Optimal Pre- and Post-Filtering in Noisy Sampled-Data Systems

Henrique Sarmento Malvar

Technical Report 519

August 1986

Massachusetts Institute of Technology

Research Laboratory of Electronics

Cambridge, Massachusetts 02139

Optimal Pre- and Post-Filtering in Noisy Sampled-Data Systems

Henrique Sarmento Malvar

Technical Report 519

August 1986

Massachusetts Institute of Technology

Research Laboratory of Electronics

Cambridge, Massachusetts 02139

This work has been supported in part by the Brazilian Government, through its Conselho Nacional de Desenvolvimento Científico e Tecnológico. It has also been supported in part by the Center for Advanced Television Studies, an industry group consisting of the American Broadcasting Company, Ampex Corporation, Columbia Broadcasting Systems, Harris Corporation, Home Box Office, Public Broadcasting Service, National Broadcasting Company, RCA Corporation, Tektronix, and the 3M Company.

OPTIMAL PRE- AND POST-FILTERING IN NOISY SAMPLED-DATA SYSTEMS

by

HENRIQUE SARMENTO MALVAR

Submitted to the Department of Electrical Engineering
and Computer Science on August 15, 1986, in partial
fulfillment of the requirements for the Degree of
Doctor of Philosophy in Electrical Engineering and Computer Science.

ABSTRACT

In this thesis we consider the problem of jointly optimizing the pre- and post-filters in a communications or storage system, with optimality considered in a weighted mean-square error sense. We adopt a system model that is general enough to be applicable to a wide variety of problems, such as broadcasting, tape recording, telemetry, and signal coding, among others. Our fundamental assumptions throughout this work are that the pre- and post-filters are linear and that all signal and noise spectra of interest are known.

We derive the optimal pre- and post-filters for three basic classes of systems, characterized by infinite impulse response (IIR), finite impulse response (FIR), and block filters. Whenever appropriate, we present filters with nearly optimal performance that can be efficiently implemented. We also derive analytic forms and a fast version for a recently introduced class of pre- and post-filters for block processing with overlapping basis functions, namely, "Lapped Orthogonal Transforms" (LOT's). In all of these classes, for typical image processing and coding applications, we obtain improvements in the weighted r.m.s. error over traditional systems on the order of 1 to 6 dB.

Some of the results of this work can be immediately used to improve existing digital signal coding systems. For example, the combination of pseudo-random noise quantization with appropriate filtering, and the use of a fast LOT, may lead to a reduction of more than 3 dB in the r.m.s. error in a block coder, with a simultaneous whitening of the noise patterns and significant reduction of the so-called "blocking effects".

Thesis Supervisor: David H. Staelin

Title: Professor of Electrical Engineering

To Regina Helena and Ana Beatriz

who make everything worthwhile

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Professor David Staelin for his guidance, supervision, sharing of ideas, dedication of time, and continuous encouragement. His good humor and ever present willingness to discuss all of my ideas, even the ones not related to this work, have made my stay as a graduate student at M.I.T. a most enjoyable experience. The useful suggestions of Prof. Jae Lim and Prof. Bernard Lévy, my thesis readers, are also acknowledged.

I want to thank also my friends in the Video Information Processing Group, for all the fun: weeks in a row of pizza for lunch with Jeff Bernstein and Brian Hinman; unusual talks relating child psychology and Weierstrass' theorems with Alain Briançon; the "let's do hardware" calls to reality of Adam Tom and Jerry Shapiro; the TeX-nical and signal coding talks, mostly at 3:00 a.m., with Ashok Popat; and the more normal conversations with Bernie Szabo, Mark Colavita, and Philippe Cassereau. Jack Barrett was always helpful in my hardware projects.

Other friends at M.I.T. will always be present in my memories of campus life. In particular, the weekly Portuguese talks over lunch at the Lobdell cafeteria with Peter Roberts, and the friendly discussions about the latest PC hardware and software with Paulo Rosman.

I would also like to thank my parents, for their constant support and encouragement. They introduced me to the basics of statistics and probability, and that has certainly influenced my statistical approach to communication theory.

My deepest thanks go to Regina and Ana Beatriz for their love and moral support. They certainly deserve as much credit as I do for completion of this work. I love you both.

Some of the equipment used for the image processing experiments was made available by Prof. D. E. Troxel, of the Cognitive Information Processing Group at M.I.T., and by Brian Hinman, of PicTel Corporation.

Finally, I would like to thank the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* of the Brazilian Government for financial support and to the *Fundação Universidade de Brasília* for the four-year leave of absence that made my studies at M.I.T. possible.

Table of Contents

Chapter 1

Introduction

One of the basic problems of communication theory is that of efficient transmission of information through a noisy channel [1]. By efficient we mean that the transmitter and receiver must be designed so that the reconstructed signal replicates the original as well as possible, given the inherent physical limitations of the systems involved. If an adequate measure of the error between the original and reconstructed signals is available, we can postulate the problem of designing an optimal communication system, in which a minimum of that error measure is sought.

There are many issues and sub-problems related to the optimal design of transmitters and receivers, ranging from appropriate signal modeling to circuit design. We will concentrate this work on the system model of Fig. 1.1. The input signal is available only after being degraded by an additive input noise. The channel is modeled as another additive noise source. In general, the input signal is available at a higher bandwidth, or higher rate, than the channel is capable of handling. Thus, some form of processing, generally a combination of sampling and filtering, has to be applied to the signal before it is sent through the channel.

One of the functions of the transmitter is to shape the input signal spectrum into some appropriate form that takes into account sampling and noise degradations. At the receiver, an approximate inverse filter is used, so that as much of the original signal as possible is recovered. These filtering operations are performed by the networks labeled pre- and post-filter in Fig. 1.1. The sampler may not always be present, in which case it is removed from Fig. 1.1. We will consider in this work

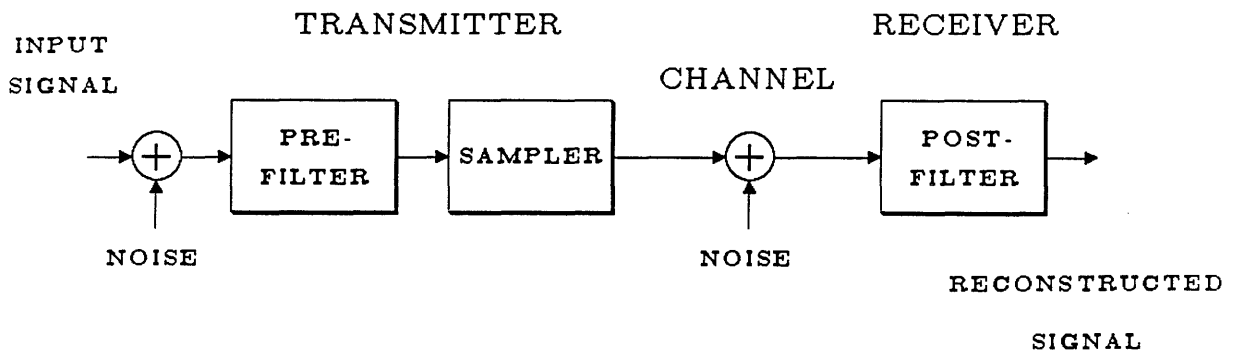


Fig. 1.1. Fundamental system model.

systems with and without samplers. For digital channels, continuous amplitude signals cannot be represented exactly, so that some quantization error is unavoidable. This error can be appropriately modeled by the additive channel noise source of Fig. 1.1.

1.1. Applications of the model

The system of Fig. 1.1 may represent a large variety of physical systems. A typical application is in the design of a broadcasting system, in which the pre- and post-filters are also referred to as the pre-emphasis and de-emphasis networks. In this case, there is no sampling. As pointed out by Cramer [2], the pre- and de-emphasis characteristics of the standard FM broadcasting system, for example, were chosen with basis on the characteristics of the modulation system. Using reciprocal transmitter and receiver filters, Cramer has shown that the SNR of an FM system can be improved by more than 7 dB relative to the standard 75μs FM pre-emphasis, for a speech spectrum. The standard AM broadcasting system, which is used not only

for commercial radio but also for the image component of commercial television, does not employ any pre- and de-emphasis circuitry [3] at all! There is certainly much room for improvement in AM systems.

Another application of the pre- and post-filtering model of Fig. 1.1 is in tape recording systems. Historically, tape recording is more recent than radio broadcasting, and that is probably the reason why the pre- and de-emphasis problem has been analyzed much more carefully in that context. The criteria used in most cases, including some of the NAB standards for tape recording, were [4]: 1) flattening of the signal spectrum, so that the signal to be recorded has approximately a white frequency distribution; 2) subjective criteria based on listening tests; and 3) equalization of the overload probability for all frequencies.

It is interesting to note that one of the most important specifications of commercial tape recorders, by which their quality is usually judged, is the CCIR-weighted signal-to-noise ratio. The weighting curve takes into account the frequency response of the human ear. This is a strong justification for the application of the results in Chapter 2 to the design of tape equalization systems. One important aspect regarding the pre-emphasis network, discussed in Chapter 2, is that the optimal pre-filter should *not* flatten the input spectrum; in fact it should perform a *half-whitening* processing. This certainly contradicts criterion 1) in the above paragraph, which sounds intuitively valid but is theoretically incorrect. Thus, we believe that the design of optimal tape equalization systems could benefit from the results of our work. With the increased usage of digital filtering techniques, as exemplified by the compact disc system [5], the optimal finite impulse response (FIR) filters of Chapter 3 should be of significant practical value.

In digital processing and coding of speech and images [6], the advantages of optimal pre- and post-filtering are clear. In particular, the design of optimal FIR filters for multidimensional signal processing can be simplified by the use of the techniques presented in Chapter 3. For block coding systems, the results of Chapter 4 allow the joint optimization of linear coders and decoders. Since the

error due to quantization is taken into account, we believe that the system models of Chapters 3 and 4 are adequate for most digital communications systems.

There are many other areas in which the system model of Fig. 1.1 could be applied. One example is telemetry, in which the outputs of a set of sensors can be optimally pre-filtered before transmission to a remote analysis site. Transmission may be carried out by a set of phone lines, for example, for which the vector channel model of Chapter 4 is appropriate. Another is sampling and interpolation of time series, where the optimal pre- and post-filters provide the best compromise between aliasing errors and distortions due to missing spectral components.

1.2. Problem Formulation and Thesis Outline

The main objective of this work is the design of jointly-optimal linear pre- and post-filters, with the distortion measure to be minimized being a weighted mean-square error between the input and reconstructed signals. The input signal and additive noise sources are zero-mean real random processes with known spectra. We are interested not only in deriving performance bounds, but also in the practical design of realizable filters, including the design of sub-optimal filters that can approximate the performance of the optimal systems at a reduced implementation cost.

Even with all the assumptions stated above, there are still several classes of filters to be considered. In Chapter 2 we derive optimal filters with infinite impulse responses (IIR). These ideal IIR filters, although not realizable, provide bounds on the system performance. Also, they provide basic prototypes, which could be used as a basis for the design of realizable filters.

The main properties of jointly-optimal IIR filters for the system in Fig. 1.1 that we derive in Chapter 2 are: 1) the optimal pre- and post-filters are band-limited, and when sampling is present they avoid aliasing; 2) the optimal pre-filter performs a 'half-whitening' operation on the input signal, so that the spectrum at

the channel input is proportional to the square root of the input spectrum. One important observation about the optimal filters in Chapter 2 is that their transfer functions are quite sensitive to the characteristics of the input and channel noises, even at high signal-to-noise ratios. Therefore, even if the noise sources have a small amplitude in a particular application, they should not be neglected.

In practice, ideal IIR filters are not realizable by means of finite lumped networks. There are two basic approaches to the design of realizable filters. The most common procedure is that in which the IIR filters are designed, and then standard approximation techniques are used [7] for the design of IIR filters with rational transfer functions, or for the design of finite impulse response filters. A more direct approach is the inclusion of the realizability constraints in the optimization procedure; this is the route that we follow in this work.

We consider two types of realizable filters: FIR filters and block filters. The first class is considered in Chapter 3, where we show that one of the advantages of the direct optimization of the FIR filters is that multidimensional filters can be designed almost as easily as their 1-D counterparts. Unlike the results in Chapter 2, we cannot derive closed-form solutions for the optimal FIR filters. It is possible, though, to obtain the optimal post-filter for a given pre-filter as the solution of a system of linear equations whose coefficients can be easily determined. The same is true for the computation of the optimal pre-filter for a given post-filter. The joint optimization algorithm is an iterative procedure that alternates between computing an optimal pre-filter and an optimal post, until there is a negligible error improvement.

Block filters are the subject of Chapter 4. One advantage of that class of filters is that they are not restricted to be shift-invariant operators. In fact, our approach towards the design of block filters is to view them as matrices, operating on a signal block that may have been obtained from the input signal in any convenient way. Thus, without any changes in the formulation, the results are applicable to systems where the input block is formed by consecutive time samples of a one-dimensional

signal, or by the elements of a vector signal, or even by a combination of both, so long as the second-order statistics of the input block are known. Closed-form solutions can be derived for the optimal block filters in most cases, although they may contain factors that require computation of the eigenvectors of positive definite matrices. In some cases they will also contain factors that require the solution of an inverse eigenvalue problem, i.e., the design of a matrix with prescribed eigenvalues. An algorithm for our version of that problem is presented in Appendix A.

In Chapter 4 we also consider a system that includes a non-linear component, which models the quantization process in digital channels. Basically, the model consists of an additive white noise source and a gain factor, both of which are functions of the signal energy. This is similar to the describing-function approach for the frequency analysis of non-linear systems [8]. In that analysis, we have obtained two important new results. First, we show that the optimality of the Karhunen-Loève Transform for block signal coding does not require the assumption of a Gaussian probability distribution for the input signal, as in earlier works [9], [10]. Second, we show that the use of pseudo-random noise (PRN) in the quantizers can improve the overall system performance.

Our research reported in Chapter 4 raises an issue that is common to block signal processing: the reconstructed signal has discontinuities at the block boundaries. This is the so-called 'blocking effect' [11]. In Chapter 5 we derive optimal pre- and post-filters for block processing based on overlapping basis functions, which virtually eliminate the blocking effects. The resulting transform operators, which we will refer to as the "Lapped Orthogonal Transform" (LOT) after Cassereau [12], are also more efficient than the non-overlapping transforms of Chapter 4, in terms of leading to lower mean-square errors.

There is one important final point relevant to the application of the models of Chapters 2–5 to real world systems that must be observed. In many applications, the characteristics of the signal change with time or space. In speech processing, for example, one cannot use the same spectral representations for voiced and unvoiced

sounds [13]. In buffered digital communications, if fewer bits are spent whenever or wherever the input signal has a decreased level of activity, then more bits will be available to represent the detailed portions of the signal. Thus, whenever possible the communications system of Fig. 1.1 must be adaptive.

The block processing models of Chapters 4 and 5 are better suited for the optimal design of adaptive systems, because all we need in order to optimize the processing of a particular signal block is a probabilistic description of that block. This inherent independence among blocks allows us, in principle, to process each block optimally, so that overall system performance is maximized. In practice, it is likely that there exists well-defined classes of typical signal blocks, so that we could have a few classes of optimal pre- and post-filters that could be applied successfully to most incoming signals. The definition of such classes for particular kinds of signals is a modeling problem that is not addressed in this work.

References

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* Chicago: University of Illinois Press, 1963.

[2] B. G. Cramer, "Optimal linear filtering of analog signals in noisy channels," *IEEE Trans. Audio Electroacoust.*, vol. AU-14, pp. 3–15, Mar. 1966.

[3] L. W. Couch II, *Digital and Analog Communications Systems.* New York: Macmillan, 1983.

[4] L. D. Fielder, "Pre- and postemphasis techniques as applied to audio recording systems," *J. Audio Eng. Soc.*, vol. 33, pp. 649–657, Sept. 1985.

[5] H. Nakajima *et al.*, *The Sony Book of Digital Audio Technology.* Blue Ridge Summit, PA: TAB Books, 1983.

[6] N. S. Jayant and P. Noll, *Digital Coding of Waveforms, Principles and Applications to Speech and Video.* Englewood Cliffs, N.J.: Prentice-Hall, 1984, chapters 4 and 12.

[7] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1975, chapter 5.

[8] P. Vidal, *Non-linear Sampled-data Systems.* New York: Gordon and Breach, 1969, chapters 4 and 9.

[9] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, Sept. 1963.

[10] A. Segall, "Bit allocation and encoding for vector sources," *IEEE Trans. Inform. Theory,* vol. IT-22, pp. 162–169, Mar. 1976.

[11] H. C. Reeve III and J. S. Lim, *Reduction of blocking effect in image coding.* in Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Boston, MA, 1983, pp. 1212–1215.

[12] P. Cassereau, *A new class of optimal unitary transforms for image processing.* S. M. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, May 1985.

[13] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Chapter 2

Optimal Filters with Infinite Impulse Response

In this chapter we study the problem of designing jointly-optimal pre- and post-filters without realizability constraints, so that the filter impulse responses extend from $-\infty$ to $+\infty$. Infinite impulse response (IIR) solutions are important from a theoretical viewpoint, since they provide performance bounds for realizable systems. In practice, we can approximate the ideal filter responses to any desired precision by rational transfer functions or by finite impulse response (FIR) filters. By increasing the order of the approximations, we can get arbitrarily close to the error bounds.

Continuous-time signals transmitted through continuous- and discrete-time channels will be analyzed in Sections 2.1 and 2.2, respectively. In Section 2.3 we extend the results to discrete-time signals. Some basic properties of optimal filter pairs that also hold for the realizable filter models to be considered in subsequent chapters will be derived. Among these properties is the concept of 'half-whitening', which means that the optimal pre-filter sends through the communications channel a signal whose spectrum is approximately the square root of the input spectrum, if the channel noise is white. This is a somewhat non-intuitive but important result that can be applied to the design of efficient sub-optimal filters.

2.1. Continuous-Time Signals and Channels

We consider here the system depicted in Fig. 2.1. The filters $F(\omega)$ and $G(\omega)$ are linear, time-invariant. The input signal $x(t)$ and the input and channel noises, $u(t)$ and $d(t)$, respectively, are stationary, uncorrelated, and zero-mean random processes with known spectra. Although the stationarity assumption may not hold for long time periods, most practical signals have slowly-varying statistics, and so we could use the results of this chapter to derive slowly-varying optimal filters. We do not assume that $F(\omega)$ and $G(\omega)$ are causal, i.e., their impulse responses extend infinitely for both positive and negative time. Therefore, zero-delay solutions are allowed.

The error signal $e(t)$ is defined in Fig. 2.2 as a filtered version of the absolute error $\hat{x}(t) - x(t)$. The filter $W(\omega)$ is a frequency weight that can be appropriately chosen according to the particular application. For example, in an audio system $W(\omega)$ would represent the frequency response of the human ear. We will refer to $W(\omega)$ as the 'observer response', and will assume $W(\omega) \neq 0$, $\forall \omega$, in order to avoid the existence of error components with zero weighting; if this restriction were not imposed we would certainly run into singularities and non-uniqueness problems.

The error measure ξ is defined as the energy, or variance, of $e(t)$,

$$\xi \triangleq \mathrm{E}[e^2(t)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{ee}(\omega) d\omega, \tag{2.1}$$

where $\Phi_{ee}(\omega)$ is the power spectrum of $e(t)$. Whenever we use the word 'error' in the following discussion, we will be referring to ξ.

The problem of jointly optimizing $F(\omega)$ and $G(\omega)$ for systems without sampling was first considered by Costas [1], without the input noise $u(t)$ and for an absolute mean-square error criterion, i.e., $w(t) = \delta(t)$, the Dirac delta function. The error improvement due to pre-filtering reported in [1] was atypically low, due

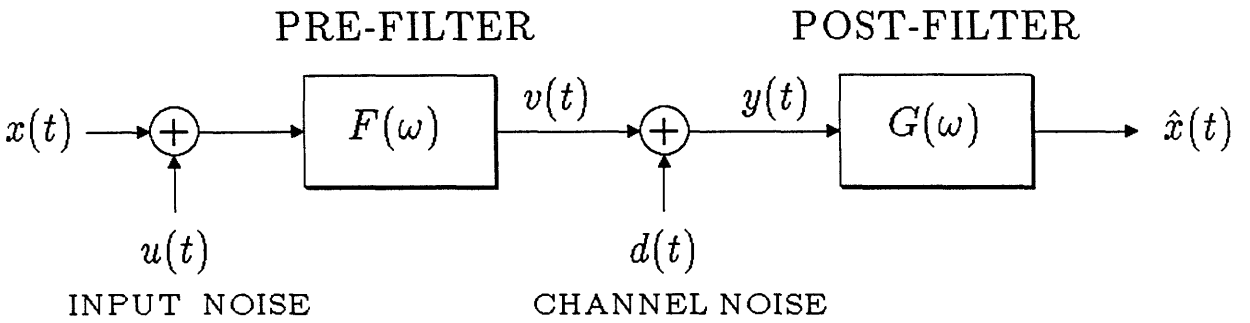


Fig. 2.1. Continuous-time system model.

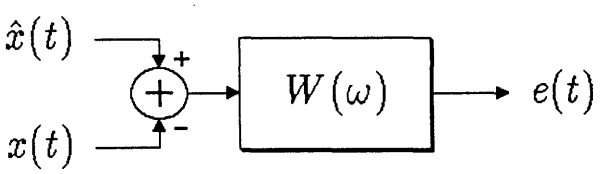


Fig. 2.2. Error signal definition.

to an unfortunate choice of system parameters. In [2] it was pointed out that significantly larger error improvements than those reported by Costas can be obtained. Cramer [3] has also studied the problem, re-deriving Costas' results and considering the constraint of reciprocal pre- and post-filters, i.e., $G(\omega) = F^{-1}(\omega)$, for which the solution depends on the channel noise spectrum but not on the amplitude of that noise. As in our present work, Cramer considered a frequency-weighted error criterion, but his analysis was purely intuitive, and although he obtained the correct

solution for the case of reciprocal filters, his approach would have led him to an erroneous conclusion for the general case, had he analyzed it.

Our work in this Section is an extension of the results of Costas and Cramer, for noisy inputs and a weighted mean-square error criterion. In what follows we derive the optimal $G(\omega)$ and $F(\omega)$, with the only restriction on the filters being a power limitation on the pre-filter output. We will start by obtaining the optimal $G(\omega)$ for a given pre-filter; that allows us to derive an error expression that depends only on $F(\omega)$. By finding the pre-filter that minimizes the new error function, we effectively obtain the jointly-optimal filter pair.

2.1.1. The Optimal Post-Filter

The problem of finding the optimal post-filter $G(\omega)$ for a given pre-filter $F(\omega)$ is in the form of Wiener's optimal estimation problem [4],[5]. Given the received signal $y(t)$, the post-filter has to generate the optimal estimate $\hat{x}(t)$ of the original signal $x(t)$. The error ξ is given by (2.1), with

$$\begin{aligned}\Phi_{ee}(\omega) =& \mid W(\omega) \mid^2 \mid F(\omega)G(\omega) - 1 \mid^2 \Phi_{xx}(\omega) \\ &+ \mid W(\omega)G(\omega) \mid^2 \Phi_{dd}(\omega) + \mid W(\omega)F(\omega)G(\omega) \mid^2 \Phi_{uu}(\omega) \ .\end{aligned} \tag{2.2}$$

We could substitute (2.2) into (2.1) and apply variational calculus to derive the optimal $G(\omega)$, but an easier route is to make use of the *orthogonality principle* [5],[6], of optimal estimation, which states that the estimation error has to be orthogonal to the received signal. This means $\mathrm{E}[e(t_1)y(t_2)] = 0$, $\forall t_1, t_2$, that is,

$$\begin{aligned}\Phi_{ey}(\omega) &= W(\omega)G(\omega)\{\Phi_{dd}(\omega) + \mid F(\omega) \mid^2 \left[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)\right]\} \\ &\quad - W(\omega)F^*(\omega)\Phi_{xx}(\omega) \\ &= 0 \ ,\end{aligned} \tag{2.3}$$

where * denotes complex conjugation. Since $W(\omega) \neq 0, \forall \omega$, we have

$$G_{\mathrm{OPT}}(\omega) = \frac{F^*(\omega)\Phi_{xx}(\omega)}{|\, F(\omega)\,|^2\,[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] + \Phi_{dd}(\omega)} \,. \tag{2.4}$$

We note that $G_{\mathrm{OPT}}(\omega)$ does not depend on the observer response, for a given $F(\omega)$, the reason being that the optimal post-filter actually minimizes $\Phi_{ee}(\omega)$ for all ω. This is, in fact, a particular case of a more general property of minimum-variance estimators in linear spaces, namely that the optimal estimator is invariant under non-singular error weighting [6].

The optimal post-filter can be factored in the form

$$G_{\mathrm{OPT}}(\omega) = \frac{1}{F(\omega)}\Big[\frac{\Phi_{xx}(\omega)}{\Phi_{xx}(\omega) + \Phi_{\bar{u}\bar{u}}(\omega)}\Big] \,, \tag{2.5}$$

which is a cascade of the inverse pre-filter response and the Wiener filter for the noise $\bar{u}(t)$. The latter is the equivalent noise source when the channel noise is mapped into the input, so that its spectrum is given by

$$\Phi_{\bar{u}\bar{u}}(\omega) = \Phi_{uu}(\omega) + \frac{\Phi_{dd}(\omega)}{|\, F(\omega)\,|^2} \,. \tag{2.6}$$

The above interpretation holds only for the frequencies for which $F(\omega) \neq 0$, since the channel noise spectrum cannot be mapped back into an equivalent input spectrum at any frequency ω_o for which $F(\omega_o) = 0$.

2.1.2. Optimal Pre-Filtering

When we use the optimal post-filter in (2.4), the error can be written as a function of the pre-filter, in the form

$$\xi = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\, W(\omega)\,|^2 \, \frac{|\, F(\omega)\,|^2\, \Phi_{uu}(\omega) + \Phi_{dd}(\omega)}{|\, F(\omega)\,|^2\, \left[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)\right] + \Phi_{dd}(\omega)}\, \Phi_{xx}(\omega)\; d\omega\,. \qquad (2.7)$$

Our objective is to minimize (2.7) by proper choice of $F(\omega)$, which would lead us to the jointly-optimal pair of filters, in view of (2.4). For every ω, the integrand in (2.7) is a monotonic function of $|\, F\,|^2$, which is minimized when $F(\omega) \rightarrow \infty$. But that would require an infinite power at the pre-filter output, which certainly makes no practical sense. Therefore, we must add a power constraint to the problem, e.g., by forcing the pre-filter output to have unit average power:

$$\mathrm{E}[v^2(t)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\, F(\omega)\,|^2\, \left[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)\right]\; d\omega \;=\; 1\;. \qquad (2.8)$$

Minimization of (2.7) is a problem of calculus of variations with variable end-points [7], since no restrictions are imposed on the values of $F(\omega)$. Therefore, $F(\omega)$ can be viewed as a free variable for any ω, and so we can make use of the Lagrange multiplier rule, which states that there must exist a scalar λ such that

$$\frac{\partial}{\partial F(\omega)}\Upsilon(F(\omega),\omega) + \lambda\, \frac{\partial}{\partial F(\omega)}\Psi(F(\omega),\omega) = 0, \qquad (2.9)$$

where $\Upsilon(F(\omega),\omega)$ and $\Psi(F(\omega),\omega)$ are the integrands in (2.7) and (2.8), respectively. By solving (2.9) we get the optimal pre-filter as

$$|\, F_{\mathrm{OPT}}(\omega)\,|^2 = \begin{cases} Z(\omega), & \text{if } Z(\omega) > 0, \\ 0, & \text{otherwise} \end{cases} \qquad (2.10)$$

where

$$Z(\omega) = \frac{\Phi_{xx}(\omega)}{\Phi_{xx}(\omega) + \Phi_{uu}(\omega)} \left[\mid W(\omega) \mid \sqrt{\frac{\Phi_{dd}(\omega)}{\lambda\left(\Phi_{xx}(\omega) + \Phi_{uu}(\omega)\right)}} - \frac{\Phi_{dd}(\omega)}{\Phi_{xx}(\omega)} \right] , \qquad (2.11)$$

with λ adjusted so that (2.8) is satisfied. We note that (2.10) specifies only the magnitude of the optimal pre-filter. The phase response is irrelevant, since it is canceled by the term $F^*(\omega)$ in (2.4).

If $\Phi_{xx}(\omega)$ decays faster than $\Phi_{dd}(\omega)$ there exists a frequency ω_c such that $F(\omega) = 0$ for $\mid \omega \mid > \omega_c$. Therefore the optimal pre- and post-filters will be band-limited to ω_c. The channel bandwidth must be at least equal to ω_c if additional errors are to be avoided. Nevertheless, even if the channel bandwidth is lower than ω_c, we can still use (2.4),(2.10), and (2.11) to compute the optimal $F(\omega)$ and $G(\omega)$ for $\omega < \omega_c$, with the integral in (2.8) being evaluated over the interval $(-\omega_c, \omega_c)$. In this case the minimum error with $F_{\mathrm{OPT}}(\omega)$ and $G_{\mathrm{OPT}}(\omega)$ is given by

$$\begin{aligned} \xi_{\min} = & \frac{1}{\pi} \int_0^{\omega_c} \mid W(\omega) \mid^2 \frac{\Phi_{xx}(\omega)\Phi_{uu}(\omega)}{\Phi_{xx}(\omega) + \Phi_{uu}(\omega)} \, d\omega \\ & + \frac{1}{\pi} \int_0^{\omega_c} \mid W(\omega) \mid^2 \Phi_{xx}(\omega) \sqrt{\frac{\lambda \Phi_{dd}(\omega)}{\Phi_{xx}(\omega) + \Phi_{uu}(\omega)}} \, d\omega \\ & + \frac{1}{\pi} \int_{\omega_c}^{\infty} \mid W(\omega) \mid^2 \Phi_{xx}(\omega) \, d\omega \, . \end{aligned} \qquad (2.12)$$

If the input noise $u(t)$ has an amplitude negligible compared to that of $x(t)$, the pre-filter output $v(t)$ has the spectrum

$$\begin{aligned} \Phi_{vv}(\omega) = & \mid F_{\mathrm{OPT}}(\omega) \mid^2 \Phi_{xx}(\omega) \\ = & \mid W(\omega) \mid \sqrt{\frac{\Phi_{dd}(\omega)\Phi_{xx}(\omega)}{\lambda}} - \Phi_{dd}(\omega) \, . \end{aligned} \qquad (2.13)$$

Intuitively, we would expect that for a white channel noise the pre-filter output should have a flat spectrum, to keep a frequency-independent SNR. As we see from

(2.13), however, even for an unweighted mean-square error criterion, ($W(\omega) \equiv 1$), the pre-filter output spectrum is not white for a white channel noise, although the square root on (2.13) makes $\Phi_{vv}(\omega)$ flatter than $\Phi_{xx}(\omega)$. On a logarithmic scale, amplitude variations on $\Phi_{xx}(\omega)$ are divided by two. This is referred to as the *half-whitening* effect [2] of the optimal pre-filter.

If the channel is noiseless, i.e. $\Phi_{dd}(\omega) = 0$, eqn. (2.10) is not applicable, since (2.7) is independent of the pre-filter. In this case the pre-filtering concept actually loses its meaning, and any pre-filter that satisfies the power constraint (2.8) can be used.

Example 2.1

In order to evaluate the improvement due to pre-filtering over optimal post-filtering only, let's consider a white channel noise, $\Phi_{dd}(\omega) = N_o$, no input noise, $\Phi_{uu}(\omega) = 0$, no error weighting, $W(\omega) = 1$, and an input signal with a first-order Butterworth spectrum

$$\Phi_{xx}(\omega) = \frac{2\omega_o}{\omega^2 + \omega_o^2} \ .$$

Let's compare two alternatives:

i) Optimal post-filtering only, $F(\omega) = 1$. From (2.7) the error is

$$\begin{aligned} \xi_1 &= \frac{1}{\pi} \int_0^\infty \frac{2\omega_o N_o}{2\omega_o + N_o(\omega^2 + \omega_o^2)} \, d\omega \\ &= \left(1 + \frac{2}{N_o \omega_o}\right)^{-1/2} . \end{aligned}$$

ii) Optimal pre- and post-filtering. Using (2.10) and (2.11), we obtain the optimal pre-filter as

$$|\, F_{\mathrm{OPT}}(\omega) \,|^2 = \begin{cases} \dfrac{N_o}{2\omega_o} \left[\sqrt{(\omega_c^2 + \omega_o^2)(\omega^2 + \omega_o^2)} - (\omega^2 + \omega_o^2) \right] , & |\, \omega \,| \leq \omega_c \ , \\ 0, \quad \text{otherwise} \end{cases}$$

where we have used the fact that the Lagrange multiplier is related to the cutoff frequency ω_c by

$$\lambda = \frac{2\omega_o}{N_o(\omega_c^2 + \omega_o^2)} ,$$

since $Z(\omega_c) = 0$ in (2.11). The minimum error, as determined by (2.12), is

$$\begin{aligned}\xi_2 &= \frac{1}{\pi}\int_0^{\omega_c} \sqrt{\frac{2\omega_o N_o}{\omega^2 + \omega_o^2}}\, d\omega + \frac{1}{\pi}\int_{\omega_c}^{\infty} \frac{2\omega_o}{\omega^2 + \omega_o^2}\, d\omega \\ &= \frac{2\omega_o}{N_o(\omega_c^2 + \omega_o^2)}\left(1 + \frac{N_o\omega_c}{\pi}\right) + 1 - \frac{2}{\pi}\tan^{-1}\left(\frac{\omega_c}{\omega_o}\right) .\end{aligned}$$

The cutoff frequency ω_c is implicitly determined by the power constraint (2.8), which leads to

$$\sqrt{\omega_c^2 + \omega_o^2}\,\log\left(\frac{\omega_c + \sqrt{\omega_c^2 + \omega_o^2}}{\omega_o}\right) - \omega_c = \frac{\pi}{N_o} .$$

As N_o is decreased, ξ_1 and ξ_2 are both decreased, but ξ_2 does so faster, so that $\xi_1 - \xi_2$ increases. A plot of the ratio ξ_1/ξ_2 as a function of ξ_1 is shown in Fig. 2.3, in a log-log scale. For channels with high SNR the improvement can be higher than 25 dB. We note, however, that the required channel bandwidth may be several times ω_o. For example, if $\xi_1 = 0.01$, the input signal and channel noise spectra have the same amplitude, N_o, at the frequency $\omega_c \simeq 14\omega_o$. The half-whitening effect of the optimal pre-filter pushes this crossover point to $\omega_c = 2100\omega_o$. In a broadcasting system, for example, such a bandwidth increase would certainly be unacceptable. A better channel noise model would be to take into account any other signal sources onto frequency-adjacent channels and include them in the channel noise spectrum. In Example 2.2 a more realistic model is considered.

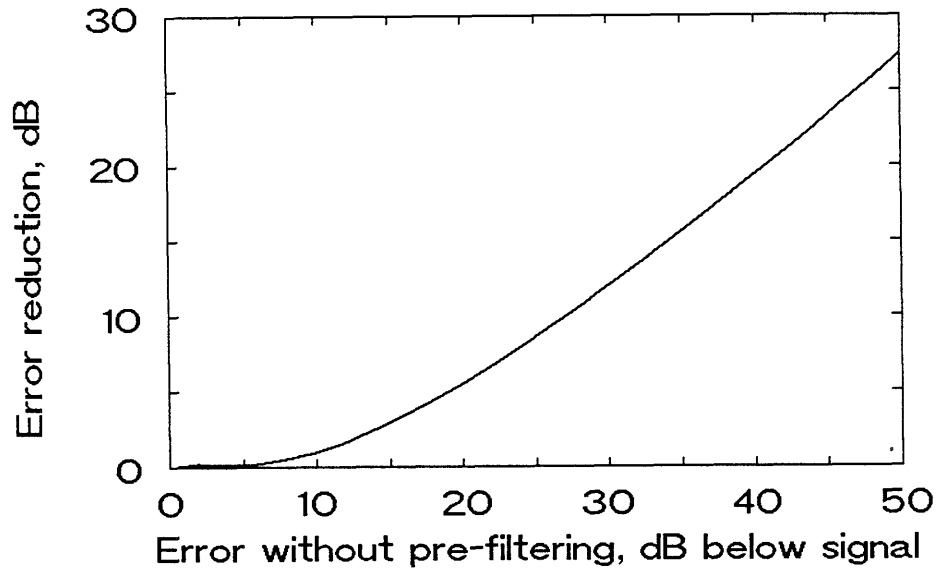


Fig. 2.3. Error improvement due to optimal pre-filtering.

2.2. Continuous-Time Signals on Discrete-Time Channels

The system model is now that of Fig. 2.4. It is a useful model for either pulse amplitude modulation (PAM) or pulse code modulation (PCM) communications systems. In any case, the post-filter is actually an interpolator, since it produces a continuous-time signal from discrete-time samples. In Fig. 2.5 we have an equivalent model, in which sampling is represented as multiplication by the periodic sampling function

$$\delta_T(t) \triangleq T \sum_{r=-\infty}^{\infty} \delta(t - rT),$$

where T is the sampling period. The gain factor T in the above equation was introduced mainly to avoid the presence of scaling factors in the periodic spectral repetitions that we will encounter later.

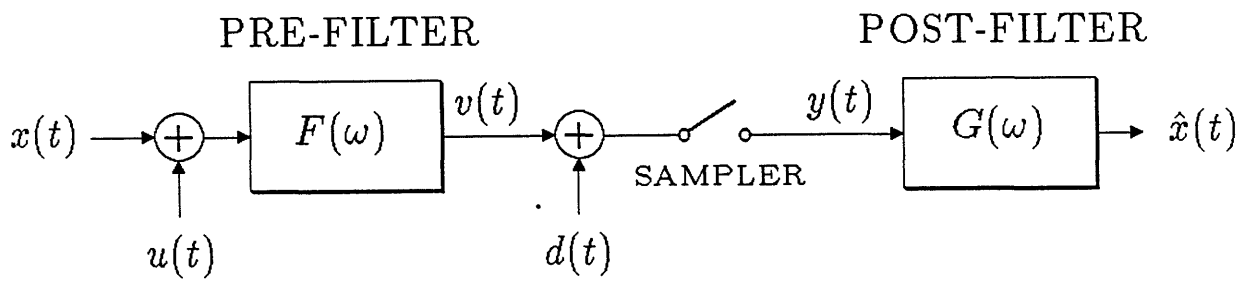


Fig. 2.4. System with a sampled-data channel.

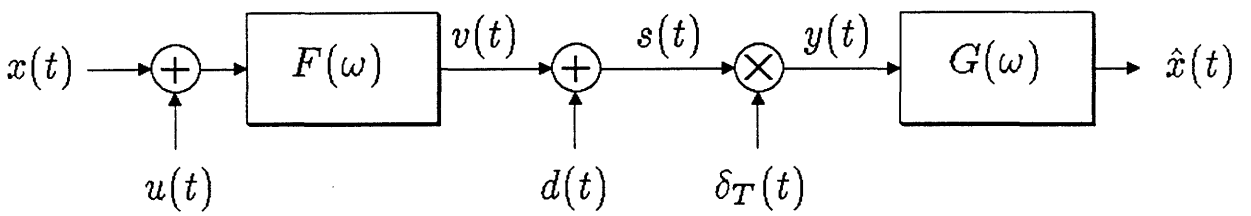


Fig. 2.5. The sampler as a multiplier.

If a PAM system is under consideration, the channel noise $d(t)$ can be assumed to be uncorrelated with $v(t)$. For a PCM system, however, the noise $d(t)$ is due mainly to quantization and will be in general correlated with $v(t)$, unless pseudo-random noise is employed [12]. If the quantizer is optimal [8] in a mean-square error

sense, its output is orthogonal (and therefore uncorrelated) to the quantization noise, for any number of quantization levels [9]. So, we can write

$$\begin{aligned} \mathrm{E}\{[v(t)+d(t)]d(t)\} = 0 \Rightarrow \mathrm{E}\{v(t)d(t)\} = -\mathrm{E}\{d(t)d(t)\} \\ \Rightarrow R_{vd}(0) = -R_{dd}(0) = -\epsilon^2 R_{vv}(0), \end{aligned} \tag{2.14}$$

where, as usual, $R_{vd}(\tau) \triangleq \mathrm{E}[v(t)d(t-\tau)]$, and ϵ^2 is the inverse of the signal-to-noise ratio (SNR) for the quantizer. Furthermore, Chan and Donaldson [10] have shown that the cross-correlation function between quantizer input and quantization noise is given, for Gaussian signals, by

$$R_{vd}(\tau) = \beta R_{vv}(\tau) \quad \Rightarrow \quad \Phi_{vd}(\omega) = \beta \Phi_{vv}(\omega) \ . \tag{2.15}$$

Combining (2.14) and (2.15) we conclude that $\beta = -\epsilon^2$.

Although a detailed evaluation of the parameter ϵ^2 for optimal quantizers has been carried out by Mauersberger [11], for several probability density functions, we must proceed under the assumption that the input signal and the input noise have Gaussian distributions. Otherwise, the p.d.f. of the pre-filter output signal would be a function of the pre-filter, and thus ϵ^2 would also depend on $F(\omega)$; such a dependence would be complicated enough to render the problem intractable.

Due to the increased interest in PCM over the last two decades, the system in Fig. 2.5 has been considered by many researchers. Stewart [13] was the first to bring up the fact that in practice signals to be sampled are never exactly band-limited to half the sampling frequency, and so by proper evaluation of the aliasing effects the reconstruction filter $G(\omega)$ can be designed to minimize the total error variance. Stewart's model was that of Fig. 2.5, with $u(t) \equiv 0$ and $F(\omega) \equiv 1$. One of the main contributions in [13] was the observation that the optimal reconstruction filter $G(\omega)$ is time-invariant, i.e., there is no time-varying impulse response $g(t,\tau)$ that leads to a lower reconstruction error than the optimal time-invariant post-filter. This

would probably not hold if the stationarity assumption for the input signal were removed. Tufts and Johnson [14] extended Stewart's result to the case when the available number of samples in $y(t)$ is finite. They pointed out the important fact that polynomial interpolation produces a mean-square error that is *always* larger than that of optimal linear time-invariant interpolation.

Spilker [15] seems to have been the first to consider the optimal design of a pre-sampling filter, for a system without input noise ($u(t) = 0$), showing that under a mean-square error criterion the optimal pre-filter must be band-limited, in order to avoid aliasing. Brown [16] studied the case of a noisy input and a noiseless channel, and noted that the optimal pre- and post-filters should still be band-limited, a fact also verified by Ericson [17]. The optimal causal pre- and post-filters for a noiseless channel and a weighted mean-square error criterion were derived by DeRusso [18] and Chang [19], with a few errors corrected in [20].

The first analysis of the system of Fig. 2.5 for the specific case of a pulse-code-modulation (PCM) system, in which $d(t)$ is generated by a minimum-mean-square-error Max quantizer [8] , was performed by Kellogg [21], [22], who was not able to jointly optimize the pre-filter, quantizer and the post-filter, and so the optimal pre-post pair was numerically derived under the assumption of a noiseless channel. Kellogg's objective was to compare the PCM system with optimal pre- and post-filtering to the rate distortion bound [23] of Information Theory. Chan and Donaldson [24] have refined Kellogg's work by taking into account the cross-correlation between $d(t)$ and $v(t)$ in an analytical optimization procedure for the pre- and post-filters. They have derived precise bounds on how close a PCM system with optimal pre- and post-filters can get to the rate distortion bound. In a later work, Noll [25] obtained performance bounds under the assumption that $G(\omega) = F^{-1}(\omega)$ (which Noll referred to as a D*PCM system).

In the literature cited above, we notice that the system of Fig. 2.5 was analyzed with either $u(t)$ or $d(t)$ or both set to zero, and in most cases for an unweighted error criterion. The most complete analysis was that of Chan and Donaldson [24],

where not only $d(t)$ was taken into account, but also a weighted mean-square error criterion. The purpose of this section is to extend the work of Chan and Donaldson to the general case of Fig. 2.5, in which both noise sources may be present, as in the case of telemetry systems on a low bit rate digital channel. As we will see later, such an extension is justified by the fact the the performance of the optimal system is strongly dependent on the input noise characteristics, even at high signal-to-noise ratios.

2.2.1. The Optimal Post-Filter

As in the previous analysis for systems without sampling, we start by deriving the optimal post-filter, or interpolator, for a fixed pre-filter. We have to be careful in terms of defining the error criterion, since the signals $y(t)$ and $\hat{x}(t)$ in Fig. 2.5 are not stationary and, therefore, cannot be characterized by a power spectrum in a conventional way. However, they are cyclostationary [26], i.e., their statistics are periodically time-varying, and so by averaging over one sampling period we can obtain time-invariant statistical descriptions that have the same meaning as those of stationary processes. Therefore, as long as autocorrelation functions are averaged over one sampling period, we can still refer to their Fourier transforms as power spectra. This concept was used in most of the literature cited in this section. Our error measure should, then, be taken as the average variance of the signal $e(t)$ over one period,

$$\xi \triangleq \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{E}[e^2(t)]\,dt\ . \tag{2.16}$$

We note that any positive weighting function could have been used in the above equation, but since the original signal is stationary it is unlikely that a time-weighting function could be of any practical value.

Calling $s(t) = v(t) + d(t)$ the input to the sampler in Fig. 2.5, we can write the autocorrelation function of $y(t)$ as

$$\begin{aligned} R_{yy}(\tau) &= \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{E}[y(t)y(t-\tau)]\,dt \\ &= R_{ss}(\tau)\,\frac{1}{T}\int_{-T/2}^{T/2} \delta_T(t)\delta_T(t-\tau)\,dt \\ &= R_{ss}(\tau)\delta_T(\tau), \end{aligned} \tag{2.17}$$

from which

$$\Phi_{yy}(\omega) = \sum_{k=-\infty}^{\infty} \Phi_{ss}(\omega + k\omega_T), \tag{2.18}$$

where $\omega_T \triangleq 2\pi/T$ is the sampling frequency. Similarly, the cross-correlation between $s(t)$ and $y(t)$ is

$$\begin{aligned} R_{sy}(\tau) &= \frac{1}{T}\int_{-T/2}^{T/2} \mathrm{E}[s(t)y(t-\tau)]\,dt \\ &= R_{ss}(\tau)\,\frac{1}{T}\int_{-T/2}^{T/2} \delta_T(t-\tau)\,dt \\ &= R_{ss}(\tau). \end{aligned} \tag{2.19}$$

where $R_{ss}(t)$ has the conventional meaning, since $s(t)$ is stationary. Thus, it is clear that periodic sampling produces periodic replication of power spectra, which may lead to aliasing, but cross-spectra are not affected. Using these two properties, it is relatively easy to extend the analysis of the previous section to include the effects of sampling.

