Optimal Bit and Power Constrained Filter Banks

Are Hjørungnes

A dissertation submitted in partial fulfillment
of the requirements for the doctoral degree of

DOKTOR INGENIØR

Department of Telecommunications
Norwegian University of Science and Technology
N-7491 Trondheim
Norway

2000

Abstract

In this dissertation, two filter banks optimization problems are studied. The first problem is the optimization of filter banks used in a subband coder under a bit constraint. In the second problem, a multiple input multiple output communication system is optimized under a power constraint. Three different cases on the filter lengths are considered: unconstrained length filter banks, transforms, and finite impulse response filter banks with arbitrary given filter lengths.

In source coding and multiple input multiple output communication systems, transforms and filter banks are used to decompose the source in order to generate samples that are partly decorrelated. Then, they are more suitable for source coding or transmission over a channel than the original source samples. Most transforms and filter banks that are studied in the literature have the perfect reconstruction property. In this dissertation, the perfect reconstruction condition is relaxed, so that the transforms and filter banks are allowed to belong to larger sets, which contain perfect reconstruction transforms and filter banks as subsets.

Jointly optimal analysis and synthesis filter banks and transforms are proposed under the bit and power constraint for all the three filter length cases. For a given number of bits used in the quantizers or for a given channel with a maximum allowable input power, the analysis and synthesis transforms and filter banks are jointly optimized such that the mean square error between the original and decoded signal is minimized. Analytical expressions are obtained for unconstrained length filter banks and transforms, and an iterative numerical algorithm is proposed in the finite impulse response filter bank case.

The channel in the communication problem is modeled as a known multiple input multiple output transfer matrix with signal independent additive vector noise having known second order statistics. A pre- and postprocessor containing modulation is introduced in the unconstrained length filter bank system with a power constraint. It is shown that the performance of this system is the same as the performance of the power constrained transform coder system

when the dimensions of the latter system approach infinity.

In the source coding problem, the results are obtained with different quantization models. In the simplest model, the subband quantizers are modeled as additive white signal independent noise sources. The proposed unconstrained length filter banks perform better than the optimal unconstrained length unitary and biorthogonal filter banks, and it is shown that the proposed transform has better performance than the Karhunen-Loève transform. Also, the proposed transform coder has the same performance as a transform coder using a reduced rank Karhunen-Loève analysis transform with jointly optimal bit allocation and Wiener synthesis transform. The proposed finite impulse response filter banks have at least as good theoretical rate distortion performance as the perfect reconstruction filter banks and the finite impulse response Wiener filter banks used in the comparisons.

A practical coding system is introduced where the coding of the subband signals is performed by uniform threshold quantizers using the centroids as representation levels. It is shown that there is a mismatch between the theoretical and practical results. Three methods for removing this mismatch are introduced. In the two first methods, the filter banks themselves are unchanged, but the coding method of the subband signals is changed. In the first of these two methods, quantizers are derived such that the additive coding noise and subband signals are uncorrelated. Subtractive dithering is the second method used for coding of the subband signals. In the third method, a signal dependent colored noise model is introduced, and this model is used to redesign the filter banks. In all three methods, good correspondence is achieved between the theoretical and practical results, and comparable or better practical rate distortion performance is achieved by the proposed methods compared to systems using perfect reconstruction filter banks and finite impulse response Wiener synthesis filter banks.

Finally, conditions for when finite impulse response filter banks are optimal are derived.

Preface

This dissertation is submitted in partial fulfillment of the requirements for the doctoral degree of *doktor ingeniør* at the Norwegian University of Science and Technology (NTNU).

The work, including the compulsory courses corresponding to full-time studies in two semesters, as well as one year of teaching assistant duties, has taken place in the period of January 1996 to January 2000. From September 1997 to August 1998, the research was carried out at the University of California, Santa Barbara, USA. The rest of the work was done at the Department of Telecommunications, NTNU.

The work has been funded by scholarships from the Research Council of Norway (NFR) and the Department of Telecommunications, NTNU.

The work has been completed under the guidance and supervision of Professor Tor A. Ramstad at the Department of Telecommunications, NTNU.

Acknowledgments

I would like to express my sincere thanks to Professor Tor A. Ramstad for his support throughout this work. With his insight and advice, he has contributed significantly to the ideas and content of this dissertation. I will also like to thank him for the great support I received during our stay in Santa Barbara.

I am grateful to Professor Sanjit K. Mitra at University of California, Santa Barbara, who gave me the opportunity to stay with his group for eleven months from September 1997 to August 1998 plus three weeks in November and December 1998. Sanjit and his group are genuinely friendly people, and the months in USA proved to be both enjoyable and rewarding with respect to my dissertation work.

My colleagues at the Department of Telecommunications have contributed tremendously towards fostering an excellent work environment. For this I am truly graceful. In particular, my office mate, Helge Coward, deserves special recognition for his willingness to discuss my research problems, his company and support during dinners in the campus canteen, and for all his support with computer related problems.

Several people have helped me in the finishing stages of the dissertation. They are as follows: Helge Coward, David Choi, Tage Røsten, Morten Eriksen, Tor A. Ramstad, Geir Øien (Chapter 5), Øyvind Kvennås, and Sigurd Pleym.

I would like to send special thanks to my family and friends. Their support and encouragement during these years are greatly appreciated.

Contents

Abstract iii

Preface v

Acknowledgments vii

Nomenclature xxi

List of Abbreviations xxxiii

1 Introduction 1
- 1.1 Scope of the Dissertation . . . 2
  - 1.1.1 Filter Banks for Compression . . . 2
  - 1.1.2 Filter Banks for Communication . . . 4
- 1.2 Problems Considered and Basic Assumptions . . . 5
  - 1.2.1 Bit Constrained Filter Banks . . . 5
  - 1.2.2 Power Constrained Filter Banks . . . 10
- 1.3 Previous Work . . . 12
  - 1.3.1 Bit Constrained Filter Banks . . . 12
  - 1.3.2 Power Constrained Filter Banks . . . 13
- 1.4 Outline of the Dissertation . . . 14
- 1.5 Contributions of the Dissertation . . . 15

2 Unconstrained Length Signal-Adaptive Filter Banks 17
- 2.1 Bit Constrained Filter Banks . . . 17
  - 2.1.1 Problem Formulation . . . 18
  - 2.1.2 Transformation to an Equivalent Problem . . . 19
  - 2.1.3 Optimal System Solution . . . 20
    - 2.1.3.1 Necessary Conditions for Optimality . . . 21
    - 2.1.3.2 Synthesis Polyphase Matrix . . . 22

2.1.3.3 Conditions for Optimality of PR in the Unconstrained Length Filter Bank Case 22
2.1.3.4 Zero Elements of the Matrix $\boldsymbol{G}(f)$ 23
2.1.3.5 The Matrix $\boldsymbol{G}(f)$ 25
2.1.3.6 Block MSE and Bit Constraint Expressions . . 30
2.1.3.7 The Elements of $\boldsymbol{G}(f)$ 31
2.1.3.8 Frequency Response Expressions 32
2.1.3.9 Aliasing Noise 34
2.1.3.10 High Rate Case 35
2.1.4 Bit Constrained Filter Bank Results 36
2.2 Power Constrained Filter Banks 38
2.2.1 Problem Formulation 39
2.2.2 Optimal Matrix System 39
2.2.3 A Combined Source-Channel Coding Problem 41
2.2.4 Power Constrained Filter Bank Results 43
2.3 Summary 45

3 Signal-Adaptive Transforms 47
3.1 Bit Constrained Transforms 47
3.1.1 Problem Formulation 48
3.1.2 Optimal Transform Coder 49
3.1.2.1 Performance Expressions 51
3.1.2.2 Linear Phase 52
3.1.3 Comparisons to the KLT 52
3.1.4 Conditions for Optimality of PR in the Transform Case 54
3.1.5 Wiener Transformed KLT 56
3.1.6 Bit Constrained Transform Results 57
3.2 Power Constrained Transforms 61
3.2.1 Problem Formulation 62
3.2.2 Alternative Derivation of the BPAM system 62
3.2.3 Performance Expressions 63
3.3 Summary 64

4 Algorithms for Finding Signal-Adaptive FIR Filter Banks 65
4.1 Bit Constrained FIR Filter Banks 66
4.1.1 Problem Formulation 66
4.1.2 Equations for Optimality with Equal Filter Lengths . . 70
4.1.3 Numerical Optimization Algorithm 72
4.1.4 Arbitrary Filter Length Optimization 72
4.1.5 Bit Constrained FIR Filter Bank Results 74
4.1.5.1 Linear Phase 78

4.1.5.2 Magnitude Response Comparisons 79
4.2 Power Constrained FIR Filter Banks 80
4.2.1 Problem Formulation 80
4.2.2 Equations for Optimality 82
4.2.3 Power Constrained FIR Filter Bank Results 84
4.2.3.1 Comparison to a Method Proposed by Honig et al. 84
4.2.3.2 Comparison to a Method Proposed by Malvar 86
4.3 Summary 89

5 Connection between BPAM and Power Constrained MIMO 91
5.1 Problem Formulation 92
5.2 The BPAM System 93
5.3 The MIMO System 94
5.4 Comparison of the BPAM and Modulated MIMO Systems . . . 95
5.4.1 BPAM with Dimensions Approaching Infinity 95
5.4.1.1 Case 1: $K \geq L$ 95
5.4.1.2 Case 2: $K < L$ 96
5.4.2 Modulated MIMO System 97
5.4.2.1 Case 1: $M \geq N$ 100
5.4.2.2 Case 2: $M < N$ 101
5.4.3 Comparison 102
5.5 Numerical Example 102
5.6 Summary 102

6 Practical Simulations for Bit Constrained FIR Filter Banks 105
6.1 Practical Coding System 105
6.2 Comparison of Theoretical and Practical Results 109
6.3 Analysis of the Reasons for the Mismatch 111
6.3.1 Bit Rate Estimation 111
6.3.2 Block MSE Estimation 113
6.4 Summary 116

7 Improvements of the Correspondence between Theory and Practice 119
7.1 Uncorrelated Subband Signals and Coding Noise 120
7.1.1 Redesigned Scalar Quantizers 121
7.1.2 Subtractive Dithering 122
7.2 Signal Dependent Colored Quantization Noise Model 123
7.2.1 Problem Formulation 124
7.2.1.1 PR Expressions when $N = M$ 127

7.2.2 Equations for Optimality . . . 128
7.2.3 FIR Wiener Synthesis Filter Bank . . . 129
7.2.4 Numerical Optimization Algorithm . . . 130
7.3 Conditions for Optimality of PR in the FIR Case . . . 131
7.4 Results and Comparisons . . . 134
7.4.1 Redesigned Scalar Quantizer . . . 134
7.4.2 Subtractive Dithering . . . 136
7.4.3 Signal Dependent Colored Quantization Noise Model . . . 136
7.4.4 Practical Performance Comparisons . . . 140
7.5 Summary . . . 143

8 Conclusions 145

A Derivation of Block MSE, Bit Constraint, and Power Constraint 151
A.1 Block MSE Derivation . . . 151
A.1.1 Block MSE for Unconstrained Length Filter Banks . . . 152
A.1.2 Block MSE for FIR Filter Banks . . . 154
A.2 Bit Constraint Derivation . . . 155
A.2.1 Bit Constraint for Unconstrained Length Filter Banks . 155
A.2.2 Bit Constraint for FIR Filter Banks . . . 156
A.3 Power Constraint Derivation . . . 157
A.3.1 Power Constraint for Unconstrained Length Filter Banks . . . 157
A.3.2 Power Constraint for FIR Filter Banks . . . 158

B Ordering Functions and Eigenvalues of the PSD Matrix 159
B.1 Eigenvalues of the PSD Matrix . . . 159
B.2 Properties of the Ordering Functions . . . 161

C Matrix Variational Calculus and Differentiation 165
C.1 Matrix Variational Calculus . . . 165
C.2 Matrix Differentiation Results . . . 168

D Correlation Matrix Elements 171
D.1 Elements of the $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ Matrix . . . 172
D.2 Elements of the $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ Matrix . . . 173

References 179

List of Figures

1.1 Subband coder model. 6
1.2 One branch of the subband coder model. 9
1.3 The power constrained MIMO block system. 11

2.1 The equivalent block diagram of the subband coder model. . . . 20
2.2 System example for the unconstrained length filter banks, using $N = 2$. (a) PSD of the input signal $S_x(f)$. (b) Frequency response of the first analysis filter $H_0(f)$. (c) Frequency response of the second analysis filter $H_1(f)$. (d) Overall frequency response. In the last three plots, the dash-dotted curves give the result with $b = 3.28$ bits/sample and SNR = 18.41 dB, while the dotted curves show the result with $b = 1.12$ bits/sample and SNR = 7.51 dB. 36
2.3 Different distortion rate performances when coding a Gaussian AR(1)-process with correlation factor 0.9 and $N = 3$ subbands. The solid cure shows the performance of the proposed unconstrained length filter bank system, the dash-dotted curve represents the optimal unconstrained length biorthogonal system, while the dashed curve shows the optimal unconstrained length unitary system performance. The dotted curve shows the distortion rate function. 38
2.4 The combined source-channel coding problem. 41
2.5 The MIMO system used to solve the combined source-channel coding problem. 42
2.6 The SNR vs. CSNR performance of the proposed system and OPTA using $N = 3$, $M = 3$, and $C(z) = 1$ is shown by the solid curves. The upper curve is OPTA. The dash-dotted curves show the performance of the proposed system and OPTA using $N = 3$, $M = 2$. The upper curve is OPTA. The input source is a Gaussian AR(3) with PSD shown in Figure 2.2 (a), and the channel noise is white and Gaussian. 43

2.7 System example using $N = 2$, $M = 2$, and $C(z) = 1$. (a) PSD of the input signal $S_x(f)$. (b) Frequency response of the first channel on the transmitter side $H_0(f)$. (c) Frequency response of the second channel on the transmitter side $H_1(f)$. (d) Total frequency response through the system. In the last three plots, the dotted curves give the result with CSNR = -0.30 dB and SNR = 4.61 dB, while the dash-dotted curves show the result with CSNR = 4.80 dB and SNR = 7.79 dB. 44

3.1 Transform coder model. 48
3.2 KLT and proposed transform coder models. 53
3.3 Reduced rank analysis KLT with Wiener synthesis transform. . 55
3.4 The distortion rate function [Berger 1971] of the Gaussian AR(3) source with PSD shown in Figure 2.7 (a) is shown by the solid curve, the performance of the proposed transform coder is shown by the dash-dotted curve, and the dotted curve shows the performance of the KLT, using $N = 2$ and $c_i = \frac{\sqrt{3}\pi}{2}$. 58
3.5 Comparison between the Wiener transformed KLT and the proposed transform model. 60
3.6 Power constrained transform model. 61

4.1 Magnitude response for the 9_7 filter banks, where $N = M = 2$, $c_i = \frac{e\pi}{6}$, $m = l = 4$, $d_v = 5$, and $d_s = 0$. The proposed filter bank is shown by the solid curves, the dash-dotted curves show linear phase biorthogonal system [Balasingham 1998] responses, and the dashed curves show the responses of the 9_7 wavelet in [Antonini, Barlaud, Mathieu & Daubechies 1992]. A Gaussian AR(1) source with correlation coefficient 0.95 is coded at 2.21 bits/sample in all cases. Row number i in the figure is subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters. 75
4.2 The upper figure shows the alias free transfer function $A_0(f)$ for the 9_7 filter bank shown in Figure 4.1. The lower figure shows the corresponding first aliasing term $A_1(f)$. The same parameters as in Figure 4.1 are used here. 77
4.3 Impulse responses of the proposed 9_7 FIR filter bank system for the same parameters as in Figure 4.1. Row number i in the figure is subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters. 78

4.4 Comparisons of magnitude responses of the proposed systems. The input signal is a Gaussian AR(1) signal with correlation coefficient 0.95, $N = M = 2$, and $c_i = \frac{e\pi}{6}$. The solid curves show the frequency response of the unconstrained length filter banks from Section 2.1, the dotted curves show the FIR responses when using the optimized 9_7 filter banks with $d_v = 5$ and $d_s = 0$ from this section, and the dash-dotted curves show the bit constrained transform magnitude responses from Section 3.1. The bit rate used is 4.21 bits/sample in all cases. Row number i in the figure is subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters. 79

4.5 SNR vs. CSNR performance of proposed power constrained FIR filter banks are shown by the $\times$-marks while the system proposed in [Honig, Crespo & Steiglitz 1992] is shown by the circles. Both the input time series and the additive channel noise are white and Gaussian, $N = M = 2$, $m = l = 2$, $o = 0$, $d_v = 1$, $d_s = 0$, and $\boldsymbol{C}(z) = \boldsymbol{I}$. 6_6 filter banks are used. 84

4.6 SNR vs. CSNR performance of proposed power constrained FIR filter banks are shown by the $\times$-marks while the system proposed in [Honig et al. 1992] is shown by the circles. The input time series is a Gaussian AR(1) time series with correlation coefficient 0.95, Gaussian white noise is added on the channel, with: $N = M = 2$, $m = l = 2$, $o = 0$, $d_v = 1$, $d_s = 0$, and $\boldsymbol{C}(z) = \boldsymbol{I}$. 5_3 filter banks are used. The upper curve is OPTA. 85

4.7 The power constrained MIMO block model with noise added to the original signal. 86

4.8 The solid curve shows the SNR vs. CSNR performance using the proposed theory with the following parameters: $N = 3$, $M = 1$, $m = l = 4$, $d_v = 3$, $d_s = 1$, and the input PSD is an AR(1) source with correlation coefficient 0.9. The circle shows the performance of the system proposed in [Malvar 1986]. ISNR = 30.0 dB in both systems. 88

5.1 Block diagram of the communication system. 92

5.2 Preprocessing used in the modulated MIMO system. 98

5.3 Illustration of PSDs in branch number 0 in the preprocessor when an AR(3) input source is used with $N = 3$. (a) PSD of the input signal $S_x(f)$. (b) PSD of the output of the filter $B_0(f)$. (c) PSD of the output of the modulator $\Psi_0(\cdot)$. (d) PSD of the output of the decimator in branch number 0, which is equal to $\lambda_0^{(N)}(f)$. (e) $\lambda_0^{(N)}(fN)$. 99

5.4 SNR vs. CSNR performance for decimated MIMO is shown by the dashed curves, modulated MIMO by the dash-dotted curves, and BPAM system by the solid curves. For both the MIMO systems the following parameters have been used: $(N, M) = (2, 1)$ and $(L, K) \in \{(2, 1), (4, 2), (8, 4), (16, 8), (32, 16)\}$ for the BPAM system. In the results for the BPAM system (solid) the values for L and K are increasing when going from bottom to top. An AR(3) source with the spectrum shown in Figure 5.3 (a) is coded. 103

6.1 Subband coding of subband number i. 107

6.2 Characteristics of midtread uniform threshold quantizer number i. The quantizer step size is Δ_i and the number of representation levels are infinite. 108

6.3 The distortion rate performance of the optimized 5_3 filter banks from Section 4.1. The performance of the practical coder is shown by the solid curve, while the dashed curve shows the theoretical performance found from the theory developed in Section 4.1. The dotted curve gives the distortion rate function for the input signal, which is a Gaussian AR(1) signal with correlation factor 0.95. $N = 2$ subbands are used. 110

6.4 The distortion rate performance when using the 9_7 PR filter bank in [Antonini et al. 1992]. The performance of the practical coder is shown by the solid curve, while the dashed curve shows the theoretical performance given by the theory developed in Section 4.1. The dotted curve gives the distortion rate function for the input signal, which is a Gaussian AR(1) signal with correlation factor 0.95. $N = 2$ subbands are used. 111

6.5 The solid curve shows the distortion rate performance of the theoretical quantization model when coding one subband signal, and the dotted curve shows the practical coding performance when coding the same Gaussian subband signal. The coding coefficient is $\frac{\pi e}{6}$ in the theoretical model. 112

6.6 Different MSE contributions per source sample as a function of coding rate with $N = 2$ and $M = 1$. The optimized 5_3 filter banks from Section 4.1 are used. The dotted curves with diamonds show the MSE contributions for the practical coding system, and the solid curves with squares show the MSE contributions for the theoretical coding system. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95. . . 115

7.1 Subtractive dithering in subband number i. 122

7.2 The coding coefficient c_i, as a function of rate b_i, when coding a Gaussian time series with a uniform threshold quantizer using centroids as representation levels. 126

7.3 Illustration of the iterative numerical algorithm for optimizing the FIR filter banks when the cross-correlation between the input signal and the additive quantization noise is included. . . . 130

7.4 FIR PR filter bank model. 132

7.5 Different MSE contributions per source sample as a function of coding rate with $N = 2$ and $M = 1$. The dotted curves with $\times$-marks show the MSE contributions for the practical coding system with the redesigned quantizers, and the solid curves with circles show the MSE contributions for the practical coding system with centroids as representation levels in the quantizers. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95. The optimized 9_7 filter banks from Section 4.1 are used. 135

7.6 Different MSE contributions per source sample as a function of coding rate using subtractive dithering when $N = 2$ and $M = 1$. The dotted curves with $\times$-marks show the MSE contributions for the practical performance obtained with subtractive dithering, and the solid curves with circles show the theoretical MSE contributions. The filter banks used are the optimized 9_7 filter banks from Section 4.1. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95. 137

7.7 Different MSE contributions per source sample as a function of coding rate when $N = 2$ and $M = 1$. The dotted curves with $\times$-marks show the MSE contributions for the practical coding system, and the solid curves with circles show the MSE contributions for the theoretical coding system with the signal dependent colored quantization noise model. The dashed curves with squares show the MSE contributions for the 9_7 PR filter bank in [Antonini et al. 1992]. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95. The filter banks used are the optimized 9_7 filter banks from Section 7.2. 138
7.8 Theoretical distortion rate performance using uniform threshold quantizers having centroid representation levels (dotted), uniform threshold quantizers with scaled centroid representation levels (dash-dotted), and subtractive dithering (solid). A Gaussian time series is coded. 142

8.1 Venn-diagram of different sets of filter banks. 146
8.2 Basic concepts of research. 148

B.1 Ordering functions $l_i(fN)$ as a function of frequency f for $N = 3$, where the input PSD $S_x(f)$ is an AR(1) with correlation coefficient 0.95. 162

List of Tables

3.1 Theoretical distortion rate performances. 59

4.1 Pseudo code of the numerical optimization algorithm. 71
4.2 Theoretical distortion rate performances. 76

7.1 Practical distortion rate performances. 139

Nomenclature

$*$	convolution operator
*	z^* means the complex conjugation of the number z
$\emptyset$	empty set
$\setminus$	set difference
$\mid$	such that
α_i	non-zero scaling constant in subband number i
β	constant which depends on the average coding rate, quantization noise variances, and coding coefficients
$\boldsymbol{\Gamma}(f)$	$N \times N$ diagonal matrix with diagonal element number k given by $[\boldsymbol{\Gamma}(f)]_{k,k} = e^{-j2\pi f \frac{k}{N}}$
$\overline{\Gamma}(\gamma)$	the set of frequencies in the fundamental period $\left(-\frac{1}{2}, \frac{1}{2}\right]$ of total length γ giving the largest values of $S_x(f)$
$\underline{\Gamma}(\gamma)$	the set of frequencies in the fundamental period $\left(-\frac{1}{2}, \frac{1}{2}\right]$ of total length γ giving the smallest values of $S_x(f)$
$\Gamma_i^{(N)}$	the frequency region of length $\frac{1}{N}$ of the fundamental period $\left(-\frac{1}{2}, \frac{1}{2}\right]$ where $S_x(f)$ has the ith largest values
γ	length of a frequency interval, $\gamma \in [0, 1]$
Δ_i	quantizer step size in uniform threshold quantizer number i
$\delta(\cdot)$	Krönecker delta function
δ_g	Gâteaux variation operator
$\boldsymbol{\epsilon}(n)$	$N \times 1$ error vector between the input vector $\boldsymbol{x}(n)$ and output vectors $\hat{\boldsymbol{x}}(n)$
$\varepsilon_{L,K}^{\mathrm{BPAM}}(\mu)$	MSE per *source symbol* in the BPAM system, using K channel samples for L source samples and Lagrange multiplier μ

$\varepsilon_{N,M}^{\mathrm{MIMO}}(\mu)$	MSE per *source symbol* in the modulated MIMO system, using M channel samples for N source samples and Lagrange multiplier μ
$\zeta_k^{(N)}(f)$	element number k of the first row of the matrix $\boldsymbol{S}_{\boldsymbol{x}}(f)$, where $k \in \mathbb{Z}_N$
$\boldsymbol{\Theta}$	$M \times M$ diagonal matrix where diagonal element number i is θ_i
θ_i	non-negative Kuhn-Tucker parameter, where $i \in \mathbb{Z}_M$
$\kappa_i^{(M)}$	eigenvalue number i of the matrix $\boldsymbol{K}_{\boldsymbol{v}}(0)$, where $i \in \mathbb{Z}_M$
$\kappa_i^{(M)}(f)$	diagonal element number i of the diagonal matrix $\boldsymbol{\Lambda}_{\boldsymbol{v}}(f)$, where $i \in \mathbb{Z}_M$
$\boldsymbol{\Lambda}_{\boldsymbol{q}}$	diagonal $M \times M$ matrix containing the variances of the additive quantization noise
$\boldsymbol{\Lambda}_{\boldsymbol{v}}$	diagonal $M \times M$ matrix containing the eigenvalues of the matrix $\boldsymbol{K}_{\boldsymbol{v}}(0)$
$\boldsymbol{\Lambda}_{\boldsymbol{v}}^{-1}(f)$	diagonal $M \times M$ matrix which contains the eigenvalues of the matrix $\boldsymbol{C}^H(f)\boldsymbol{S}_{\boldsymbol{v}}^{-1}(f)\boldsymbol{C}(f)$
$\boldsymbol{\Lambda}_{\boldsymbol{x}}$	$N \times N$ diagonal matrix containing the eigenvalues of $\boldsymbol{K}_{\boldsymbol{x}}(0)$ in descending order
$\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)}$	$M \times M$ diagonal matrix containing the first M eigenvalues of $\boldsymbol{K}_{\boldsymbol{x}}(0)$ in descending order
$\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$	$N \times N$ diagonal matrix containing the eigenvalues of $\boldsymbol{S}_{\boldsymbol{x}}(f)$ in descending order
$\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$	$N \times N$ diagonal matrix given by $\boldsymbol{\Pi}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{\Pi}^H(f)$
$\bar{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$	identical to $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$ except for swapping of the jth and the kth diagonal element at frequencies in the set $\mathcal{F}_{j,k}$
$\lambda_i^{(N)}$	eigenvalue number i of the matrix $\boldsymbol{K}_{\boldsymbol{x}}(0)$, where $i \in \mathbb{Z}_N$
$\lambda_i^{(N)}(f)$	eigenvalue number i of the matrix $\boldsymbol{S}_{\boldsymbol{x}}(f)$, where $i \in \mathbb{Z}_N$
$\tilde{\lambda}_i^{(N)}(f)$	diagonal element number i of the matrix $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$, where $i \in \mathbb{Z}_N$
$\bar{\lambda}_i^{(N)}(f)$	diagonal element number i of the matrix $\bar{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$, where $i \in \mathbb{Z}_N$
μ	Lagrange multiplier
ν	small non-negative constant
$\boldsymbol{\Pi}(f)$	$N \times N$ frequency dependent permutation matrix

π	mathematical constant
$\rho_i(f)$	$e^{-j2\pi\frac{\left(f+l_i^{(N)}(f)\right)}{N}}$, where $i \in \mathbb{Z}_N$
$\boldsymbol{\Sigma_y}$	$M \times M$ diagonal matrix containing the variances of the subband signals
σ_q^2	variance of the noise added by the quantizers when all the quantization noise variances are equal
$\sigma_{q_i}^2$	variance of the noise added by quantizer number $i \in \mathbb{Z}_M$
σ_u^2	variance of the additive input noise
σ_v^2	variance of the additive channel noise
σ_x^2	variance of the input time series $x(n)$
$\sigma_{y_i}^2$	variance of subband signal number $i \in \mathbb{Z}_M$
$\tilde{\sigma}_{y_i}^2$	part of the subband variance when using the $\bar{G}_{i,i}(f)$ filters, Equation (2.45) gives the definition
$\boldsymbol{\Phi}$	$M \times M$ diagonal matrix which depends on the matrices $\boldsymbol{\Sigma_y}$ and $\boldsymbol{\Theta}$
$\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$	$(l+1)M \times (l+1)M$ autocorrelation matrix of the $(l+1)M \times 1$ vector $\boldsymbol{q}(n)_{\text{l}}$
$\boldsymbol{\Phi}_{\boldsymbol{u}}^{(p,N)}$	$(p+1)N \times (p+1)N$ autocorrelation matrix of the $(p+1)N \times 1$ vector $\boldsymbol{u}(n)_{\text{l}}$, where $p \in \{m, m+o+l\}$
$\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}$	$(l+1)M \times (l+1)M$ autocorrelation matrix of the $(l+1)M \times 1$ vector $\boldsymbol{v}(n)_{\text{l}}$
$\boldsymbol{\Phi}_{\boldsymbol{x}}^{(p,N)}$	$(p+1)N \times (p+1)N$ autocorrelation matrix of the $(p+1)N \times 1$ vector $\boldsymbol{x}(n)_{\text{l}}$, where $p \in \{0, m, m+l, m+o+l\}$
$\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$	$(m+l+1)N \times (l+1)M$ cross-correlation matrix between the input vector $\boldsymbol{x}(n)_{\text{l}}$ and the additive quantization vector $\boldsymbol{q}(n)_{\text{l}}$
$\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,M)}$	$(l+1)M \times (l+1)M$ autocorrelation matrix of the $(l+1)M \times 1$ vector $\hat{\boldsymbol{y}}(n)_{\text{l}}$
$\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s)$	$(l+1)M \times N$ cross-correlation matrix between the additive quantization vector $\boldsymbol{q}(n)_{\text{l}}$ and the input vector $\boldsymbol{x}_{d_s}(n-d_v)$
$\boldsymbol{\phi}_{\boldsymbol{x}}^{(p,N)}(d_v, d_s)$	$(p+1)N \times N$ covariance matrix, which is a submatrix of the $(p+1)N \times (p+1)N$ autocorrelation matrix $\boldsymbol{\Phi}_{\boldsymbol{x}}^{(p,N)}$, the vector delay d_v and scalar delay d_s decide how this submatrix can be found, and $p \in \{m+l, m+o+l\}$

$\boldsymbol{\phi}_{\boldsymbol{x},\hat{\boldsymbol{y}}}^{(l,N,M)}(d_v, d_s)$	$N \times (l+1)M$ covariance matrix between the $N \times 1$ vector $\boldsymbol{x}_{d_s}(n-d_v)$ and the $(l+1)M \times 1$ vector $\hat{\boldsymbol{y}}(n)$
ϕ_i	diagonal element number i of the matrix $\boldsymbol{\Phi}$
$\chi_{q_i}(a)$	characteristic function of the stochastic variable $q_i(n)$ evaluated at the argument a
$\chi_{x_i(n),q_k(m)}(a,b)$	joint characteristic function of the stochastic variables $x_i(n)$ and $q_k(m)$ evaluated at the arguments a and b
$\boldsymbol{\Psi}(f)$	$N \times N$ DFT matrix with columns permuted according to the ordering functions $l_i^{(N)}(f)$
$\Psi_i(\cdot)$	modulator number i in the preprocessor
$\boldsymbol{\mathcal{A}}_{\_}$	$M \times (m+1)N$ matrix containing ones at the positions corresponding to where the analysis filter bank $\boldsymbol{E}_{\_}$ contains free parameters and zeros where $\boldsymbol{E}_{\_}$ must contain zeros
$A_i(f)$	ith alias function in a filter bank, where $i \in \mathbb{Z}_N$
a	real variable
$a_k^{(i)}$	difference between the centroid and the midpoint in decision interval number k in the ith uniform threshold quantizer
a_i	real variable number i
$B_i(f)$	frequency response of the filter number i in the preprocessor
b	1) average number of bits used for coding N source samples 2) real variable
b_i	number of bits used in quantizer number i, where $i \in \mathbb{Z}_M$
b'	constant used to simplify the bit allocation expressions
^{C}	$\mathcal{F}_{j,k}^C$ is the complementary set of $\mathcal{F}_{j,k}$ in the frequency interval $\left(-\frac{1}{2}, \frac{1}{2}\right]$
$C(z)$	scalar channel transfer function
$\boldsymbol{C}(z)$	$M \times M$ transfer matrix of the MIMO channel
$\boldsymbol{C}_{\_}$	$M \times (o+1)M$ row-expansion of the FIR channel matrix $\boldsymbol{C}(z)$
$(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\searrow}$	$(l+1)M \times (m+o+l+1)N$ matrix used to express the total polyphase matrix in the power constrained FIR problem
$\boldsymbol{c}(n)$	$M \times M$ impulse response matrix sequence of the MIMO channel
c_i	coding coefficient for coding subband number $i \in \mathbb{Z}_M$
$D_{i,i}$	ith diagonal element of the matrix $\boldsymbol{D}$, where $i \in \mathbb{Z}_M$

