\mathbf{e}^{\mathrm{T}} =$

$[1\ 0\ldots\ 1\ 0\ldots\ 1\ \mathbf{R}]$. We can also obtain expressions for the matrices $G_i, i = 0, \ldots, M$, the matrix S and the vector q. The total number of variables in our problem is equal to $M = N(N+1)/2 + N$. The *dual* problem associated with the *primal* semi definite program (SDP) (4.57) is also a semi definite program in the form :

$$\begin{aligned} & \max_{z,Z} \quad \mathbf{q}^{\mathrm{T}} z - Tr(F_0 Z) \\ \text{subject to} \quad & (S^T z)_j + Tr(F_j Z) = t_j,\ j = 1, \ldots, M \\ & Z \succeq_K 0 \end{aligned} \tag{4.58}$$

where Z is a symmetric matrix $\in R^{N \times N}$ and z is an $N \times 1$ vector. Perhaps the most important advantage of semi definite programs is that they can be solved very efficiently both in theory and practice. Theoretical efficiency can be proved from the convexity of a SDP together with the fact that we can construct, in polynomial time, a cutting plane for the constraint set through any given infeasible point and then solve the primal problem in polynomial time using the ellipsoid method. The Ellipsoid algorithm is however slow and practical efficiency is actually due to the fact that an SDP can be solved using interior point methods [53]. Currently, a wide variety of papers as well as software packages for solving SDP based on different interior point methods is available (See for example [83, 3] and the references therein as well as the excellent book which has brought to our attention the availability of efficient software that handle LMI constraints [9]). We have used two different programs in this chapter: the first one is the one written by Vandenberghe and Boyd [81] which uses a particular primal-dual interior point method. The second one uses the MATLAB LMI toolbox that implements the projective algorithm of Nesterov and Nemirovski [53]. Primal and dual interior point methods solve simultaneously problems (4.57) and (4.58) by generating a sequence of primal and dual feasible points y^k and Z^k. Among other benefits (see [83, 82]), a primal-dual algorithm possesses an *extremely* important property, namely the following: Because the program is minimizing the duality gap (the difference between the primal and dual solutions), it knows *apriori* that the optimum should be zero (assuming that the Slater conditions are satisfied). Therefore, the program can monitor the duality gap and according to a specified gap tolerance, say $\epsilon = 10^{-4}$ decides to terminate. The *elegance* is that in this case, this given accuracy of 10^{-4} is an upper bound of how far we are from the *true* optimum. In other terms, the primal and dual program *guarantee* upon terminating that we are ϵ sub-optimal and therefore explicitly allows the user to control numerical sub-optimality. Contrast this with other methods that solve either the primal or dual problem. These methods terminate after a certain number of iterations without really knowing how good the obtained result is compared to the true *unknown* optimum. For more theoretical background, the reader is referred to section 5 in [83].

4.5.2 The MATLAB programs

As we have mentioned above, we have written two alternative programs that are currently running well. The first program uses the MATLAB LMI control toolbox while the other is based on the SP package written by Vandenberghe and Boyd [81]. In order to remain in the MATLAB environment, the program of Vandenberghe and Boyd needs to be used with the the SDP solver/parser written by Shao-Po Wu and Stephen Boyd [89]. The two programs are actually lumped into a single package called SDPSOL. The package is extremely easy to learn and use, can be called from within MATLAB and the program input is specified in the convenient form described by (4.50) and (4.51). The SDPSOL program with the corresponding documentation is generously made available by Professor Stephen Boyd at *http://www-isl.stanford.edu/people/boyd/SDPSOL.html*. All the energy compaction filter design programs as well as the corresponding documentation can be found in our home page *http://www.systems.caltech.edu/tuqan*.

4.6 Numerical results

All design examples described below were obtained using the MATLAB LMI control toolbox program. The user should get the exact same result if using the SDPSOL package.

Example 2. AR(1) process. Assume that the input $x(n)$ is a zero mean AR(1) process with an autocorrelation sequence in the form $R_{xx}(k) = \rho^{|k|}$ where $0 < \rho < 1$. Let $M = 2$. The optimum compaction gain curves for $N = 2$ and 3 as a function of ρ are shown in Fig. 4.4. The curve for $N = 3$ coincide with the theoretical compaction gain formula $G_{comp}(2,3) = 1 + \dfrac{2\rho}{\sqrt{3+\rho^2}}$ derived in [43]. The precise difference is actually in the order of 10^{-5}. The last curve denotes the compaction gain when $N = \infty$ (ideal low pass filter case). A closed form expression for the compaction gain can be obtained by evaluating the integral in (4.1) using the fact that the integrand is a Poisson kernel [13, page 308]. The final result is $G_{comp}(2,\infty) = \dfrac{4}{\pi} \arctan \dfrac{1+\rho}{1-\rho}$. From Fig. 4.4, it is therefore very clear that for an AR(1) process, we do not loose much by using short filters. Assume now that $\rho = 0.9$, $N = 3$, $M = 2$ and set $d = 10^{-6}$. The coefficients of the theoretical optimum $F_{th}(z)$ (obtained from [43]) are displayed in the first column of Table 4.1. The coefficients of $F_{opt}(z)$ obtained through the newly proposed method are given in the second column whereas the coefficients resulting from convolving $h_{min}(n)$ with its flipped version can be found in the third column. The compaction gains corresponding to each case are also illustrated. As the numbers clearly show, we are extremely close to the theoretical optimum. The minimum phase filter in this case is given by:

$$H_{min}(z) = 0.49390514439923 + 0.82792470898754z^{-1} + 0.22818468063421z^{-2} - 0.13612540851745z^{-3}$$

The positivity of $F(z)$ is demonstrated by the double roots shown in the Z-plane plot of Fig. 4.5(a).

The spectral factorization is also quite accurate. The compaction gain remains almost the same and the positivity of $H_{min}(z)$ is not lost as we can clearly see from Fig. 4.5(b).

Example 3. Low pass FIR filter design. As we pointed out in the introduction, we can also design FIR filters with maximum passband energy using the new proposed approach. Assume therefore that $W(e^{j\omega}) = rect(\omega/\omega_c)U(e^{j\omega})$ where $U(e^{j\omega})$ is a positive function of ω. The input autocorrelation sequence is therefore a weighted sinc function. With $M = 2$, $\omega_c = 0.45$, $U(e^{j\omega}) = 1$, $R_{xx}(k) = \sin\frac{0.45\pi k}{\pi k}$, the resulting low pass filters of order $N = 7, 17$ and 27 are shown in Fig. 4.6. Again, the positivity of the resulting $F(z)$ is shown in Fig. 4.7 for the case of $N = 27$. The coefficients of the minimum phase low pass filters together with the corresponding passband energy (compaction gain) can be found in Table 4.2. The above technique can be easily extended for the design of high pass and multiband FIR filters.

Example 4. Multiband AR(5) process. Assume that the input $x(n)$ is a zero mean multiband AR(5) process with a spectrum shown in Fig. 4.8. The magnitude squared responses of the resulting optimum compaction filters are shown in Fig. 4.9 for $N = 7$, 17 and 27 and $M = 2$. The corresponding compaction gains are 1.52434653018097, 1.56325346934405 and 1.57478692527542. Similarly, the magnitude squared responses of the resulting optimum compaction filters with $N = 7$, 17 and 27 and $M = 3$ are shown in Fig. 4.10. In this case, the compaction gains are 1.86679555840553, 2.00733453932405 and 2.04490921478611. A plot of G_{comp} as a function of M is shown in Fig. 4.11 clearly indicating a non monotonic behavior. The KLT gain is the compaction gain obtained from a filter of order $N < M$. In this case, the filter is the eigen vector corresponding to the largest eigen value of the Toeplitz symmetric autocorrelation matrix with first row $[R(0)R(1)\ldots R(N)]$. The maximum eigen value (compaction gain) serves therefore as an upper bound on the compaction gain. The overall incremental behavior of the compaction gain should be intuitively acceptable because as M increases with N fixed, the constraints on the filter coefficients become less stringent. This is actually quite clear when kM is increased to $(k+1)M$ and in the extreme case ($N < M$) where the Nyquist constraint reduces to a unit energy constraint.

Example 5. A "noisy" version of the compaction filtering problem. Consider the scheme shown in Fig. 4.12 which describes a similar scheme as in Fig. 4.3 except that the original input input $x(n)$ is replaced by its noisy version $z(n) = x(n) + q(n)$ where $q(n)$ is additive noise.

In this example, we will assume that the signal $x(n)$ and the noise $q(n)$ are uncorrelated. Each sequence is individually a zero mean WSS process with corresponding power spectrums $S_{xx}(e^{j\omega})$ and $S_{qq}(e^{j\omega})$. Given the above set up, we would like to optimize the FIR filter $H(z)$ such that the Signal to noise ratio is maximized. In general, this is not an FIR compaction filtering problem because the noise power at the output needs to be minimized. It is also not an FIR Wiener filtering problem because the filter is trying to "approximate" the process $x(n)$ in the least squared sense. Furthermore, the magnitude response of the filter is constrained to be Nyquist(M). We will consider two different

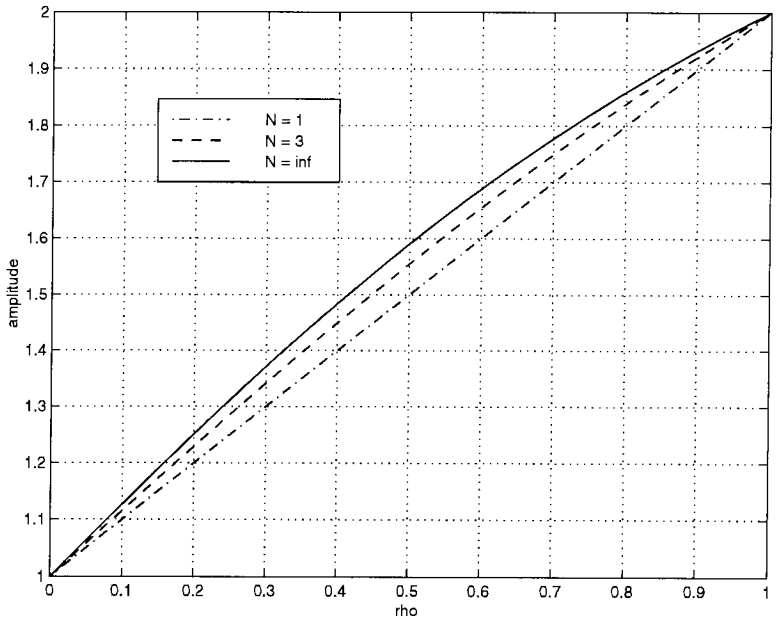

Fig. 4.4: Compaction gain curves for an AR(1) process for $N = 2, 3$ and ∞ with $M = 2$.