In order to derive the optimal post-filter by means of the orthogonality principle, we need the spectrum $\Phi_{ey}(\omega)$ corresponding to our new error definition in (2.16). According to our previous discussion about quantization noise, we assume $\Phi_{vd}(\omega) = \beta\Phi_{vv}(\omega)$. Then, $\Phi_{ey}(\omega)$ is given by

$$\Phi_{ey}(\omega) = W(\omega)\{G(\omega)\Phi_{yy}(\omega) - F^*(\omega)[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)]\}\ , \tag{2.20}$$

where

$$\Phi_{yy}(\omega) = \sum_{k=-\infty}^{\infty} \Phi_{dd}(\omega + k\omega_T) + \sum_{k=-\infty}^{\infty} |\, F(\omega + k\omega_T)\,|^2\, (1+2\beta)\big[\Phi_{xx}(\omega + k\omega_T) + \Phi_{uu}(\omega + k\omega_T)\big]\ . \tag{2.21}$$

For a PAM system without quantization, β may be set to zero. We recall from the previous section that the orthogonality principle states that $G_{\mathrm{OPT}}(\omega)$ is the one for which $\Phi_{ey}(\omega) = 0$. Since $W(\omega) \neq 0, \forall\omega$, (2.20) leads to

$$G_{\mathrm{OPT}}(\omega) = \frac{F^*(\omega)[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)]}{\Phi_{yy}(\omega)}\ . \tag{2.22}$$

When $d(t) = 0$ and $\beta = 0$, the result above is the same as that obtained by Brown [16]. With $G(\omega) = G_{\mathrm{OPT}}(\omega)$, the error is

$$\begin{aligned}\xi &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{ee}(\omega)\ d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\, W(\omega)\,|^2\, \Phi_{xx}(\omega)\ d\omega \\ &\quad - \frac{1}{2\pi}\int_{-\infty}^{\infty} |\, W(\omega)\,|^2 |\, F(\omega)\,|^2\, \frac{[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)]^2}{\Phi_{yy}(\omega)} d\omega\ .\end{aligned} \tag{2.23}$$

2.2.2. The optimal pre-filter

The problem now is that of maximizing the second integral in (2.23), under the power constraint (2.8). A slight complication in (2.23) is that $F(\omega)$ affects the aliasing components of $\Phi_{yy}(\omega)$. Let's consider some frequency ω_o and the set of its aliasing images $A_o = \{\omega_o + k\omega_T,\ k = \pm1, \pm2, \ldots\}$, with ω_o leading to the maximum pre-filter power output, i.e.,

$$|\, F(\omega_o)\,|^2\, [\Phi_{xx}(\omega_o) + \Phi_{uu}(\omega_o)] \geq |\, F(\omega)\,|^2\, [\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] \forall\ \omega \in A_o\ .$$

Since $\Phi_{xx}(\omega) + \Phi_{uu}(\omega) \geq 0$, $\forall\ \omega$, the power constraint (2.8) states that if we increase $\mid F(\omega) \mid$ for any $\omega \in A_o$, we will have to reduce $F(\omega_o)$. But this forces a strict reduction in the second integrand in (2.23), which we want to maximize. Therefore, we have just proved the following

Lemma. *Assume $F(\omega)$ and $G(\omega)$ are jointly-optimal for the system in Fig. 2.5. Then, the pre-filter $F(\omega)$ avoids aliasing, i.e.,*

$$F(\omega_o) \neq 0 \Rightarrow F(\omega) = 0,\ \omega = \omega_o + k\omega_T,\ k = \pm 1, \pm 2, \ldots$$

From (2.22), $F(\omega_s) = 0 \Rightarrow G(\omega_s) = 0$, for any ω_s. So, the jointly-optimal pre- and post-filters have identical passbands, with a total bandwidth (including negative frequencies) of ω_T. If $\Phi_{xx}(\omega)$ decays faster than both noise spectra, the bandwidth of the optimal filter pair is $\mid \omega \mid < \omega_T/2$. In this case the error expression assumes the form

$$\begin{aligned}\xi = &\frac{1}{\pi}\int_{\omega_T/2}^{\infty} \mid W(\omega) \mid^2 \Phi_{xx}(\omega)\ d\omega \\ &+ \frac{1}{\pi}\int_{0}^{\omega_T/2} Q(\omega) \mid F(\omega) \mid^2 \{\Phi_{xx}(\omega)\Phi_{uu}(\omega) - \beta^2[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)]^2\}\ d\omega \\ &+ \frac{1}{\pi}\int_{0}^{\omega_T/2} Q(\omega)\Phi_{xx}(\omega)\tilde{\Phi}_{dd}(\omega)\ d\omega\ ,\end{aligned} \tag{2.24}$$

where

$$Q(\omega) \triangleq \frac{\mid W(\omega) \mid^2}{\mid F(\omega) \mid^2 (1+2\beta)[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] + \tilde{\Phi}_{dd}(\omega)}\ , \tag{2.25}$$

and

$$\tilde{\Phi}_{dd}(\omega) \triangleq \sum_{k=-\infty}^{\infty} \Phi_{dd}(\omega + k\omega_T)\ . \tag{2.26}$$

The last two integrals in (2.24) are similar to (2.7), with $\tilde{\Phi}_{dd}(\omega)$ replacing $\Phi_{dd}(\omega)$, and the introduction of the parameter β. The optimal pre-filter is

$$| F_{\mathrm{OPT}}(\omega) |^2 = \begin{cases} Z(\omega), & \text{if } | \omega | \leq \omega_T/2, \\ 0, & \text{otherwise} \end{cases} \tag{2.27}$$

with $Z(\omega)$ given by

$$\begin{aligned} Z(\omega) &= \frac{\Phi_{xx}(\omega)}{(1+2\beta)[\Phi_{xx}(\omega)+\Phi_{uu}(\omega)]} \\ &\times \left\{ \left[(1+\beta) + \beta \frac{\Phi_{uu}(\omega)}{\Phi_{xx}(\omega)} \right] | W(\omega) | \sqrt{\frac{\tilde{\Phi}_{dd}(\omega)}{\lambda(\Phi_{xx}(\omega)+\Phi_{uu}(\omega))}} - \frac{\tilde{\Phi}_{dd}(\omega)}{\Phi_{xx}(\omega)} \right\}, \end{aligned} \tag{2.28}$$

Thus, by using (2.10), (2.28) and (2.22), we can design the jointly-optimal filter pair. It is interesting to note that if $u(t) = 0$ and $d(t) = 0$, the first integral in (2.23) vanishes and the error is independent of the shape of $F(\omega)$ within the passband. The error is then due entirely to the missing signal spectra for $| \omega | > \omega_T/2$. If we had $F(\omega) = 1$ for all ω, with $u(t) = 0$ and $d(t) = 0$, the error with optimal post-filtering would be due entirely to aliasing. In the worst case, aliasing leads to twice the error level as missing signal spectra for $| \omega | > \omega_T/2$ [21],[22]. Therefore, optimal pre-filtering can reduce the error in a noiseless system by 3 dB, at most. For a noisy channel, i.e., $d(t) \not\equiv 0$, the error improvement can be somewhat larger, as in the next example.

Example 2.2

We conclude this section with the following design example for the system in Fig. 2.5: let's consider again an input signal with a Butterworth spectrum $\Phi_{xx}(\omega) = 2\omega_o/(\omega^2 + \omega_o^2)$, a white input noise, $\Phi_{uu}(\omega) = \sigma_u^2$, the error weighting

$$| W(\omega) |^2 = \frac{1 + a\omega^2/\omega_o^2}{1 + b\omega^2/\omega_o^2} ,$$

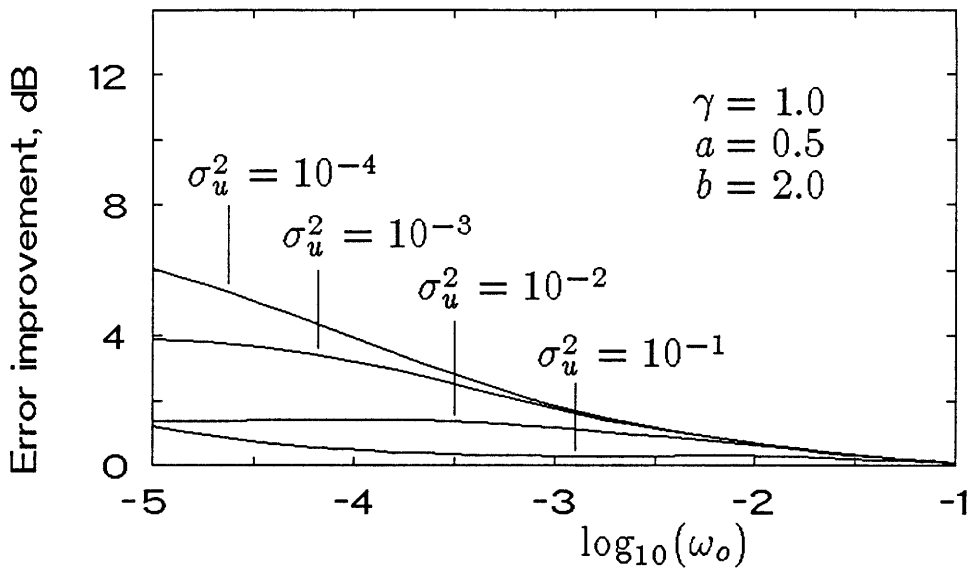
and let's consider a FSK (frequency-shift keying) channel, in which each sample of $v(t)$ determines the frequency of a carrier during an interval T. We assume a sampling period $T = 2\pi \rightarrow \omega_T = 1$. Because of the frequency modulation, the channel noise spectrum has the typical 6 dB/octave high-frequency rise given by $\Phi_{dd}(\omega) = \gamma\omega^2$ [27], where γ is a constant.

Unlike the previous example, instead of comparing the jointly-optimal filter pair to a system without pre-filtering, it would be more interesting to compare the optimal system to the one in which an ideal pre-filter is used, with the frequency response

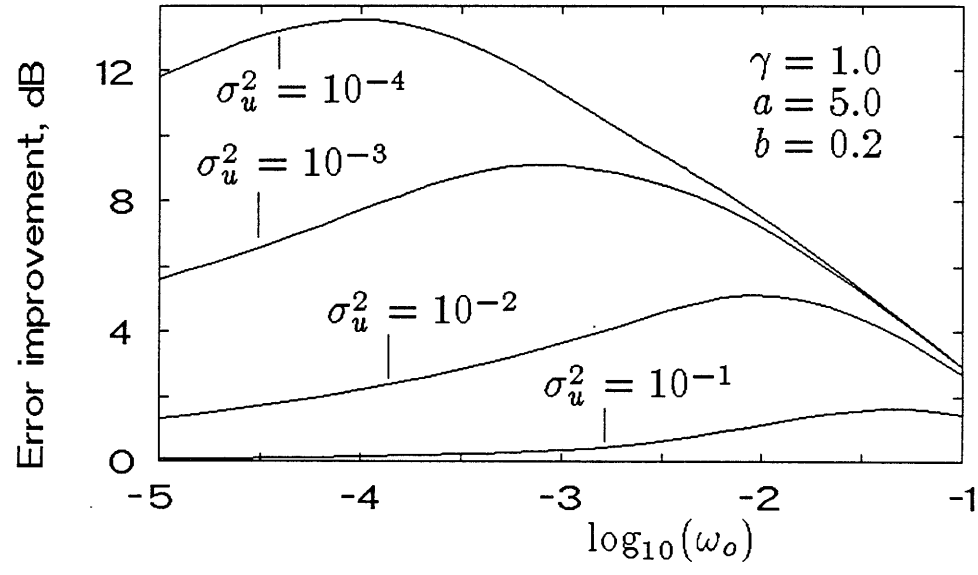
$$F(\omega) = \begin{cases} \alpha, & |\omega| \leq 1/2, \\ 0, & \text{otherwise}, \end{cases}$$

with α adjusted so that the pre-filter output has unit power. We cannot derive close-form expressions for computing the error as a function of the parameters a, b, γ, and σ_u^2, because the error integrals contain fourth-order rational integrands. Thus, our comparison is based on the numerical evaluation of the integrals.

In Fig. 2.6 we have plotted the error improvement by using the optimal pre- and post-filters, as compared to the flat pre-filter and its corresponding optimal post-filter. We have fixed $\gamma = 1$, so that a significant level of channel noise is present. In Fig. 2.6(a) we have $a = 5.0$ and $b = 0.2$, and so the weighting function has a strong high-frequency emphasis, whereas in Fig. 2.6(b) the values of a and b produce a low-frequency emphasis. In both cases ω_o was varied from 10^{-5} to 10^{-1}, and the input noise level set at four different values. We note that as ω_o is decreased, the error improvement increases, until a point is reached where the input spectrum becomes so strongly peaked that the input noise has less influence over the total noise. Thus, the error reduction due to optimal pre-filtering is less significant. In Fig. 2.6(b) the error improvements are much lower. This is because the frequency weighting is stronger at lower frequencies, where the channel noise has a lower amplitude, which reduces the benefits of optimal pre-filtering.



(b)



(a)

Fig. 2.6. Error reduction due to optimal pre-filtering (as compared to ideal low-pass pre-filtering with a constant passband gain) for a FSK channel.

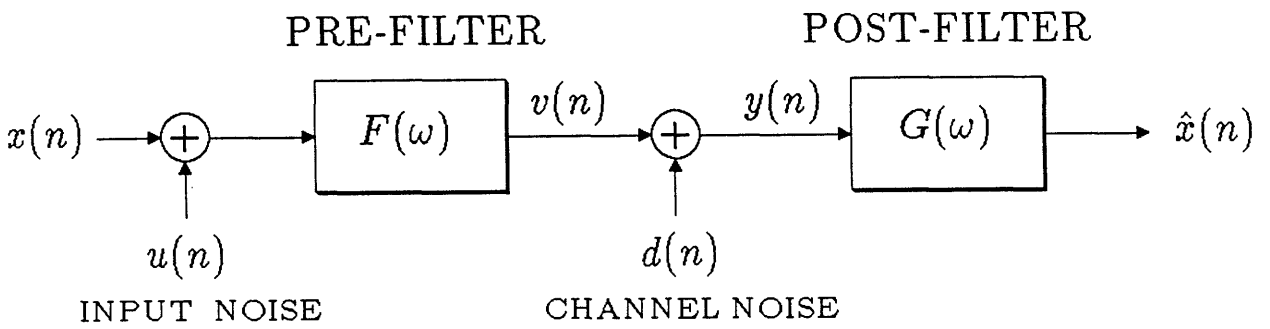


Fig. 2.7. Discrete-time system model.

The strong influence of the input noise on the performance of the optimal system is clear from Fig. 2.6. Thus, even if that noise has a low amplitude, it should always be taken into account in the design of optimal systems.

2.3. Discrete-Time Signals and Channels

In this section, we assume that the signal to be transmitted is available only in sampled form, so that the complete system works in discrete time. The system model is now that of Fig. 2.7, which is basically the same as the previous model in Fig. 2.1, except that the signals are functions of the discrete-time index n, and $F(\omega)$ and $G(\omega)$ are the Fourier transforms of the pre- and post-filter impulse response sequences $f(n)$ and $g(n)$, respectively.

If the sampling rate of the channel is lower than that of the original signal $x(n)$, as it is often the case, the output of the pre-filter must be down-sampled before

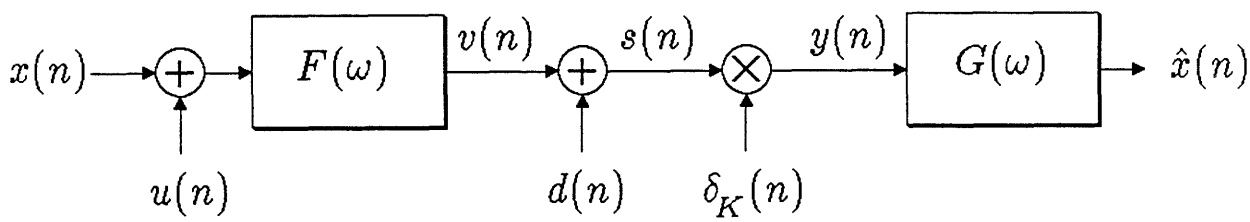


Fig. 2.8. Channel with down- and up-sampling by a factor of K.

transmission, and the received signal must be up-sampled before post-filtering. In this case, the model in Fig. 2.8 applies. Assuming that the channel sampling frequency is K times lower than that of the input signal, we can model the down- and up-sampling process in the channel as multiplication by the periodic sampling sequence

$$\begin{aligned} \delta_K(n) &\triangleq K \sum_{r=-\infty}^{\infty} \delta(n - rK) \\ &= \sum_{r=0}^{K-1} \exp\left(j\frac{2\pi rn}{K}\right), \end{aligned} \tag{2.29}$$

where $\delta(n)$ is the unit-sample (or impulse) sequence [28]. In practice, we can make use of polyphase filter structures [29] so that pre-filtering and down-sampling can be performed by a single block, as well as up-sampling and post-filtering.

Intuitively, we would expect that the results of the Sections 2.1 and 2.2 could be applied to the systems in Fig. 2.7 and Fig. 2.8, respectively, with little modification. This is indeed the case, as we will see in what follows. One important

difference between the systems for continuous- and discrete-time signals is that the system in Fig. 2.7 can be obtained from that in Fig. 2.8 simply by setting $K = 1$, whereas the systems in Fig. 2.4 and Fig. 2.1 cannot be made equivalent by letting $T \rightarrow 0$. This is a direct consequence of the fact that the unit-sample sequence contains no singularities, unlike the Dirac delta function. Therefore, we can focus on the system of Fig. 2.8, and any results will be applicable to Fig. 2.7 by setting $K = 1$.

In the general case where $K \neq 1$, the periodic sampling operation by $\delta_K(n)$ leads to periodic autocorrelation functions, with period K, i.e., the sampled signals are cyclostationary. So, as in the previous section, the error criterion should be an average of the the error signal variance over K samples,

$$\begin{aligned} \xi &\triangleq \frac{1}{K} \sum_{n=0}^{K-1} \mathrm{E}[e^2(n)] \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_{ee}(\omega)\, d\omega \ . \end{aligned} \tag{2.30}$$

Similarly, the correlation functions of interest should also be averaged over K samples, so that we can write

$$\begin{aligned} R_{yy}(n) &= \frac{1}{K} \sum_{r=0}^{K-1} \mathrm{E}[y(r)y(r-n)] \\ &= R_{ss}(n) \frac{1}{K} \sum_{r=0}^{K-1} \delta_K(r)\delta_K(r-n) \\ &= R_{ss}(n)\delta_K(n) \end{aligned} \tag{2.31}$$

and

$$\begin{aligned} R_{sy}(n) &= \frac{1}{K} \sum_{l=0}^{K-1} \mathrm{E}[s(r)y(r-n)] \\ &= R_{ss}(n) \frac{1}{K} \sum_{r=0}^{K-1} \delta_K(r-n) \\ &= R_{ss}(n) \ . \end{aligned} \tag{2.32}$$

The first of the two equations above implies

$$\Phi_{yy}(\omega) = \sum_{r=0}^{K-1} \Phi_{ss}(\omega + r\omega_K), \tag{2.33}$$

where $\omega_K \triangleq 2\pi/K$, that is, down-sampling and up-sampling by a factor K leads to K spectral replications, unlike the continuous-time case, where the number of replications is infinite. This is a well-known property of discrete-time systems [29].

As in the previous section, we can apply the orthogonality principle to the system in Fig. 2.8, in order to derive the optimal post-filter. Then, we can minimize the resulting error function to obtain the optimal pre-filter. The result is that (2.10), (2.22), and (2.28) can be used directly here, i.e., the optimal post-filter is given by

$$G_{\mathrm{OPT}}(\omega) = \frac{F^*(\omega)[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)]}{\Phi_{yy}(\omega)} \tag{2.34}$$

and the optimal pre-filter by

$$| F_{\mathrm{OPT}}(\omega) |^2 = \begin{cases} Z(\omega), & \text{if } Z(\omega) > 0, \\ 0, & \text{otherwise} \end{cases} \tag{2.35}$$

where $Z(\omega)$ is determined by (2.28), with

$$\begin{aligned} \Phi_{yy}(\omega) = & \sum_{r=0}^{K-1} \Phi_{dd}(\omega + r\omega_K) \\ & + \sum_{r=0}^{K-1} | F(\omega + r\omega_K) |^2 \left[(1+2\beta)\Phi_{xx}(\omega + r\omega_K) + \Phi_{uu}(\omega + r\omega_K)\right] , \end{aligned} \tag{2.36}$$

and

$$\tilde{\Phi}_{dd}(\omega) \triangleq \sum_{r=0}^{K-1} \Phi_{dd}(\omega + r\omega_K) \ . \tag{2.37}$$

The optimal filters will be band-limited, with a total bandwidth of ω_K. As we have noted before, by setting $K = 1$ in (2.36) and (2.37), we can compute the optimal filters for the system of Fig. 2.7; in this case $\omega_K = 2\pi$, and the optimal filters are not necessarily band-limited, since there is no aliasing.

Example 2.3

In this last example for this chapter, we use the previous analysis to derive a new bound for the error improvement that jointly-optimal pre- and post-filtering provides over optimal post-filtering only. We assume that the system is noiseless, i.e., $d(n) = u(n) = 0$, no error weighting is assumed, $W(\omega) = 1$, and the signal is modeled as a first-order Gauss-Markov process. We also show that in this case the optimal post-filter corresponding to no pre-filtering has a finite impulse response.

We have seen before that for the case of a continuous-time input signal and a discrete-time channel, the error reduction with optimal pre-filtering is bounded at 3 dB. A simple proof of that, presented in [9] and [21], is that if we use no pre-filter and an ideal low-pass post-filter we get exactly twice the error of that obtained with both the pre- and post-filters having an ideal low-pass response (the jointly-optimal solution for the noiseless case). This bound is not tight, however, since (2.22) shows that if the pre-filter is not band-limited the corresponding optimal post-filter will not be band-limited, either.

If the input signal is a raster-scanned image, for example, the first-order Gauss-Markov autocorrelation

$$R_{xx}(n) = \rho^{|n|}$$

$$\rightarrow \quad \Phi_{xx}(\omega) = \frac{1-\rho^2}{1+\rho^2-2\rho\cos\omega}$$

is a good model for the one-dimensional raster signal, with ρ being a function of the sampling resolution [30].

When no pre-filtering is employed, $F(\omega) = 1$, and (2.34) and (2.36) define the corresponding optimal post-filter. From (2.30), the error is given by

$$\xi_1 = 1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Phi_{xx}^2(\omega)}{\Phi_{yy}(\omega)} d\omega \ ,$$

where

$$\begin{aligned}\Phi_{yy}(\omega) &= \sum_{r=0}^{K-1} \Phi_{xx}(\omega - r\frac{2\pi}{K}) \\ &= \frac{K(1-\rho^{2K})}{1+\rho^{2K}-2\rho\cos(K\omega)} \ .\end{aligned}$$

When we substitute the expressions for $\Phi_{xx}(\omega)$ and $\Phi_{yy}(\omega)$ in the error equation above, we obtain

$$\xi_1 = 1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{(1-\rho^2)^2[1+\rho^{2K}-2\rho\cos(K\omega)]}{K(1-\rho^{2K})(1+\rho^2-2\rho\cos\omega)^2} d\omega \ .$$

The easiest approach towards deriving a closed-form expression for the error is by converting the integral above to the time domain

$$\xi_1 = 1 - \varsigma(0) \ ,$$

where $\varsigma(n)$ is given, according to the convolution theorem, by

$$\varsigma(n) = R_{xx}(n) * * R_{xx}(n) * * \mathcal{F}^{-1}\{\Phi_{yy}(\omega)\} \ ,$$

where $**$ denotes convolution and $\mathcal{F}^{-1}$ denotes the inverse Fourier transform operator. Since

$$\mathcal{F}^{-1}\{\Phi_{yy}(\omega)\} = \begin{cases} \dfrac{1+\rho^{2K}}{K(1-\rho^{2K})}, & n = 0 \ , \\ -\dfrac{\rho^K}{K(1-\rho^{2K})}, & n = \pm K \ , \\ 0, \quad \text{otherwise}, \end{cases}$$

the convolution that defines $\varsigma(n)$ can be easily evaluated directly in the time domain, with the final result

$$\xi_1 = \frac{1+\rho^{2K}}{1-\rho^{2K}} - \frac{1}{K}\left(\frac{1+\rho^2}{1-\rho^2}\right).$$

It is interesting to note that the corresponding optimal post-filter is given by

$$\begin{aligned} G_{\mathrm{OPT}}(\omega) &= \frac{\Phi_{xx}(\omega)}{\Phi_{yy}(\omega)} \\ &= \frac{1-\rho^2}{K(1-\rho^{2K})} \quad \frac{1+\rho^{2K}-2\rho\cos(K\omega)}{1+\rho^2-2\rho\cos\omega}, \end{aligned}$$

which has the inverse Fourier transform

$$g_{\mathrm{OPT}}(n) = \begin{cases} \dfrac{\rho^{|n|}-\rho^{2K-|n|}}{K(1-\rho^{2K})}, & -K \le n \le K, \\ 0, \quad \text{otherwise}. \end{cases}$$

So, the optimal interpolator is actually an FIR filter. As $\rho \to 1$, $g_{\mathrm{OPT}}(n)$ decreases linearly from the value $1/K$ at $n = 0$ to 0 at $n = K$.

The optimal pre- and post-filters have an ideal low-pass response with the cutoff frequency at π/K, and the error is entirely due to the missing spectral components for $|\omega| > \pi/K$, that is,

$$\begin{aligned} \xi_2 &= \frac{1}{\pi}\int_{\pi/K}^{\pi} \Phi_{xx}(\omega)\,d\omega \\ &= \frac{1}{\pi}\int_{\pi/K}^{\pi} \frac{1-\rho^2}{1+\rho^2-2\rho\cos\omega}\,d\omega \\ &= 1 - \frac{2}{\pi}\tan^{-1}\left[\left(\frac{1+\rho}{1-\rho}\right)\tan\left(\frac{\pi}{2K}\right)\right]. \end{aligned}$$

If $\rho \to 0$ the input spectrum is approximately white, and we have $\xi_1 = \xi_2 = 1 - K^{-1}$. As $\rho \to 1$, we can use Taylor series expansions for ξ_1 and ξ_2, around $\rho = 1$, to obtain

$$\lim_{\rho \to 1} \xi_1 = \frac{1-\rho}{3}\left(\frac{K^2-1}{K}\right)$$

and

$$\lim_{\rho \to 1} \xi_2 = \frac{1-\rho}{\pi}\left[\tan\left(\frac{\pi}{2K}\right)\right]^{-1} .$$

The error improvement due to pre-filtering, ξ_1/ξ_2, is a monotonically increasing function of both ρ and K. Therefore, the supremum of the improvement is approached as $\rho \to 1$ and $K \to \infty$, and is given by

$$\sup_{\substack{0<\rho<1 \\ K \geq 1}} \frac{\xi_1}{\xi_2} = \lim_{\substack{\rho \to 1 \\ K \to \infty}} \frac{\frac{1-\rho}{3}\left(\frac{K^2-1}{K}\right)}{\frac{1-\rho}{\pi}\left[\tan\left(\frac{\pi}{2K}\right)\right]^{-1}}$$

$$= \frac{\pi^2}{6} .$$

Thus, the error improvement with optimal pre- and post-filtering over optimal post-filtering only cannot be larger than $\pi^2/6$, which is approximately equal to 2.16 dB. This is somewhat lower than the previous 3 dB bound for general spectra [9].

2.4. Summary

We have derived in this chapter the ideal IIR solutions to the jointly-optimal pre- and post-filtering problem for continuous- and discrete-time systems, with or without sampling. Our results are an extension of previous work cited in the references, in the sense that we have taken into account the presence of both input and channel noises, the error criterion was a weighted mean-square measure, and the cross-correlation between channel signal and noise was taken into account for the case of quantization noise for Gaussian signals. The results already available in the literature can be obtained as special cases of our equations, by zeroing the appropriate noise source(s) and/or setting $W(\omega) \equiv 1$.

An adaptive system can be designed by including a spectral estimator in the transmitter. Such an estimator could, for example, match the incoming signal spectrum to a member of a set of 'typical' spectra, and an optimal pair of pre- and post-filters could be selected from a table of pairs optimized for each member of that set. Such an approach would make more sense if we use the FIR filters of Chapter 3 or the block filters of Chapter 4, since in practice we cannot realize the ideal IIR filters.

In Fig. 2.6 we have noticed that even when the input signal $x(t)$ or $x(n)$ is available with a high SNR, e.g., 30 dB or more, the performance of the optimal system may be significantly influenced by the characteristics of the input noise. Therefore, we believe that inclusion of both noise sources allows better modeling of a real system, mainly when the input signal spectrum is strongly peaked. We should point out that Fig. 2.6 represented a more realistic model of a noisy communications system than Fig. 2.3, which indicates that improvements on the order of many decibels may be expected in some circumstances, and only a few dB in others.

The theoretically attainable improvements of 20 dB or more in Fig. 2.3 will not occur in practice, in general, since they can only be attained at the expense of an increased bandwidth. Nevertheless, error reductions in the range of 2–10 dB are generally significant in most applications, so that joint optimization of the pre- and post-filters has a good potential for the enhancement of existing systems.

A simple application of the results of this chapter, in Example 2.3, has resulted in a new tight bound for the maximum pre-filtering error improvement for first-order Gauss-Markov processes in a noiseless system. The supremum of the improvement was found to be 2.16 dB, which is somewhat lower than the 3 dB bound for general signals.

We have seen that the optimal filters are generally band-limited, even if there is no sampling in the channel. Furthermore, their frequency responses depend on factors that are square roots of rational spectra. Hence, the jointly-optimal pre-

and post-filters cannot be represented, in general, by rational transfer functions, and thus they are not realizable by finite lumped networks (analog or digital). One approach towards designing realizable filters would be to consider only rational and causal IIR responses for the pre- and post-filters, as done by Schott [31]. The design variables would be the poles and residues of the pre-filter response. The resulting optimization problem is virtually intractable analytically, so that numerical techniques must be applied without much insight about convergence, distribution of local minima, and other issues.

In the next chapter we adopt a different approach to the direct design of realizable filters: we impose a finiteness constraint on the region of support of the filter impulse responses. With the availability of fast integrated circuits for FIR filtering, such an approach is of significant practical interest.

References

[1] J. P. Costas, "Coding with linear systems," *Proc. IRE,* vol. 40, pp. 1101–1103, Sept. 1952.

[2] L. M. Goodman and P. R. Drouilhet, Jr., "Asymptotically optimum pre-emphasis and de-emphasis networks for sampling and quantizing," *Proc. IEEE,* vol. 54, pp. 795–796, May 1966.

[3] B. G. Cramer, "Optimum linear filtering of analog signals in noisy channels," *IEEE Trans. Audio Electroacoust.,* vol. AU-14, pp. 3–15, Mar. 1966.

[4] H. L. Van Trees, *Detection, Estimation and Modulation Theory,* part I. New York: Wiley, 1968, chapter 6.

[5] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* New York: Mc.Graw–Hill, 1965, chapter 11.

[6] D. G. Luenberger, *Optimization by vector space methods.* New York: Wiley, 1969, chapter 4.

[7] G. A. Bliss, *Lectures on the Calculus of Variations.* Chicago: The University of Chicago Press, 1963, chapter VII.

[8] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory,* vol. IT-6, pp. 7–12, March 1960.

[9] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984, chapter 4.

[10] D. Chan and R. W. Donaldson, "Correlation functions and reconstruction error for quantized Gaussian signals transmitted over discrete memoryless channels," *IEEE Trans. Inform. Theory,* vol. IT-18, pp. 519–523, July 1972.

[11] W. Mauersberger, "An analytic function describing the error performance of optimum quantizers," *IEEE Trans. Inform. Theory,* vol. IT-27, pp. 519–521, July 1981.

[12] L. G. Roberts, "Picture coding using pseudo-random noise," *IEEE Trans. Inform. Theory,* vol. IT-8, pp. 145–154, Feb. 1962.

[13] R. M. Stewart, "Statistical design and evaluation of filters for restoration of sampled data," *Proc. IRE,* vol. 44, pp. 253–257, Feb. 1956.

[14] D. W. Tufts and N. Johnson, "Methods for recovering a random waveform from a finite number of samples," *IEEE Trans. Circuit Theory,* vol. CT-12, pp. 32–39, March 1965.

[15] J. J. Spilker, Jr., "Theoretical bounds on the performance of sampled data communications systems," *IEEE Trans. Circuit Theory,* vol. CT-7, pp. 335–341, Sept. 1960.

[16] W. M. Brown, "Optimum prefiltering of sampled data," *IEEE Trans. Inform. Theory,* vol. IT-7, pp. 269–270, Oct. 1961.

[17] T. Ericson, "Optimum PAM filters are always band limited," *IEEE Trans. Inform. Theory,* vol. IT-19, pp. 570–573, Oct. 1973.

[18] P. M. DeRusso, "Optimum linear filtering of signals prior to sampling," *AIEEE Trans. (Appl. Ind.),* vol. 79, pp. 549–555, Jan. 1961.

[19] S. S. L. Chang, "Optimum transmission of continuous signal over a sampled data link," *AIEEE Trans. (Appl. Ind.),* vol. 79, pp. 538–542, Jan. 1961.

[20] W. C. Kellogg, "Jointly optimum realizable linear pre- and postfilters for systems with samplers," *Proc. IEEE,* vol. 53, pp. 623–624, 1965.

[21] ————, *Numerical Operations on Random Functions.* Ph. D. dissertation, Div. of Engrg. and Appl. Physics, Harvard University, Cambridge, Mass., 1966.

[22] ————, "Information rates in sampling and quantization," *IEEE Trans. Inform. Theory,* vol. IT-13, pp. 506–511, July 1967.

[23] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[24] D. Chan and R. W. Donaldson, "Optimum pre- and postfiltering of sampled signals with applications to pulse modulation and data compression systems," *IEEE Trans. Commun. Technol.,* vol. COM-19, pp. 141–156, Apr. 1971.

[25] P. Noll, "On predictive quantizing schemes," *Bell Syst. Tech. J.,* vol. 57, pp. 1499–1532, May-June 1978.

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

[27] L. W. Couch, III, *Digital and Analog Communication Systems.* New York: Macmillan, 1983, chapter 7.

[28] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1975, chapter 1.

[29] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1983, chapter 3.

[30] R. J. Clarke, *Transform Coding of Images.* London: Academic Press, 1985, chapter 2.

[31] W. Schott, *Joint optimization of causal transmitter and receiver in pulse amplitude modulation.* Proc. of First European Signal Processing Conference (EUSIPCO-80), Lausanne, Switzerland, pp. 545–549, Sept. 1980.

Chapter 3

Optimal FIR Filters

This chapter is focused on the discrete-time pre- and post-filtering system, with the added assumption that the pre- and post-filters have a finite impulse response (FIR). The system model is that of Fig. 3.1. One approach towards the design of an 'almost optimal' system with FIR filters would be a two-step procedure: first, we would use the results of the previous chapter to derive the ideal pair of filters, and then we could design FIR filters that approximate the ideal ones as closely as possible. This is, in fact, the standard approach in digital filter design, and several FIR filter design techniques for approximating ideal filters have been developed [1]. The ideal low-pass filter with a constant passband gain, which is the optimal one for pre- and post-filtering on a noiseless system, is the standard prototype for FIR filter design.

Approximating the ideal pre- and post-filters by FIR responses has one strong disadvantage: for any particular design technique, e.g., windowing, it is difficult to predict *a priori* the effect of the approximation on the weighted mean-square reconstruction error, except by computing that error explicitly *after* the realizable filters are designed. Under the assumption that minimization of the weighted mean-square error is the goal, a better approach is to reformulate the joint optimization problem in order to absorb the finiteness of the filter impulse responses.

The frequency-domain error analysis that we have developed previously cannot be brought into this chapter without changes, since the frequency responses of FIR filters are limited to a subspace of of all frequency responses. The modifications necessary to accommodate FIR filters are discussed in the next section.

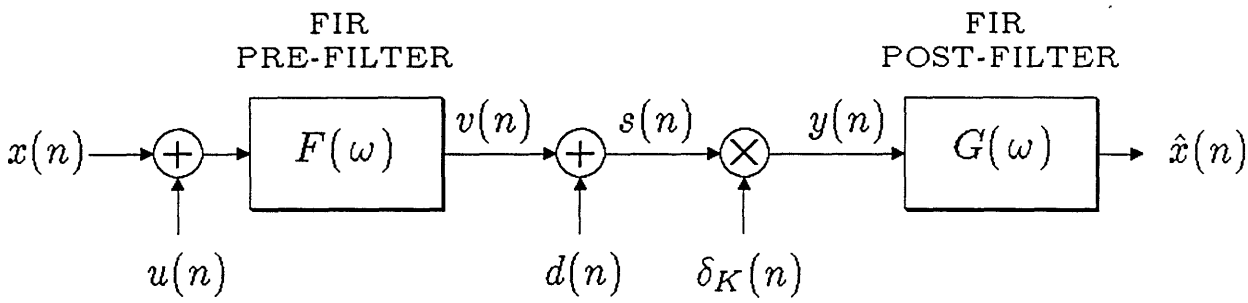


Fig. 3.1. Discrete-time system model with FIR filters.

In Section 3.2 we consider the independent optimization of the pre- or post-filter, whereas Section 3.3 presents an algorithm for their joint optimization. In Section 3.4 the performance of optimal FIR filter pairs is compared to that of optimal IIR filters and some commonly-used FIR filters designed under different criteria.

The use of FIR filters in multidimensional signal processing applications is common practice nowadays, due to the ever increasing availability of fast and inexpensive integrated circuits for data processing and storage. Therefore, the design of multidimensional filters has received significant attention in recent years. We will extend our results on optimal FIR pre- and post-filter design to the multidimensional case in Section 3.5.

3.1. Analysis of Systems with FIR Filters

The spectral representations of the previous chapter are independent of the finiteness of the pre- and post-filter impulse responses. Thus, the error measure is defined

in the same way as in Chapter 2, namely,

$$\begin{aligned}\xi &= \frac{1}{K}\sum_{n=0}^{K-1} \mathrm{E}[e^2(n)] \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{ee}(\omega)\, d\omega \ .\end{aligned} \tag{3.1}$$

with $\Phi_{ee}(\omega)$ given by

$$\begin{aligned}\Phi_{ee}(\omega) =& \mid W(\omega) \mid^2 [\Phi_{xx}(\omega) + \mid G(\omega) \mid^2 \Phi_{yy}(\omega)] \\ & - 2 \mid W(\omega) \mid^2 \mathrm{Re}\{G(\omega)F(\omega)\}[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)] \ ,\end{aligned} \tag{3.2}$$

where

$$\begin{aligned}\Phi_{yy}(\omega) &= \tilde{\Phi}_{dd}(\omega) \\ &+ \sum_{r=0}^{K-1} F^2(\omega + r\omega_K)[(1+2\beta)\Phi_{xx}(\omega + r\omega_K) + \Phi_{uu}(\omega + r\omega_K)] \ ,\end{aligned} \tag{3.3}$$

with $\omega_K \triangleq 2\pi/K$, and

$$\tilde{\Phi}_{dd}(\omega) \triangleq \sum_{r=0}^{K-1} \Phi_{dd}(\omega + r\omega_K) \ . \tag{3.4}$$

The parameter β has the same meaning as in the previous chapter, i.e., it is either zero for an uncorrelated channel noise or a negative number when the channel noise is due to quantization. For a non-zero β we assume that the input signal and noise are Gaussian.

For any given magnitude responses for the filters $F(\omega)$ and $G(\omega)$, the term $\mathrm{Re}\{G(\omega)F(\omega)\}$ in (3.2) is maximized when the phases of $F(\omega)$ and $G(\omega)$ are both equal to zero, for all ω. Therefore, we shall concentrate on zero phase filters.

Specifically, we impose the following constraints on their impulse responses:

$$\begin{aligned} f(-n) &= f(n) \ , \\ g(-n) &= g(n) \ , \\ f(n) &= 0, \quad \text{if } \mid n \mid > L \ , \text{ and} \\ g(n) &= 0, \quad \text{if } \mid n \mid > M \ . \end{aligned} \tag{3.5}$$

Under these assumptions we can rewrite the error spectrum as

$$\begin{aligned} \Phi_{ee}(\omega) = & \mid W(\omega) \mid^2 [\Phi_{xx}(\omega) + G^2(\omega)\Phi_{yy}(\omega)] \\ & - 2 \mid W(\omega) \mid^2 G(\omega)F(\omega)[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)] \ . \end{aligned} \tag{3.6}$$

Our objective in this chapter is the minimization of (3.6) under the constraints in (3.5). We cannot work directly with $F(\omega)$ and $G(\omega)$, since we do not have enough degrees of freedom to arbitrarily set their values for all frequencies. One approach towards incorporating the FIR constraints into (3.6) is to convert to the time domain all terms in which $F(\omega)$ and $G(\omega)$ appear; this leads to

$$\begin{aligned} \xi = & \frac{1}{2\pi} \int_{-\pi}^{\pi} \mid W(\omega) \mid^2 \Phi_{xx}(\omega) \ d\omega \\ & - 2 \sum_{l=-L}^{L} \sum_{m=-M}^{M} f(l)g(m)a(l-m) + \sum_{l=-M}^{M} \sum_{m=-M}^{M} g(l)g(m)b(l-m) \\ & + \sum_{l=-L}^{L} \sum_{m=-L}^{L} f(l)f(m) \sum_{r=-M}^{M} \sum_{s=-M}^{M} g(r)g(s) \\ & \times \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} w(u)w(v)\delta_K(r-s+u-v)c(l-m+r-s+u-v) \ , \end{aligned} \tag{3.7}$$

where $w(n)$ is the observer impulse response and the sequences $a(n)$, $b(n)$, and $c(n)$

have the Fourier transforms

$$\begin{aligned} A(\omega) &= \mid W(\omega) \mid^2 [(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)] \ , \\ B(\omega) &= \mid W(\omega) \mid^2 \tilde{\Phi}_{dd}(\omega) \ , \text{ and} \\ C(\omega) &= (1+2\beta)\Phi_{xx}(\omega) + \Phi_{uu}(\omega) \ . \end{aligned} \tag{3.8}$$

Our optimization problem could be formulated as the minimization of (3.7) as a function of the vector $[f(0)\ f(1)\ \ldots\ f(L)\ g(0)\ g(1)\ \ldots\ g(M)]$, but it would be virtually impossible to analyze such issues as convexity and convergence since (3.7) is a quartic form. However, if we fix the pre(post)-filter coefficients, then the error is a quadratic form on the post(pre)-filter coefficients, which is easier to minimize. Thus, we shall follow the following route towards derivation of a jointly-optimal filter pair: obtaining first independent solutions for the pre- and post-filters, and then combining them in an iterative procedure that computes the jointly-optimal pair.

In the previous chapter, the solution for the optimal post-filter was used to derive an error expression as a function of the pre-filter only. The latter expression was then minimized in order to obtain the jointly-optimal filter pair. Unfortunately, we cannot follow the same approach here; the difficulty is that the error obtained with an optimal FIR post-filter cannot be written as a function of the FIR pre-filter coefficients, as we will see in the next section. Therefore, closed-form solutions for a jointly-optimal filter pair cannot be obtained, except for trivial cases, e.g. $L = M = 1$, which will not be specifically considered.

3.2. Optimization of a Single Filter

We focus our attention now on the derivation of both the optimal post-filter for a given pre-filter, and the optimal pre-filter for a given post-filter. In the next section

we will combine the two results in an algorithm that computes the jointly-optimal pair.

The design of the post-filter (or interpolator) has received much more attention in the literature than the pre-filter design. Oetken *et. al.* [2] derived the optimal interpolator without a pre-filter, for band-limited input signals and noiseless samples. Polydoros and Protonotarios [3] assumed a statistical description of the input signal, as in our work, and derived the optimal interpolator without a pre-filter. As in [2], they have considered a noiseless system, but with the added restriction of zero intersymbol interference. Keys [4] used cubic convolution kernels, derived from cubic splines, to determine the impulse response of the interpolator; his main concern was the alleviation of sampling artifacts in image processing.

Interpolation of a stochastic signal from noisy samples with an FIR filter has been considered by Kay [5] and more recently by Radbel and Marks [6]. The solution in [6] applies to the system in Fig. 3.1 for the case $F(\omega) \equiv 1$ and $u(n) \equiv 0$. Our results for the optimal interpolator in this chapter are essentially a generalization of [6] for any pre-filter and input noise spectrum.

The design of optimal FIR pre-filters has received little attention in the literature. Chevillat and Ungerboeck [7] derived optimal pre- and post-filters for a discrete-time input signal and a continuous-time band-limited channel. Their results apply directly to modem design, for example, but they cannot be used in our case, since we have a discrete-time channel. Hummel [8] has considered the problem of designing an optimal pre-filter when the interpolator is a spline function, and the system is noiseless. He showed that the optimal pre-filter in that case is also a spline function. Ratzel [9] has derived optimal Gaussian pre-filters for digitized images, based on subjective experiments. Recently, Faubert [10] has determined the jointly-optimal pre-and post-filters for a noiseless system. Our work in this chapter can be viewed as a one-dimensional extension of Faubert's results for the noisy system in Fig. 3.1.

3.2.1. The Optimal Post-filter

For a fixed pre-filter, we can rewrite (3.7) explicitly as a function of the post-filters coefficients, in the form

$$\begin{aligned}\xi = & \frac{1}{2\pi}\int_{-\pi}^{\pi} |\, W(\omega)\,|^2\, \Phi_{xx}(\omega)\, d\omega \\ & + \sum_{l=-M}^{M}\sum_{m=-M}^{M} g(l)g(m)\psi(l-m) - 2\sum_{l=-M}^{M} g(l)\theta(l)\ ,\end{aligned} \tag{3.9}$$

where $\psi(n)$ and $\theta(n)$ are the inverse Fourier transforms of $\Psi(\omega)$ and $\Theta(\omega)$, respectively, which are defined by

$$\Psi(\omega) \triangleq |\, W(\omega)\,|^2\, \Phi_{yy}(\omega)\ , \tag{3.10}$$

and

$$\Theta(\omega) \triangleq |\, W(\omega)\,|^2\, F(\omega)[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)]\ . \tag{3.11}$$

The first-order necessary condition for $g(n)$ to be an optimal post-filter is that $\partial\xi/\partial g(l) = 0,\ \forall\, l$, which leads to the system of linear equations

$$\begin{aligned}\sum_{m=-M}^{M} g(m)\psi(l-m) &= \theta(l) \\ l &= -M, -M+1, \ldots, M\ .\end{aligned} \tag{3.12}$$

Since $\Psi(\omega)$ is a valid power spectrum, the matrix whose entries are $\psi(l-m)$, for $l, m = -M, \ldots, M$ is at least positive semidefinite [15]. With the mild assumption that $\Psi(\omega) > 0\ \forall\ \omega$, the matrix is positive definite, and the error is then a strictly convex function of the post-filter coefficients. Thus, the unique solution to (3.12) globally minimizes the error, for a fixed pre-filter.

The equations in (3.12) have a Toeplitz structure, and so they can be solved in $O[(2M+1)^2]$ operations by means of Levinson's recursion, which is well explained in [11]. If M is very large, there are algorithms that have $O(2M+1)[\log(2M+1)]^2$ complexity, [12],[13]. These algorithms are considerably more difficult to implement than Levinson's recursion. It is interesting to note that the symmetry constraint imposed on the pre-filter forces $\Theta(\omega)$ to be a real function, so that $\theta(n)$ is a symmetric sequence. Therefore, the solution to (3.12) necessarily leads to a symmetric sequence $g(n)$ that satisfies (3.5). We could exploit this symmetry to convert (3.12) to a Toeplitz-plus-Hankel system of only $M+1$ equations, which could also be efficiently solved, as discussed in [14].