$\boldsymbol{D}$	$M \times M$ diagonal matrix used in the Wiener transform when the analysis transform is a reduced rank KLT matrix
$\boldsymbol{D}(f)$	$M \times M$ diagonal matrix
d_s	scalar delay through the FIR analysis and synthesis filter bank
d_v	vector delay through the FIR analysis and synthesis filter bank
$\mathcal{E}(f)$	the integrand in the block MSE $\mathcal{E}_{N,M}$ at the frequency f, when using the $\tilde{G}_{i,i}(f)$ filters
$\bar{\mathcal{E}}(f)$	the integrand in the block MSE $\mathcal{E}_{N,M}$ at the frequency f, when using the $\bar{G}_{i,i}(f)$ filters
$\mathcal{E}_{N,M}$	block MSE by coding the vector $\boldsymbol{x}(n)$ using unconstrained length filter banks or transform matrices
$\mathcal{E}_{N,M}\,(d_v, d_s)$	block MSE by coding the vector $\boldsymbol{x}(n)$ using FIR filter banks with vector delay d_v and scalar delay d_s
$\mathcal{E}^{(\boldsymbol{q})}_{N,M}\,(d_v, d_s)$	quantization block MSE by coding the vector $\boldsymbol{x}(n)$ using FIR filter banks with vector delay d_v and scalar delay d_s
$\mathcal{E}^{(\boldsymbol{x})}_{N,M}\,(d_v, d_s)$	signal block MSE by coding the vector $\boldsymbol{x}(n)$ using FIR filter banks with vector delay d_v and scalar delay d_s
$\mathcal{E}^{(\boldsymbol{x},\boldsymbol{q})}_{N,M}\,(d_v, d_s)$	crossterm block MSE contribution by coding the vector $\boldsymbol{x}(n)$ using FIR filter banks with vector delay d_v and scalar delay d_s
E	expected value operator
$E_{m,n}(z)$	element in row number m and column number n of the matrix $\boldsymbol{E}(z)$
$\boldsymbol{E}$	$M \times N$ analysis matrix which is used in the transform coder
$\boldsymbol{E}(z)$	$M \times N$ analysis polyphase matrix
$\boldsymbol{E}_{\_}$	$M \times (m+1)N$ row-expansion of the FIR analysis polyphase matrix $\boldsymbol{E}(z)$
$\hat{\boldsymbol{E}}_{\_}$	$M \times (m+o+1)N$ row-expansion of the combination of the transmitter FIR polyphase matrix $\boldsymbol{E}(z)$ and the channel FIR polyphase matrix $\boldsymbol{C}(z)$

$\boldsymbol{E}_{\lceil}$	1) $(l+1)M \times (l+m+1)N$ block matrix used to express the convolution between the FIR analysis and synthesis polyphase matrices used in bit constrained filter banks 2) $(o+1)M \times (o+m+1)N$ block matrix used to express the convolution between the FIR analysis and channel polyphase matrices in power constrained filter banks
e	base of natural logarithm
$\boldsymbol{e}(n)$	$M \times N$ analysis impulse response matrix of lag n
$\hat{\boldsymbol{e}}(n)$	$M \times N$ impulse response matrix of lag n of the convolution of the transmitter FIR polyphase matrix $\boldsymbol{E}(z)$ and channel FIR polyphase matrix $\boldsymbol{C}(z)$
$\mathrm{ED}_i(\cdot)$	entropy decoder operating on subband number i, where $i \in \mathbb{Z}_M$
$\mathrm{EE}_i(\cdot)$	entropy encoder operating on subband number i, where $i \in \mathbb{Z}_M$
$\mathcal{F}_{j,k}$	set of frequencies in $\left(-\frac{1}{2}, \frac{1}{2}\right]$ satisfying Equation (2.35), which depends on the indices j and k
$\mathcal{F}_{j,k}^{(i)}$	$i \in \{1,2\}$, the sets $\mathcal{F}_{j,k}^{(1)}$ and $\mathcal{F}_{j,k}^{(2)}$ are defined in Equations (2.39) and (2.40), respectively
$F_i(f)$	synthesis filter number i, where $i \in \mathbb{Z}_M$
f	1) relative frequency 2) continuous real integration variable
$f_{\boldsymbol{y}(n)}$	pdf of subband vector signal $\boldsymbol{y}(n)$
$f_{y_i}(\cdot)$	pdf of subband signal y_i, where $i \in \mathbb{Z}_M$
$f_{x_i(n),q_k(m)}(\cdot,\cdot)$	joint two-dimensional pdf of the stochastic variables $x_i(n)$ and $q_k(m)$
$G_{m,m}$	diagonal element number m of the matrix $\boldsymbol{G}$
$G_{m,n}(f)$	element in row number m and column number n of the matrix $\boldsymbol{G}(f)$
$\tilde{G}_{i,i}(f)$	diagonal element number i of the matrix $\tilde{\boldsymbol{G}}(f)$
$\bar{G}_{i,i}(f)$	filters used in connection with $\bar{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$
$\boldsymbol{G}$	$M \times N$ analysis matrix used in the transform coder
$\boldsymbol{G}(f)$	$M \times N$ analysis polyphase matrix in the equivalent system
$\tilde{\boldsymbol{G}}(f)$	$M \times N$ analysis polyphase matrix with non-zero elements only on the main diagonal

$g(\cdot,\cdot,\cdot,\cdot,\cdot)$	function that is used to express some of the terms that are part of an integral
^{H}	$\boldsymbol{A}^H$ is the conjugate transpose of the matrix $\boldsymbol{A}$
$H_i(f)$	analysis filter number i, where $i \in \mathbb{Z}_M$
$\mathcal{I}$	index set used in the estimation of MSE contributions in the practical subband coder
$\boldsymbol{I}$	identity matrix
$\boldsymbol{I}_p$	the $p \times p$ identity matrix
i	index
j	1) imaginary unit 2) index
K	number of channel samples representing L source samples in the BPAM system
$\boldsymbol{K}_{\boldsymbol{q}}(m)$	$M \times M$ autocorrelation matrix for the quantization noise vector $\boldsymbol{q}(n)$ at lag m
$\boldsymbol{K}_{\boldsymbol{v}}(m)$	$M \times M$ autocorrelation matrix for the channel noise vector $\boldsymbol{v}(n)$ at lag m
$\boldsymbol{K}_{\boldsymbol{x}}(m)$	$N \times N$ autocorrelation matrix for the source vector $\boldsymbol{x}(n)$ at lag m
$\boldsymbol{K}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(0)$	$N \times M$ cross-correlation matrix between the vectors $\boldsymbol{x}(n)$ and $\hat{\boldsymbol{y}}(n)$
$\boldsymbol{K}_{\boldsymbol{y}}(m)$	$M \times M$ autocorrelation matrix for the vector $\boldsymbol{y}(n)$ at lag m
$\boldsymbol{K}_{\hat{\boldsymbol{y}}}(m)$	$M \times M$ autocorrelation matrix for the vector $\hat{\boldsymbol{y}}(n)$ at lag m
k	index
L	1) number of source samples in a source vector in the BPAM system 2) number of levels used in a scalar quantizer
$\mathcal{L}$	Lagrange function
l	1) order of the FIR synthesis polyphase matrix 2) index
$l_i^{(N)}(f)$	ordering function number i, where $i \in \mathbb{Z}_N$
ℓ_{F_i}	filter length of synthesis filter $F_i(f)$, where $i \in \mathbb{Z}_M$
ℓ_{H_i}	filter length of analysis filter $H_i(f)$, where $i \in \mathbb{Z}_M$

M	1) number of quantizers receiving a positive number of bits 2) number of inputs and outputs of the channel in a MIMO system
m	1) time and summation index 2) order of the FIR analysis polyphase matrix $\boldsymbol{E}(z)$
N	number of source samples in the source vector $\boldsymbol{x}(n)$
$\mathbb{N}$	natural numbers $\{0, 1, 2, \dots\}$
n	time and summation index
o	order of the FIR channel polyphase matrix $\boldsymbol{C}(z)$
P	power used by the input vector $\boldsymbol{y}(n)$ to the channel
$P_k^{(i)}$	estimate of the probability for the index k to occur in the ith quantizer
$\mathcal{P}_{L,K}^{\text{BPAM}}(\mu)$	power used per *source symbol* in the BPAM system, using K channel samples for L source samples and Lagrange multiplier μ
$\mathcal{P}_{N,M}^{\text{MIMO}}(\mu)$	power used per *source symbol* in the modulated MIMO system, using M channel samples for N source samples and Lagrange multiplier μ
Pr	product of the diagonal elements of a matrix
p	1) summation index 2) positive integer 3) number of times the inequality $\sigma_{y_i}^2 \geq \sigma_q^2$ holds with equality
p_1	summation index
p_2	summation index
$Q_i(\cdot)$	quantizer operating on the samples in subband number i, where $i \in \mathbb{Z}_M$
$Q_i^{-1}(\cdot)$	inverse quantizer operating on the quantization indices in subband number i, where $i \in \mathbb{Z}_M$
$q_i(n)$	component number i of the vector $\boldsymbol{q}(n)$, where $i \in \mathbb{Z}_M$
$\boldsymbol{q}(n)$	$M \times 1$ vector containing the additive quantization noise of the M quantizers
$\boldsymbol{q}(n)_{\text{I}}$	column-expanded quantizer vector of dimension $(l+1)M \times 1$
$R_x(m)$	autocorrelation function of the input signal $x(n)$ at lag m
$R_{x_i,y_k}(n-m)$	cross-correlation function between $x_i(n)$ and $y_k(m)$

$R_{y_i,y_k}(n-m)$	cross-correlation function between $y_i(n)$ and $y_k(m)$
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^+$	the set $(0,\infty)$
$R_{n,m}(z)$	element in row number n and column number m of the matrix $\boldsymbol{R}(z)$
$\boldsymbol{R}$	$N \times M$ synthesis matrix which is used in the transform coder
$\boldsymbol{R}(z)$	$N \times M$ synthesis polyphase matrix
$\boldsymbol{R}_{\text{Wiener}}(f)$	$N \times M$ Wiener synthesis polyphase matrix
$\boldsymbol{R}_{\mathsf{I}}$	$(l+1)N \times M$ column-expansion of the FIR synthesis polyphase matrix $\boldsymbol{R}(z)$
$\boldsymbol{R}_{\_}$	$N \times (l+1)M$ row-expanded FIR synthesis matrix
$\hat{\boldsymbol{R}}(z)$	$N \times N$ synthesis polyphase matrix which is part of an FIR PR filter bank
$\hat{\boldsymbol{R}}_{\_}$	$N \times (l+1)N$ row-expanded FIR synthesis matrix, which is part of an FIR PR filter bank
$\boldsymbol{r}(n)$	$N \times M$ synthesis impulse response matrix at lag n
$r_k^{(i)}$	representation level number k in uniform threshold quantizer number i
$\hat{r}_k^{(i)}$	representation level number k in uniform threshold quantizer number i which has uncorrelated input and additive quantization noise
$S_x(f)$	power spectral density function of the input signal $x(n)$
$\boldsymbol{S}_{\boldsymbol{v}}(f)$	$M \times M$ power spectral density matrix of the additive noise vector $\boldsymbol{v}(n)$
$\boldsymbol{S}_{\boldsymbol{x}}(f)$	$N \times N$ power spectral density matrix of the source vector $\boldsymbol{x}(n)$
$\boldsymbol{S}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(f)$	$N \times M$ cross power spectral density matrix of the vectors $\boldsymbol{x}(n)$ and $\hat{\boldsymbol{y}}(n)$
$\boldsymbol{S}_{\hat{\boldsymbol{y}}}(f)$	$M \times M$ power spectral density matrix of the vector $\hat{\boldsymbol{y}}(n)$
$\boldsymbol{\mathcal{S}}_{\mathsf{I}}$	$(l+1)N \times M$ matrix containing ones at the positions corresponding to where the synthesis filter bank $\boldsymbol{R}_{\mathsf{I}}$ contains free parameters and zeros where $\boldsymbol{R}_{\mathsf{I}}$ must contain zeros
$\boldsymbol{\mathcal{S}}_{\_}$	$N \times (l+1)M$ matrix containing ones at the positions corresponding to where the synthesis filter bank $\boldsymbol{R}_{\_}$ contains free parameters and zeros where $\boldsymbol{R}_{\_}$ must contain zeros

^{T}	$\boldsymbol{A}^T$ is the transpose of the matrix $\boldsymbol{A}$
$T_{i,i}$	diagonal element number i of the matrix $\boldsymbol{T}$
$T_{n,m}(f)$	element in row number n and column number m of the matrix $\boldsymbol{T}$
$\boldsymbol{T}$	$N \times M$ synthesis transform matrix in the transform coder
$\boldsymbol{T}(f)$	$N \times M$ synthesis polyphase matrix in the equivalent system
$\tilde{\boldsymbol{T}}(f)$	$N \times M$ synthesis polyphase matrix with non-zero elements only on the main diagonal
$\mathcal{T}\{\cdot\}$	operator which produces an $(l+1)N \times (m+1)N$ block Toeplitz matrix from an $N \times (m+l+1)N$ matrix
$\mathcal{T}_1\{\cdot\}$	operator which produces an $(l+1)N \times (m+o+1)N$ block Toeplitz matrix from an $N \times (m+o+l+1)N$ matrix
$\mathcal{T}_2\{\cdot\}$	operator which produces an $(o+1)M \times (m+1)N$ block Toeplitz matrix from an $M \times (m+o+1)N$ matrix
Tr	trace of a matrix
$\boldsymbol{U}$	$N \times N$ matrix containing the eigenvectors of $\boldsymbol{K}_{\boldsymbol{x}}(0)$
$\boldsymbol{U}_{(M)}$	$N \times M$ matrix containing the first M eigenvectors of $\boldsymbol{K}_{\boldsymbol{x}}(0)$
$\boldsymbol{U}(f)$	$N \times N$ matrix containing the eigenvectors of $\boldsymbol{S}_{\boldsymbol{x}}(f)$
$\boldsymbol{U}_i(f)$	$N \times 1$ eigenvector number i of $\boldsymbol{S}_{\boldsymbol{x}}(f)$ corresponding to eigenvalue $\lambda_i^{(N)}(f)$
$\boldsymbol{u}(n)$	$N \times 1$ vector containing noise which is added to the original signal vector
$\boldsymbol{u}(n)_{\mathbf{I}}$	column-expanded vector containing noise which is added to the original signal vector of dimension $(p+1)N \times 1$, where $p \in \{m, m+o+l\}$
$\boldsymbol{V}$	$M \times M$ matrix containing the eigenvectors of the matrix $\boldsymbol{K}_{\boldsymbol{v}}^{-1}$
$\boldsymbol{V}(f)$	$M \times M$ matrix containing the eigenvectors of the matrix $\boldsymbol{C}^H(f)\boldsymbol{S}_{\boldsymbol{v}}^{-1}(f)\boldsymbol{C}(f)$
$v_i(n)$	component number i of the vector $\boldsymbol{v}(n)$, where $i \in \mathbb{Z}_M$
v_1	continuous real integration variable
v_2	continuous real integration variable
$\boldsymbol{v}(n)$	1) $K \times 1$ vector containing the additive channel noise 2) $M \times 1$ vector containing the additive channel noise

$\boldsymbol{v}(n)_{\shortmid}$	column-expanded channel noise vector of dimension $(l+1)M \times 1$
$\boldsymbol{W}(f)$	$N \times N$ polyphase matrix through the overall system
$\boldsymbol{W}_{\text{-}}$	1) $N \times (m+l+1)N$ row-expanded FIR overall transfer matrix 2) $M \times (m+o+1)N$ row-expanded FIR matrix used to define the operator $\mathcal{T}_2$
$\boldsymbol{w}(n)$	$N \times N$ synthesis impulse response matrix sequence through the overall system
$w_i(n)$	uniformly distributed pseudo-random sequence over the interval $\left(-\frac{\Delta_i}{2}, \frac{\Delta_i}{2}\right)$ used in subtractive dithering of subband number i, where $i \in \mathbb{Z}_M$
x	continuous real variable
$x(n)$	time series to be compressed or transmitted
$\hat{x}(n)$	reconstructed time series
$x_i(n)$	component number i of the vector $\boldsymbol{x}(n)$, where $i \in \mathbb{Z}_N$
$\hat{x}_i(n)$	component number i of the vector $\hat{\boldsymbol{x}}(n)$, where $i \in \mathbb{Z}_N$
$\hat{x}_{\text{quant}}(n)$	part of the reconstructed time series $\hat{x}(n)$ generated by the additive quantization noise
$\hat{x}_{\text{sig}}(n)$	part of the reconstructed time series $\hat{x}(n)$ put out by the synthesis filter bank when the quantizers are removed
$\boldsymbol{x}(n)$	1) $L \times 1$ vector containing L source samples 2) $N \times 1$ vector containing N source samples
$\boldsymbol{x}_{d_s}(n)$	$N \times 1$ vector containing N consecutive source samples, see Equation (4.5) for the definition
$\hat{\boldsymbol{x}}(n)$	1) $L \times 1$ vector containing L reconstructed samples 2) $N \times 1$ vector containing N reconstructed samples
$\boldsymbol{x}(n)_{\shortmid}$	column-expanded vector of dimension $(p+1)M \times 1$, where $p \in \{m, m+l, m+o+l\}$
$y_i(n)$	component number i of the vector $\boldsymbol{y}(n)$, where $i \in \mathbb{Z}_M$
$\hat{y}_i(n)$	component number i of the vector $\hat{\boldsymbol{y}}(n)$, where $i \in \mathbb{Z}_M$
$\boldsymbol{y}(n)$	1) $K \times 1$ vector containing the K subband signal coefficients 2) $M \times 1$ vector containing the M subband signal coefficients
$\boldsymbol{y}(n)_{\shortmid}$	column-expanded vector of dimension $(l+1)M \times 1$

$\hat{\boldsymbol{y}}(n)$	1) $K \times 1$ vector containing the M reconstructed subband signal coefficients 2) $M \times 1$ vector containing the M reconstructed subband signal coefficients
$\hat{\boldsymbol{y}}(n)_{\mathsf{I}}$	column-expanded vector of dimension $(l+1)M \times 1$
$\mathbb{Z}$	the set of integers
$\mathbb{Z}_N$	the set $\{0, 1, \ldots, N-1\}$
z	the variable in the z transform

List of Abbreviations

AR	autoregressive
BPAM	block pulse amplitude modulation
bit	binary digit
CCITT	International Telegraph and Telephone Consultative Committee
CSNR	channel signal to noise ratio
DCT	discrete cosine transform
DSP	digital signal processing
dB	decibel
FIR	finite impulse response
FFT	fast Fourier transform
IEC	International Electrotechnical Commission
IEEE	The Institute of Electrical and Electronics Engineers
ISNR	input signal to noise ratio
ISO	International Standards Organization
IS	Interim Standard
i.i.d.	independent and identically distributed
KLT	Karhunen-Loève transform
MIMO	multiple input multiple output
MSE	mean square error
OPTA	optimal performance theoretically attainable
pdf	probability density function
PR	perfect reconstruction
PSD	power spectral density
SISO	single input single output
SNR	signal to noise ratio
URL	uniform resource locator
WSS	wide sense stationary

Chapter 1

Introduction

The number of people with access to the Internet and mobile communication units is increasing rapidly. Networks and data terminals have become suitable for multimedia processing, and their processing power is increasing steadily. In multimedia units, very large amounts of data can be generated, received, and processed. Thus, the trend is that the amount of data to be stored and transmitted is increasing, and the necessity of efficient signal compression and communication is ever more evident.

In communication systems, these large amounts of data are sent over a channel with limited capacity. Examples of channels are twisted pairs, coaxial cables, satellite and terrestrial wireless communication links, and optical fibers. The capacity of a channel is given by the available bandwidth, the noise level, and the maximum allowed transmitter power. When the signal is sent over a noisy channel, it is important to design the communication system such that it optimally utilizes the available power and bandwidth resources. Additionally, the source and channel characteristics must be taken into consideration when the system is optimized.

Digital signal processing (DSP) is an important area in the development of modern technologies, such as: Communication, multimedia, and Internet systems. Multirate filter banks are one of the tools of DSP. Filter banks will be studied in this dissertation in order to find efficient algorithms for data compression and communication.

Applications for filter banks can be found in compression, communications, filtering, restoration, signal analysis, etc. Filter banks are popular in compression and communication applications because most signal sources, e.g. audio signals, still images, and video signals, are highly correlated. Signal decomposition in terms of filter banks can partly decorrelate and adapt the signal for subsequent quantization or transmission over a channel. This processing

of the signal can make the compression or communication of the signal more efficient.

1.1 Scope of the Dissertation

Filter banks or transform coders are important constituents in source compression algorithms, and they can also be used in communication problems. Most of the theory developed in the literature concentrates on optimizing these systems while imposing the perfect reconstruction (PR) property [Vaidyanathan 1993] on the filter banks and transforms. In this dissertation, the PR constraint is relaxed, and the analysis and synthesis filter banks and transforms are jointly optimized to minimize the mean square error (MSE) between the reconstructed and original signal for a given bit rate or channel bandwidth, noise, and transmitter power. If the PR condition is dropped, better distortion rate performance can be achieved. The reason is that the often employed PR condition reduces the number of free parameters that can be used in the filter bank optimization. In the proposed filter banks and transforms, where no PR constraint is imposed, the set of filter banks and transforms used in the optimization includes the PR filter banks and transforms as a subset. Thus, the results from the optimization will be at least as good as all PR filter banks and transforms for all rates and for all sources when the filter lengths are the same in both systems. For high rates or very good channels, the optimal solution should be close to PR, but for low rates or poor channels this is not necessarily the case. Uniform filter banks are treated in this dissertation and therefore, the analysis filter bank structure will generate maximally decimated and equal bandwidth subbands.

1.1.1 Filter Banks for Compression

Filter banks and transforms are parts of a subband coder, and their operation is as follows. The *analysis* filter bank or the memoryless analysis transform matrix using decimation factor N decomposes the input signal such that N source samples produce M *subband coefficients* or *transform coefficients*. The subband signals or transform coefficients are then encoded using source coding, e.g., bit allocation with scalar quantizers, entropy constrained scalar quantization, etc. In the decoder, the M approximate subband coefficients or transform coefficients are derived from the bit representation. For every block of M approximate subband coefficients that are applied to the *synthesis* filter bank or transform matrix, N approximate source samples are reconstructed.

Two classes of filter banks possessing the PR property are the unitary and

the biorthogonal filter banks. Unitary filter banks are a subset of biorthogonal filter banks. Hence, biorthogonal filter banks perform at least as well as unitary filter banks. If the PR constraint is relaxed and no constraints are imposed on the filter banks and transforms, the most general class is obtained. This class of filter banks and transforms is used in this dissertation.

The basic assumptions are as follows: The system consists of uniform filter banks or transforms, the commonly used high rate model for the scalar quantizer [Jayant & Noll 1984] is used for all rates, and the input signal is modeled as a wide sense stationary (WSS) time series with a *known* power spectral density (PSD) function.

The goal is to optimize the distortion rate performance of this system for three different cases: Unconstrained length filter banks, transform coders, and finite impulse response (FIR) filter banks.

In the first case, the filters are allowed to be non-causal with unconstrained filter lengths, implying that the filters might be non-realizable. The results provide upper bounds for the performance of filter banks working under the additive white signal independent noise model, which is equal to the high rate model used for scalar quantizers.

The second case is a transform coder system. The Karhunen-Loève transform (KLT) is optimal in the distortion rate sense when the synthesis matrix is restricted to the inverse of the analysis matrix [Gersho & Gray 1992]. If the PR assumption is relaxed, it is shown that better system performance can be achieved. The jointly optimal analysis and synthesis transform matrix in the distortion rate sense is found.

In the third case, FIR filter banks are treated. An iterative numerical optimization algorithm for jointly optimizing the FIR analysis and synthesis filter banks is proposed. Causal FIR filters are assumed to be used, but the same methodology could be used for non-causal and anti-causal FIR filters. The solutions based on signal-adaptive jointly optimized transforms and unconstrained length filter banks give a lower and upper bound, respectively, for the signal to noise ratio (SNR) vs. channel signal to noise ratio (CSNR) performance of the FIR filter banks that are found.

When using the high rate quantization model, the FIR filter banks are designed under the assumption that the subband signals are uncorrelated with the additive coding noise. It will be shown that this assumption is not correct for low rates, causing a mismatch between the theoretical performance and the performance found by a practical subband coder. Therefore, quantizers having uncorrelated input and additive quantization noise are proposed, and subtractive dithering is used as a method of coding the subbands. Filter banks are also designed under the assumption that the subband signals and the ad-

ditive coding noise are correlated, and therefore, a signal dependent colored quantization noise model is introduced.

1.1.2 Filter Banks for Communication

The problem of transmitting a vector time series with continuous amplitude vectors over a continuous amplitude vector channel is investigated. It is assumed that the channel is of discrete time and corrupted by additive signal independent noise with zero mean and known second order statistics, but the probability density function (pdf) of the noise is arbitrary. Furthermore, it is assumed that the channel input vectors can only have limited power, and that the vector channel transfer function is known. The input vector time series is assumed to be of discrete time and to have known second order statistics with arbitrary pdf.

The dimension N of the vectors in the original time series and the dimension M of the input time series to the channel may in general be different, resulting in a discrete time multiple input multiple output (MIMO) system. The transmitter and receiver are represented by polyphase matrices.

For given values of N and M, the transmitter and receiver will be jointly optimized with respect to the block MSE between the system input and output vector time series, subject to a channel power constraint. As in the bit constrained problem, three different cases are treated.

In the first case, the filters are assumed to have infinite lengths, and they are allowed to be non-causal. The discrete time jointly optimal transmitter and receiver filter banks are deduced from the corresponding solution for continuous time. For the unconstrained case, the channel transfer matrix can have unconstrained order as well. Explicit expressions are found for the unconstrained case.

Transforms are also considered for the problem of communication of vector time series. This problem has previously been treated in [Lee & Petersen 1976]. An alternative proof for finding jointly optimal transmitter and receiver based on the unconstrained length solution is proposed, and explicit expressions are found. In the transform case, it is assumed that the channel transfer matrix is given by the identity matrix.

In the third case, FIR transmitter and receiver filter banks are jointly optimized, and an iterative numerical algorithm is proposed based on formulas for finding the optimal transmitter polyphase matrix for a given receiver matrix, and vice versa. The channel transfer matrix is assumed to have a finite order in the FIR case.

The optimized block based MIMO systems can be used to solve a combined source-channel coding problem where a continuous amplitude discrete

time scalar time series is transmitted over a continuous amplitude discrete time scalar channel. When working with this problem, compression is defined by sample reduction from the signal to the channel representation, that is when $M < N$. Very good results have been obtained for identity channel matrix and Gaussian white channel noise by using nonlinear methods for the compression case [Vaishampayan 1989, Fuldseth & Ramstad 1997]. In this dissertation however, only linear systems are examined, and the minimum MSE linear filter bank solution is found.

1.2 Problems Considered and Basic Assumptions

As mentioned in the two previous subsections, three different cases of the filter lengths in the linear filter banks are treated in this dissertation. However, in *this introduction chapter* no restrictions are imposed on the filters, i.e., they are allowed to be non-causal with infinitely long impulse responses. This means that they might not always be implementable and that the frequency responses can be equal to zero in certain frequency intervals. Since this is the most general case, the results obtained in this section can later be specialized to transform coders. However, a different treatment will be needed for FIR filter banks. Since causal FIR filters are assumed, the delay through the FIR filter banks must be taken into consideration.

The infinite length filters assumed in this section will be called *unconstrained length* filters because infinite impulse response (IIR) is usually preserved for implementable filters, i.e., filters which can be described by rational transfer functions.

1.2.1 Bit Constrained Filter Banks

The bit constrained filter bank model considered in this dissertation is shown in Figure 1.1, where the analysis and synthesis filter banks are given in polyphase form [Bellanger, Bonnerot & Coudreuse 1976]. The analysis and synthesis polyphase matrices are denoted $\boldsymbol{E}(z)$ and $\boldsymbol{R}(z)$, respectively. The analysis filter in subband number m can be expressed as

$$H_m(z) = \sum_{n=0}^{N-1} E_{m,n}\left(z^N\right) z^{-n}, \tag{1.1}$$

where $E_{m,n}(z)$ is the element in row number m and column number n of the analysis polyphase matrix $\boldsymbol{E}(z)$. The numbering of the rows and columns starts with zero. In the same way, the synthesis filter in subband number m

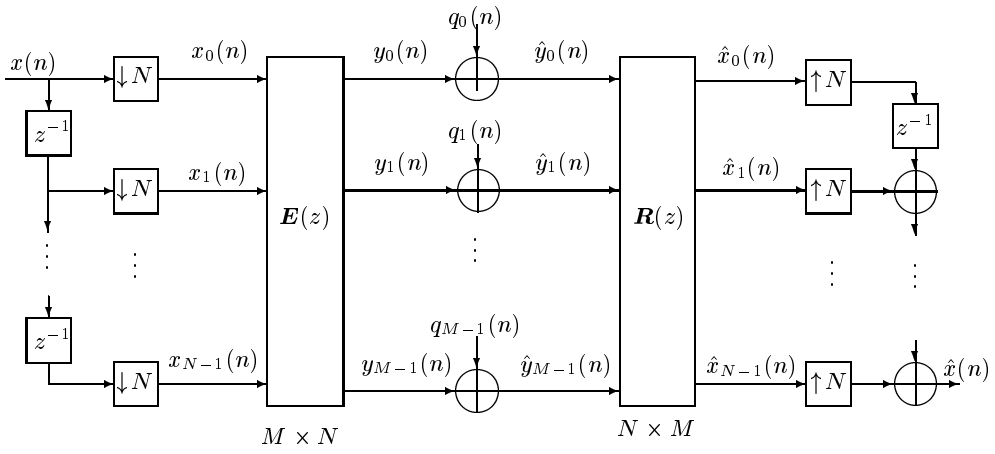

Figure 1.1 Subband coder model.

can be expressed as

$$F_m(z) = \sum_{n=0}^{N-1} R_{n,m}\left(z^N\right) z^{-(N-1-n)}, \tag{1.2}$$

where $R_{n,m}(z)$ is the element in row number n and column number m of the synthesis polyphase matrix. M is the number of quantizers receiving a positive number of bits. Therefore, $N - M$ quantizers do not receive any bits and their output is zero. In general, the polyphase matrices are rectangular, and it is assumed that $N \geq M$ when unconstrained length filter banks and transforms with a bit constraint are studied.

The quantizers are modeled as additive noise sources described by the noise vector

$$\boldsymbol{q}(n) = [q_0(n), q_1(n), \ldots, q_{M-1}(n)]^T. \tag{1.3}$$

In the first quantization noise model studied, it is assumed that the quantizer noise can be modeled as signal independent zero mean white noise [Jayant & Noll 1984], and that the noise in one subband channel is uncorrelated with the noise in all the other subband channels. Therefore, the $M \times M$ PSD matrix of the noise $\boldsymbol{\Lambda_q}$ is a diagonal frequency independent matrix, given by

$$\boldsymbol{\Lambda_q} = E\left[\boldsymbol{q}(n)\boldsymbol{q}^H(n)\right], \tag{1.4}$$

where the superscript ^{H} and the operator $E[\cdot]$ represent the Hermitian, i.e., transpose conjugate and expectation operators, respectively. The matrix $\boldsymbol{\Lambda_q}$ is also equal to the autocovariance matrix of the $\boldsymbol{q}(n)$ vector at lag zero. For non-zero lags, the quantization noise autocovariance matrix $\boldsymbol{K_q}(m)$ is equal to zero. Therefore, $\boldsymbol{K_q}(m)$ can be expressed as: $\boldsymbol{K_q}(m) = \delta(m)\boldsymbol{\Lambda_q}$. In Chapter 7, a signal dependent colored noise model is introduced.

The output time series $\hat{x}(n)$ of the subband coder system shown in Figure 1.1 is cyclostationary [Sathe & Vaidyanathan 1993]. By studying a more general vector system, a MIMO system, the problem becomes more tractable. In the MIMO system the inputs are given by vector time series, and these are *stationary vector time series* if $x(n)$ is a WSS time series [Sathe & Vaidyanathan 1993].

The input vector signal $\boldsymbol{x}(n)$ to the polyphase matrix $\boldsymbol{E}(z)$ and the output vector signal $\hat{\boldsymbol{x}}(n)$ of the polyphase matrix $\boldsymbol{R}(z)$ are given by

$$\begin{aligned} \boldsymbol{x}(n) &= [x_0(n), x_1(n), \ldots, x_{N-1}(n)]^T \text{ and} \\ \hat{\boldsymbol{x}}(n) &= [\hat{x}_0(n), \hat{x}_1(n), \ldots, \hat{x}_{N-1}(n)]^T. \end{aligned} \tag{1.5}$$

Using the decimating and interpolating structure shown in Figure 1.1, the following relations can be derived:

$$x_i(n) = x(nN - i) \quad \text{and} \quad \hat{x}_i(n) = \hat{x}(nN + N - 1 - i), \tag{1.6}$$

where $i \in \{0, 1, \ldots, N-1\}$.

The subband signals in Figure 1.1 are represented by the $M \times 1$ vector $\boldsymbol{y}(n) = [y_0(n), y_1(n), \ldots, y_{M-1}(n)]^T$ given by

$$\boldsymbol{y}(n) = \boldsymbol{e}(n) * \boldsymbol{x}(n), \tag{1.7}$$

where $*$ is the convolution operator, and $\boldsymbol{e}(n)$ is the $M \times M$ impulse response matrix sequence of the analysis polyphase filter bank. The z-transform of $\boldsymbol{e}(n)$ is equal to $\boldsymbol{E}(z)$.

The vector time series $\boldsymbol{x}(n)$ and $\boldsymbol{q}(n)$ are assumed to represent jointly WSS vector series [Sathe & Vaidyanathan 1993], which are mutually uncorrelated and have zero mean.

A partial statistical description of the signal $\boldsymbol{x}(n)$ is given by the following $N \times N$ PSD matrix:

$$\boldsymbol{S_x}(f) = \sum_{m=-\infty}^{\infty} \boldsymbol{K_x}(m) e^{-j2\pi f m}, \tag{1.8}$$

where $\boldsymbol{K_x}(m) = E\left[\boldsymbol{x}(n+m)\boldsymbol{x}^H(n)\right]$ is the $N \times N$ autocorrelation matrix of the vector time series $\boldsymbol{x}(n)$ at lag m. This matrix is also equal to the

autocovariance matrix at lag m, since the mean of the vector time series $\boldsymbol{x}(n)$ is assumed to be zero.