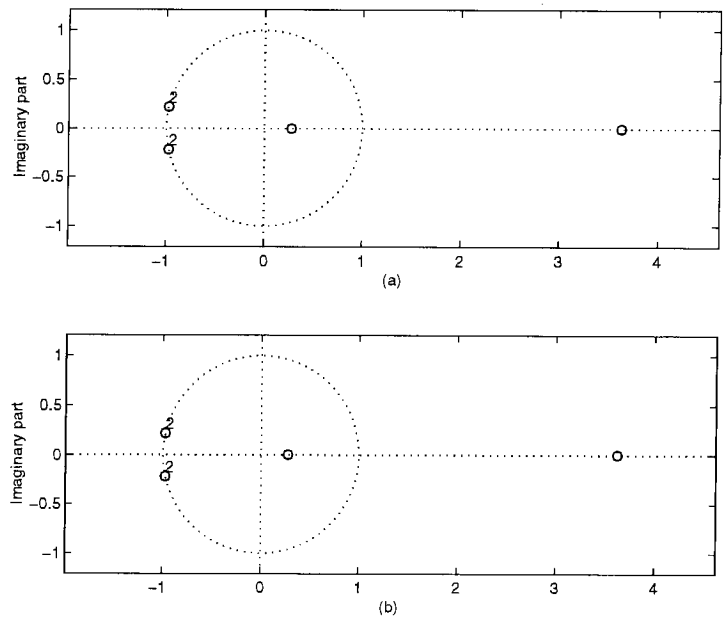

Fig. 4.5: Double roots on the unit circle indicating the positivity of the product filter $F(z)$ (a) as the output of the program (b) as a result of convolving $h_{min}(n)$ with its flipped version.

$\rho = 0.9$

n	$F_{th}(z)$	$F_{opt}(z)$	$H_{min}(z)H_{min}(z^{-1})$
-3	-0.06723300781871	-0.06723303955021	-0.06723303955021
-2	0	0	-0.00000000159429
-1	0.56677425591171	0.56677428160920	0.56677427538491
0	1	1	0.99999999073630
1	0.56677425591171	0.56677428160920	0.56677427538491
2	0	0	-0.00000000159429
3	-0.06723300781871	-0.06723303955021	-0.06723303955021
compaction gain	1.92216793524141	1.92216793523235	1.92216792144589

Table 4.1: The product filter coefficients with the corresponding compaction gains for an AR(1) process with $\rho = 0.9$. The filter order is $N = 3$ and the number of channels is $M = 2$.

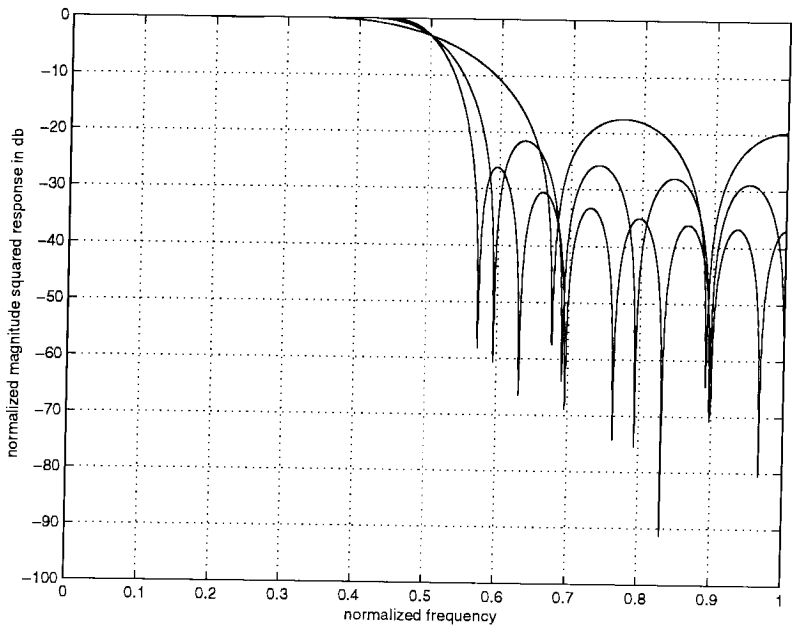

Fig. 4.6: Normalized magnitude squared responses for the low pass filters of orders $N = 7$, 17 and 27 with $M = 2$.

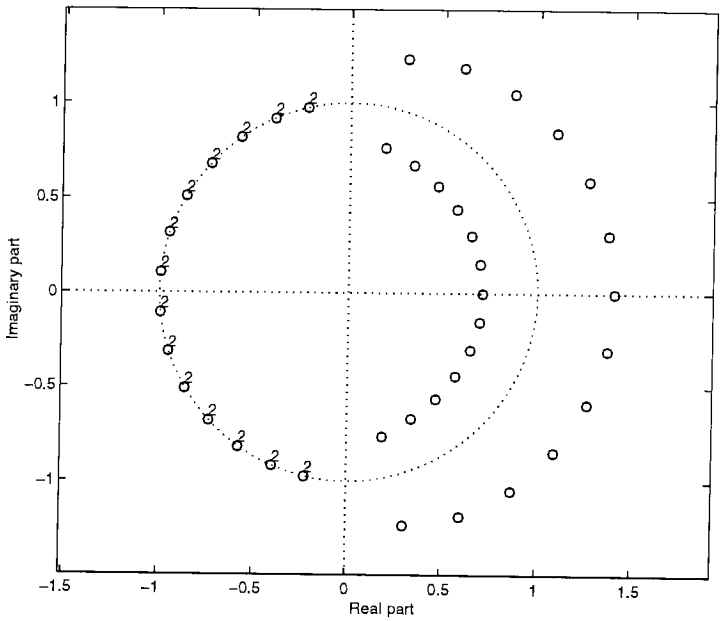

Fig. 4.7: Zeros of the product filter $F(z)$ with $N = 27$. The zeros of $H_{min}(z)H_{min}(z^{-1})$ are exactly the same.

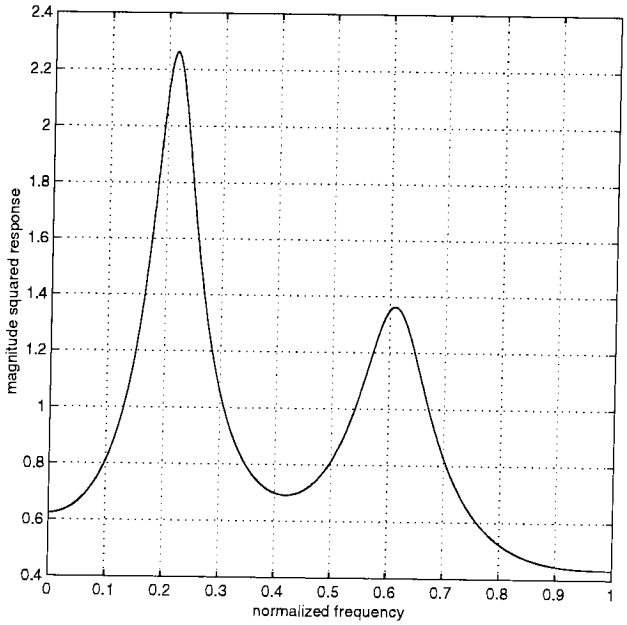

Fig. 4.8: The AR(5) multiband power spectrum

n	N = 7	N = 17	N = 27
0	0.43774556860797	0.27089534996412	0.17914567313110
1	0.74261793188355	0.64109787056304	0.52614310963751
2	0.43526111009482	0.61246074670445	0.66250904342908
3	-0.09987127825081	0.12173710961729	0.31521610704904
4	-0.20098601432762	-0.24788410493153	-0.16824391041704
5	0.04904590185372	-0.12156430184572	-0.23779713024009
6	0.10434139639111	0.13529262755166	0.04997857013165
7	-0.06150536110779	0.09145991062907	0.16997482174936
8	0.00000000000000	-0.08685093858740	-0.01359193635327
9	0.00000000000000	-0.06489278869468	-0.12495706064442
10	0.00000000000000	0.06154843265026	0.00210610415061
11	0.00000000000000	0.04346519555900	0.09413912340052
12	0.00000000000000	-0.04645417334889	0.00047932830380
13	0.00000000000000	-0.02575893986209	-0.07184826421845
14	0.00000000000000	0.03648781694090	0.00027258799740
15	0.00000000000000	0.00948233507099	0.05492121161694
16	0.00000000000000	-0.02843979245180	-0.00214590101366
17	0.00000000000000	0.01202293163233	-0.04158642321392
18	0.00000000000000	0.00000000000000	0.00419088222859
19	0.00000000000000	0.00000000000000	0.03081057810040
20	0.00000000000000	0.00000000000000	-0.00601288715256
21	0.00000000000000	0.00000000000000	-0.02193024780939
22	0.00000000000000	0.00000000000000	0.00750717680080
23	0.00000000000000	0.00000000000000	0.01437861077583
24	0.00000000000000	0.00000000000000	-0.00874725119727
25	0.00000000000000	0.00000000000000	-0.00720886447623
26	0.00000000000000	0.00000000000000	0.00965436737872
27	0.00000000000000	0.00000000000000	-0.00328720102111
compaction gains	1.85122508103806	1.89315244307063	1.89883368304497

Table 4.2: low pass filter coefficients for $N = 7,\ 17$ and 27 with the corresponding compaction gains

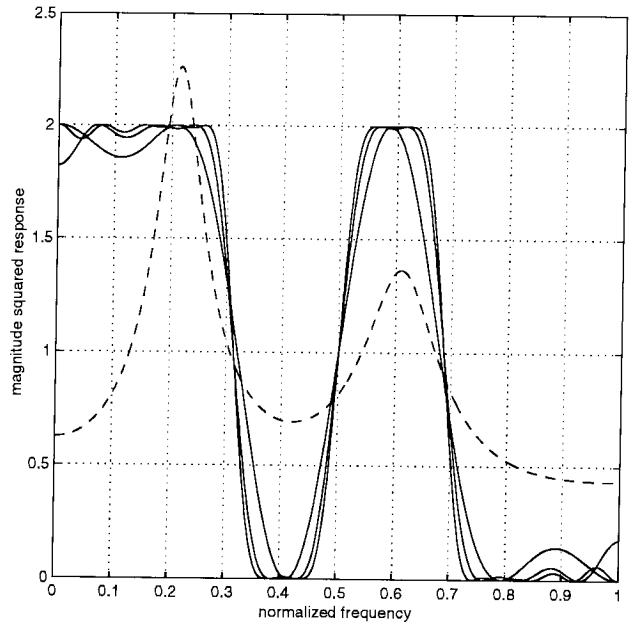

Fig. 4.9: The magnitude squared responses of the optimum compaction filters corresponding to the multiband AR(5) process of order $N = 7,\ 17$ and 27 with $M = 2$.