With the optimal post-filter, (3.9) can be simplified to

$$\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\, W(\omega)\,|^2\, \Phi_{xx}(\omega)\; d\omega - \sum_{l=-M}^{M} g(l)\theta(l)\;. \tag{3.13}$$

It is not possible, however, to write (3.13) in terms of the pre-filter coefficients, since Toeplitz forms are not, in general, analytically invertible [15].

3.2.2. The Optimal Pre-filter

Now we assume that the post-filter is fixed. Then, the error expression in (3.7) can be simplified to

$$\begin{aligned}\xi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\, W(\omega)\,|^2\, [\Phi_{xx}(\omega) + \tilde{\Phi}_{dd}(\omega)G^2(\omega)]\; d\omega \\ + \sum_{l=-L}^{L}\sum_{m=-L}^{L} f(l)f(m)\gamma(l-m) - 2\sum_{l=-L}^{L} f(l)\vartheta(l)\;,\end{aligned} \tag{3.14}$$

where $\gamma(n)$ and $\vartheta(n)$ are the inverse Fourier transforms of $\Gamma(\omega)$ and $\Upsilon(\omega)$, respectively, which are defined by

$$\Gamma(\omega) \triangleq [(1+2\beta)\Phi_{xx}(\omega) + \Phi_{uu}(\omega)] \sum_{r=0}^{K-1} \mid W(\omega - r\omega_K) \mid^2 G^2(\omega - r\omega_K) \ , \qquad (3.15)$$

and

$$\Upsilon(\omega) \triangleq \mid W(\omega) \mid^2 G(\omega)[(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)] \ . \qquad (3.16)$$

At this point, in view of our final objective of jointly optimizing the pre- and post-filter, we introduce a power constraint on the pre-filter output $v(n)$. The necessity of such constraint is clear from (3.7); if we multiply all $f(n)$ by a constant α and divide all $g(n)$ by α, with $\mid \alpha \mid > 1$, the error is reduced, since the matrix formed by the elements $b(l-m)$ is at least positive semidefinite. However, since we are interested only in the optimal pre-filter in this section, we have to consider the possibility that an unconstrained optimal pre-filter may lead to a signal power at the channel input that is less than unity.

Thus, the power constraint must be an inequality, i.e.,

$$P \triangleq \mathrm{E}[v^2(n)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mid F(\omega) \mid^2 \left[\Phi_{xx}(\omega) + \Phi_{uu}(\omega)\right] d\omega \ \leq \ 1 \ , \qquad (3.17)$$

which can also be written in the time domain as

$$P = \sum_{l=-L}^{L} \sum_{m=-L}^{L} f(l)f(m)[R_{xx}(l-m) + R_{uu}(l-m)] \ \leq \ 1 \ . \qquad (3.18)$$

An optimal pre-filter has to be a stationary point of the Lagrangian [16] corresponding to the objective function (3.14) and the constraint (3.18), i.e., there must exist a Lagrange multiplier λ such that

$$\frac{\partial \xi}{\partial f(l)} + \lambda \frac{\partial P}{\partial f(l)} = 0, \ \forall \, l \ . \qquad (3.19)$$

The Lagrange multiplier also has the properties

$$\begin{aligned} \lambda &\geq 0 \\ \lambda(P-1) &= 0 \,, \end{aligned} \tag{3.20}$$

that is, if the power constraint is not satisfied by equality the value of the Lagrange multiplier is zero, since the constraint is not binding. The Lagrange multiplier is non-negative, since the inequality is $P \leq 1$. A proof of (3.20) for the general non-linear optimization problem can be found in [16].

From (3.19) we obtain

$$\begin{aligned} \sum_{m=-L}^{L} f(m)\{\gamma(l-m)+\lambda[R_{xx}(l-m)+R_{uu}(l-m)]\} &= \vartheta(l) \\ l &= -L,-L+1,\ldots,L \,. \end{aligned} \tag{3.21}$$

We have again a symmetric Toeplitz system of linear equations to be solved. So, our discussion about fast algorithms in the previous sub-section applies here, too. We note also that $\Gamma(\omega)$ is non-negative for all ω, which means that $\gamma(n)$ is a valid autocorrelation function, and so the matrices formed by the elements $\gamma(l-m)$ and $\gamma(l-m)+\lambda[R_{xx}(l-m)+R_{uu}(l-m)]$, $l,m=-L,\ldots,L$, are at least positive semidefinite. Thus (3.14) is a convex function of the pre-filter coefficients, and a solution to (3.21) is a global minimum. If we use Sugiyama's algorithm [17] to solve (3.21), we can obtain a solution for the pre-filter coefficients even if the matrix $\gamma(l-m)+\lambda[R_{xx}(l-m)+R_{uu}(l-m)]$ is singular.

There is still a problem in solving (3.21), which is the fact that we don't know *a priori* the value of the Lagrange multiplier λ. There is a simple approach, however, that we can apply to the solution of (3.21): first, we set $\lambda = 0$ and solve (3.21); if the solution satisfies $P < 1$, we're done; otherwise, the power constraint must be active, and we repeatedly solve (3.21) with λ updated by some technique for zeros of one-dimensional functions, e.g. Newton-Raphson's method [18], until we obtain a solution for which $P = 1$.

3.3. Jointly-optimal Solution

In the previous section we derived the optimal post-filter for any given pre-filter, and *vice versa*. The availability of those solutions suggests using them alternately, until they converge to an optimal pair. Formally, this corresponds to the following

Algorithm

Step 1 – Set $i \leftarrow 0$, and $f_0(n) \leftarrow \alpha\delta(n)$, with α chosen so that (3.18) is satisfied.

Step 2 – Use (3.10)-(3.12) with $f(n) = f_i(n)$ and solve for the optimal post-filter, $g(n)$. Set $g_i(n) = g(n)$.

Step 3 – Set $\lambda = 0$ and use (3.15)-(3.21) to compute an optimal pre-filter $f(n)$. Evaluate (3.18). If $P < 1$, go to Step 5, otherwise go to the next step.

Step 4 – Set λ to some positive value, solve (3.21), and update λ by means of some technique for zeros of functions, e.g. Newton-Raphson's method [18]. Repeat the process until $P \simeq 1$

Step 5 – Compute Δ by

$$\Delta \triangleq \max_{-L \le l \le L} | f_i(n) - f_{i-1}(n) | .$$

If Δ is sufficiently small, stop: the optimal pre- and post-filter are $f_i(n)$ and $g_i(n)$, respectively. Otherwise, set $i \leftarrow i+1$ and go back to Step 2. Alternatively, we could monitor the error level and stop whenever the error reduction from Step 4 is small enough.

The above algorithm is in the class of 'coordinate descent' algorithms for minimization of functions of several variables [20],[19], since at each step it finds

the unique global minimum of the error, with either the pre- or the post-filter coefficients kept fixed. Therefore, the algorithm necessarily converges to a stationary point of the Lagrangian [19], with a monotonic decrease in the error at each step. Unfortunately, there is no guarantee that the attained stationary point will be a global minimum; it could be a local minimum or a saddle point.

Our practical experience with the above algorithm has pointed out that stationary points tend to be well separated from each other, with large differences in their corresponding values of the error. This suggests a relatively simple way to check for the 'likelihood of global optimality' of a computed solution: we use the results in the previous chapter to compute the minimum error obtained with IIR filters; if the above algorithm leads to an error level close to the lower bound determined by the IIR solution, and the shapes of the impulse responses of the FIR pre- and post-filters resemble those of their IIR counterparts, we accept the FIR filters, otherwise we try a different starting point.

With the initial guess for the pre-filter suggested in Step 1, we have never failed to obtain a correct solution for the optimal FIR filters, with several different signal and noise spectra, but we did experience non-convergence problems if the observer frequency response $W(\omega)$ got too close to zero on some frequency range, since this leads to ill-conditioned matrices in (3.12) and (3.21).

The algorithm described above has a rate of convergence typical of coordinate descent methods, i.e., a weakly linear convergence [20] that is somewhat slower than that of the steepest descent algorithm [19]. Faster convergence, in terms of number of iterations, could be obtained by using the steepest descent or Newton's methods. In either of these two alternative approaches, however, additional information would have to be computed, namely the gradient of the error for the steepest descent method, and both the gradient and the Hessian for Newton's. For example, the number of operations required by the coordinate descent approach with $L = M = 8$ is approximately 6,000 per iteration, whereas Newton's method requires about 200,000 operations per iteration (assuming, in both cases, that convolutions are

performed by means of FFT's). Typically, the coordinate descent algorithm would have converged before a single iteration of Newton's method could be performed. Another advantage of the coordinate descent method besides its simplicity is that at any iteration we have at the end of Step 5 a 'partially-optimal' solution, in the sense that at least the pre-filter is optimal for the current post-filter, which is in turn optimal for the previous post-filter.

We end this section by deriving an expression for the value of the Lagrange multiplier λ at a jointly-optimal solution; knowledge of this value can accelerate λ's convergence in Step 4. Using (3.12) we can write the error as

$$\xi_g = \xi_o - \sum_{l=-L}^{L} f(l) \sum_{m=-M}^{M} g(m)\varsigma(l-m) \ , \tag{3.22}$$

where

$$\xi_o \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} | W(\omega) |^2 \ \Phi_{xx}(\omega) \ d\omega \ , \tag{3.23}$$

and $\varsigma(n)$ is the inverse Fourier transform of

$$| W(\omega) |^2 \ [(1+\beta)\Phi_{xx}(\omega) + \beta\Phi_{uu}(\omega)] \ .$$

From (3.21) we obtain

$$\xi_f = \xi_g + \frac{1}{2\pi} \int_{-\pi}^{\pi} | W(\omega) |^2 \ G^2(\omega)\tilde{\Phi}_{dd}(\omega) \ d\omega \ - \lambda P \ , \tag{3.24}$$

where P is the pre-filter output power. Since a jointly optimal pair satisfies both (3.12) and (3.21), we must have $\xi_g = \xi_f$, and so

$$\lambda_{\mathrm{OPT}} = \frac{1}{2\pi P} \int_{-\pi}^{\pi} | W(\omega) |^2 \ G^2(\omega)\tilde{\Phi}_{dd}(\omega) \ d\omega \ . \tag{3.25}$$

Thus, the optimal value of the Lagrange multiplier has a noise-to-signal ratio interpretation: it is the ratio of the filtered channel noise at the interpolator output (weighted by the observer response) to the available pre-filter power.

3.4. Performance of Optimal FIR Filters

In this section we are concerned about a few basic issues related to optimal FIR filters: i) how sensitive are their responses to the observer weighting function, ii) what is the error improvement due to the use of jointly-optimal pre- and post-filters, as compared with an optimal interpolator only, iii) how large should we make L and M in order to get an error level close to that obtained with the ideal IIR filters, and iv) how much error reduction is to be expected by using optimal FIR pre- and post-filters instead of a filter pair designed under a different criterion. Each of the following sub-sections addresses one of the topics above.

3.4.1. Sensitivity to the Observer Response

In order to evaluate the effect of the observer $W(\omega)$ on the optimal filters, let's consider the peaked low-pass response in Fig. 3.2, which is an approximation to a Gaussian bandpass response, with $w(n)$ limited to the range $-6 \leq n \leq 6$. An FIR observer with a short duration allows the use of Discrete Fourier Transforms of short length for the computation of convolutions.

The responses of the optimal pre- and post-filters without and with the observer weighting are shown in Fig. 3.3 and Fig. 3.4, respectively. The design parameters were the following: a first-order Gauss-Markov input spectrum with unit energy and an inter-sample correlation coefficient $\rho = 0.95$, $L = M = 8$, a down-sampling factor $K = 4$, white input and channel noise with energies of

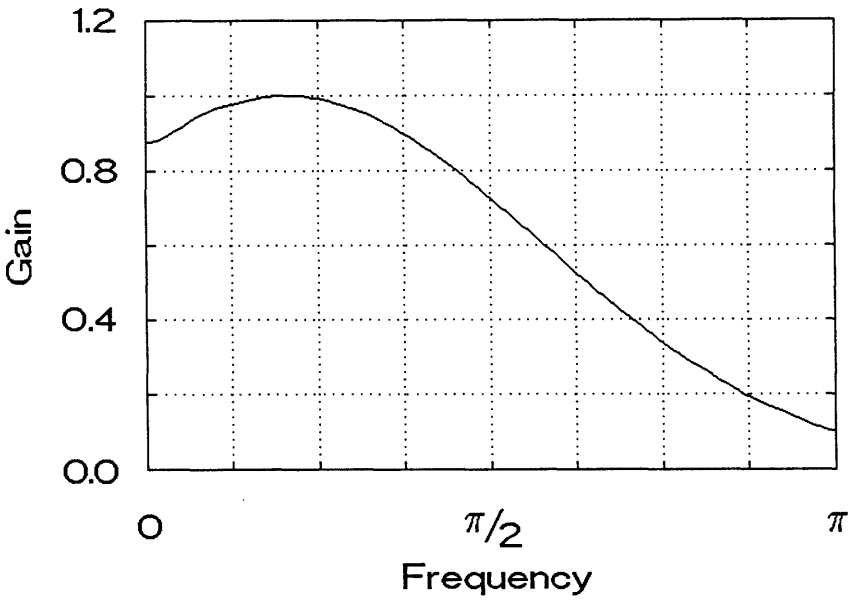
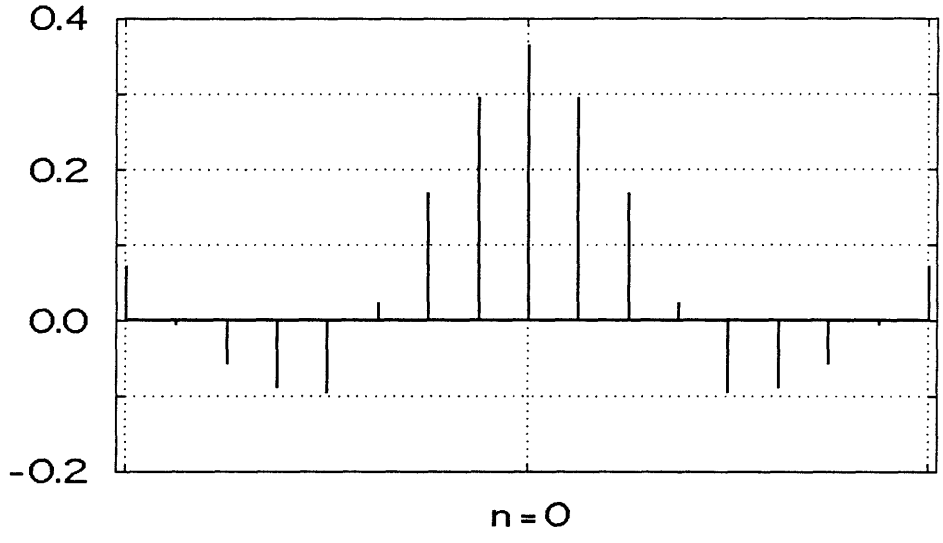


Fig. 3.2. A low-pass observer $W(\omega)$.

4×10^{-5} (i.e., $-44\ dB$), and $\beta = 0$. These noise levels correspond approximately to those produced by 8-bit quantization of Gaussian random variables. Quantization at 8 bits is common practice in digital image processing, since a finer gray scale resolution cannot be detected by the human eye. Although β should not be zero, strictly speaking, since we assume that the noise is due to quantization, the approximation $\beta = 0$ leads to a negligible error for an SNR of 44 dB, and allows the simplification of most of the equations.

We note that, since our sampling function $\delta_K(n)$ has a gain of K for each sample, according to its previous definition, the interpolator will have a d.c. gain close to unity. Therefore, the pre-filter should also have a d.c. gain close to unity, since the optimal filters will certainly keep the d.c. reconstruction error at a low level.



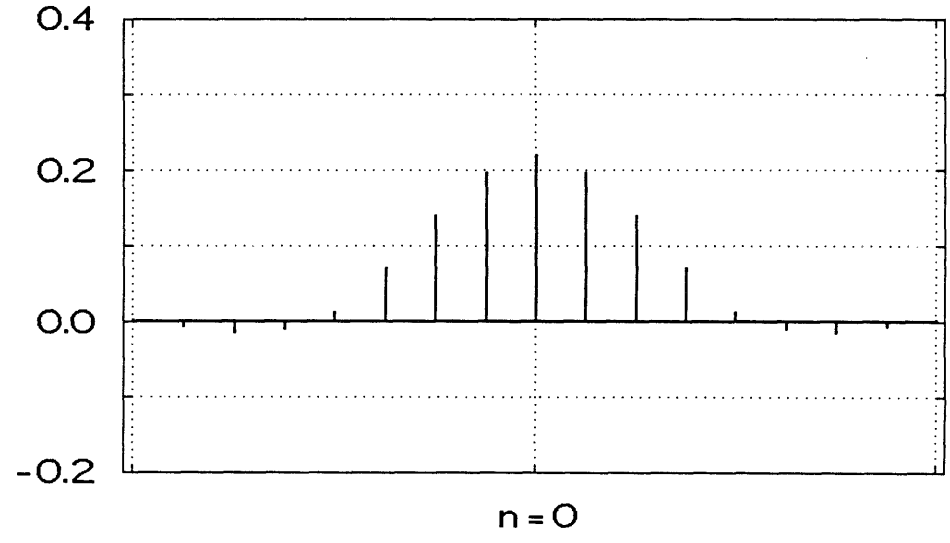
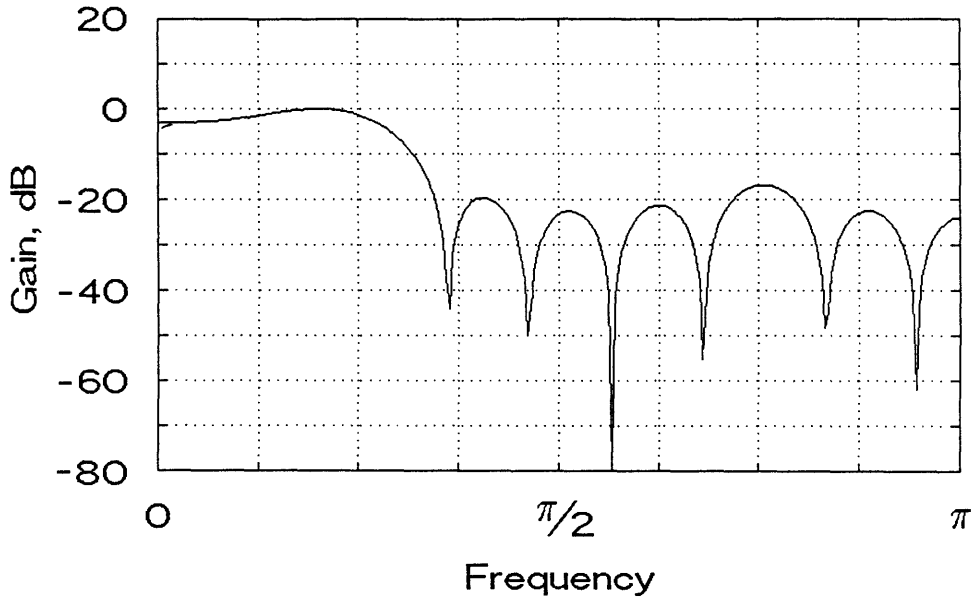
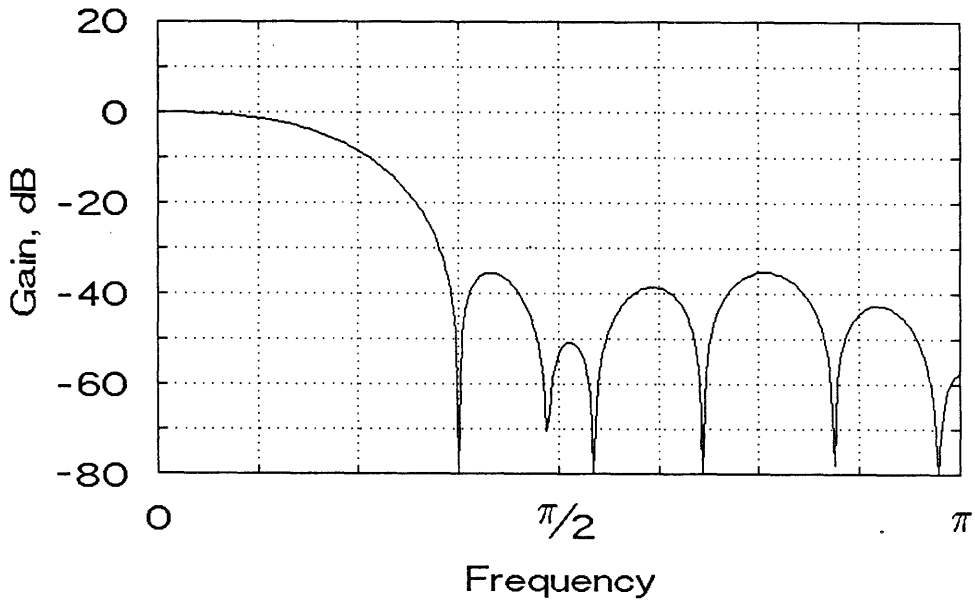


Fig. 3.3. Impulse responses of optimal filters without error frequency-weighting ($W(\omega) \equiv 1$): (a) pre; and (b) post.

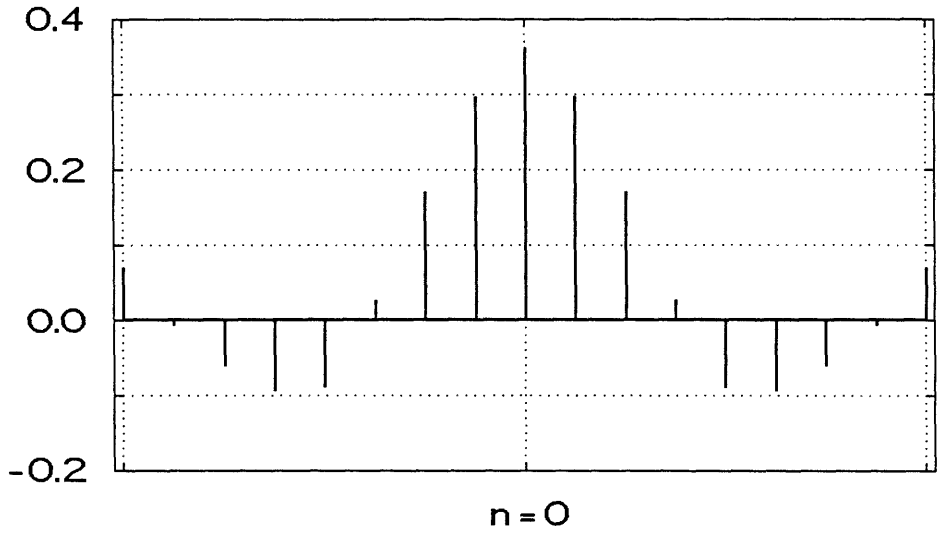


(c)

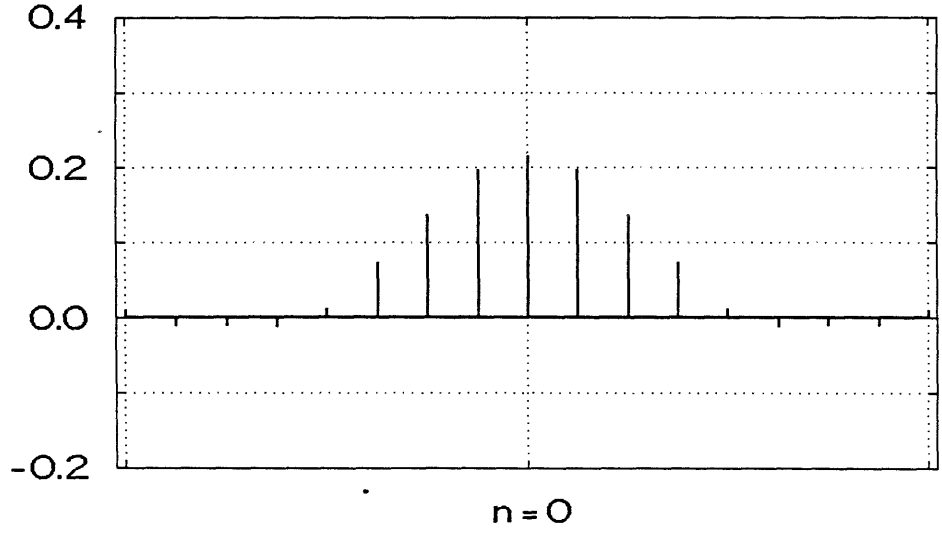


(d)

Fig. 3.3. Continued, frequency responses: (c) pre; and (d) post.

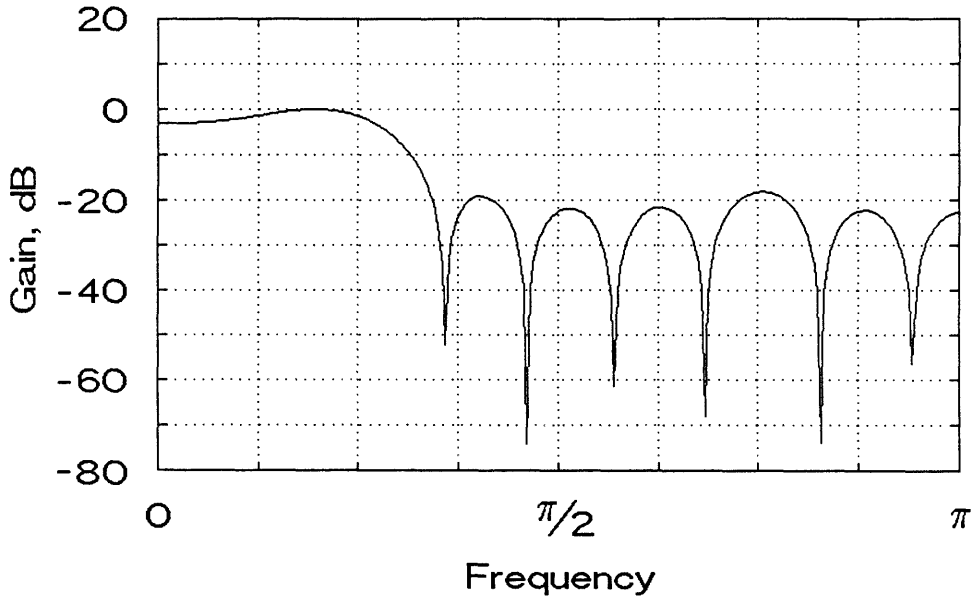


(a)

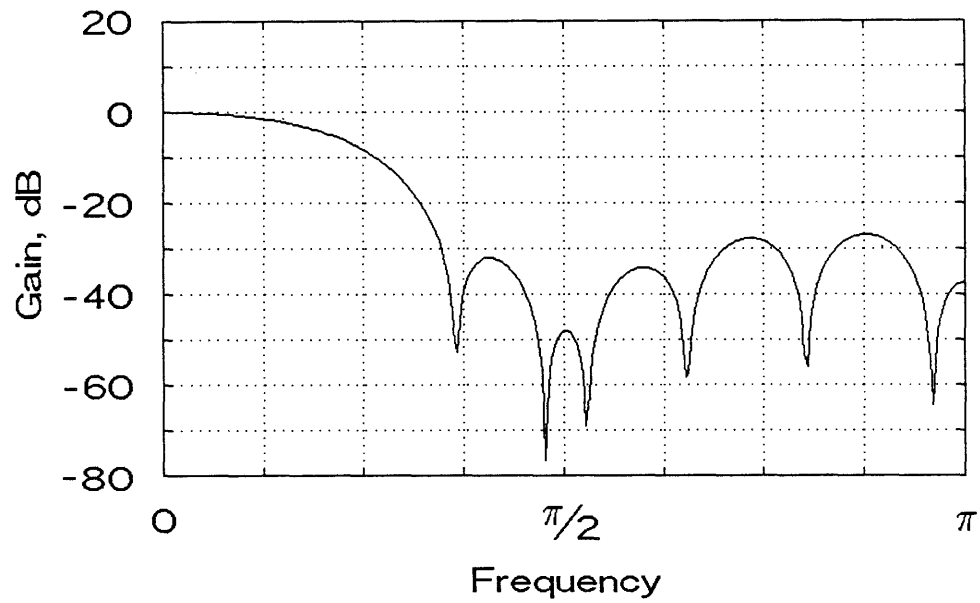


(b)

Fig. 3.4. Impulse responses of optimal filters for the low-pass observer of Fig. 3.2: (a) pre; and (b) post.



(c)



(d)

Fig. 3.4. Continued, frequency responses: (c) pre; and (d) post.

In Fig. 3.3 (a), (b) the impulse responses of the optimal pre- and post-filter for $W(\omega) \equiv 1$ are depicted, respectively, with the corresponding frequency responses shown in Fig. 3.3 (c) and (d). The magnitude responses start decaying after $\omega = \pi/4$, which is the cutoff frequency of the ideal IIR pre- and post-filters. The transition bandwidth from passband to stopband is approximately equal to $\pi/8$; it can be reduced by using longer impulse responses.

The pre-filter impulse response has stronger side-lobes than the post-filter; in the frequency domain that corresponds to a slight high-frequency boost on the pre-filter response. That was expected from our previous discussion of the half-whitening effect. We note, however, that the high-frequency boost in the pre-filter is actually much weaker than what an IIR half-whitening responses should have. This is due to the fact that the optimal pre-filter should also be band-limited, and so the FIR response achieves the best compromise between ideal half-whitening and ideal band-limiting. The minimum stopband gains of the pre- and post-filters are markedly different, -17 dB and -34 dB, respectively.

An optimal pair of filters for the band-pass observer is shown in Fig. 3.4. The main effect of the observer is to increase the maximum stopband gain of the post-filter from -35 dB to -28 dB. The pre-filter response is virtually unaltered. Thus, the influence of the observer responses on the optimal FIR filters is much weaker than what it is on the optimal IIR filters of Chapter 2, where the pre-filter frequency response is proportional to $\sqrt{W(\omega)}$.

3.4.2. Error Improvement with Optimal Pre-filtering

We consider now the error improvement with optimal pre- and post-filtering over optimal post-filtering only (i.e., no pre-filter). In Fig. 3.5 the error reduction is plotted as a function of the correlation coefficient. We kept the same system parameters as described in the previous sub-section, but we varied the input correlation

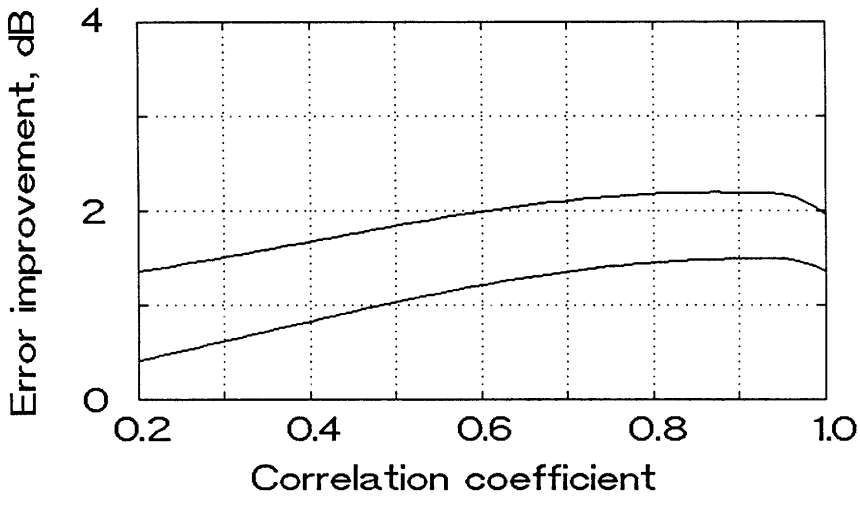


Fig. 3.5. Error improvement with optimal pre-filtering. Top curve: low-pass observer; bottom: flat observer.

coefficient ρ from 0.2 to 1.0. With the low-pass observer the loss of the signal high-frequency components has a lower weight than the in-band aliasing errors. Therefore, the error improvement with optimal pre-filtering is larger.

3.4.3. Choosing the Lengths of the Impulse Responses

As we discussed at the beginning of this section, it would be interesting to evaluate the error for several values of the parameters L and M, which determine the length of the filter responses. This is done in Fig. 3.6, for $W(\omega) = 1$, input and channel SNR's of 30 dB, and $\rho = 0.95$. In Fig. 3.6 (a) M is set to 12, and L varied from 0 to 12. It is interesting that the error reduces up to $L = 6$, when it assumes a minimum value of 0.065, which cannot be reduced further by an increase in L. We

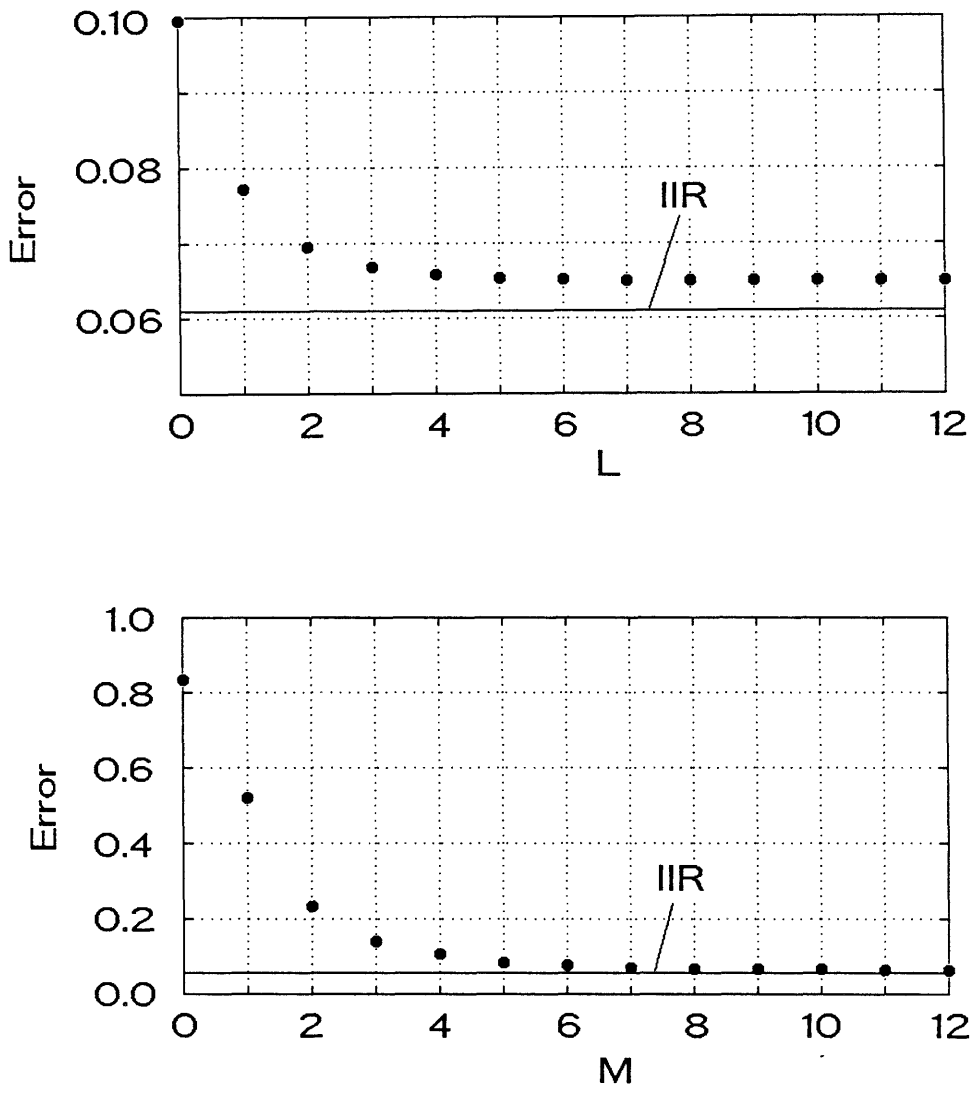


Fig. 3.6. Error as a function of the lengths of the optimal filters: (a) varying L for $M = 12$; (b) varying M for $L = 6$.

note that the minimum error with ideal IIR filters is 0.061. In Fig. 3.6 (b) L is set to 6, and M varied from 0 to 12. For $M < K$ the reconstructed signal $\hat{x}(n)$ has zero samples in it, and so the error is large. In both figures we see that with L and M on the order of $2K$ it is possible to obtain an error level close to that of the ideal IIR filter pair.

The above result was unexpected to some extent, given our experience with

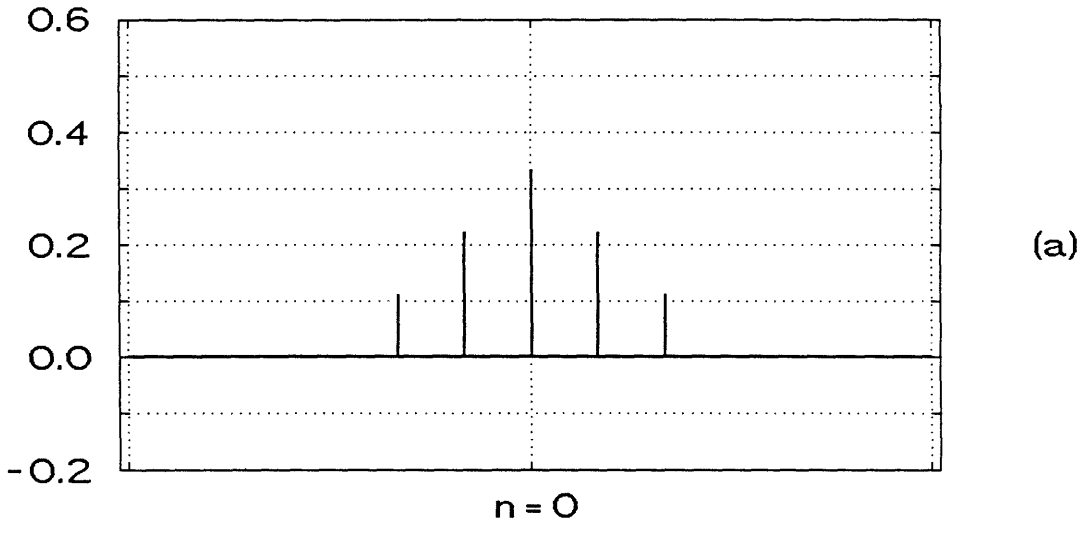
other filter design techniques. For example, when the approximation criterion is minimizing some error measure between the FIR response and that of the ideal IIR filter, as in the case of the Parks-McClellan method [21], the error can always be significantly reduced by a large increase in the filter length. In terms of the mean-square signal reconstruction error ξ, however, the performance is bounded by that of the ideal IIR filter, and so it is natural to expect a reduced sensitivity of the error to the filter lengths as they increase, but it is somewhat surprising that even for very short lengths we can get relatively low error levels.

3.4.4. Comparison with Other Filters

There are several approaches to the design of FIR filters for decimation and interpolation [23], of which the majority are based on approximating the ideal low-pass responses of the optimal IIR filters. A natural question that arises at this point is how the optimal FIR pre- and post-filters that we have derived compare to those other types of filters in practice. Some comparisons that may help answering that question are described below.

We have carried out some decimation and interpolation experiments with the filters depicted in Fig. 3.7 and Fig. 3.8, taking as pre- and post-filter pairs (a)-(b), (c)-(d), (e)-(f), and (g)-(h). The parameters in common for the various filters were $K = 3$ and $L = M = 6$, with the exception of the linear filter of Fig. 3.7 (a), which by definition has a nonzero impulse response only for $\mid n \mid < K$.

The linear filter is commonly used in image processing, since it can be easily implemented. Cubic convolution splines were suggested by Keys [4], who had as a main objective the reduction of the blur caused by linear interpolators. The pair (c)-(d) is a combination of two optimal filters designed under different criteria. The Parks-McClellan filter [21] minimizes the maximum absolute deviation between the frequency responses of the FIR and the ideal IIR low-pass filters. Designed



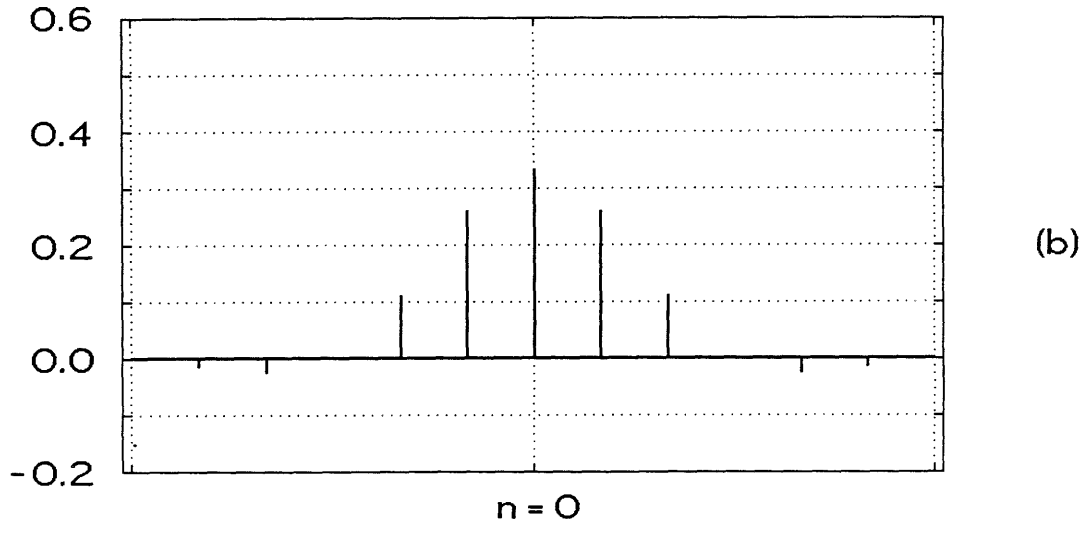
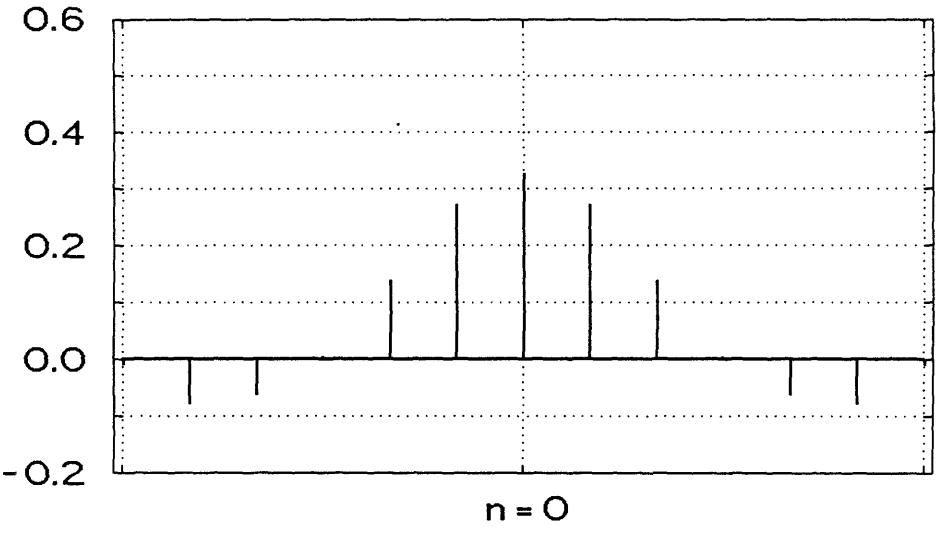
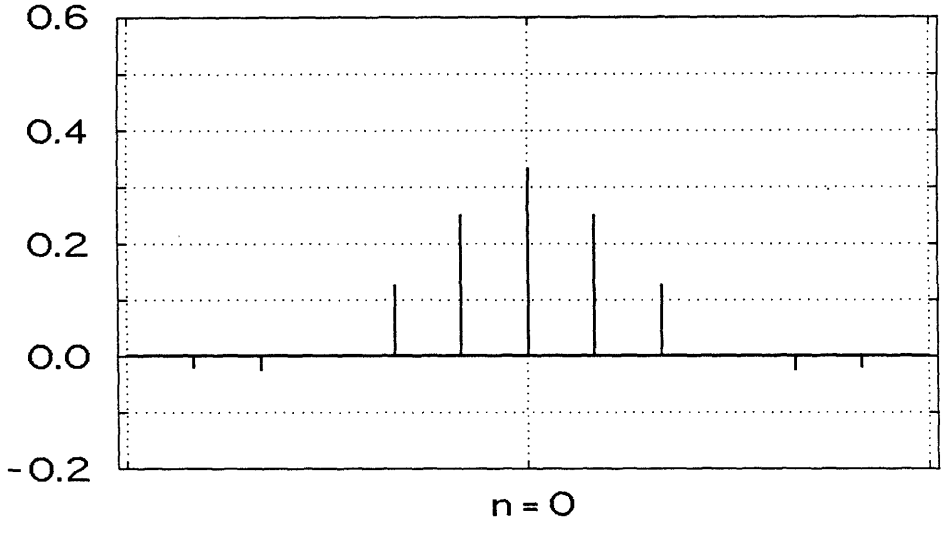


Fig. 3.7. Impulse responses of filters used for the experiments: (a) linear; and (b) cubic convolution [4].

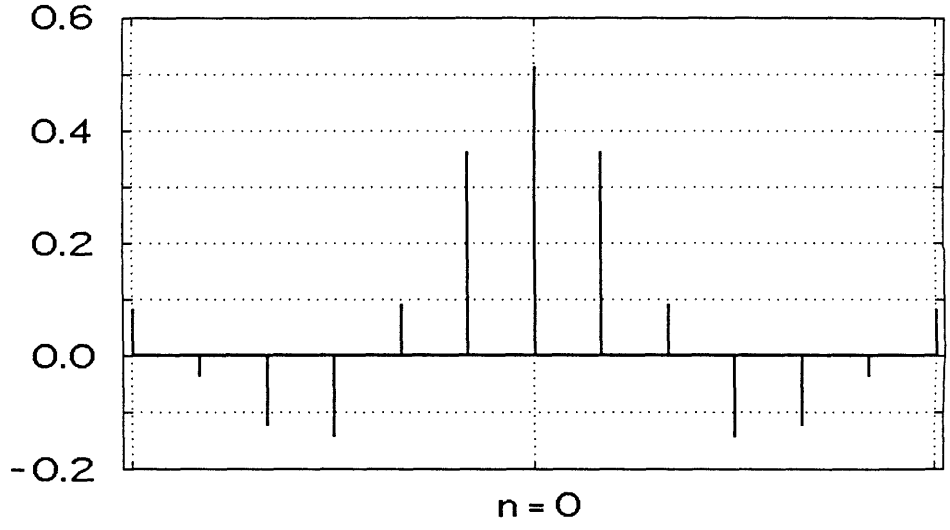


(c)

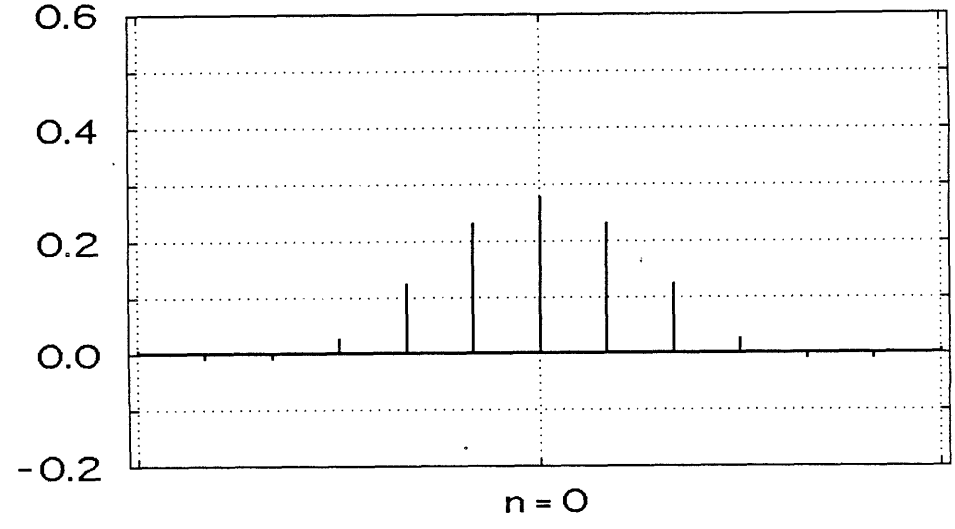


(d)

Fig. 3.7. Continued: (c) Parks-McClellan [21]; and (d) Oetken, Parks, and Schüssler [2].