The error vector $\boldsymbol{\epsilon}(n)$ between the input and output vectors in Figure 1.1 is defined as

$$\boldsymbol{\epsilon}(n) = \hat{\boldsymbol{x}}(n) - \boldsymbol{x}(n). \tag{1.9}$$

Then, from Figure 1.1,

$$\boldsymbol{\epsilon}(n) = \sum_{m=-\infty}^{\infty} \boldsymbol{w}(n-m)\boldsymbol{x}(m) - \boldsymbol{x}(n) + \sum_{m=-\infty}^{\infty} \boldsymbol{r}(n-m)\boldsymbol{q}(m), \tag{1.10}$$

where $\boldsymbol{w}(n)$ is the impulse response matrix sequence of the overall system given by

$$\boldsymbol{w}(n) = \boldsymbol{r}(n) * \boldsymbol{e}(n). \tag{1.11}$$

The z-transform of $\boldsymbol{r}(n)$ is equal to $\boldsymbol{R}(z)$.

The block MSE, used as the performance measure, is defined as

$$\mathcal{E}_{N,M} = \mathrm{Tr}\left(E\left[\boldsymbol{\epsilon}(n)\boldsymbol{\epsilon}^H(n)\right]\right), \tag{1.12}$$

where Tr is the trace operator. In the notation $\mathcal{E}_{N,M}$, N is the decimation factor and M is the number of quantizers receiving a positive number of bits.

MSE is a performance criterion which does not follow the human perception system very closely, but it is a measure which can be treated mathematically. One feature of MSE is that it can decide when two signals are equal. The performance measure can be improved by using weighted MSE, as was done in [Lee & Petersen 1976, Vandendorpe 1991, Gosse, Pothier & Duhamel 1995, Gosse & Duhamel 1997]. In this dissertation, a weighted MSE is *not* used, but all the theory developed should possibly be extended to include weighted MSE.

If high rates are assumed, the variance of the noise in quantizer number $i \in \{0, 1, \ldots, M-1\}$ can be modeled as [Jayant & Noll 1984]

$$\sigma_{q_i}^2 = c_i \sigma_{y_i}^2 2^{-2b_i}, \tag{1.13}$$

where $\sigma_{y_i}^2$ is the variance of the corresponding subband signal $y_i(n)$, see Figure 1.1, b_i is the number of bits used in quantizer number i, and c_i is the coding coefficient.[1] This traditional quantizer model will be used in Chapters 2 through 4 even though this model is not very accurate at low rates.

[1] The coding coefficient c_i depends on the pdf of the coded signal and the coding method used. If pdf optimized scalar quantizers are used, c_i is equal to the constant ϵ_*^2 defined on page 121 in [Jayant & Noll 1984]. If entropy coded scalar quantizers are used, c_i depends on the relative entropy of the pdf of the coded signal. If the pdf is Gaussian and pdf optimized scalar quantizers are used, $c_i = \frac{\sqrt{3}\pi}{2}$, and if entropy coded scalar quantizers are used, $c_i = \frac{e\pi}{6}$.

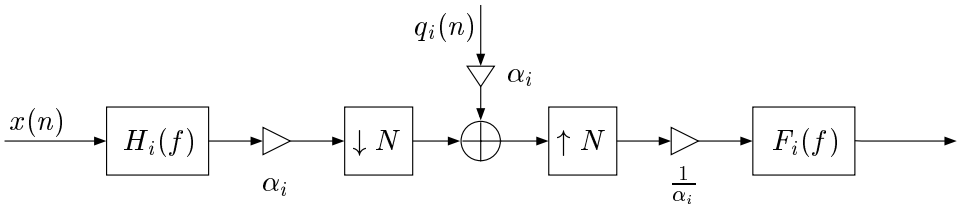

Figure 1.2 One branch of the subband coder model.

Rearranging Equation (1.13) gives

$$b_i = \frac{1}{2}\log_2\left(\frac{c_i\sigma_{y_i}^2}{\sigma_{q_i}^2}\right). \tag{1.14}$$

The bit constraint can be expressed as

$$b = \frac{1}{N}\sum_{i=0}^{M-1} b_i, \tag{1.15}$$

where b is the average number of bits used on N source samples by the M quantizers.

If a positive number of bits is spent in quantizer number i and the centroids are used as representation levels, it is impossible that $\sigma_{q_i}^2 > \sigma_{y_i}^2$. Therefore, if optimized quantizers are used, the following constraints must be satisfied:

$$\sigma_{y_i}^2 \geq \sigma_{q_i}^2, \quad \forall i \in \{0, 1, \ldots, M-1\}. \tag{1.16}$$

The goal is to find the jointly optimal filter banks and bit distribution which minimize the block MSE for a given total bit rate. With the model chosen for the quantizers, see Equation (1.14), the quantizers are fully specified by the relationship between the variances $\sigma_{y_i}^2$ and $\sigma_{q_i}^2$, when the coding coefficients c_i are known. In Figure 1.2, the operations in subband number i is shown. $H_i(f)$ and $F_i(f)$ are the frequency response of the respective analysis and synthesis filters in subband i. This subsystem has the same performance for all non-zero scaling factors α_i. Therefore, these degrees of freedom can be used to set the values of $\sigma_{q_i}^2$ equal to an arbitrary positive value without loss of generality. By assuming that the quantizer noise variances are known, the optimization of the filter banks also decides the bit distribution through the resulting values of $\sigma_{y_i}^2$. Another approach would be to scale the subband variances $\sigma_{y_i}^2$ and optimize the values of $\sigma_{q_i}^2$.

The degrees of freedom which exist because of the scaling of the subbands, as explained above, are used to choose all the diagonal elements of the matrix $\boldsymbol{\Lambda_q}$ such that $\sigma_{q_0}^2 = \sigma_{q_1}^2 = \cdots = \sigma_{q_{M-1}}^2 = \sigma_q^2$. Therefore, the PSD matrix of the quantization noise vector $\boldsymbol{q}(n)$ can be expressed as

$$\boldsymbol{\Lambda_q} = \sigma_q^2 \boldsymbol{I}. \tag{1.17}$$

Throughout this dissertation, σ_q^2 has been chosen equal to 1 in all the numerical results.

Although the numbering of the subband signals is arbitrary, in the theory and results presented in this dissertation, the subbands are numbered in accordance with decreasing values of the subband variances, i.e.,

$$\sigma_{y_0}^2 \geq \sigma_{y_1}^2 \geq \cdots \geq \sigma_{y_{M-1}}^2. \tag{1.18}$$

1.2.2 Power Constrained Filter Banks

The power constrained filter bank system operating on vectors is shown in Figure 1.3. By comparing Figures 1.3 and 1.1, it is seen that a vector formulation is used and that a channel transfer matrix $\boldsymbol{C}(z)$ is included in the power constrained problem. It is assumed that the $M \times M$ channel matrix $\boldsymbol{C}(z)$ is known. This matrix can, for example, model crosstalk between channels and frequency dependent attenuation. Since the additive channel noise can have different characteristics from the quantization noise in the bit constrained case, another symbol is chosen for the noise. The noise vector in power constrained filter bank systems is denoted $\boldsymbol{v}(n)$, and this vector time series is not necessarily white.

In this subsection, the transmitter and receiver filter banks consist of linear unconstrained length non-causal filters. Transforms and FIR filter banks will be treated in later chapters. The dimensions of the matrices in the figure show that a block of N source samples is transmitted by a block of M channel samples. The values of N and M are assumed to be known, and if a scalar time series is sent over a scalar channel, they indicate the available bandwidth. In this system, the objective is to design the encoder matrix $\boldsymbol{E}(z)$ and decoder matrix $\boldsymbol{R}(z)$ such that the block MSE between the signals $\boldsymbol{x}(n)$ and $\hat{\boldsymbol{x}}(n)$ is minimized. In addition, there is a constraint on the power used on the channel.

It is assumed that the vector time series $\boldsymbol{x}(n)$ and $\boldsymbol{v}(n)$ represent jointly WSS vector series [Sathe & Vaidyanathan 1993] which are mutually uncorrelated and have zero mean.

The block MSE used as the optimization criterion is defined in the same way as in the bit constrained case, and it is given by Equation (1.12), where the error vector $\boldsymbol{\epsilon}(n)$ is defined by Equation (1.9). Since the $M \times M$ channel

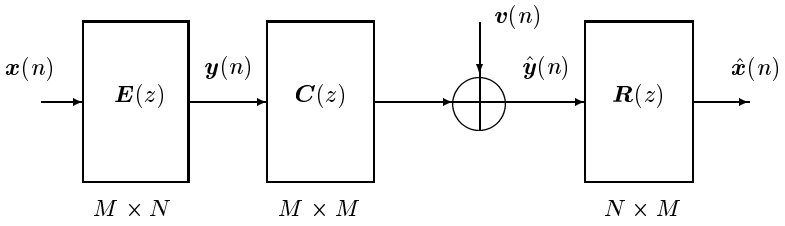

Figure 1.3 The power constrained MIMO block system.

impulse response matrix sequence $\boldsymbol{c}(n)$ must be taken into consideration, the $N \times N$ impulse response matrix sequence of the overall system $\boldsymbol{w}(n)$ is given by

$$\boldsymbol{w}(n) = \boldsymbol{r}(n) * \boldsymbol{c}(n) * \boldsymbol{e}(n), \tag{1.19}$$

where the z-transform of the channel impulse response matrix sequence $\boldsymbol{c}(n)$ equals $\boldsymbol{C}(z)$, i.e.,

$$\boldsymbol{C}(z) = \sum_{k=-\infty}^{\infty} \boldsymbol{c}(k) z^{-k}. \tag{1.20}$$

It is assumed that the matrix $\boldsymbol{C}(z)$ is known.

The statistical descriptions of the vector time series $\boldsymbol{x}(n)$ and $\boldsymbol{v}(n)$ are given by Equation (1.8) and the following PSD matrix, respectively:

$$\boldsymbol{S}_{\boldsymbol{v}}(f) = \sum_{m=-\infty}^{\infty} \boldsymbol{K}_{\boldsymbol{v}}(m) e^{-j2\pi f m}, \tag{1.21}$$

where the matrix $\boldsymbol{K}_{\boldsymbol{v}}(m) = E\left[\boldsymbol{v}(n+m)\boldsymbol{v}^H(n)\right]$ is the autocorrelation matrix of the vector time series $\boldsymbol{v}(n)$ at lag m. By using the inverse Fourier transform of the PSD matrices in Equations (1.8) and (1.21), $\boldsymbol{K}_{\boldsymbol{x}}(m)$ and $\boldsymbol{K}_{\boldsymbol{v}}(m)$ can be found.

The power constraint is imposed on the input vector $\boldsymbol{y}(n)$ to the channel transfer matrix, see Figure 1.3. Let $\sigma_{y_i}^2$ be the variance of the ith vector component of the vector $\boldsymbol{y}(n)$. The power used by the vector $\boldsymbol{y}(n)$ can be expressed as the following sum: $\sum_{i=0}^{M-1} \sigma_{y_i}^2$. If the average allowed power of the vector $\boldsymbol{y}(n)$ is denoted P, the power constraint can be expressed:

$$\sum_{i=0}^{M-1} \sigma_{y_i}^2 = P. \tag{1.22}$$

The matrix $E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]$ has $\sigma_{y_i}^2$ as the ith diagonal element, and by means of the Tr operator, the power constraint can be rewritten as

$$\mathrm{Tr}\left\{E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]\right\} = P. \tag{1.23}$$

1.3 Previous Work

In this section, some relevant literature treating the two main problems in the dissertation is mentioned. Since there has been an enormous amount of research activity in the area of multirate filter banks, the review is not intended to be exhaustive[2].

1.3.1 Bit Constrained Filter Banks

In the field of unconstrained length filter banks with a bit constraint, most of the literature treats PR systems. Optimal unitary filter banks with unconstrained lengths are proposed in [Vaidyanathan 1998], and these filter banks are equal to the optimal signal-adaptive, unitary, principal component filter banks with unconstrained filter lengths derived in [Tsatsanis & Giannakis 1995]. For a discussion on the connection between principal component filter banks and optimal unitary filter banks, see [Kıraç 1998]. In [Tuqan & Vaidyanathan 1997], a unitary filter bank is combined with optimal pre- and post-filters. Optimal unconstrained length biorthogonal filter banks have been studied in [Aase & Ramstad 1995, Aas & Mullis 1996, Vaidyanathan & Kıraç 1998, Moulin, Anitescu & Ramchandran 2000], where the total noise is minimized while maintaining PR. Some preliminary results on unconstrained length minimum MSE filter banks are presented in [Moulin, Anitescu & Ramchandran 1998].

A description of many existing transforms can be found in [Jain 1989]. The KLT is the optimal transform if PR is imposed [Hotelling 1933, Jain 1989]. In [Huang & Schultheiss 1963, Segall 1976, Jain 1989], it is shown that KLT is optimal if it is assumed that the input time series is Gaussian, scalar Lloyd-Max quantizers are used, and the analysis transform produces a vector with uncorrelated components. If Lloyd-Max vector quantizers are used, it is shown in [Vaidyanathan & Chen 1994] that KLT is the optimal transform. The discrete cosine transform (DCT) [Ahmed, Natarajan & Rao 1974] is often used as a transform because it is close to the KLT [Scharf & Tufts 1987] when the input can be modeled as a Gauss-Markov source with high correlation [Jain 1989].

[2] Many references on filter banks can be found in [Vaidyanathan 1993, Vetterli & Kovačević 1995, Donoho, Vetterli, DeVore & Daubechies 1998]

In the field of FIR filter banks, no closed form expressions have been found for unitary, biorthogonal, or filter banks without any restrictions. Some PR FIR filter banks that are frequently used can be found in [Le Gall & Tabatabai 1988], [Rodrigues, da Silva & Diniz 1997], [Balasingham 1998], [Antonini et al. 1992], and [Tsai, Villasenor & Chen 1996]. In [Malvar 1992], the Lapped Orthogonal Transforms (LOT) are proposed, which are low complexity FIR filter banks having the PR property. For a given analysis FIR filter bank, the optimal FIR synthesis filter bank has been derived in different ways [Honig et al. 1992, Gosse & Duhamel 1997, Delopoulos & Kollais 1996], when the same filter length are used in all the filters and the delay through the filter bank is $N - 1 + d_v N$, where d_v is a positive integer. In [Aase 1993], near PR FIR filter banks are found using a weighted error function which includes PR, coding gain, blocking effects, and scaling of the synthesis filters.

1.3.2 Power Constrained Filter Banks

Jointly optimal transmitter and receiver filters with single input single output (SISO) were derived in [Costas 1952, Berger & Tufts 1967, Tufts & Berger 1967, Berger 1971] for both continuous and discrete time. In this system, the receiver is the well known SISO Wiener filter. In the continuous time case, jointly optimal analysis and synthesis filter banks with a power constraint are presented in [Yang & Roy 1994]. The receiver filter bank is an unconstrained length Wiener filter bank, which is also treated in [Vaidyanathan & Chen 1994].

In [Lee & Petersen 1976], a jointly optimal transform coding system was derived with a constraint on the power used on the channel. This is a non-PR transform, which is related to the KLT.

In the field of power constrained FIR filter banks, the analysis and synthesis filter banks are jointly optimized by random search in [Song & Ritcey 1997]. Formulas for finding the optimal synthesis filter bank for a given analysis filter bank and vice versa are proposed in [Honig et al. 1992]. In [Malvar & Staelin 1988], an algorithm is given for finding jointly optimized, linear phase, FIR pre- and post-filters with decimator and expander for communication over a channel with additive noise. FIR Wiener filter banks references are mentioned in Subsection 1.3.1. In [Chen, Lin & Chen 1995], a multirate Kalman synthesis filtering approach was proposed for solving the FIR problem for a fixed FIR analysis filter bank.

1.4 Outline of the Dissertation

There are eight chapters in this dissertation, and the upcoming chapters are organized as follows:

Chapter 2: Optimal bit and power constrained filter banks with unconstrained filter lengths are derived.

Chapter 3: Optimal bit and power constrained transforms are found based on the results from Chapter 2.

Chapter 4: Numerical algorithms for joint optimization of analysis and synthesis filter banks under bit and power constraints are proposed for FIR filter banks.

Chapter 5: It is shown that there is a connection between the performance of the power constrained transform in Chapter 3 and the unconstrained length filter banks with a power constraint presented in Chapter 2, when the dimensions of the transform system approach infinity and a modulation/demodulation unit is used in the unconstrained filter banks.

Chapter 6: A practical source coder is introduced, and it is shown that there exists a mismatch between the theoretical and practical results. The nature of this mismatch is analyzed.

Chapter 7: Three ways of improving the connection between theoretical and practical results are presented. In the first two methods, the subband signals and the additive coding noise are forced to be uncorrelated, and in the third method, a signal dependent colored noise model is proposed.

Chapter 8: The conclusions of the work presented in this dissertation are given.

There are four appendices, and these are organized as follows:

Appendix A: The block MSE, bit constraint, and power constraint expressions used in the optimization problems of the dissertation are derived for the unconstrained length and the FIR cases.

Appendix B: The eigenvalues of the PSD matrix are derived and the so-called ordering functions are introduced. These functions are used in connection with unconstrained length filter banks. Some properties of these functions are derived.

Appendix C: Formulas for matrix variational calculus and matrix differentiation are derived. These formulas are used in solving some of the optimization problems in the dissertation.

Appendix D: Formulas for finding the correlations of the signal dependent colored noise model are found.

1.5 Contributions of the Dissertation

The main contributions in this dissertation for the bit constrained problem are as follows: Jointly optimal analysis and synthesis filter banks with unconstrained lengths are derived in Section 2.1, while jointly optimal analysis and synthesis transforms are found in Section 3.1. An iterative numerical algorithm for finding jointly optimized analysis and synthesis FIR filter banks is proposed in Section 4.1. This includes the derivation of the FIR Wiener filter solution for arbitrary given filter lengths and delay through the filter bank is $N-1+d_vN+d_s$, where d_v and d_s are appropriate integers. The analysis FIR filters can also have arbitrary given filter lengths.

The main contributions of this dissertation for the power constrained problem are as follows: Jointly optimal transmitter and receiver filter banks for discrete time having unconstrained filter lengths are deduced from a corresponding solution for the continuous time case in Section 2.2. An alternative derivation of the jointly optimal transmitter and receiver transforms is presented in Section 3.2. Formulas for finding a jointly optimal transmitter and receiver in the FIR case are given.

There are also other topics treated in the dissertation, and the main contributions of these parts are as follows: Scalar quantizers with uncorrelated input and quantization noise are proposed in Subsection 7.1.1, and they are used in a practical subband coder. Subtractive dithering is used in a practical subband coder in Chapter 7, and it is shown that good correspondence is achieved between practical and theoretical performance results. A signal dependent colored quantization noise model is introduced in Section 7.2. Formulas for finding all the correlations in the subband coder, when using midtread uniform threshold quantizers having infinite dynamic range and using centroid representation levels, are derived in Appendix D. Conditions for optimality of FIR PR filter banks are derived in Section 7.3.

Chapter 2

Unconstrained Length Signal-Adaptive Filter Banks

In this chapter, no constraints are set on the filter lengths, so the goal is to find the polyphase matrices

$$\begin{aligned} \boldsymbol{E}(z) &= \sum_{k=-\infty}^{\infty} \boldsymbol{e}(k) z^{-k}, \\ \boldsymbol{R}(z) &= \sum_{k=-\infty}^{\infty} \boldsymbol{r}(k) z^{-k}, \end{aligned} \tag{2.1}$$

which jointly minimize the block MSE under a bit and power constraint.

This chapter consists of two main parts. The first main part treats the problem of finding jointly optimal analysis and synthesis filter banks under a bit constraint, and this is presented in Section 2.1. The second main part, which is given in Section 2.2, treats jointly optimal transmitter and receiver filter banks under a power constraint for transmission over a known linear channel with additive noise. Finally, a short summary is given in Section 2.3.

The first part of this chapter is partly based on [Hjørungnes & Ramstad 1998*a*, Hjørungnes & Ramstad 1999*b*, Hjørungnes & Ramstad 1999*c*], while the second part is partly based on [Hjørungnes & Ramstad 1997].

2.1 Bit Constrained Filter Banks

This section is organized as follows: The problem is formulated in Subsection 2.1.1. In Subsection 2.1.2, an equivalent system with diagonal PSD matrix for the input vectors is presented, while in Subsection 2.1.3, the optimal

solution is derived and some properties of the optimal solution are presented. In Subsection 2.1.4, some results using the proposed filter banks are given.

2.1.1 Problem Formulation

It is shown in Appendix A that the block MSE can be expressed in the frequency domain as

$$\mathcal{E}_{N,M} = \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \Big\{ [\boldsymbol{I} - \boldsymbol{R}(f)\boldsymbol{E}(f)]\, \boldsymbol{S_x}(f)\, [\boldsymbol{I} - \boldsymbol{R}(f)\boldsymbol{E}(f)]^H + \boldsymbol{R}(f)\boldsymbol{\Lambda_q}\boldsymbol{R}^H(f) \Big\}\, df \right). \quad (2.2)$$

$\boldsymbol{E}(f)$ and $\boldsymbol{R}(f)$ are the polyphase analysis and synthesis matrices, respectively, evaluated on the unit circle, $z = e^{j2\pi f}$. Strictly speaking, the notation $\boldsymbol{E}\left(e^{j2\pi f}\right)$ should be used, but to simplify, $\boldsymbol{E}(f)$ is used instead.

The integrand in Equation (2.2) is composed of two main terms. The first term is a signal distortion term. Systems where $\boldsymbol{R}(f)\boldsymbol{E}(f) = \boldsymbol{I}$ have the PR property. In this case, the signal distortion term is zero. In general, the signal distortion can be classified as amplitude, phase, and aliasing distortions [Vaidyanathan 1993]. The second term in Equation (2.2) is due to quantization noise, and it is independent of the analysis filter bank.

By using the quantizer model of Equation (1.13), it is shown in Appendix A that the bit constraint of Equation (1.15) can be expressed as

$$\mathrm{Pr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{E}(f)\boldsymbol{S_x}(f)\boldsymbol{E}^H(f)\, df \right) = 2^{2Nb} \cdot \prod_{i=0}^{M-1} \frac{\sigma_{q_i}^2}{c_i} \stackrel{\mathrm{def}}{=} \beta, \quad (2.3)$$

where the operator Pr multiplies the elements of the main diagonal of the matrix, and the constant β is defined for convenience. The bit constraint is independent of the synthesis filter bank $\boldsymbol{R}(f)$.

Since all the quantization variances are chosen equal, see Subsection 1.2.1, the constraint in Equation (1.16) can now be rewritten as:

$$\sigma_{y_i}^2 \geq \sigma_q^2, \quad \forall i \in \{0, 1, \ldots, M-1\}. \quad (2.4)$$

The problem is to find jointly optimal analysis and synthesis filter banks which minimize Equation (2.2), subject to the bit constraints in Equations (2.3) and (2.4).

The only term in Equation (2.2) which is dependent on the bit rates in the quantizers is the last term. For PR filter banks, the total block MSE is

given by this term only, since the signal distortion is zero for PR filter banks. Therefore, the condition for optimal rate allocation for the filter banks treated in this section is the same as the condition for PR filter banks. In [Moulin et al. 2000], jointly optimal analysis and synthesis biorthogonal filter banks are treated, and it is shown that optimal rate allocation is achieved when the product of the quantization noise variance and the squared norm of the synthesis filter in the same subband is constant for all subbands. Because of this optimality condition, the scaling of the quantization variances is equivalent to setting the norm of all the synthesis filters equal, which is the scaling used in [Aase & Ramstad 1995].

2.1.2 Transformation to an Equivalent Problem

To simplify the analysis and optimization further, the system shown in Figure 1.1 is transformed into an equivalent system with new matrices $\boldsymbol{G}(f)$ and $\boldsymbol{T}(f)$. These transforms are given by

$$\begin{aligned} \boldsymbol{E}(f) &= \boldsymbol{G}(f)\boldsymbol{U}^H(f), \\ \boldsymbol{R}(f) &= \boldsymbol{U}(f)\boldsymbol{T}(f). \end{aligned} \tag{2.5}$$

$\boldsymbol{U}(f)$ is a unitary matrix which diagonalizes the input PSD matrix $\boldsymbol{S_x}(f)$ given in Equation (1.8), i.e.,

$$\boldsymbol{S_x}(f)\boldsymbol{U}(f) = \boldsymbol{U}(f)\boldsymbol{\Lambda_x}(f). \tag{2.6}$$

In Equation (2.6), $\boldsymbol{\Lambda_x}(f)$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{S_x}(f)$. Since the matrix $\boldsymbol{S_x}(f)$ is Hermitian [Vaidyanathan 1993], its eigenvalues will be real [Young 1990]. In addition, the elements of $\boldsymbol{\Lambda_x}(f)$ are ordered as follows:

$$\lambda_0^{(N)}(f) \geq \lambda_1^{(N)}(f) \geq \cdots \geq \lambda_{N-1}^{(N)}(f), \quad \forall f. \tag{2.7}$$

By substituting the results from Equations (2.5) and (2.6) into Equation (2.2), the block MSE can be expressed as

$$\begin{aligned} \mathcal{E}_{N,M} = \mathrm{Tr}\Bigg(\int_{-\frac{1}{2}}^{\frac{1}{2}} \Big\{ &\left[\boldsymbol{I} - \boldsymbol{T}(f)\boldsymbol{G}(f)\right] \boldsymbol{\Lambda_x}(f) \left[\boldsymbol{I} - \boldsymbol{T}(f)\boldsymbol{G}(f)\right]^H + \\ &\boldsymbol{T}(f)\boldsymbol{\Lambda_q}\boldsymbol{T}^H(f) \Big\}\, df \Bigg), \end{aligned} \tag{2.8}$$

where the trace identity $\mathrm{Tr}\left(\boldsymbol{A}_1\boldsymbol{A}_2\right) = \mathrm{Tr}\left(\boldsymbol{A}_2\boldsymbol{A}_1\right)$, where $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are matrices, and the fact that the integral and the trace operators commute have been used.

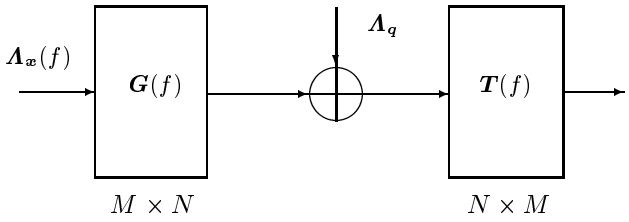

Figure 2.1 The equivalent block diagram of the subband coder model.

The bit constraint (2.3) is transformed into

$$\Pr\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f)\,df\right) = \beta. \tag{2.9}$$

The additive noise components remain uncorrelated. The original and new equivalent system have the same subband variances since

$$\begin{aligned}\left(\boldsymbol{E}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{E}^H(f)\right)_{i,i} &= \left(\boldsymbol{G}(f)\boldsymbol{U}^H(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{U}(f)\boldsymbol{G}^H(f)\right)_{i,i} \\ &= \left(\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f)\right)_{i,i},\end{aligned} \tag{2.10}$$

where the notation $(\cdot)_{i,i}$ means element in row and column number i. Subband variance number i of the original system in Figure 1.1 is found by integrating the left hand side of Equation (2.10), while the ith subband variance of the equivalent system is found by integrating the right hand side of Equation (2.10). Since the two integrands are equal for all values of f, the subband variances of the two systems are equal. The PSD matrices of the vectors containing the *subband* signals are equal in the two systems, so this vector will be called $\boldsymbol{y}(n)$ in both systems. The constraints $\sigma_{y_i}^2 \geq \sigma_q^2$ have to be satisfied for all $i \in \{0, 1, \ldots, M-1\}$ as well. By comparing Equations (2.8) and (2.9) to Equations (2.2) and (2.3), it is seen that the original system has been transformed into an equivalent system where the PSD matrix of the input vector time series is diagonal. The equivalent system is shown in Figure 2.1.

Zero bits in a quantizer are obtained by reducing the value of M. For a given value of N and bit rate, the value used for M must be optimized through a discrete optimization, where M is chosen from the set $\{0, 1, 2, \ldots, N\}$.

2.1.3 Optimal System Solution

Having transformed the original problem into an equivalent one, the objective is to minimize the block MSE given by Equation (2.8), subject to the bit constraint given by Equation (2.9) and the constraints $\sigma_{y_i}^2 \geq \sigma_q^2$, where $i \in \{0, 1, \ldots, M-1\}$, with respect to the new variables $\boldsymbol{G}(f)$ and $\boldsymbol{T}(f)$.

2.1.3.1 Necessary Conditions for Optimality

In order to find necessary conditions for optimality, the Kuhn-Tucker conditions [Luenbereger 1984] can be calculated for the problem, and the differentiation is done by *variational calculus* since the unknowns are functions instead of scalars or vectors. The unconstrained objective function is first written as a function of the elements of the matrices $\boldsymbol{G}(f)$ and $\boldsymbol{T}(f)$, and then the Gâteaux variation [Troutman 1996, Magnus & Neudecker 1988] is calculated with respect to these elements. If the resulting equations are written in matrix form, the necessary conditions are found. In Appendix C, it is shown how this can be done, and the necessary Kuhn-Tucker conditions can be written as:

$$\boldsymbol{T}(f) = \boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f)\left[\boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda_q}\right]^{-1}, \tag{2.11}$$

$$\boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{G}(f) + \left(\mu\boldsymbol{\Sigma_y}^{-1} - \boldsymbol{\Theta}\right)\boldsymbol{G}(f) = \boldsymbol{T}^H(f), \tag{2.12}$$

where $\mu \in \mathbb{R}^+ = (0, \infty)$ is a Kuhn-Tucker parameter for the equality constraint (2.9). The matrix $\boldsymbol{\Sigma_y}$ is an $M \times M$ diagonal matrix with diagonal element number m given by

$$\sigma_{y_m}^2 = [\boldsymbol{\Sigma_y}]_{m,m} = \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{N-1} |G_{m,i}(f)|^2 \lambda_i^{(N)}(f)\, df, \quad m \in \mathbb{Z}_M, \tag{2.13}$$

where the set $\mathbb{Z}_M$ is defined as $\mathbb{Z}_M = \{0, 1, \ldots, M-1\}$. The matrix $\boldsymbol{\Theta}$ is a diagonal matrix where diagonal element number i is the parameter in the Kuhn-Tucker conditions for the inequality constraint $\sigma_{y_i}^2 \geq \sigma_q^2$. This parameter will be denoted θ_i and it must be non-negative, i.e., $\theta_i \geq 0$. Furthermore, the following equations must hold:

$$\left(\sigma_q^2 - \sigma_{y_m}^2\right)\theta_m = 0, \quad m \in \mathbb{Z}_M. \tag{2.14}$$

To simplify the expressions, the following $M \times M$ diagonal matrix is introduced:

$$\boldsymbol{\Phi} = \left[\mu\boldsymbol{\Sigma_y}^{-1} - \boldsymbol{\Theta}\right]^{-1}, \tag{2.15}$$

where diagonal element number i of the matrix $\boldsymbol{\Phi}$ is named ϕ_i.

The diagonal elements in $\boldsymbol{\Sigma_y}$ represent the subband powers (and also the variances since the mean is assumed to be zero). The constraints in Equation (2.4) assure that M of the subband variances are positive, so the inverse of the matrix $\boldsymbol{\Sigma_y}$ will always exist.

2.1.3.2 Synthesis Polyphase Matrix

It will now be explained why the synthesis polyphase matrix represents a polyphase Wiener matrix.

By using the results from Equations (2.5), (2.6), and (2.11), the synthesis polyphase matrix can be expressed as

$$\boldsymbol{R}(f) = \boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{E}^H(f)\left[\boldsymbol{E}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{E}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]^{-1}. \tag{2.16}$$

Using the vector notation introduced in Figure 1.1, the standard unconstrained length polyphase Wiener matrix [Vaidyanathan & Chen 1994] $\boldsymbol{R}_{\text{Wiener}}(f)$ can be expressed as

$$\boldsymbol{R}_{\text{Wiener}}(f) = \boldsymbol{S}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(f)\boldsymbol{S}_{\hat{\boldsymbol{y}}}^{-1}(f), \tag{2.17}$$

where $\boldsymbol{S}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(f)$ is the cross PSD matrix between the vectors $\boldsymbol{x}(n)$ and $\hat{\boldsymbol{y}}(n)$, and $\boldsymbol{S}_{\hat{\boldsymbol{y}}}(f)$ is the PSD matrix of the input vector $\hat{\boldsymbol{y}}(n)$ to the synthesis matrix.

Since the additive quantization noise is assumed to be uncorrelated with the input signal, $\boldsymbol{S}_{\hat{\boldsymbol{y}}}(f) = \boldsymbol{E}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{E}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}$. It can be shown that $\boldsymbol{S}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(f) = \boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{E}^H(f)$ [Vaidyanathan 1993]. By inserting these two results in Equation (2.17), it is seen that the optimal synthesis polyphase matrix in Equation (2.16) is a polyphase Wiener matrix with unconstrained length.

It can also be verified that the optimal receiver matrix $\boldsymbol{T}(f)$ given in Equation (3.10), is a polyphase Wiener matrix for the system shown in Figure 2.1.

For unconstrained length non-causal filter banks, the derivation of the polyphase Wiener matrix has been obtained in [Vaidyanathan & Chen 1994], using the orthogonality principle [Therrien 1992].