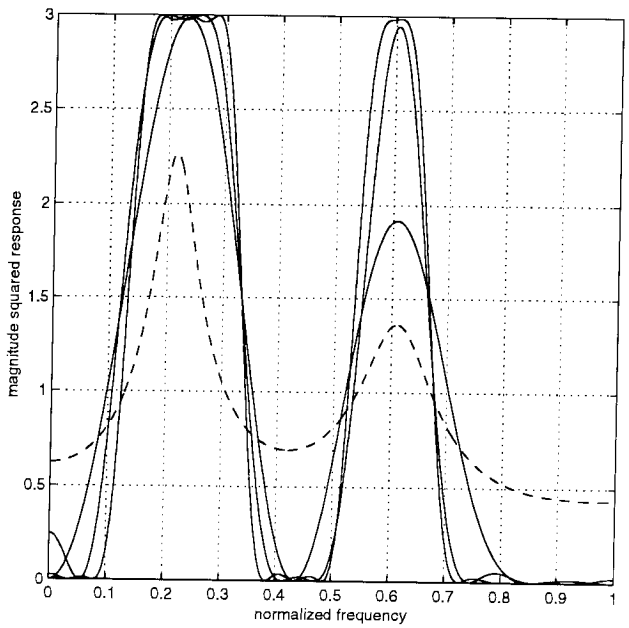

Fig. 4.10: The magnitude squared responses of the optimum compaction filters corresponding to the multiband AR(5) process of order $N = 7,\ 17$ and 27 with $M = 3$.

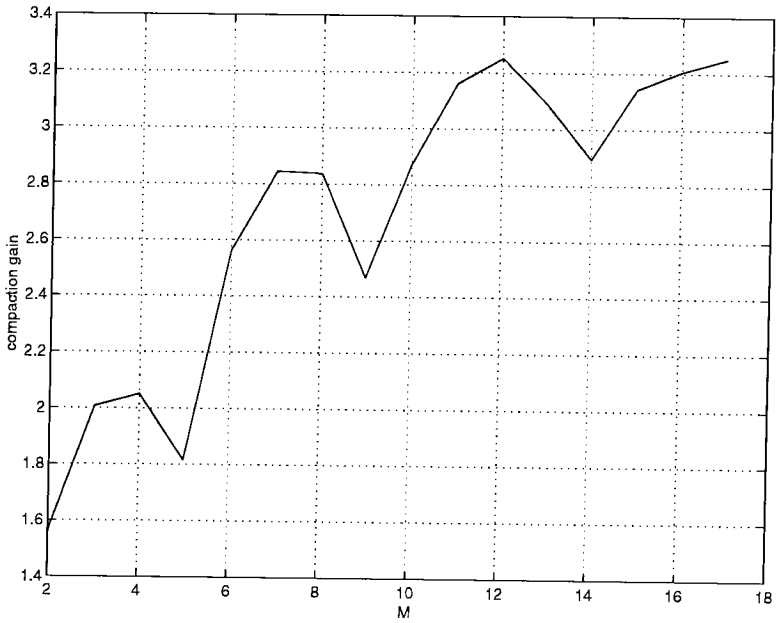

Fig. 4.11: The non-monotone behavior of the compaction gain as a function of the number of channels M with a filter of fixed order $N = 17$.

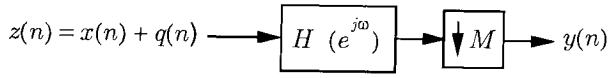

Fig. 4.12: Schematic of the noisy FIR energy compaction problem.

cases depending on the statistical nature of the noise. In the first case, we assume that the noise is white, i.e., $S_{qq}(e^{j\omega}) = \sigma_q^2$ for all ω. Using the linearity of the filter, it can be easily shown that the objective function to be maximized is equal to $C_d\mathbf{R}^T$ subject to the same constraints (4.10) and (4.11). The noisy version in this case simplifies to the original problem and the compaction filter is the optimum filter. Note that this argument holds whether the subband filters are FIR, IIR or ideal. Fig. 4.14 validates the analysis perfectly for the FIR case. The noise source is assumed to be white with variance $\sigma_q^2 = 0.4$, 0.7 and 1 respectively. We have set $N = 3$, $M = 2$ and the input $x(n)$ is an AR(5) process with low pass spectrum. The resulting optimum filter is the same regardless of the noise variance and is equal to the optimum compaction filter. The signal to noise ratio decreases according to the noise increase only. The result is quite intuitive because the whiteness of the noise does not "bias" any frequency region over the other. All frequencies are equally "bad" and the optimum filter tries to reduce the approximation error as much as possible. The second and more interesting case corresponds to colored noise. In this case, the objective function to be maximized is $C_d{\mathbf{R_{xx}}}^T - C_d{\mathbf{R_{qq}}}^T$ subject to the usual constraints. We have performed a similar experiment as in the white noise case, namely, starting with the no noise case, we have gradually increased the noise amplitude. When the noise level is low, the filter response remains very close to the compaction filter one (as one would expect) until a certain point where the response take the multiband shape shown in Fig. 4.13. One explanation might be that the noise power in the passband ($[0, \pi/2]$) becomes too significant and in order for the (constrained) filter to reduce this excessive noise, it "compromises" and amplifies some stopband noise. The total mean squared error (due to both approximation error and output noise variance), as the noise level increases, is given by 0.07667707204384, 0.27978660959974, 0.48275850975035, 0.68557203510958 and 0.83526496202582. We finally note that, for the last case, a low pass filter (almost similar to the ones shown in Fig. 4.13) gives a mean squared error equal to 0.88936649071035 which is higher than the one obtained by the multiband filter.

4.7 A 2-channel IIR optimum compaction filter

The aim of this section is to statistically optimize a two channel orthonormal filter bank when subband quantizers are present at the lowest possible cost. As we have argued in section 4.2, this is equivalent to designing an optimum compaction filter. Two channel orthonormal filter banks are of special interest because they form a basic building block in the design of wavelet transforms. Low cost filters are quite attractive in image and audio coding applications. The requirement for a very efficient two channel system motivates the investigation of filter banks based on IIR filters rather than FIR ones. Furthermore, to the best of our knowledge, previous work on finite order compaction filters and/or finite order two channel optimum orthonormal filter banks has been only dedicated to the FIR case. To meet the above requirements, we propose the optimization of a class of two channel IIR

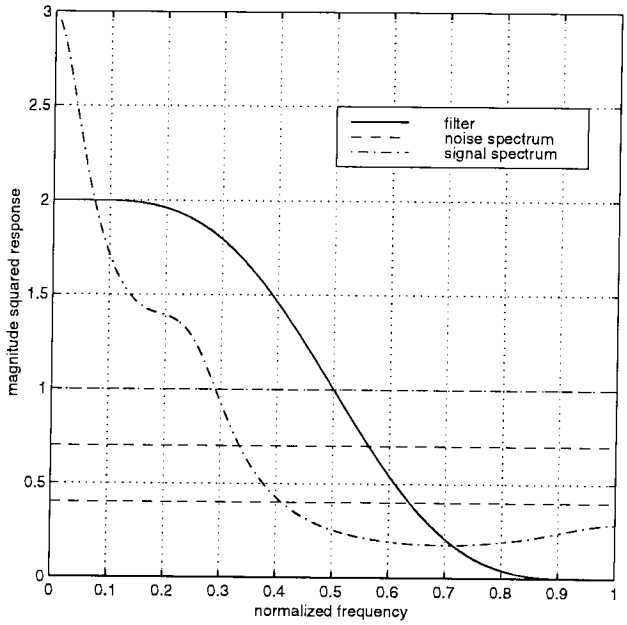

Fig. 4.13: The magnitude squared responses of the optimum filters remain the same in presence of white noise.

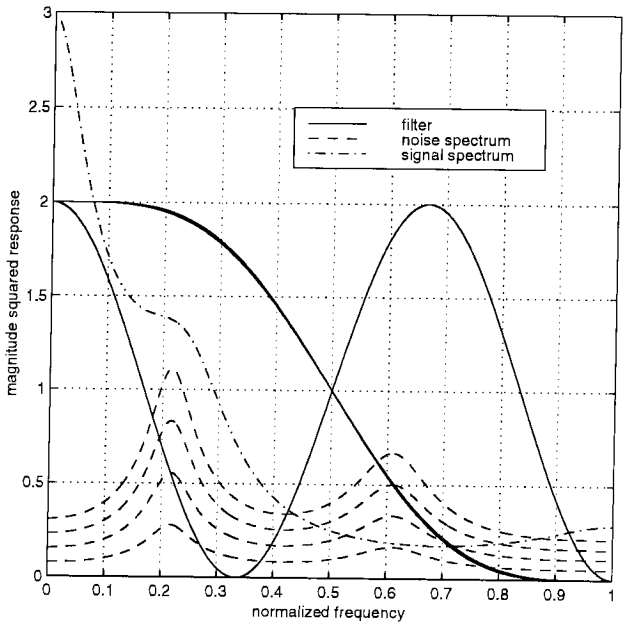

Fig. 4.14: The magnitude squared responses of the optimum filters as the colored noise level is increased.

orthonormal filter banks based on the sum of two all pass filters ([73]). In specific, consider the single coefficient system:

$$H_0(z) = \frac{1}{\sqrt{2}}(z^{-2} + z^{-1}\frac{z^{-2} - \alpha}{1 - \alpha z^{-2}}) \ , \ H_1(z) = H_0(-z) \tag{4.59}$$

where α is real and $-1 < \alpha < 1$. The synthesis filters are given by $F_0(e^{j\omega}) = H_0^*(e^{j\omega})$ and $F_1(e^{j\omega}) = H_1^*(e^{j\omega})$. The two channel system is shown in Fig. 4.15.

Note that since the polyphase components of the analysis filters are stable causal all pass filters, their reciprocals will produce unstable synthesis filters. To overcome this difficulty, Ramstad [57] proposed to implement the inverse filters as anti-causal stable IIR filters. Although (4.59) is a seemingly restrictive case, the proposed form of the filter $H_0(z)$ is a special case of the more general structure introduced recently by Phoong et al. [55]. It has been shown that this type of filter provides several excellent advantages [55]. For example, the filter $H_0(z)$ (and therefore $H_1(z)$) can have a very good frequency response. Furthermore, the special form of the filter assure the existence of a zero at π which can be important for wavelet applications. For the purpose of this paper, our results indicate that for the cases where we can obtain high compaction gain with the special filters in (4.59), using higher order filters does not increase the compaction gain.