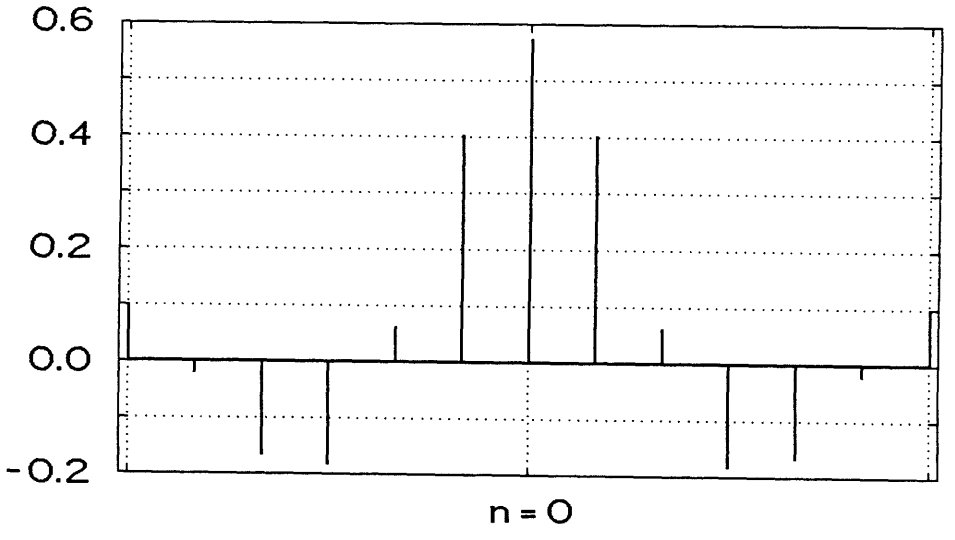


(e)

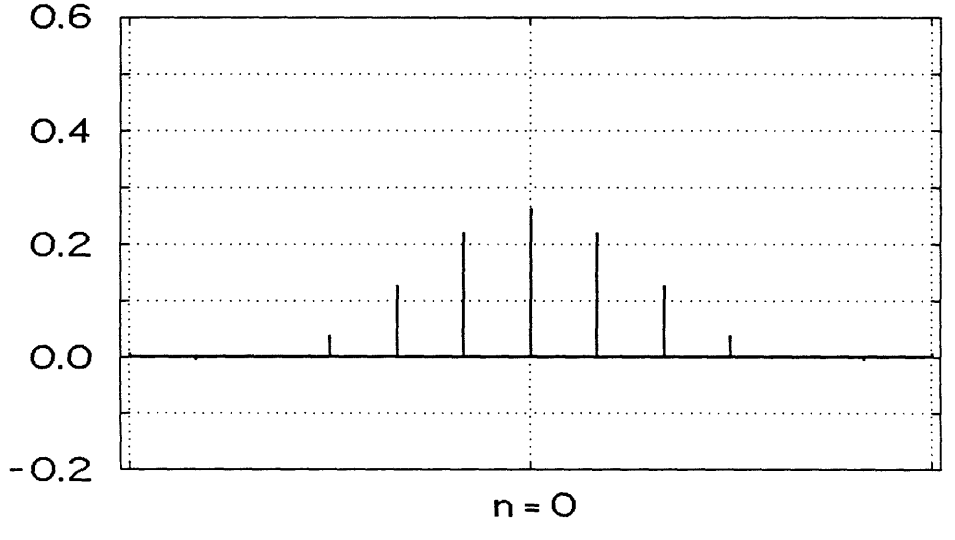


(f)

Fig. 3.7. Continued. Jointly-optimal filters for a flat observer: (e) pre; and (f) post.

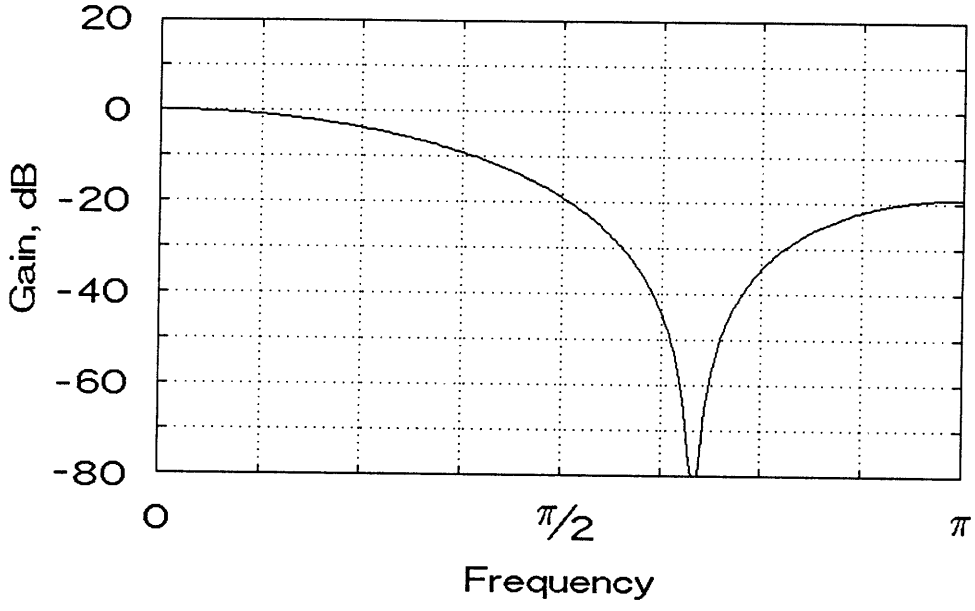


(g)

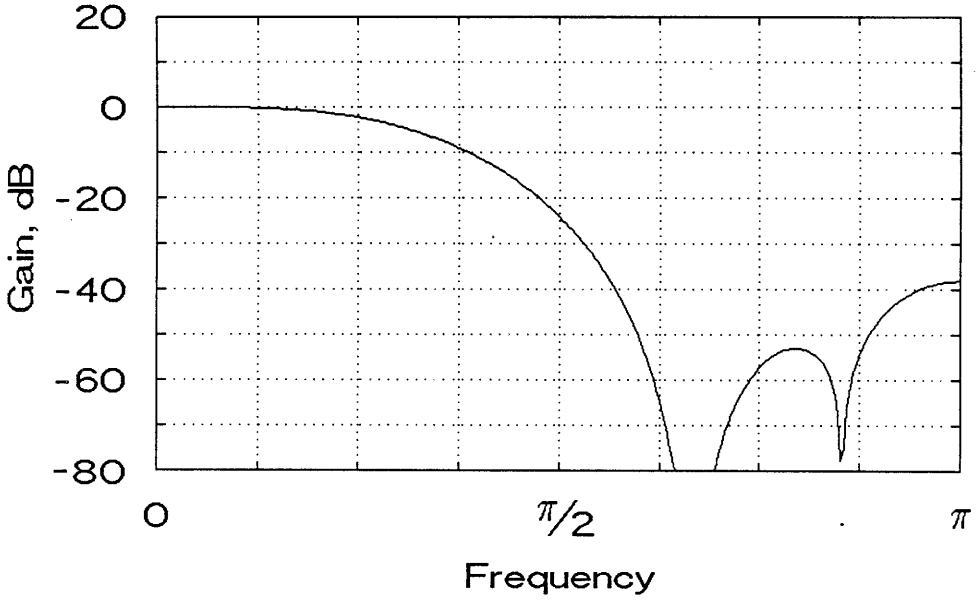


(h)

Fig. 3.7. Continued. Jointly-optimal filters for a bandpass observer: (g) pre; and (h) post.

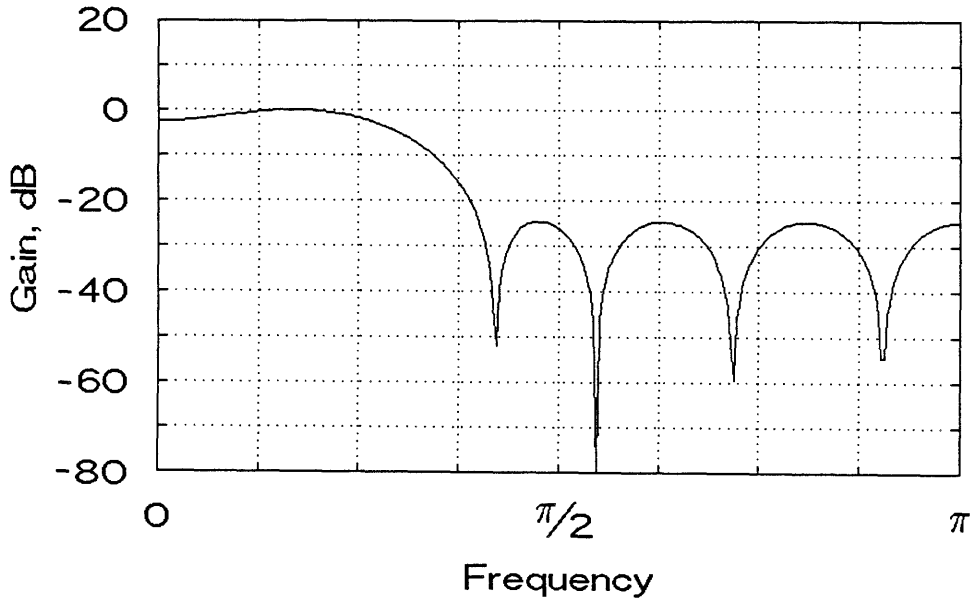


(a)

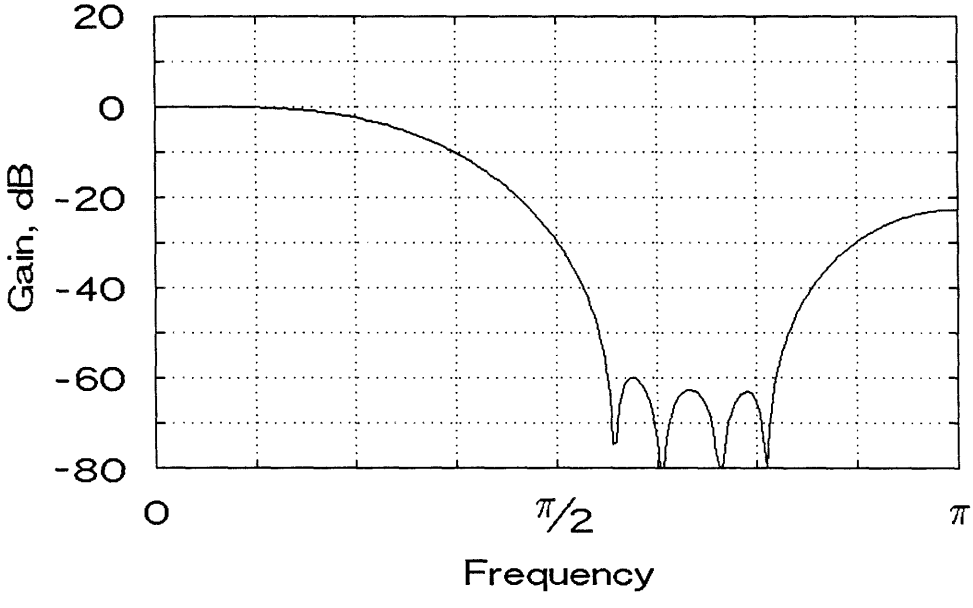


(b)

Fig. 3.8. Frequency responses of the filters in Fig. 3.7, in the same order.

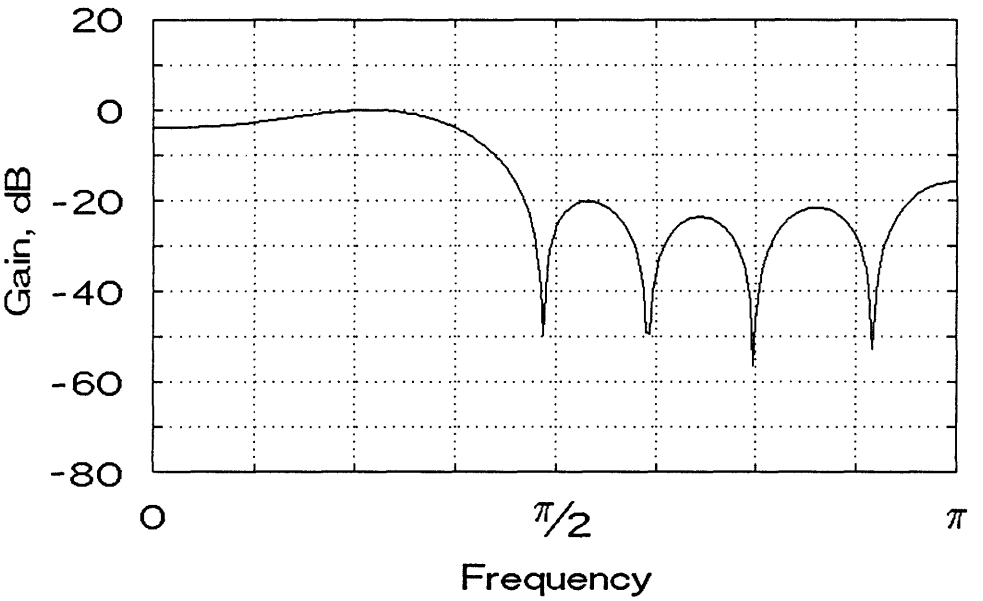


(c)

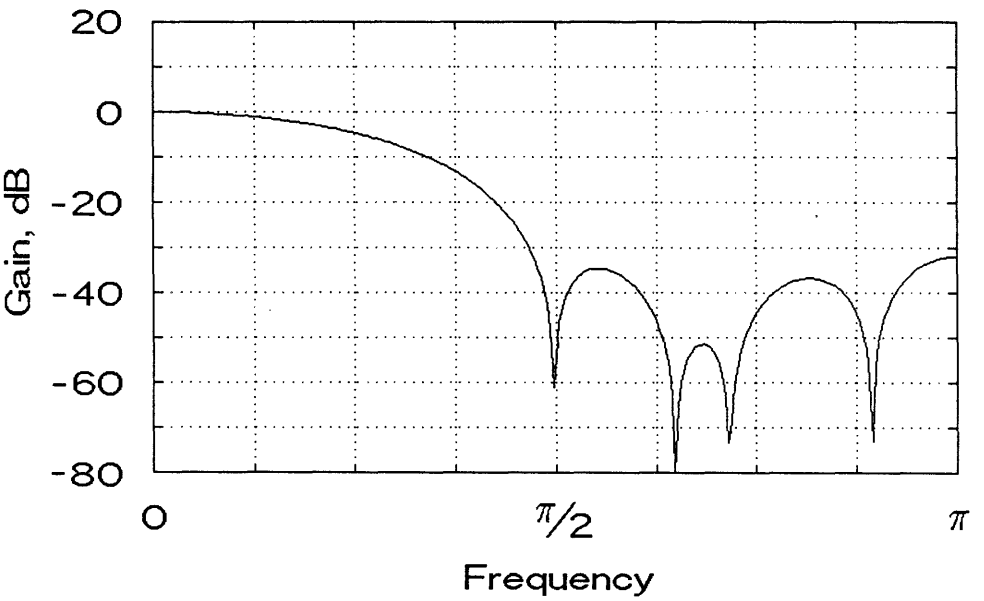


(d)

Fig. 3.8. Continued.

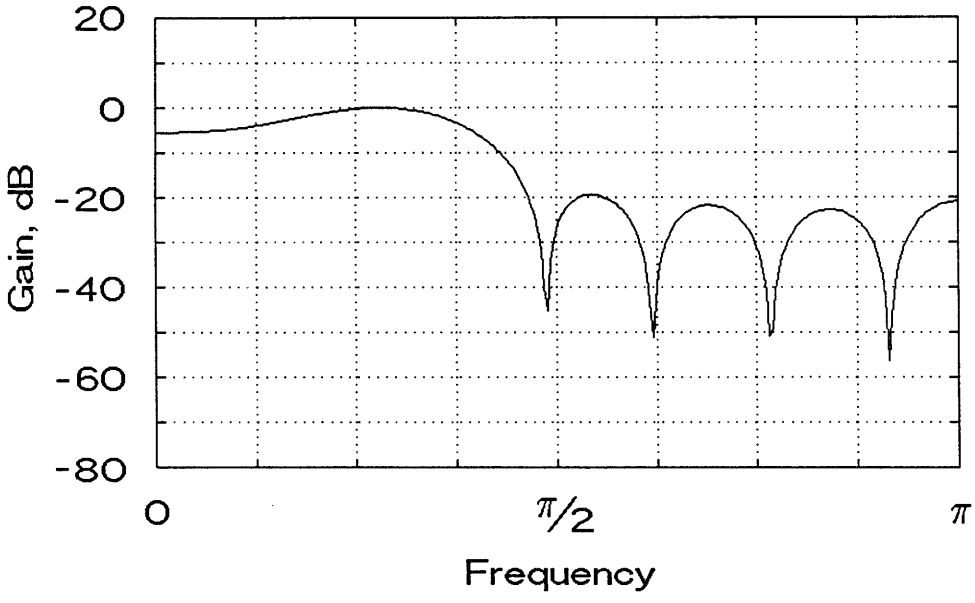


(e)

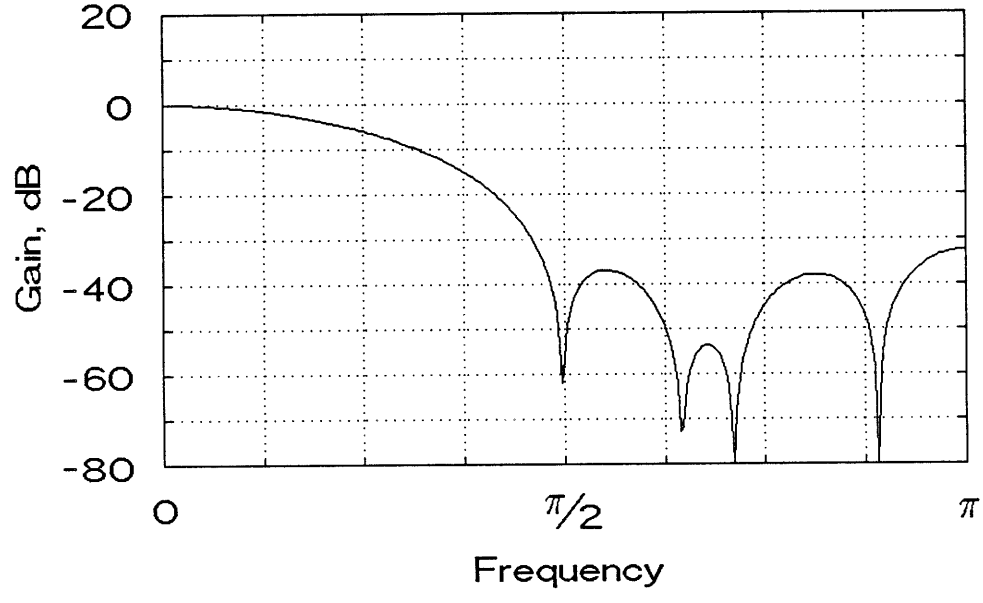


(f)

Fig. 3.8. Continued.



(g)



(h)

Fig. 3.8. Continued.

specifically for interpolation, the Oetken filter [22] minimizes the mean-square reconstruction error for band-limited signals. The Parks-McClellan filter in (c) was designed with a passband ripple of 0.15, stopband ripple of 0.065, and a transition band from 0.267π to 0.4π, which are appropriate for the values of K and L.

The optimal pairs (e)-(f) and (g)-(h) were designed with the algorithm of the previous section, assuming input and channel SNR's of 30 dB, and a first-order Gauss-Markov input signal with $\rho = 0.90$. The first pair corresponds to a constant observer response, and the second to the low-pass response of Fig. 3.2. As in the previous subsection, we see that the observer has little influence on the optimal filters.

We have used the original "KID" image of Fig. 3.9 (a) for our image processing experiment. The sampling grid is 256×240 pixels (picture elements), and each pixel is represented digitally with 8 bits. The original image was processed with the filter pairs of Fig. 3.7, using two-dimensional separable filters designed from the one-dimensional responses. The down-sampling factor K was equal to three in both the horizontal and vertical directions. The results are shown in Fig. 3.9 (b)–(d). We have not presented the image resulting from the use of the optimal filters for the low-pass observer, because it was virtually identical to the one obtained with the unweighted optimal filters; the differences would be lost in the reproduction process. The r.m.s. errors are indicated in Fig. 3.9 as a percentage that represents the ratio of the error variance to the signal variance.

The optimal filters led to an error improvement of 4.6 dB when compared to the spline–cubic convolution pair, and 1.7 dB when compared to the Parks-McClellan–Oetken pair. If we had chosen higher band-edge frequencies for the Parks-McClellan pre-filter, for example, the mean-square error in Fig. 3.9 (c) would have been higher. In general, a good choice for the parameters of the Parks-McClellan filter may require a trial-and-error approach.

Although the optimal filters have produced a sharp output image, the sampling artifacts are somewhat stronger than in the other images. These artifacts

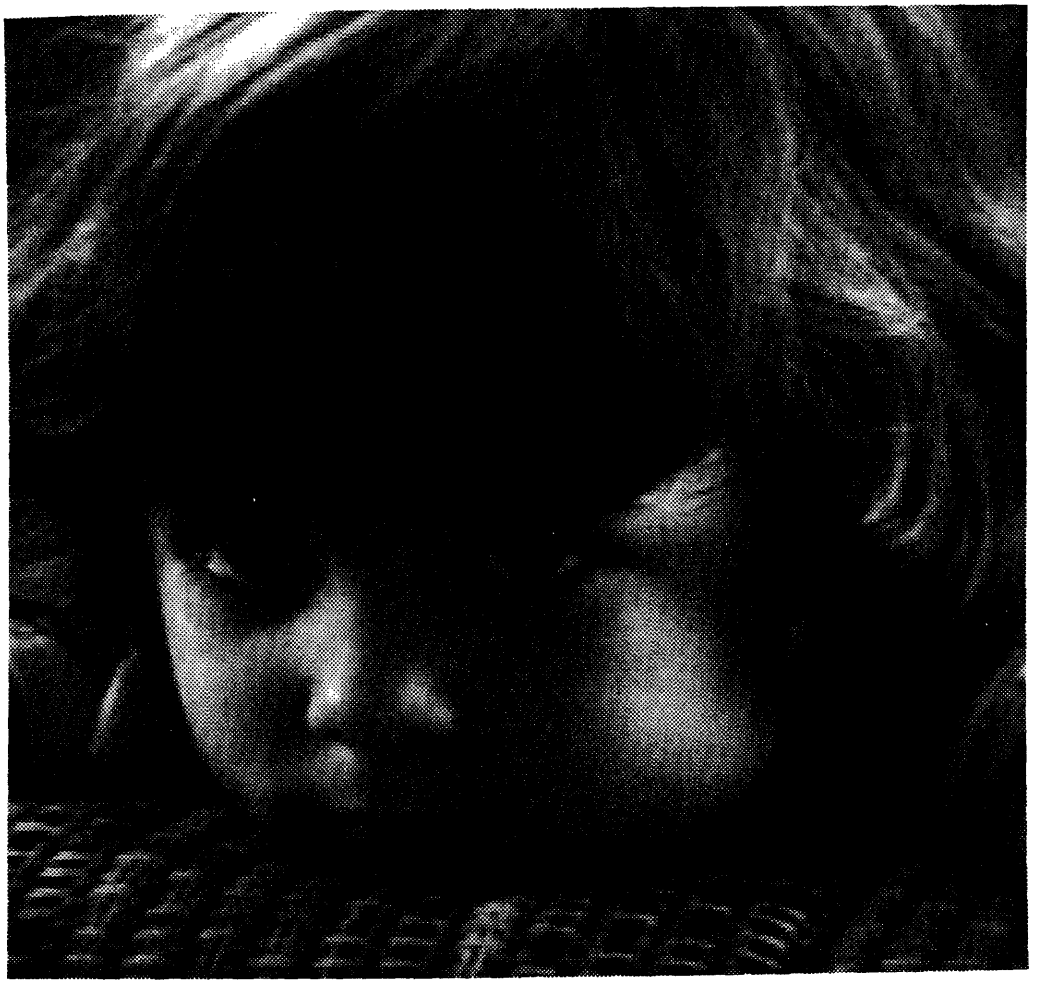


Fig. 3.9 (a) Original "KID" image, 256x240 pixels.

Fig. 3.9 (b) "KID" processed with the linear pre-filter and cubic convolution post-filter of Fig. 3.7 (a) and (b), respectively. R.m.s. error = 19.3%.

Fig. 3.9 (c) "KID" processed with the Parks-McClellan pre-filter and Oetken-Parks-Schüssler post-filter of Fig. 3.7 (c) and (d), respectively. R.m.s. error = 13.7%.

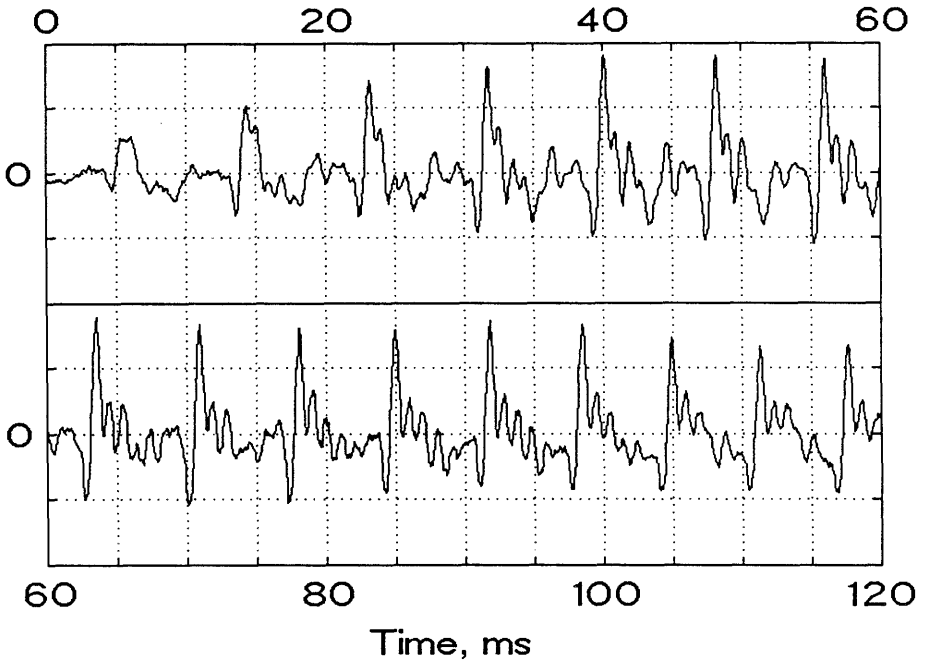
Fig. 3.9 (d) "KID" processed with the mean-square-optimal pre- and post-filters of Fig. 3.7 (e) and (f), respectively. R.m.s. error = 11.3%.

do not have a pronounced effect on the mean-square error, but they are somewhat unpleasing to the human eye. This trade-off between blur and sampling artifacts is typical of image interpolation [25].

It is clear, then, that for image processing applications, minimization of the mean-square reconstruction error might be of better value when coupled with some other relevant criteria. One approach towards this goal would be, for example, to force the pre- and post-filters to have exactly zero intersymbol interference. This could be accomplished by setting $f(rK) = g(sK) = 0$, where r and s are any integers, and re-deriving the optimal responses. For the design of the interpolator only, this approach was adopted in [3].

We have also processed a speech segment with the filters in Fig. 3.7. The original segment of 120 ms duration shown in Fig. 3.10 (a) corresponds to the vowel 'a' spoken by a male person. In Fig. 3.10 (b)–(d) we have the error signals, magnified by a factor of six, due to processing the original segment with the pre- and post-filter pairs: linear–cubic convolution, Parks-McClellan–Oetken, and mean-square optimal, respectively. We note that the optimal filters lead to a r.m.s. error of about half of the other two filter pairs, the main reason for that being the virtual absence of low-frequency errors. Since speech signals do not have sharp discontinuities as images do, there is no evidence of sampling artifacts on the waveforms of Fig. 3.10 (b)–(d).

This speech processing example is also a verification of the robustness of the optimal filters with respect to the input spectrum, since a first-order Gauss-Markov process with a correlation coefficient of 0.95 is not a good model for speech signals [26]. The results obtained with the input spectrum estimated from the incoming signal are not significantly better; this is an indication that optimal pre- and pos-filters are not much sensitive to variations on the input spectrum.



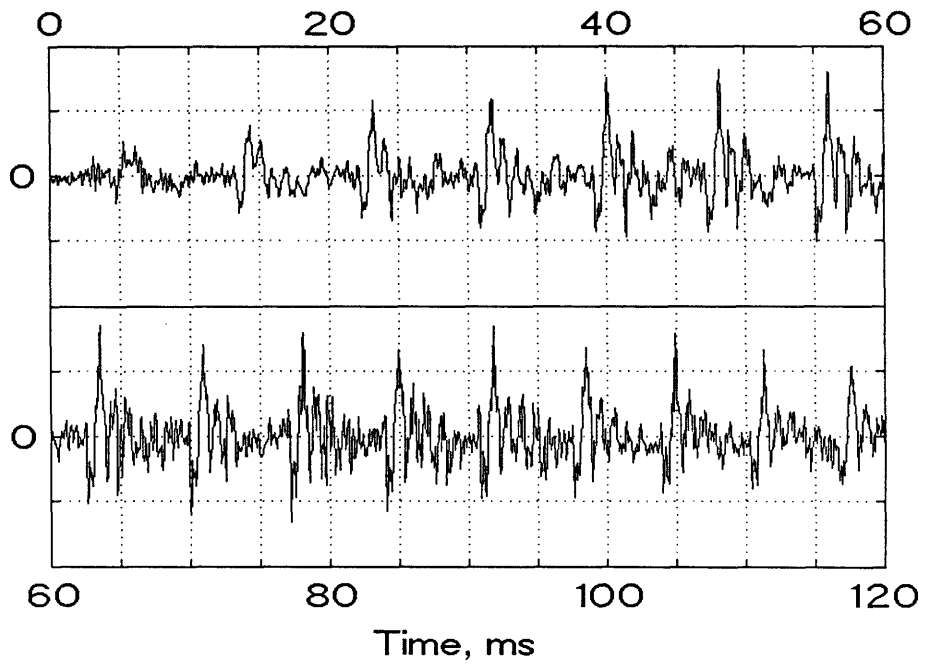
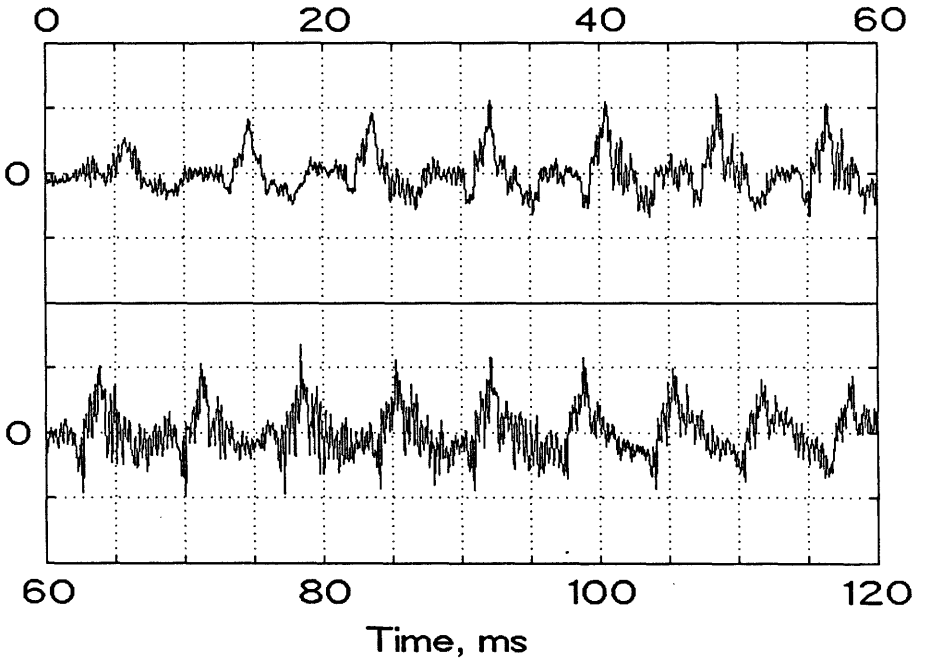
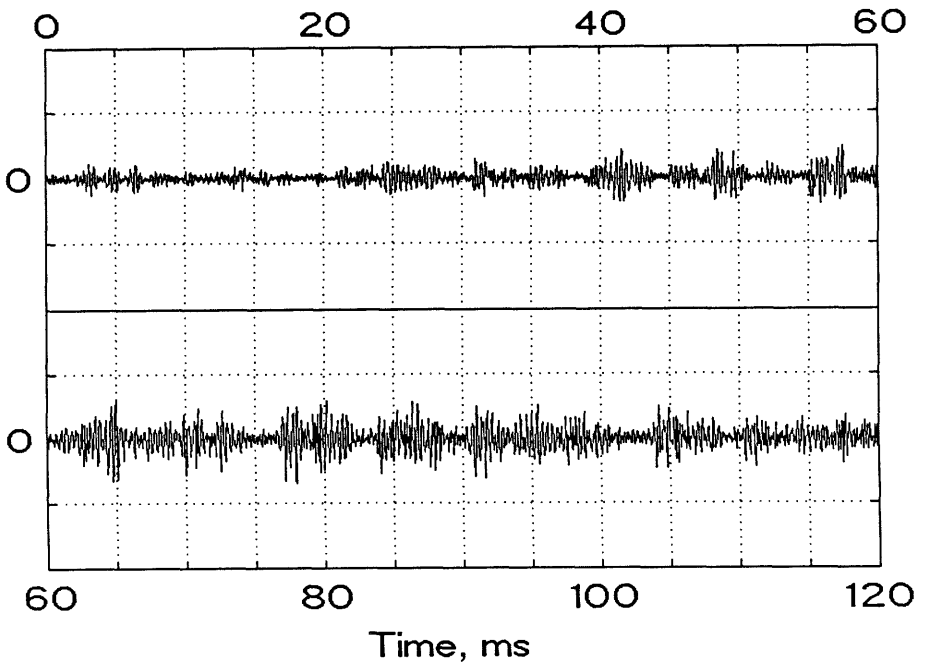


Fig. 3.10. (a) 120 ms speech segment for the vowel 'a', male speaker. (b) Error signal (x 6) for the filter pair in Fig. 3.7 (a)-(b), r.m.s. amplitude = 12.6%.



(c)



(d)

Fig. 3.10. Continued. Error signals (x 6) for: (c) filters of Fig. 3.7 (c)-(d), r.m.s. amplitude = 11.5%; and (d) optimal filters of Fig. 3.7 (e)-(f), r.m.s. amplitude = 5.5%.

3.5. Multidimensional filters

There are a number of signal processing applications in which multidimensional filters are used, for example, image processing, seismology, and multichannel telemetry, among others. There are cases where only a single filter must be designed, e.g., for sensor equalization, but in most cases we are interested in a multidimensional communications or storage system, which can be modeled by Fig. 3.1.

The design of optimal multidimensional FIR filters is somewhat more complicated than its unidimensional counterpart. Specifically, there is no simple way to generalize the Parks-McClellan algorithm to several dimensions, mainly because uniqueness of the optimal Chebyshev approximation is not guaranteed in several dimensions [27]. Although algorithms for the Chebyshev approximation of two-dimensional filters have been derived [27], they generally require a large number of iterations.

In our case, although the main objective is the simultaneous design of two filters, the fact that the error criterion is a weighted mean-square error makes the problem much simpler than that of Chebyshev approximation. In fact, the analysis of Section 3.3 can be easily extended to the multidimensional case. The details of such an extension are considered in this subsection.

The system model for a noisy multidimensional communications system is that of Fig. 3.11, in which the shift index $\mathbf{n}$ is a vector in $\Re^N$, where N is the number of dimensions, i.e.,

$$\mathbf{n} = [n_1, n_2, \ldots, n_N]' \,, \tag{3.26}$$

where $'$ denotes transposition. We assume that the input signal and noise sources in Fig. 3.11 are zero-mean stationary processes with known spectra.

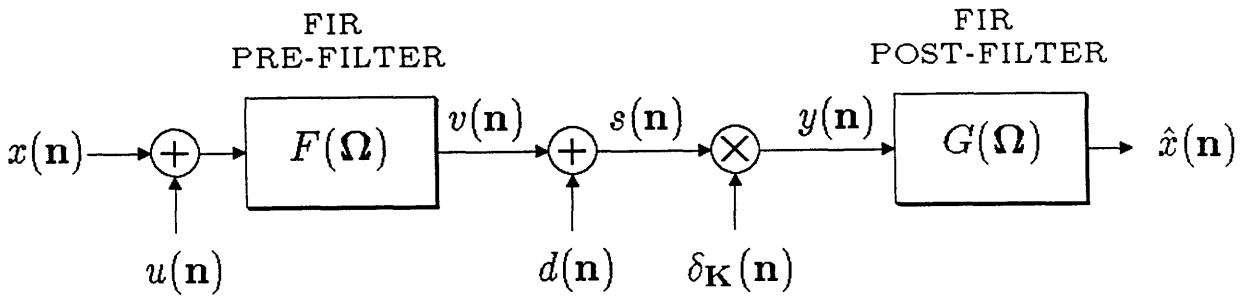


Fig. 3.11. Discrete-time multidimensional system with FIR filters.

The periodic sampling function $\delta_{\mathbf{K}}(\mathbf{n})$ is equivalent to its 1-D counterpart, with the exception that $\mathbf{K}$ is a matrix that determines the sampling structure. Of particular interest in 2-D are the rectangular and hexagonal sampling geometries of Fig. 3.12, because they are optimal sampling patterns for band-limited signals with rectangular and circular passbands, respectively [28]. Rectangular sampling corresponds to a diagonal $\mathbf{K}$, whereas for the hexagonal case the matrix $\mathbf{K}$ is determined by

$$\mathbf{K} = \begin{pmatrix} k_1 & k_1 \\ k_2 & -k_2 \end{pmatrix} \tag{3.27}$$

The term hexagonal sampling comes from the fact that a hexagonal lattice is obtained when $k_2 = b/\sqrt{3}$. Since in our case the original signal is already in discrete-time, k_1 and k_2 are integers, and an exact hexagonal lattice cannot be obtained. The region labeled $\mathcal{H}$ in Fig. 3.12 is one of the many possible regions that, when replicated periodically according to $\mathbf{K}$, covers the whole plane. The number of samples in $\mathcal{H}$ equals $|\det \mathbf{K}|$.

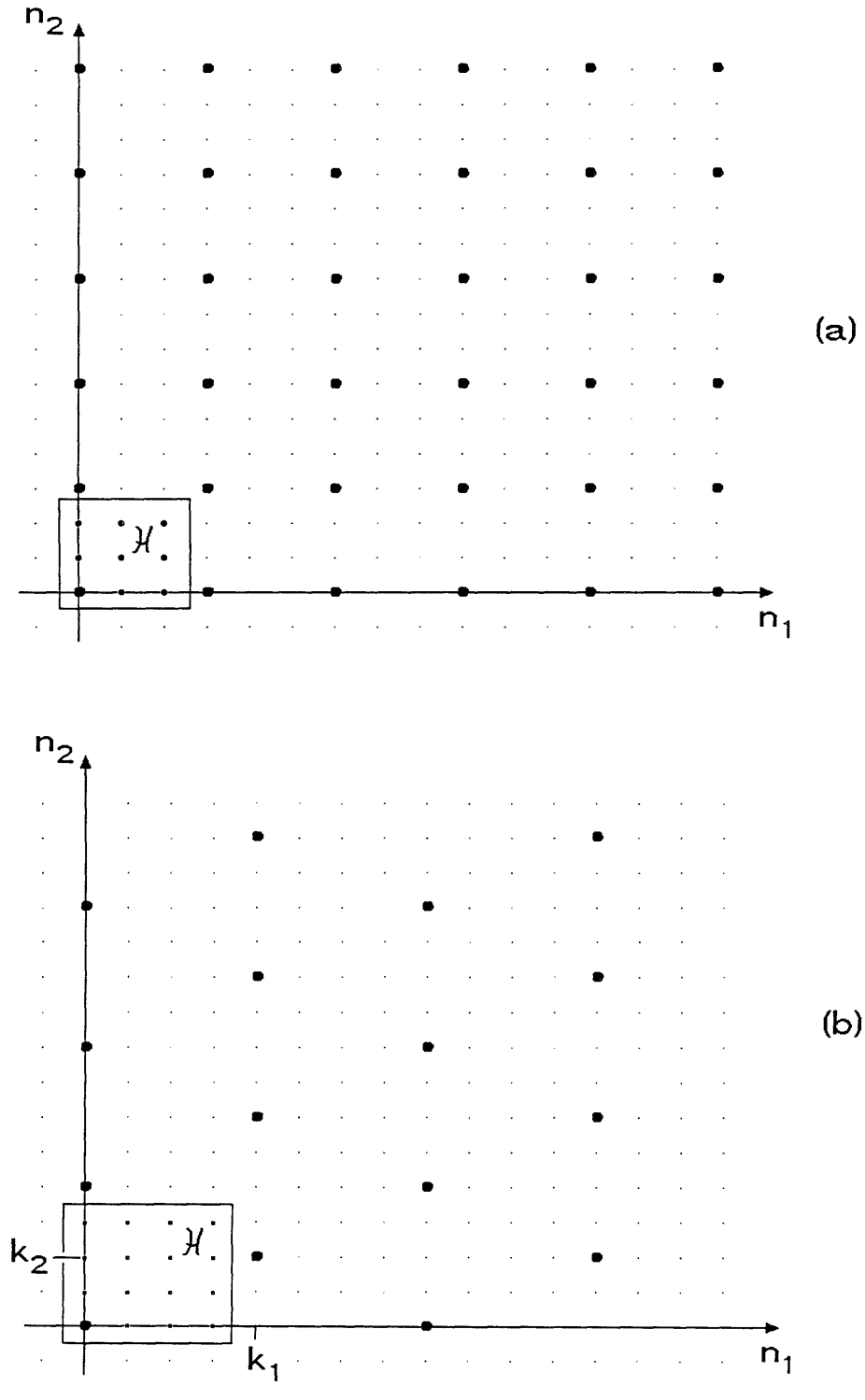


Fig. 3.12. (a) Rectangular sampling pattern in 2-D. (b) Hexagonal sampling lattice in 2-D.

We can write the periodic sampling function $\delta_{\mathbf{K}}(\mathbf{n})$ in the form

$$\begin{aligned}\delta_{\mathbf{K}}(\mathbf{n}) &= \mid \det \mathbf{K} \mid \sum_{\mathbf{r}} \delta(\mathbf{n} + \mathbf{K}\mathbf{r}) \\ &= \mid \det \mathbf{K} \mid \sum_{r_1=-\infty}^{\infty} \cdots \sum_{r_N=-\infty}^{\infty} \delta(n + Kr) \ ,\end{aligned} \tag{3.28}$$

By using the multidimensional Discrete Fourier Transform [27], we can represent the periodic sampling function as a finite sum of complex exponentials,

$$\begin{aligned}\delta_{\mathbf{K}}(\mathbf{n}) &= \sum_{\mathbf{r}\in\mathcal{H}} \exp(j2\pi \mathbf{r}'\mathbf{K}^{-1}\mathbf{n}) \\ &= \sum_{\mathbf{r}\in\mathcal{H}} \exp\left[j(\mathbf{W}_{\mathbf{K}}\mathbf{r})'\mathbf{n}\right] \ ,\end{aligned} \tag{3.29}$$

where

$$\mathbf{W}_{\mathbf{K}} \triangleq 2\pi(\mathbf{K}^{-1})' \tag{3.30}$$

is the sampling frequency matrix; it is analogous to the scalar sampling frequency ω_K in 1-D. We stress that (3.29) holds for any valid choice for the region $\mathcal{H}$.

Due to the sampling operation performed by $\delta_{\mathbf{K}}(\mathbf{n})$, the signals $y(\mathbf{n})$ and $\hat{x}(\mathbf{n})$ in Fig. 3.11 are cyclostationary, with their correlation functions having a periodicity pattern determined by the sampling matrix $\mathbf{K}$. A detailed analysis of sampling of multidimensional random processes was presented by Petersen and Middleton [28]. If we redefine the correlation functions as averages of expected values over the region $\mathcal{H}$, e.g.,

$$R_{sy}(\mathbf{n}) \triangleq \frac{1}{\mid \det \mathbf{K} \mid} \sum_{\mathbf{r}\in\mathcal{H}} \mathrm{E}[s(\mathbf{r})y(\mathbf{n}-\mathbf{r})] \ ,$$

we obtain non-periodic correlation functions that can be represented in the frequency domain as power spectra.

In order to optimize the filters in Fig. 3.11 we need an error measure equivalent to the 1-D measure in (3.1). The basic idea is still to average the variance of

the error signal over one sampling region. So, our error measure should be

$$\xi = \frac{1}{|\det \mathbf{K}|} \sum_{\mathbf{n} \in \mathcal{H}} \mathrm{E}[e^2(\mathbf{n})] \ , \tag{3.31}$$

or, in the frequency domain,

$$\xi = \frac{1}{(2\pi)^N} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \Phi_{ee}(\mathbf{\Omega}) \ d\mathbf{\Omega} \ , \tag{3.32}$$

where $\mathbf{\Omega}$ is the N-dimensional frequency variable.