2.1.3.3 Conditions for Optimality of PR in the Unconstrained Length Filter Bank Case

Now, conditions for when the optimal unconstrained length filter bank possesses the PR property will be stated for a given invertible analysis filter bank, and it will be showed that this is never the case for the quantization model assumed in this chapter.

In [Vaidyanathan & Chen 1994], it was shown that for a given invertible analysis filter bank $\boldsymbol{E}(z)$, the optimal unconstrained length filter bank system has the PR property if, and only if, the following condition is satisfied:

$$E\left[\hat{\boldsymbol{y}}(m)\boldsymbol{q}(n)^H\right] = \boldsymbol{0}, \quad \forall n, m, \tag{2.18}$$

where the vector $\hat{\boldsymbol{y}}(n)$ is the $M \times 1$ input vector of the synthesis polyphase filter bank matrix, that is $\hat{\boldsymbol{y}}(n) = [\hat{y}_0(n), \hat{y}_1(n), \ldots, \hat{y}_{M-1}(n)]^T$, see Figure 1.1.

If PR should be possible $M = N$. The condition in Equation (2.18) is very strict, and it will almost never occur in a practical coding system.

In this chapter, it is assumed that the quantization noise $\boldsymbol{q}(n)$ is uncorrelated with the input signal to the quantizers $\boldsymbol{y}(n)$, and it is also assumed that the additive quantization noise is white with uncorrelated components in each quantizer. With this quantization noise model, the cross-correlation matrix in Equation (2.18) is given by:

$$E\left[\hat{\boldsymbol{y}}(m)\boldsymbol{q}(n)^H\right] = E\left[\left(\boldsymbol{y}(n)+\boldsymbol{q}(n)\right)\boldsymbol{q}(n)^H\right] = \delta(n-m)\boldsymbol{\Lambda_q} \neq \boldsymbol{0}, \quad \forall n, m. \tag{2.19}$$

Therefore, with the quantization noise model used in this chapter, a PR system will never be optimal.

2.1.3.4 Zero Elements of the Matrix $\boldsymbol{G}(f)$

Here, it will be shown that the optimal matrix $\boldsymbol{G}(f)$ has at most one non-zero element in each row and column.

Post-multiplication of Equation (2.12) by $\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f)$ gives

$$\begin{aligned}\boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Phi}^{-1}\boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) \\ = \boldsymbol{T}^H(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f).\end{aligned} \tag{2.20}$$

By rearranging Equation (2.11), one gets

$$\boldsymbol{T}(f)\boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) = \boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) - \boldsymbol{T}(f)\boldsymbol{\Lambda_q}. \tag{2.21}$$

Pre-multiplying Equation (2.21) by $\boldsymbol{T}^H(f)$ and subtracting the resulting expression from Equation (2.20) renders

$$\boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) = \boldsymbol{\Phi}\boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{\Lambda_q}. \tag{2.22}$$

Because the left hand side of Equation (2.22) is Hermitian, the right hand side is also Hermitian, i.e., $\boldsymbol{\Phi}\boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{\Lambda_q} = \boldsymbol{\Lambda_q}\boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{\Phi}$. Because of the choice in Equation (1.17), $\boldsymbol{\Lambda_q}$ can be dropped from this equation, and it can be rewritten as $\boldsymbol{\Phi}\boldsymbol{T}^H(f)\boldsymbol{T}(f) = \boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{\Phi}$.

If it is assumed that

$$\phi_m \neq \phi_n \text{ whenever } n \neq m, \tag{2.23}$$

where ϕ_m is the mth diagonal element in $\boldsymbol{\Phi}$, the right hand side of Equation (2.22) can be Hermitian only if the matrix $\boldsymbol{T}^H(f)\boldsymbol{T}(f)$ is diagonal. This is

shown by considering the off-diagonal elements of the equation $\boldsymbol{\Phi}\boldsymbol{T}^H(f)\boldsymbol{T}(f) = \boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{\Phi}$.

Since all the matrices on the right hand side of Equation (2.22) are diagonal, the matrix $\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f)$ must also be diagonal.

If the assumption in Equation (2.23) is not valid, i.e., p of the ϕ_i components are equal, one can add a small positive quantity $(p-1)\nu$ to the ϕ_i component with the smallest index, $(p-2)\nu$ to the ϕ_i component with the next increasing index, and so on until zero is added to the ϕ_i component with the largest index among these p equal ϕ_i components. These ϕ_i components are now unequal, and by using the same reasoning as above, it is shown that the matrix product $\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f)$ is diagonal. Since the Lagrange function for the problem considered is continuous in the elements of $\boldsymbol{\Phi}$, see Equation (C.1), this reasoning is valid.

Equation (2.11) can be rewritten as

$$\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f) = \left[\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]\boldsymbol{T}^H(f). \tag{2.24}$$

By substituting $\boldsymbol{T}^H(f)$, given by Equation (2.12), into Equation (2.24), one obtains the following result:

$$\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f) = \left[\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]\left[\boldsymbol{T}^H(f)\boldsymbol{T}(f) + \boldsymbol{\Phi}^{-1}\right]\boldsymbol{G}(f). \tag{2.25}$$

Post-multiplying Equation (2.25) by $\boldsymbol{\Lambda}_{\boldsymbol{x}}^l(f)\boldsymbol{G}^H(f)$, where $l \in \mathbb{N} = \{0, 1, 2, \dots\}$, one gets

$$\begin{aligned}\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}^{l+1}(f)\boldsymbol{G}^H(f) = &\left[\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right] \\ &\cdot \left[\boldsymbol{T}^H(f)\boldsymbol{T}(f) + \boldsymbol{\Phi}^{-1}\right]\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}^l(f)\boldsymbol{G}^H(f).\end{aligned} \tag{2.26}$$

It was already proven that the matrices $\boldsymbol{T}^H(f)\boldsymbol{T}(f)$ and $\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f)$ are diagonal. Then, it follows by induction that the matrices $\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}^l(f)\boldsymbol{G}^H(f)$ are diagonal for $l \in \mathbb{N}$. This will now be used to show that each column in the matrix $\boldsymbol{G}(f)$ has at most one non-zero element.

Substitution of Equation (2.11) into Equation (2.12) leads to

$$\begin{aligned}\Big(&\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}^2(f)\boldsymbol{G}^H(f)\left[\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]^{-1} + \\ &\left[\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]\boldsymbol{\Phi}^{-1}\Big)\cdot\boldsymbol{G}(f) = \boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f).\end{aligned} \tag{2.27}$$

This equation is of the form $\boldsymbol{D}(f)\boldsymbol{G}(f) = \boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$, where the matrix $\boldsymbol{D}(f)$ is a *diagonal* $M \times M$ matrix, since it is given by sums and products of diagonal matrices. Assume that all the elements in $\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$ are different. Since $\boldsymbol{D}(f)$ is

diagonal, M of the elements of $\boldsymbol{\Lambda_x}(f)$ will be found on the diagonal of the matrix $\boldsymbol{D}(f)$. By studying the equation $\boldsymbol{D}(f)\boldsymbol{G}(f) = \boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)$ and using the assumption that all the elements of $\boldsymbol{\Lambda_x}(f)$ are different, it is seen that every column in the matrix $\boldsymbol{G}(f)$ has at most one non-zero element.

By using the same reasoning as above, it can be seen from the equation $\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) = \boldsymbol{G}^H(f)\boldsymbol{D}(f)$ that each row of the matrix $\boldsymbol{G}(f)$ has at most one non-zero element.

If p of the eigenvalues are equal, one can add a small quantity $(p-1)\nu$ to the eigenvalue with the smallest index, $(p-2)\nu$ to the eigenvalue with the next increasing index, and so on until zero is added to the eigenvalue with the largest index of these p equal eigenvalues. These eigenvalues are now unequal, and by using the reasoning above it is seen that the matrix $\boldsymbol{G}(f)$ has at most one non-zero element in each row and column. By letting ν approach zero from the right, the desired result is obtained. The Lagrange function for the problem is continuous in the elements of $\boldsymbol{\Lambda_x}(f)$, see Equation (C.1), so this reasoning holds. However, for $\nu = 0$ the optimal matrix $\boldsymbol{G}(f)$ is not unique.

2.1.3.5 The Matrix $\boldsymbol{G}(f)$

Now, it will be shown that there is no loss in optimality by letting $\boldsymbol{G}(f)$ be a diagonal matrix. Only the case $N \geq M$ will be treated. From Subsection 2.1.3.4, it is known that the optimum matrix $\boldsymbol{G}(f)$ can be expressed as

$$\boldsymbol{G}(f) = \tilde{\boldsymbol{G}}(f)\boldsymbol{\Pi}(f), \tag{2.28}$$

where the matrix $\boldsymbol{\Pi}(f)$ is an $N \times N$ permutation matrix, and $\tilde{\boldsymbol{G}}(f)$ is an $M \times N$ matrix with non-zero elements on the main diagonal. From Equation (2.11), it can be seen that the matrix $\boldsymbol{T}(f)$ can have non-zero elements only where the matrix $\boldsymbol{G}^H(f)$ has non-zero elements. The reason is that the matrix $\boldsymbol{T}(f)$ is equal to $\boldsymbol{G}^H(f)$ pre- and post-multiplied by diagonal matrices. Therefore, the matrix $\boldsymbol{T}(f)$ can be expressed as

$$\boldsymbol{T}(f) = \boldsymbol{\Pi}^H(f)\tilde{\boldsymbol{T}}(f), \tag{2.29}$$

where $\tilde{\boldsymbol{T}}(f)$ is an $N \times M$ matrix with non-zero elements on the main diagonal. The permutation matrix $\boldsymbol{\Pi}(f)$ is unitary, so the inverse of this matrix is $\boldsymbol{\Pi}^H(f)$. If $\boldsymbol{G}(f)$ and $\boldsymbol{T}(f)$ given in Equation (2.28) and (2.29) are substituted into Equation (2.8), one obtains the following expression for the block

MSE:

$$\mathcal{E}_{N,M} = \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \left\{ \left[\boldsymbol{I} - \tilde{\boldsymbol{T}}(f)\tilde{\boldsymbol{G}}(f)\right] \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f) \left[\boldsymbol{I} - \tilde{\boldsymbol{T}}(f)\tilde{\boldsymbol{G}}(f)\right]^H + \tilde{\boldsymbol{T}}(f)\boldsymbol{\Lambda}_{\boldsymbol{q}}\tilde{\boldsymbol{T}}^H(f) \right\} df \right), \tag{2.30}$$

where $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f) = \boldsymbol{\Pi}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{\Pi}^H(f)$ is a diagonal matrix. The bit constraint given in Equation (2.9) can be rewritten as

$$\Pr\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \tilde{\boldsymbol{G}}(f)\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)\tilde{\boldsymbol{G}}^H(f)\, df\right) = \beta. \tag{2.31}$$

Equation (2.11) now becomes

$$\tilde{\boldsymbol{T}}(f) = \tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)\tilde{\boldsymbol{G}}^H(f)\left[\tilde{\boldsymbol{G}}(f)\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)\tilde{\boldsymbol{G}}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]^{-1}. \tag{2.32}$$

Using the result from (2.32) to rewrite the block MSE in Equation (2.30), one obtains

$$\mathcal{E}_{N,M} = \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{M-1} \frac{\tilde{\lambda}_i^{(N)}(f)\sigma_q^2}{\left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f) + \sigma_q^2}\, df + \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=M}^{N-1} \tilde{\lambda}_i^{(N)}(f)\, df. \tag{2.33}$$

The first sum in Equation (2.33) represents the M subbands that will receive bits, and the last sum represents the remaining $N - M$ subbands that do not receive bits. Therefore, if $M = N$, the last sum of Equation (2.33) is equal to zero.

The bit constraint given in Equation (2.31) can be rewritten as

$$\sum_{i=0}^{M-1} \ln \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f)\, df = \ln(\beta). \tag{2.34}$$

Through Equations (2.33) and (2.34), the original system has been transformed into an equivalent diagonal system, which is less complex then the original system. It will be shown below that no optimality is lost by choosing $\boldsymbol{\Pi}(f) = \boldsymbol{I}$. Therefore, the optimal $\boldsymbol{G}(f)$ is *diagonal*.

Assume that $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$ is not identical to $\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$ in the ordering of the first N diagonal elements. Then, there exists at least one pair of indices $j < k$ such that for some f

$$\tilde{\lambda}_j^{(N)}(f) < \tilde{\lambda}_k^{(N)}(f). \tag{2.35}$$

Let the set of frequencies in $\left(-\frac{1}{2}, \frac{1}{2}\right]$ satisfying Equation (2.35) be denoted $\mathcal{F}_{j,k}$.

Now, consider the matrix $\bar{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$, which is identical to $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$ except for swapping of the jth and the kth diagonal element for the frequencies $f \in \mathcal{F}_{j,k}$, i.e.,

$$\bar{\lambda}_j^{(N)}(f) = \tilde{\lambda}_k^{(N)}(f) \quad \text{and} \quad \bar{\lambda}_k^{(N)}(f) = \tilde{\lambda}_j^{(N)}(f), \quad \forall f \in \mathcal{F}_{j,k}. \tag{2.36}$$

Let the filters used in connection with $\bar{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$ be called $\bar{G}_{i,i}(f)$, and let these filters be chosen equal to the $\tilde{G}_{i,i}(f)$ filters for all i and all f, except for $i \in \{j,k\}$ and $f \in \mathcal{F}_{j,k}$.

For the indices j and k there are three possibilities

$$\begin{aligned} &\text{(i)} \quad j < k \leq M-1, \\ &\text{(ii)} \quad M \leq j < k, \\ &\text{(iii)} \quad j \leq M-1 < k. \end{aligned} \tag{2.37}$$

First assume (i). The difference between the integrands of the block MSE using the $\tilde{G}_{i,i}(f)$ and the $\bar{G}_{i,i}(f)$ filters, respectively, is given by

$$\mathcal{E}(f) - \bar{\mathcal{E}}(f) = \begin{cases} \frac{\tilde{\lambda}_j^{(N)}(f)\sigma_q^2}{\left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f)+\sigma_q^2} + \frac{\tilde{\lambda}_k^{(N)}(f)\sigma_q^2}{\left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f)+\sigma_q^2} \\ \quad - \left(\frac{\bar{\lambda}_j^{(N)}(f)\sigma_q^2}{\left|\bar{G}_{j,j}(f)\right|^2 \bar{\lambda}_j^{(N)}(f)+\sigma_q^2} + \frac{\bar{\lambda}_k^{(N)}(f)\sigma_q^2}{\left|\bar{G}_{k,k}(f)\right|^2 \bar{\lambda}_k^{(N)}(f)+\sigma_q^2} \right), & f \in \mathcal{F}_{j,k}, \\ 0, & f \in \mathcal{F}_{j,k}^C, \end{cases} \tag{2.38}$$

where the set $\mathcal{F}_{j,k}^C$ is the complementary set of $\mathcal{F}_{j,k}$ in the frequency interval $\left(-\frac{1}{2}, \frac{1}{2}\right]$.

If there exist filters $\bar{G}_{k,k}(f)$ and $\bar{G}_{j,j}(f)$ such that the right hand side of Equation (2.38) is non-negative, and if they use no more bits than the filters $\tilde{G}_{k,k}(f)$ and $\tilde{G}_{j,j}(f)$, then it is shown that the performance of the system corresponding to $\bar{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$ is at least as good as the system corresponding to $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{x}}(f)$.

Let the following frequency sets be defined:

$$\mathcal{F}_{j,k}^{(1)} = \left\{ f \in \mathcal{F}_{j,k} \;\middle|\; \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f) > \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f) \right\} \text{ and} \tag{2.39}$$

$$\mathcal{F}_{j,k}^{(2)} = \left\{ f \in \mathcal{F}_{j,k} \;\middle|\; \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f) \leq \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f) \right\}. \tag{2.40}$$

From the definitions above, it is seen that the sets $\mathcal{F}_{j,k}^{(1)}$ and $\mathcal{F}_{j,k}^{(2)}$ are disjoint and their union is equal to $\mathcal{F}_{j,k}$, i.e., $\mathcal{F}_{j,k}^{(1)} \cap \mathcal{F}_{j,k}^{(2)} = \emptyset$ and $\mathcal{F}_{j,k}^{(1)} \cup \mathcal{F}_{j,k}^{(2)} = \mathcal{F}_{j,k}$.

Assume that the filters $\bar{G}_{i,i}(f)$ are chosen equal to the filters $\tilde{G}_{i,i}(f)$ for all i and all frequencies, except for

$$\begin{aligned} \left|\bar{G}_{j,j}(f)\right|^2 \bar{\lambda}_j^{(N)}(f) &= \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f) \quad \text{and} \\ \left|\bar{G}_{k,k}(f)\right|^2 \bar{\lambda}_k^{(N)}(f) &= \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f), \quad \forall f \in \mathcal{F}_{j,k}^{(1)} \end{aligned} \tag{2.41}$$

and except for

$$\begin{aligned} \left|\bar{G}_{j,j}(f)\right|^2 \bar{\lambda}_j^{(N)}(f) &= \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f) \quad \text{and} \\ \left|\bar{G}_{k,k}(f)\right|^2 \bar{\lambda}_k^{(N)}(f) &= \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f), \quad \forall f \in \mathcal{F}_{j,k}^{(2)}. \end{aligned} \tag{2.42}$$

Since there is a strict inequality in Equation (2.35), it is seen from Equations (2.36) and (2.41) that the filters $\bar{G}_{j,j}(f)$ and $\tilde{G}_{j,j}(f)$ and the filters $\bar{G}_{k,k}(f)$ and $\tilde{G}_{k,k}(f)$ are not equal in $\mathcal{F}_{j,k}^{(1)}$. The choice in Equation (2.42) can be rewritten as $\left|\bar{G}_{j,j}(f)\right| = \left|\tilde{G}_{k,k}(f)\right|$ and $\left|\bar{G}_{k,k}(f)\right| = \left|\tilde{G}_{j,j}(f)\right|$ for all frequencies in $\mathcal{F}_{j,k}^{(2)}$, because of Equation (2.36).

The left hand side of the bit constraint given in Equation (2.34) can be rewritten in the following way when the $\tilde{G}_{i,i}(f)$ filters are used,

$$\begin{aligned} &\sum_{i=0}^{M-1} \ln \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f)\, df \\ &= \sum_{i=0}^{M-1} \ln \left(\int_{\mathcal{F}_{j,k}^C} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f)\, df \right. \\ &\qquad \left. + \int_{\mathcal{F}_{j,k}^{(1)}} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f)\, df + \int_{\mathcal{F}_{j,k}^{(2)}} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f)\, df \right). \end{aligned} \tag{2.43}$$

When the $\bar{G}_{i,i}(f)$ filters are used, the bit constraint can be rewritten as

$$\begin{aligned} &\sum_{i=0}^{M-1} \ln \int_{-\frac{1}{2}}^{\frac{1}{2}} \left|\bar{G}_{i,i}(f)\right|^2 \bar{\lambda}_i^{(N)}(f)\, df \\ &= \sum_{i=0}^{M-1} \ln \left(\int_{\mathcal{F}_{j,k}^C} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}_i^{(N)}(f)\, df \right. \\ &\qquad \left. + \int_{\mathcal{F}_{j,k}^{(1)}} \left|\bar{G}_{i,i}(f)\right|^2 \bar{\lambda}_i^{(N)}(f)\, df + \int_{\mathcal{F}_{j,k}^{(2)}} \left|\bar{G}_{i,i}(f)\right|^2 \bar{\lambda}_i^{(N)}(f)\, df \right), \end{aligned} \tag{2.44}$$

where the fact that $\bar{G}_{i,i}(f)$ is equal to $\tilde{G}_{i,i}(f)$ for frequencies not belonging to $\mathcal{F}_{j,k}$ is used.

To verify that the expression in Equation (2.43) is greater than or equal to the expression in Equation (2.44), the strategy is to assume that the number of bits used with the $\tilde{G}_{i,i}(f)$ filters is greater than or equal to the number of bits spent by using the $\bar{G}_{i,i}(f)$ filters. Then, it is shown that this inequality is equivalent to a valid inequality.

Define the following quantity

$$\tilde{\sigma}^2_{y_i} = \int_{\mathcal{F}^C_{j,k} \cup \mathcal{F}^{(1)}_{j,k}} \left|\tilde{G}_{i,i}(f)\right|^2 \tilde{\lambda}^{(N)}_i(f)\, df, \quad i \in \{j,k\}. \tag{2.45}$$

Assume that the expression given in Equation (2.43) is greater than or equal to the similar expression in Equation (2.44). It is observed that the terms where $i \in \{0,1,\ldots,M-1\} \setminus \{j,k\}$ in Equations (2.43) and (2.44) are equal. These terms are then canceled. The assumed inequality in the number of bits used is equivalent to

$$\begin{aligned}\ln\left(\tilde{\sigma}^2_{y_j} + \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}^{(N)}_j(f)\, df\right) + \ln\left(\tilde{\sigma}^2_{y_k} + \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}^{(N)}_k(f)\, df\right) \geq \\ \ln\left(\tilde{\sigma}^2_{y_j} + \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}^{(N)}_k(f)\, df\right) + \ln\left(\tilde{\sigma}^2_{y_k} + \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}^{(N)}_j(f)\, df\right),\end{aligned} \tag{2.46}$$

where the results from Equations (2.41) and (2.42) are used. Equation (2.46) can be rewritten as the following,

$$\left(\tilde{\sigma}^2_{y_j} - \tilde{\sigma}^2_{y_k}\right)\left(\int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}^{(N)}_k(f)\, df - \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}^{(N)}_j(f)\, df\right) \geq 0. \tag{2.47}$$

The expression in the second factor is non-negative because from the definition of the set $\mathcal{F}^{(2)}_{j,k}$, in Equation (2.40), it follows that

$$\int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}^{(N)}_k(f)\, df \geq \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}^{(N)}_j(f)\, df. \tag{2.48}$$

Because of the numbering of the subband signals in Equation (1.18), $\sigma^2_{y_j} \geq \sigma^2_{y_k}$, and this is equivalent to

$$\tilde{\sigma}^2_{y_j} + \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}^{(N)}_j(f)\, df \geq \tilde{\sigma}^2_{y_k} + \int_{\mathcal{F}^{(2)}_{j,k}} \left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}^{(N)}_k(f)\, df. \tag{2.49}$$

By adding Equations (2.48) and (2.49) the following inequality is found:

$$\tilde{\sigma}_{y_j}^2 \geq \tilde{\sigma}_{y_k}^2. \tag{2.50}$$

Because of the result in Equation (2.50), the first factor in Equation (2.47) is also non-negative, and the assumed inequality holds.

By inserting the choice made in Equations (2.41) and (2.42) into Equation (2.38), the difference between the integrands of the block MSE can be rewritten as

$$\mathcal{E}(f) - \bar{\mathcal{E}}(f) = \begin{cases} \left(\tilde{\lambda}_k^{(N)}(f) - \tilde{\lambda}_j^{(N)}(f)\right) \cdot \\ \left(\frac{\sigma_q^2}{\left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f) + \sigma_q^2} - \frac{\sigma_q^2}{\left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f) + \sigma_q^2}\right), & f \in \mathcal{F}_{j,k}^{(1)}, \\ 0, & f \in \mathcal{F}_{j,k}^{C} \cup \mathcal{F}_{j,k}^{(2)}. \end{cases} \tag{2.51}$$

The right hand side of Equation (2.51) is non-negative due to Equation (2.35) and because when $f \in \mathcal{F}_{j,k}^{(1)}$,

$$\frac{\sigma_q^2}{\left|\tilde{G}_{k,k}(f)\right|^2 \tilde{\lambda}_k^{(N)}(f) + \sigma_q^2} > \frac{\sigma_q^2}{\left|\tilde{G}_{j,j}(f)\right|^2 \tilde{\lambda}_j^{(N)}(f) + \sigma_q^2}. \tag{2.52}$$

Now, case (i) is completely treated.

Next assume case (ii). From Equations (2.33) and (2.34), it is seen that the systems using $\tilde{\lambda}_i^{(N)}(f)$ and $\bar{\lambda}_i^{(N)}(f)$ have identical performance. Thus, no optimality is lost in case (ii).

For case (iii), using related arguments as in case (i), it can be shown that the performance by using $\bar{\lambda}_i^{(N)}(f)$ instead of $\tilde{\lambda}_i^{(N)}(f)$ renders no loss in optimality.

For every pair of indices satisfying Equation (2.35), the process shown above can be performed, and this can be repeated until the ordering is as in Equation (2.7), i.e., $\boldsymbol{\Pi}(f) = \boldsymbol{I}$. Therefore, it is proven that the optimal solution can be found among diagonal matrices $\boldsymbol{G}(f)$.

2.1.3.6 Block MSE and Bit Constraint Expressions

Considering diagonal $\boldsymbol{G}(f)$ matrices, one can rewrite the block MSE as

$$\mathcal{E}_{N,M} = \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{M-1} \frac{\lambda_i^{(N)}(f)\sigma_q^2}{|G_{i,i}(f)|^2 \lambda_i^{(N)}(f) + \sigma_q^2}\, df + \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=M}^{N-1} \lambda_i^{(N)}(f)\, df. \tag{2.53}$$

When the matrix $\boldsymbol{G}(f)$ is diagonal, the bit constraint can be rewritten as

$$\sum_{i=0}^{M-1} \ln \int_{-\frac{1}{2}}^{\frac{1}{2}} |G_{i,i}(f)|^2 \lambda_i^{(N)}(f)\, df = \ln(\beta), \tag{2.54}$$

where β is the constant defined in Equation (2.3), while the inequality constraints can be expressed as

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} |G_{i,i}(f)|^2 \lambda_i^{(N)}(f)\, df \geq \sigma_q^2, \quad i \in \{0, 1, \ldots, M-1\}. \tag{2.55}$$

2.1.3.7 The Elements of $\boldsymbol{G}(f)$

The matrix $\boldsymbol{G}(f)$ is an $M \times N$ diagonal matrix. The diagonal elements of $\boldsymbol{G}(f)$ can be found by the Kuhn-Tucker conditions, with objective function given in Equation (2.53) and constraints in Equations (2.54) and (2.55). By using variational calculus on this problem, the squared magnitudes of the diagonal elements of the matrix $\boldsymbol{G}(f)$ are obtained as

$$|G_{i,i}(f)|^2 = \max\left(0, \sqrt{\frac{\sigma_{y_i}^2 \sigma_q^2}{\left(\mu - \theta_i \sigma_{y_i}^2\right) \lambda_i^{(N)}(f)}} - \frac{\sigma_q^2}{\lambda_i^{(N)}(f)}\right), \quad i \in \{0, 1, \ldots, M-1\}, \tag{2.56}$$

where μ is a Kuhn-Tucker parameter for the bit constraint in Equation (2.54), θ_i is the Kuhn-Tucker parameter corresponding to constraint number i in Equation (2.55), and $\sigma_{y_i}^2$ is the variance in the ith subband, which can be found by solving the following equation:

$$\sigma_{y_i}^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} \max\left(0, \sqrt{\frac{\sigma_{y_i}^2 \sigma_q^2 \lambda_i^{(N)}(f)}{\mu - \theta_i \sigma_{y_i}^2}} - \sigma_q^2\right) df, \quad i \in \{0, 1, \ldots, M-1\}. \tag{2.57}$$

It can be shown that Equation (2.57) in general has three solutions, but because of the inequality constraint in Equation (2.55), only solutions greater than σ_q^2 are valid. The non-negative parameters θ_i have to satisfy Equation (2.14).

The phase of the elements in $\boldsymbol{G}(f)$, see Equation (2.56), can be chosen arbitrarily.

If the expression for $|G_{i,i}(f)|^2$ given in Equation (2.56) is substituted into the block MSE expression of Equation (2.53), the following result is obtained:

$$\mathcal{E}_{N,M} = \sum_{i=0}^{M-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} \min\left(\lambda_i^{(N)}(f), \sqrt{\frac{\left(\mu - \theta_i \sigma_{y_i}^2\right) \sigma_q^2 \lambda_i^{(N)}(f)}{\sigma_{y_i}^2}}\right) df + \sum_{i=M}^{N-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} \lambda_i^{(N)}(f)\, df. \tag{2.58}$$

For a given target average bit rate b, the objective is to find the parameters μ and $\theta_i \geq 0$, given implicitly through Equations (2.14), (2.56), and (2.57), that satisfy Equation (2.54). These values of μ and θ_i are then inserted into Equation (2.58) to calculate the block MSE.

2.1.3.8 Frequency Response Expressions

From [Vaidyanathan & Mitra 1988], it can be shown that eigenvector number i of the PSD matrix $\boldsymbol{S_x}(f)$ in Equation (1.8) can be chosen as

$$\boldsymbol{U}_i(f) = \frac{1}{\sqrt{N}} \left[1, \rho_i(f), \rho_i^2(f), \ldots, \rho_i^{N-1}(f)\right]^T, \quad i \in \mathbb{Z}_N, \tag{2.59}$$

where $\rho_i(f) = e^{-j2\pi \frac{\left(f + l_i^{(N)}(f)\right)}{N}}$, and $l_i^{(N)}(f)$ is the ith *ordering function*. The ordering functions $l_i^{(N)} : \mathbb{R} \longrightarrow \mathbb{Z}_N$ have been introduced to make sure that the ordering in Equation (2.7) is maintained for all frequencies. In Appendix B, it is shown that the ordering functions $l_i^{(N)}(f)$ satisfy the following equation:

$$\lambda_i^{(N)}(f) = S_x\left(\frac{f + l_i^{(N)}(f)}{N}\right), \quad i \in \mathbb{Z}_N. \tag{2.60}$$

Some of the properties of the ordering functions are derived in Appendix B, and from these properties the ordering functions can be calculated. The vector given in Equation (2.59) is column number i of the unitary matrix $\boldsymbol{U}(f)$ in Equation (2.5), so the norm of this vector is 1.

If the noble identities [Vaidyanathan 1993] are used, the decimators and the expanders can be moved after the analysis and before the synthesis polyphase matrix, respectively, see Figure 1.1. Then, the frequency responses of the analysis filter $H_m(f)$ and synthesis filter $F_m(f)$ can be found from Equations (2.5)

and (2.59). The analysis filter $H_m(f)$ can then be expressed as

$$\begin{aligned} H_m(f) &= \sum_{k=0}^{N-1} \frac{G_{m,m}(fN)}{\sqrt{N}} e^{j2\pi k\left(\frac{fN+l_m^{(N)}(fN)}{N}\right)} e^{-j2\pi k} \\ &= \frac{G_{m,m}(fN)}{\sqrt{N}} \sum_{k=0}^{N-1} e^{j2\pi k \frac{l_m^{(N)}(fN)}{N}} \\ &= \sqrt{N} G_{m,m}(fN) \frac{1}{N} \sum_{k=0}^{N-1} e^{j2\pi k \frac{l_m^{(N)}(fN)}{N}} \\ &= \sqrt{N} G_{m,m}(fN) \delta(l_m^{(N)}(fN)), \quad m \in \mathbb{Z}_M, \end{aligned} \tag{2.61}$$

where $\delta(\cdot)$ is the Krönecker delta function.

For high rates, it can be seen from Equations (2.56) and (2.60) that the analysis filters in Equation (2.61) approach half-whitening filters. Half whitening filters are treated in [Jayant & Noll 1984, Aase & Ramstad 1995].

Using similar arguments, it can be shown that $F_m(f)$ can be expressed as

$$F_m(f) = \sqrt{N} G_{m,m}^*(fN) \delta(l_m^{(N)}(fN)) \frac{\lambda_m^{(N)}(fN) e^{-j2\pi f(N-1)}}{|G_{m,m}(fN)|^2 \lambda_m^{(N)}(fN) + \sigma_q^2}, \quad m \in \mathbb{Z}_M, \tag{2.62}$$

where the superscript * denotes complex conjugation.

In Equation (2.61), the analysis filters have the same phase as the $G_{m,m}(fN)$ filters, and from Equation (2.62), it is seen that the phase of the synthesis filters is given by the phase of the $G_{m,m}^*(fN)$ filters plus the extra linear phase given by the fraction in Equation (2.62). As mentioned before, the phase of the $G_{m,m}(f)$ filters in Equation (2.56) can be chosen arbitrarily. From the filter expressions in Equations (2.61) and (2.62), it is concluded that the group delay through the system is independent of the phase of the $G_{m,m}(f)$ filters and is always $N-1$.

In order to make a comparison to the optimal unitary filter banks, the frequency response of the optimal filter can be derived from [Vaidyanathan 1998]. Using the notation introduced in this dissertation, analysis filter number m is given by:

$$H_m(f) = \sqrt{N} \delta(l_m^{(N)}(fN)), \quad m \in \mathbb{Z}_M \tag{2.63}$$

and synthesis filter number m by:

$$F_m(f) = \sqrt{N} \delta(l_m^{(N)}(fN)) e^{-j2\pi f(N-1)}, \quad m \in \mathbb{Z}_M. \tag{2.64}$$

If Equations (2.63) and (2.64) are compared to Equations (2.61) and (2.62), respectively, it is seen that the filters have the same non-zero frequency regions, but the shaping of the proposed filter banks are very different from the optimal unitary filter bank. In the passbands, the amplitude response of the optimal unitary filters are constant, while the proposed filter banks have non-constant amplitude response in the passband, see Equations (2.61) and (2.62).

The frequency passbands of the optimal biorthogonal filter banks are the same as the frequency passbands in the optimal unitary filter banks and the proposed filter banks, but the shaping of the passbands are equal to half-whitening, see [Aase & Ramstad 1995, Aas & Mullis 1996, Vaidyanathan & Kıraç 1998, Moulin et al. 2000] for details.

2.1.3.9 Aliasing Noise

It will now be shown that the proposed filter banks are alias free.