4.7.1 The analytical results

Consider the set up shown in Fig. 4.15 where the input signal $x(n)$ is a zero mean wide-sense stationary (WSS) random process with a power spectrum $S_{xx}(e^{j\omega})$. Each subband quantizer, labeled by $\mathcal{Q}$, represents a scalar uniform (PCM) quantizer and is modeled as an additive zero mean white noise source $q(n)$ with variance

$$\sigma_q^2 = c2^{-2b}\sigma_{x_i}^2 \tag{4.60}$$

where σ_q^2 is the quantization noise variance, c is a constant that depends on the statistical distribution of the subband signal $x_i(n)$ and the overflow probability, and $\sigma_{x_i}^2$ is the variance of the ith subband signal. The subband coding problem reduces to finding the optimum coefficient α_{opt} that maximizes the compaction gain (alternatively the subband variance) at the output of one of the subband. The specific form of the analysis filters given in (4.60) guarantee automatically the Nyquist property and transforms the constrained optimization problem into an unconstrained one. A closed form expression for the compaction gain is given next.

Theorem 24 *Consider the scheme of Fig. 4.15 under all the previous filter and quantizer assumptions. The compaction gain at the output of one of the subband filters, say $H_0(z)$, can be expressed as follows:*

$$G_{comp}(2) \quad = \quad 1 + \frac{R_{xx}(1)}{R_{xx}(0)} - (\alpha + \alpha^2)\frac{R_{xx}(1)}{R_{xx}(0)}$$

$$+ \quad \frac{(\alpha - \alpha^3)}{R_{xx}(0))} \sum_{n=0}^{\infty} \alpha^n R_{xx}(2n+3) \tag{4.61}$$

The proof is a straightforward algebraic derivation and is omitted. The infinite summation in (4.61) is the result of the IIR nature of the filter bank. The above equation was written specifically in the above form in order to emphasize the following points: First, when α is equal to zero, the compaction gain is equal to $1 + R_{xx}(1)/R_{xx}(0)$, which is simply the 2×2 KLT compaction gain. This indeed makes sense since the structure of Fig. 4.15 reduces to the 2×2 universal KLT. Second, when the input signal is white noise, i.e. $R_{xx}(k) = \delta(k)$, the compaction gain is equal to one. Finally, observe that the above equation involves only the odd samples of the autocorrelation sequence $R_{xx}(k)$, due to the Nyquist constraint on $|H_0(e^{j\omega})|^2$. Therefore, if the input signal $x(n)$ is such that its power spectrum $S_{xx}(z)$ takes the form $S(z^2)$, the compaction gain is equal to unity.

The goal now is to find the optimum coefficient α_{opt} that maximizes (4.61). In general, it is difficult to obtain analytical solutions due to the complexity of the expression in (4.61). We will therefore present analytical solutions for the optimum coefficient α_{opt}, the compaction gain $G_{comp}(2)$ and the coding gain $\mathcal{G}_{SBC}(2)$ only for specific examples of the input $x(n)$ such as the MA(1) and AR(1) processes. For a general random process $x(n)$, the optimum coefficient α_{opt} is obtained numerically through a MATLAB program.

Example 6. *MA(1) process.* Assume that the input $x(n)$ is a zero mean MA(1) process with an autocorrelation sequence in the form

$$R_{xx}(k) = \begin{cases} 1 & k = 0 \\ \theta/1+\theta^2 & k = 1, -1 \\ 0 & \text{otherwise} \end{cases} \tag{4.62}$$

where θ is between -1 and 1. It can be shown that, for this case,

$$\alpha_{opt} = -0.5, \quad G_{comp}(2) = 1 + 5R_{xx}(1)/4R_{xx}(0), \quad \mathcal{G}_{SBC}(2) = \frac{1}{\sqrt{1 - 25R_{xx}^2(1)/16R_{xx}^2(0)}} \tag{4.63}$$

It is interesting to note that the optimum coefficient α_{opt} is independent of the signal statistics.

Example 7. *AR(1) process.* Assume now that the input $x(n)$ is a zero mean AR(1) process with an autocorrelation sequence in the form $R_{xx}(k) = \rho^{|k|}$ where ρ is between -1 and 1. It can also be shown that, for this case,

$$\begin{aligned} \alpha_{opt} &= (1 - \sqrt{(1+\rho^2)})/\rho^2 \\ G_{comp}(2) &= 1 + \rho + \rho(1-\rho^2)\alpha_{opt}^2 \\ \mathcal{G}_{SBC}(2) &= \frac{1}{\sqrt{1 - \rho^2(1 + \alpha_{opt}^2(1-\rho^2))^2}} \end{aligned} \tag{4.64}$$

We note that the optimum coefficient in this case is independent of the sign of ρ, is always negative and between $1-\sqrt{2}$ and -0.5 (the case where $\rho = 0$). Furthermore, one can show that there is a negligible loss of compaction gain even when α_{opt} is implemented using very small number of binary shift and add operations. As ρ approaches unity, the scheme is asymptotically equivalent to the 2×2 universal KLT.

4.7.2 Examples for more general inputs

We give several examples where the optimum coefficient α_{opt} is computed numerically through a MATLAB program. The program uses the compaction gain expression derived previously with input $R_{xx}(k)$ and output α_{opt}. Although written in MATLAB, the program converges in fractions of a second. This is an order of magnitude faster than previously described techniques used to design high order FIR compaction filters to achieve similar compaction gain. We vary the input $x(n)$ such that the power spectral density shape spans a variety of choices from "low pass" to multiband with energy concentrated in a specific region to multiband with more even energy distribution. The magnitude squared response of the optimum IIR compaction filter together with the ideal optimum compaction filter magnitude squared response and the input power spectral density are shown in Fig. 4.16, Fig. 4.17, Fig. 4.18 and Fig. 4.19. We adopt the following convention for all the plots : the **solid curve** denotes the input power spectral density, the **dash-dot** curve denotes the magnitude squared response of the optimum IIR compaction filter and the **dashed curve** represents the magnitude response of the ideal optimum compaction filter. This last curve is obtained by optimizing an FIR filter with order equal to 65 using a linear programming approach.

4.8 Conclusion

In this chapter, we proposed a new approach for the design of an FIR filter of arbitrary order N subject to a Nyquist(M) constraint on its magnitude squared response. The problem is quite important because of its appearance in a wide variety of applications. Unlike the standard discretization methods used to solve this type of problems, the positivity of the product filter is guaranteed over all frequencies. Furthermore, because of the convexity of the new formulation, the resulting filter is the global optimum. The method will work for all M and any input sequence $\mathbf{R}$ (not necessarily a positive definite one). Finally, an additional spectral factorization step is not required to obtain $H(z)$. We provide the reader with a qualitative comparison between most of the previous methods (at least the ones we are aware of) in Table 4.3.

To put our results in the correct context, we would like also to point out the two *current* disadvantages of the state space approach: First, there is the question of how to choose the weight scalar d or even the weight matrix W in the multi-objective formulation. Although there is no closed form