By making use of two basic properties of the periodic sampling function, namely

$$\frac{1}{|\det \mathbf{K}|} \sum_{\mathbf{r} \in \mathcal{H}} \delta_{\mathbf{K}}(\mathbf{r} - \mathbf{n}) = 1 \tag{3.33}$$

and

$$\frac{1}{|\det \mathbf{K}|} \sum_{\mathbf{r} \in \mathcal{H}} \delta_{\mathbf{K}}(\mathbf{r}) \delta_{\mathbf{K}}(\mathbf{r} - \mathbf{n}) = \delta_{\mathbf{K}}(\mathbf{n}) \ , \tag{3.34}$$

we obtain the relationships

$$R_{yy}(\mathbf{n}) = R_{ss}(\mathbf{n}) \delta_{\mathbf{K}}(\mathbf{n}) \ , \tag{3.35}$$

and

$$R_{sy}(\mathbf{n}) = R_{ss}(\mathbf{n}) \ . \tag{3.36}$$

Combining (3.29) and (3.35), we get

$$\Phi_{yy}(\mathbf{\Omega}) = \sum_{\mathbf{r} \in \mathcal{H}} \Phi_{ss}(\mathbf{\Omega} + \mathbf{W}_{\mathbf{K}} \mathbf{r}) \ . \tag{3.37}$$

We can see that up to this point the 1-D analysis of the previous chapter extends immediately to the multidimensional case. If the pre- and post-filter impulse responses satisfy the symmetry and finiteness constraints

$$\begin{aligned} f(-\mathbf{n}) &= f(\mathbf{n}) \ , \\ g(-\mathbf{n}) &= g(\mathbf{n}) \ , \\ f(\mathbf{n}) &= 0, \quad \mathbf{n} \notin \mathcal{R}_f \ , \\ g(\mathbf{n}) &= 0, \quad \mathbf{n} \notin \mathcal{R}_g \ , \end{aligned} \tag{3.38}$$

where $\mathcal{R}_f$ and $\mathcal{R}_g$ are the regions of support for the pre- and post-filter, respectively, the equations for the optimal filters of Sections 3.2 and 3.3 hold also for the multidimensional case. Thus the optimal post-filter is given by the solution to the set of linear equations

$$\sum_{\mathbf{m} \in \mathcal{R}_g} g(\mathbf{m}) \psi(\mathbf{l} - \mathbf{m}) = \theta(\mathbf{l}) \qquad \mathbf{l} \in \mathcal{R}_g \ , \tag{3.39}$$

and the optimal pre-filter is given by the solution to

$$\sum_{\mathbf{m} \in \mathcal{R}_f} f(\mathbf{m}) \{ \gamma(\mathbf{l} - \mathbf{m}) + \lambda [R_{xx}(\mathbf{l} - \mathbf{m}) + R_{uu}(\mathbf{l} - \mathbf{m})] \} = \vartheta(\mathbf{l}) \qquad \mathbf{l} \in \mathcal{R}_f \ , \tag{3.40}$$

where λ is still a scalar Lagrange multiplier associated with the power constraint, and the sequences $\psi(\mathbf{n})$, $\theta(\mathbf{n})$, $\gamma(\mathbf{n})$, and $\vartheta(\mathbf{n})$ are trivial extensions of their 1-D counterparts. Thus, (3.10), (3.11), and (3.16) can be directly applied with $\mathbf{\Omega}$ replacing the scalar ω, and $\Gamma(\mathbf{\Omega})$ is given by

$$\Gamma(\mathbf{\Omega}) \triangleq [(1 + 2\beta)\Phi_{xx}(\mathbf{\Omega}) + \Phi_{uu}(\mathbf{\Omega})] \sum_{\mathbf{r} \in \mathcal{H}} | \, W(\mathbf{\Omega} - \mathbf{W_K r}) \, |^2 \, G^2(\mathbf{\Omega} - \mathbf{W_K r}) \ . \tag{3.41}$$

By appropriate index mappings, the multidimensional systems of equations in (3.39) and (3.40) can be converted to standard matrix equations, which can then be easily solved. The only difference from the 1-D case is that the matrices will be block-Toeplitz instead of Toeplitz. There are fast algorithms for block-Toeplitz system of equations, which are basically extensions of the Levinson recursion [29]. In practice, solving (3.39) is not a major issue, since most of the computer time is actually spent in computing the sequences $\psi(\mathbf{n})$, $\theta(\mathbf{n})$, $\gamma(\mathbf{n})$, and $\vartheta(\mathbf{n})$.

The algorithm in Section 3.3 applies directly to the multidimensional case, with (3.39) and (3.40) being alternately solved until convergence is attained. Under the mild assumption that $\Phi_{xx}(\mathbf{\Omega})$ and $\Phi_{dd}(\mathbf{\Omega})$ are strictly positive for all $\mathbf{\Omega}$, the functions $\Psi(\mathbf{\Omega})$ and $\Gamma(\mathbf{\Omega})$ will also be strictly positive for all $\mathbf{\Omega}$, and thus the systems of equations in (3.40) and (3.39) will have unique solutions. Hence, the multidimensional algorithm will also be a valid coordinate descent method, guaranteed to converge at least to a stationary point of the error measure ξ.

As in the 1-D case, with the use of a unit-sample as the initial guess for the pre-filter we have never had any convergence problem with the multidimensional version of the algorithm for joint optimization of the filters. Another logical choice for the initial guess is a $\sin(x)/x$ function, which leads to the same answer as the unit-sample, in about the same number of iterations. For the examples described below, the algorithm has converged to a precision of 10^{-3} in the filter coefficients in $4 - 8$ iterations.

Examples

In Fig. 3.13 and Fig. 3.14 we have two examples of optimal pre- and post-filter pairs, for rectangular and hexagonal sampling, respectively. The parameters for the rectangular case were $\mathbf{K} = \text{diag}(4,4)$, no input noise, a channel SNR of 30 dB, and a Gauss-Markov circularly symmetric input spectrum with a radial correlation coefficient of 0.9. The regions of support for the pre- and post-filter were set as

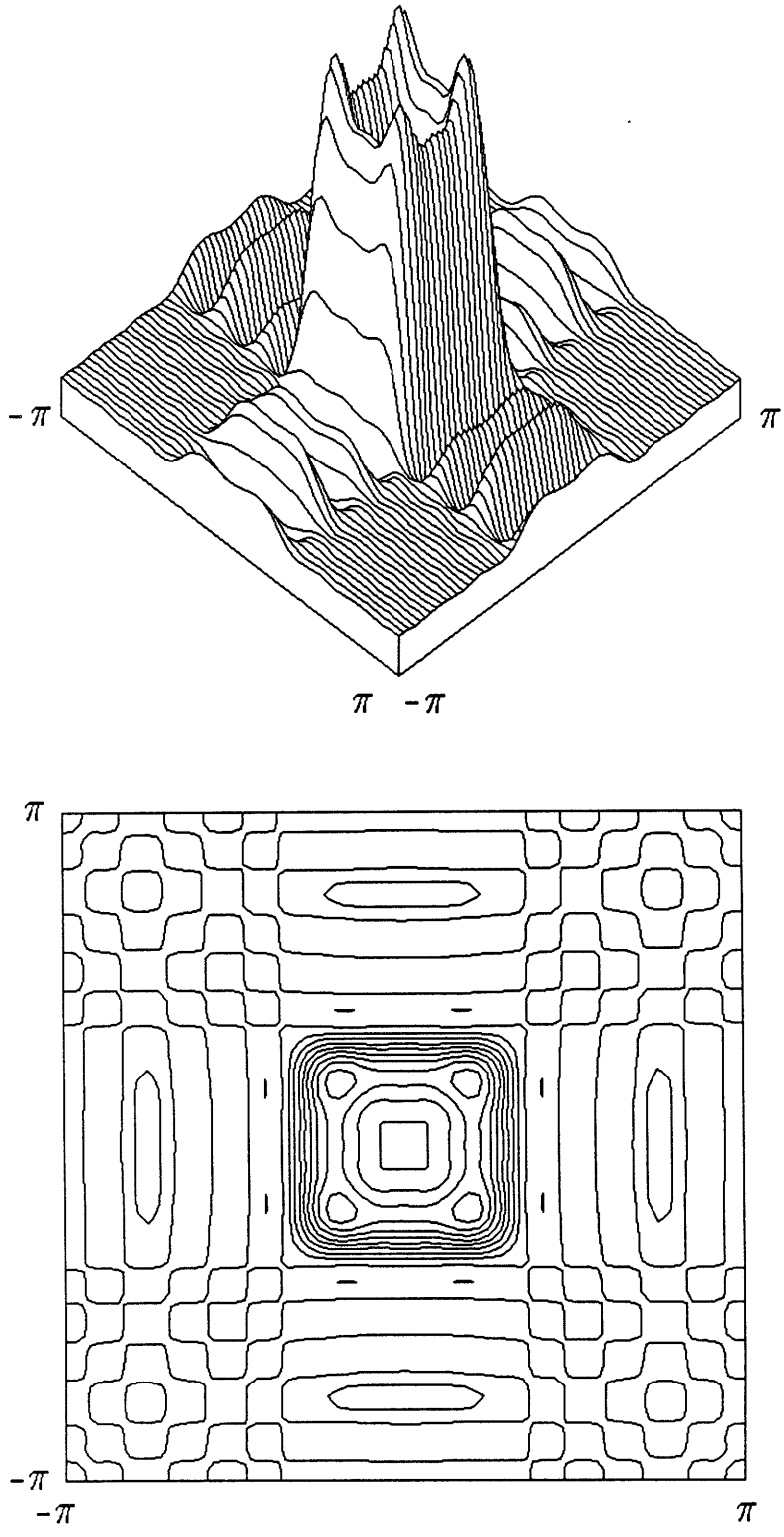
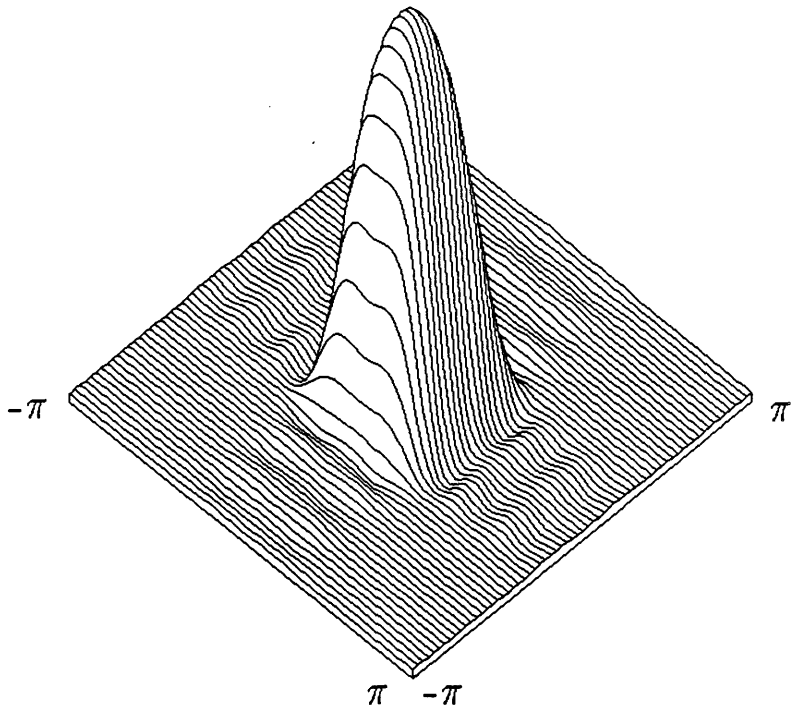


Fig. 3.13. (a) Frequency response (linear amplitude scale) of an optimal pre-filter with a square region of support. (b) Contours of constant amplitude, at intervals of 0.2.



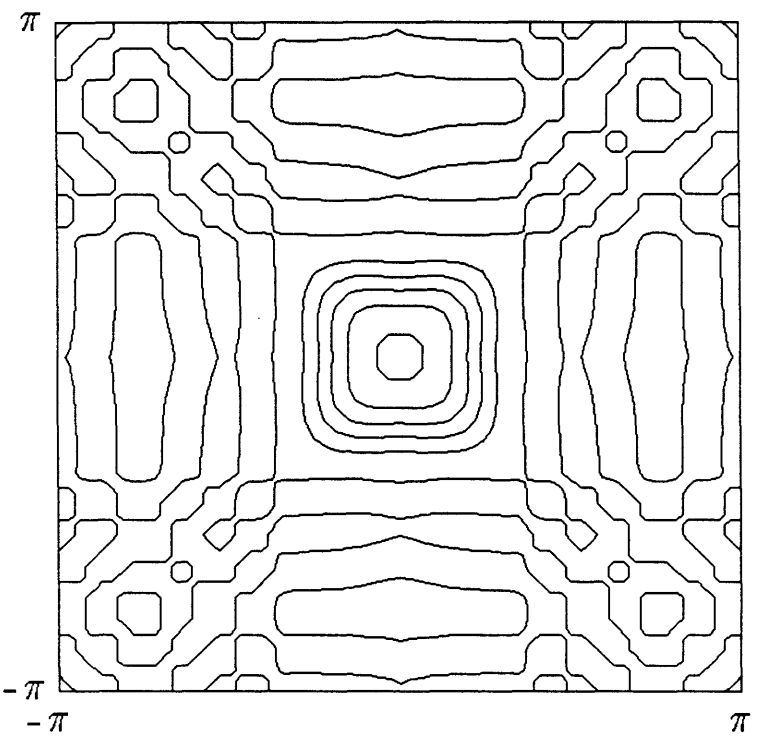


Fig. 3.13. Continued. (c) Frequency response of the optimal post-filter with a square region of support, corresponding to the optimal pre-filter in Fig. 3.13 (a). (d) Contours of constant amplitude, at intervals of 0.2.

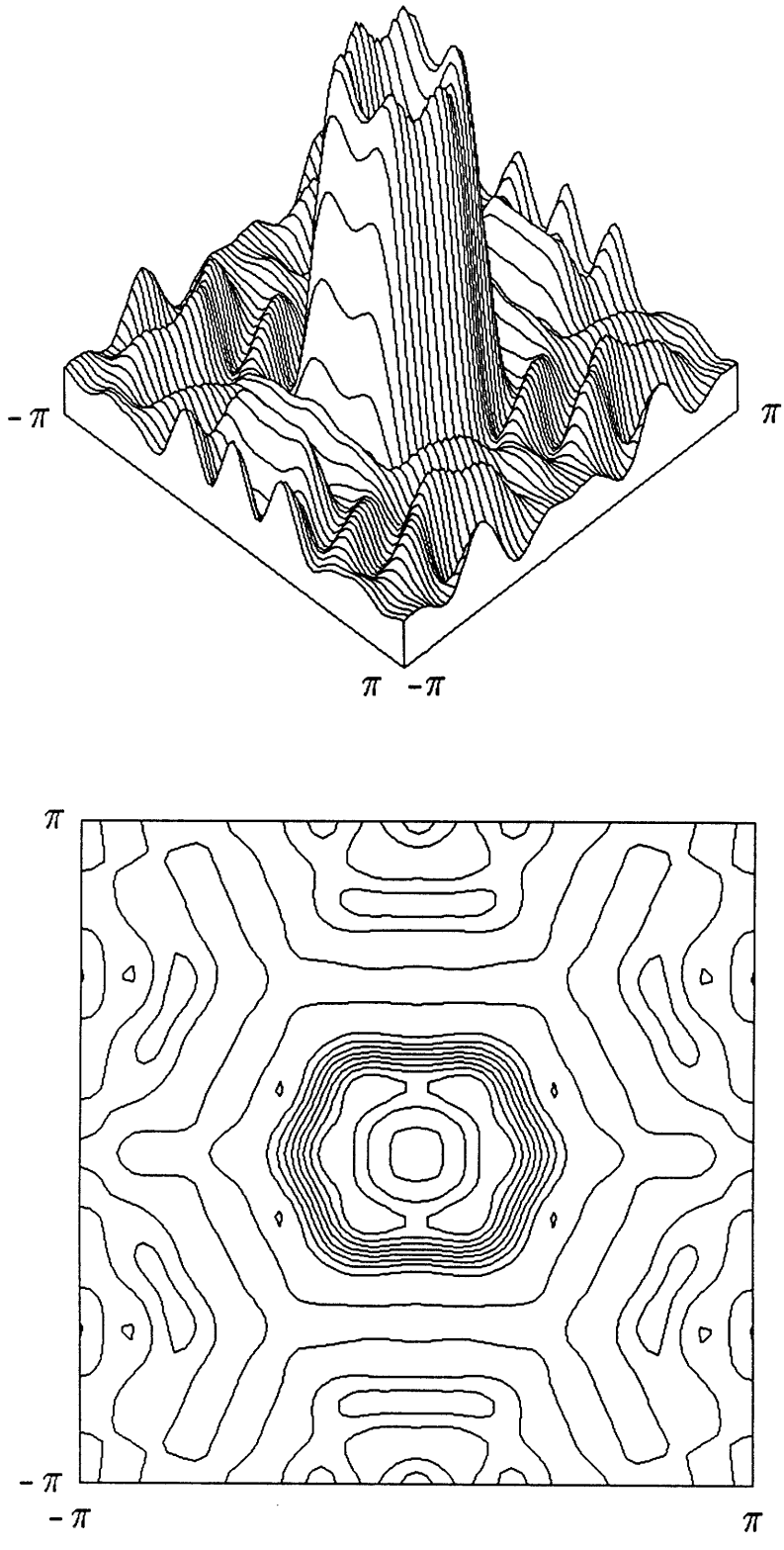


Fig. 3.14. (a) Frequency response (linear amplitude scale) of an optimal pre-filter with a hexagonal region of support. (b) Contours of constant amplitude, at intervals of 0.2.

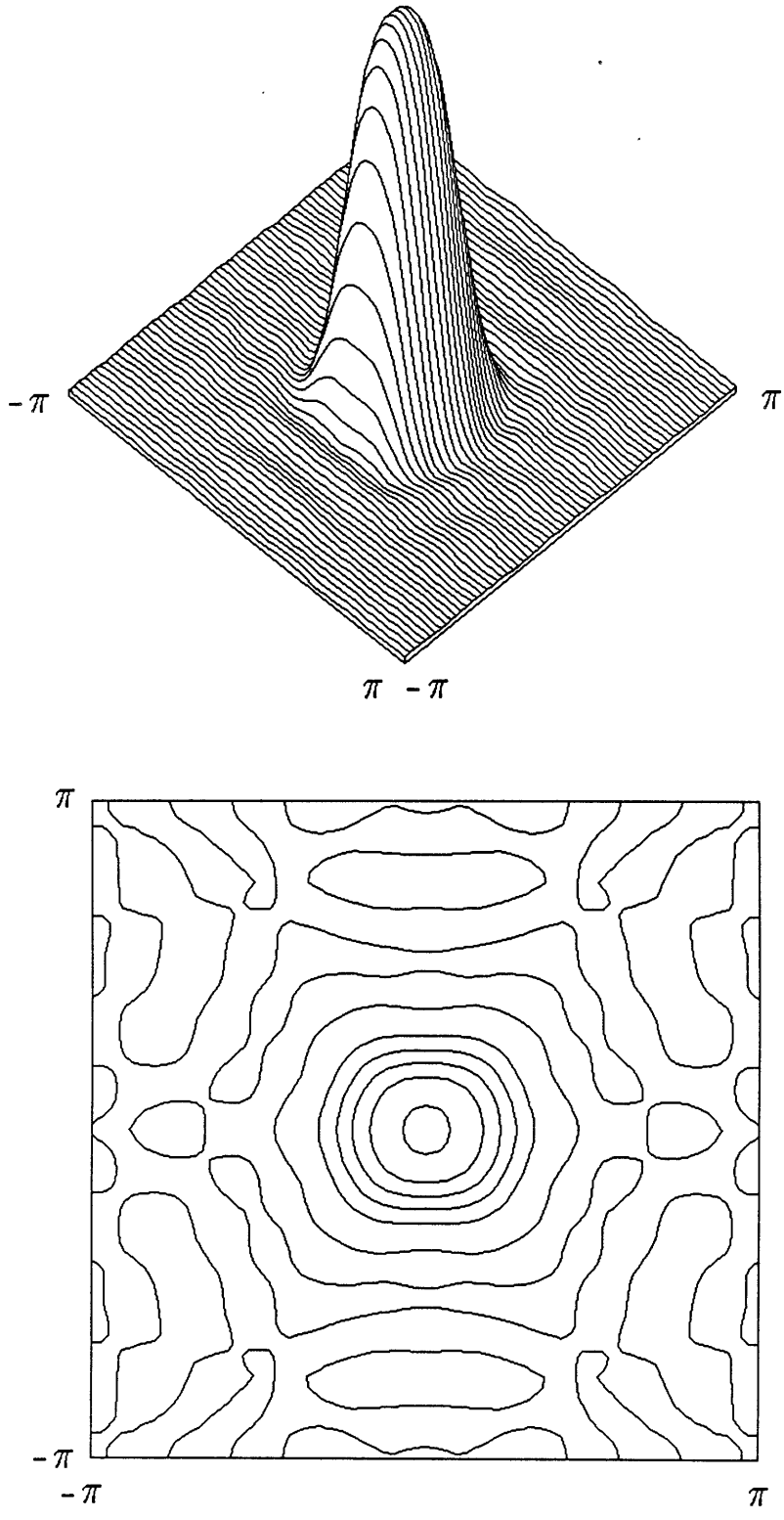


Fig. 3.14. Continued. (c) Frequency response of the optimal post-filter with a hexagonal region of support, corresponding to the optimal pre-filter in Fig. 3.14 (a). (d) Contours of constant amplitude, at intervals of 0.2.

a square region centered at the origin, with sides of 17 samples. The hexagonal filters were designed with the same parameters, except that the sampling matrix had $k_1 = 4$ and $k_2 = 2$, and the region of support of the pre- and post-filters were approximately hexagon-shaped, with corners at (0,8), (8,4), and their symmetrical reflections. For easier viewing, we have kept a linear scale for the frequency response plots in Fig. 3.13 and Fig. 3.14. We note that the difference between the pre- and post-filter responses is even more pronounced than in 1-D.

3.6. Summary

We have derived in this chapter an iterative algorithm for the design of jointly-optimal FIR pre- and post-filters under a weighted mean-square error criterion. Unlike the previous chapter, it is not possible to derive closed-form expressions for the optimal pre- and post-filters. As a by-product, we have also obtained the solutions to the independent optimization of either the pre- or the post-filter. It is possible to design pre- and post-filters with short impulse responses (on the order of twice the down-sampling ratio) that lead to a weighted mean-square error that is just slightly higher than that of the optimal ideal IIR filters.

The optimal pre- and post filter responses are significantly different from each other. The optimal pre-filter has much higher stopband ripples than the optimal post-filter. The former also has slight high-frequency boost in the passband, whereas the latter has a monotonically-decaying passband response. This is an indication that, even if sub-optimal filters are adopted, the decimation and interpolation filters should not be designed to have the same response.

The good practical performance of the optimal FIR filters has been verified by means of some image and speech processing examples, which suggest that the mean-square reconstruction error criterion is a reasonable one for practical filter design. This is in contrast with another use of the mean-square error for the design

of FIR filters, in which the m.s. error is measured between the frequency responses of the FIR filter and that of an ideal prototype. Generally, this latter approach leads to filters with large ripples both in the passband and the stopband [1],[27], and so it is of limited practical value.

We have also derived the optimal multidimensional FIR pre- and post-filters, through a simple extension of the 1-D analysis. If a single filter is to be optimized, either the pre- or the post-filter, the design problem reduces to solving a system of linear equations on the filter impulse response coefficients, as in the 1-D case. When both filters must be optimized, an iterative algorithm that alternates between finding the optimal pre-filter and finding the optimal post-filter should be employed. The algorithm is in essence the same as that for 1-D filters.

Applications

The optimal FIR filters described in this chapter will probably be most useful for sampling and interpolation systems where hardware cost is strongly dependent on the number of operations per second required by the filters, so that short-length FIR filters are a must. A good example is in the design of an image acquisition/display board that converts between different scanning resolutions in real time. Given the maximum FIR filter order supported by the hardware, the optimal set of filter coefficients can be designed using the techniques presented in this chapter.

Another example is in the design of interpolation filters for Compact Disc (CD) systems [30]. Presently, most CD players use analog filters for reconstruction of the PCM-coded signal. An alternative is to use an FIR interpolator to up-sample the PCM signal from 44.5 kHz to, say, 133.5 kHz. This up-sampled signal would then be converted to analog form by means of a D/A converter, which could be followed by a very simple analog filter. If such an approach is adopted, an optimal FIR interpolator could be designed with basis on the results of Section 3.2.

References

[1] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1975, chapter 5.

[2] G. Oetken, T. W. Parks and H. W. Schüssler, "New results in the design of digital interpolators," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-23, pp. 301–309, June 1975.

[3] A. D. Polydoros and E. N. Protonotarios, "Digital interpolation of stochastic signals," *IEEE Trans. Circuits Syst.,* vol. CAS-26, pp. 916–922, Nov. 1979.

[4] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-29, pp. 1153–1160, Dec. 1981.

[5] S. Kay, "Some new results in linear interpolation theory," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-31, pp. 746–749, June 1983.

[6] D. Radbel and R. J. Marks, II, "An FIR estimation filter based on the sampling theorem," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 455–460, Apr. 1985.

[7] P. R. Chevillat and G. Ungerboeck, "Optimum FIR transmitter and receiver filters for data transmission over band-limited channels," *IEEE Trans. Commun.,* vol. COM-30, pp. 1909–1915, Aug. 1982.

[8] R. Hummel, "Sampling for spline reconstruction," *SIAM J. Appl. Math.,* vol. 43, pp. 278–288, Apr. 1983.

[9] J. N. Ratzel, *The Discrete Representation of Spatially Continuous Images.* Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1980.

[10] P. Faubert, *Optimisation Conjointe du Pré-filtre et du Post-filtre pour la Decimation et L'Interpolation des Signaux Numeriques Multidimensionnels.* M.Sc. Thesis, University of Québec, Québec, Canada, 1985.

[11] R. E. Blahut, *Fast Algorithms for Digital Signal Processing.* Reading, MA: Addison-Wesley, 1985, chapter 11.

[12] F. G. Gustavson and D. Y. Y. Yun, "Fast algorithms for rational Hermite approximation and solution of Toeplitz systems," *IEEE Trans. Circuits Syst.,* vol. CAS-26, pp. 750–755, Sept. 1979.

[13] R. Kumar, "A fast algorithm for solving a Toeplitz system of equations," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, pp. 254–267, Feb. 1985.

[14] C. Manolakis, N. Kalouptsidis and G. Carayannis, "Efficient determination of FIR Wiener filters with linear phase," *Electron. Lett.*, vol. 18, pp. 429–431, May 13, 1982.

[15] U. Grenander and G. Szegō, *Toeplitz forms and their applications.* Los Angeles: University of California Press, 1958, chapter 10.

[16] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods.* New York: Academic Press, 1982, chapter 2.

[17] Y. Sugiyama, "An algorithm for solving discrete-time Wiener-Hopf equations based upon Euclid's algorithm," *IEEE Trans. Inform. Theory,* vol. IT-32, pp. 394–409, May 1986.

[18] G. Dahlquist and A. Björk, *Numerical Methods.* Englewood Cliffs, NJ: Prentice-Hall, 1974, chapter 6.

[19] D. G. Luenberger, *Linear and Non-linear Programming.* Reading, MA: Addison-Wesley, 1984, chapter 7.

[20] T. Abatzoglou and B. O'Donnel, "Minimization by coordinate descent," *J. Optimiz. Theory Appl.*, vol. 36, pp. 163–174, Feb. 1982.

[21] J. H. McClellan, T. W. Parks, and L. R. Rabiner, *FIR Linear Phase Filter Design Program, in Programs for Digital Signal Processing.* New York: IEEE Press, 1979, section 5.1.

[22] G. Oetken, *A Computer Program for Digital Interpolator Design.* ibid., Section 8.1.

[23] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1983, chapter 4.

[24] R. J. Clarke, *Transform Coding of Images.* London: Academic Press, 1985, chapter 6.

[25] W. K. Pratt, *Digital Image Processing.* New York: Wiley, 1978, chapters 4 and 23.

[26] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, N.J.: Prentice-Hall, 1978, chapter 8.

[27] D. E. Dudgeon and R. M. Mersereau, *Multidimensional Digital Signal Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1984, chapter 3.

[28] D. P. Petersen and D. Middleton, "Sampling and reconstruction of wave-number-limited functions in N-dimensional Euclidean spaces," *Inform. and Control,* vol. 5, pp. 279–323, 1962.

[29] N. Kalouptsidis, D. Manolakis, and G. Carayannis, "A family of computationally efficient algorithms for multichannel signal processing – a tutorial review," *Signal Processing,* vol. 5, pp. 5–19, 1983.

[30] H. Nakajima *et al.*, *The Sony Book of Digital Audio Technology.* Blue Ridge Summit, PA: TAB Books, 1983.

Chapter 4

Optimal Filters for Block Processing

Our objective in this chapter is to derive jointly-optimal pre- and post-filters for the system in Fig. 4.1, which is the block-processing equivalent to the systems studied in the previous chapters. Although there is no explicit indication of a sampler in Fig. 4.1, the model does allow for sampling, as discussed below. The input vector $\mathbf{x}$, also called a block, may be formed by a collection of samples from an unidimensional signal at different time instants, a single observation of a multidimensional signal, or a combination of both.

We will divide the analysis of the system in Fig. 4.1 into two cases: analog channels and digital channels. In the first case, considered in the next section, we assume that the channel noise is uncorrelated with the signal. In the second, which is the subject of Section 4.2, we shall study in more detail the effects of quantization, which is always present in digital channels.

In our study of pre- and post-filters for digital channels, we derive two important new results. First, we show that the optimal pre-filter must contain a Karhunen-Loève transform as its last factor, *without* the assumption that the input signal and noise sources are Gaussian. This is a generalization of the classical property of block coding of Gaussian signals with scalar quantizers, namely, that the vector to be quantized must have uncorrelated components. Second, we demonstrate that the overall system performance can be improved by the inclusion of dithering (also called pseudo-random noise) in the quantization process.

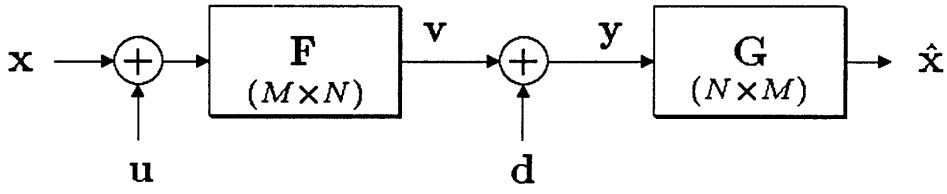


Fig. 4.1. Block processing system.

The input and reconstructed signals $\mathbf{x}$ and $\hat{\mathbf{x}}$ are vectors in the N-dimensional real space $\mathbb{R}^N$; the channel input and output, $\mathbf{v}$ and $\mathbf{y}$, are vectors in $\mathbb{R}^M$. The pre-filter $\mathbf{F}$ and post-filter $\mathbf{G}$ are $M \times N$ and $N \times M$ matrices, respectively. If the pre-filter has fewer rows than columns, i.e., if $M < N$, it is also a sampler, with a down-sampling factor of N/M. Since the sampling resolution is generally limited by the channel, we will consider $M \leq N$ throughout this chapter. As in the previous chapters, we assume that the input noise $\mathbf{u}$ is uncorrelated with the input signal, but the channel noise $\mathbf{d}$ may be correlated with the channel input $\mathbf{v}$. All signals and noises have zero mean.

We cannot adopt a frequency-domain framework here, as we did in the previous two chapters, since that would not be applicable to the case where the input signal is a block of samples from an unidimensional signal, for example, because in that context the block filters $\mathbf{F}$ and $\mathbf{G}$ are time-variant. Thus, in order to keep the generality of the model of Fig. 4.1, we work exclusively in the time domain (or space domain, whichever is appropriate), without making use of Fourier transforms.

Within the formulation that the signals and filters in Fig. 4.1 are vectors and matrices, respectively, our problem of pre- and post-filter optimization is similar to

that of A-optimal Bayesian experiment design, in Statistics [1],[2]. On one hand, we have a more general formulation, since we do not assume that the channel noise is white and uncorrelated with the signal, and we also have an input noise source. On the other hand, our model is more restricted because we do not work with general constraint sets for the pre-filter; we are interested only in transmitted-power limitations.

4.1. Optimal Filters for Analog Channels

We proceed in this section under the assumption that the channel input $\mathbf{v}$ and noise $\mathbf{d}$ are uncorrelated. This is a reasonable assumption for analog channels, where no quantization is present. If the additional restriction $\mathbf{u} \equiv \mathbf{0}$ were imposed, the system of Fig. 4.1 would be in the form studied by Lee and Petersen [3]. Our work in this section is not only an extension of their results for the more general system model of Fig. 4.1, but we also derive sub-optimal solutions that lead to fast filter structures. With these sub-optimal filters, an error level that is typically less than 0.1 dB higher than that of the optimal filters can be obtained

The signal estimate $\hat{\mathbf{x}}$ in Fig. 4.1 is related to the original signal $\mathbf{x}$ by

$$\hat{\mathbf{x}} = \mathbf{GF}(\mathbf{x} + \mathbf{u}) + \mathbf{Gd} \ , \tag{4.1}$$

where $\mathbf{x} \in \Re^N$ and $\mathbf{d} \in \Re^M$. We want to find the filters $\mathbf{F}$ and $\mathbf{G}$ such that the error measure

$$\xi \triangleq N^{-1} \mathrm{E}[\, \|\mathbf{W}(\hat{\mathbf{x}} - \mathbf{x})\|^2 \,] \tag{4.2}$$

is minimized. The factor N^{-1} normalizes the error, and so it can be viewed as the average weighted mean-square error per element of the vector $\mathbf{x}$. The $N \times N$ weighting matrix $\mathbf{W}$ is the observer filter. In order to avoid the existence of subspaces in which the error weighting is zero, we assume that $\det \mathbf{W} \neq \mathbf{0}$.

The error measure ξ can also be written in the form

$$\xi = N^{-1}\,\mathrm{E}\left[(\hat{\mathbf{x}}-\mathbf{x})'\mathbf{W}'\mathbf{W}(\hat{\mathbf{x}}-\mathbf{x})\right] . \tag{4.3}$$

where the prime denotes transposition. Since $\mathbf{W}$ is non-singular, the matrix $\mathbf{W}'\mathbf{W}$ above is positive definite.

In this section we assume that both additive noise sources in Fig. 4.1 are uncorrelated with the signals. Then, using (4.1) and (4.3), we get

$$\begin{aligned}\xi = {} & N^{-1}\,\mathrm{tr}\left\{\mathbf{W}'\mathbf{W}(\mathbf{I}-\mathbf{GF})\mathbf{R}_{\mathbf{xx}}(\mathbf{I}-\mathbf{GF})' + \mathbf{W}'\mathbf{W}\mathbf{G}\mathbf{F}\mathbf{R}_{\mathbf{uu}}\mathbf{F}'\mathbf{G}'\right\} \\ & + N^{-1}\,\mathrm{tr}\left\{\mathbf{W}'\mathbf{W}\mathbf{G}\mathbf{R}_{\mathbf{dd}}\mathbf{G}'\right\} ,\end{aligned} \tag{4.4}$$

where the autocorrelation matrices have their usual meanings, e.g.,

$$\mathbf{R}_{\mathbf{xx}} \triangleq \mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] ,$$

and $\mathrm{tr}\{\cdot\}$ is the trace operator.

Our goal is to find the pair of matrices $\mathbf{F}$ and $\mathbf{G}$ such that (4.4) is minimized. In what follows we shall adopt the same approach as in Chapter 2: we first derive the optimal post-filter, and obtain an expression for the error as a function of the pre-filter, which is then optimized. In the next subsection we obtain the optimal $\mathbf{G}$, whereas the optimal $\mathbf{F}$ will be derived in the following two subsections, for two different types of power restrictions: maximum total power and identical uncorrelated channels.

4.1.1. The Optimal Post-filter

Deriving the optimal post-filter for a given pre-filter $\mathbf{F}$ is a relatively simple exercise of the concepts of optimal estimation and Wiener filtering [4], [5]. Using the techniques in [6], [7], we can take the derivative of ξ with respect to the post-filter matrix, with the result

$$\frac{\partial \xi}{\partial \mathbf{G}} = 2\ N^{-1}\mathbf{W}'\mathbf{W}\{\mathbf{G}[\mathbf{R_{dd}} + \mathbf{F}(\mathbf{R_{xx}} + \mathbf{R_{uu}})\mathbf{F}'] - \mathbf{R_{xx}}\mathbf{F}'\}\ . \tag{4.5}$$

Since $\mathbf{W}$ is non-singular by assumption, $\det(\mathbf{W}'\mathbf{W}) \neq 0$. Thus, the unique solution to $\partial\xi/\partial\mathbf{G} = \mathbf{0}$ is

$$\mathbf{G_{OPT}} = \mathbf{R_{xx}}\mathbf{F}'[\mathbf{R_{dd}} + \mathbf{F}(\mathbf{R_{xx}} + \mathbf{R_{uu}})\mathbf{F}']^{-1}\ , \tag{4.6}$$

where the inverse of the term within brackets is assumed to exist. This is a mild assumption, which is satisfied if $\mathbf{R_{dd}}$ is not singular. We see that the optimal post-filter does not depend on the weighting function, for a fixed pre-filter. This independence was also verified for the optimal IIR post-filter in Chapter 2. We note that (4.6) is a standard result of linear estimation theory; it is the unique optimal signal estimator given the received noisy measurements. We recognize in (4.6) the two classic factors of an optimal estimator: the cross-correlation between the input vector and the received vector, and the autocorrelation of the received vector.

The post-filter in (4.6) is the unique stationary point of the error, and therefore the unique candidate for the minimum. However, there is no *a priori* guarantee, strictly speaking, that the error attains a minimum, since $\mathbf{G}$ does not belong to a compact set. Although the error expression in (4.4) looks like a quadratic form, this is not so obvious as it was in the previous chapters. Thus, we believe that a little effort should be dedicated to the verification of global optimality of (4.6). This is accomplished by the following theorem.

Theorem. *Assume that* $\mathbf{R_{dd}} + \mathbf{F}(\mathbf{R_{xx}} + \mathbf{R_{uu}})\mathbf{F}'$ *is invertible. Then, the post-filter* $\mathbf{G_{OPT}}$ *leads to the unique global minimum of the error.*

Proof: We need to use the vec operator [7], [8], defined by

$$\text{vec}\,\mathbf{A} \triangleq \begin{pmatrix} \mathbf{A}_{\cdot 1} \\ \mathbf{A}_{\cdot 2} \\ \mathbf{A}_{\cdot n} \end{pmatrix} , \tag{4.7}$$

where $\mathbf{A}_{\cdot i}$ is the i-th column vector of the matrix $\mathbf{A}$, and the Kronecker product

$$\mathbf{A} \otimes \mathbf{B} \triangleq \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{pmatrix} . \tag{4.8}$$

Using those two operators, we can rewrite the error as

$$\begin{aligned} \xi = {} & N^{-1}\,\text{tr}\,\{\mathbf{W}'\mathbf{W}\mathbf{R_{xx}}\} - 2N^{-1}(\,\text{vec}\,\mathbf{G}')'\,\text{vec}\,(\mathbf{F}\mathbf{R_{xx}}\mathbf{W}'\mathbf{W}) \\ & + N^{-1}(\,\text{vec}\,\mathbf{G}')'\{(\mathbf{W}'\mathbf{W}) \otimes [\mathbf{R_{dd}} + \mathbf{F}(\mathbf{R_{xx}} + \mathbf{R_{uu}})\mathbf{F}']\}\,\text{vec}\,\mathbf{G}' , \end{aligned} \tag{4.9}$$

which is now clearly a quadratic form on vec $\mathbf{G}'$. Since $\mathbf{W}'\mathbf{W}$ and $\mathbf{R_{dd}} + \mathbf{F}(\mathbf{R_{xx}} + \mathbf{R_{uu}})\mathbf{F}'$ are positive definite by assumption, the Kronecker product above leads to a positive definite matrix [8]. Therefore, the error is a strictly convex function of the post-filter coefficients, and thus the solution in (4.6) is the unique global minimum. ■

With the optimal post-filter in (4.6), the error can be written as a function of $\mathbf{F}$ only, in the form

$$\begin{aligned} \xi &= N^{-1}\,\text{tr}\,\{\mathbf{W}'\mathbf{W}\mathbf{R_{xx}} - \mathbf{W}'\mathbf{W}\mathbf{R_{xx}}\mathbf{F}'[\mathbf{F}(\mathbf{R_{xx}} + \mathbf{R_{uu}})\mathbf{F}' + \mathbf{R_{dd}}]^{-1}\mathbf{F}\mathbf{R_{xx}}\} \\ &= N^{-1}\,\text{tr}\,\{\mathbf{W}'\mathbf{W}[\mathbf{R_{xx}^{-1}} + \mathbf{F}'(\mathbf{F}\mathbf{R_{uu}}\mathbf{F}' + \mathbf{R_{dd}})^{-1}\mathbf{F}]^{-1}\} . \end{aligned} \tag{4.10}$$

If we set $\mathbf{F} = \alpha\mathbf{F_o}$, for any given $\mathbf{F_o}$, the error is a monotonically decreasing function of α. Thus, we need to include a power limitation on the pre-filter output,

as we did in the previous chapters; this is the subject of the next subsection. When there is no channel noise, such a power constraint is not required, and the optimal pre- and post-filter can be obtained from Kazakos's results on optimal estimation on prescribed subspaces [9].

4.1.2. Optimal Pre-filters for a Total Power Constraint

We derive here the optimal pre-filter under the constraint that its output power

$$P = \mathrm{tr}\,\{\mathbf{F}(\mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{uu}})\mathbf{F}'\} \tag{4.11}$$

must be bounded. One way of setting such a bound is to limit the average power per element of the vector $\mathbf{v}$ to be at most equal to one, i.e.,

$$M^{-1}P = M^{-1}\,\mathrm{tr}\,\{\mathbf{F}(\mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{uu}})\mathbf{F}'\} \leq 1\ . \tag{4.12}$$

Our problem is then the minimization of (4.10) under the constraint in (4.12). We could build the corresponding Lagrangian functional and then compute its stationary point(s). However, taking the derivative of the error in (4.10) with respect to $\mathbf{F}$ would lead to a virtually untractable expression. A much easier approach, suggested by the work of of Lee and Petersen [3], is to decompose $\mathbf{F}$ into factors such that the autocorrelation matrices in (4.10) are diagonalized. In this way, we can convert our optimization problem to a form studied in great detail by Başar [10]. Thus, we factor the pre-filter as

$$\mathbf{F} \triangleq \mathbf{HBUWV}\ , \tag{4.13}$$

where $\mathbf{V}$ is the Wiener filter for the input noise,

$$\mathbf{V} = \mathbf{R}_{xx}(\mathbf{R}_{xx} + \mathbf{R}_{uu})^{-1} ,$$

$\mathbf{U}$ is an orthogonal matrix ($\mathbf{U'U} = \mathbf{UU'} = \mathbf{I}$) formed by the eigenvectors of the autocorrelation matrix for the signal $\mathbf{WV}(\mathbf{x}+\mathbf{u})$ (i.e., the Karhunen-Loève transform for $\mathbf{WV}(\mathbf{x}+\mathbf{u})$),

$$\begin{aligned} \Lambda &= \mathbf{UWV}(\mathbf{R}_{xx} + \mathbf{R}_{uu})\mathbf{V'W'U'} \\ &= \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\} , \\ &\quad \lambda_1 \geq \lambda_2 \geq \ldots \lambda_N \geq 0 , \end{aligned} \tag{4.14}$$

and $\mathbf{H}$ diagonalizes the channel noise autocorrelation matrix, that is, $\mathbf{H'H} = \mathbf{HH'} = \mathbf{I}$ and

$$\begin{aligned} \Gamma &= \mathbf{H'R_{dd}H} \\ &= \text{diag}\{\gamma_1, \gamma_2, \ldots, \gamma_M\} , \\ &\quad 0 \leq \gamma_1 \leq \gamma_2 \leq \ldots \gamma_M \quad . \end{aligned} \tag{4.15}$$

The factor $\mathbf{B}$ in (4.13) is then the matrix that is free to be designed. We note that the matrices $\mathbf{V}$, $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{H}$ are non-singular, and so for any $\mathbf{F}$ there is a $\mathbf{B}$ such that the factorization in (4.13) holds. The indicated ordering of the eigenvalues is just a matter of convenience; it avoids the appearance of permutation matrices as factors of the optimal filters.

With $\mathbf{F}$ as in (4.13), the error assumes a simpler form:

$$\begin{aligned} \xi &= N^{-1} \,\text{tr}\,\{\mathbf{W'WR}_{xx} - \mathbf{\Lambda^2 B'}[\mathbf{B\Lambda B'} + \mathbf{\Gamma}]^{-1}\mathbf{B}\} \\ &= \xi_o + N^{-1} \,\text{tr}\,\{[\mathbf{\Lambda}^{-1} + \mathbf{B'\Gamma}^{-1}\mathbf{B}]^{-1}\} , \end{aligned} \tag{4.16}$$

where ξ_o is a constant term defined by

$$\xi_o \triangleq N^{-1} \,\text{tr}\,\{\mathbf{W'WR}_{xx} - \mathbf{\Lambda}\} , \tag{4.17}$$

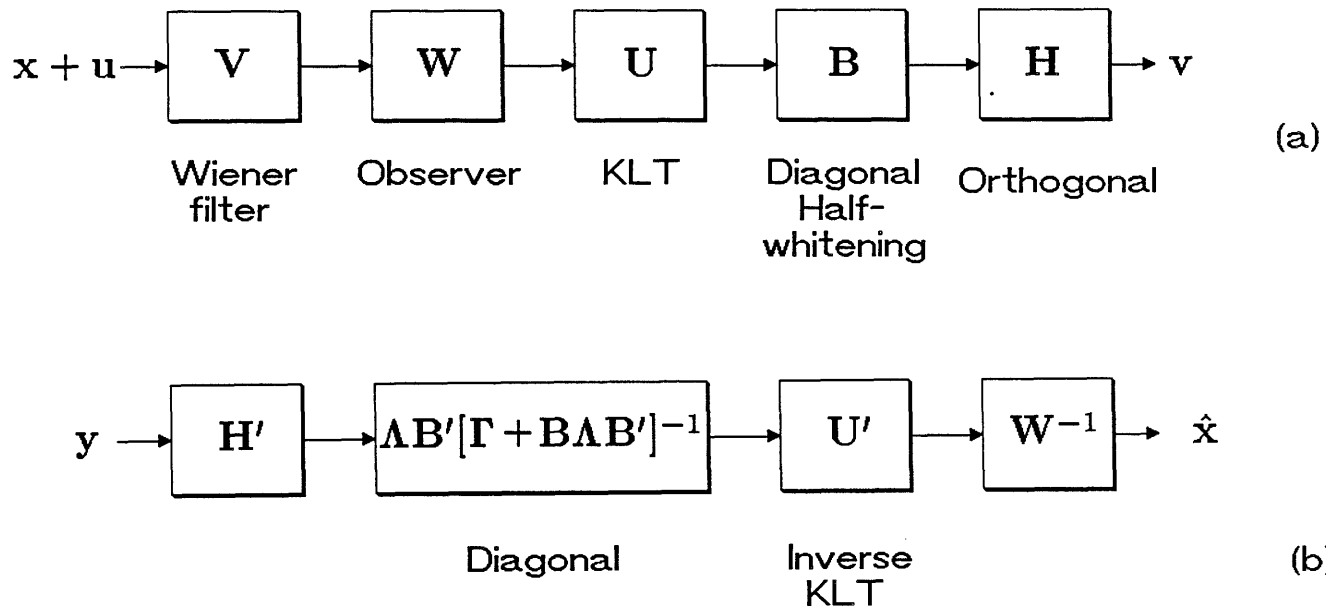


Fig. 4.2. Factors of the optimal (a) pre- and (b) post-filter.

and the power constraint becomes

$$M^{-1}\,\mathrm{tr}\,\{\mathbf{B}\boldsymbol{\Lambda}\mathbf{B}'\} \leq \mathbf{1}\ . \tag{4.18}$$

The optimal post-filter can also be written as a function of the factors of $\mathbf{F}$, as

$$\mathbf{G}_{\mathbf{OPT}} = \mathbf{W}^{-1}\mathbf{U}'\boldsymbol{\Lambda}\mathbf{B}'[\boldsymbol{\Gamma} + \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}']^{-1}\mathbf{H}'\ . \tag{4.19}$$

At this point it is interesting to note that if the channel noise autocorrelation matrix $\mathbf{R_{dd}}$ has repeated eigenvalues, the choice of $\mathbf{H}$ is not unique, but any one leads to exactly the same error. There will exist, then, an infinitude of optimal pre- and post-filter pairs. The factors of the optimal pre- and post-filters are depicted in Fig. 4.2, in which the matrix $\mathbf{B}$ is diagonal. This is a key result that is verified in the following lemma.