A necessary and sufficient condition for aliasing cancellation is that the total polyphase matrix $\boldsymbol{W}(f) = \boldsymbol{R}(f)\boldsymbol{E}(f)$ is pseudocirculant [Vaidyanathan & Mitra 1988]. The results in Equation (2.5) lead to

$$\boldsymbol{W}(f) = \boldsymbol{R}(f)\boldsymbol{E}(f) = \boldsymbol{U}(f)\boldsymbol{T}(f)\boldsymbol{G}(f)\boldsymbol{U}^H(f). \tag{2.65}$$

From Equation (2.11), it is seen that $\boldsymbol{T}(f)$ is a diagonal matrix since it is a product of diagonal matrices. Therefore, the product $\boldsymbol{T}(f)\boldsymbol{G}(f)$ is also a diagonal matrix. If the input signal to the analysis filter bank is a WSS scalar time series, it is shown in [Sathe & Vaidyanathan 1993] that the PSD matrix $\boldsymbol{S_x}(f)$ will be pseudocirculant. In [Vaidyanathan & Mitra 1988], it is shown that the eigenvalue matrix of every pseudocirculant matrix can be chosen as

$$\boldsymbol{U}(f) = \boldsymbol{\Gamma}(f)\boldsymbol{\Psi}(f), \tag{2.66}$$

where $\boldsymbol{\Gamma}(f)$ is a diagonal matrix with $[\boldsymbol{\Gamma}(f)]_{k,k} = e^{-j2\pi f\frac{k}{N}}$ and $\boldsymbol{\Psi}(f)$ is an $N \times N$ DFT matrix with the columns permuted according to the ordering function $l_n^{(N)}(f)$, i.e., $[\boldsymbol{\Psi}(f)]_{m,n} = \frac{1}{\sqrt{N}} e^{-j2\pi\frac{m \cdot l_n^{(N)}(f)}{N}}$. It can be seen that this is equivalent to Equation (2.59), where the ith column of $\boldsymbol{U}(f)$ is given. Therefore, the matrix $\boldsymbol{W}(f)$ can be expressed as

$$\boldsymbol{W}(f) = \boldsymbol{\Gamma}(f)\boldsymbol{\Psi}(f)\boldsymbol{T}(f)\boldsymbol{G}(f)\boldsymbol{\Psi}^{-1}(f)\boldsymbol{\Gamma}^{-1}(f). \tag{2.67}$$

From [Vaidyanathan & Mitra 1988], it can be shown that matrices which can be factored as the right hand side of Equation (2.67) are pseudocirculant. This is just a renumbering of the case which is considered in [Vaidyanathan & Mitra 1988]. This concludes the proof that there is no aliasing error in the proposed filter banks.

2.1.3.10 High Rate Case

The high rate case is equivalent to $\sigma_q^2 \to 0^+$. In the high rate case, synthesis filter number $m \in \mathbb{Z}_M$ is given by

$$\lim_{\sigma_q^2 \to 0^+} F_m(f) = \sqrt{N} \frac{\delta(l_m^{(N)}(fN))}{G_{m,m}(fN)} e^{-j2\pi f(N-1)}. \tag{2.68}$$

By comparing Equations (2.61) and (2.68), it is seen that the synthesis filters are scaled and delayed versions of the pseudo-inverse [Jain 1989] analysis filters. Since high rates are used, all the quantizers will receive bits, so $M = N$.

It is already shown that the unconstrained length filter banks have no alias error. Therefore, the overall transfer function through the system is well defined. In the high rate case, the overall transfer function can be expressed as [Vaidyanathan 1993]

$$\lim_{\sigma_q^2 \to 0^+} \frac{1}{N} \sum_{m=0}^{N-1} H_m(f) F_m(f) = \frac{1}{N} \sum_{m=0}^{N-1} N \delta(l_m^{(N)}(fN)) e^{-j2\pi f(N-1)} = e^{-j2\pi f(N-1)}, \tag{2.69}$$

where it is used that $\sum_{m=0}^{N-1} \delta(l_m^{(N)}(fN)) = 1$ for all frequencies f. This follows from the definition of the ordering functions, see Appendix B. From Equation (2.69), it is seen that in the high rate case the overall transfer function represents a PR system with group delay $N-1$.

In this dissertation, no PR constraints are set on the filter banks. However, in the high rate case ($\sigma_q^2 \to 0^+$), it is shown above that the optimal solution is found among the PR filters. In [Aas & Mullis 1996, Vaidyanathan & Kıraç 1998], it was conjectured that the optimal PR analysis filter bank has the structure of the polyphase matrix $\boldsymbol{U}^H(f)$ followed by a *diagonal* polyphase matrix $\boldsymbol{G}(f)$, and that the optimal PR synthesis filter bank is given by the inverse filter bank, i.e., the *diagonal* polyphase matrix $\boldsymbol{G}^{-1}(f)$ followed by the polyphase matrix $\boldsymbol{U}(f)$. In [Moulin et al. 2000], it is shown that this is indeed optimal among PR filter banks. The optimal PR filter banks are rate independent if optimal bit allocation is used because, in this case, the problem of minimizing the MSE for a given number of bits is equivalent to maximizing the coding gain, and the coding gain is rate independent for PR filter banks [Vaidyanathan & Kıraç 1998]. From Subsections 2.1.3.4, 2.1.3.5, and the current subsection, it can be concluded that the results found in this section give an alternative proof of the conjecture.

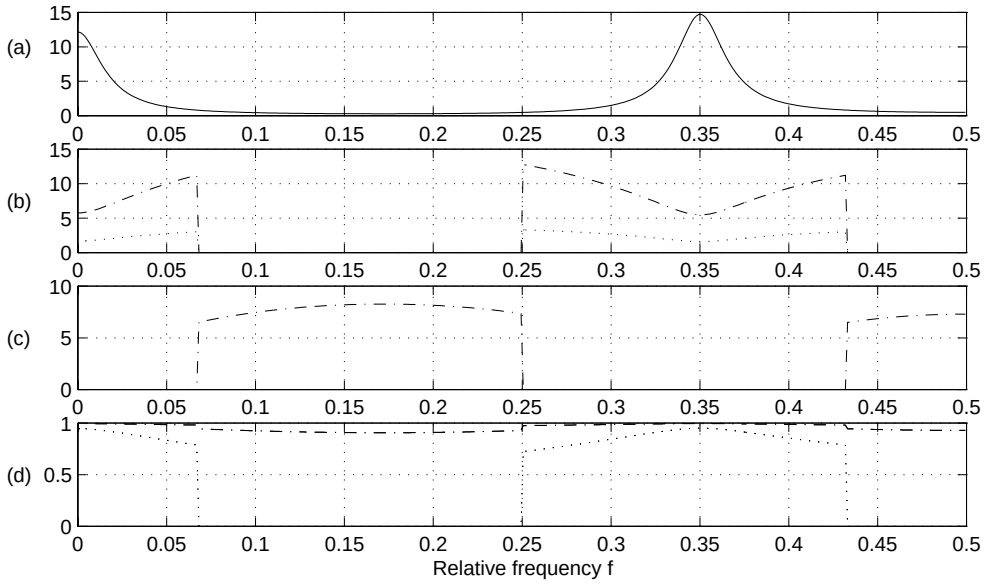

Figure 2.2 System example for the unconstrained length filter banks, using $N = 2$. (a) PSD of the input signal $S_x(f)$. (b) Frequency response of the first analysis filter $H_0(f)$. (c) Frequency response of the second analysis filter $H_1(f)$. (d) Overall frequency response. In the last three plots, the dash-dotted curves give the result with $b = 3.28$ bits/sample and SNR = 18.41 dB, while the dotted curves show the result with $b = 1.12$ bits/sample and SNR = 7.51 dB.

The analysis and synthesis filter bank defined by the polyphase matrices $\boldsymbol{U}^H(f)$ and $\boldsymbol{U}(f)$, respectively, are called *principal component filter banks* [Tsatsanis & Giannakis 1995]. These filter banks are also equal to the optimal unitary filter banks [Vaidyanathan 1998] with unconstrained length.

2.1.4 Bit Constrained Filter Bank Results

In all the results presented in this subsection, the following choice has been made: $\sigma_q^2 = 1$.

To illustrate some of the results derived in this section, consider the two channel case, $N = 2$. The input signal is a Gaussian AR(3) process with the poles located at 0.9, $0.9e^{-j2\pi 0.35}$, and $0.9e^{j2\pi 0.35}$, whose PSD is shown in Figure 2.2 (a). Since the source is Gaussian and the filter bank is a linear operator, the pdf of the inputs to the quantizers are also Gaussian. In all the results presented in this section, pdf optimized scalar quantizers are used to

quantize the subband signals, thus the coding coefficient is $c_i = \frac{\sqrt{3}\pi}{2}$ [Jayant & Noll 1984].

The frequency responses of the first and second analysis filters are shown in Figure 2.2 (b) and (c), respectively. The dash-dotted curves show the filters when $b = 3.28$ bits/sample, and the dotted curves represent the responses when $b = 1.12$ bits/sample. The phases of the $G_{m,m}(f)$ filters have been chosen to be zero, so the analysis filters have zero phase and the synthesis filters will have linear phase, see Subsection 2.1.3.8.

Observe from Figure 2.2 (b) that both filters extract two frequency bands. From parts (b) and (c) of the figures, it is seen that for a fixed frequency, only one of the filter responses in the same filter bank is different from zero. It was shown in Subsection 2.1.3.9 that the proposed unconstrained length filter banks are free from alias error. Therefore, the overall transfer function through the system is well defined, and this function is shown in part (d) of the figure. Figure 2.2 (d) reveals that the optimal filter banks do not have the PR property. It should also be observed that in frequency intervals where the input signal has low energy, no signal is transmitted if the average bit rate is low. It can be proven that the N-fold decimation of the analysis frequency responses $H_i(f)$ does not create aliasing.

Figure 2.3 shows the performance of four systems in terms of SNR per source sample (in dB) vs. rate (in bits per sample) with $N = 3$ subbands. The input signal in this case is a Gaussian AR(1) process with correlation coefficient 0.9. The distortion rate function is found in [Berger 1971], while the performance of the optimal biorthogonal system is found using the theory in [Aas & Mullis 1996, Vaidyanathan & Kıraç 1998, Moulin et al. 2000]. The performance of the optimal unitary system is obtained from the theory developed in [Vaidyanathan 1998]. From the figure, it can be observed that the proposed unconstrained length filter bank system outperforms the optimal unconstrained length unitary and biorthogonal systems at all rates.

With the white signal independent noise model used for the quantizers, see Equation (1.13), and the constraint $\sigma_{y_i}^2 \geq \sigma_q^2$, the smallest number of bits that can be allocated to *one* quantizer is $\frac{1}{2}\log_2 c_i$, see Equation (1.14). Therefore, the minimum *average* number of bits having a positive SNR is $\frac{1}{2N}\log_2 c_i$. When coding a Gaussian input signal using pdf optimized scalar quantizers and $N = 3$, an SNR $= 0$ dB is obtained when the average bit rate is less than $\frac{1}{2N}\log_2 c_i \approx 0.24$ bits/sample. This is in accordance with Figure 2.3. From the figure, it can be seen that the performance curves are not smooth for all rates. When the average bit rate is decreased, the number of quantizers that receive a positive number of bits M is reduced, and this gives rise to the unsmooth characteristic of the performance curves.

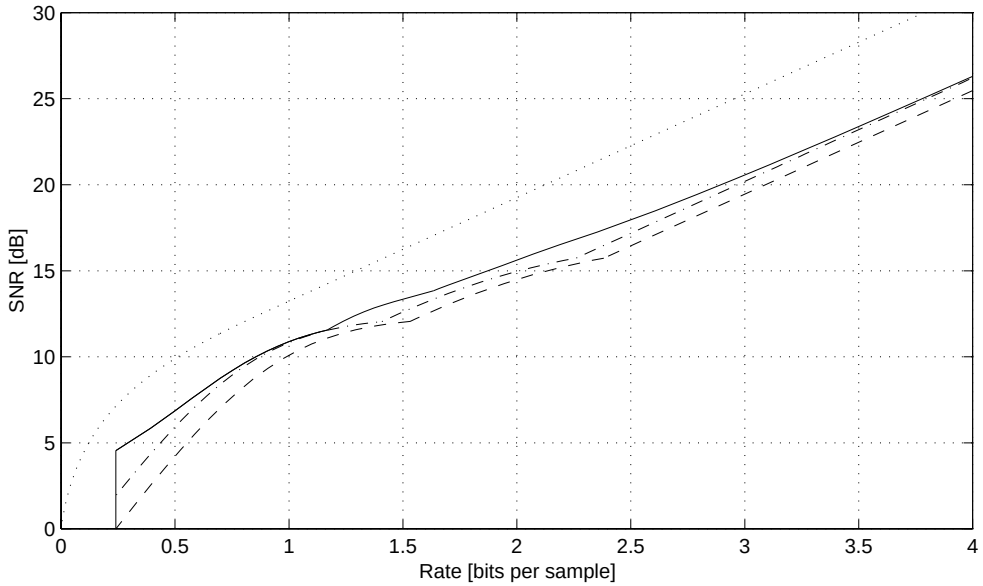

Figure 2.3 Different distortion rate performances when coding a Gaussian AR(1)-process with correlation factor 0.9 and $N = 3$ subbands. The solid cure shows the performance of the proposed unconstrained length filter bank system, the dash-dotted curve represents the optimal unconstrained length biorthogonal system, while the dashed curve shows the optimal unconstrained length unitary system performance. The dotted curve shows the distortion rate function.

At high rates, Figure 2.3 illustrates that the performance of the optimal biorthogonal system is approximately the same as the proposed unconstrained length filter bank system. This is because, at high rates, the quantization noise is small, and therefore, the filter banks should be close to PR, see Subsection 2.1.3.10.

2.2 Power Constrained Filter Banks

In this section, the problem of transmitting a vector time series over a power constrained vector channel with a known transfer matrix and signal independent noise is considered. The goal is to find jointly optimal transmitter and receiver polyphase matrices which minimize the block MSE between the original and reconstructed vector time series when only a limited amount of power is used. The polyphase matrices are allowed to be non-causal with infinite lengths.

This section is organized as follows: In Subsection 2.2.1 the problem is formulated. Subsection 2.2.2 contains an explanation of how the optimal solution can be obtained. A combined source-channel coding problem is introduced in Subsection 2.2.3, and some results obtained by the proposed filter banks in the combined source-channel coding problem are presented in Subsection 2.2.4.

2.2.1 Problem Formulation

In order to state the problem, expressions are needed for the block MSE and the power used by the channel input vector.

By manipulating the formulas as was done in the bit constrained case in Appendix A for infinite length filters, the following expression for the block MSE in the frequency domain is obtained

$$\mathcal{E}_{N,M} = \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \Big\{ [\boldsymbol{I} - \boldsymbol{R}(f)\boldsymbol{C}(f)\boldsymbol{E}(f)]\, \boldsymbol{S}_{\boldsymbol{x}}(f)\, [\boldsymbol{I} - \boldsymbol{R}(f)\boldsymbol{C}(f)\boldsymbol{E}(f)]^H + \boldsymbol{R}(f)\boldsymbol{S}_{\boldsymbol{v}}(f)\boldsymbol{R}^H(f) \Big\}\, df \right), \tag{2.70}$$

where $\boldsymbol{E}(f)$, $\boldsymbol{C}(f)$, and $\boldsymbol{R}(f)$ are the Fourier transform of the transmitter, channel, and receiver impulse response matrices, respectively, evaluated on the unit circle. The matrix $\boldsymbol{C}(f)$ is assumed to be known. In the notation $\mathcal{E}_{N,M}$, the numbers N and M refers to the dimensions of the vectors used in Figure 1.3. The values of N and M are arbitrary in this section.

In Appendix A, the following expression for the power constraint in the frequency domain is derived:

$$\mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{E}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{E}^H(f)\, df \right) = P, \tag{2.71}$$

where P is the average power used by the channel input vector $\boldsymbol{y}(n)$.

The objective is to minimize the block MSE given by Equation (2.70) with respect to $\boldsymbol{E}(f)$ and $\boldsymbol{R}(f)$, subject to the power constraint in Equation (2.71).

2.2.2 Optimal Matrix System

In this section, the joint optimal transmitter matrix $\boldsymbol{E}(z)$ and receiver matrix $\boldsymbol{R}(z)$ are derived.

The problem which is considered here corresponds to the problem solved in [Yang & Roy 1994], Section II, B, where a MIMO *band-limited continuous time* system is studied. The results below are the discrete time expressions for

the corresponding discrete problem. The objective function in Equation (2.70) and the constraint in Equation (2.71) are equivalent to Equations (10) and (11), respectively, in [Yang & Roy 1994]. This equivalence is obtained if the parameter T used in [Yang & Roy 1994] is chosen as $T = 1$, and the noise PSD matrix used in [Yang & Roy 1994] must be changed from the continuous time case to the discrete time PSD matrix in Equation (1.21). The optimal discrete time solution can therefore be found from [Yang & Roy 1994] with these modifications. The optimal solution for the discrete time unconstrained length jointly optimal transmitter and receiver filter banks for communication over a power constrained linear vector channel with additive noise will now be given.

The optimal transmitter and receiver matrices can be written as

$$\begin{aligned} \boldsymbol{E}(f) &= \boldsymbol{V}(f)\boldsymbol{G}(f)\boldsymbol{U}^H(f), \\ \boldsymbol{R}(f) &= \boldsymbol{U}(f)\boldsymbol{T}(f)\boldsymbol{V}^H(f)\boldsymbol{C}^H(f)\boldsymbol{\Lambda}_{\boldsymbol{v}}^{-1}(f). \end{aligned} \tag{2.72}$$

$\boldsymbol{U}(f)$ and $\boldsymbol{V}(f)$ are unitary matrices which diagonalize the input PSD matrix $\boldsymbol{S}_{\boldsymbol{x}}(f)$ and the Hermitian matrix $\boldsymbol{C}^H(f)\boldsymbol{S}_{\boldsymbol{v}}^{-1}(f)\boldsymbol{C}(f)$, respectively, i.e.,

$$\begin{aligned} \boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{U}(f) &= \boldsymbol{U}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f), \\ \boldsymbol{C}^H(f)\boldsymbol{S}_{\boldsymbol{v}}^{-1}(f)\boldsymbol{C}(f)\boldsymbol{V}(f) &= \boldsymbol{V}(f)\boldsymbol{\Lambda}_{\boldsymbol{v}}^{-1}(f). \end{aligned} \tag{2.73}$$

In Equation (2.73), $\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$ and $\boldsymbol{\Lambda}_{\boldsymbol{v}}^{-1}(f)$ are diagonal matrices that contain the eigenvalues of $\boldsymbol{S}_{\boldsymbol{x}}(f)$ and $\boldsymbol{C}^H(f)\boldsymbol{S}_{\boldsymbol{v}}^{-1}(f)\boldsymbol{C}(f)$, respectively. In addition, the elements of $\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$ and $\boldsymbol{\Lambda}_{\boldsymbol{v}}(f)$ are ordered as follows:

$$\begin{aligned} &\lambda_0^{(N)}(f) \geq \lambda_1^{(N)}(f) \geq \cdots \geq \lambda_{N-1}^{(N)}(f), \quad \text{and} \\ &\kappa_0^{(M)}(f) \leq \kappa_1^{(M)}(f) \leq \cdots \leq \kappa_{M-1}^{(M)}(f), \end{aligned} \tag{2.74}$$

respectively. The matrix $\boldsymbol{G}(f)$ is an $M \times N$ diagonal matrix where the magnitude of the diagonal elements are given by the square root of

$$|G_{i,i}(f)|^2 = \max\left(0, \sqrt{\frac{\kappa_i^{(M)}(f)}{\mu\lambda_i^{(N)}(f)}} - \frac{\kappa_i^{(M)}(f)}{\lambda_i^{(N)}(f)}\right), i \in \{0, 1, \ldots, \min(M,N) - 1\}, \tag{2.75}$$

where μ is a Lagrange multiplier for the constraint optimization problem. The phase of the elements in $\boldsymbol{G}(f)$ can be chosen arbitrarily.

The matrix $\boldsymbol{T}(f)$ in Equation (2.72) can be expressed as

$$\boldsymbol{T}(f) = \boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f)\left[\boldsymbol{G}(f)\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda}_{\boldsymbol{v}}(f)\right]^{-1}\boldsymbol{\Lambda}_{\boldsymbol{v}}(f). \tag{2.76}$$

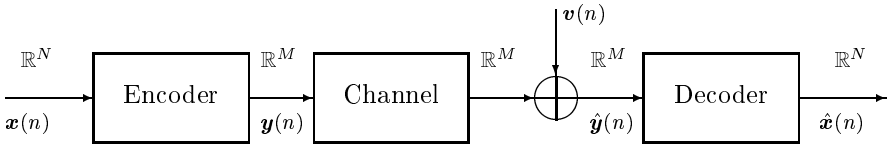

Figure 2.4 The combined source-channel coding problem.

The performance of the optimal system measured by block MSE, see Equation (1.12), is found to be

$$\mathcal{E}_{N,M} = \begin{cases} \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{N-1} \frac{\lambda_i^{(N)}(f)\kappa_i^{(M)}(f)}{|G_{i,i}(f)|^2\lambda_i^{(N)}(f) + \kappa_i^{(M)}(f)}\, df, & \text{if } M \geq N, \\ \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{M-1} \frac{\lambda_i^{(N)}(f)\kappa_i^{(M)}(f)}{|G_{i,i}(f)|^2\lambda_i^{(N)}(f) + \kappa_i^{(M)}(f)}\, df + \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=M}^{N-1} \lambda_i^{(N)}(f)\, df, & \text{if } M < N. \end{cases} \tag{2.77}$$

The constraint on the power used per channel input vector can be expressed in the following way:

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{i=0}^{\min(M,N)-1} |G_{i,i}(f)|^2 \lambda_i^{(N)}(f)\, df = P. \tag{2.78}$$

For a given target power P in Equation (2.78), the objective is to find the Lagrange multiplier μ, given implicitly through Equation (2.75), which satisfies Equation (2.78). This value of μ is then inserted in Equation (2.77) and the block MSE is found.

2.2.3 A Combined Source-Channel Coding Problem

Combined source-channel coding is a promising topic in the search for optimal communication systems. This in spite of "Shannon's separation theorem" [Shannon 1948, Shannon 1959, Vembu, Verdú & Steinberg 1995], which states that the source and channel coders can be optimized separately. The reason why there is still hope for improvements in practical systems is that the separation theorem requires infinite delay and thus infinite complexity. For finite delays and complexities, the situation is more complex.

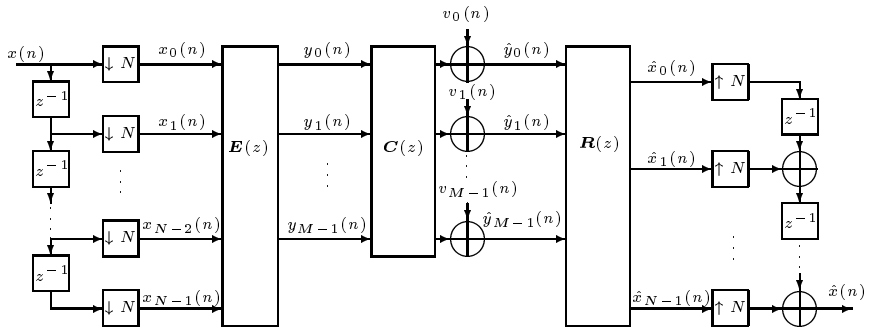

Figure 2.5 The MIMO system used to solve the combined source-channel coding problem.

In this subsection, the MIMO system developed in Subsection 2.2.2 is used to solve a combined source-channel coding problem, which is illustrated in Figure 2.4. The problem is to transmit a *scalar* time series over a *scalar* channel having the scalar transfer function $C(z)$ and additive signal independent Gaussian white noise. N source samples are transmitted using M channel samples. The optimal solution of the combined source-channel coding problem is nonlinear. However, in this dissertation, only the optimal *linear* filter bank solutions are studied. In the current chapter, the delay and complexity complexity of the unconstrained length filter banks are infinite. The study of these filter banks is mostly of theoretical interest. Their performance provide an upper bound on the SNR vs. CSNR performance of the linear transform and FIR solutions found in Sections 3.2 and 4.2, respectively.

The linear MIMO system developed in Subsection 2.2.2 can be applied to the combined source-channel coding problem in Figure 2.4, by redrawing Figure 1.3, as shown in Figure 2.5. The system input signal $x(n)$ is multiplexed into vectors $\boldsymbol{x}(n)$, according to

$$\boldsymbol{x}(n) = [x(nN), x(nN-1), \ldots, x(nN-(N-1))]^T. \tag{2.79}$$

The channel noise vector is given by

$$\boldsymbol{v}(n) = [v_0(n), v_1(n), \ldots, v_{M-1}(n)]^T, \tag{2.80}$$

where $v_i(n)$ is Gaussian white noise with known variance σ_v^2 when considering the combined source-channel coding problem. In Figure 2.5, the channel transfer matrix $\boldsymbol{C}(z)$ can be found from the scalar transfer function $C(z)$, by letting the first row in $\boldsymbol{C}(z)$ be the polyphase components of $C(z)$. The matrix $\boldsymbol{C}(z)$ should be pseudocirculant, therefore, the rest of the rows in $\boldsymbol{C}(z)$ are given by the first row, see Section 10.1 in [Vaidyanathan 1993] for details.

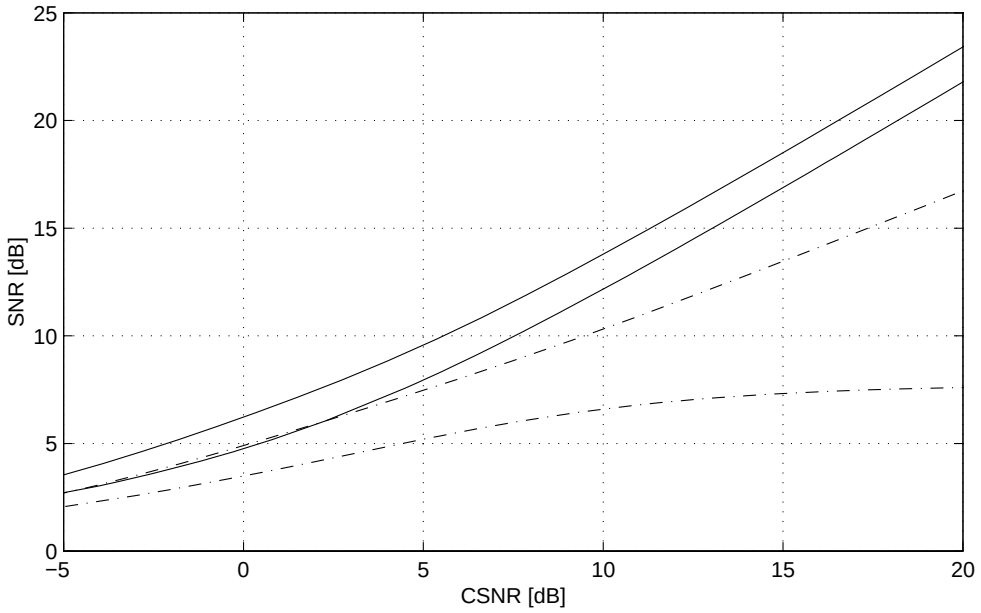

Figure 2.6 The SNR vs. CSNR performance of the proposed system and OPTA using $N = 3$, $M = 3$, and $C(z) = 1$ is shown by the solid curves. The upper curve is OPTA. The dash-dotted curves show the performance of the proposed system and OPTA using $N = 3$, $M = 2$. The upper curve is OPTA. The input source is a Gaussian AR(3) with PSD shown in Figure 2.2 (a), and the channel noise is white and Gaussian.

2.2.4 Power Constrained Filter Bank Results

As an example of using the optimized MIMO system found in Subsection 2.2.2 for solving the combined source-channel coding problem in Subsection 2.2.3, consider the cases $N = 3$, $M = 2$ and $N = 3$, $M = 3$. The source which is transmitted is a Gaussian AR(3) process for which the PSD is given in Figure 2.2 (a), and the channel noise is assumed to be white Gaussian with unit variance, i.e., $\sigma_v^2 = 1$. The components of the additive noise vector are therefore uncorrelated. The channel transfer function is $C(z) = 1$.

The system performance is shown in Figure 2.6. In the figure, CSNR is the channel signal to noise ratio $(P/M)/\sigma_v^2$, and SNR is the signal to overall reconstruction noise ratio $\sigma_x^2/(\mathcal{E}_{N,M}/N)$. Both SNR and CSNR are expressed in dB in the figure. In Figure 2.6, the results obtained from the theory developed in Subsection 2.2.2 are compared to the optimal performance theoretically attainable (OPTA) curve for a Gaussian signal. The OPTA curve is found by

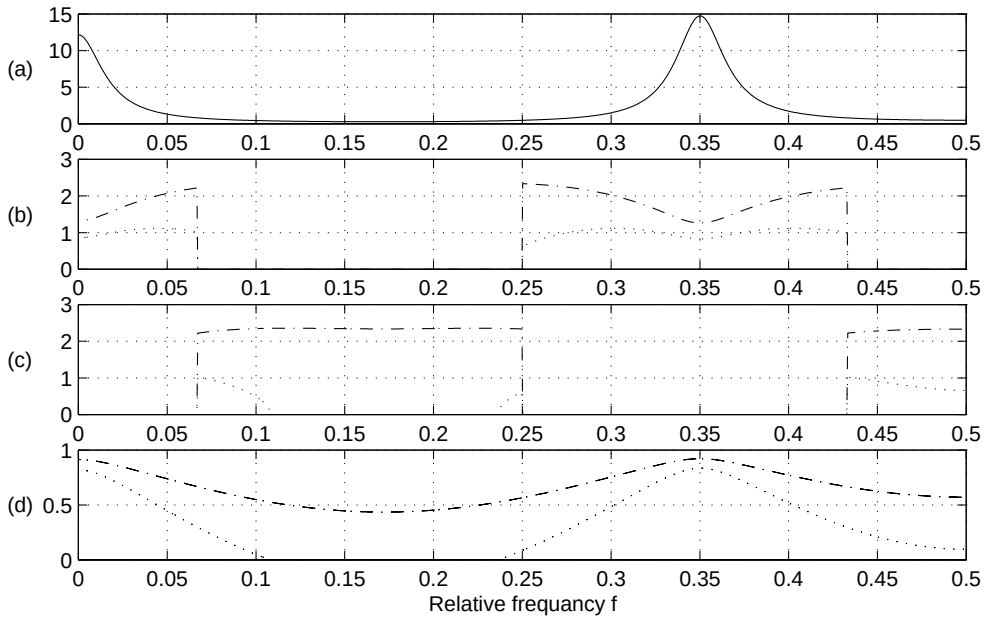

Figure 2.7 System example using $N = 2$, $M = 2$, and $C(z) = 1$. (a) PSD of the input signal $S_x(f)$. (b) Frequency response of the first channel on the transmitter side $H_0(f)$. (c) Frequency response of the second channel on the transmitter side $H_1(f)$. (d) Total frequency response through the system. In the last three plots, the dotted curves give the result with CSNR $= -0.30$ dB and SNR $= 4.61$ dB, while the dash-dotted curves show the result with CSNR $= 4.80$ dB and SNR $= 7.79$ dB.

evaluating the distortion rate function of a source at the channel capacity function [Berger 1971]. From Figure 2.6, it can be observed that the MIMO system performs well for poor channels and not well for CSNR > 10 dB, when the number of channel symbols are less than the number of source symbols, i.e., $N > M$. When $N = 3$, $M = 3$ the system performs well for all CSNRs.

When the number of channel samples per source sample is reduced, the least important transfer functions are set to zero and only the most important ones are used. The reason why the performance of $N = 3$, $M = 3$ case is not equal to the $N = 3$, $M = 2$ case for low CSNR, is that CSNR is measured as power used per *channel* symbol. When the number of channel samples is greater than the number of source samples, i.e., $N < M$, $M - N$ of the filters are set to zero. This is the best one can achieve using the linear structure chosen here.

By using the noble identities, the decimators can be moved after the

polyphase matrix on the transmitter side. The frequency responses of the filters in the transmitter filter banks can then be found by taking the delay chain in front of the polyphase matrix into account. This was done for the bit constrained filter banks in Subsection 2.1.3.8.

Figure 2.7 (a) shows the PSD of the Gaussian AR(3) process having poles located at 0.9, $0.9e^{-j2\pi 0.35}$, and $0.9e^{j2\pi 0.35}$. The channel noise is assumed to be white Gaussian with uncorrelated noise components with unit variance. The channel transfer function is $C(z) = 1$. The frequency responses of the first and second transmitter filter are shown in Figure 2.7 (b) and (c), respectively, using $N = M = 2$. The dash-dotted curves show the transmitter filters when CSNR = 4.80 dB, and the dotted curves represent the responses when CSNR = -0.30 dB. It can shown in the same way as in Subsection 2.1.3.9 that the power constrained proposed unconstrained length filter banks are free from alias error when using the identity channel matrix. Therefore, the overall transfer function through the system is well defined, and this function is shown in part (d) of the figure.

Notice from Figure 2.7 (b) and (c) that both filters extract two frequency bands. It can be shown that the N-fold decimation of the transmitter frequency responses $H_i(f)$ does not create aliasing. From Figure 2.7 (d), it is seen that the optimal filter banks do not have the perfect reconstruction property. The phase of the analysis filters are arbitrary, but can be chosen linear, then the synthesis filters will also have linear phase. It is also seen from the figure that in frequency intervals where the input signal has low energy, no signal is sent through the system if the average transmitted power used is low.