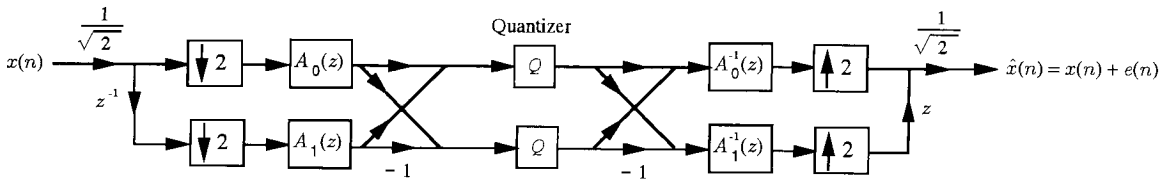

Fig. 4.15: The class of two channels IIR filter banks under consideration

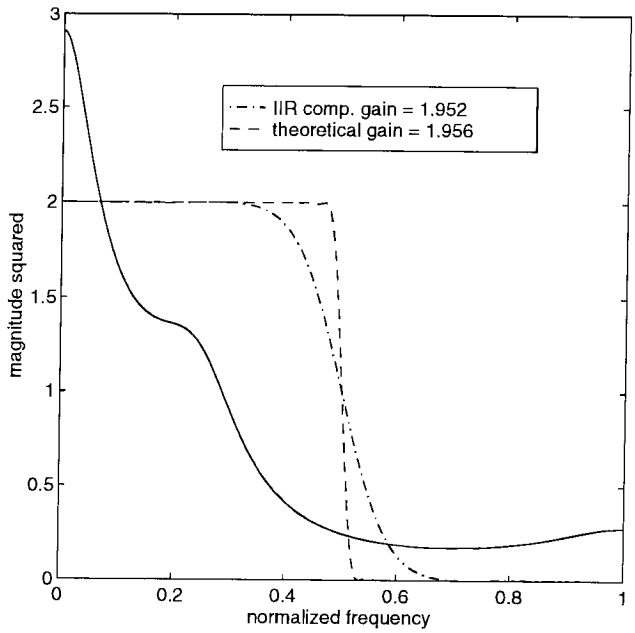

Fig. 4.16: Case of a low pass AR(5) process with IIR FB coding gain = 5.10 db

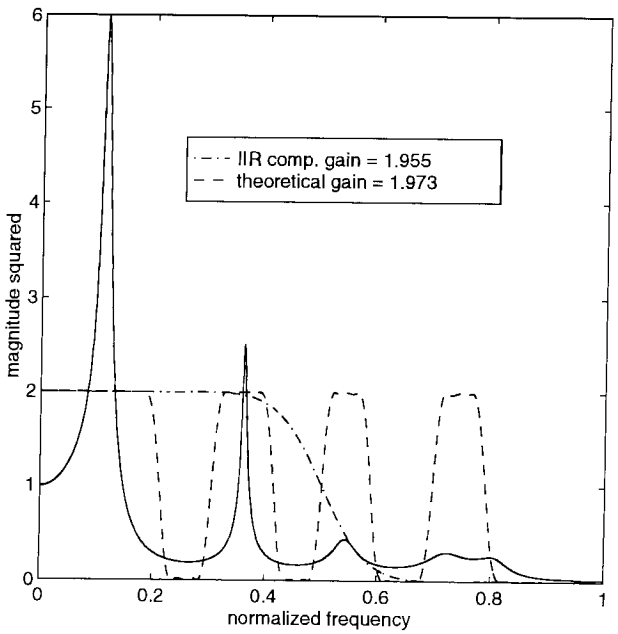

Fig. 4.17: Case of a multiband AR(12) process with IIR FB coding gain = 5.14 db

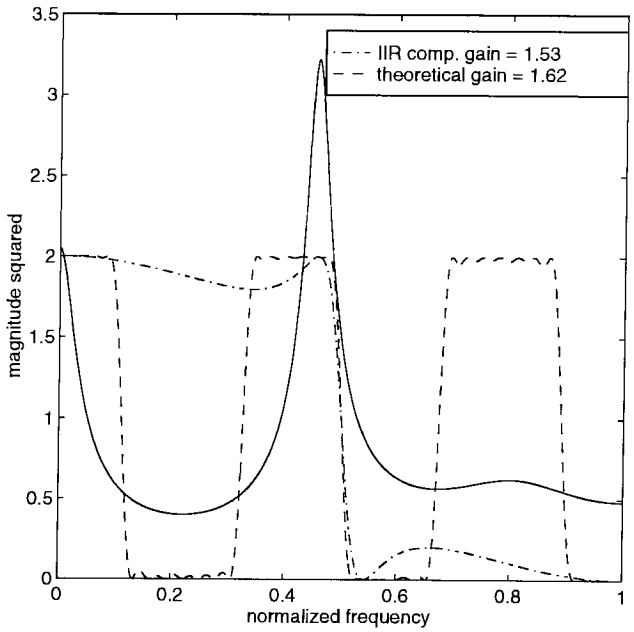

Fig. 4.18: Case of a multiband AR(10) process with IIR FB coding gain = 0.67 db

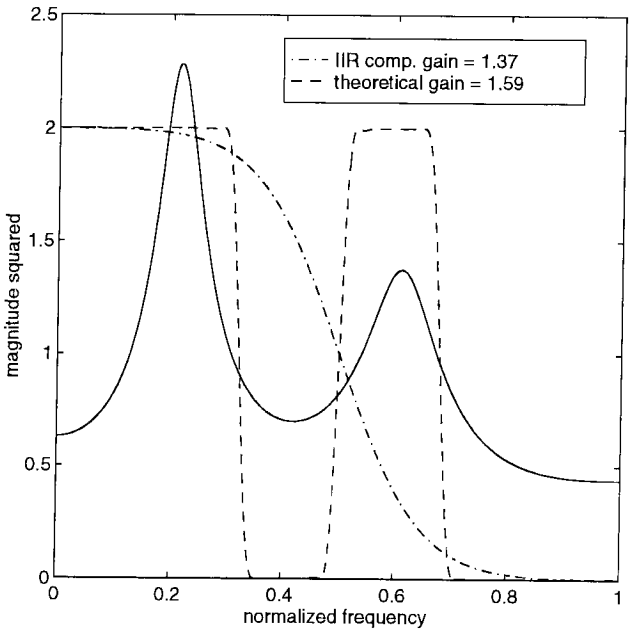

Fig. 4.19: Case of a multiband AR(5) process with IIR FB coding gain = 0.34 db

authors	Approach	M	objective function	problem type	solution	Nyquist constraint	Positivity constraint	spectral fact.
[10],[84],[39]	QCQP	2	compaction filter	non linear non convex	local optimum	Yes	No	No
[11]	QCQP	arbitrary	ideal low pass filter	non linear non convex	local optimum	Yes	No	No
[71],[90],[91]	FIR Lattice	2	compaction filter	non linear non convex	local optimum	No	No	No
[72]	eigen filter [a]	arbitrary	ideal low pass filter	iterative [b]	suboptimum	Yes	No	Yes
[42]	window method [a]	arbitrary	compaction filter	iterative [b]	suboptimum	Yes	Yes	Yes
[52],[54]	product filter	2	compaction filter	linear convex	global optimum [c]	Yes	Yes	Yes
[1]	analytical	2	ideal low pass filter	Non iterative	global optimum	Yes	No	Yes
[43]	analytical	2	compaction filter	non iterative	global optimum [d]	Yes	No	Yes
New method	state space	arbitrary	compaction filter	non linear convex	global optimum	Yes	Yes	No

[a] The eigen filter and window methods are special cases of the product filter approach where the product filter is assumed to be a cascade of two filters

[b] Since the product filter is assumed to be a cascade of two filters, the optimization procedure alternate between the two in an iterative manner

[c] Over only the defined discrete set of frequencies

[d] Only for a certain class of random processes

Table 4.3: Qualitative comparison between the different FIR design methods

expression or formal proof that our chosen weight function will work for *any* sequence $\mathbf{R}$, we have not experienced any problems in all the simulations we have performed. In fact, once we settled on the choice of $W = I$ and $d = 10^{-6}$, we did not worry about changing it again. The second disadvantage is that the algorithm becomes extremely slow for filter orders larger than 20. From a practical point of view, other suboptimum proposed algorithms like the semi-infinite programming method [52] and/or the window method [43] might be of use. The main point we would like to establish is that our approach is sound. The LMI control toolbox implements state of the art interior point LMI solvers which are significantly faster than classical convex optimization algorithms but still not fast enough to compete with discretization techniques. Nevertheless, research on LMI optimization is currently quite active and substantial speed-ups can be expected in the future.

The two channel IIR optimum filter bank provides some good advantages that are not automatically available in the FIR case. First, the Nyquist property is satisfied automatically because of the special form of the filters. Second, in the FIR compaction filter design, although we have eliminated the usual spectral factorization step to obtain $H(e^{j\omega})$, the exact accuracy of the coefficients of $H(e^{j\omega})$ depends on the accuracy of $P_{d_{min}}$ which, in turn, is a function of the weight d. *In this particular IIR scheme, these concerns do not exist* since $H_0(z)$ is directly found. Third, the form of the filters assure the existence of a zero at π which can be important for some wavelet applications. Fourth, the filters have only one coefficient which can be quantized without a major sacrifice in compaction gain. Finally, the compaction gain obtained is high and very close to two (ideal case) for low pass spectrums, high pass spectrums and certain cases of multiband spectrum. The only weakness of the 2-channel filter bank under consideration is its poor performance for the case of general multiband spectrums. This is mainly due to the monotone property of the phase of an all pass function. For such cases, α_{opt} can be set to zero to obtain the 2×2 universal KLT.

Appendix A. Proof of the existence of a spectral factor

Substituting equations (4.15-4.17) in the expression of $F(z)$, we have:

$$\begin{aligned}
F(z) &= D(z) + D(z^{-1}) \\
&= D_d + D_d^T + C_d(zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}C_d^T \\
&= W^TW + B_d^TPB_d + (B_d^TPA_d + W^TL^T)(zI - A_d)^{-1}B_d \\
&+ B_d^T(z^{-1}I - A_d^T)^{-1}(A_d^TPB_d + LW) \\
&= W_d^TW_d + B_d^TP_dB_d + W_d^TL_d^T(zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}LW \\
&+ B_d^TP_dA_d(zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}A_d^TP_dB_d \\
&= W_d^TW_d + W_d^TL_d^T(zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}L_dW_d
\end{aligned}$$

$$
\begin{aligned}
&+ \quad B_d^T[P_d + P_d A_d(zI - A_d)^{-1} + (z^{-1}I - A_d^T)^{-1}A_d^T P_d]B_d \\
&= \quad W_d^T W_d + W_d^T L_d^T (zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}L_d W_d B_d^T(z^{-1}I - A_d^T)^{-1} \\
&\quad [(z^{-1}I - A_d^T)P_d(zI - A_d) + (z^{-1}I - A_d^T)P_d A_d + A_d^T P_d(zI - A_d)](zI - A_d)^{-1}B_d \\
&= \quad W_d^T W_d + W_d^T L_d^T (zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}L_d W_d \\
&+ \quad B_d^T(z^{-1}I - A_d^T)^{-1}[P - A_d^T P A_d](zI - A_d)^{-1}B_d \\
&= \quad W_d^T W_d + W_d^T L_d^T (zI - A_d)^{-1}B_d + B_d^T(z^{-1}I - A_d^T)^{-1}L_d W_d \\
&+ \quad B_d^T(z^{-1}I - A_d^T)^{-1}L_d L_d^T (zI - A_d)^{-1}B_d \\
&= \quad (W_d^T + B_d^T(z^{-1}I - A_d^T)^{-1}L_d)(W_d + L_d^T(zI - A_d)^{-1}B_d)
\end{aligned}
$$

Appendix B. Proof of the discrete time minimum phase spectral factor form

Starting with $D_c + C_c(sI - A_c)^{-1}B_c$ and letting $s = (z-1)/(z+1)$, we get :

$$
\begin{aligned}
D_c + C_c((z-1)(z+1)^{-1}I - A_c)^{-1}B_c &= D_c + C_c(z+1)((z-1)I - (z+1)A_c)^{-1}B_c \\
&= D_c + C_c(z+1)(z(I - A_c) - (I + A_c))^{-1}B_c \\
&= D_c + C_c(z(I - A_c) - (I + A_c))^{-1}B_c \\
&+ D_c + C_c((I - A_c) - z^{-1}(I + A_c))^{-1}B_c
\end{aligned}
$$

By applying the matrix inversion lemma [40] on $((I - A_c) - z^{-1}(I + A_c))^{-1}$, we get :

$$
\begin{aligned}
&\quad D_c + C_c[z(I - A_c) - (I + A_c)]^{-1}B_c \\
&+ \quad C_c\{(I - A_c)^{-1} - (I - A_c)^{-1}[(I + A_c)(I - A_c)^{-1} - zI]^{-1}(I + A_c)(I - A_c)^{-1}\}B_c \\
&= \quad D_c + C_c(I - A_c)^{-1}B_c + C_c[z(I - A_c) - (I + A_c)]^{-1}B_c \\
&+ \quad C_c\{(I - A_c)^{-1}[zI - (I + A_c)(I - A_c)^{-1}]^{-1}(I + A_c)(I - A_c)^{-1}\}B_c \\
&= \quad D_c + C_c(I - A_c)^{-1}B_c + C_c[z(I - A_c) - (I + A_c)]^{-1}B_c \\
&+ \quad C_c[z(I - A_c) - (I + A_c)]^{-1}(I + A_c)(I - A_c)^{-1}B_c \\
&= \quad D_c + C_c(I - A_c)^{-1}B_c + C_c[z(I - A_c) - (I + A_c)]^{-1}[I + (I + A_c)(I - A_c)^{-1}]B_c \\
&= \quad D_c + C_c(I - A_c)^{-1}B_c + C_c[z(I - A_c) - (I + A_c)]^{-1}[2(I - A_c)^{-1}]B_c \\
&= \quad D_c + C_c(I - A_c)^{-1}B_c + C_c[z - (I - A_c)^{-1}(I + A_c)]^{-1}2(I - A_c)^{-2}B_c
\end{aligned}
$$

which can be put in the form $W_d + L_d(sI - A_d)^{-1}B_d$ by the choice (4.30).