Lemma. *If* $\mathbf{F}_{\mathrm{OPT}}$ *is an optimal pre-filter, then its corresponding matrix* $\mathbf{B}_{\mathrm{OPT}}$ *is diagonal in a generalized sense, i.e.,* $b_{ij} = 0$ if $i \neq j, \forall i, j$.

Proof: Let's define a matrix $\mathbf{C}$ by

$$\mathbf{C} \triangleq \mathbf{\Gamma}^{-1/2}\mathbf{B}\mathbf{\Lambda}^{-1/2} \ . \tag{4.20}$$

Substituting (4.20) into (4.16) and (4.18), our optimization problem assumes the form

$$\begin{aligned} \min \ \ \xi &= \xi_o + \frac{1}{N}\,\mathrm{tr}\,\{\mathbf{\Lambda}[\mathbf{I} + \mathbf{C}'\mathbf{C}]^{-1}\} \\ \text{subjectto} \ \ \frac{1}{M}\,\mathrm{tr}\,\{\mathbf{\Gamma}\mathbf{C}\mathbf{C}'\} &\leq \mathbf{1} \ , \end{aligned} \tag{4.21}$$

The above trace minimization problem was analyzed in detail by Başar [10], who proved that if $\mathbf{C_o}$ is optimal then the only non-zero entries in $\mathbf{C_o}$ are at the diagonal positions. But since $\mathbf{B}$ is obtained from $\mathbf{C}$ by pre- and post-multiplication by diagonal matrices, the off-diagonal elements of $\mathbf{B}$ must also be zero. ∎

Since $\mathbf{B}$ must be diagonal, we can remove the effect of its non-diagonal elements on (4.16) and (4.18). Then, our problem reduces to

$$\begin{aligned} \min \ \ \xi &= \xi_o + \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda_i}{1 + b_{ii}^2\lambda_i/\gamma_i} \\ \text{s.t.} \ \ \frac{1}{M}\sum_{i=1}^{M} b_{ii}^2\lambda_i &\leq 1 \ . \end{aligned} \tag{4.22}$$

This is a simple resource allocation problem. We note that the power constraint must be binding, otherwise if we select j such that $\lambda_j > 0$, we can replace b_{jj}^2 by $b_{jj}^2 + \epsilon$, with $\epsilon > 0$, which would lead to a strict error reduction without violation of the power constraint. Thus, we know that an optimal solution must be a stationary

point of the Lagrangian of (4.22), i.e., there exists a nonzero Lagrange multiplier α such that

$$-\frac{\lambda_i^2 b_{ii}/\gamma_i}{(1+b_{ii}^2\lambda_i/\gamma_i)^2}+\alpha b_{ii}\lambda=0\ , \tag{4.23}$$

from which we get

$$b_{ii}^2=\max\left\{0,\sqrt{\frac{\gamma_i}{\alpha\lambda_i}}-\frac{\gamma_i}{\lambda_i}\right\}\ . \tag{4.24}$$

The ordering of the eigenvalues in (4.14) and (4.15) implies that there exists an index i_o such that

$$b_{ii}^2=\begin{cases}\sqrt{\gamma_i/\alpha\lambda_i}-\gamma_i/\lambda_i, & \text{if } i\le i_o\ ,\\ 0, & \text{otherwise.}\end{cases} \tag{4.25}$$

The relationship between i_o and α is

$$\alpha=\left(\frac{\sum_{i=1}^{i_o}\sqrt{\lambda_i\gamma_i}}{M+\sum_{i=1}^{i_o}\gamma_i}\right)^2 \tag{4.26}$$

and

$$\begin{aligned}\sqrt{\frac{\gamma_{i_o}}{\alpha\lambda_{i_o}}}-\frac{\gamma_{i_o}}{\lambda_{i_o}}&\ge 0\ ,\\ \sqrt{\frac{\gamma_{i_o+1}}{\alpha\lambda_{i_o+1}}}-\frac{\gamma_{i_o+1}}{\lambda_{i_o+1}}&<0\ .\end{aligned} \tag{4.27}$$

Thus, we can find i_o as the minimum value of l, with $l<\min(M,N)$ such that

$$\frac{M+\sum_{i=1}^{l}\gamma_i}{\sum_{i=1}^{l}\sqrt{\lambda_i\gamma_i}}<\sqrt{\frac{\gamma_{l+1}}{\lambda_{l+1}}}\ . \tag{4.28}$$

If no such l exists, $i_o = \min(M, N)$.

With the optimal $\mathbf{B}$ the error is given by

$$\xi_{\min} = \xi_o + \frac{1}{N} \sum_{i=i_o+1}^{N} \lambda_i \quad + \frac{1}{N} \frac{\left(\sum_{i=1}^{i_o} \sqrt{\lambda_i \gamma_i} \right)^2}{M + \sum_{i=1}^{i_o} \gamma_i} , \tag{4.29}$$

where the first summation is zero if $i_o \geq N$.

It is interesting to compare (4.24) with the optimal IIR pre-filters in Chapter 2. If we look at the eigenvalues λ_i and γ_i as spectral representations of their corresponding matrices, and i as the equivalent of a frequency index, the analogy between the optimal b_{ii}'s and the optimal IIR pre-filters is immediate. The similarity can be extended through the observation that if the channel noise has small amplitude, i.e., if the γ_i's are small, then b_{ii} is proportional to $\lambda_i^{-1/4}$, so that the matrix $\mathbf{B}$ performs a half-whitening operation on its input signal.

4.1.3. Optimal Pre-filters for Independent Identical Sub-channels

The total power constraint of the previous subsection is reasonable if the channel works by pulse amplitude modulation, or some other technique in which the average power that is physically transmitted is proportional to the square magnitude of the vector $\mathbf{v}$ in Fig. 4.1. In several applications, however, the physical channel is better modeled as a set of M independent sub-channels, e.g., when it is a multi-track tape recorder. In this case the correlation between noises in different tracks is virtually zero, but all tracks have to operate at the same power level.

Thus, we assume that

$$\mathbf{R_{dd}} = \sigma_d^2 \mathbf{I} , \tag{4.30}$$

and replace the power constraint in (4.12) by

$$[\mathbf{HB\Lambda B'H'}]_{ii} = 1 \ , \qquad i = 1, 2, \ldots, M \ , \tag{4.31}$$

where $[\cdot]_{ii}$ denotes the i-th diagonal element. We have kept the pre-filter factorization in (4.13). Since the channel noise autocorrelation matrix is a scaled identity, the matrix $\mathbf{H}$ in (4.15) can be chosen as any orthogonal matrix, with $\gamma_1 = \gamma_2 = \ldots = \gamma_M = \sigma_d^2$.

We note that the minimum error that can be attained under the constraint of unit power for all sub-channels is bounded from below by the minimum error with an average unit power per sub-channel, since any pair of filters satisfying the former constraint also satisfies the latter. If we relax the sub-channel power constraint in (4.31) for a moment and use the average power constraint in (4.18), we could apply the results of the previous subsection to derive the jointly-optimal filters. But since the matrix $\mathbf{H}$ can be chosen as any orthogonal matrix, with no effect on the total pre-filter output power and the error level, a natural question arises: with $\mathbf{B}$ computed under the average power constraint, can $\mathbf{H}$ be chosen so that (4.31) is satisfied? The answer is yes, in view of Horn's conditions for the diagonal entries of a symmetric matrix [11], [12]. Such an $\mathbf{H}$ would certainly lead to an optimal solution, because we hit the error lower bound.

Specifically, we want to find an orthogonal matrix $\mathbf{H}$ such that the symmetric matrix $\mathbf{HB\Lambda B'H'}$ has all of its diagonal entries equal to one, and its eigenvalues equal to the entries of the diagonal matrix $\mathbf{B\Lambda B'}$. Since all prescribed diagonal entries are equal, Horn's conditions [11] for the existence of $\mathbf{H}$ reduce to the trivial requirement that the sum of the prescribed eigenvalues must be equal to the sum of the diagonal entries. This is discussed further in Appendix A.

We conclude, therefore, that the optimal filters for identical independent sub-channels can be obtained by using the same matrix $\mathbf{B}$ as computed in the

previous section. The matrix $\mathbf{H}$ can be designed by means of an efficient algorithm recently developed by Chan and Li [13], which applies $M-1$ plane rotations and permutations to the diagonal matrix until a symmetric matrix with the prescribed diagonal is obtained. The details of the algorithm are explained in Appendix A.

In general, the choice of $\mathbf{H}$ is not unique, and in some cases there is a simple solution: suppose we can find an orthogonal matrix $\mathbf{H}$ such that

$$| h_{ij} | = \sqrt{1/M} \ , \quad \forall i,j \ . \tag{4.32}$$

Then, it is clear that for any diagonal matrix $\mathbf{C}$ it holds

$$[\mathbf{HCH}']_{ii} = \frac{1}{M} \sum_{l=1}^{M} c_{ii} \ . \tag{4.33}$$

If M equals two or a multiple of four, a Hadamard matrix multiplied by $\sqrt{1/M}$ satisfies (4.32). When M is a power of two, a relatively common case, a Hadamard matrix can be easily constructed [14], [15].

Error Improvement

In order to evaluate the error improvement with a jointly-optimal filter pair, as opposed to an optimal post-filter only, let's consider an example similar to the ones studied in previous chapters. The input $\mathbf{x}$ is a gauss-Markov signal with $[\mathbf{R}_{\mathbf{xx}}]_{ij} = \rho^{|i-j|}$, $N = M$, $\mathbf{W} \equiv \mathbf{I}$, and $\mathbf{u} \equiv \mathbf{0}$. In Fig. 4.3 the error improvement is plotted as a function of the signal-to-noise ratio (SNR) of the channel, for $\rho = 0.95$. We note that significant error reductions can be achieved even with low-SNR channels.

In the limit when $N \rightarrow \infty$, any autocorrelation matrix is asymptotically circulant [16], and thus the Karhunen-Loève transforms in Fig. 4.2 reduce to Fourier transforms, with the eigenvalues being samples of the signal power spectrum. Thus, the error improvement for $N \rightarrow \infty$ can be computed from the results in Chapter 2.

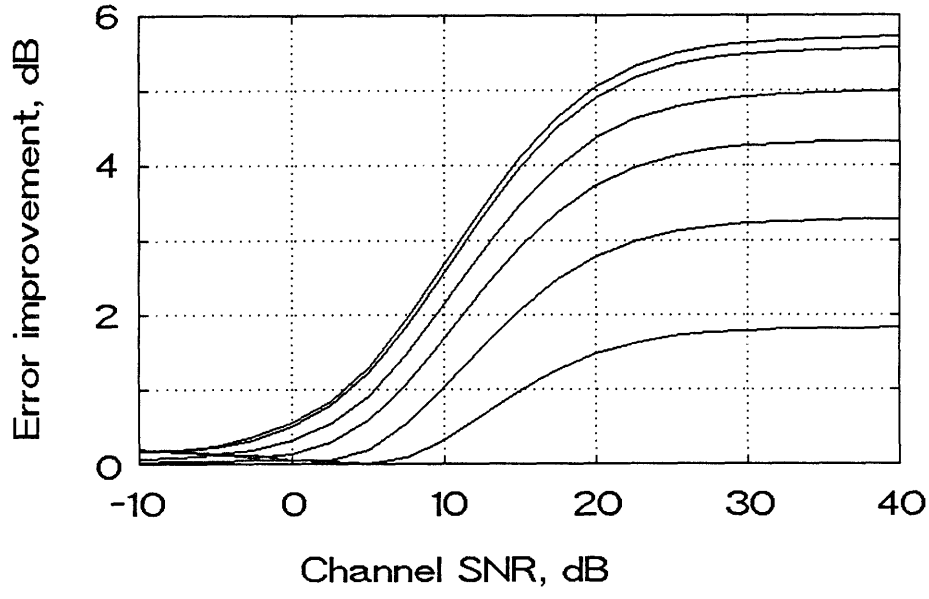


Fig. 4.3. Error improvement due to optimal pre- and post-filtering over optimal post-filtering only, for a first-order Gauss-Markov signal with $\rho = 0.95$. The curves correspond to block sizes of, from top to bottom: 256, 64, 16, 8, 4, and 2.

4.1.4. Sub-optimal solutions

The block diagram in Fig. 4.2 is simply a result of our factorization of the pre-filter matrix $\mathbf{F}$, but it also suggests an approach to the design of sub-optimal filters that leads to fast computation algorithms. The basic idea is to replace the Karhunen-Loève transform (KLT) by the discrete cosine transform (DCT), which can be implemented by fast algorithms. Substitution of KLT's by DCT's is common practice in block signal coding, because the DCT spectra of speech and image signals is close to their KLT spectra [15].

Let's consider the following typical case for a block processing system: assume that $\mathbf{x}$ is a first-order Gauss-Markov process with $[\mathbf{R}_{\mathbf{xx}}]_{ij} = \rho^{|i-j|}$, $\mathbf{u}$ and $\mathbf{d}$ are white noise processes with autocorrelation matrices $\sigma_u^2\mathbf{I}$ and $\sigma_d^2\mathbf{I}$, and $\mathbf{W} = \mathbf{I}$. We

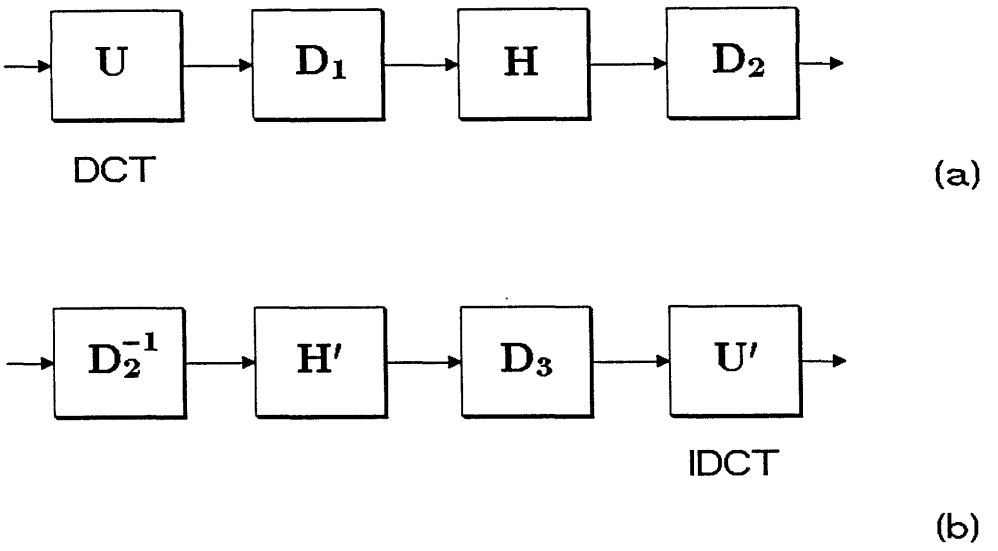


Fig. 4.4. Sub-optimal system.

use the sub-optimal structure of Fig. 4.4, which is derived through the procedure described below.

First, the optimal pre- and post-filters are computed, and the diagonal factors of $\mathbf{F}$ and $\mathbf{G}$ are obtained. If we perform the Wiener filter after the KLT it becomes diagonal (when the input noise is white the Wiener filter in the KLT transform domain is diagonal [17]), lets's call it $\mathbf{V_o}$; we then define the diagonal matrices $\mathbf{D_1} \triangleq \mathbf{V_o B}$ and $\mathbf{D_3} \triangleq \mathbf{\Lambda B}(\sigma_d^2 \mathbf{I} + \mathbf{B\Lambda B}')^{-1}$. Next, we replace the KLT by a DCT and introduce a diagonal matrix $\mathbf{D_2}$ such that the power constraints are satisfied (since they were violated by the replacement of the KLT by a DCT); $\mathbf{D_2^{-1}}$ then becomes a factor of the post-filter.

We note that the matrix $\mathbf{H}$, if obtained by Chan and Li's algorithm, is a cascade of M plane rotations and M permutations, alternately, and so it requires exactly M butterflies to be implemented. The modules that require the largest number of operations in Fig. 4.4 are the direct and inverse DCT's, which can be

performed in $O(N \log N)$ operations [18]. When a non-uniform observer has to be taken into account, i.e., $\mathbf{W} \neq \mathbf{I}$, it is likely that $\mathbf{W}$ can be approximated by a circular convolution, which corresponds to a time-invariant (or space-invariant) observer. In this case, $\mathbf{W}$ is circulant, and multiplication by $\mathbf{W}$ or $\mathbf{W}^{-1}$ can be performed in the Fourier domain by means of diagonal matrices. Thus, the overall system complexity would still be of $O(N \log N)$.

The performance of the sub-optimal system is evaluated in Fig. 4.5 where the SNR gain due to optimal pre- and post-filtering is plotted as a function of the inter-sample correlation coefficient ρ, as well as the increase in error due to the use of the sub-optimal system of Fig. 4.4, for an input noise level of -40 dB and a channel noise of -40 dB. We note that for all values of ρ the sub-optimal filters are effectively as good as the optimal ones. As $\rho \rightarrow 1$, the error gap goes to zero, since the KLT converges to the DCT [19].

In Fig. 4.6 (a) the original "KID" image is degraded by a a white channel noise at a level of -15 dB, which corresponds to a 3.2 % r.m.s. error (this percentage reflects the ratio of the error variance to the signal variance). When an optimal post-filter only is used, with $N = M = 16$, we obtain the image in Fig. 4.6 (b), which has a r.m.s. error of 2.0 %. With the sub-optimal pre- and post-filters of Fig. 4.4, the resultant image is that in Fig. 4.6 (c), which has a r.m.s. error of 1.0 %, i.e., a 6 dB error reduction. In both cases the filters were designed for a first-order Gauss-markov process with $\rho = 0.92$. This value of ρ was estimated from the original image.

The example above illustrates the robustness of the optimal and sub-optimal filters, in the sense that a precise knowledge of the input spectrum is not essential for a good performance. A reasonable model for it usually suffices. It also shows that a pre- and post-filtering system with complexity $O(N \log N)$ can be designed with a good performance, in practice. We note also that the image in Fig. 4.6 (c) is slightly sharper than that in Fig. 4.6 (b). This is because when there is no pre-filter, the high-frequency components of the input signal fall below the noise level, and

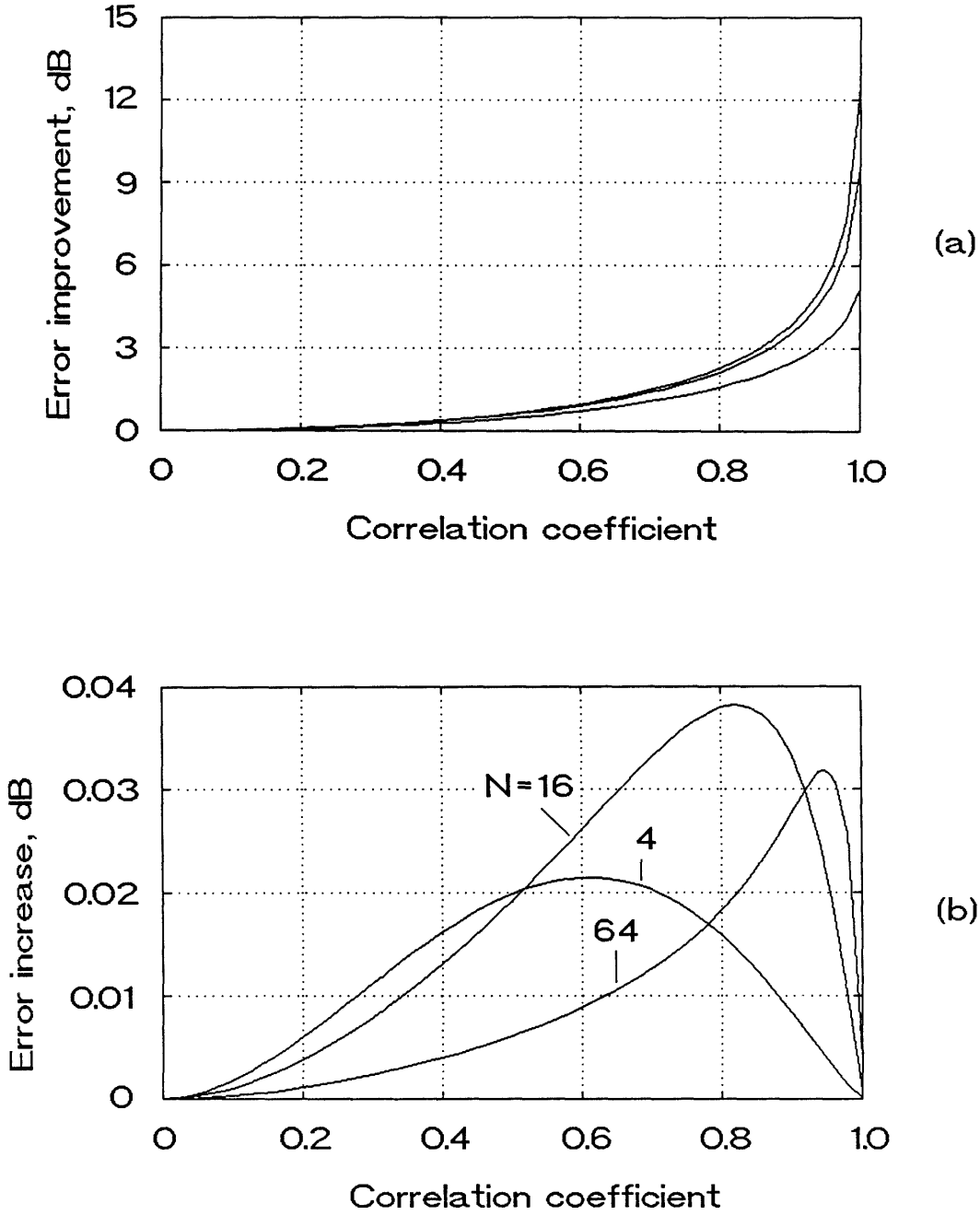


Fig. 4.5. (a) Error improvement due to optimal pre- and post-filtering over optimal post-filtering only, for a first-order Gauss-Markov signal. From top to bottom, the curves correspond to block sizes of 64, 16, and 4. (b) Error increase with the sub-optimal filters of Fig. 4.4.

Fig. 4.6. (a) "KID" image obtained with a white Gaussian channel noise, no pre- or post-filtering, r.m.s. error 3.2%.

Fig. 4.6. (b) "KID" image for the same channel as in Fig. 4.6 (a); no pre-filter, optimal post-filter.

Fig. 4.6. (c) "KID" image for the same channel as in Fig. 4.6 (a); optimal pre- and post-filter.

so the optimal post-filter sets a gain close to zero to those components. The high-boosting effect of the optimal pre-filter alleviates this problem significantly. Finally, we note in Fig. 4.6 (b) and (c) the presence of 'blocking effects', that is, the block boundaries are visible in the reconstructed images. This is a common problem in block signal processing. In Chapter 5 we will study some pre- and post-filtering techniques that help reduce those effects.

4.2. Optimal Filters for Digital Channels

When the channel is digital in a block-processing system, there is an added flexibility: if the M elements of the vector $\mathbf{v}$ in Fig. 4.1 have different variances, it is more efficient to quantize each one with a different number of bits. In this way, the elements with higher energies are represented by a larger number of possible quantization levels. In order to optimize the pre- and post-filters, we have not only to take into account the bit allocation on the channel, but also the fact that quantization noise is correlated with the signal. Unlike the previous chapters, though, we don't have to assume that the input signal and noise are multivariate Gaussian random variables, thanks to a simple but accurate model of the quantization process, which is valid for *any* probability density function.

Without loss of generality, we could set $M = N$ throughout this section, because there is no limit to the number of scalar variables that can be transmitted through the channel; the total number of bits spent on those variables is what counts. However, when the average channel rate (in bits per element) is low, it is likely that the optimal bit assignment will allocate zero bits to several of the v_i. The output of a zero-bit quantizer is always equal to zero, which is the expected value of the variable to be quantized. Thus, the simple additive noise model of Fig. 4.1 does not apply, since the zero-bit elements are not transmitted at all. Therefore, we

shall set M equal to the number of elements that receive a non-zero bit assignment, and so the matrices $\mathbf{B}$ and $\mathbf{C}$ may not be square.

In the next subsection we will derive the optimal matrices in Fig. 4.1, under the assumption that noise $\mathbf{d}$ results from scalar quantization of the elements of $\mathbf{v}$. We further assume that the quantizers are optimal in a mean-square error sense, i.e., that they are Lloyd–Max quantizers [20], [21]. In Subsection 4.2.2 we will consider the use pseudo-random noise (dither) in the quantization process, which actually leads to a reduction in the overall mean-square reconstruction error.

4.2.1. Optimal Filters for Max Quantizers

Our analysis here can be greatly simplified if we make use of the pre- and post-filter factorizations of the previous section. In Fig. 4.7 the system of Fig. 4.1 is partitioned in such a way that the pre- and post-filtering operations are each performed in two steps. The pre-filter generates first the intermediate signal $\mathbf{w}$, which is then transformed by the matrix $\mathbf{B}$ and transmitted through the channel. The post-filter builds an estimate $\hat{\mathbf{w}}$ of $\mathbf{w}$, from which the final input estimate $\hat{\mathbf{x}}$ is generated. The factors $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{U}$ are the same as in Section 4.1.

From Fig. 4.7 it is clear that

$$\mathbf{w} = \mathbf{UWV}(\mathbf{x} + \mathbf{u}) \ , \tag{4.34}$$

and

$$\hat{\mathbf{w}} = \mathbf{UW}\hat{\mathbf{x}}. \tag{4.35}$$

Thus, the absolute mean-square error between $\hat{\mathbf{w}}$ and $\mathbf{w}$ is given by

$$\xi_w \triangleq N^{-1}\,\mathrm{E}\left[\|\hat{\mathbf{w}} - \mathbf{w}\|^2\right] = N^{-1}\,\mathrm{tr}\left\{\mathbf{W}'\mathbf{W}\,\mathrm{E}\left[(\hat{x} - z)(\hat{x} - z)'\right]\right\} \ , \tag{4.36}$$

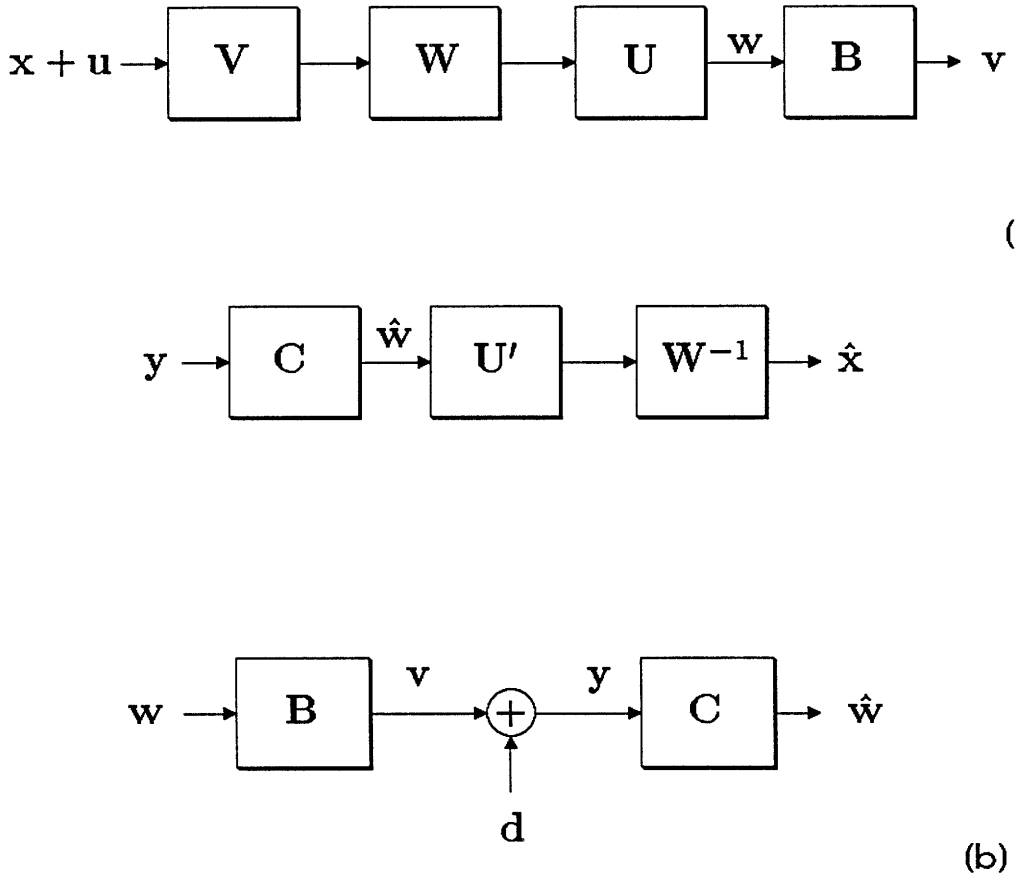


Fig. 4.7. (a) The pre- and post-filtering system, with emphasis on the intermediate signals z and w. (b) The subsystem to be optimized.

where

$$\mathbf{z} \triangleq \mathbf{V}(\mathbf{x}+\mathbf{u}) \ . \tag{4.37}$$

If the matrices **B** and **C** in Fig. 4.7 minimize the absolute m.s. error between $\hat{\mathbf{w}}$ and **w**, then the overall system of Fig. 4.7 (a) leads to a minimum weighted mean-square error between $\hat{\mathbf{x}}$ and **z**. But this is not quite our objective, since we want $\hat{\mathbf{x}}$ to be the best estimate of **x**, not **z**. However, since **V** is the Wiener filter for the input noise **u**, the signal **z** is actually the best estimate of **x**, if all elements of $\mathbf{x}+\mathbf{u}$ are given. Thus, if $\hat{\mathbf{x}}$ is an optimal estimate of **z** it is also an optimal estimate of **x**.

Our problem, therefore, is that of finding the matrices **B** and **C** in Fig. 4.7 (b) that minimize ξ_w. This is a classical problem of information theory, usually referred

to as optimal block quantization or optimal block coding [22]–[24]. Under the assumption that the output of the pre-filter, $\mathbf{v}$, has uncorrelated components, and that $M = N$, Huang and Schultheiss [23] have shown that the optimal post-filter matrix $\mathbf{C}$ must satisfy $\mathbf{C} = \mathbf{B}^{-1}$, if the quantizers are optimal *non-uniform* Max quantizers for Gaussian random variables. This result was also verified by Segall [24], who derived more precise formulas for the optimal bit allocation.

In the following analysis we present an alternative derivation of the optimal $\mathbf{B}$ and $\mathbf{C}$, without the assumptions in [23] and [24], namely that $\mathbf{v}$ has a diagonal autocorrelation, that the quantizers are *non-uniform*, and that the input signal is Gaussian. Our approach will also simplify the derivation of optimal filters for quantizers with dithering, which is presented in the next subsection.

We could make use of the cross-correlation between $\mathbf{v}$ and $\mathbf{d}$ to derive an expression for the error $\xi_{\mathbf{w}}$ as a function of the matrices $\mathbf{B}$ and $\mathbf{C}$. However, this would lead to matrix equations that would be difficult to manipulate. A much easier approach is to use the 'gain plus additive noise' model of scalar quantization. This model, depicted in Fig. 4.8, is derived and justified in Appendix B. The quantizer output, $\mathbf{y}$, is given by

$$\mathbf{y} = \mathbf{\Psi v} + \tilde{\mathbf{d}} \ , \tag{4.38}$$

where $\tilde{\mathbf{d}}$ is a noise source with uncorrelated elements, and $\mathbf{\Psi}$ is a diagonal matrix. The elements of $\mathbf{\Psi}$ and $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ depend on the autocorrelation $\mathbf{R}_{\mathbf{vv}}$, as explained in Appendix B.

With the relationship between $\mathbf{v}$ and $\mathbf{y}$ above, we can rewrite (4.36) in the form

$$\begin{aligned} \xi_w &= N^{-1} \operatorname{tr}\{ \mathrm{E}[(\mathbf{C\Psi Bw} + \mathbf{Cd} - \mathbf{w})(\mathbf{C\Psi Bw} + \mathbf{Cd} - \mathbf{w})']\} \\ &= N^{-1} \operatorname{tr}\{\mathbf{\Lambda} + \mathbf{C\Psi B\Lambda B'\Psi C'} + \mathbf{CR_{dd}C'} - \mathbf{2C\Psi B\Lambda}\} \ . \end{aligned} \tag{4.39}$$

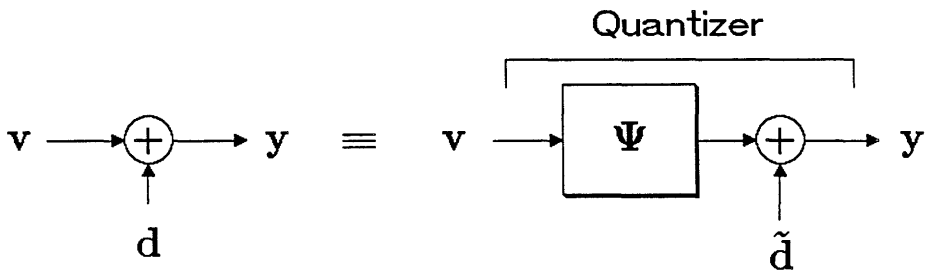


Fig. 4.8. A model for the quantizers. See Appendix B for details.

For any given $\mathbf{B}$, the optimal $\mathbf{C}$ can be obtained by setting $\partial\xi/\partial\mathbf{C} = \mathbf{0}$, which leads to

$$\mathbf{C}_{\mathrm{OPT}} = \mathbf{\Lambda B'\Psi}(\mathbf{B\Lambda B'} + \mathbf{R}_{\mathbf{vd}})(\mathbf{\Psi B\Lambda B'\Psi} + \mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}})^{-1} \ . \tag{4.40}$$

Substituting (4.40) into (4.39), we obtain

$$\begin{aligned} \xi_w &= N^{-1}\,\mathrm{tr}\,\{\mathbf{\Lambda} - \mathbf{\Lambda B'\Psi}(\mathbf{\Psi B\Lambda B'\Psi} + \mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}})^{-1}\mathbf{\Psi B\Lambda}\} \\ &= N^{-1}\,\mathrm{tr}\,\{[\mathbf{\Lambda}^{-1} + \mathbf{B'\Psi R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}^{-1}\mathbf{\Psi B}]^{-1}\} \ . \end{aligned} \tag{4.41}$$

The two equations above are a direct consequence of classical estimation theory [5], since the optimal matrix $\mathbf{C}$ is an optimal post-filter, i.e., an optimal signal estimator.

Although finding the optimal $\mathbf{B}$ in the above equation seems to be a simple task, we must recall from the model in Fig. 4.8 that $\mathbf{\Psi}$ and $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ depend on $\mathbf{B}$. In Appendix C we show that an optimal $\mathbf{B}$ is given by a generalized identity matrix, i.e.,

$$\mathbf{B} = [\ \mathbf{I}\ \mathbf{0}\ ]\ , \tag{4.42}$$

where the number of rows of $\mathbf{B}$, M, is equal to the number of elements that receive a non-zero bit assignment. From (4.19), the optimal post-filter is obtained as

$$\mathbf{C}_{\mathrm{OPT}} = \begin{pmatrix} \mathbf{C_o} \\ 0 \end{pmatrix} , \tag{4.43}$$

where $\mathbf{C_o}$ is a diagonal matrix of order M, with entries

$$\begin{aligned} c_{o_i} &= \frac{\lambda_i \psi_i}{\psi_i^2 \lambda_i + \sigma_{\tilde{d}_i}^2} \\ &= 1 , \\ & \qquad i = 1, \ldots, M . \end{aligned} \tag{4.44}$$

The last equality is easily obtained from the values of ψ_i and $\sigma_{\tilde{d}_i}^2$ derived in Appendix B.

Thus, an optimal system in Fig. 4.7 (b) is obtained by letting the M elements of $\mathbf{w}$ with the largest variances be transmitted directly. The corresponding minimum error ξ_w is

$$\xi_w = \frac{1}{N} \sum_{i=1}^{M} \epsilon_i^2 \lambda_i + \frac{1}{N} \sum_{i=M+1}^{N} \lambda_i , \tag{4.45}$$

where we recall from Appendix B that $\epsilon_i^2 = 1 - \psi_i$.

The remaining problem now is that of the optimal bit assignment among the elements of $\mathbf{v}$. A reasonable model for the ϵ_i's in Appendix B is [15], [23]

$$\epsilon_i = \alpha 2^{-k_i} , \tag{4.46}$$

where k_i is the number of bits assigned to the i-th element of $\mathbf{v}$. The factor α actually depends on k, being equal to one for $k = 0$ and increasing slightly with k.

A good approximation is to set α constant, on the order of $\sim$ 1.3 for non-uniform Max quantization of Gaussian random variables [21], for example.

If we want to quantize the pre-filter output $\mathbf{v}$ at an average rate of K bits per element, we must have $\sum k_i = NK$, or

$$\prod_{i=1}^{M} \epsilon_i = \alpha^M 2^{-NK} \ . \tag{4.47}$$

Minimizing (4.45) under the constraint in (4.47) is the classical bit assignment problem [23],[24], which is in fact a special case of the general resource allocation problem of non-linear optimization [5]. The solution is given in [24],

$$k_i = \max\{0, k_o + \frac{1}{2}\log_2(\lambda_i)\} \ , \tag{4.48}$$

where k_o is chosen so that (4.47) is satisfied. We note that, according to our previous definition, M should be set equal to the number of non-zero k_i's.

The above equation is the so-called "log-variance rule". Segall has derived more accurate bit-assignment equations than (4.48), based on better models for the dependence of ϵ_i^2 on k_i. The error reductions with those more precise bit assignments are virtually negligible, though, given that the k_i's must be rounded to integers, as discussed below.

In practice, the number of bits assigned to the i-th element of $\mathbf{v}$ must be integer, and thus some rounding has to be applied to (4.48). Although integer optimization problems cannot, in general, be solved by rounding the optimal real solutions [25], Segall has reported [24] that the exact integer solutions to the bit assignment problem lead virtually to the same error as those obtained by rounding (4.48).

From the above analysis, we conclude that an optimal pre-filter for the system in Fig. 4.7 process the input signal, in order, by: 1) a Wiener filter $\mathbf{V}$, which minimizes the error due to the input noise, 2) the observer matrix $\mathbf{W}$, which effectively

maps the signal into the observer domain, and 3) a Karhunen-Loève transform **U**, which decorrelates the elements of the signal, prior to quantization. The post-filter is then just a cascade of the inverse KLT and the inverse observer.

The system of Fig. 4.7 (a) is frequently used for block coding of images [26], [27], but generally without the matrices **V** and **W**. Although the optimal pre- and post-filters are not very sensitive to variations in the signal spectrum, as in the systems of the previous chapters, a good knowledge of the input spectrum is essential. The reason is that quantization errors increase rapidly if the quantizers are not matched to the signal variances [15], [19]. In fact, short-space spectral variations within blocks of the same image are generally strong enough to produce significant degradations in system performance. Thus, it is common practice to make the quantizers adaptive, e.g., by estimating the input spectrum for each block and spending a certain fraction of the available bits to code a few parameters that describe the spectral variations.

4.2.2. Quantization with Pseudo-random Noise

When the channel operates at a low bit rate, it is likely that the optimal bit allocation in (4.48) will assign only one or two bits to several elements of **v**. The reconstructed waveform for those elements will have a staircase appearance, which produces visible noise patterns on images [28], and spurious tones in speech processing [29]. A simple idea that virtually eliminates those artifacts is to add a deterministic noise-like waveform to the signal, immediately prior to quantization, and subtract the noise pattern at the receiver [30]. These noise patterns are called 'dither' or 'pseudo-random noise' (PRN) waveforms. They can be obtained, for example, from a congruential pseudo-random number generator [31].

If the PRN waveform is carefully designed, it is possible to achieve the same r.m.s. quantization error as would be obtained without it [15], with the advantage

that the error is then a white noise waveform, which is subjectively much less objectionable [28], [29]. This subjective quality improvement has been the traditional reason for using quantization with PRN in signal coding systems.

There is one important consequence of the application of PRN to the channel quantizers in Fig. 4.7: optimality is lost. We recall from the previous subsection that our derivation of the optimal pre- and post-filters has taken into account the cross-correlation between quantized signal and quantization error. In fact, it is precisely this non-zero cross-correlation that sets the optimal post-filter as the inverse of the pre-filter (except for the Wiener factor $\mathbf{V}$).

Thus, a natural question arises: when PRN is used, can we redesign the pre- and post-filters so that the overall system minimizes the mean-square reconstruction error? The answer is not only yes, but the resultant system is actually better than the one without PRN, i.e., it leads to a lower signal reconstruction error. In what follows we derive the optimal pre- and post-filters for a channel with PRN quantizers. To the best of our knowledge, no similar analysis of a system with PRN quantization exists in the literature.

When PRN is employed, The matrix $\mathbf{\Psi}$ in Fig. 4.8 is equal to the identity matrix, and the autocorrelation $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ remains unchanged, according to Appendix B. The result in Appendix C is still valid, i.e., the optimal $\mathbf{B}$ matrix is given by (4.42). Thus, the optimal post-filter is given by (4.43), with the entries

$$c_{o_i} = \frac{1}{1 + \epsilon_i^2(1 - \epsilon_i^2)} \ . \tag{4.49}$$

The corresponding minimum error is

$$\xi_w = \frac{1}{N} \sum_{i=1}^{M} \frac{\epsilon_i^2 \lambda_i}{1 + \epsilon_i^2} + \frac{1}{N} \sum_{i=M+1}^{N} \lambda_i \ . \tag{4.50}$$

We note that if the ϵ_i^2's are large, i.e., when the average bit rate is low, the error in (4.50) may be significantly lower than that in (4.45), because of the factors

$1+\epsilon_i^2$. These factors also preclude the validity of the log-variance rule, and so the optimal bit assignment must be rederived. Using the relationship between ϵ_i and k_i in (4.46), we seek to minimize (4.50) under the constraint, in (4.47), that the average number of bits per element must be K. This is an instance of the resource allocation problem of non-linear optimization [5], for which an optimal solution must satisfy either

$$\frac{\lambda_i}{(1+\epsilon_i^2)^2} - \eta\frac{q}{\epsilon_i^2} = 0 \,, \tag{4.51}$$

if the corresponding k_i is non-negative, or

$$k_i = 0 \,, \tag{4.52}$$

where η is a Lagrange multiplier and $q \triangleq \alpha^M 2^{-2NK}$. The solution to (4.51) is

$$\epsilon_i^2 = \frac{\lambda_i}{2\eta q} - 1 - \sqrt{\left(1-\frac{\lambda_i}{2\eta q}\right)-1} \,, \tag{4.53}$$

where the Lagrange multiplier η must be adjusted so that (4.47) is satisfied. Combining (4.51)–(4.53), we get the assignment rule

$$k_i = \max\left\{0, \frac{1}{2}\log_2(\alpha) - \frac{1}{2}\log_2\left[\frac{\lambda_i}{2\eta q} - 1 - \sqrt{\left(1-\frac{\lambda_i}{2\eta q}\right)-1}\right]\right\} \,. \tag{4.54}$$

The two basic points about quantization with PRN are: 1) as in the case without PRN, the optimal matrix $\mathbf{B}$ allows the M components of $\mathbf{w}$ with the largest variance to be directly quantized, but the optimal $\mathbf{C}$ now performs a diagonal filtering operation, according to (4.49), and 2) the optimal bit assignment does not lead to a log-variance rule. These differences are minor at high bit rates, where $1+\epsilon_i^2 \simeq 1$.

Performance of Systems with PRN Quantizers

In order to evaluate the potential advantage of using PRN, let's assume that the channel is composed of uniform Max quantizers for Gaussian random variables. In practice, uniform quantizers are frequently used in block coding, since they minimize the entropy of the quantized signal, for a given reconstruction error [32]. The PRN waveform must have a uniform p.d.f. in $(-\Delta/2, \Delta/2)$ [15]. Using the previous analyses, we computed the reconstruction error with and without PRN, for a first-order Gauss-Markov input, and $\mathbf{u} \equiv \mathbf{0}$. The results are shown in Fig. 4.9 and Fig. 4.10.

In Fig. 4.9 (a) we have varied the correlation coefficient ρ from zero to one, in intervals of 0.05, for an average channel rate of one bit per pixel. For easier viewing, we have connected the points with straight lines. The jaggedness in the curves comes from the integer approximation for the optimal bit assignment. The 0 dB reference is the minimum error level that could be attained by any channel, for each given rate, according to the rate-distortion bound [33]. We see that the use of PRN leads to an error improvement of about 1–2 dB, which is significant. At a channel rate of three bits per element, we see in Fig. 4.9 (b) that the use of PRN cannot be strongly justified, since the error improvement is very low.

In Fig. 4.10 (a) we have varied the rate from 0.25 to 5.0 bits per element, in steps of 0.25 bits, for $\rho = 0.6$. Similar curves are presented in Fig. 4.10 (b) for $\rho = 0.9$. In both cases, we see that error reductions on the order of 1.5 dB can be attained at bit rates below two bits per element.