2.3 Summary

In the first part of this chapter, bit constrained jointly optimal analysis and synthesis filter banks having unconstrained filter lengths were derived. A subband coder structure was optimized with respect to the minimum block MSE between the output and the input signals under a bit constraint. The filters were allowed to be non-causal with infinite impulse responses. To simplify the optimization, an optimal MIMO system was first derived.

The second part of the chapter contains a derivation of the jointly optimal discrete time power constrained transmitter and receiver filter bank having unconstrained filter lengths. This is deduced from the corresponding continuous time solution found in [Yang & Roy 1994]. This system was applied to a combined source-channel coding problem.

Chapter 3

Signal-Adaptive Transforms

A linear transform can be viewed as a special case of a filter bank [Vaidyanathan 1993] since it functions like a filter bank with a memoryless polyphase matrix. When using transform coding, the input time series $x(n)$ is divided into blocks of length N, and each block is transformed into a block of length M by a memoryless analysis transform matrix.

The problem of jointly optimizing the analysis and synthesis transforms under a bit and power constraint is treated in this chapter. No PR conditions are imposed on the transforms.

The current chapter is organized as follows: In Section 3.1, the problem of jointly optimal analysis and synthesis transforms under a bit constraint is treated. An alternative derivation of the optimal solution under a power constraint is described in Section 3.2. Finally, a brief summary is given in Section 3.3.

This chapter is partly based on [Hjørungnes & Ramstad 1998*b*, Hjørungnes & Ramstad 1999*c*].

3.1 Bit Constrained Transforms

The Karhunen-Loève transform (KLT) has optimal distortion rate performance among PR transforms for a given WSS signal, but in this section a transform that outperforms the KLT is proposed. In the proposed transform, the PR constraint is relaxed. The resulting transform coding system performs at least as well as the KLT coding system for all rates and sources.

Transform coders are used in image and video compression standards, e.g., [CCITT Rec. T.81 1992, ISO/IEC IS 11172 1995, ISO/IEC IS 13818 1998]. It is therefore an important task to evaluate their performance and derive optimal solutions.

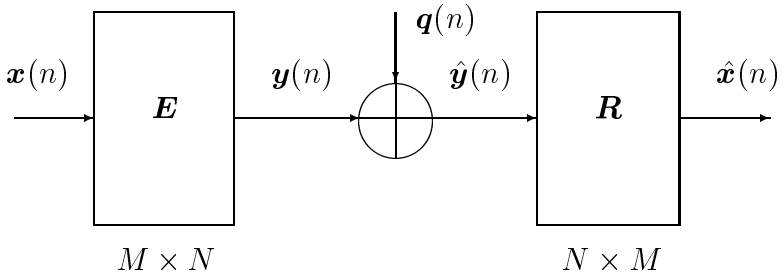

Figure 3.1 Transform coder model.

This section is organized as follows: The assumptions and the problem treated are stated in Subsection 3.1.1, while in Subsection 3.1.2, the optimal solution is derived. In Subsection 3.1.3, the optimal transform is compared to the KLT. Conditions for PR being the optimal transform are stated in Subsection 3.1.4. Analytical expressions for jointly optimal Wiener transform and bit allocation are derived in Subsection 3.1.5, when the analysis transform is a reduced rank KLT. Finally, Subsection 3.1.6 contains results using the proposed transform.

3.1.1 Problem Formulation

In a transform coder model, the analysis and synthesis polyphase matrices are memoryless, so the system treated in this section is shown in Figure 3.1. The analysis and synthesis transform matrices are denoted $\boldsymbol{E}$ and $\boldsymbol{R}$, respectively, and the dimensions of these matrices are $M \times N$ and $N \times M$. These polyphase matrices are independent of z, contrary to unconstrained length and FIR filter banks, see Chapters 2 and 4. It is assumed that $N \geq M$.

The same quantizer model as introduced in Chapter 1 will be used here.

Since the mean of the input is assumed to be zero, the autocovariance matrix at lag zero for the vector $\boldsymbol{x}(n)$ is equal to the autocorrelation matrix at lag zero, and this matrix is given by:

$$\boldsymbol{K}_{\boldsymbol{x}}(0) = E\left[\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right]. \tag{3.1}$$

If the memoryless matrices $\boldsymbol{E}$ and $\boldsymbol{R}$ are inserted into Equation (2.2), the block MSE for the transform coder can be expressed as

$$\mathcal{E}_{N,M} = \mathrm{Tr}\left\{(\boldsymbol{I} - \boldsymbol{RE})\,\boldsymbol{K}_{\boldsymbol{x}}(0)\,(\boldsymbol{I} - \boldsymbol{RE})^H + \boldsymbol{R}\boldsymbol{\Lambda}_{\boldsymbol{q}}\boldsymbol{R}^H\right\}, \tag{3.2}$$

where it has been used that $\boldsymbol{K}_{\boldsymbol{x}}(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S}_{\boldsymbol{x}}(f)\, df$, which follows from the inverse Fourier transformation of Equation (1.8).

Using the same reasoning as above, the bit constraint in Equation (2.3) can be rewritten as

$$\Pr\left(\boldsymbol{E}\boldsymbol{K}_{\boldsymbol{x}}(0)\boldsymbol{E}^H\right) = \beta. \tag{3.3}$$

The constraints given in Equation (2.4) have to be satisfied as well.

The bit constrained transform problem is to find the bit distribution and matrices $\boldsymbol{E}$ and $\boldsymbol{R}$ which minimize the block MSE in Equation (3.2), subject to the bit constraint in Equation (3.3) and the constraints in Equation (2.4).

3.1.2 Optimal Transform Coder

In this subsection, the optimal transform is derived from the results for unconstrained length filter banks given in Section 2.1.

Assume that the input signal to the unconstrained length filter bank found in Section 2.1 can be characterized by the following autocovariance matrix sequence:

$$\boldsymbol{K}_{\boldsymbol{x}}(m) = E\left[\boldsymbol{x}(n+m)\boldsymbol{x}^H(n)\right] = \boldsymbol{K}_{\boldsymbol{x}}(0)\delta(m), \tag{3.4}$$

where $\delta(m)$ is the Krönecker delta function. Equation (3.4) indicates that the input *vectors* are uncorrelated, but there may be correlation between the components within a vector. By inserting Equation (3.4) into Equation (1.8) the PSD matrix of the input is found as

$$\boldsymbol{S}_{\boldsymbol{x}}(f) = \boldsymbol{K}_{\boldsymbol{x}}(0). \tag{3.5}$$

That is, the PSD matrix is *independent* of frequency.

To specify the optimal unconstrained length filter bank solution, the eigenvalues and eigenvectors of the PSD matrix are needed. Let $\boldsymbol{U}$ be an $N \times N$ unitary matrix which diagonalizes the input covariance matrix $\boldsymbol{K}_{\boldsymbol{x}}(0)$, i.e.,

$$\boldsymbol{K}_{\boldsymbol{x}}(0)\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}_{\boldsymbol{x}}. \tag{3.6}$$

In Equation (3.6), $\boldsymbol{\Lambda}_{\boldsymbol{x}}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{K}_{\boldsymbol{x}}(0)$. The elements of $\boldsymbol{\Lambda}_{\boldsymbol{x}}$ are ordered as follows:

$$\lambda_0^{(N)} \geq \lambda_1^{(N)} \geq \cdots \geq \lambda_{N-1}^{(N)}. \tag{3.7}$$

Again, it is observed that since the PSD matrix is frequency independent, so are the eigenvalues and eigenvectors.

An important part of the optimal unconstrained length filter bank is the $G_{i,i}(f)$ filters given in Equation (2.56). The eigenvalues of the input PSD matrix are independent of frequency, thus from Equation (2.56), the $G_{i,i}(f)$ filters are also frequency independent. The optimal polyphase matrix $\boldsymbol{G}(f)$ is an $M \times N$ frequency independent diagonal matrix. The diagonal element number i of this matrix is found by solving the following equation:

$$|G_{i,i}|^2 = \max\left(0, \sqrt{\frac{\sigma_{y_i}^2 \sigma_q^2}{\left(\mu - \theta_i \sigma_{y_i}^2\right) \lambda_i^{(N)}}} - \frac{\sigma_q^2}{\lambda_i^{(N)}}\right), \quad i \in \{0, 1, \ldots, M-1\}, \tag{3.8}$$

where μ is a Lagrange multiplier for the bit constraint and θ_i is the Kuhn-Tucker parameter corresponding to constraint number i in Equation (2.4).

Since both the matrices $\boldsymbol{G}(f)$ and $\boldsymbol{U}(f)$ are frequency independent, it is realized by inspection of Equations (2.5) and (2.11), that the optimal unconstrained length analysis and synthesis filter bank is independent of frequency when the input PSD matrix is given by Equation (3.5). The optimal unconstrained length filter banks can be found from Equation (2.5), and are given by the following frequency independent matrices:

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{G}\boldsymbol{U}^H, \\ \boldsymbol{R} &= \boldsymbol{U}\boldsymbol{T}. \end{aligned} \tag{3.9}$$

From Equation (2.11), it is seen that the matrix $\boldsymbol{T}$ is given by

$$\boldsymbol{T} = \boldsymbol{\Lambda_x} \boldsymbol{G}^H \left[\boldsymbol{G}\boldsymbol{\Lambda_x}\boldsymbol{G}^H + \boldsymbol{\Lambda_q}\right]^{-1}. \tag{3.10}$$

From this equation, it is observed that the matrix $\boldsymbol{T}$ is frequency independent and diagonal because it is a product and sum of frequency independent diagonal matrices.

The autocovariance matrix of the subband coefficients at lag zero is given by

$$\begin{aligned} \boldsymbol{K_y}(0) &= E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right] = E\left[\boldsymbol{G}\boldsymbol{U}^H\boldsymbol{x}(n)\boldsymbol{x}^H(n)\boldsymbol{U}\boldsymbol{G}^H\right] \\ &= \boldsymbol{G}\boldsymbol{U}^H\boldsymbol{K_x}(0)\boldsymbol{U}\boldsymbol{G}^H = \boldsymbol{G}\boldsymbol{\Lambda_x}\boldsymbol{G}^H. \end{aligned} \tag{3.11}$$

Since the input vectors are uncorrelated, see Equation (3.4), $\boldsymbol{K_y}(0)$ is also equal to the PSD matrix of the subband vector $\boldsymbol{y}(n)$. Equation (3.11) shows that the subband coefficients are uncorrelated since the matrix $\boldsymbol{K_y}(0)$ is diagonal. From Equation (3.11), it is deduced that subband variance number i is given

by $\sigma_{y_i}^2 = |G_{i,i}|^2 \lambda_i^{(N)}$. If this expression for the subband variance is inserted into Equation (3.8), the expression for $|G_{i,i}|^2$ can be simplified to

$$|G_{i,i}|^2 = \max\left(0, \sqrt{\frac{|G_{i,i}|^2 \sigma_q^2}{\mu - \theta_i |G_{i,i}|^2 \lambda_i^{(N)}}} - \frac{\sigma_q^2}{\lambda_i^{(N)}}\right), \quad i \in \{0, 1, \ldots, M-1\}. \tag{3.12}$$

The solution stated above, which is frequency *independent*, is the optimal unconstrained length filter bank minimizing the block MSE given in Equation (2.2) under the constraints given in Equations (2.3) and (2.4), with input PSD matrix given in Equation (3.5). By comparing Equations (2.2) and (2.3) using $\boldsymbol{S_x}(f) = \boldsymbol{K_x}(0)$, with Equations (3.2) and (3.3), respectively, it is seen that the optimal unconstrained length filter bank solution has to be the optimal transform matrices as well. The reason is that the optimization of the unconstrained length problem is performed over a set that includes the whole set of transform coders as a proper subset, and when it turns out that the optimal unconstrained length filter bank is indeed a transform, this is the optimal transform.

It can be shown that the optimal synthesis matrix $\boldsymbol{R}$ given in Equation (3.9), is a Wiener transform matrix [Vaidyanathan & Chen 1994], i.e., it can be written as:

$$\boldsymbol{R} = \boldsymbol{K}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(0) \boldsymbol{K}_{\hat{\boldsymbol{y}}}^{-1}(0), \tag{3.13}$$

where the cross-correlation matrix $\boldsymbol{K}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(0) = E\left[\boldsymbol{x}(n)\hat{\boldsymbol{y}}^H(n)\right]$, and the autocorrelation matrix $\boldsymbol{K}_{\hat{\boldsymbol{y}}}(0) = E\left[\hat{\boldsymbol{y}}(n)\hat{\boldsymbol{y}}^H(n)\right]$. In [Vaidyanathan & Chen 1994], the derivation of the Wiener transform was obtained by using the orthogonality principle [Therrien 1992].

3.1.2.1 Performance Expressions

The performance evaluated by the block MSE when using jointly optimal analysis and synthesis transforms can be found from Equation (2.53), and it can be expressed as

$$\mathcal{E}_{N,M} = \sum_{i=0}^{M-1} \frac{\lambda_i^{(N)} \sigma_q^2}{|G_{i,i}|^2 \lambda_i^{(N)} + \sigma_q^2} + \sum_{i=M}^{N-1} \lambda_i^{(N)}. \tag{3.14}$$

The first sum in Equation (3.14) represents the M transform coefficients that will receive bits, and the last sum represents the remaining $N - M$ transform coefficients that do not receive bits. Therefore, if $M = N$, the last sum of Equation (3.14) is equal to zero.

The expression for the bit constraint in the optimal transform case can be derived from Equation (2.54), and it is given by

$$\sum_{i=0}^{M-1} \ln\left(|G_{i,i}|^2 \lambda_i^{(N)}\right) = \ln(\beta), \tag{3.15}$$

where β is the constant defined in Equation (2.3), and $|G_{i,i}|^2$ is found by solving Equation (3.12).

3.1.2.2 Linear Phase

The impulse responses of the ith analysis and synthesis filter are given by the ith row and the reversed ith column of the matrices $\boldsymbol{E}$ and $\boldsymbol{R}$ given in Equation (3.9), respectively. Since the matrices $\boldsymbol{G}$ and $\boldsymbol{T}$ are diagonal matrices, the phase of the impulse responses will be decided by the phase of the column vectors of the matrix $\boldsymbol{U}$, see Equation (3.9). The columns of the matrix $\boldsymbol{U}$ contain the eigenvectors of the autocovariance matrix $\boldsymbol{K_x}(0)$, see Equation (3.6). The autocovariance matrix $\boldsymbol{K_x}(0)$ is a double symmetric matrix for real input signals $x(n)$, and in [Makhoul 1981], it is shown that it is always possible to choose the eigenvectors of double symmetric matrices either symmetric or skew symmetric[1]. Therefore, it is always possible to choose the impulse responses in the optimal transform to have linear phase.

3.1.3 Comparisons to the KLT

Figure 3.2 shows a block diagram of the KLT and the proposed transform coder model. From the figure, it is seen that the differences between the two systems are the diagonal matrices $\boldsymbol{G}$ and $\boldsymbol{T}$.

From Equation (3.9), it is seen that the analysis filter $H_i(z)$ and the synthesis filter $F_i(z)$ of the transform filters are given by:

$$\begin{aligned} H_i(z) &= G_{i,i} \sum_{k=0}^{N-1} U_{k,i}^* z^{-k}, \\ F_i(z) &= T_{i,i} \sum_{k=0}^{N-1} U_{N-1-k,i} z^{-k}, \end{aligned} \tag{3.16}$$

where $i \in \{0, 1, \ldots, M-1\}$, $U_{k,i}$ is the element in row number k and column number i of the matrix $\boldsymbol{U}$, and $G_{i,i}$ and $T_{i,i}$ are diagonal elements number i in the diagonal matrices $\boldsymbol{G}$ and $\boldsymbol{T}$, respectively.

[1] A skew symmetric impulse response corresponds to an odd impulse response, and therefore, the phase of the filter is linear.

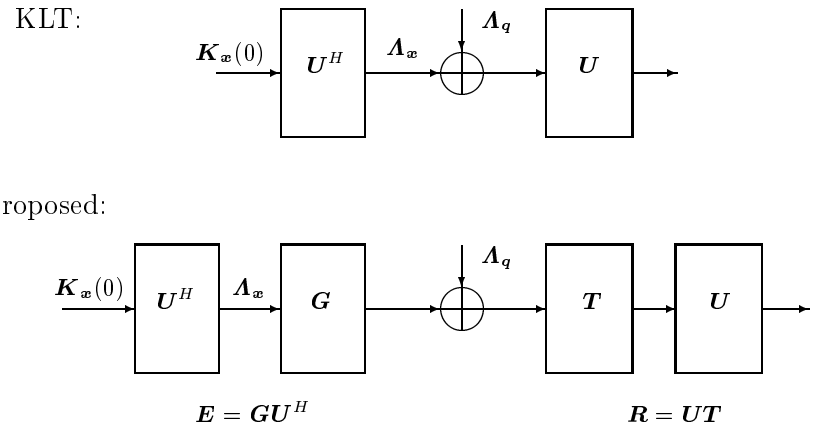

Figure 3.2 KLT and proposed transform coder models.

From Equation (3.16), it is observed that for real input signals, the zeros of analysis filter number i have the exact same positions as the zeros of synthesis filter number i, since the eigenvectors of the matrix $\boldsymbol{K}_{\boldsymbol{x}}(0)$ can always be chosen symmetric or skew symmetric for real input signals, see Subsection 3.1.2.2. By studying Equation (3.16), it is seen that the zeros have the same positions as for the KLT filters. The filters in the KLT and the proposed transform coder are therefore equal except the gain factor in each filter, see Equation (3.16). These gain factors are chosen in an optimal manner in the proposed transform coder.

If M quantizers are used to code N source samples, it can be shown that the bit optimal allocation for the KLT can be expressed as:

$$b_i = b' + \frac{1}{2}\log_2 \frac{c_i \lambda_i^{(N)}}{\left(\prod_{k=0}^{M-p-1} c_k \lambda_k^{(N)}\right)^{\frac{1}{M}}}, \tag{3.17}$$

where $b' = \frac{N}{M-p}b - \frac{1}{2(M-p)} \sum_{i=M-p}^{M-1} \log_2 c_i$, and $p \leq M$ is the number of times the inequality $\sigma_{y_i}^2 \geq \sigma_q^2$ holds with equality. When KLT is used, $\sigma_{y_i}^2 = \lambda_i^{(N)}$. If $p = 0$, the last sum in the formula of b' should be set equal to zero. The number of quantizers M must be found through a discrete optimization.

It can be shown that in the proposed transform coder, the optimal bit

allocation is given by

$$b_i = b' + \frac{1}{2}\log_2 \frac{c_i \lambda_i^{(N)}}{\left(\prod_{k=0}^{M-p-1} c_k \lambda_k^{(N)}\right)^{\frac{1}{M}}} + \frac{1}{2}\log_2 \frac{|G_{i,i}T_{i,i}|^2}{\left(\prod_{k=0}^{M-p-1} |G_{k,k}T_{k,k}|^2\right)^{\frac{1}{M}}}, \quad (3.18)$$

where b' and p are defined as above. From Equations (3.17) and (3.18), it is seen that the difference in bit allocation in the two systems is due to the last term in Equation (3.18). The KLT coder has the PR property, so if the matrices $\boldsymbol{G}$ and $\boldsymbol{T}$ were included in the KLT system, $|G_{i,i}T_{i,i}| = 1$ for all i. Therefore, the last term in Equation (3.18) would be equal to zero.

The total transfer matrix $\boldsymbol{RE}$ is Hermitian in the proposed transforms. This can be proved by considering

$$\begin{aligned}(\boldsymbol{RE})^H &= \boldsymbol{U}\boldsymbol{G}^H \left[\boldsymbol{G}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{G}^H + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]^{-1} \boldsymbol{G}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{U}^H \\ &= \boldsymbol{U}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{G}^H \left[\boldsymbol{G}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{G}^H + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]^{-1} \boldsymbol{G}\boldsymbol{U}^H = \boldsymbol{RE}, \end{aligned} \quad (3.19)$$

where it is used that the $N \times N$ matrices $\boldsymbol{G}^H \left[\boldsymbol{G}\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{G}^H + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right]^{-1} \boldsymbol{G}$ and $\boldsymbol{\Lambda}_{\boldsymbol{x}}$ commute since they are diagonal.

It can be shown that when the input time series $x(n)$ is real, the total transfer matrix through the transform coder $\boldsymbol{RE}$ is symmetric around both the main and the secondary diagonal, so the total transfer matrix is centrosymmetric [Makhoul 1981]. This can be shown by finding expressions for the elements of the matrix $\boldsymbol{RE}$ and comparing elements from opposite sides of the secondary diagonal. The fact that the columns of the matrix $\boldsymbol{U}$ can be chosen symmetric or skew symmetric must be utilized when comparing the expressions.

3.1.4 Conditions for Optimality of PR in the Transform Case

In this subsection, conditions for when PR is optimal will be stated for a given invertible analysis transform, and it will be shown that this is never the case for the quantization model assumed in this chapter.

In [Vaidyanathan & Chen 1994], it was shown for a given invertible analysis transform $\boldsymbol{E}$, the optimal transform system has the PR property if, and only if, the following condition is satisfied:

$$E\left[\hat{\boldsymbol{y}}(n)\boldsymbol{q}^H(n)\right] = \boldsymbol{0} \ \ \forall n, \quad (3.20)$$

where the vector $\hat{\boldsymbol{y}}(n)$ is the input of the synthesis transform matrix, see Figure 3.1. It can be shown that the condition in Equation (3.20) is satisfied if

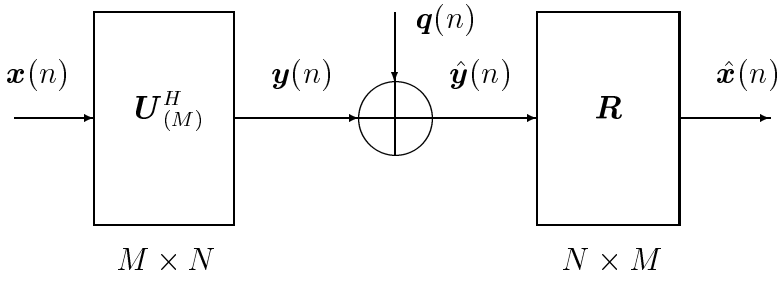

Figure 3.3 Reduced rank analysis KLT with Wiener synthesis transform.

vector quantizers are used for coding of the subband vector $\boldsymbol{y}(n)$, and if the centroid of the pdf $f_{\boldsymbol{y}(n)}$ is used as the representation level of the Voronoi region [Gersho & Gray 1992, Vaidyanathan & Chen 1994].

In this section, it is assumed that the quantization noise $\boldsymbol{q}(n)$ is uncorrelated with the input signal to the quantizers $\boldsymbol{y}(n)$, and it is also assumed that the additive quantization noise is white with uncorrelated components in different subbands. With this quantization noise model, the cross-correlation matrix in Equation (3.20) is given by:

$$E\left[\hat{\boldsymbol{y}}(n)\boldsymbol{q}^H(n)\right] = E\left[\left(\boldsymbol{y}(n) + \boldsymbol{q}(n)\right)\boldsymbol{q}^H(n)\right] = \boldsymbol{\Lambda}_{\boldsymbol{q}} \neq \boldsymbol{0}. \tag{3.21}$$

Therefore, with the quantization noise model used in this section a PR transform system will never be optimal for finite bit rates.

In [Huang & Schultheiss 1963], it is assumed that scalar Lloyd-Max quantizers are used to quantize each transform coefficient and that the input time series is Gaussian. Furthermore, it is assumed that the analysis transform will transform the Gaussian input vector $\boldsymbol{x}(n)$ to a vector containing uncorrelated vector components. In the mentioned reference, it is showed that the KLT is optimal. The KLT has the PR property. This result is in accordance with the conditions for PR being an optimal transform, given in Equation (3.20). The reason for this is that when a Gaussian vector is linearly transformed to a vector with uncorrelated components, the components are statistically independent. Then, the Lloyd-Max *vector quantizer* and the Lloyd-Max scalar quantizers will be equivalent provided that the decision regions of the vector quantizers are given by the Cartesian product [Truss 1991] of the scalar quantizers' decision intervals. The condition in Equation (3.20) is therefore satisfied for the assumptions made in [Huang & Schultheiss 1963]. Extensions of the results found in [Huang & Schultheiss 1963] can be found in [Segall 1976, Jain 1989], and these results are also in accordance with Equation (3.20).

3.1.5 Wiener Transformed KLT

In this subsection, analytical expressions for jointly optimal Wiener transform and bit allocation is derived when a reduced rank analysis KLT [Scharf & Tufts 1987] is used. The system studied is shown in Figure 3.3. The analysis transform is given by the $M \times N$ matrix $\boldsymbol{U}_{(M)}^H$. The matrix $\boldsymbol{U}_{(M)}$ contains the first M eigenvectors of the matrix $\boldsymbol{K}_{\boldsymbol{x}}(0)$, i.e., the M first columns of the matrix $\boldsymbol{U}$, which is given in Equation (3.6). The following equation will be useful:

$$\boldsymbol{K}_{\boldsymbol{x}}(0)\boldsymbol{U}_{(M)} = \boldsymbol{U}_{(M)}\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)}, \tag{3.22}$$

where the $M \times M$ matrix $\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)}$ is a diagonal matrix containing the M largest eigenvalues of $\boldsymbol{K}_{\boldsymbol{x}}(0)$.

First, an expression is derived for the Wiener transform given in Equation (3.13) when using reduced rank analysis KLT. The cross-covariance matrix $\boldsymbol{K}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(0)$ is given by

$$\begin{aligned}\boldsymbol{K}_{\boldsymbol{x},\hat{\boldsymbol{y}}}(0) &= E\left[\boldsymbol{x}(n)\hat{\boldsymbol{y}}^H(n)\right] = E\left[\boldsymbol{x}(n)\left(\boldsymbol{q}(n) + \boldsymbol{U}_{(M)}^H\boldsymbol{x}(n)\right)^H\right] \\ &= \boldsymbol{K}_{\boldsymbol{x}}(0)\boldsymbol{U}_{(M)} = \boldsymbol{U}_{(M)}\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)},\end{aligned} \tag{3.23}$$

where $E\left[\boldsymbol{x}(n)\boldsymbol{q}^H(n)\right] = \boldsymbol{0}$ and the result from Equation (3.22) has been utilized. The other matrix which is specifying the Wiener transform is $\boldsymbol{K}_{\hat{\boldsymbol{y}}}(0)$, and this matrix is given by

$$\boldsymbol{K}_{\hat{\boldsymbol{y}}}(0) = E\left[\left(\boldsymbol{q}(n) + \boldsymbol{U}_{(M)}^H\boldsymbol{x}(n)\right)\left(\boldsymbol{q}(n) + \boldsymbol{U}_{(M)}^H\boldsymbol{x}(n)\right)^H\right] = \boldsymbol{\Lambda}_{\boldsymbol{q}} + \boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)}, \tag{3.24}$$

where the diagonal matrix $\boldsymbol{\Lambda}_{\boldsymbol{q}} = E\left[\boldsymbol{q}(n)\boldsymbol{q}^H(n)\right]$ does *not* necessarily have equal values on the main diagonal when using reduced rank analysis KLT with Wiener synthesis transform. By substituting the results from Equations (3.23) and (3.24) into Equation (3.13), the Wiener transform matrix $\boldsymbol{R}$ can be expressed as:

$$\boldsymbol{R} = \boldsymbol{U}_{(M)}\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)}\left(\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)} + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right)^{-1} = \boldsymbol{U}_{(M)}\boldsymbol{D}, \tag{3.25}$$

where $\boldsymbol{D} = \boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)}\left(\boldsymbol{\Lambda}_{\boldsymbol{x}}^{(M)} + \boldsymbol{\Lambda}_{\boldsymbol{q}}\right)^{-1}$ is a $M \times M$ diagonal matrix, and the ith diagonal element is denoted $D_{i,i}$. Since $\boldsymbol{D}$ is given by a product of diagonal

matrices, $D_{i,i}$ is given by

$$D_{i,i} = \frac{\lambda_i^{(N)}}{\lambda_i^{(N)} + \sigma_{q_i}^2}, \quad i \in \{0, 1, \ldots, M-1\}. \tag{3.26}$$

It follows from Equation (3.25) that the norm of the ith synthesis filter in the Wiener transform is given by $|D_{i,i}|$. If optimal bit allocation is used, it follows from [Moulin et al. 2000], that the product of the quantization variance $\sigma_{q_i}^2$ and the squared norm of the synthesis filters $|D_{i,i}|^2$ should be constant for each subband. If this non-negative constant is named μ, this condition can be expressed mathematically as:

$$|D_{i,i}|^2 \sigma_{q_i}^2 = \mu, \quad i \in \{0, 1, \ldots, M-p-1\}, \tag{3.27}$$

where $p \leq M$ indicates the number of times the inequality $\sigma_{y_i}^2 \geq \sigma_q^2$ holds with equality.

If $|D_{i,i}|$ is eliminated from Equations (3.26) and (3.27) for $i \in \{0, 1, \ldots, M-p-1\}$, the following expression for $\sigma_{q_i}^2$ is obtained:

$$\sigma_{q_i}^2 = \lambda_i^{(N)} \frac{\lambda_i^{(N)} - 2\mu - \sqrt{\left(\lambda_i^{(N)}\right)^2 - 4\lambda_i^{(N)}\mu}}{2\mu}, \quad i \in \{0, 1, \ldots, M-p-1\}, \tag{3.28}$$

where it has been used that $\sigma_{q_i}^2 \geq 0$ when choosing the sign in front of the square-root operator.

Since the Wiener transform can be written as shown in Equation (3.25), it can be shown that optimal bit allocation in this case can be expressed as

$$b_i = b' + \frac{1}{2}\log_2 \frac{c_i \lambda_i^{(N)}}{\left(\prod_{k=0}^{M-p-1} c_k \lambda_k^{(N)}\right)^{\frac{1}{M}}} + \frac{1}{2}\log_2 \frac{|D_{i,i}|^2}{\left(\prod_{k=0}^{M-p-1} |D_{k,k}|^2\right)^{\frac{1}{M}}}, \tag{3.29}$$

where b' and p are defined as in Equation (3.17).

The system introduced in this section will be called *Wiener transformed KLT*.

3.1.6 Bit Constrained Transform Results

Results for KLT with reduced rank are obtained from [Scharf & Tufts 1987], and the bit allocation is given in Equation (3.17). The DCT is taken from [Jain 1989].

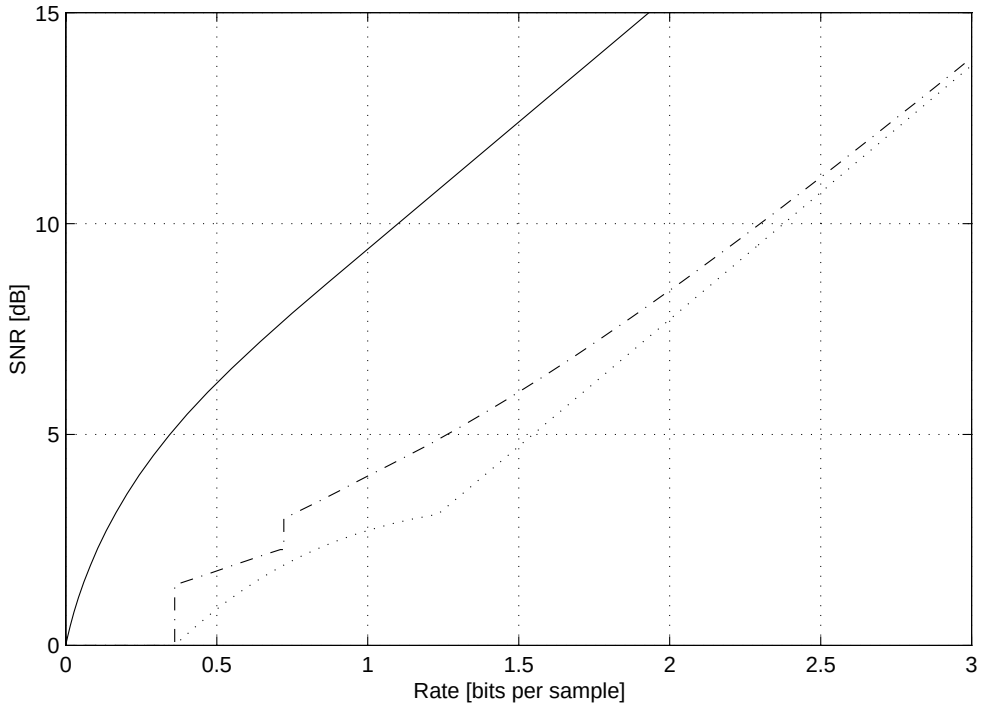

Figure 3.4 The distortion rate function [Berger 1971] of the Gaussian AR(3) source with PSD shown in Figure 2.7 (a) is shown by the solid curve, the performance of the proposed transform coder is shown by the dash-dotted curve, and the dotted curve shows the performance of the KLT, using $N = 2$ and $c_i = \frac{\sqrt{3}\pi}{2}$.

Figure 3.4 shows the distortion rate performance of the proposed transform coder when coding the Gaussian AR(3) source with PSD shown in Figure 2.2 (a). It is assumed that pdf optimized scalar quantizers are used, so $c_i = \frac{\sqrt{3}\pi}{2}$. The performance is compared to the KLT with reduced rank [Scharf & Tufts 1987] and the distortion rate function for the source [Berger 1971]. It is seen from Figure 3.4 that the proposed transform coder outperforms the KLT for all rates. However, the performance is far away from the distortion rate function of the source, which is expected due to the simplicity of these systems.

From Figure 3.4, it is seen that there are sudden changes in the performance of the proposed transform for bit rates at 0.72 bits/sample and 0.36 bits/sample. These bit rates correspond to the minimum positive bit rates that are possible when using $M = 2$ and $M = 1$, respectively, because the minimum average bit rate b when M quantizers receive a positive number

Table 3.1 Theoretical distortion rate performances.