Appendix C. Proof of minimality

Recall that a state space realization is minimal if and only if it is jointly observable and controllable. Assuming the minimality of the triple $\{A_c,\ B_c,\ C_c\}$, we use the (PBH) test [40, pages 135–136] to prove the minimality of the triple $\{A_d,\ B_d,\ C_d\}$ given by (4.30). In particular, since $(A_c,\ B_c)$ is controllable, then, there does not exist a row vector $q \neq 0$ such that $qB_c = 0$ and $qA_c = \mu q$. Now, assume that $(A_d,\ B_d)$ is not controllable. Then, there exists a row vector $x \neq 0$ such that $xB_d = x(I - A_c)^{-2}B_c = 0$ and $x(I - A_c)^{-1}(I + A_c) = \lambda x$. Let $y = x(I - A_c)^{-2}$. Then, $yB_c = 0$ and $y(I - A_c)(I + A_c) = \lambda y(I - A_c)^2$. By observing that the matrices $(I - A_c)(I + A_c)$ commute, the last expression therefore simplifies to $y(I + A_c) = \lambda y(I - A_c)$. This in turn implies that $yA = \dfrac{(\lambda - 1)}{(\lambda + 1)}y$. If $\lambda = -1$, then, $y = x = 0$ which is a contradiction. If $\lambda \neq -1$, then the assumption that $(A_d,\ B_d)$ is not controllable implies that $(A_c,\ B_c)$ is also not controllable, which is again a contradiction. The observability of $(A_d,\ C_d)$ can be established in a similar way.

Appendix D. Simplifying equation (4.37)

It is not difficult to see that by multiplying the matrices in (4.37), we get the following matrix inequality:

$$\begin{bmatrix} -P_cA_c - A_c^TP_c & C_c^T - P_cB_c - (P_cA_c + A_c^TP_c)(I + A_d)^{-1}B_d \\ C_c - B_c^TP_c - B_d^T(I + A_d^T)^{-1}(P_cA_c + A_c^TP_c) & X \end{bmatrix} \succeq_K 0$$

where

$$\begin{aligned} X &= D_c + D_c^T + (C_c - B_c^TP_c)(I + A_d)^{-1}B_d \\ &+ B_d^T(I + A_d^T)^{-1}(C_c^T - P_cB_c) + B_d^T(I + A_d^T)^{-1}(P_d - A_d^TP_dA_d)(I + A_d)^{-1}B_d \end{aligned}$$

Making the substitutions (4.38) and (4.39), the first term in the above matrix becomes:

$$\begin{aligned} -P_cA_c - A_c^TP_c &= P_c(A_d + I)^{-1}(I - A_d) + (I - A_d^T)(A_d^T + I)^{-1}P_c \\ &= (A_d^T + I)(A_d^T + I)^{-1}P_c(A_d + I)^{-1}(I - A_d) \\ &+ (I - A_d^T)(A_d^T + I)^{-1}P_c(I + A_d)^{-1}(I + A_d) \\ &= \frac{1}{2}\{(A_d^T + I)P_d(I - A_d) + (I - A_d^T)P_d(I + A_d^T)\} \\ &= P_d - A_d^TP_dA_d \end{aligned}$$

Similarly, the second term simplifies as follows:

$$\begin{aligned}
& C_c^T - P_c B_c - (P_c A_c + A_c^T P_c)(I + A_d)^{-1} B_d \\
&= C_d^T - 2P_c(I + A_d)^{-2} B_d - (P_d - A_d^T P_d A_d)(I + A_d)^{-1} B_d \\
&= C_d^T - (I + A_d^T) P_d (I + A_d)^{-1} B_d - (P_d - A_d^T P_d A_d)(I + A_d)^{-1} B_d \\
&= C_d^T - \{(I + A_d^T) P_d - P_d + A_d^T P_d A_d\}(I + A_d)^{-1} B_d \\
&= C_d^T - \{A_d^T P_d + A_d^T P_d A_d\}(I + A_d)^{-1} B_d \\
&= C_d^T - A_d^T P_d B_d
\end{aligned}$$

The third term is simply the transpose of the second term. Finally, the X term reduces to:

$$\begin{aligned}
X &= B_d^T (I + A_d^T)^{-1} (P_d - A_d^T P_d A_d)(I + A_d)^{-1} B_d \\
&+ C_d (I + A_d)^{-1} B_d - 2B_d^T (I + A_d^T)^{-2} P_c (I + A_d)^{-1} B_d \\
&+ B_d^T (I + A_d^T)^{-1} C_d^T 2B_d^T (I + A_d^T)^{-1} P_c (I + A_d)^{-2} B_d \\
&+ D_d + D_d^T - C_d (I + A_d)^{-1} B_d - B_d^T (I + A_d^T)^{-1} C_d^T \\
&= D_d + D_d^T + B_d^T (I + A_d^T)^{-1} \{P_d - A_d^T P_d A_d - P_d (I + A_d) - (I + A_d^T) P_d\}(I + A_d)^{-1} B_d \\
&= D_d + D_d^T - B_d^T (I + A_d^T)^{-1} \{(I + A_d^T) P_d (I + A_d)\}(I + A_d)^{-1} B_d \\
&= D_d + D_d^T - B_d^T P_d B_d
\end{aligned}$$

Appendix E. Proof of condition (4.36)

Before going through the derivation, we first note the following identity:

$$\begin{aligned}
(R - B_d^T P_d B_d)^{-1} &= R^{-1} + R^{-1} B_d^T (P_d^{-1} - B_d R^{-1} B_d^T)^{-1} B_d R^{-1} \\
&= R^{-1} + R^{-1} B_d^T P_d B_d R^{-1} \\
&+ R^{-1} B_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} B_d^T P_d B_d R^{-1} \qquad (4.65)
\end{aligned}$$

provided P_d and R are non singular. We actually assume in this proof that P_d and R are positive definite. The previous identity is obtained by applying twice the matrix inversion lemma. To prove that $P_{d_{min}}$ is a solution to equation (4.36), we show that by making the suggested substitution for A_1 and R, we get (4.35) for which $P_{d_{min}}$ is a solution. We can therefore write:

$$\begin{aligned}
P_d &= (A_d - B_d R^{-1} C_d)^T P_d (A_d - B_d R^{-1} C_d) + C_d^T R^{-1} C_d \\
&+ (A_d - B_d R^{-1} C_d)^T P_d B_d (R - B_d^T P_d B_d)^{-1} B_d^T P_d (A_d - B_d R^{-1} C_d) \\
&= C_d^T R^{-1} C_d + C_d^T R^{-1} B_d^T P_d B_d R^{-1} C_d
\end{aligned}$$

$$
\begin{aligned}
&+ \quad C_d^T R^{-1} B_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} B_d^T P_d B_d R^{-1} C_d + \text{ other terms} \\
&= \quad C_d^T (R - B_d^T P_d B_d)^{-1} C_d + \text{ other terms}
\end{aligned}
$$

where the last equation follows from (4.65). Substituting now the other terms, we get:

$$
\begin{aligned}
P_d &= A_d^T P_d A_d + C_d^T (R - B_d^T P_d B_d)^{-1} C_d A_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} \\
&- A_d^T P_d B_d R^{-1} C_d - A_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} B_d^T P_d B_d R^{-1} C_d \\
&- C_d^T R^{-1} B_d^T P_d A_d - C_d^T R^{-1} B_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} B_d^T P_d A_d \\
&= A_d^T P_d A_d + C_d^T (R - B_d^T P_d B_d)^{-1} C_d A_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} \\
&- A_d^T P_d B_d \{R^{-1} + (R - B_d^T P_d B_d)^{-1} B_d^T P_d B_d R^{-1}\} C_d \\
&- C_d^T \{R^{-1} + R^{-1} B_d^T P_d B_d (R - B_d^T P_d B_d)^{-1}\} B_d^T P_d A_d
\end{aligned} \tag{4.66}
$$

We are only interested in simplifying the cross terms (last two lines of (4.66)). Simplifying one cross term is actually enough because they are the transpose each other. So, recalling that $R \succ_k 0$ (by assumption), we can then write:

$$
\begin{aligned}
&C_d^T \{R^{-1} + R^{-1} B_d^T P_d B_d (R - B_d^T P_d B_d)^{-1}\} B_d^T P_d A_d \\
&= C_d^T R^{-1/2} \{I + R^{-1/2} B_d^T P_d B_d (R - B_d^T P_d B_d)^{-1} R^{1/2}\} R^{-1/2} B_d^T P_d A_d \\
&= C_d^T R^{-1/2} \{I + R^{-1/2} B_d^T P_d B_d R^{-1/2} (I - R^{-1/2} B_d^T P_d B_d R^{-1/2})^{-1}\} R^{-1/2} B_d^T P_d A_d \\
&= C_d^T R^{-1/2} \{(I - R^{-1/2} B_d^T P_d B_d R^{-1/2}) + R^{-1/2} B_d^T P_d B_d R^{-1/2}\} \\
&\quad (I - R^{-1/2} B_d^T P_d B_d R^{-1/2})^{-1} R^{-1/2} B_d^T P_d A_d \\
&= C_d^T R^{-1/2} (I - R^{-1/2} B_d^T P_d B_d R^{-1/2})^{-1} R^{-1/2} B_d^T P_d A_d \\
&= C_d^T (R - B_d^T P_d B_d)^{-1} B_d^T P_d A_d
\end{aligned} \tag{4.67}
$$

By substituting (4.67) and its transpose into (4.66), (4.35) is easily obtained.

Chapter 5

Concluding remarks

We would like to conclude this dissertation with some brief remarks on two of the various future research directions. The chapter is divided into two main sections. In the first section, we propose possible extensions of some of the results of the thesis and cite some of the remaining open problems related to the topic of optimal subband coding. In the second section, we discuss the potential application of some of the main ideas of chapter 3 to the area of multiuser communications such as in code division multiple access (CDMA) and discrete multi-tone modulation (DMT).

5.1 Optimal subband coders

To quickly recapitulate the main ideas, recall that in the absence of quantization, the theory and design of so-called perfect reconstruction (PR) filter banks is well established. In the presence of quantizers, perfect reconstruction is not possible and the FB output $\hat{x}(n)$ in this case is the original input $x(n)$ plus a filtered version of the quantization noise denoted by $e(n)$. Given a fixed budget of b bits for the subband quantizers, the aim is to minimize the average variance of $e(n)$. This problem involves optimizing the analysis and synthesis filters and choosing a subband bit allocation strategy.

For the subband coder case and with order unconstrained filters, theoretical results on the statistical optimization of two channel orthonormal filter banks were given by Unser [68]. A general set of necessary and sufficient conditions for the optimality of an M-channel maximally decimated orthonormal uniform filter bank were independently developed in [76]. Recently, some issues pertaining to the optimization of a subband coder over the class of biorthogonal filter banks have been addressed in [2] and [80] but the optimization of biorthogonal filter banks, in its entire generality, remains an open problem. Furthermore, the extension of the results of chapter 2 to the M-channel case remain an open problem. Some numerical results were actually obtained by Gosse and Duhamel [26] but no analytical solution is currently available. Another problem is the *optimal* generalization of Djokovic

and Vaidyanathan's scheme. If fully optimized, the more general scheme of Fig. 2.16 should clearly outperform the scheme proposed by Djokovic and Vaidyanathan. An easy way to see this is to simply put a Wiener filter at the output of the half whitening filter sandwiching an optimum orthonormal PRFB. Even in this suboptimum procedure, the mean square reconstruction error cannot increase.