We have processed the "KID" image of Fig. 4.6 (a) with the optimal pre- and post-filters and uniform Max quantizers with and without PRN. We have used a block size of 16×16, and the autocorrelation matrix $\mathbf{R}_{xx}$ was estimated from the 256 blocks of the image. The results are shown in Fig. 4.11 (a) and (b), for a channel rate of one bit per pixel. We note that the use of PRN led to a reduction of 2 dB in the r.m.s. error, and also to some alleviation of the blocking effects. There

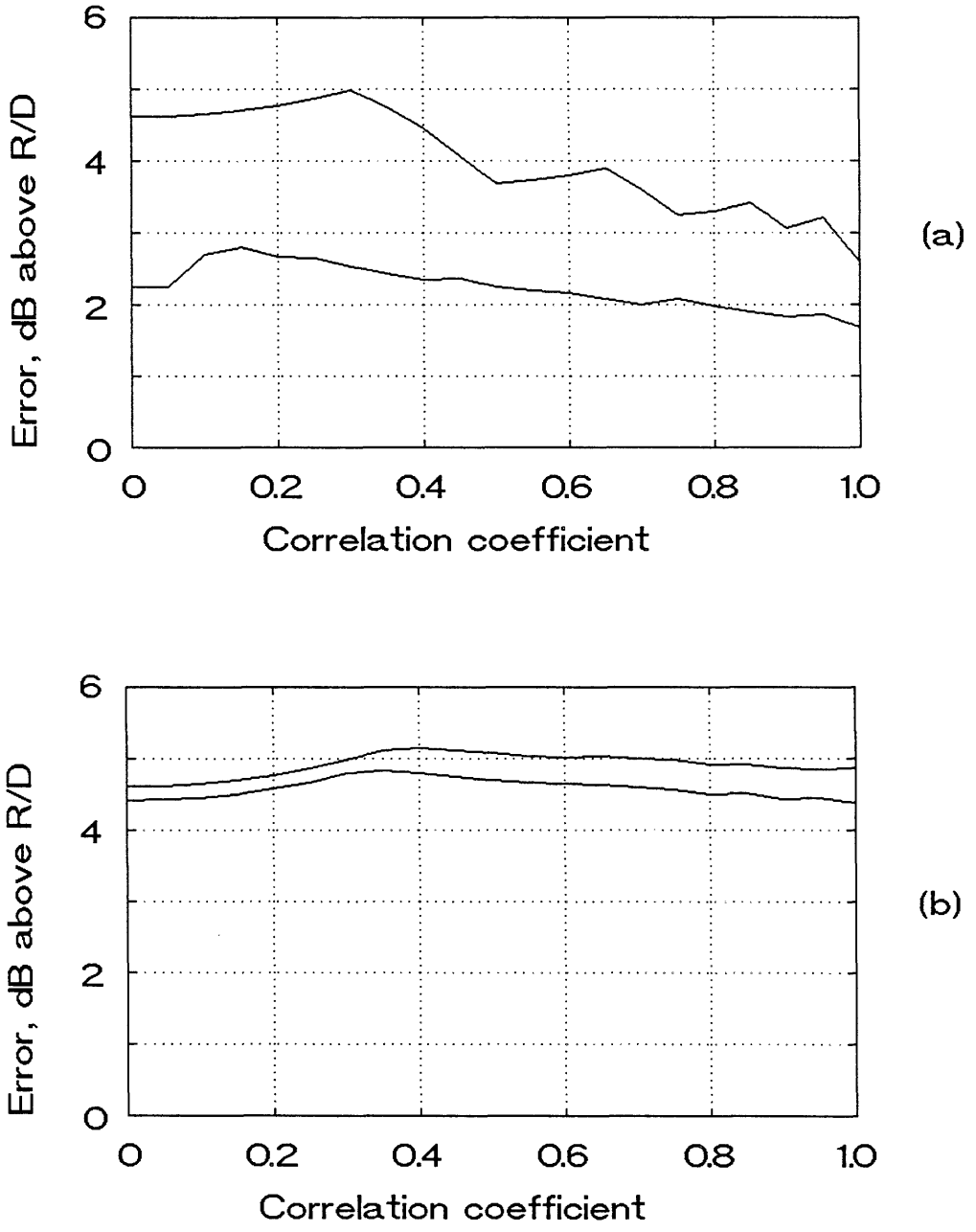
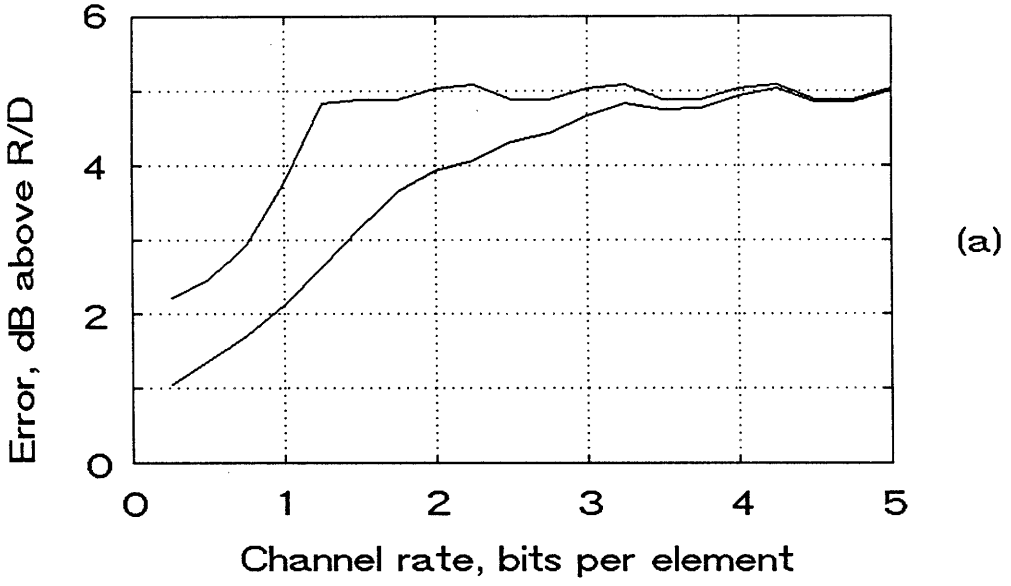


Fig. 4.9. Reconstruction error for a Gauss-Markov signal, with optimal pre- and post-filters and uniform Max quantizers, as a function of the inter-sample correlation coefficient ρ. The channel rates are: (a) one bit per sample, and (b) three bits per sample. In both case, the top and bottom curves correspond to quantization without and with PRN, respectively.



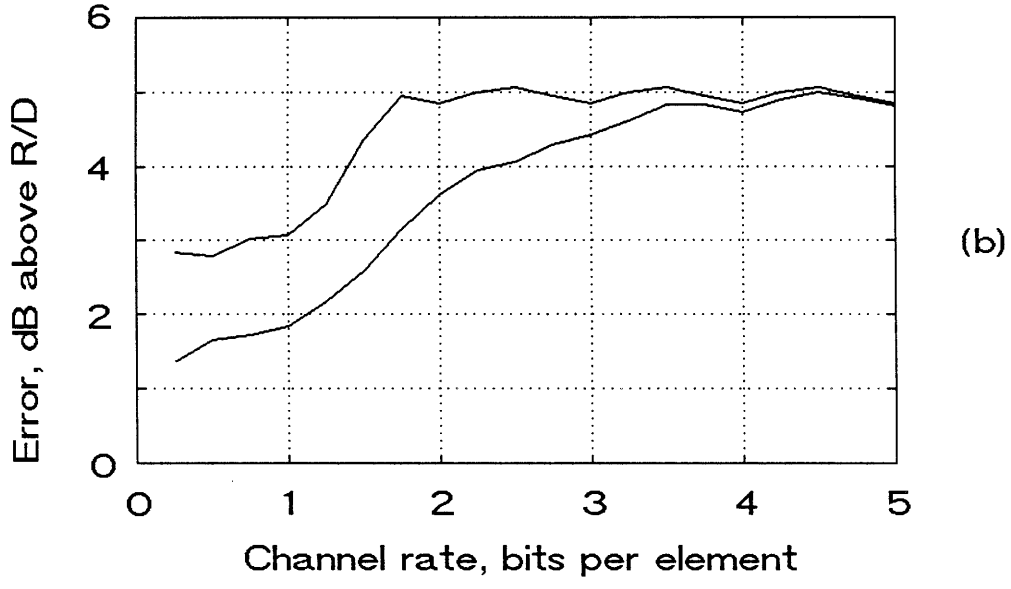


Fig. 4.10. Reconstruction error for a Gauss-Markov signal, with optimal pre- and post-filters and uniform Max quantizers, as a function of the channel rate. The correlation coefficients are: (a) 0.6, and (b) 0.9. In both case, the top and bottom curves correspond to quantization without and with PRN, respectively.

Fig. 4.11. (a) "KID" processed at 1.0 bit per element, uniform quantizers without PRN. R.m.s. error: 15.3 %.

Fig. 4.11. (b) "KID" processed at 1.0 bit per element, uniform quantizers with PRN. R.m.s. error: 12.1 %.

is no difference in sharpness between the two pictures; this is mainly because the zero-bit assignments (which produce low-pass filtering) did not change significantly with the inclusion of PRN.

It is important to observe that no adaptation was employed in the processing that generated the images in Fig. 4.11 (a) and (b), i.e., the same filters and the same bit allocation was used, in each case, for all the blocks of the image. It is unclear, at this point, how much improvement the use of PRN could bring to an adaptive image coding system.

Finally, we note that PRN can also be employed even if the quantizers are non-uniform. In this case, there will be some residual correlation between signal and quantization noise, but some improvement is still potentially attainable.

4.3. Summary

The optimal filters for the system in Fig. 4.1 were derived in this chapter, for channels with and without quantization. By breaking up the filters into basic factors, the algebraic complexity of our analysis was kept low. For an analog channel composed of M identical sub-channels, with M a power of two, the factors of an optimal pre-filter include both a Karhunen-Loève transform (KLT) and a Hadamard transform.

The factorization of the optimal filters has also enabled us to suggest sub-optimal structures that perform within 0.1 dB of the optimal ones. The sub-optimal filters are obtained by replacing the KLT by a discrete cosine transform and adjusting some diagonal factors, so that the pre- and post-filtering operations can be computed with $O(N \log N)$ complexity.

For channels with quantization, we have derived the optimal pre- and post-filters without the assumption that the signals have Gaussian probability distributions. Such an analysis was possible by means of a simple but accurate model of

scalar quantizers. To the best of our knowledge, this is the first derivation of an optimal block coding system without the Gaussian assumption. Basically, we have shown that the optimal reconstruction matrix in a block coding system must be the inverse of the coding matrix, which, in turn, must be a KLT. In practice, it is already known that the use of the KLT leads to good performance even without the Gaussian assumption [34]. We have, therefore, produced a theoretical explanation of why this is so. We have also considered the use of pseudo-random noise in the quantization process, which actually leads to better overall performance.

Further reductions in the reconstruction error for digital channels can be obtained by means of vector quantization (VQ) [34]. When VQ is employed, the cross-correlation between signal and quantization noise is difficult to evaluate and model, and so our previous analysis cannot be applied to systems with VQ. A recent contribution to the design of optimal block coding systems employing VQ can be found in [35], for example.

Applications

One of the main advantages of block signal processing is that the pre- and post-filters can be changed from block to block, thus allowing for a high degree of adaptability. In block quantization, the bit pattern can also be changed from block to block. Full exploitation of this adaptation potential requires the development of good estimators for the signal autocovariance matrix; this is an important practical problem, but it is outside of the scope of this thesis. A clustering approach to this problem can be found in [36].

In block coding systems that use entropy-coded quantization, it is likely that the quantizers are uniform [15]. In this case, we have shown the the use of pseudo-random noise (PRN) leads not only to a complete decorrelation of the quantization error, but also to a *reduction* of the total mean-square error, as long as the filters *and* bit assignment are re-optimized according to Section 4.2.2. At low bit rates

(one bit per sample), the use of PRN can result in improvements in the r.m.s. error in the 0.5–2 dB range.

References

[1] K. Chaloner, "Optimal Bayesian experiment design for linear models," *Ann. Statist.*, vol. 12, pp. 283–300, Mar. 1984.

[2] N. N. Chan, "A-Optimality for regression designs," *J. Math. Anal. Appl.*, vol. 87, pp. 45–50, 1982.

[3] K. -W. Lee and D. P. Petersen, "Optimal linear coding for vector channels," *IEEE Trans. Commun.*, vol. COM-24, pp. 1283–1290, Dec. 1976.

[4] F. C. Schweppe, *Uncertain Dynamic Systems.* Englewood Cliffs, N.J.: Prentice-Hall, 1973, chapter 5.

[5] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: Wiley, 1969, chapters 4 and 7.

[6] M. Athans and F. C. Schweppe, *Gradient matrices and matrix calculations.* Technical Note 1965-53, M.I.T. Lincoln Laboratory, Cambridge, MA, 1965.

[7] A. Graham, *Kronecker Products and Matrix Calculus: with Applications.* Chichester, England: Ellis Horwood, 1981.

[8] J. W. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 772–781, Sept. 1978.

[9] D. Kazakos, "Optimal constrained representation and filtering of signals," *Signal Processing*, vol. 5, pp. 347–353, 1983.

[10] T. Başar, "A trace minimization problem with applications in joint estimation and control under nonclassical information," *J. Optimiz. Theory Appl.*, vol. 31, pp. 343–359, July 1980.

[11] A. Horn, "Doubly-stochastic matrices and the diagonal of a rotation matrix," *Amer. J. Math.*, vol. 76, pp. 620–630, 1954.

[12] S. Friedland, "The reconstruction of a symmetric matrix from the spectral data," *J. Math. Anal. Appl.*, vol. 71, pp. 412–422, 1979.

[13] N. N. Chan and K.-H. Li, "Diagonal elements and eigenvalues of a real symmetric matrix," *J. Math. Anal. Appl.*, vol. 91, pp. 562–566, 1983.

[14] M. Vetterli, "Tree structures for orthogonal transforms and applications to the Hadamard Transform," *Signal Processing*, vol. 5, pp. 473–484, 1983.

[15] N. S. Jayant and P. Noll, *Digital Coding of Waveforms, Principles and Applications to Speech and Video.* Englewood Cliffs, N.J.: Prentice-Hall, 1984, chapters 4 and 12.

[16] P. J. Davis, *Circulant Matrices.* New York: Wiley, 1979, chapter 3.

[17] W. K. Pratt, "Generalized Wiener filter computation techniques," *IEEE Trans. Comput.*, vol. C-21, pp. 636–641, July 1972.

[18] H. S. Malvar, "Fast computation of the discrete cosine transform through fast Hartley transform," *Electron Lett.*, vol. 22, pp. 352–353, Mar. 27, 1986.

[19] R. J. Clarke, "On the relation between the Karhunen-Loève and cosine transforms," *IEE Proc.* **F**, *Commun., Radar, and Signal Processing*, vol. 6, pp. 359–360, 1981.

[20] S. P. Lloyd, "Least-squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129–136, Mar. 1982.

[21] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.

[22] H. P. Kramer and M. V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 41–46, Mar. 1956.

[23] J. J. Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, Sept. 1963.

[24] A. Segall, "Bit allocation and encoding for vector sources," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 162–169, Mar. 1976.

[25] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: Algorithms and Complexity.* Englewood Cliffs, N.J.: Prentice-Hall, 1982, chapter 14.

[26] W. K. Pratt, *Digital Image Processing.* New York: Wiley, 1978, chapters 21–24.

[27] R. J. Clarke, *Transform Coding of Images.* London: Academic Press, 1985, chapters 3 and 4.

[28] D. E. Troxel, "Application of pseudorandom noise to DPCM," *IEEE Trans. Commun.*, vol. COM-29, pp. 1763–1766, Dec. 1981.

[29] J. L. Flanagan, *et al.*, "Speech coding," *IEEE Trans. Commun.*, vol. COM-27, pp. 710–736, Apr. 1979..

[30] L. G. Roberts, "Picture coding using pseudo-random noise," *IEEE Trans. Inform. Theory*, vol. IT-8, pp. 145–154, Feb. 1962.

[31] G. Dahlquist and A. Björk, *Numerical Methods.* Englewood Cliffs, N.J.: Prentice-Hall, 1974, chapter 11.

[32] R. C. Wood, "On optimum quantization," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 248–252, Mar. 1969.

[33] B. J. Bunin, "Rate-distortion functions for Gaussian Markov processes," *Bell Syst. Tech. J.*, vol. 48, pp. 3059–3074, Nov. 1969.

[34] J. Makhoul *et al*, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, pp. 1551–1588, Nov. 1985.

[35] B. Mazor and W. A. Pearlman, "An optimal transform trellis code with applications to speech," *IEEE Trans. Commun.*, vol. COM-33, pp. 1109–1116, Oct. 1985.

[36] A. M. Kirkland, *Image block classification for adaptive transform coding.* S. M. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, Sept.1985.

Chapter 5

Block Processing with Overlapping Functions

The optimal system of the previous chapter can easily be made adaptive, since the signal is broken into blocks. If a local estimate of the block autocorrelation matrix can be obtained, the optimal filters and quantizers can be adjusted for each block. A sophisticated system could even take into account, for example, variations in the observer response with the input signal. Therefore, the block-processing structure of Chapter 4 can form the basis of a very efficient signal coding system. Nevertheless, such adaptability, if fully exploited, may also lead to stronger blocking effects.

It is easy to conceive a situation in which one block of an image contains part of a human face, and an immediate neighbor block is part of a background that contains mostly low-frequency components. In this case a fully adaptive system might process the two neighboring blocks quite differently, and that could potentially produce accentuated discontinuities at the boundaries. With the use of pseudo-random noise in the quantization process and its associated optimal filter, the blocking effects may become the most visible degradation in the reconstructed image, as we have verified in the previous chapter.

Our objective here is to develop an optimal set of pre- and post-filters for block processing with virtually no blocking effects. Sub-optimal implementations that lead to fast algorithms, with increases in the error level of less than 0.1 dB will also be derived. Recently, a few basic approaches have been suggested for the reduction of blocking effects. Reeve and Lim [1] have introduced two techniques: one in which the blocks overlap by a few samples, and the other based on adaptive

post-filtering of the reconstructed signal. Both techniques are effective in reducing the blocking effect, but at the expense of a slightly higher bit rate in the first method and a slightly noticeable blurring across the block boundaries in the second.

Hinman, Bernstein, and Staelin [2] presented a multidimensional extension of the short-time Fourier transform [3], which was referred to as the "Short-space Fourier Transform" (SSFT). Although inherently free of blocking effects, the SSFT introduces strong ringing problems due to the infinite extent of its basis functions. Specifically, whenever a coefficient is omitted in an SSFT representation, the reconstructed signal contains long-duration ripples that are similar to the Gibbs phenomenon in truncated Fourier series representations.

Cassereau [4] has introduced a new concept in which the reconstructed signal is obtained as a weighted combination of basis vectors of length greater than the block size. Those basis functions referred to collectively as the "Lapped Orthogonal Transform" (LOT), decay slowly towards zero after the block boundaries, in such a way that a smooth transition is obtained in the reconstructed signal from one block to another. He has successfully applied his new transforms to block image coding, with a significant reduction in the blocking effects. Cassereau's basis functions were derived numerically, by means of a recursive procedure in which a series of non-linear optimization problems were solved in order to obtain a new basis function (we will use the terms basis function and basis vector interchangeably, as it is common practice in the block coding literature). This approach unfortunately is prone to error propagation, which may lead to sub-optimal solutions for the high-order functions, since an error in basis number r produces an inexact formulation of the optimization problem number $r + 1$. Furthermore, very little can be said about the underlying structure of the basis vectors, and the development of fast algorithms for the LOT becomes virtually impossible.

In this chapter we present an analytical derivation of an optimal LOT, in which all the basis functions of the LOT are obtained as the eigenvectors of a modified autocorrelation matrix (as in the previous chapters, we assume that the signals

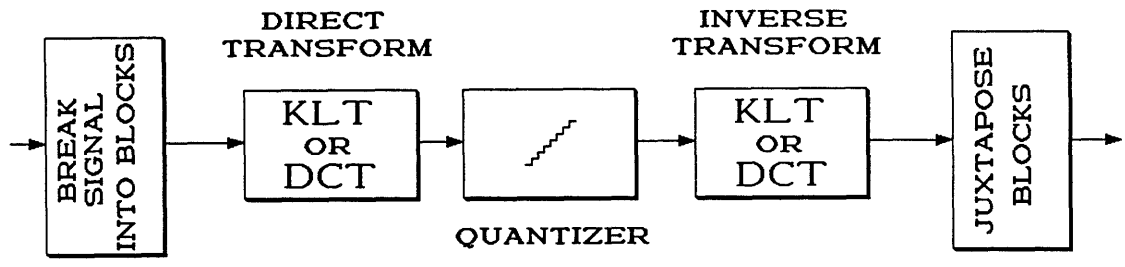


Fig. 5.1. A block coding system.

of interest are all stationary and zero-mean, so that the terms autocorrelation and autocovariance can be interchangeably used). The analysis immediately suggests a simple approximation to the optimal functions that leads to a fast LOT implementation, so that the direct and inverse LOT transformations of a signal can be performed in $O(N \log N)$ operations.

5.1. Basic Properties of Lapped Orthogonal Transforms

We recall in Fig. 5.1 the optimal block coding system of the previous chapter, under the assumption that the input signal is noiseless and the error weighting **W** is the identity matrix. We have included explicitly in Fig. 5.1 the operations of breaking the input vector into blocks at the transmitter, and concatenation of the reconstructed blocks at the receiver. As we have seen in the previous chapter, the Karhunen-Loève transform (KLT) can be replaced by the discrete cosine transform

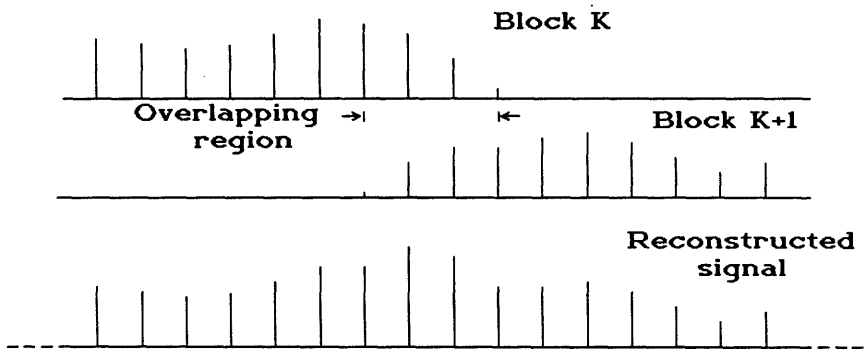


Fig. 5.2. Signal reconstruction with overlapping blocks.

(DCT) with virtually no loss of optimality, and with the advantage of the fast computability of the DCT.

We want to study here the LOT class of transforms, for which the first operator in Fig. 5.1 decomposes the input signal into *overlapping* blocks of length L, with $L > N$, where N is the transform size, as in the previous chapter. At the receiver the reconstructed blocks are superimposed, with an overlapping region of $L - N$ samples between adjacent blocks. This is illustrated in Fig. 5.2.

There are several basic properties that must be satisfied by the set of N basis vectors used to represent the signal for each block. First, in order that the representation be unique for any incoming block, the N basis vectors that are the columns of the inverse transform matrix must be orthogonal (by definition of a basis). Second, the representation must also be independent of the samples of the neighboring blocks that fall into the overlapping areas. Therefore, the basis vectors must also

be orthogonal to the tails of the vectors from the neighboring blocks. Calling $\mathbf{P}$ the inverse transform matrix (which is the optimal post-filter $\mathbf{U}'$, in the notation of Chapter 4), we must have

$$\mathbf{P}'\mathbf{P} = \mathbf{I} \tag{5.1}$$

and

$$\mathbf{P}'\mathbf{W}\mathbf{P} = \mathbf{P}'\mathbf{W}'\mathbf{P} = \mathbf{0} \ , \tag{5.2}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{W}$ is defined by

$$\mathbf{W} \triangleq \begin{pmatrix} \mathbf{0} & I \\ 0 & 0 \end{pmatrix} \ , \tag{5.3}$$

where the identity matrix is of order $L-N$. If $L = N$ there is no overlap, and (5.2) is trivially satisfied.

We note that, because we have fewer basis functions per block than the length of the block, each block cannot be reconstructed exactly, even for a noiseless channel. However, the subspace that is orthogonal to that spanned by the basis functions of a block is exactly that spanned by the combined basis functions of the neighboring blocks. Thus, each signal block can be represented exactly through an LOT, *after* the superposition of the blocks. This aspect is further discussed in [4].

Besides the required orthogonality conditions above, we should expect additional properties to hold for a good LOT matrix $\mathbf{P}$, based on our knowledge of the DCT and KLT. We recall from the previous chapter that a good model for the input signal statistics, in the case of a raster-scanned picture, is the simple first-order Gauss-Markov process. The autocorrelation matrix for such a process has the form

$$\mathbf{R}_{xx} = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^L \\ \rho & 1 & \rho & \cdots & \rho^{L-1} \\ \vdots & & \ddots & & \vdots \\ \rho^{L-1} & & \rho & 1 & \rho \\ \rho^L & \cdots & \rho^2 & \rho & 1 \end{pmatrix} \ , \tag{5.4}$$

where ρ is the inter-sample correlation coefficient. Since the above matrix is symmetric and Toeplitz, its eigenvectors (which compose the KLT) are either symmetric or antisymmetric vectors [6], [7], i.e.,

$$\mathbf{R}_{xx}\mathbf{y} = \lambda\mathbf{y} \quad \Rightarrow \quad \mathbf{J}\mathbf{y} = \mathbf{y} \quad \text{or} \quad \mathbf{J}\mathbf{y} = -\mathbf{y} \ , \tag{5.5}$$

where $\mathbf{J}$ is the 'counter-identity'

$$\mathbf{J} \triangleq \begin{pmatrix} 0 & \cdots & & 0 & 1 \\ 0 & \cdots & 0 & 1 & 0 \\ \vdots & & & & \vdots \\ 1 & 0 & & \cdots & 0 \end{pmatrix} . \tag{5.6}$$

We note that $\mathbf{J}$ is a permutation matrix, and thus it is orthogonal. It is easy to show that half of the eigenvalues of $\mathbf{J}$ are equal to -1 and the other half equal to 1. It also turns out that half of the eigenvectors of $\mathbf{R}_{xx}$ are symmetric, i.e., $\mathbf{J}\mathbf{y} = \mathbf{y}$, and the other half are antisymmetric, $\mathbf{J}\mathbf{y} = -\mathbf{y}$ [7]. It is reasonable to expect that the LOT should also have this kind of symmetry, i.e., it should be formed by $N/2$ symmetric (or even) vectors and $N/2$ antisymmetric (or odd) vectors. The DCT functions have this even-odd symmetry.

Another important property of a good set of basis vectors is smoothness, i.e., the vectors should be approximately sampled sinusoids, so that sharp variations are avoided. The eigenvectors of $\mathbf{R}_{xx}$ in (5.4) are exactly sampled sinusoids [5], for any value of ρ, as well as the DCT basis functions.

With basis on the discussion above, we can safely assume that half of the basis functions that compose the inverse transform matrix $\mathbf{P}$ are even, and the other half odd. So, we can write $\mathbf{P}$ as

$$\mathbf{P} = [\ \mathbf{P}_e\ \mathbf{P}_o\ ] \ , \tag{5.7}$$

where $\mathbf{P_e}$ and $\mathbf{P_o}$ are $L \times N/2$ matrices whose columns contain the even and odd vectors of $\mathbf{P}$, respectively, that is

$$\begin{aligned} \mathbf{JP_e} &= \mathbf{P_e} \ , \\ \mathbf{JP_o} &= -\mathbf{P_o} \ . \end{aligned} \tag{5.8}$$

We can also decompose $\mathbf{P}$ as

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} \\ \mathbf{JQK} \end{pmatrix} \ , \tag{5.9}$$

where $\mathbf{K}$ is a $N \times N$ matrix defined by

$$\mathbf{K} \triangleq \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \ , \tag{5.10}$$

and $\mathbf{Q}$ is a $L/2 \times N$ matrix that is partitioned as

$$\mathbf{Q} \triangleq [\ \mathbf{Q_e}\ \mathbf{Q_o}\ ] \ . \tag{5.11}$$

Here $\mathbf{Q_e}$ and $\mathbf{Q_o}$ are not composed of even and odd vectors. The subscripts are just to remind that $\mathbf{P_e}$ is obtained from $\mathbf{Q_e}$ and $\mathbf{P_o}$ from $\mathbf{Q_o}$. Combining (5.7) and (5.11), we obtain

$$\mathbf{P_e} = \begin{pmatrix} \mathbf{Q_e} \\ \mathbf{JQ_e} \end{pmatrix} \tag{5.12}$$

and

$$\mathbf{P_o} = \begin{pmatrix} \mathbf{Q_o} \\ -\mathbf{JQ_o} \end{pmatrix} \ . \tag{5.13}$$

The necessary orthogonality conditions can be written in terms of the column vectors of $\mathbf{Q}$. Using the column indexing

$$\mathbf{Q} = [\ \mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_N\ ] \ , \tag{5.14}$$

or

$$\begin{aligned} \mathbf{Q_e} &= [\ \mathbf{q}_1\ \mathbf{q}_2\ \cdots\ \mathbf{q}_{N/2}\ ]\ , \\ \mathbf{Q_o} &= [\ \mathbf{q}_{N/2+1}\ \mathbf{q}_{N/2+2}\ \cdots\ \mathbf{q}_N\ ]\ , \end{aligned} \tag{5.15}$$

we can write the orthogonality conditions as

$$\begin{aligned} &\mathbf{q}_i'\mathbf{q}_i = 1/2, \qquad i = 1,2,\ldots,N, \\ &\mathbf{q}_i'\mathbf{q}_j = 0, \qquad i \le N/2 \text{ and } j \le N/2,\ i \ne j, \\ &\qquad\qquad \text{or } \ i > N/2 \text{ and } j > N/2,\ i \ne j, \\ &\mathbf{q}_i'\mathbf{J}\mathbf{q}_j = 0, \qquad i,j = 1,2,\ldots,N. \end{aligned} \tag{5.16}$$

For $N > 4$, there are fewer constraints in (5.16) than the number of unknowns, so that an infinite number of feasible solutions exist. In the next section we will derive an optimal choice for the $\mathbf{q}_i$'s.

5.2. An Optimal LOT

A fundamental property of LOT's is that if $\mathbf{P}$ is a valid LOT matrix, i.e., if it satisfies (5.1) and (5.2), then the matrix

$$\mathbf{P}_2 = \mathbf{P}\mathbf{Z} \tag{5.17}$$

is also a valid LOT, for any orthogonal $\mathbf{Z}$, since

$$\begin{aligned} \mathbf{P}_2'\mathbf{P}_2 &= \mathbf{Z}'\mathbf{P}'\mathbf{P}\mathbf{Z} = \mathbf{Z}'\mathbf{Z} = \mathbf{I}\ , \\ \mathbf{P}_2'\mathbf{W}\mathbf{P}_2 &= \mathbf{Z}'\mathbf{P}'\mathbf{W}\mathbf{P}\mathbf{Z} = \mathbf{0}\ . \end{aligned} \tag{5.18}$$

So, given any performance measure, one way to derive an optimal LOT is to start with a feasible $\mathbf{P}$ and seek a $\mathbf{Z}$ in (5.17) such that the resulting $\mathbf{P}_2$ maximizes

the performance. As we have seen for the system model for digital channels in Chapter 4, the optimal transform, the KLT, leads to a minimum coding error because it maximizes the spread of the variances of the transformed coefficients. Thus, calling $\mathbf{v}$ the transformed vector resulting from applying block $\mathbf{P}_2$ to the input block $\mathbf{x}$, we have

$$\mathbf{v} = \mathbf{P}_2'\mathbf{x} \ . \tag{5.19}$$

According to Chapter 4, the total r.m.s. error in a block coding system is minimized by the use of the Karhunen-Loève transform (KLT), because the KLT produces a maximum spread of the variance of the transform coefficients. Thus, in order to optimize the LOT we should seek the maximum of the 'energy compaction' measure, which is defined as the ratio of the arithmetic to the geometric mean of the variances of the elements of $\mathbf{v}$, that is,

$$\gamma \triangleq \frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_{v_i}^2}{\left(\prod_{i=1}^{N}\sigma_{v_i}^2\right)^{1/N}} \ , \tag{5.20}$$

where the autocorrelation matrix for the transformed vector $\mathbf{v}$ is given by

$$\mathbf{R}_{\mathbf{vv}} = \mathbf{P}_2'\mathbf{R}_{\mathbf{xx}}\mathbf{P}_2 \ . \tag{5.21}$$

The variances $\sigma_{v_i}^2$ are the diagonal entries of $\mathbf{R}_{\mathbf{vv}}$. We can express $\mathbf{R}_{\mathbf{vv}}$ as a function of $\mathbf{Z}$ by substituting (5.17) into (5.21), with the result

$$\mathbf{R}_{\mathbf{vv}} = \mathbf{Z}'\mathbf{R}_o\mathbf{Z} \ , \tag{5.22}$$

where

$$\mathbf{R}_o \triangleq \mathbf{P}'\mathbf{R}_{\mathbf{xx}}\mathbf{P} \ . \tag{5.23}$$

For any given $\mathbf{P}$, $\mathbf{R_o}$ is a valid autocorrelation matrix. The optimal choice for $\mathbf{Z}$ is then, according to Chapter 4, the KLT corresponding to $\mathbf{R_o}$, i.e., the matrix formed by the orthogonal eigenvectors of $\mathbf{R_o}$. Thus, we have an exact solution for an optimal LOT.

It is important to point out that our optimization approach leads to an optimal LOT that is tied to the choice of the initial matrix $\mathbf{P}$. For any such a choice, we have a matrix $\mathbf{P}$ whose N columns are orthogonal, by definition. Since each column of $\mathbf{P}$ has L elements, with $L > N$, they span an N-dimensional subspace of $\mathbb{R}^L$. For any $\mathbf{Z}$, the matrix $\mathbf{PZ}$ will always belong to that subspace, and so will the optimal LOT. However, if $\mathbf{Z}$ is not orthogonal then $\mathbf{PZ}$ is not a valid LOT. On the other hand, there may exist an LOT $\hat{\mathbf{P}}$ that does not belong to the subspace spanned by the columns of $\mathbf{P}$. Therefore, the set of feasible LOT's in $\mathbb{R}^L$ does not form a subspace of dimensionality N.

Thus, an optimal LOT derived by the procedure above may not be the globally optimal LOT, in the sense of maximizing the energy compaction γ. However, as we will see later, our choice for $\mathbf{P}$ suggested in what follows is good enough, since we obtain the same energy compaction as Cassereau's functions, which are aimed to be globally optimal (actually, we have obtained slightly higher γ's than Cassereau; this is probably due to some error propagation in Cassereau's algorithm).

We will assume from now on that $L = 2M$, so that the length of the basis functions is twice the block size, and there is a 50 percent overlap area between blocks. The reasons for this choice will become clear later. We have to make now an initial choice for $\mathbf{Q}$ in (5.9), and then seek a value for $\mathbf{Z}$ in (5.17) such that γ is maximized. Without loss of generality, we can write $\mathbf{Q}$ in the form

$$\mathbf{Q} = \mathbf{DA} \ , \tag{5.24}$$

where $\mathbf{D}$ is the matrix whose columns are the DCT basis functions of length N. The entries of $\mathbf{D}$ are given by

$$d_{ij} = c(j)\cos\left((2i-1)r_j\frac{\pi}{2N}\right), \quad i,j = 1,2,\ldots,...,N, \tag{5.25}$$

where

$$c(j) = \begin{cases} 1/\sqrt{N}\,, & j = 1 \\ \sqrt{2/N}\,, & j > 1 \end{cases} \tag{5.26}$$

and

$$r_j = \begin{cases} 2(j-1)\,, & j \le N/2 \\ 2(j-N/2)-1\,, & j > N/2\,. \end{cases} \tag{5.27}$$

The index r_j is chosen so that the first $N/2$ columns of $\mathbf{D}$ are even vectors, and the last $N/2$ columns are odd. Calling the matrices formed by the even and odd vectors $\mathbf{D_e}$ and $\mathbf{D_o}$, respectively, we can write

$$\mathbf{D} = [\ \mathbf{D_e}\ \ \mathbf{D_o}\ ]\,. \tag{5.28}$$

The reason for the above factorization is that we would like, as stated before, to obtain smooth functions for the columns of $\mathbf{P}$, which implies smooth functions for the columns of $\mathbf{Q}$. Since the columns of $\mathbf{D}$ are sampled sinusoids, we can generate smooth functions for $\mathbf{Q}$ if we try to keep as few as possible non-zero entries on each column of $\mathbf{A}$. In order to determine feasible choices for $\mathbf{A}$, we must rewrite (5.16) in terms of the columns of $\mathbf{A}$, with the result

$$\begin{aligned} &\mathbf{a}_i'\mathbf{a}_i = 1/2, \qquad i = 1,2,\ldots,N, \\ &\mathbf{a}_i'\mathbf{a}_j = 0, \qquad i \le N/2 \text{ and } j \le N/2,\ i \ne j, \\ &\qquad\qquad \text{or } \ i > N/2 \text{ and } j > N/2,\ i \ne j, \\ &\mathbf{a}_i'\mathbf{K}\mathbf{a}_j = 0, \qquad i,j = 1,2,\ldots,N. \end{aligned} \tag{5.29}$$

The last equality above comes from the fact that $\mathbf{D'JD} = \mathbf{K}$. We can also rewrite (5.29) in terms of summations

$$\begin{aligned} \sum_{r=1}^{N} a_{ir}^2 &= 1/2, \quad i = 1,2,\ldots,N, \\ \sum_{r=1}^{N} a_{ir}a_{jr} &= 0, \qquad i \le N/2 \text{ and } j \le N/2,\ i \ne j, \\ &\text{or } i > N/2 \text{ and} \\ \sum_{r=1}^{N/2} a_{ir}a_{jr} - \sum_{r=N/2+1}^{N} a_{ir}a_{jr} &= 0, \qquad i,j = 1,2,\ldots,N. \end{aligned} \tag{5.30}$$

It would be desirable, on the basis of our smoothness arguments, to choose the columns of $\mathbf{Q}$ to be sampled sinusoids. It is clear, however, that this is not possible, since it would require that each column of $\mathbf{A}$ had only a single non-zero entry, and that would violate the third equation in (5.30). The next logical choice would then be to synthesize each column of $\mathbf{Q}$ with only two frequencies. This turns out to be a feasible alternative, and we can choose $\mathbf{A}$ as

$$\mathbf{A} = \frac{1}{2}\begin{pmatrix} \mathbf{I} & \mathbf{I} \\ -\mathbf{I} & -\mathbf{I} \end{pmatrix}. \tag{5.31}$$

With $\mathbf{A}$ as above we obtain a feasible LOT $\mathbf{P}$ as

$$\mathbf{P} = \frac{1}{2}\begin{pmatrix} \mathbf{D_e} - \mathbf{D_o} & \mathbf{D_e} - \mathbf{D_o} \\ \mathbf{J}(\mathbf{D_e} - \mathbf{D_o}) & -\mathbf{J}(\mathbf{D_e} - \mathbf{D_o}) \end{pmatrix}. \tag{5.32}$$

The matrix $\mathbf{R_o}$ can then be computed from (5.23) and the optimal $\mathbf{Z}$ obtained as the eigenvectors of $\mathbf{R_o}$. The resulting basis vectors, which are the columns of $\mathbf{PZ}$, are shown in Fig. 5.3, for $\rho = 0.95$. The functions are not much sensitive to variations in ρ, so that the results for $\rho = 0.8$, for example, are virtually the same as those in Fig. 5.3. These functions are very similar to those obtained by

Cassereau [4], except that in [4] the basis functions are obtained as the solution of a non-linear optimization problem, which is prone to local maxima. The higher order functions in [4] are significantly different from those in Fig. 5.3. In fact, the performance factor for Cassereau's functions was $\gamma = 9.37$, whereas the functions in Fig. 5.3 lead to $\gamma = 9.49$. It seems that the algorithm in [4] may have converged to local minima when computing the high-order functions.

We note in Fig. 5.3 that the orthogonality constraints for the functions belonging to neighboring blocks lead to basis functions that decay towards zero at their boundaries. The first basis function, for example, has a boundary value that is 5.83 times lower than its value at the center. So, the discontinuity from zero to the boundary value is much lower than that of the standard DCT functions.

There are two basic properties of the LOT functions in Fig. 5.3 that are a dirct consequence of the choice $L = 2M$. First, if the lower order basis functions for a group of consecutive blocks are superimposed, the resultant sequence has a constant d.c. value, except for the first and last blocks. This is one important and desirable characteristic, since it implies that a flat field can be reproduced with only one transform coefficient per block. In an adaptive coding system, this means that more bits can be allocated to blocks that contain more image detail. Second, the fact that the right boundaries of the basis functions for block r are immediately adjacent to the left boundaries of the functions for block $r + 2$. If L were smaller than $2M$, there would be two different positions in block $r + 1$ were discontinuities from block r and block $r + 2$ might occur. It seems, therefore, that $L = 2M$ is a good choice for the length of the functions.

The factor $\mathbf{Z}$ of the optimal LOT matrix $\mathbf{PZ}$ is formed by the eigenvectors of $\mathbf{R_o}$, and so it may not be factorable in $N \log(N)$ butterfly stages. This is exactly the same deficiency of the optimal KLT for block coding without overlapping. In the next section we discuss an approximation to the optimal LOT that can be implemented through a fast algorithm, just as the DCT is a good approximation to the KLT with a fast algorithm [8].

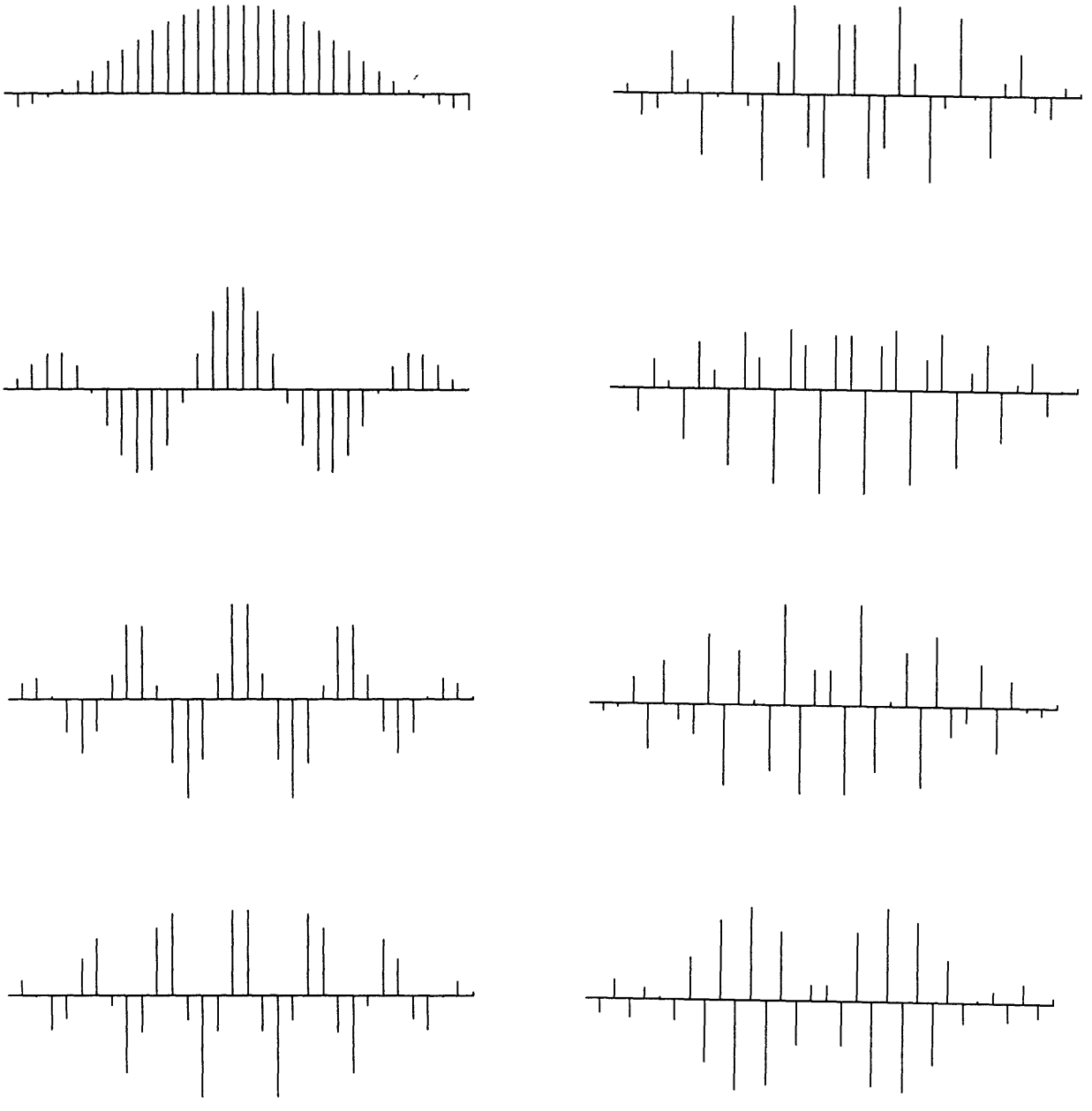


Fig. 5.3(a). An optimal LOT for $N = 16$, $L = 32$, and $\rho = 0.95$, even functions.

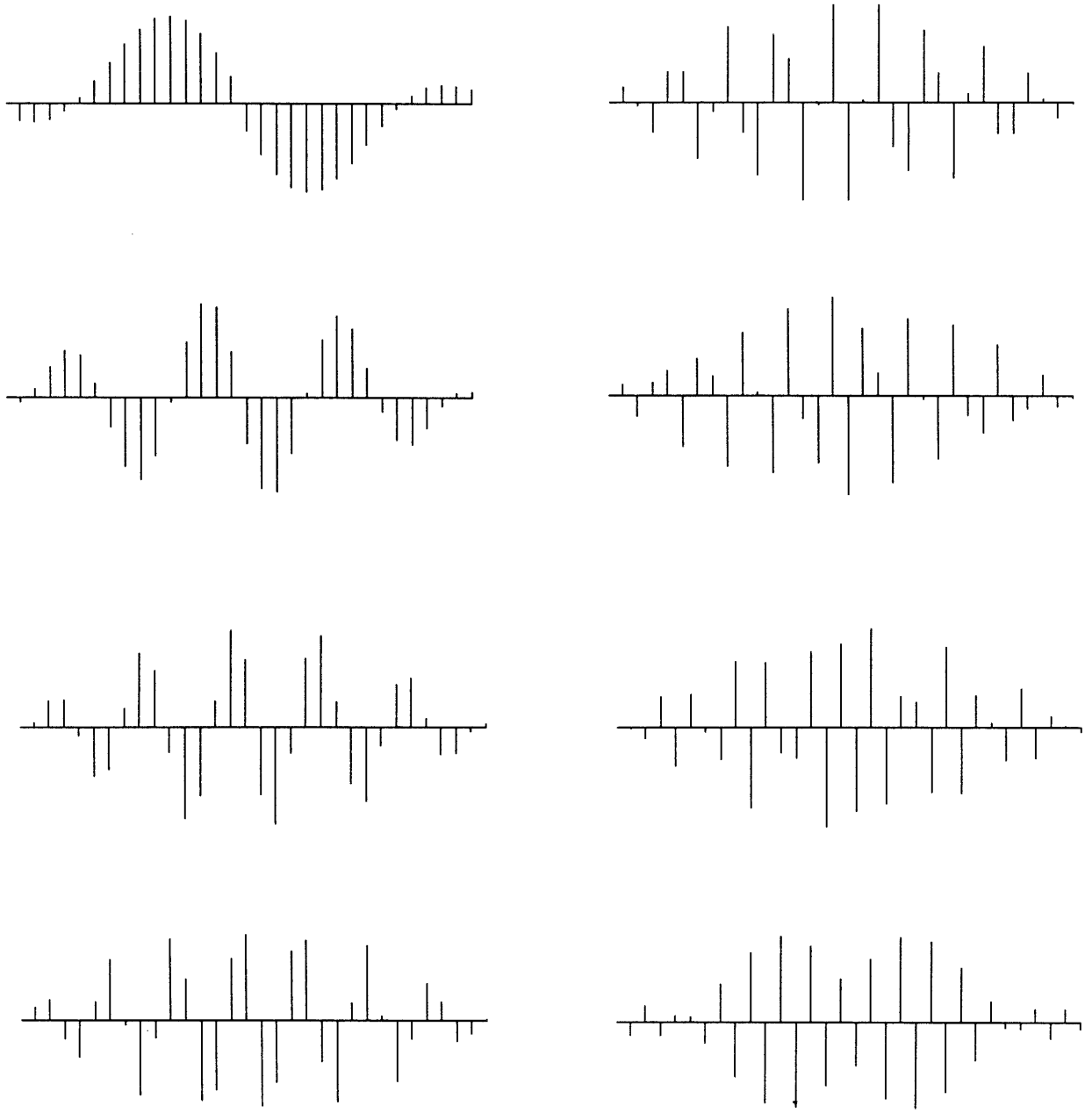


Fig. 5.3(b). An optimal LOT for $N = 16$, $L = 32$, and $\rho = 0.95$, odd functions.