$N = 8$. The input source is Gaussian AR(3) with PSD shown in Figure 2.2 (a), and $c_i = \frac{\sqrt{3}\pi}{2}$.

Type of transform	Bit rate [bits per sample]			
	0.50	1.00	2.00	3.00
	SNR [dB]			
DCT [Jain 1989]	2.37	4.99	9.20	15.02
KLT [Scharf & Tufts 1987]	2.86	5.99	10.39	15.84
Wiener transformed KLT	**3.62**	**6.30**	**10.99**	**16.13**
Proposed transform	**3.62**	**6.30**	**10.99**	**16.13**

of bits is

$$b = \frac{1}{N} \sum_{i=0}^{M-1} \frac{1}{2} \log_2(c_i). \tag{3.30}$$

The minimum average bit rate is obtained when the constraints in Equation (2.4) are satisfied with equality for all $i \in \{0, 1, \ldots, M-1\}$. If the eigenvalues of the autocovariance matrix $\boldsymbol{K}_{\boldsymbol{x}}(0)$ are almost equal, which is the case in Figure 3.4, this sudden breakdown in performance appears. However, if the eigenvalues are very different in size, this breakdown phenomenon does not occur. An example of an input PSD giving eigenvalues that are very different in size is an AR(1) specter with correlation coefficient 0.95.

The performance of the DCT, the KLT, the Wiener transformed KLT, and the proposed transform coder is shown in Table 3.1 using $N = 8$, with the input source being the same Gaussian AR(3) source coded in Figure 3.4.

From Table 3.1, it is seen that the performance of the proposed transform coder is better than the DCT and KLT for all rates. The Wiener transformed KLT has the same performance as the proposed transform. The reason for this is as follows: Since the matrices $\boldsymbol{G}$ and $\boldsymbol{T}$ in the proposed system are diagonal matrices with dimensions $M \times N$ and $N \times M$, respectively, the last $N - M$ rows in $\boldsymbol{U}^H$ in the proposed analysis transform and the last $N - M$ columns in $\boldsymbol{U}$ in the proposed synthesis transform are multiplied by zeros. Therefore, the first and last matrices in the proposed system and the Wiener transformed KLT system are equivalent. The synthesis transform in both systems is a Wiener transform, and no loss of optimality is caused by the Wiener transform. The essential difference between the two systems is therefore shown in Figure 3.5: In the proposed system, the output of the analysis KLT matrix is multiplied by a

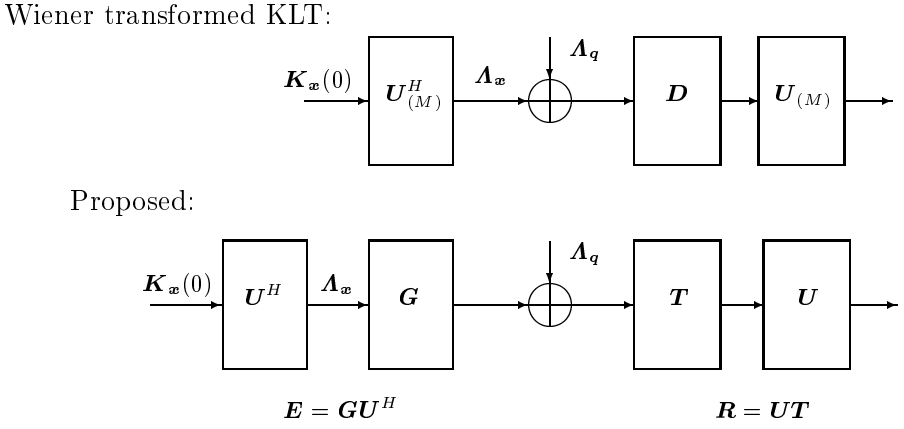

Figure 3.5 Comparison between the Wiener transformed KLT and the proposed transform model.

diagonal matrix $\boldsymbol{G}$, and noise vectors with the same variance in each component are added. In the Wiener transformed KLT system, noise vectors with not necessarily equal variance components are added to the output of the analysis KLT matrix. In the proposed transform, the matrix $\boldsymbol{G}$ is optimized while the elements of $\boldsymbol{\Lambda_q}$ are kept constant, and in the Wiener transformed KLT, the elements of $\boldsymbol{\Lambda_q}$ are optimized while the output of the analysis KLT matrix is unchanged. Since the matrices $\boldsymbol{\Lambda_q}$ and $\boldsymbol{G}$ are diagonal, the optimization of one of them while keeping the other constant is equivalent to keeping the latter constant and optimizing the first matrix.

The proposed system covers the whole set of transforms. In the Wiener transformed KLT, this is not the case since the analysis filters are given by the KLT filters. In the proposed system, the optimization is performed over the set of all transforms, since in the optimization, the matrix $\boldsymbol{G}$ can be any matrix, not only a diagonal matrix. However, the optimization shows that no optimality is lost by constraining the matrix $\boldsymbol{G}$ to be diagonal, and therefore, the optimization is performed over the same set of transforms which is covered by the Wiener transformed KLT system. Since both systems are optimal, they are equivalent.

The proposed transform does not have PR. As an example, let $N = 3$, and let the input signal be the AR(3) source with PSD shown in Figure 2.2 (a).

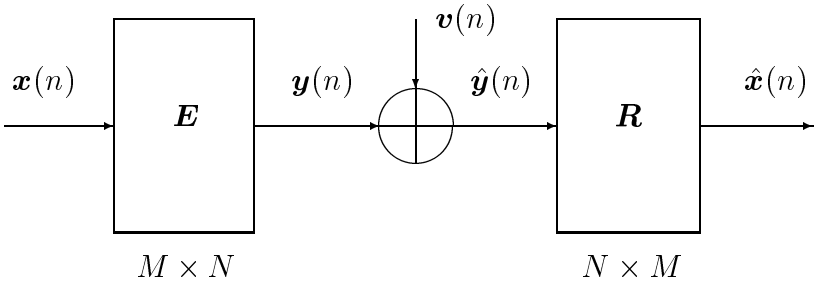

Figure 3.6 Power constrained transform model.

Then, the total transfer matrix through the system can be written as

$$\boldsymbol{RE} = \begin{bmatrix} 0.9999531 & -0.0000054 & 0.0000027 \\ -0.0000054 & 0.9999527 & -0.0000054 \\ 0.0000027 & -0.0000054 & 0.9999531 \end{bmatrix}, \tag{3.31}$$

at 7.92 bit/sample, and

$$\boldsymbol{RE} = \begin{bmatrix} 0.7457125 & -0.0432553 & 0.0181861 \\ -0.0432553 & 0.7386143 & -0.0432553 \\ 0.0181861 & -0.0432553 & 0.7457125 \end{bmatrix}, \tag{3.32}$$

at 1.50 bit/sample. From Equations (3.31) and (3.32), it is seen that the optimal transform coder is not alias free. This is because the total transfer matrix through the system is not pseudocirculant [Vaidyanathan & Mitra 1988]. From Equations (3.31) and (3.32), it is also seen that the total transfer matrix is not Toeplitz.

3.2 Power Constrained Transforms

The problem of power constrained transforms is illustrated in Figure 3.6. The goal is to minimize the block MSE between the output vector $\hat{\boldsymbol{x}}(n)$ and the input vector $\boldsymbol{x}(n)$ with respect to the transform matrices $\boldsymbol{E}$ and $\boldsymbol{R}$, and at the same time, only a limited amount of power must be used by the channel input vector $\boldsymbol{y}(n)$. The input vector $\boldsymbol{x}(n)$ and the channel noise vector $\boldsymbol{v}(n)$ are assumed to represent jointly stationary vector time series with known second order statistics and zero mean.

This problem was solved in [Lee & Petersen 1976], and the system will be called the block pulse amplitude modulation (BPAM) system. An alternative derivation of the BPAM solution will be given, based on the unconstrained length filter banks introduced in Section 2.2.

3.2.1 Problem Formulation

If the memoryless matrices $\boldsymbol{E}$ and $\boldsymbol{R}$ are inserted into Equation (2.70), the block MSE for the BPAM system can be expressed as

$$\mathcal{E}_{N,M} = \mathrm{Tr}\left\{(\boldsymbol{I} - \boldsymbol{RE})\,\boldsymbol{K_x}(0)\,(\boldsymbol{I} - \boldsymbol{RE})^H + \boldsymbol{R}\boldsymbol{K_v}(0)\boldsymbol{R}^H\right\}, \tag{3.33}$$

where it has been used that $\boldsymbol{K_x}(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S_x}(f)\,df$ and $\boldsymbol{K_v}(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S_v}(f)\,df$. In the BPAM system, it is assumed that the channel transfer matrix is equal to the identity matrix, and this is used in the derivation of Equation (3.33).

Using the same reasoning as above, the power constraint in Equation (2.71) can be rewritten as

$$\mathrm{Tr}\left(\boldsymbol{E}\boldsymbol{K_x}(0)\boldsymbol{E}^H\right) = P. \tag{3.34}$$

The problem is to minimize the block MSE given by Equation (3.33) with respect to $\boldsymbol{E}$ and $\boldsymbol{R}$, subject to the power constraint in Equation (3.34).

3.2.2 Alternative Derivation of the BPAM system

The optimal transform coder will be derived by first finding the optimal unconstrained length power constrained filter banks for special input and noise PSD matrices. This derivation follows the same procedure as used in Subsection 3.1.2.

Assume that the PSD matrix of the input vector time series $\boldsymbol{x}(n)$ to the *unconstrained length* power constrained filter bank with channel transfer matrix $\boldsymbol{C}(z) = \boldsymbol{I}$ is given by Equation (3.4). Furthermore, assume that the PSD matrix of the additive channel noise is given by:

$$\boldsymbol{S_v}(f) = \boldsymbol{K_v}(0), \tag{3.35}$$

where $\boldsymbol{K_v}(0) = E\left[\boldsymbol{v}(n)\boldsymbol{v}^H(n)\right]$ is the autocovariance matrix at lag zero of the additive channel noise vector $\boldsymbol{v}(n)$.

In order to specify the optimal unconstrained length filter bank solution, the frequency independent eigenvector and eigenvalue matrices, represented by $\boldsymbol{U}$ and $\boldsymbol{\Lambda_x}$, respectively, are needed. These matrices can be found from Equations (3.6) and (3.7).

Let $\kappa_i^{(M)}$ be eigenvalue number i of the $M \times M$ autocovariance matrix $\boldsymbol{K_v}(0)$, i.e.,

$$\boldsymbol{K_v}(0)\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda_v}, \tag{3.36}$$

where the matrix $\boldsymbol{\Lambda_v}$ is a diagonal matrix with $\kappa_i^{(M)}$ as its ith diagonal element, and the $M \times M$ matrix $\boldsymbol{V}$ contains the eigenvectors of the matrix $\boldsymbol{K_v}(0)$. The diagonal elements of the matrix $\boldsymbol{\Lambda_v}$ are ordered according to

$$\kappa_0^{(M)} \leq \kappa_1^{(M)} \leq \cdots \leq \kappa_{M-1}^{(M)}. \tag{3.37}$$

By studying the optimal power constrained unconstrained length filter banks given by Equations (2.72) through (2.76), it is realized that the matrix $\boldsymbol{G}(f)$ is frequency independent. Therefore, the optimal polyphase matrix $\boldsymbol{G}$ is an $M \times N$ frequency independent diagonal matrix. The diagonal element number i of this matrix is found by solving the following equation:

$$|G_{i,i}|^2 = \max\left(0, \sqrt{\frac{\kappa_i^{(M)}}{\mu\lambda_i^{(N)}}} - \frac{\kappa_i^{(M)}}{\lambda_i^{(N)}}\right), i \in \{0, 1, \ldots, \min(M, N) - 1\}, \tag{3.38}$$

where μ is a Lagrange multiplier for the power constraint.

The jointly optimal transmitter and receiver unconstrained length polyphase matrices can be expressed as:

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{V}\boldsymbol{G}\boldsymbol{U}^H, \\ \boldsymbol{R} &= \boldsymbol{U}\boldsymbol{T}\boldsymbol{V}^H\boldsymbol{\Lambda}_{\boldsymbol{v}}^{-1}, \end{aligned} \tag{3.39}$$

and all the matrices in Equation (3.39) are frequency independent.

The matrix $\boldsymbol{T}$ in Equation (3.39) can be expressed by

$$\boldsymbol{T} = \boldsymbol{\Lambda_x}\boldsymbol{G}^H\left[\boldsymbol{G}\boldsymbol{\Lambda_x}\boldsymbol{G}^H + \boldsymbol{\Lambda_v}\right]^{-1}\boldsymbol{\Lambda_v}. \tag{3.40}$$

The unconstrained length jointly optimal power constrained transmitter and receiver filter bank given in Equation (3.39) is frequency *independent*. By comparing Equations (2.70) and (2.71), using $\boldsymbol{S_x}(f) = \boldsymbol{K_x}(0)$, $\boldsymbol{S_v}(f) = \boldsymbol{K_v}(0)$ and $\boldsymbol{C}(z) = \boldsymbol{I}$, with Equations (3.33) and (3.34), respectively, it is seen that the optimal unconstrained length filter bank solution has to be the optimal transform matrices as well. The same reasoning as in the bit constrained case is used, see Subsection 3.1.2.

3.2.3 Performance Expressions

The performance of the jointly optimal transmitter and receiver transform can be expressed by the block MSE and power. The block MSE of the optimal

power constrained transform system is derived from Equation (2.77), using the modifications stated above

$$\mathcal{E}_{N,M} = \begin{cases} \displaystyle\sum_{i=0}^{N-1} \frac{\lambda_i^{(N)}\kappa_i^{(M)}}{|G_{i,i}|^2\lambda_i^{(N)} + \kappa_i^{(M)}}, & \text{if } M \geq N, \\ \displaystyle\sum_{i=0}^{M-1} \frac{\lambda_i^{(N)}\kappa_i^{(M)}}{|G_{i,i}|^2\lambda_i^{(N)} + \kappa_i^{(M)}} + \sum_{i=M}^{N-1} \lambda_i^{(N)}, & \text{if } M < N. \end{cases} \tag{3.41}$$

From Equation (2.78), the constraint on the power used per channel input vector can be found:

$$\sum_{i=0}^{\min(M,N)-1} |G_{i,i}|^2 \lambda_i^{(N)} = P. \tag{3.42}$$

In [Lee & Petersen 1976], the performance expressions are formulated in a slightly different manner, but it is straightforward to verify that the solution found above corresponds to the solution found in [Lee & Petersen 1976].

3.3 Summary

In Section 3.1, the problem of finding the jointly optimal analysis and synthesis transform matrices under a bit constraint was studied, and analytical expressions for the optimal transforms and bit allocation were found. Differences between the proposed transform and the KLT were pointed out.

In Section 3.2, the optimal power constrained transform matrices were treated. The optimal solution was derived in [Lee & Petersen 1976], and an alternative derivation of the optimal transform was given based on the unconstrained length power constrained solution found in Section 2.2.

The proposed transforms for both the bit and power constrained problems are related to the KLT since they are given by the KLT multiplied by a diagonal matrix. Therefore, the complexity of the proposed transforms is of the same order as for KLTs. Other well known transforms like DCT and FFT have lower complexity.

Chapter 4

Algorithms for Finding Signal-Adaptive FIR Filter Banks

For FIR filter banks, the analysis and synthesis polyphase matrices denoted $\boldsymbol{E}(z)$ and $\boldsymbol{R}(z)$, respectively, are expressed as

$$\begin{aligned} \boldsymbol{E}(z) &= \sum_{k=0}^{m} \boldsymbol{e}(k) z^{-k} \\ \boldsymbol{R}(z) &= \sum_{k=0}^{l} \boldsymbol{r}(k) z^{-k}, \end{aligned} \tag{4.1}$$

where $\{\boldsymbol{e}(k), 0 \leq k \leq m\}$ and $\{\boldsymbol{r}(k), 0 \leq k \leq l\}$ are matrix sequences of unknown $M \times N$ and $N \times M$ matrices, respectively. The problems considered in this chapter are to minimize the block MSE with respect to the matrices $\boldsymbol{E}(z)$ and $\boldsymbol{R}(z)$, subject to a bit or a power constraint.

The known solutions for signal-adaptive jointly optimized transforms, see Chapter 3, and infinite length filter banks, see Chapter 2, give a lower and upper bound, respectively, for the SNR vs. rate or CSNR performance of the FIR filter banks that will be found in this chapter.

Causal FIR filters are assumed, but the same methodology could be used for non-causal and anti-causal FIR filters. Neither PR nor linear phase is presumed in this chapter. Since the set of filter banks considered in this chapter includes PR filter banks, at least as good performance compared to PR filter banks will be achieved.

The notation introduced in this chapter is closely related to the notation used in [Honig et al. 1992], but it has been further developed. In [Gosse

& Duhamel 1997], another notation is used, but the matrices containing the filter bank coefficients in [Gosse & Duhamel 1997] can be obtained by the matrices $\boldsymbol{E}_{\_}$ and $\boldsymbol{R}_{\mathsf{l}}$, which will be introduced in this chapter, by rearranging the matrices' elements.

The current chapter is organized as follows: A numerical algorithm for finding jointly optimal FIR analysis and synthesis filter banks under a bit constraint is proposed in Section 4.1, and under a power constraint in Section 4.2. Section 4.3 contains a brief summary.

This chapter is partly based on [Hjørungnes & Ramstad 1999*c*, Hjørungnes & Ramstad 1999*a*, Hjørungnes, Coward & Ramstad 1999].

4.1 Bit Constrained FIR Filter Banks

Figure 1.3 indicates that the dimensions of the source and channel vectors are $N \times 1$ and $M \times 1$, respectively. No assumptions on the values of N and M are made.

This section is organized as follows: The assumptions and the problem treated are stated in Subsection 4.1.1. In Subsection 4.1.2, equations for optimality are derived, and in Subsection 4.1.3, the proposed algorithm is presented. Optimization for arbitrary given filter lengths is explained in Subsection 4.1.4. Results using the proposed theory are presented in Subsection 4.1.5, where comparisons to other filter bank solutions are given.

4.1.1 Problem Formulation

It is impossible to use the expressions for the performance developed in Chapter 1 for FIR filter banks since the delay through the system must be taken into account. In this section, an alternative method for finding expressions for the performance of the filter bank system is used.

In the following results, some additional matrices are needed. They are introduced here. A *row-expanded* matrix $\boldsymbol{E}_{\_}$ is an $M \times (m+1)N$ matrix given by

$$\boldsymbol{E}_{\_} = [\boldsymbol{e}(0)|\boldsymbol{e}(1)|\ldots|\boldsymbol{e}(m)], \tag{4.2}$$

and the *column-expanded* matrix $\boldsymbol{R}_{\mathsf{l}}$ is an $(l+1)N \times M$ matrix given by

$$\boldsymbol{R}_{\mathsf{l}} = \begin{bmatrix} \boldsymbol{r}(l) \\ \boldsymbol{r}(l-1) \\ \vdots \\ \boldsymbol{r}(1) \\ \boldsymbol{r}(0) \end{bmatrix}. \tag{4.3}$$

Row- and column-expansions are defined for any matrix sequence as shown in Equations (4.2) and (4.3), respectively.

Row number i of the row-expanded matrix $\boldsymbol{E}_{\_}$ in Equation (4.2) contains the impulse response of analysis filter number i. The indices of the impulse response increase while going from the left to the right in the matrix. The first column of the matrix $\boldsymbol{e}(0)$ contains the first impulse response coefficients of the analysis filters. Column number i of the column-expanded matrix $\boldsymbol{R}_{|}$ in Equation (4.3) contains the impulse response of synthesis filter number i. The indices of the impulse response increase while going from the bottom to the top of the matrix. The last row of the matrix $\boldsymbol{r}(0)$ contains the first impulse response coefficients of the synthesis filters.

Another matrix that will be useful for expressing the performance of the filter bank is the matrix $\boldsymbol{E}_{\ulcorner}$, which is used to express convolution of two matrix sequences. For example, if $m = 2$ and $l = 3$, the row-expansion of the convolution of the matrices $\boldsymbol{R}(z)$ and $\boldsymbol{E}(z)$ is given by the matrix-product $\boldsymbol{R}_{\_}\boldsymbol{E}_{\ulcorner}$, where $\boldsymbol{R}_{\_}$ is an $N \times (l+1)M$ matrix and $\boldsymbol{E}_{\ulcorner}$ is an $(l+1)M \times (l+m+1)N$ matrix given by

$$\boldsymbol{E}_{\ulcorner} = \begin{bmatrix} \boldsymbol{e}(0) & \boldsymbol{e}(1) & \boldsymbol{e}(2) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{e}(0) & \boldsymbol{e}(1) & \boldsymbol{e}(2) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{e}(0) & \boldsymbol{e}(1) & \boldsymbol{e}(2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{e}(0) & \boldsymbol{e}(1) & \boldsymbol{e}(2) \end{bmatrix}, \tag{4.4}$$

where $\mathbf{0}$ is the $M \times N$ zero matrix. The matrix $\boldsymbol{E}_{\ulcorner}$ is a *block Sylvester* matrix [Lütkepohl 1996], and it is related to the transpose of the matrix $\boldsymbol{A}$ given in Equation (30) used in the method proposed in [Nayebi, Barnwell & Smith 1992].

The input vector signal $\boldsymbol{x}(n)$ to the analysis polyphase matrix $\boldsymbol{E}(z)$ and the output vector signal $\hat{\boldsymbol{x}}(n)$ of the synthesis matrix $\boldsymbol{R}(z)$ are given by Equations (1.5) and (1.6). Theory will be developed for an arbitrary given positive integer delay through the total filter bank system. For this, the following vector is needed:

$$\boldsymbol{x}_{d_s}(n) = [x(nN - d_s), x(nN - d_s - 1), \ldots, x(nN - d_s - (N-1))]^T, \tag{4.5}$$

where $d_s \in \{0, 1, \ldots, N-1\}$ will be called the *scalar delay* through the system. This delay is included such that the theory developed is as general as possible, and it adds an extra parameter d_s to the optimization of the filter banks.

The subband signals in Figure 1.1 are represented by the $M \times 1$ vector $\boldsymbol{y}(n) = [y_0(n), y_1(n), \ldots, y_{M-1}(n)]^T$ given by

$$\boldsymbol{y}(n) = \sum_{k=0}^{m} \boldsymbol{e}(k)\boldsymbol{x}(n-k). \tag{4.6}$$

The same quantizer model as introduced in Chapter 1 will be used here.

Let p be a positive integer. The column-expansion of a vector time series $\boldsymbol{x}(n)$ of dimension $(p+1)N \times 1$ is defined as:

$$\boldsymbol{x}(n)_{\shortmid} = \begin{bmatrix} \boldsymbol{x}(n) \\ \boldsymbol{x}(n-1) \\ \vdots \\ \boldsymbol{x}(n-p) \end{bmatrix}. \tag{4.7}$$

The column-expansion of other vector time series are found in the same manner as shown by Equation (4.7). A partial statistical description of the $p \times 1$ vector $\boldsymbol{x}(n)_{\shortmid}$ and the $p \times 1$ vector $\boldsymbol{q}(n)_{\shortmid}$ is given by the following covariance matrices:

$$\boldsymbol{\Phi}_{\boldsymbol{x}}^{(p,N)} = E\left[\boldsymbol{x}(n)_{\shortmid}\boldsymbol{x}^H(n)_{\shortmid}\right], \tag{4.8}$$

$$\boldsymbol{\Phi}_{\boldsymbol{q}}^{(p,M)} = E\left[\boldsymbol{q}(n)_{\shortmid}\boldsymbol{q}^H(n)_{\shortmid}\right] = \sigma_q^2 \boldsymbol{I}_{(p+1)M}, \tag{4.9}$$

where the superscript $^{(p,N)}$ means that it is a $(p+1)N \times (p+1)N$ matrix, and where $\boldsymbol{I}_{(p+1)M}$ is the $(p+1)M \times (p+1)M$ identity matrix.

The covariance matrix $\boldsymbol{\phi}_{\boldsymbol{x}}^{(p,N)}(d_v, d_s)$ of dimension $(p+1)N \times N$ is defined as:

$$\boldsymbol{\phi}_{\boldsymbol{x}}^{(p,N)}(d_v, d_s) = E\left[\boldsymbol{x}(n)_{\shortmid}\boldsymbol{x}_{d_s}^H(n-d_v)\right], \tag{4.10}$$

where $d_v \in \{0, 1, \ldots, p\}$ is a number specifying the *vector delay* through the filter banks. If $d_v = p$, d_s must be equal to zero in order to ensure that PR is possible. PR should be structurally possible, because for high rates, PR is asymptotically optimal. For other values of d_v, d_s can be chosen from the set $\{0, 1, \ldots, N-1\}$. The total delay from the input $x(n)$ to the output $\hat{x}(n)$, see Figure 1.1, is given by $N - 1 + d_v N + d_s$. The term $N - 1$ comes from the delay chain before and after the decimators and expanders, respectively, while the terms $d_v N$ and d_s are due to the vector and scalar delays through the combined analysis and synthesis polyphase filter bank, respectively. d_v is equal to m_0 and d_s is equal to r in Equation (5.6.7) in [Vaidyanathan 1993].

By studying Equations (4.8) and (4.10), it can be seen that the matrix $\boldsymbol{\phi}_{\boldsymbol{x}}^{(p,N)}(d_v, d_s)$ is given by a submatrix of the matrix $\boldsymbol{\Phi}_{\boldsymbol{x}}^{(p,N)}$, consisting of column number $d_s + d_v N$ through column number $d_s + d_v N + N - 1$.

The delay through the filter banks can be optimized through a discrete optimization, where the optimal values of d_v and d_s are obtained by comparing the performances for all allowable delay values.

By rewriting the convolution sum with the notation introduced earlier in this section, it is possible to express the vector $\hat{\boldsymbol{x}}(n)$ as follows:

$$\hat{\boldsymbol{x}}(n) = \boldsymbol{R}\_\boldsymbol{E}\_\boldsymbol{x}(n)_{\shortmid} + \boldsymbol{R}\_\boldsymbol{q}(n)_{\shortmid}, \tag{4.11}$$

where $\boldsymbol{R}_{\_}$ is an $N \times (l+1)M$ matrix, $\boldsymbol{x}(n)_{|}$ is an $(m+l+1)N \times 1$ vector, and $\boldsymbol{q}(n)_{|}$ is an $(l+1)M \times 1$ vector.

In this section, it is assumed that there is no cross-correlation between the quantization noise signal and the input signal. Therefore, the cross-correlation between the quantization noise and the signal vector is set to zero. By substitution of the covariance matrices introduced in Equations (4.8) through (4.10) into Equation (A.8) found for the block MSE in Appendix A, the block MSE for FIR filter banks can be expressed as

$$\begin{aligned}\mathcal{E}_{N,M}(d_v, d_s) = \mathrm{Tr}\Big\{ & \boldsymbol{R}_{\_}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{\_}^{H} \\ & -\boldsymbol{R}_{\_}\boldsymbol{E}_{\Gamma}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s) \\ & -\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s)\right)^{H}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{\_}^{H} \\ & +\boldsymbol{\Phi}_{\boldsymbol{x}}^{(0,N)} + \boldsymbol{R}_{\_}\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}\boldsymbol{R}_{\_}^{H}\Big\}. \end{aligned} \tag{4.12}$$

The last term in Equation (4.12) is the quantization error, and the rest of the terms are signal distortions. Signal distortions can further be classified as amplitude, phase, and alias distortions.

By using the quantizer model introduced in Equation (1.14), it is shown in Appendix A that the bit constraint in the FIR filter bank case can be written as

$$\Pr\left\{\boldsymbol{E}_{\_}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}\boldsymbol{E}_{\_}^{H}\right\} = \frac{2^{2Nb}\sigma_q^{2M}}{\prod_{i=0}^{M-1} c_i}, \tag{4.13}$$

where the operator Pr returns the product of the elements on the main diagonal of the matrix.

The problem is to minimize the MSE given in Equation (4.12) with respect to the analysis and synthesis filter banks, under the constraint given in Equation (4.13), while at the same time the constraints in Equation (2.4) must be satisfied.

The objective function for the unconstrained optimization problem can be expressed as

$$\mathcal{E}_{N,M}(d_v, d_s) + \mu \sum_{k=0}^{M-1} \ln \sigma_{y_k}^2 - \sum_{k=0}^{M-1} \theta_k \sigma_{y_k}^2, \tag{4.14}$$

where μ is the Lagrange multiplier for the bit constraint, and θ_k is the non-negative Kuhn-Tucker parameter [Luenbereger 1984] for the inequality constraint $\sigma_{y_k}^2 \geq \sigma_q^2$.

4.1.2 Equations for Optimality with Equal Filter Lengths

By using matrix differentiation, see Appendix C, and setting the derivative of the objective function given in Equation (4.14) with respect to the synthesis filter bank $\boldsymbol{R}_{-}$ to zero, it can be shown that the optimal synthesis filter bank for a given analysis filter bank, vector, and scalar delay can be expressed as:

$$\boldsymbol{R}_{-} = \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s)\right)^H \boldsymbol{E}_{\Gamma}^H \left(\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^H + \boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}\right)^{-1}. \tag{4.15}$$

The result in Equation (4.15) is an FIR Wiener synthesis filter bank. In Subsection 7.2.3, the orthogonality principle [Therrien 1992] will be used to derive the FIR Wiener filter bank.

Using the matrix differentiation formulas found in Appendix C, the objective function in Equation (4.14) can be differentiated with respect to the analysis filter bank $\boldsymbol{E}_{-}$. In this way, it is possible to derive equations for the optimal analysis filter bank for a given synthesis filter bank, vector, and scalar delay. These equations are nonlinear, and are given by:

$$\begin{aligned}\boldsymbol{R}_{\mathsf{I}}^H \mathcal{T}\left\{\boldsymbol{R}_{-}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)} - \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s)\right)^H\right\} \\ = \left(\boldsymbol{\Theta} - \mu\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\right)\boldsymbol{E}_{-}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)},\end{aligned} \tag{4.16}$$

where the matrix $\boldsymbol{\Theta}$ is an $M \times M$ diagonal matrix, where diagonal element number i is the Kuhn-Tucker parameter θ_i for the inequality given in Equation (2.4). $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ in an $M \times M$ diagonal matrix, where diagonal element number i is given by the variance of subband signal $\sigma_{y_i}^2$, which depends on the input statistics and the analysis filter bank $\boldsymbol{E}_{-}$. In Equation (4.16), the operator $\mathcal{T} : \mathbb{R}^{N\times(m+l+1)N} \rightarrow \mathbb{R}^{(l+1)N\times(m+1)N}$ produces a rectangular block Toeplitz matrix of dimension $(l+1)N \times (m+1)N$ from an $N \times (m+l+1)N$ matrix. Let $\boldsymbol{W}_{-}$ be an $N \times (m+l+1)N$ matrix, where the ith $N \times N$ block is given by $[\boldsymbol{W}_{-}]_i$, $i \in \{0, 1, \ldots, m+l\}$. Then, the operator $\mathcal{T}$ is defined as follows:

$$\mathcal{T}\{\boldsymbol{W}_{-}\} = \begin{bmatrix} [\boldsymbol{W}_{-}]_l & [\boldsymbol{W}_{-}]_{l+1} & \cdots & [\boldsymbol{W}_{-}]_{m+l} \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{W}_{-}]_1 & [\boldsymbol{W}_{-}]_2 & \cdots & [\boldsymbol{W}_{-}]_{m+1} \\ [\boldsymbol{W}_{-}]_0 & [\boldsymbol{W}_{-}]_1 & \cdots & [\boldsymbol{W}_{-}]_m \end{bmatrix}. \tag{4.17}$$

The Kuhn-Tucker parameters θ_i must be non-negative, i.e., $\theta_i \geq 0$, and Equation (2.14) must be satisfied.

Table 4.1 Pseudo code of the numerical optimization algorithm.

Step 1: Initialization
Choose values for μ, θ_i, N, M, m, and l
Initialize the analysis filter bank

Step 2: Delay Processing
for $d_v := 0, 1, \ldots, m + l$
  if $d_v = m + l$
    $d_s := 0$
    Perform the optimization procedure in Step 3
  else
    for $d_s := 0, 1, \ldots, N - 1$
      Perform the optimization procedure in Step 3
    end
  end
end
Go to Step 4

Step 3: Filter Bank Optimization Procedure
Step i: Synthesis Filter Optimization
For the current value of the analysis filter bank, find the corresponding synthesis filter bank from Equation (4.15)
Step ii: Analysis Filter Optimization
For the current value of the synthesis filter bank, find the corresponding analysis filter bank by solving Equation (4.16)
Step iii: Convergence Check
if the analysis filter bank has converged
  Store the current filter bank as the optimized one for the current values of d_v, d_s, μ, and θ_i
  Procedure finished
else
  Go to Step i
end

Step 4: Rate Check
Find the filter banks with the best performance among all the calculated filter banks. The corresponding values of the vector delay d_v, the scalar delay d_s, and the best performing filter banks are the optimized values for the current values of μ and θ_i
Calculate the rate of the current optimized filter banks
if the rate is the desired target rate
  The current values of the analysis and synthesis filter banks and the corresponding values of d_v and d_s contain the optimized values
  Stop
else
  Adjust μ and θ_i, and go to Step 2
end

4.1.3 Numerical Optimization Algorithm

An important task in the numerical optimization algorithm is to find a good initial condition for the analysis filter bank. For high rates, the jointly optimized filter banks should be close to having the PR property since the quantization noise is very small in this case. Therefore, for high rates, the signal-adaptive gain optimized linear phase biorthogonal filter banks found in [Balasingham 1998] and the PR 10_18 filter banks found in [Tsai et al. 1996] were used as the initial filter banks. The optimization was then carried out for decreasing rates, with the filter bank found for a higher rate used as an initial filter bank.