In the finite order case (FIR and IIR filters), almost all cases remain unsolved. In specific, globally optimal solutions for finite order filters (FIR and IIR) are not known except for the special case of a transform coder (KLT solution) and for the FIR two channel orthonormal filter bank (chapter 4 of this thesis). A good starting point might be the extension of the results of chapter 4 for the M-channel case. The issue of the existence of a principal component filter bank with FIR filters of arbitrary order N is quite important and can be also pursued.

In tree structure configurations (uniform, wavelet type and wavelet packets), the optimal design of both ideal and finite order subband filters adapted to input signal statistics are not known and represent interesting problems to be investigated.

The various problems stated above assume the standard high bit rate noise model, use the mean square error criterion and can therefore be interpreted as the "classical" approach to design optimal subband coders. A more realistic approach is to develop a quantizer model that matches better the rate distortion curve at low bit rates. A recent paper by Mallat and Falzon discuss this issue in some detail [47] but it is somehow premature to draw any strong conclusions at this moment in time. Another important topic is the optimization of subband coders according to perceptual measures. Compression schemes of video, images and audio based on models of human perception rather the conventional mean square error use subband and transform coders as a major building block. The main reason is that filter banks seem to be qualitatively matched to the properties and methods of human perception (e.g. human auditory system). With perceptual masking in the subbands and using some quantitative perceptual criterion, it is not yet clear how to design the filter bank for "optimum" performance. In fact, it appears that most of the research is currently experimental in nature. Therefore, good theoretical results can provide insight and suggest different ways to optimize filter banks for best perceptual performance.

5.2 Multirate signal processing applications in communications.

In chapter 3, we have shown that for a class of non bandlimited signals that are modeled as the output of a single or several interpolation filters, we can greatly reduce the quantization noise variance by taking advantage of the oversampled nature of the signals. In the time domain, the oversampled property of these signals is expressed by an excess (redundant) amount of samples. It is that extra amount of available samples that produce a decrease in the noise variance. That same idea is actually

the essence of channel coding (introduce redundancy to combat error) and it is therefore not surprising that novel CDMA schemes based on similar multirate structures has been recently an active area of research. Discrete multi-tone modulation (DMT) schemes based on NON-DFT filter banks have also been proposed to reduce the cross talk (ICI) between adjacent channels and the intersymbol interference (ISI) within each individual channel.

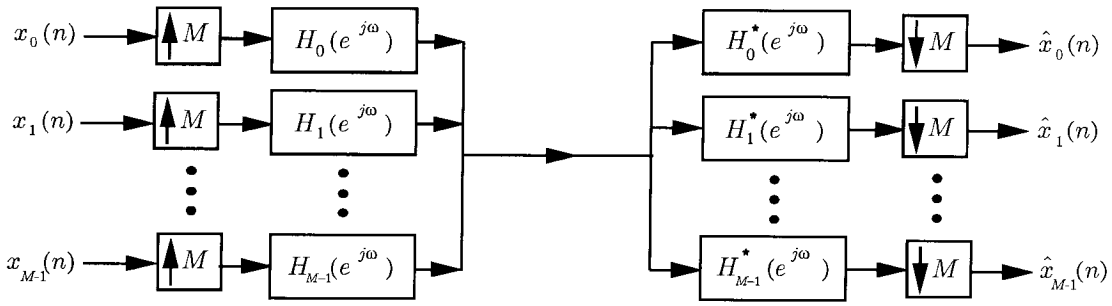

Fig. 5.1: A discrete time transmultiplexer with an ideal channel

From a *multirate digital signal processing* perspective, the modeling of the above problems reduces to the design of a digital transmultiplexer, shown in Fig. 5.1. The transmultiplexer of Fig. 5.1 can be interpreted as the dual of the subband coder. The multi-input *synthesis* section is now connected to the multi-output *analysis* section by a single channel. Contrast this with the subband coder case where the single input analysis section is connected to the single output synthesis section by multiple "channels". Analogous to the subband coder problem with and without quantizers, the theory of "perfect reconstruction" transmultiplexers is well known, assuming an ideal channel. In fact, the design of a PR transmultiplexer is equivalent to designing a paraunitary filter bank. Cross talk elimination results from the orthogonality of the filters whereas intersymbol interference cancelation follows from the Nyquist property of the magnitude squared of each individual filter. Future challenges are in the design of "optimum" transmultiplexers over additive white Gaussian noise (AWGN) channels, flat fading channels and frequency selective fading channels. This includes general theoretical solutions and optimum bounds as well as efficient and implementable schemes. The theory and design of general (not wavelet based) non uniform transmultiplexers is also important and has not been investigated. Non uniform transmultiplexers correspond to the situation where different types of information e.g. voice versus data are sent over the channel at different rates.

Bibliography

[1] Aas, K. C., Duell, K. C. and Mullis, C. T., "Synthesis of extremal wavelet-generating filters using Gaussian quadrature", IEEE Transactions on Signal Processing, Vol. 43, pp. 1045-1057, May 1995.

[2] Aas, K. C. and Mullis, C. T., "Minimum mean-squared error transform coding and subband coding", IEEE Transactions on Information Theory, Vol. 42, pp. 1179-1192, July 1996.

[3] Alizadeh, F., "Interior point methods in semidefinite programming with applications to combinatorial optimization", Siam journal on Optimization, Vol. 5., pp. 13-51, February 1995.

[4] Anderson, B. D. O. and Vongpanitlerd, S., *Network analysis and synthesis : A modern system theory approach*, Englewood Cliffs, NJ: Prentice Hall, 1973.

[5] Anderson, B. D. O., "Algebraic properties of minimal degree spectral factors", Automatica, Vol. 9, pp. 491-500, June 1973.

[6] Aziz, P., Sorensen, H. and Van Der Spiegel, J., "An overview of sigma-delta converters", IEEE Signal Processing Magazine, Vol. 13, pp. 61-84, January 1996.

[7] Berger, T., *Rate distortion theory : a mathematical basis for data compression*, Englewood Cliffs, NJ: Prentice Hall, 1971.

[8] Berger, T. and Tufts, Donald, "Optimum pulse amplitude modulation part I : transmitter-receiver design and bounds from information theory", IEEE Transactions on Information Theory, pp. 196-208, April 1967.

[9] Boyd S., El Ghaoui L., Feron E. and Balakrishnan V., *Linear matrix inequalities in system and control theory*, Vol. 15 of studies in applied mathematics, SIAM, Philadelphia, PA, 1994.

[10] Caglar, H., Liu, Y. and Akansu, A., "Statistically optimized PR-QMF design", SPIE, Visual Communication and Image Processing, Vol. 1605, pp. 86-94, 1991.

[11] Chevillat, P. R. and Ungerboeck, G., "Optimum FIR transmitter and receiver filters for data transmission over bandlimited channels", IEEE Transactions on Communications, Vol. 30, pp. 1909-1915, August 1982.

[12] Chan, D. and Donaldson, R. W., "Optimum pre-and postfiltering of sampled signals with application to pulse modulation and data compression systems", IEEE Transactions on Communication Technology, pp. 141-157, April 1971.

[13] Churchill, R. V., *Complex variables and applications*, McGraw-Hill Inc., 1990.

[14] Costas, J. P., "Coding with linear systems", Proceedings of the IRE, pp. 1101-1103, September 1952.

[15] Crochiere, R. E., Weber, S. A., and Flanagan, J. L., "Digital coding of speech in subbands", Bell Systems Technical Journal, Vol. 55, pp. 1069-1085, October 1976.

[16] Croisier, A., Esteban, D., and Galand, C., "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques", International Symposium on Information, Circuits and Systems, Patras, Greece, 1976.

[17] Delopoulos, A. and Kolias, S., "Optimal filterbanks for signal reconstruction from noisy subband components", IEEE Transactions on Signal Processing, Vol. 44, pp. 212-224, February 1996.

[18] Delsarte, P., Macq, B. and Slock, D., "Signal-adapted multiresolution transform for image coding", IEEE Transactions on Information Theory, Vol. 38, pp. 897-904, March 1992.

[19] Demeure, C. J. and Mullis, C. T., "The Euclid algorithm and the fast computation of cross-covariance and auto-covariance sequences", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, pp. 545-552, April 1989.

[20] Djokovic, I. and Vaidyanathan, P. P., "On optimal analysis/synthesis filters for coding gain maximization", IEEE Transactions on Signal Processing, Vol. 44, pp. 1276-1279, May 1996.

[21] Franklin, G. F., Powell, J. D. and Workman, M. L., *Digital control of dynamic systems*, Addison-Wesley publishing company, Inc., 1990.

[22] Gardner, W. A. and Franks, L. E., "Characterization of cyclostationary random processes", IEEE Transactions on Information Theory, Vol. 21, pp. 4-14, January 1975.

[23] Gardner, W. A., "Stationarizable random processes", IEEE Transactions on Information Theory, Vol. 24, pp. 8-22, January 1978.

[24] Gelfand, I. M. and Fomin, S. V. *Calculus of variations,* Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1963.

[25] Gladyshev, E. G., "Periodically correlated random sequences", Soviet Mathematics, Vol. 2, pp. 385-388, 1961.

[26] Gosse, K. and Duhamel, P., "Perfect reconstruction versus MMSE filter banks in source coding", IEEE Transactions on Signal Processing, Vol. 45, pp. 2188-2202, September 1997.

[27] Goodman, L. M. and Drouillet, P. R., "Asymptotically optimum pre-emphasis and de-emphasis networks for sampling and quantizing", Proceedings of the IEEE, pp. 795-796, May 1963.

[28] Gradshteyn, I. S. and Ryzhik, I. M. *Table of integrals, series, and products,* fifth edition, Academic press Inc., San Diego, California, 1994.

[29] Hayes, M. *Statistical Digital Signal Processing and Modeling*, John Wiley and Sons, Inc., 1996.

[30] Hemami, S. S. and Gray, R. M., "Subband filters optimized for lost coefficient reconstruction", IEEE Transactions on Signal Processing, Vol. 45, pp. 763-767, March 1997.

[31] Hitz, B. E. L. and Anderson, B. D. O., "Discrete positive-real functions and their application to system stability", Proceedings of the IEE, Vol. 116, pp. 153-155, January 1969.