5.3. Fast Implementation of an LOT

The key to the realization of a fast LOT is the approximation of the factor $\mathbf{Z}$ by a matrix that can be expressed as a product of a few simple factors. Actually, this is the main reason why we have chosen the DCT matrix $\mathbf{D}$ in the factorizations in (5.9) and (5.24). With the initial LOT matrix $\mathbf{P}$ in (5.32), it is easy to derive an expression for the transformed correlation matrix $\mathbf{R_o}$. In order to simplify the notation, let's refer to the Gauss-Markov autocorrelation matrix in (5.4) as $\mathbf{R}(2N,\rho)$, where the first parameter represents the matrix order. We can relate $\mathbf{R}(2N,\rho)$ to $\mathbf{R}(N,\rho)$ by

$$\mathbf{R}(2N,\rho) = \begin{pmatrix} \mathbf{R}(N,\rho) & \mathbf{B} \\ \mathbf{B} & \mathbf{R}(N,\rho) \end{pmatrix} , \tag{5.33}$$

where

$$\mathbf{B} = \rho \mathbf{J} \mathbf{r} \mathbf{r}' , \tag{5.34}$$

and $\mathbf{r} = [\ 1\ \rho\ \rho^2\ \cdots\ \rho^N\ ]'$.

Combining (5.32) and (5.33), we obtain, after a few manipulations,

$$\mathbf{R_o} = \begin{pmatrix} \mathbf{R_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R_2} \end{pmatrix} , \tag{5.35}$$

where the diagonal blocks $\mathbf{R_1}$ and $\mathbf{R_2}$ are given by

$$\mathbf{R_1} = \mathbf{D_e'} \mathbf{R}(N,\rho) \mathbf{D_e}' + \mathbf{D_o'} \mathbf{R}(N,\rho) \mathbf{D_o}' + \rho \mathbf{D_e'} \mathbf{r} \mathbf{r}' \mathbf{D_e} + \rho \mathbf{D_o'} \mathbf{r} \mathbf{r}' \mathbf{D_o} , \tag{5.36}$$

and

$$\mathbf{R_2} = \mathbf{D_e'} \mathbf{R}(N,\rho) \mathbf{D_e}' + \mathbf{D_o'} \mathbf{R}(N,\rho) \mathbf{D_o}' - \rho \mathbf{D_e'} \mathbf{r} \mathbf{r}' \mathbf{D_e} - \rho \mathbf{D_o'} \mathbf{r} \mathbf{r}' \mathbf{D_o} , \tag{5.37}$$

If we let the correlation coefficient ρ approach unity, the matrices $\mathbf{D_e}$ and $\mathbf{D_o}$ will contain the asymptotic even and odd eigenvectors of $\mathbf{R}(N,\rho)$, respectively, since the DCT is the limit of the KLT as $\rho \to 1$, as we have seen in Chapter 4. Thus, the terms $\mathbf{D_e'R}(N,\rho)\mathbf{D_e}'$ and $\mathbf{D_o'R}(N,\rho)\mathbf{D_o}'$ are asymptotically diagonal, with positive entries. Also, as $\rho \to 1$ the vector $\mathbf{r}$ will have all of its entries equal to one, i.e., it will be an even vector. Thus, the term $\mathbf{D_o'rr'D_o}$ goes to zero. Furthermore, since the vector $[\ 1\ 1\ \cdots\ 1\ ]'$ is equal to $\sqrt{N}$ times the first column of $\mathbf{D_e}$, it follows that

$$\mathbf{D_e'rr'D_e} \to \begin{pmatrix} N & 0 & 0 & \cdots & 0 \\ 0 & 0 & & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & & 0 \end{pmatrix} . \tag{5.38}$$

Thus, it is clear that $\mathbf{R_1}$ will asymptotically be a diagonal matrix with positive diagonal entries. The factor $\mathbf{R_2}$, however, may not have a dominant diagonal, because the third term in (5.37) is subtracted from the others. Nevertheless, we can expect the following approximation to hold as ρ gets closer to one

$$\mathbf{Z} \simeq \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \tilde{Z} \end{pmatrix} , \tag{5.39}$$

where $\tilde{\mathbf{Z}}$ is of order $N/2$. Although $\mathbf{R_2}$ may not have a strongly dominant diagonal, we should expect some diagonal dominance, so that $\tilde{\mathbf{Z}}$ should not be far from the identity matrix. In Fig. 5.4 we have the optimal $\mathbf{Z}$ for $N = 16$. We note that $\tilde{\mathbf{Z}}$ is actually far from being diagonal, but the magnitudes of its entries decay reasonably fast for indices far from the diagonal. This fact, coupled with the observation that the signs of the entries below the diagonal are negative and the ones above are positive suggests the approximation of $\tilde{\mathbf{Z}}$ by a cascade of $N/2 - 1$ plane rotations, as

$$\tilde{\mathbf{Z}} = \mathbf{T}_1\mathbf{T}_2\cdots\mathbf{T}_{N/2-1} , \tag{5.40}$$

12.824	-0.114	-0.018	-0.006	-0.002	-0.001	-0.001	-0.000	0.000	-0.000	-0.000	0.000	-0.000	-0.000	-0.000	-0.000
-0.114	0.480	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000
-0.018	-0.000	0.144	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000
-0.006	-0.000	-0.000	0.073	-0.000	-0.000	-0.000	-0.000	0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
-0.002	-0.000	-0.000	-0.000	0.047	-0.000	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000	-0.000	0.000
-0.001	-0.000	-0.000	-0.000	-0.000	0.035	-0.000	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	0.000
-0.001	-0.000	-0.000	-0.000	-0.000	-0.000	0.029	-0.000	0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.000	-0.000
-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.026	-0.000	-0.000	0.000	0.000	0.000	0.000	-0.000	-0.000
0.000	-0.000	-0.000	0.000	-0.000	-0.000	0.000	-0.000	1.524	-0.577	-0.159	-0.071	-0.038	-0.021	-0.011	-0.004
-0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000	-0.000	-0.577	0.461	-0.006	-0.002	-0.001	-0.000	-0.000	-0.000
-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000	0.000	-0.159	-0.006	0.143	-0.000	-0.000	-0.000	-0.000	-0.000
0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000	-0.071	-0.002	-0.000	0.072	-0.000	-0.000	-0.000	-0.000
-0.000	0.000	0.000	-0.000	0.000	0.000	-0.000	0.000	-0.038	-0.001	-0.000	-0.000	0.047	-0.000	-0.000	-0.000
-0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.000	0.000	-0.021	-0.000	-0.000	-0.000	-0.000	0.035	-0.000	-0.000
-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.000	-0.000	-0.011	-0.000	-0.000	-0.000	-0.000	-0.000	0.029	-0.000
-0.000	-0.000	-0.000	-0.000	0.000	0.000	-0.000	-0.000	-0.004	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	0.026

Fig. 5.4. Optimal $\mathbf{Z}$ for $N = 16$ and $\rho = 0.95$

where each plane rotation is defined as

$$\mathbf{T}_i \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Y}_{\theta_i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} , \tag{5.41}$$

The matrix $\mathbf{Y}_{\theta_i}$ is a 2×2 butterfly,

$$\mathbf{Y}_{\theta_i} = \begin{pmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{pmatrix} , \tag{5.42}$$

where θ_i is the rotation angle, and the top left identity factor in (5.41) is of order $i - 1$. If we apply the transpose of each $\mathbf{T}_i$ to $\tilde{\mathbf{Z}}$ in the reverse order of

0.995	-0.057	0.044	0.041	0.035	0.029	0.021	0.010
0.053	0.987	-0.089	0.067	0.067	0.057	0.043	0.021
-0.041	0.084	0.983	-0.094	0.081	0.078	0.060	0.030
-0.050	-0.059	0.091	0.981	-0.088	0.088	0.074	0.037
-0.040	-0.077	-0.068	0.086	0.982	-0.075	0.086	0.046
-0.027	-0.062	-0.086	-0.077	0.073	0.985	-0.061	0.058
-0.015	-0.039	-0.061	-0.082	-0.080	0.056	0.987	-0.062
-0.006	-0.016	-0.027	-0.041	-0.056	-0.058	0.055	0.994

Fig. 5.5. Resulting matrix when $\tilde{\mathbf{Z}}$ in Fig. 5.4 is multiplied by an appropriate cascade of plane rotations.

(5.41), we should obtain the identity matrix. The resulting matrix for $[\theta_1 \cdots \theta_7] =$ [0.42 0.53 0.53 0.5 0.44 0.35 0.23 0.11] is shown in Fig. 5.5. That matrix is close enough to the identity for us to accept the approximation in (5.40). With the butterfly angles indicated above, the energy compaction with the approximated functions is $\gamma = 9.32$, which is close to the value $\gamma = 9.49$ corresponding to the exact solution.

The flowgraph of a fast LOT based on the approximation above is shown in Fig. 5.6. Besides the two DCT's of length N and the trivial $+1/-1$ butterflies, we need $N/2 - 1$ butterflies with nontrivial angles. Most of the computation for the direct or inverse transform is in the DCT modules. We note that the fast LOT was justified under the assumption that ρ is large. Nevertheless, just as the DCT, we should expect good performance of the fast LOT even for signals modeled by a relatively low ρ, say 0.6–0.8.

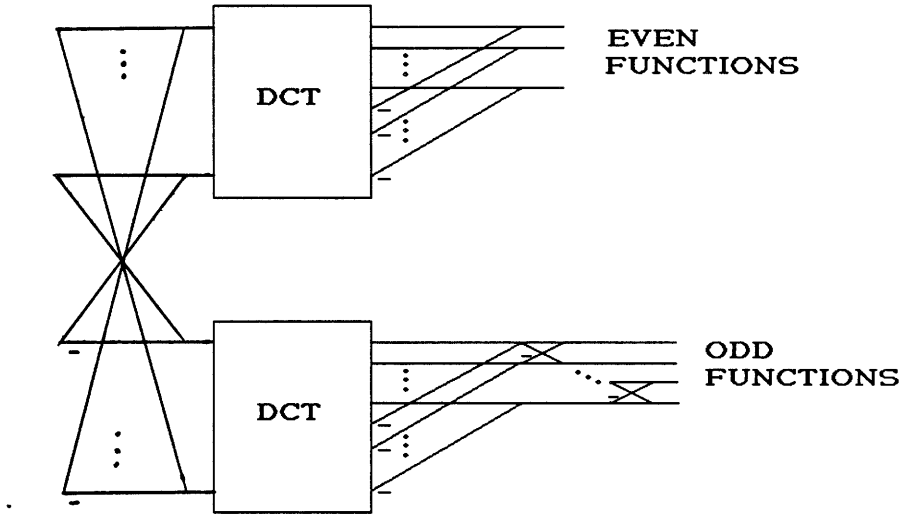


Fig. 5.6. Flowgraph of the fast LOT.

5.4. LOT Performance

The effectiveness of the LOT in reducing blocking effects has been demonstrated by Cassereau [4], for adaptive and non-adaptive coding. In this section we consider the use of the LOT within an optimal block coding system using quantizers with pseudo-random noise (PRN), as suggested in the previous chapter. A typical image processing example is shown in Fig. 5.7. In Fig. 5.7(a) we have an original image, "CAMERA", at a resolution of 256×240 samples, at 8 bits per sample. The right half of the image was replaced by a magnified version of a region of the left half, so that the effects of processing over that particular area of the image could be better

observed.

In Fig. 5.7(b), the image was coded at an average rate of 0.5 bits per sample with the optimal system of Chapter 4, using the DCT as an approximation to the KLT, for a block size $N = 16$. The blocking effect in the magnified area is strong enough to be annoying. In Fig. 5.7(c) the DCT was replaced by the LOT, as derived in the previous section, at the same rate of 0.5 bits per sample. The blocking effects are reduced to a level where they can barely be detected. The coding noise pattern is virtually unaffected by the LOT, being mainly a function of the quantization process. The use of PRN has kept that noise at a relatively low amplitude for the rate of 0.5 bits per pixel. The r.m.s. error was slightly lower with the LOT, the main reason being that the compaction γ of the LOT is somewhat larger than that of the DCT, for the same value of N.

5.5. Summary

We have derived an optimal set of overlapping basis function, which comprise the Lapped Orthogonal Transform, LOT. We have obtained basically the same functions as those reported previously by Cassereau [4]. Unlike the derivation in [4], where the basis functions are obtained recursively as the solutions to a series of non-linear optimization problems, we have obtained the LOT as the solution to a simple eigenvalue problem. Therefore, we have derived an exact representation for the LOT. By approximating one of the factors of the optimal LOT by a product of plane rotations, it was possible to derive an efficient implementation for the LOT, which makes use of two DCT's and a few extra butterflies.

A typical image processing example has shown the efficiency of the LOT in reducing the blocking effect, in agreement with the experiments reported by Cassereau. Since Cassereau's work the LOT has proved to be a good alternative to block image processing, with its only disadvantage to date being the absence of a

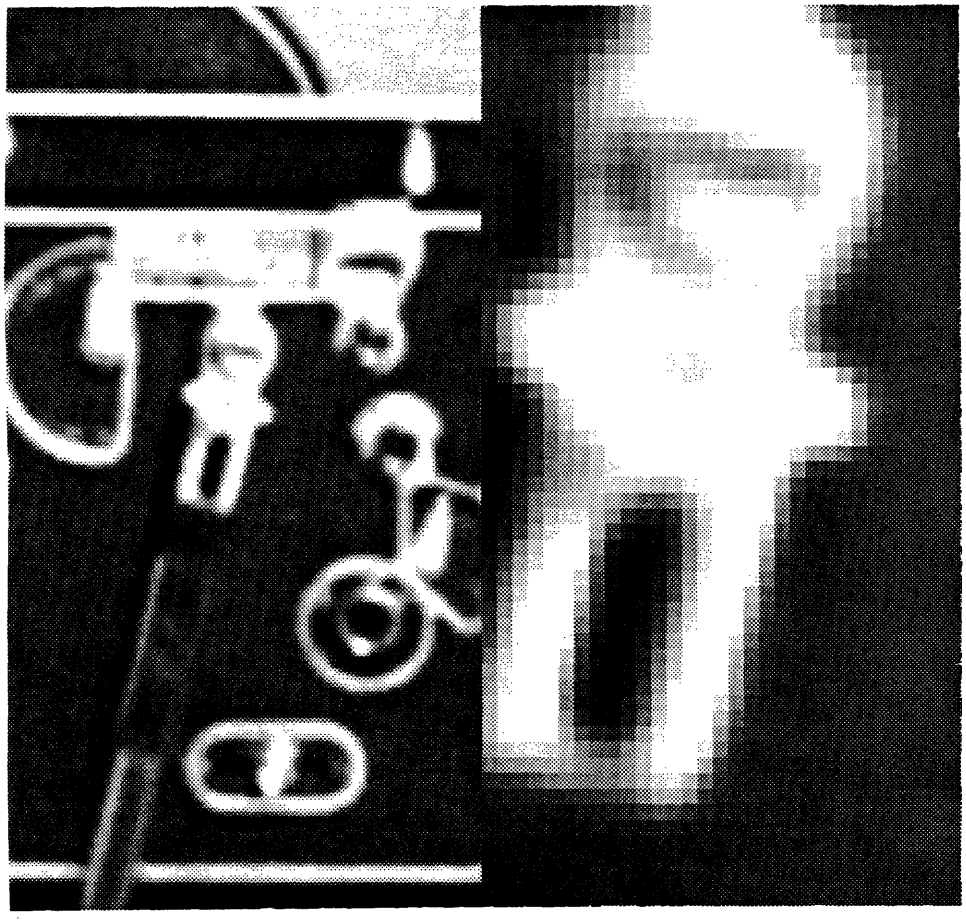


Fig. 5.7(a). Original "CAMERA" image. The right side is a magnified view of a segment from the left side.

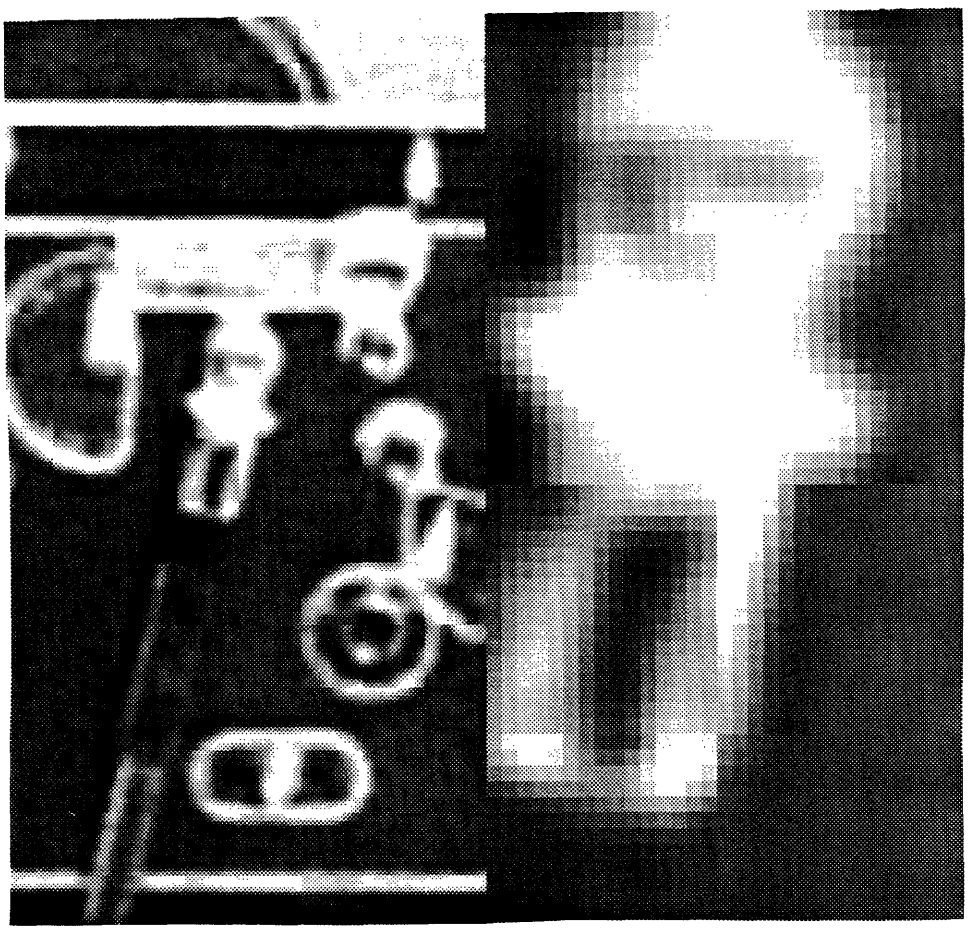


Fig. 5.7(b). "CAMERA" coded at 0.5 bits per sample with the DCT, with a block size $N = 16$. R.m.s. error = 12.1 %

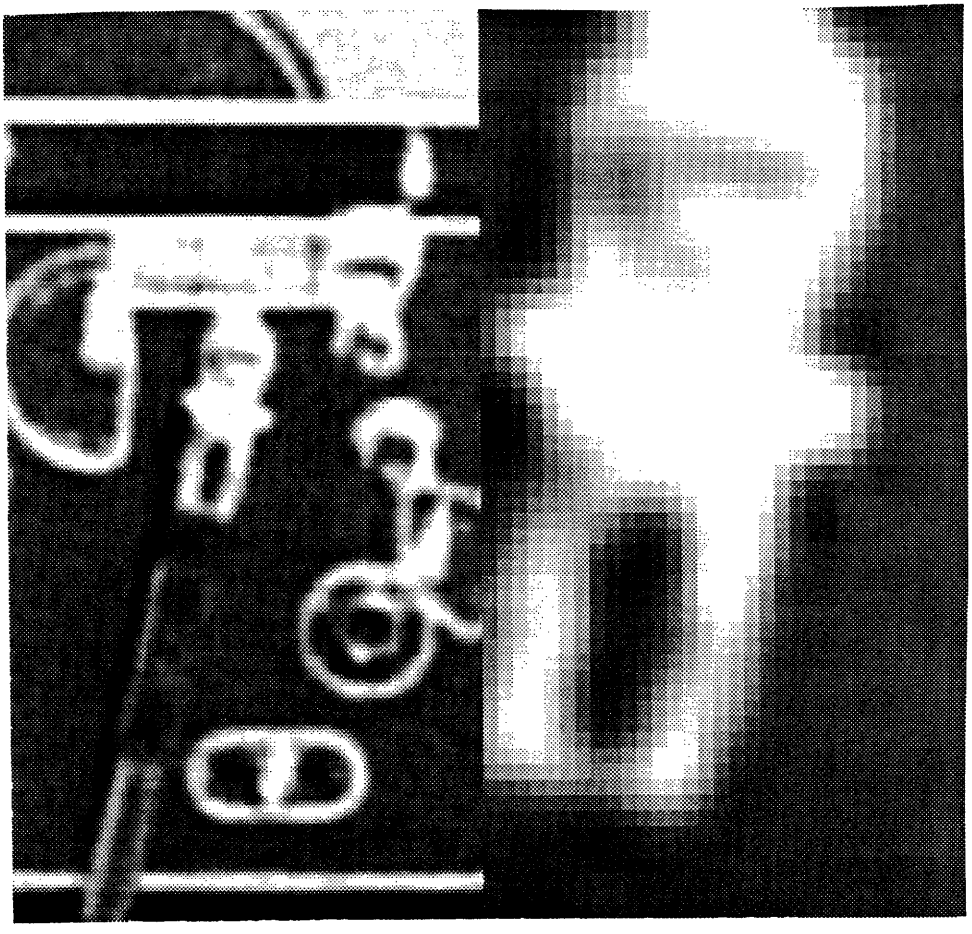


Fig. 5.7(b). "CAMERA" coded at 0.5 bits per sample with the fast LOT, with a block size $N = 16$. R.m.s. error = 10.5 %

fast algorithm. We believe that, with the fast LOT described in this chapter, it is possible to implement block coding systems that significantly outperform traditional block processing at low bit rates (below 1.5 bits per sample) with non-overlapping basis functions.

References

[1] H. C. Reeve III and J. S. Lim, *Reduction of blocking effect in image coding.* in Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Boston, MA, 1983, pp. 1212–1215.

[2] B. L. Hinman, J. G. Bernstein and D. H. Staelin, *Short-space Fourier transform image processing.* in Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, San Diego, CA, 1984, pp. 4.8.1–4.8.4.

[3] M. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-28, pp. 55–69, Feb. 1980.

[4] P. Cassereau, *A new class of optimal unitary transforms for image processing.* S. M. Thesis, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, May 1985.

[5] W. K. Pratt, *Digital Image Processing.* New York: Wiley, 1978, chapter 10.

[6] J. Makhoul, "On the eigenvectors of symmetric Toeplitz matrices," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-29, pp. 868–872, Aug. 1981.

[7] A. Cantoni and P. Butler, "Eigenvalues and eigenvectors of symmetric centrosymmetric matrices," *Lin. Algebra Appl.,* vol. 13, pp. 275–288, 1976.

[8] H. S. Malvar, "Fast computation of the discrete cosine transform through fast Hartley transform," *Electron Lett.,* vol. 22, pp. 352–353, Mar. 27, 1986.

Chapter 6

Conclusions and Suggestions for Further Work

In this thesis we have presented the solutions to the problem of jointly optimizing the pre- and post-filters of the general communications or storage system of Fig. 1.1. Optimality is considered as the minimization of a weighted mean-square signal reconstruction error. We have considered three basic classes of filters: IIR, FIR, and block filters. The optimal solutions for each of the three classes were presented in Chapters 2, 3, and 4, respectively. In Chapter 5 we studied the class of "Lapped Orthogonal Transforms" (LOT) for block processing with overlapping basis functions, and we derived an optimal LOT. Except for the optimal FIR filters of Chapter 3, all the filters derived in this thesis are obtained either in closed form or by means of a finite numerical procedure.

Summary of the Thesis

The system model of Fig. 1.1 is general enough to be applicable to a large class of communications or storage systems. Part of this generality comes from the fact that we considered two different kinds of noise in the channel: additive random noise and quantization errors. We have derived in Appendix B a model for the quantization noise that is not strictly linear but is simple enough to allow an analytical solution for the optimal pre- and post-filtering problem.

The ideal IIR filters derived in Chapter 2 are of theoretical interest only, since they are not realizable by finite lumped networks. Nevertheless, knowledge

of the ideal optimal solutions is fundamental for a thorough understanding of the basic issues related to the pre- and post-filtering problem, for example the concept that an optimal pre-filter should perform a 'half-whitening' operation on the input signal spectrum. The main objective accomplished in Chapter 2 was to group some scattered results already available in the literature for the design of optimal communication systems that are special cases of the general system in Fig. 1.1.

In Chapter 3 we have derived the first results that have immediate practical applications, since FIR filters are easily implementable. The motivation behind the study in Chapter 3 was the point that designing FIR filters that approximate ideal IIR frequency responses is not the correct way to approach the filter design problem, in the author's opinion. If it is known in advance that the filters to be used in practice must be FIR, the signal communications problem must be reformulated with the inclusion of the FIR constraints. This is exactly what we have accomplished in Chapter 3. Although the jointly-optimal filters were derived by means of an iterative procedure that is not guaranteed to converge to a globally optimal solution, we have found the algorithm to be robust, in the sense of providing the correct answer to well-posed problems.

Since FIR filters are frequently employed in multidimensional signal processing, we have also derived in Chapter 3 optimal multidimensional pre- and post-filters. It is interesting to note that even in applications where a single filter is to be designed, for example, in deriving the optimal transmitter for a given receiver (a very important problem in broadcasting), the techniques in Chapter 3 can be employed. In fact, we have shown that a closed-form solution exists when only one filter can be optimized. The image processing examples performed in that chapter have demonstrated the usefulness of our filter design technique to digital image processing.

In Chapter 4 we have derived the optimal block filters for the general communication system of Fig. 1.1. We have extended earlier results available in the

literature in order to include both noise sources. Block filters have a strong advantage over FIR filters: since each signal block is processed independently, it is relatively simple to make a block processing system adaptive. As soon as some strong change in the statistics of the incoming signal is detected, the processing of the next incoming block can be modified.

One strong disadvantage of block filters, in the other hand, is that each filter is in fact a matrix whose elements can be independently determined. For a block size of N samples, filtering the incoming data means performing a matrix multiplication, which has $O(N^2)$ complexity. This problem has led us to search for sub-optimal solutions with $O(N \log N)$ complexity. Borrowing from the block coding literature the concept of replacing the Karhunen-Loève transform (KLT) by the discrete cosine transform (DCT), we have shown that the DCT can also be used to replace the KLT matrix, which is one of the factors of the optimal pre- and post-filters, with an increase in the reconstruction error that is typically less than 0.05 dB. Other factors of the optimal filters are either diagonal matrices or orthogonal matrices that can be implemented by N plane rotations.

Another contribution of Chapter 4 was the idea of using pseudo-random noise (PRN) in the quantization process for digital channels. Although PRN has been employed in practical systems solely to produce less objectionable noise patterns, we have shown that if the system is re-optimized with the PRN characteristics taken into account, the total mean-square error can be reduced significantly (by up to 2 dB), at low bit rates (below two bits per sample). An image processing example has demonstrated the effectiveness of PRN coding.

In Chapter 5 we have studied the LOT class of overlapping transforms, which has been recently introduced as a promising alternative to traditional block processing. The main advantage of the LOT is that it virtually eliminates the blocking effect, which is a major problem in block processing at low bit rates. The optimal LOT for a first-order Gauss-Markov process was only known numerically, until now, and its major deficiency was the absence of a fast computation algorithm. We

have not only analytically derived an optimal LOT as the solution of an eigenvector computation problem, but our analysis also led to the derivation of a fast algorithm for the optimal LOT. Our examples in Chapter 5 suggest that in general the fast LOT should be preferred over the commonly-used DCT for block signal coding at low bit rates.

Suggestions for Further Research

Although our basic system model of Fig. 1.1 is quite general and applicable to a wide range of signal processing environments, there are some points in which it could certainly be improved. For example, we have used a maximum transmitted power constraint in order to derive the jointly-optimal IIR and FIR filters. There are some applications in which signal clipping may be present, e.g., in tape recording, so that the inclusion of maximum probability of clipping constraints might lead to a better model, at the expense of an increased complexity.

Another research opprtunity is the derivation of jointly-optimal filters for different performance criteria. Although the weighted mean-square error criteria is generally adequate for signal communication problems, it is quite possible that systems optimized under other criteria may lead to improved performance, with the meaning of 'improved' being highly dependent on the application.

In Chapter 4, the derivation of the optimal system using PRN-based quantization has assumed the use of uniform quantizers. The main reason for this restriction was that optimal PRN waveforms are not known for non-uniform quantizers. Therefore, if adequate PRN waveforms are derived for a larger class of quantizers, the pre- and post-filter optimization problem could be reformulated with basis on the characteristic of the new PRN waveforms.

Finally, in Chapter 5 the fast LOT has been suggested as an alternative to the DCT for block signal coding. Although we have derived a fast LOT that is a good approximation to an optimal LOT, part of that derivation process of the fast

algorithm was heuristic. An interesting research problem is to use more rigorous techniques to search for even more efficient factorizations of the LOT.

When the communication system has been fully optimized, with data, voice and images transmitted at the greatest fidelity, the most important problem will still remain: improving the quality of what *is being communicated.*

Peter J. Roberts

Appendix A

We want to solve the problem of finding an orthogonal matrix $\mathbf{H} \in \mathbb{R}^{M \times M}$ such that

$$d_i \triangleq [\mathbf{HZH}']_{ii} = 1 \ , \tag{A.1}$$

where $\mathbf{Z} = \text{diag}\{z_1, z_2, \ldots, z_M\}$, with $z_1 \geq z_2 \geq \ldots \geq z_M > 0$, and

$$\frac{1}{M} \sum_{i=1}^{M} z_i = 1 \ . \tag{A.2}$$

The matrix $\mathbf{Z}$ here corresponds to $\mathbf{B\Lambda B}'$ in Chapter 4.

Since the eigenvalues of $\mathbf{HZH}'$ are $z_1, \ldots, z_M$, we have an instance of the inverse eigenvalue problem for real symmetric matrices, i.e., we want to find a positive definite matrix $\mathbf{HZH}'$ with prescribed eigenvalues and diagonal entries. Horn's conditions for the existence of such a matrix are [1]

$$\sum_{i=k}^{M} d_i \geq \sum_{i=k}^{M} z_i \ , \qquad k = 2, 3, \ldots, M \tag{A.3}$$

and

$$\sum_{i=1}^{M} d_i = \sum_{i=1}^{M} z_i \ . \tag{A.4}$$

The second condition, (A.4), is satisfied trivially, in view of (A.1) and (A.2). In order to see that (A.3) also holds, we note that we must have $(z_1 + z_2)/2 \geq 1$, since

the average of the two largest numbers in a set cannot be smaller than the average of all the numbers in the same set. By induction, we must have

$$\sum_{i=1}^{k-1} z_i \geq k-1 \ , \tag{A.5}$$

and thus

$$\sum_{i=k}^{M} z_i = M - \sum_{i=1}^{k-1} z_i \leq M-k+1 = \sum_{i=k}^{M} d_i \ , \tag{A.6}$$

where the last equality comes from the fact that $d_i = 1$, $\forall i$. Thus, Horn's conditions are met, and the existence of $\mathbf{H}$ is assured. In general, $\mathbf{H}$ will not be unique.

Now, we present a recursive algorithm for the computation of a feasible $\mathbf{H}$ that has a particularly useful structure. The algorithm is a simplified version of the procedure devised by Chan and Li [1], in which the d_i's may be different. Since $z_1 \geq z_2 \geq \ldots \geq z_M$ and the average of all z_i equals one, we must have $z_1 \geq 1$, and there must exist and index j such that $1 \geq z_j \geq \ldots \geq z_M$. Calling our original $\mathbf{Z}$ matrix $\mathbf{Z}_1$, we can always find a permutation matrix $\mathbf{P}_1$ that exchanges the second and j-th diagonal elements of $\mathbf{Z}_1$,

$$\mathbf{P}_1 \mathbf{Z}_1 \mathbf{P}_1' = \mathrm{diag}\{z_1, z_j, z_2, z_3, \ldots, z_{j-1}, z_{j+1}, \ldots, z_M\} \ . \tag{A.7}$$

Consider the orthogonal matrix

$$\mathbf{S}_1 \triangleq \begin{pmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \ , \tag{A.8}$$

where identity is of order $M-2$, and the 2×2 orthogonal matrix $\mathbf{Q}_1$ is given by

$$\mathbf{Q}_1 = \frac{1}{\sqrt{z_j - z_1}} \begin{pmatrix} \sqrt{1-z_j} & -\sqrt{z_1 - 1} \\ \sqrt{z_1 - 1} & \sqrt{1-z_j} \end{pmatrix} \ . \tag{A.9}$$

If $z_1 = z_j = 1$, we set $\mathbf{Q}_1 = \mathbf{I}$. Then, we can write

$$\mathbf{S}_1\mathbf{P}_1\mathbf{Z}_1\mathbf{P}_1'\mathbf{S}_1' = \begin{pmatrix} 1 & \mathbf{b}' \\ \mathbf{b} & \mathbf{Z}_2 \end{pmatrix} , \tag{A.10}$$

where $\mathbf{Z}_2$ is a diagonal matrix of order $M-1$ whose trace equals $M-1$. Therefore, we have now the problem of finding orthogonal matrices corresponding to $\mathbf{Z}_2$. By induction, we compute matrices $\mathbf{P}_2$ and $\mathbf{S}_2$ using the procedure described above, obtain a diagonal matrix $\mathbf{Z}_3$, and so on. The matrix $\mathbf{Z}_M$ will be a scalar, and the procedure stops.

The orthogonal matrix $\mathbf{H}$ is, therefore, the product of all the $\mathbf{P}_i$ and $\mathbf{S}_i$,

$$\mathbf{H} = \mathbf{S}_M\mathbf{P}_M\mathbf{S}_{M-1}\mathbf{P}_{M-1}\cdots\mathbf{S}_1\mathbf{P}_1 . \tag{A.11}$$

We note that the permutation factors $\mathbf{P}_i$ are just index manipulations. Furthermore, the orthogonal matrices $\mathbf{S}_i$ are actually plane rotations, according to (A.8), and so each one can be implemented with a single butterfly structure. Thus, multiplication of a vector by the matrix $\mathbf{H}$ in (A.11) requires $O(M)$ operations.

Reference

[1] N. N. Chan and K.-H. Li,"Diagonal elements and eigenvalues of a real symmetric matrix," *J. Math. Anal. Appl.*, vol. **91**, pp. 562–566, 1983.

Appendix B

In this appendix we derive a signal-dependent linear model for the scalar quantization process. Let's call v and y the input and output for a scalar quantizer, respectively. We can always write

$$y = v + d \,, \tag{B.1}$$

where d is the quantization noise, which depends on the input signal. If the number of levels in the quantizer is large, the correlation between the noise d and the signal v may be negligible. At low bit rates, however, that correlation may be significant.

When a Max quantizer is employed and v has a probability distribution with zero mean, the cross-correlation between v and d is given by [1], [2],

$$\mathrm{E}\,[vd] = -\sigma_d^2 = -\epsilon^2 \sigma_v^2 \,, \tag{B.2}$$

where the first equality is a consequence of the quantizer optimality, and the second is the definition of the parameter ϵ^2, which is called the 'quantizer performance factor'. We use the letter σ to indicate standard deviations. A lower bound for ϵ^2 is 2^{-2k}, where k is the number of bits used to represent y, according to Rate Distortion Theory [1].

It is interesting to note that an almost trivial proof of (B.2) can be derived, based solely on linear estimation theory, without taking into account the specific

level assignments for the quantizer. Such a proof, which we now present, is based on the following lemma.

Lemma. *Assume that v is a random variable with a zero mean. If y is obtained from v by means of an unbiased scalar quantizer with a minimum mean-square error, then*

$$\mathrm{E}[vy] = \mathrm{E}[yy] \ . \tag{B.3}$$

Proof: Suppose we use the the measurement y to obtain the best linear unbiased estimator of v, say $\hat{v}$. Since v and y are zero-mean by assumption, $\hat{v}$ is given by the classical formula [3]

$$\hat{v} = \frac{\mathrm{E}[vy]}{\mathrm{E}[yy]}\, y \ . \tag{B.4}$$

However, under the assumption that the quantizer leads to a minimum mean-square error, we must have $\hat{v} = y$, which implies (B.3). ∎

Using (B.3), we get

$$\begin{aligned} \mathrm{E}[vy] &= \mathrm{E}[v(v+d)] = \sigma_v^2 + \mathrm{E}[vd] \\ &= \mathrm{E}[yy] = \mathrm{E}[(v+d)(v+d)] = \sigma_v^2 + 2\,\mathrm{E}[vd] + \sigma_d^2 \ , \end{aligned} \tag{B.5}$$

from which $\mathrm{E}[vd] = -\sigma_d^2$

Our analysis of the block processing system with scalar quantizers in Chapter 4 can be simplified if we can use a quantization model in which the additive noise is uncorrelated with the signal. One idea is to replace (B.1) by

$$y = \psi v + \tilde{d} \ , \tag{B.6}$$

which is the so-called 'gain plus additive noise' model presented in [1]. The above equation can also be viewed as the decompositon of the noise into two additive

components: one that can be estimated from v and an uncorrelated component, i.e.,

$$y = v + \mathrm{E}[d \mid v] + \tilde{d} \,. \tag{B.7}$$

It is easy to verify the equivalence of (B.6) and (B.7).

For a given performance factor ϵ^2, we can show that the models in (B.1) and (B.6) are equivalent if

$$\psi = 1 - \epsilon^2 \tag{B.8}$$

and

$$\sigma_{\tilde{d}}^2 = \psi(1-\psi)\sigma_v^2 \,, \tag{B.9}$$

with $\mathrm{E}[v\tilde{d}] = 0$.

We note that the model in (B.6) is accurate for all bit rates. In fact, it is even correct for zero-bit quantization, since in this case $\epsilon^2 = 1$ and the gain ψ is zero, which produces the correct output $y = 0$. When (B.6) is applied to all elements of the pre-filter output $\mathbf{v}$, we obtain the model of Fig. 4.8, in which the matrix $\mathbf{\Psi}$ is diagonal and the noise autocorrelation $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ is also diagonal. The entries of those matrices depend on the performance factor ϵ_i^2, which is a function not only of the bit assignment, but also of the p.d.f. of each element of $\mathbf{v}$.

References

[1] N. S. Jayant and P. Noll, *Digital Coding of Waveforms, Principles and Applications to Speech and Video*, Englewood Cliffs, N.J.: Prentice-Hall, 1984, chapter 4 and appendix D.

[2] K. Sayood and J. Gibson,"Explicit additive noise models for uniform and nonuniform quantization". *Signal Processing*, vol. **7**, pp. 407–414, 1984.

[3] A. Gelb, *Applied Optimal Estimation*, Cambridge, MA: MIT Press, 1974, chapter 2.

Appendix C

Our objective in this appendix is the minimization of

$$\xi_w = N^{-1} \operatorname{tr}\{[\mathbf{\Lambda}^{-1} + \mathbf{B}'\mathbf{\Psi}\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}^{-1}\mathbf{\Psi}\mathbf{B}]^{-1}\} , \tag{C.1}$$

where the diagonal matrices $\mathbf{\Psi}$ and $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ depend on $\mathbf{B}$. The entries of $\mathbf{\Psi}$ and $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ are determined by (B.7) and (B.8).

Since there are no restrictions on the matrix $\mathbf{B}$, a minimum of (C.1) must be a stationary point, i.e.,

$$\frac{\partial \xi_w}{\partial \mathbf{B}} = 0 \Big|_{\mathbf{B}=\mathbf{B}_{OPT}} . \tag{C.2}$$

Because of the dependence of $\mathbf{\Psi}$ and $\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}$ on $\mathbf{B}$, we cannot make use of the table of matrix derivatives in [1] and [2]. Therefore, we must compute the derivative explicitly, using the techniques in [2].

Let's introduce a diagonal matrix

$$\mathbf{Z} \triangleq \mathbf{\Psi}^{-1}\mathbf{R}_{\tilde{\mathbf{d}}\tilde{\mathbf{d}}}\mathbf{\Psi}^{-1} = \operatorname{diag}\{\mathbf{z}_1, \ldots, \mathbf{z}_M\} , \tag{C.3}$$

which can be used to rewrite (C.1) in the form

$$\xi_w = N^{-1} \operatorname{tr}\{[\mathbf{\Lambda}^{-1} + \mathbf{B}'\mathbf{Z}^{-1}\mathbf{B}]^{-1}\} . \tag{C.4}$$

Using (B.7) and (B.8) from Appendix B, we conclude that the entries of $\mathbf{Z}$ must be given by

$$z_i = \frac{\sigma_{\tilde{d}_i}^2}{\psi_i^2} = \frac{1-\psi_i}{\psi_i}\,\sigma_{v_i}^2 \ , \tag{C.5}$$

where ψ depends on the parameters of the i-th quantizer (input p.d.f., number of bits, and uniformity of the levels). Since

$$\sigma_{v_i}^2 = [\mathbf{B}\mathbf{\Lambda}\mathbf{B}']_{ii} \ , \tag{C.6}$$

we can write

$$\mathbf{Z} = \sum_{i=1}^{M} \alpha_i \, \mathbf{E}_{ii}\mathbf{B}\mathbf{\Lambda}\mathbf{B}'\mathbf{E}_{ii} \ , \tag{C.7}$$

where

$$\alpha_i \triangleq \frac{1-\psi_i}{\psi_i} \tag{C.8}$$

and $\mathbf{E}_{kl}$ is the elementary matrix that has a one on the (k,l)-th position, and zeros elsewhere [2].

Let's further define

$$f(\mathbf{B}) \triangleq [\mathbf{\Lambda}^{-1} + \mathbf{B}'\mathbf{Z}^{-1}\mathbf{B}] \ , \tag{C.9}$$

so that $N\xi_w = \text{tr}\,\{f^{-1}(\mathbf{B})\}$. Taking the matrix derivative in (C.2) elementwise [2], we get

$$\frac{\partial N\xi_w}{\partial \mathbf{B}} = \sum_{i,j} \mathbf{E}_{ij} \ \text{tr}\left\{\frac{\partial f^{-1}(\mathbf{B})}{\partial b_{ij}}\right\} \ . \tag{C.10}$$

Using the basic properties in [2], we obtain

$$\frac{\partial f^{-1}(\mathbf{B})}{\partial b_{ij}} = -f^{-1}(\mathbf{B})\frac{\partial f(\mathbf{B})}{\partial b_{ij}} f^{-1}(\mathbf{B}) \ . \tag{C.11}$$

Substituting (C.9) into (C.11), we get

$$\begin{aligned}\frac{\partial f(\mathbf{B})}{\partial b_{ij}} &= \frac{\partial}{\partial b_{ij}}[\mathbf{B}'\mathbf{Z}^{-1}\mathbf{B}] \\ &= \mathbf{E}_{ij}\mathbf{Z}^{-1}\mathbf{B} + \mathbf{B}'\mathbf{Z}^{-1}\,\mathbf{E}_{ij}\mathbf{B}'\frac{\partial \mathbf{Z}^{-1}}{\partial \mathrm{b}_{ij}}\mathbf{B}\ .\end{aligned} \tag{C.12}$$

The derivative of $\mathbf{Z}^{-1}$ with respect to each b_{ij} can be obtained from the tables in [1], [2], as

$$\frac{\partial \mathbf{Z}^{-1}}{\partial b_{ij}} = -\mathbf{Z}^{-1}\left[\sum_{l=1}^{M}\alpha_l^2\,\mathbf{E}_{ll}(\mathbf{E}_{ij}\Lambda B' + B\Lambda\,\mathbf{E}_{ji})\,\mathbf{E}_{ll}\right]\mathbf{Z}^{-1}\ . \tag{C.13}$$

From the definition of the elementary matrices, it is clear that $\mathbf{E}_{ll}\,\mathbf{E}_{ij} = 0$ and $\mathbf{E}_{ji}\,\mathbf{E}_{ll} = 0$, if $l \neq i$. Thus, (C.13) can be simplified to

$$\frac{\partial \mathbf{Z}^{-1}}{\partial b_{ij}} = -\mathbf{Z}^{-1}(\mathbf{E}_{ij}\mathbf{A}\Lambda\mathbf{B}' + \mathbf{B}\Lambda\mathbf{A}\,\mathbf{E}_{ji})\mathbf{Z}^{-1}\ , \tag{C.14}$$

where $\mathbf{A} \triangleq \mathrm{diag}\{\alpha_1, \ldots, \alpha_M, 0, \ldots, 0\}$ is a matrix of order N.

Now we back substitute (C.14) into (C.12), into (C.11), and finally into (C.10), with the result

$$\frac{\partial N\xi_w}{\partial \mathbf{B}} = -2\ \mathbf{Z}^{-1}\mathbf{B}f^{-2}(\mathbf{B})[\mathbf{I} + \mathbf{B}'\mathbf{Z}^{-1}\mathbf{B}\Lambda\mathbf{A}]\ . \tag{C.15}$$

An optimal $\mathbf{B}$ must satisfy $\partial N\xi_w/\partial\mathbf{B} = 0$, which implies

$$\mathbf{B}'\mathbf{Z}^{-1}\mathbf{B} = \Lambda^{-1}\mathbf{A}^{-1} \tag{C.16}\ ,$$

and for any such $\mathbf{B}$ the error is given by

$$\xi_w = N^{-1}\,\mathrm{tr}\,\{\Lambda[\mathbf{I} + \mathbf{A}^{-1}]^{-1}\}\ . \tag{C.17}$$

Thus, any $\mathbf{B}$ that satisfies (C.16) leads to a minimum of the error. It is easy to verify that one such $\mathbf{B}$ is a generalized identity matrix, i.e.,

$$\mathbf{B} = [\mathbf{I}\ \mathbf{0}]\ . \tag{C.18}$$

It is important to note that the optimality of $\mathbf{B}$ in (C.18) does not depend on the matrix $\mathbf{A}$, that is, it does not depend on the performance factor of each quantizer. Hence, we don't need specific assumptions about the p.d.f.'s of the quantizer inputs and the particular bit allocation in order to have (C.18) valid. Even the inclusion of PRN on the quantizers does not affect (C.18). We conclude that there is some separability between the problems of quantizer design and pre- and post-filter design.

References

[1] M. Athans and F. C. Schweppe, *Gradient matrices and matrix calculations.* Technical Note 1965-53, M.I.T. Lincoln Laboratory, Cambridge, MA, 1965.

[2] A. Graham, *Kronecker Products and Matrix Calculus: with Applications* Chichester, England: Ellis Horwood, 1981.