[32] Hoang, P.-Q. and Vaidyanathan, P. P., "Non-uniform multirate filter banks: theory and design", Proceedings ISCAS, pp. 371-374, Portland, Oregon 1989.

[33] Horn, R. A. and Johnson, C. R., *Matrix Analysis,* Cambridge University Press, 1985.

[34] Huang, Y. and Schultheiss, P. M., "Block quantization of correlated Gaussian random variables", IEEE Transactions on Communications Systems, pp. 289-296, September 1963.

[35] Hurd, H. L., "Stationarizing properties of random shifts", SIAM Journal of Applied Mathematics, Vol. 26, pp. 203-212, January 1974.

[36] Jain, A. K., *Fundamentals of digital image processing,* Prentice Hall, Inc., Englewood Cliffs, 1989.

[37] Jayant, N. S. and Noll, P. *Digital coding of waveforms,* Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1984.

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

[39] Jin, Q., Luo, Z.-Q. and Wong, K. M., "Optimum filter banks for signal decomposition and its applications in adaptive echo cancelation", IEEE Transactions on Signal Processing, Vol. 44, pp. 1669-1689, July 1996.

[40] Kailath, T., *Linear systems*, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1980.

[41] Kirac, A. and Vaidyanathan, P. P., "Theory and design of optimum FIR compaction filters", submitted to the IEEE Transactions on Signal Processing, 1996.

[42] Kirac, A. and Vaidyanathan, P. P., "FIR compaction filters : new design methods and properties", Proceedings ICASSP, Vol. 3, pp. 2229-2232, Munich, 1997.

[43] Kirac, A. and Vaidyanathan, P. P., "Analytical method for 2-channel optimum orthonormal filter banks", Proceedings ISCAS, Honk-Kong, 1997.

[44] Kirk, D. E., *Optimal Control Theory: an introduction*, Prentice Hall, Inc., Englewood Cliffs, New Jersey, 1970.

[45] Kovacevic, J. and Vetterli, M., "Perfect reconstruction filter banks with rational sampling rate changes", Proceedings ICASSP, pp. 1785-1788, Toronto 1991.

[46] Kovacevic, J., "Subband coding systems incorporating quantizer models", IEEE Transactions on Image Processing, Vol. 4, pp. 543-553, May 1995.

[47] Mallat, S. and Falzon, F., "Understanding image transform codes", submitted to the IEEE Transactions on Signal Processing, special issue on Wavelets and Filter Banks, 1997.

[48] Malvar, H. S. and Staelin, D. H.,"Optimal FIR pre- and postfilters for decimation and interpolation of random signals", IEEE Transactions on Communications, pp. 67-74, January 1988.

[49] Marshall, A. W. and Olkin, I., *Inequalities: Theory of majorization and its applications*, Academic press, Inc., 1979.

[50] Mintzer, F., "Filters for distortion-free two band multirate filter banks", IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 33, pp. 626-630, June 1985.

[51] Moulin, P., Anitescu, M., Kortanek, K. O. and Potra, F.,"Design of signal-adapted FIR paraunitary filter banks", Proceedings ICASSP, Vol. 3., pp. 1519-1522, Atlanta, 1996.

[52] Moulin, P.,"A new look at signal-adapted QMF bank design", Proceedings ICASSP, Vol. 5., pp. 1312-1315, Detroit, 1995.

[53] Nesterov, Y and Nemirovskii, A., *Interior point polynomial algorithms in convex programming*, Vol. 13 of studies in applied mathematics, SIAM, Philadelphia, PA, 1994.

[54] Pesquet, J. C. and Combettes, P. L., "Wavelet synthesis by alternating projections", IEEE Transactions on Signal Processing, Vol. 44, pp. 728-732, March 1996.

[55] Phoong, S.-M., Kim, C. W., Vaidyanathan, P. P. and Ansari, R., "A new class of two channel biorthogonal filter banks and wavelet bases", IEEE Transactions on Signal Processing, Vol. 43, pp. 649-665, March 1995.

[56] Ramana, M. V., "An exact duality theory for semidefinite programming and complexity applications", Mathematical programming, Vol. 77, pp. 129-162, May 1997.

[57] Ramstad, T. A., "IIR filterbank for subband coding of images", Proceedings ISCAS, pp. 827-830, June 1988.

[58] Sathe, V. S. and Vaidyanathan, P. P., "Effects of multirate systems on the statistical properties of random signals", IEEE Transactions on Signal Processing, Vol. 41, pp. 131-146, January 1993.

[59] Segall, A., "Bit allocation and encoding for vector sources", IEEE Transactions on Information Theory, pp. 162-169, March 1976.

[60] Smith, M. J. T. and Barnwell, T. P., "Exact reconstruction techniques for tree structured subband coders", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 34, pp. 434-441, June 1986.

[61] Soman, A. K. and Vaidyanathan, P. P.," Coding gain in paraunitary analysis/synthesis systems", IEEE Transactions on Signal Processing, Vol. 41, pp. 1824-1835, May 1993.

[62] Taubman, D. and Zakhor, A., "A multi-start algorithm for signal adaptive subband systems", Proceedings ICASSP, Vol. 3, pp. 213-216, San Francisco, 1992.

[63] Troutman, J. L., *Variational calculus with elementary convexity*, Springer-Verlag New York Inc., 1983.

[64] Tsatsanis, M. K. and Giannakis, G. B., "Time-varying system identification and model validation using Wavelets", IEEE Transactions on Signal Processing, Vol. 41, pp. 3512-3523, December 1993.

[65] Tsatsanis, M. K. and Giannakis, G. B., "Principal component filter banks for optimal multiresolution analysis", IEEE Transactions on Signal Processing, pp. 1766-1777, Vol. 43, August 1995.

[66] Tuqan, J. and Vaidyanathan, P. P., "Oversampling PCM Techniques and Optimum Noise Shapers for Quantizing a Class of Nonbandlimited Signals", submitted to the IEEE Transactions on Signal Processing, 1996.

[67] Unser, M., "An extension of the Karhunen-Loeve transform for wavelets and perfect reconstruction filter banks", SPIE mathematical imaging, Vol. 2034, pp. 45-56, 1993.

[68] Unser, M., "On the optimality of ideal filters for pyramid and wavelet signal approximation", IEEE Transactions on Signal Processing, Vol. 41, pp. 3591-3596, December 1993.

[69] Uzun, N. and Haddad, R. A., "Cyclostationary Modeling, analysis and optimal compensation of quantization errors in subband codecs", IEEE Transactions on Signal Processing, Vol. 43, pp. 2109-2119, September 1995.

[70] Vaidyanathan, P. P., "Theory and design of M-channel maximally decimated quadrature mirror filters with arbitrary M, having the perfect reconstruction property", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 35, pp. 476-492, April 1987.

[71] Vaidyanathan, P. P. and Hoang, P.-Q., "Lattice structures for optimal design and robust implementation of two channel perfect reconstruction of QMF banks", IEEE Transactions on Signal Processing, Vol. 36, pp. 81-94, January 1988.

[72] Vaidyanathan, P. P., Nguyen, T. Q., Doganata, Z. and Saramaki, T., "Improved technique for design of perfect reconstruction FIR QMF banks with lossless polyphase matrices", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, pp. 1042-1056, July 1989.

[73] Vaidyanathan, P. P. *Multirate systems and filter banks,* Prentice Hall, Inc., Englewood Cliffs, 1993.

[74] Vaidyanathan, P. P., "Orthonormal and biorthonormal filter banks as convolvers, and convolutional coding gain", IEEE Transactions on Signal Processing, Vol. 41, pp. 2110-2130, June 1993.

[75] Vaidyanathan, P. P. and Chen, T., "Statistically Optimal synthesis banks for subband coders reconstruction", Proceedings 28th Annual Asilomar conf. Sig., Sys. and Comp.", Oct-Nov. 1994.

[76] Vaidyanathan, P. P., "Optimal orthonormal filterbanks", submitted to IEEE Transactions on Signal Processing, 1995.

[77] Vaidyanathan, P. P. and Phoong, S.-M., "Reconstruction of sequences from non uniform samples", Proceedings ISCAS, pp. 601-604, Seattle 1995.

[78] Vaidyanathan, P. P. and Phoong, S.-M., "Discrete time signals which can be recovered from samples", Proceedings ICASSP, pp. 1448-1451, Detroit 1995.

[79] Vaidyanathan, P. P., "Theory of optimal orthonormal filterbanks", Proceedings ICASSP, Vol. 3, pp. 1487-1490, Atlanta, May 1996.

[80] Vaidyanathan, P. P. and Kirac, A., "Results on optimal biorthogonal filter banks", to be submitted to the IEEE Transactions on Signal Processing, October 1997.

[81] Vandenberghe, L. and Boyd, S., "SP: software for semi definite programming. User's guide. Beta version.", K. U. Leuven and Stanford university, October 1994.

[82] Vandenberghe, L. and Boyd, S., "A primal-dual potential reduction method for problems involving matrix inequalities", Mathematical programming, Vol. 69, pp. 205-236, July 1995.

[83] Vandenberghe, L. and Boyd, S., "Semidefinite programming", Siam review, Vol. 38, pp. 49-95, March 1996.

[84] Vandendorpe, L., "CQF filter banks matched to signal statistics", Signal Processing, Vol. 29, pp. 237-249, December 1992.

[85] Vetterli, M., "Filter banks allowing for perfect reconstruction", Signal Processing, Vol. 10, pp. 219-244, April 1986.

[86] Vetterli, M. and Kovacevic, J., *Wavelets and subband coding*, Prentice-Hall, Inc., Englewood Cliffs, 1995.

[87] Walter, G. G., "A sampling theorem for wavelet subspaces", IEEE Transactions on Information Theory, pp. 881-884, March 1992.

[88] Willems, J. C., "Least squares stationary optimal control and the algebraic Ricatti equation", IEEE Transactions on Automatic Control, Vol. 7, pp. 621-634, December 1971.

[89] Wu, S.-P. and Boyd, S., "Sdpsol: A parser/solver for semi definite programming and determinant maximization problems with matrix structure. User's guide, version beta.", Stanford university, June 1996.

[90] Xuan, D. and Bamberger, R., "Multidimensional paraunitary principal component filter banks", Proceedings ICASSP, Vol. 2., pp. 1488-1491, Detroit, 1995.

[91] Xuan, D. and Bamberger, R., "Complete FIR principal component filter banks", Proceedings ISCAS, Vol. 2., pp. 417-420, Atlanta, 1996.

[92] Xuan, D. and Bamberger, R., "2D factorizable FIR principal component filter banks", Proceedings ICASSP, Vol. 3., pp. 1538-1541, Atlanta, 1996.

[93] Youla, D. C., "On the factorization of rational matrices", IRE Transactions on Information Theory, Vol. 16, pp. 172-189, July 1961.