There are other possibilities for finding initial values for the optimization algorithm. The optimal bit constrained transform matrices found in Section 3.1 can be used to find initial values. This can be done by setting $\boldsymbol{e}(k) = \mathbf{0}$ for all $k \in \{0, 1, 2, \ldots, m\}$ except one, which is set equal to the analysis transform matrix found in Section 3.1. Another possibility is to truncate the unconstrained length filter banks found in Section 2.1 to the appropriate length and to use this as an initial value. Other promising FIR solutions found in the literature can also be used as initial values.

The numerical optimization algorithm for finding the jointly optimized analysis and synthesis FIR filter bank is summarized in Table 4.1.

For $m = l = 0$ the iterative numerical algorithm converges to the optimal transform proposed in Section 3.1.

Equations (4.15) and (4.16) should be solved simultaneously, and in Table 4.1 this is achieved by an iterative procedure. Another possibility would have been to solve these equations simultaneously by using numerical methods. This has been tried with the Matlab Optimization Toolbox [Coleman, Branch & Grace 1999], but this leads to convergence difficulties. The results obtained in this way were not as good as the results obtained by the algorithm shown in Table 4.1.

If different initial values are used for the analysis filter bank in the optimization algorithm, the resulting filter banks will not necessarily be the same. This shows that the global optimum is not necessarily found by the iterative FIR optimization algorithm. However, the results obtained by the algorithm show that very good performance is achieved.

4.1.4 Arbitrary Filter Length Optimization

Theory for jointly optimized FIR filter banks with analysis filter lengths $(m + 1)N$ and synthesis filter lengths $(l + 1)N$ was developed earlier in this section. Here, this theory is extended to include the case where the filters can have

arbitrary given filter lengths, such that unequal filter lengths are also possible.

Row number i from *left* to *right* in the matrix $\boldsymbol{E}_{\_}$ represents the impulse response of analysis filter number i, $H_i(f)$. Column number i from *bottom* to *top* in the matrix $\boldsymbol{R}_{|}$ represents the impulse response of synthesis filter number i, $F_i(f)$. Since the filter lengths in the filter banks are not necessarily equal, the matrices $\boldsymbol{E}_{\_}$ and $\boldsymbol{R}_{|}$ may contain impulse response coefficients that are forced to zero.

The *length* of the impulse response of an FIR filter is defined as the number of impulse response coefficients between the first non-zero impulse response coefficient and the last non-zero impulse response coefficient, even though some of the coefficients in between may be equal to zero. Let ℓ_{H_i} and ℓ_{F_i} be the length of the impulse response of analysis filter H_i and synthesis filter F_i, respectively. Assume that the decimation factor used is N, and let m and l be the smallest non-negative integers such that

$$\begin{aligned} &\max\{\ell_{H_i} | i \in \{0, 1, \ldots, N-1\}\} \leq N(m+1) \quad \text{and} \\ &\max\{\ell_{F_i} | i \in \{0, 1, \ldots, N-1\}\} \leq N(l+1), \end{aligned} \tag{4.18}$$

respectively.

Let $\boldsymbol{\mathcal{A}}_{\_}$ be an $M \times (m+1)N$ matrix containing ones at the positions corresponding to where the analysis filter bank $\boldsymbol{E}_{\_}$ contains free parameters and zeros where $\boldsymbol{E}_{\_}$ must contain zeros. In the same way, let $\boldsymbol{\mathcal{S}}_{|}$ be an $(l+1)N \times M$ matrix containing ones at the positions corresponding to where the synthesis filter bank $\boldsymbol{R}_{|}$ contains free parameters and zeros where $\boldsymbol{R}_{|}$ must contain zeros. Analogous to the above definition, let $\boldsymbol{\mathcal{S}}_{\_}$ be an $N \times (l+1)M$ row-expanded matrix corresponding to the matrix $\boldsymbol{R}_{\_}$.

By using Lagrange multipliers [Luenbereger 1984], it can be shown that the equations for finding jointly optimized analysis and synthesis filter banks with *arbitrary given filter lengths* can be found by picking out the equations from Equations (4.16) and (4.15), respectively, corresponding to the positions where $\boldsymbol{\mathcal{A}}_{\_}$ and $\boldsymbol{\mathcal{S}}_{\_}$ are different from zero. In addition, in the positions corresponding to where $\boldsymbol{\mathcal{A}}_{\_}$ and $\boldsymbol{\mathcal{S}}_{\_}$ are equal to zero, the old equations in these positions are replaced with equations stating that the corresponding filter coefficients are equal to zero. In this method, the fixed filter coefficients could be set to an *arbitrary constant value*, not only zero, and this can be done for any coefficients in the impulse response.

Since the matrices $\boldsymbol{\mathcal{A}}_{\_}$ and $\boldsymbol{\mathcal{S}}_{|}$ may contain zeros and ones at arbitrary positions, the above procedure can be used to find jointly optimized analysis and synthesis filter banks with arbitrary given filter lengths. This is done by choosing an appropriate shape of the matrices $\boldsymbol{\mathcal{A}}_{\_}$ and $\boldsymbol{\mathcal{S}}_{|}$. While choosing the shape of these matrices, it is important to remember that the delay through

each branch of the analysis/synthesis filter bank combination must be the same if the filter bank is to possess the PR property. At high rates, it is asymptotically optimal to have PR filter banks, so the structure of the matrices $\mathcal{A}_-$ and $\mathcal{S}_|$ must be chosen carefully.

An example of how the matrices $\mathcal{A}_-$ and $\mathcal{S}_|$ can be selected in the 5_3 case will be given. The notation 5_3 means that $N = 2$. The analysis lowpass filter and the synthesis highpass filter have filter length $\ell_{H_0} = \ell_{F_1} = 5$, while the analysis highpass filter and the synthesis lowpass filter have filter length $\ell_{H_1} = \ell_{F_0} = 3$. The given values lead to $m = l = 2$. If the vector delay is $d_v = 1$ and the scalar delay is $d_s = 0$, then the following matrices can be used:

$$\mathcal{A}_- = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}, \tag{4.19}$$

and

$$\mathcal{S}_| = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{4.20}$$

To make sure that the delay through each subband branch is the same, the zeros have been put at the top of the $\mathcal{S}_|$ matrix. This example is designed such that linear phase should be possible in the optimized filter bank.

4.1.5 Bit Constrained FIR Filter Bank Results

In all the results presented in this subsection, the following choice is made: $\sigma_q^2 = 1$.

Figure 4.1 shows the magnitude responses for different 9_7 filter banks when coding a Gaussian AR(1) source with correlation coefficient 0.95 at 2.67 bits/sample. In the figure, the following choices have been made: $N = M = 2$, $d_v = 5$, and $d_s = 0$. Row number i in the figure represents subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters. The magnitude responses in Figure 4.1 show that the proposed 9_7 filter bank has different shaping both in the pass- and stopband regions from the 9_7 filter bank in [Balasingham 1998] and the 9_7 wavelet in [Antonini et al. 1992].

Table 4.2 shows the distortion rate performance of different systems when coding a Gaussian AR(1) source with correlation coefficient 0.95. Since

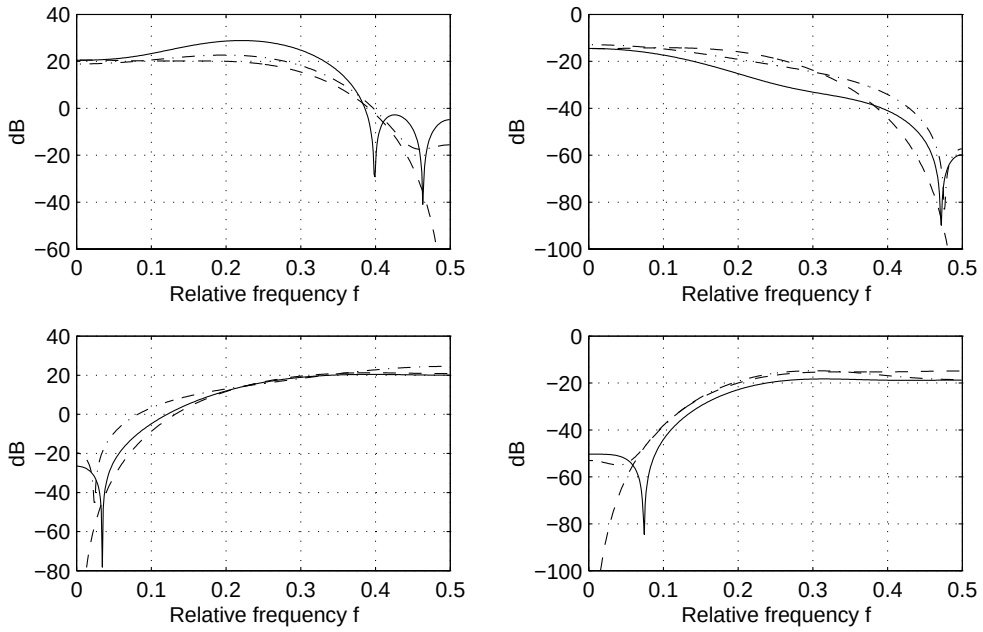

Figure 4.1 Magnitude response for the 9_7 filter banks, where $N = M = 2$, $c_i = \frac{e\pi}{6}$, $m = l = 4$, $d_v = 5$, and $d_s = 0$. The proposed filter bank is shown by the solid curves, the dash-dotted curves show linear phase biorthogonal system [Balasingham 1998] responses, and the dashed curves show the responses of the 9_7 wavelet in [Antonini et al. 1992]. A Gaussian AR(1) source with correlation coefficient 0.95 is coded at 2.21 bits/sample in all cases. Row number i in the figure is subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters.

it is assumed that uniform entopy constrained scalar quantizers are used, the coding coefficients are given by $c_i = \frac{e\pi}{6}$ [Gersho & Gray 1992] for all $i \in \{0, 1, \ldots, M-1\}$. The performances of the four well known wavelets 5_3 [Le Gall & Tabatabai 1988], 6_6 [Rodrigues et al. 1997][1], 9_7 [Antonini et al. 1992], and 10_18 [Tsai et al. 1996], are included in the table as well as the distortion rate function [Berger 1971], optimal unconstrained length filter banks from Section 2.1, and optimal transforms from Section 3.1. Gain optimized biorthogonal filter banks with linear phase [Balasingham 1998] are also shown in the table. In the table, results are also included for systems

[1]The filter coefficients in this case were found at the following URL: `http://www.wavelet.org/wavelet/digest_06/digest_06.05.html`.

Table 4.2 Theoretical distortion rate performances.

$N = 2$. The source is a Gaussian AR(1) with correlation coefficient 0.95. d_v indicates the vector delay used in the system. $d_s = 0$ in all cases, and $c_i = \frac{\pi e}{6}$.

Type of coding system	Bit rate [bits per sample]			
	0.50	1.00	2.00	3.00
	SNR [dB]			
Dist. rate func. [Berger 1971]	12.70	16.13	22.15	28.17
Infinite order, Section 2.1	9.55	13.38	19.75	24.93
Transform, Section 3.1	5.52	9.80	16.96	21.90
5_3 [Le Gall & Tabatabai 1988], $d_v = 1$	5.43	10.68	16.79	22.81
Biorth. 5_3 [Balasingham 1998], $d_v = 1$	5.46	10.72	16.82	22.84
5_3 Wiener, $d_v = 1$	6.48	**10.94**	17.71	23.12
Proposed 5_3, $d_v = 1$	**6.50**	**10.94**	**18.21**	**23.16**
6_6 [Rodrigues et al. 1997], $d_v = 2$	4.93	10.25	16.53	22.55
Biorth. 6_6 [Balasingham 1998], $d_v = 2$	5.57	10.76	16.72	22.74
6_6 Wiener, $d_v = 2$	8.10	11.83	17.79	23.06
Proposed 6_6, $d_v = 2$	**8.17**	**12.11**	**18.31**	**23.51**
9_7 [Antonini et al. 1992], $d_v = 5$	4.45	9.89	16.45	22.45
Biorth. 9_7 [Balasingham 1998], $d_v = 5$	6.22	11.35	17.01	23.03
9_7 Wiener, $d_v = 5$	8.33	12.14	18.06	23.36
Proposed 9_7, $d_v = 5$	**8.53**	**12.28**	**18.75**	**23.94**
10_18 [Tsai et al. 1996], $d_v = 6$	4.99	10.37	16.76	22.78
10_18 Wiener, $d_v = 6$	8.84	11.97	17.97	23.13
Proposed 10_18, $d_v = 6$	**9.31**	**13.08**	**19.41**	**24.54**

using a PR analysis filter bank and FIR Wiener synthesis filter bank with the same filter lengths as the PR FIR synthesis filter bank. In the 5_3, 6_6, and 9_7 cases, the analysis filter banks are found in [Balasingham 1998], while in the 10_18 case, the analysis filter bank in [Tsai et al. 1996] is used. The bit allocation is the same as the bit allocation used if PR filter banks were used, i.e., the bits are distributed such that the product of the quantization noise variance and the squared norm of the *PR* synthesis filter is constant for each branch of the filter bank.

The proposed filter banks perform better than all the other systems in Table 4.2. Most of the gain obtained over PR filter banks is achieved by using a Wiener synthesis filter bank. However, this requires that the analysis filter bank available is well suited for coding the PSD of the input time series $x(n)$. If PSDs with bandpass characteristics were used, the situation would be different. In Table 4.2, the results are obtained by PR filter banks adapted to PSDs with lowpass characteristics, but the situation would have been different if a PSD

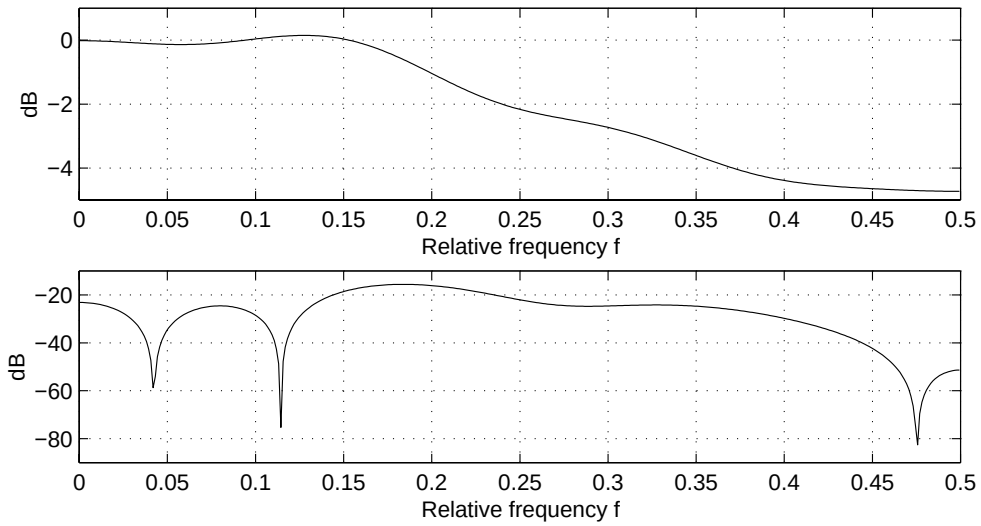

Figure 4.2 The upper figure shows the alias free transfer function $A_0(f)$ for the 9_7 filter bank shown in Figure 4.1. The lower figure shows the corresponding first aliasing term $A_1(f)$. The same parameters as in Figure 4.1 are used here.

such as the one shown in Figure 2.2 (a) was used.

From the table, it is seen that the proposed system 6_6 system performs better than the 9_7 PR filter bank in [Antonini et al. 1992] as well as the 9_7 PR filter bank in [Balasingham 1998], even though the filter lengths are shorter in the proposed filter bank.

Figure 4.2 shows the alias free transfer function $A_0(f)$ and the first aliasing term $A_1(f)$ [Vaidyanathan 1993, pp. 225–226] for the filter bank shown in Figure 4.1. The alias free transfer function of the cascaded analysis/synthesis system is approximately 0 dB for $f = 0$, and approximately -4.7 dB for $f = \frac{1}{2}$. The maximum value of the alias free transfer function is approximately 0.15 dB, and it is achieved at a relative frequency of 0.13. The gain for the first aliasing term is approximately -23 dB at $f = 0$ and approximately -51 dB at $f = \frac{1}{2}$, where f is a relative frequency. The maximum value of the first aliasing term is -15.6 dB at $f = 0.18$. It is seen from the figure that the proposed filter bank does not have the PR property, because for PR filter banks the alias free transfer function is equal to 0 dB and the first alias term is equal to $-\infty$ dB for all frequencies.

The corresponding impulse responses of the magnitude responses shown in Figure 4.1, are shown in Figure 4.3. From the figure, it is seen that the impulse

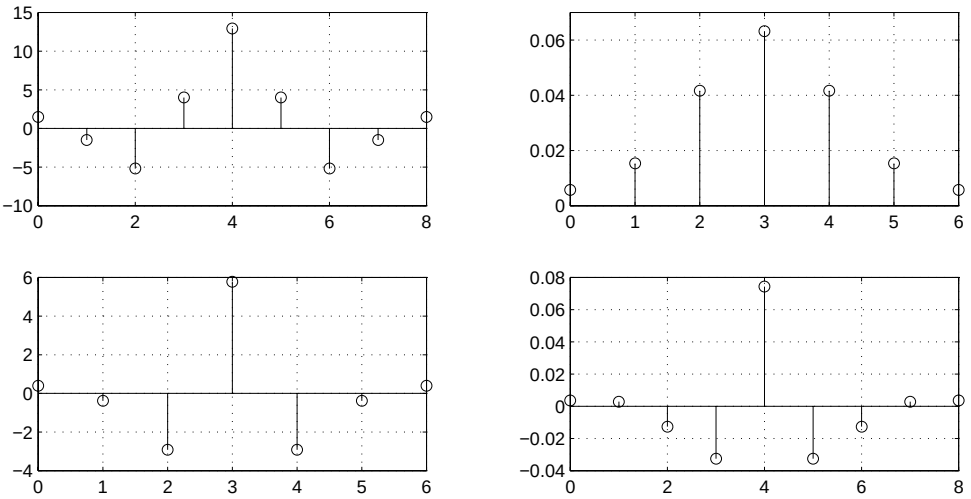

Figure 4.3 Impulse responses of the proposed 9_7 FIR filter bank system for the same parameters as in Figure 4.1. Row number i in the figure is subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters.

responses have linear phase.

4.1.5.1 Linear Phase

If the optimization is performed for all the different delays, it has been found that the optimal values of the delays d_v and d_s are rate dependent, so the optimal delay for one particular rate need not be optimal for a different rate.

If it is desired that the filter bank have the PR property, which is an advantage for high rates, the delay through each branch of the filter bank must be the same. If in addition, all the filters must have linear phase, this also decides the vector and scalar delay through the filter bank, if the filter lengths are given. However, the delay through the filter bank can be chosen in such a manner that linear phase is impossible for the given matrices $\boldsymbol{\mathcal{A}}_-$ and $\boldsymbol{\mathcal{S}}_{\mathsf{I}}$. If the initial filter bank does not have linear phase, the resulting optimized filter bank might not have linear phase.

If the delays are chosen such that linear phase is possible, and the initial filter bank does have linear phase, the optimized filter bank usually also has linear phase. However, in some optimizations, the optimized filter bank does not have linear phase even in this case. In the passbands, the phase is approx-

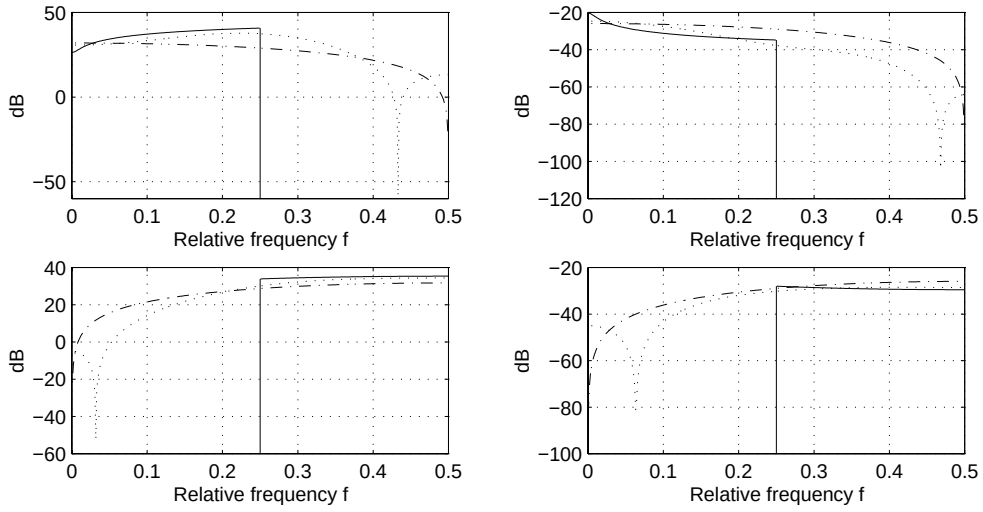

Figure 4.4 Comparisons of magnitude responses of the proposed systems. The input signal is a Gaussian AR(1) signal with correlation coefficient 0.95, $N = M = 2$, and $c_i = \frac{e\pi}{6}$. The solid curves show the frequency response of the unconstrained length filter banks from Section 2.1, the dotted curves show the FIR responses when using the optimized 9_7 filter banks with $d_v = 5$ and $d_s = 0$ from this section, and the dash-dotted curves show the bit constrained transform magnitude responses from Section 3.1. The bit rate used is 4.21 bits/sample in all cases. Row number i in the figure is subband number i in the subband coder. The first column shows the analysis filters, while the second column shows the synthesis filters.

imately linear even if the filters do not have linear phase for all frequencies. It is not surprising that linear phase might be suboptimal, because if linear phase is imposed, this approximately halves the number of free optimization parameters.

4.1.5.2 Magnitude Response Comparisons

Figure 4.4 compares the magnitude responses of the unconstrained length filter banks found in Section 2.1, the bit constrained transforms found in Section 3.1, and the FIR filter banks presented in the current section, when coding a Gaussian AR(1) source with correlation coefficient 0.95 at 4.21 bits/sample. The SNR values of the three systems are as follows: Unconstrained length filter banks: SNR = 32.0 dB, transforms: SNR = 28.9 dB, and 9_7 FIR filter banks: SNR = 30.9 dB.

From Figure 4.4, it is seen that the shaping of the stopband is very different in the three cases. In the unconstrained length case, there is infinite attenuation for all frequencies in the stopbands. In the transform and FIR filter bank cases, the shaping of the stopbands is different, and in the FIR case there are zeros on the unit circle in the passband on positions different from $z = -1$. More attenuation is therefore achieved in the FIR case than the transform case.

In the passbands, the FIR case resembles the unconstrained length filters more than the transform filters. For the unconstrained length and FIR analysis filters, it is observed that the shaping of the passbands resembles half-whitening of the input PSD function. In the transform case, each filter has only one zero since $N = 2$, and from the magnitude responses, it is seen that this zero is either placed at $z = -1$ or $z = 1$ for lowpass and highpass filters, respectively.

4.2 Power Constrained FIR Filter Banks

The system considered in this case is shown in Figure 1.3. It is assumed that the input vector signal to the transmitter $\boldsymbol{x}(n)$ is uncorrelated with the additive channel noise vector $\boldsymbol{v}(p)$ for all values of n and p. Figure 1.3 indicates that the dimensions of the source and channel vectors are $N \times 1$ and $M \times 1$, respectively. There are no assumptions on the values of N and M. Therefore, this is a generalization of the results found in [Honig et al. 1992], where it was assumed that $N = M$. The results presented are also valid for $M > N$. In this section, it is assumed that the transmitter $\boldsymbol{E}(z)$ and receiver filter bank $\boldsymbol{R}(z)$ are FIR, i.e., they are given by Equation (4.1).

This section is organized as follows: The problem treated is stated in Subsection 4.2.1. In Subsection 4.2.2, necessary conditions for optimality will be derived and the optimization algorithm will be explained. Subsection 4.2.3 contains results using the proposed filter banks, and comparisons are made with results found in the literature.

4.2.1 Problem Formulation

In this subsection, the problem is formulated for the FIR power constrained case. Let the channel transfer function be described by an FIR matrix polynomial of order o, i.e., $\boldsymbol{C}(z) = \sum_{i=0}^{o} \boldsymbol{c}(i) z^{-i}$. This matrix has dimension $M \times M$, and the matrix is assumed to be known.

In order to formulate the problem, an expression for the block MSE must be obtained. By comparing Figures 1.1 and 1.3, it is seen that if the analysis filter bank $\boldsymbol{E}(z)$ in the bit constrained case is equal to the convolution of the channel

transfer matrix and the transmitter filter matrix in the power constrained case, that is $\boldsymbol{C}(z)\boldsymbol{E}(z)$, the two systems are equivalent. Therefore, the block MSE for the power constrained FIR case, shown in Figure 1.3, can be derived from the block MSE for the bit constrained FIR case, by substituting the analysis polyphase filter bank $\boldsymbol{E}(z)$ in the bit constrained case with the convolution of the channel transfer matrix and the analysis polyphase matrix. The block MSE in the FIR bit constrained case is given in Equation (4.12).

The row-expansion of the convolution of the channel transfer matrix and the transmitter polyphase matrix can be expressed as $\hat{\boldsymbol{E}}_{\_} = \boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil}$, which is an $M \times (m+o+1)N$ matrix. The matrix $\boldsymbol{C}_{\_}$ is an $M \times (o+1)M$ matrix. Notice that the dimensions of the matrix $\boldsymbol{E}_{\lceil}$ are not the same as in Section 4.1 since it will be used when finding the convolution of the channel transfer matrix and the transmitter polyphase matrix. The dimensions of the matrix $\boldsymbol{E}_{\lceil}$ are $(o+1)M \times (o+m+1)N$ in this section. The row-expanded matrix which is used to express the total transfer function through the system is given by $\boldsymbol{R}_{\_}(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\diagdown}$, and the dimensions of this matrix product are $N \times (m+o+l+1)N$. The matrix $(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\diagdown}$ has dimensions $(l+1)M \times (m+o+l+1)N$, and it is defined as:

$$(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\diagdown} = \begin{bmatrix} \hat{\boldsymbol{e}}(0) & \hat{\boldsymbol{e}}(1) & \hat{\boldsymbol{e}}(2) & \cdots & \hat{\boldsymbol{e}}(m+o) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{e}}(0) & \hat{\boldsymbol{e}}(1) & \cdots & \hat{\boldsymbol{e}}(m+o-1) & \hat{\boldsymbol{e}}(m+o) & \cdots & \mathbf{0} \\ \vdots & & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\boldsymbol{e}}(0) & \hat{\boldsymbol{e}}(1) & \cdots & \hat{\boldsymbol{e}}(m+o-1) & \hat{\boldsymbol{e}}(m+o) \end{bmatrix}, \tag{4.21}$$

where the matrix $\hat{\boldsymbol{E}}_{\_} = [\hat{\boldsymbol{e}}(0)|\hat{\boldsymbol{e}}(1)|\ldots|\hat{\boldsymbol{e}}(m+o)]$.

From Equation (4.21), it is seen that the operator $(\cdot)_{\diagdown}$ operates on row-expanded matrices producing a block matrix with shifted versions of the row-expanded matrix. Using this operator, it can be seen from Equation (4.4) that $\boldsymbol{E}_{\lceil} = (\boldsymbol{E}_{\_})_{\diagdown}$.

If the matrix $\boldsymbol{R}_{\_}(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\diagdown}$ is used to express the row-expansion of the total transfer matrix in Equation (4.12), if m is replaced with $m+o$, and if the symbol of the additive noise vector is changed from $\boldsymbol{q}$ to $\boldsymbol{v}$, the following

expression is obtained for the block MSE in the power constrained FIR case:

$$\begin{aligned}\mathcal{E}_{N,M}\left(d_v,d_s\right) = \mathrm{Tr}\Big\{ & \boldsymbol{R}_{-}(\boldsymbol{C}_{-}\boldsymbol{E}_{\ulcorner})_{\diagdown}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}\left((\boldsymbol{C}_{-}\boldsymbol{E}_{\ulcorner})_{\diagdown}\right)^{H}\boldsymbol{R}_{-}^{H} \\ & -\boldsymbol{R}_{-}(\boldsymbol{C}_{-}\boldsymbol{E}_{\ulcorner})_{\diagdown}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s) \\ & -\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s)\right)^{H}\left((\boldsymbol{C}_{-}\boldsymbol{E}_{\ulcorner})_{\diagdown}\right)^{H}\boldsymbol{R}_{-}^{H} \\ & +\boldsymbol{\Phi}_{\boldsymbol{x}}^{(0,N)}+\boldsymbol{R}_{-}\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}\boldsymbol{R}_{-}^{H}\Big\}, \end{aligned} \tag{4.22}$$

where the matrix $\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}$ is the $(l+1)M \times (l+1)M$ autocovariance matrix of the additive noise vector $\boldsymbol{v}(n)_{\mathsf{l}}$. $\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}$ is defined as

$$\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)} = E\left[\boldsymbol{v}(n)_{\mathsf{l}}\,\boldsymbol{v}^{H}(n)_{\mathsf{l}}\right]. \tag{4.23}$$

Note that $\boldsymbol{v}(n)_{\mathsf{l}}$ is an $(l+1)M \times 1$ vector.

The value of the block MSE in Equation (4.22) for power constrained FIR filter banks reduces to the block MSE for bit constrained FIR filter banks in Equation (4.12), when $\boldsymbol{C}(z) = \boldsymbol{I}$ and $\boldsymbol{v}(n) = \boldsymbol{q}(n)$.

The expression for the power used by the input vector $\boldsymbol{y}(n)$ to the channel is derived in Appendix A:

$$\mathrm{Tr}\left\{\boldsymbol{E}_{-}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}\boldsymbol{E}_{-}^{H}\right\} = P. \tag{4.24}$$

The problem is to minimize the block MSE given by Equation (4.22) with respect to the transmitter matrix $\boldsymbol{E}(z)$ and the receiver matrix $\boldsymbol{R}(z)$, subject to the power constraint given in Equation (4.24). The means of all the signals are assumed to be zero, and the second order statistics of the vector time series $\boldsymbol{x}(n)$ and $\boldsymbol{v}(n)$ are assumed to be known.

4.2.2 Equations for Optimality

In this subsection, equations for optimality are found, and a comparison is made for the corresponding equations presented in [Honig et al. 1992].

The constrained optimization problem stated in Subsection 4.2.1 can be converted to an unconstrained optimization problem by using a Lagrange multiplier. The unconstrained objective function can be expressed as

$$\mathcal{E}_{N,M}\left(d_v,d_s\right) + \mu\,\mathrm{Tr}\left\{\boldsymbol{E}_{-}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}\boldsymbol{E}_{-}^{H}\right\}, \tag{4.25}$$

where μ is the Lagrange multiplier. Necessary conditions for optimality are found by matrix differentiation of the objective function in Equation (4.25).

For a given transmitter filter bank, the equations for the optimal receiver filter bank are found by matrix differentiation of the objective function in Equation (4.25) with respect to the matrix $\boldsymbol{R}_{\_}$. By using the formulas given in Appendix C, it can be shown that the optimized receiver filter bank for a given analysis filter bank and delay values d_v and d_s can be found by solving the following equations:

$$\begin{aligned}\boldsymbol{R}_{\_} = &\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s)\right)^H \left((\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil}){}_{\searrow}\right)^H \\ &\cdot \left((\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil}){}_{\searrow}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}\left((\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil}){}_{\searrow}\right)^H + \boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}\right)^{-1}. \end{aligned} \tag{4.26}$$

This is a generalization of the result as found in [Honig et al. 1992], to include a scalar delay d_s, and not only a vector delay d_v.

For a given receiver matrix, the equations for finding the optimized transmitter matrix are obtained by matrix differentiation of the unconstrained objective function in Equation (4.25) with respect to the transmitter matrix $\boldsymbol{E}_{\_}$. By using the results from Appendix C, it can be shown that these equations can be expressed as

$$\begin{aligned}\boldsymbol{C}_{\lfloor}^H\mathcal{T}_2\left\{\boldsymbol{R}_{\lfloor}^H\mathcal{T}_1\left\{\boldsymbol{R}_{\_}(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil}){}_{\searrow}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)} - \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s)\right)^H\right\}\right\} \\ = -\mu\boldsymbol{E}_{\_}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}, \end{aligned} \tag{4.27}$$

where the operator $\mathcal{T}_1 : \mathbb{R}^{N\times(m+o+l+1)N} \rightarrow \mathbb{R}^{(l+1)N\times(m+o+1)N}$ produces an $(l+1)N\times(m+o+1)N$ block Toeplitz matrix from an $N\times(m+o+l+1)N$ matrix, and the operator $\mathcal{T}_2 : \mathbb{R}^{M\times(m+o+1)N} \rightarrow \mathbb{R}^{(o+1)M\times(m+1)N}$ produces an $(o+1)M\times(m+1)N$ block Toeplitz matrix from an $M\times(m+o+1)N$ matrix, see Appendix C, where both $\mathcal{T}_1$ and $\mathcal{T}_2$ are defined.

Equation (4.27) is not the same equation found in [Honig et al. 1992]. In Subsection 4.2.3.2, it will be shown by design examples that the formula presented in [Honig et al. 1992] does not give the same results as the proposed formula for a correlated source, but the same results are found for an uncorrelated source.

The optimization algorithm used for optimizing the power constrained FIR filter banks can be derived from the algorithm presented in the bit constrained case in Subsection 4.1.3. It is straightforward to obtain the algorithm from Table 4.1, and the power constrained algorithm will not be presented here.

Optimization for arbitrary given filter lengths in the FIR power constrained problem can be done with a similar procedure as shown in Subsection 4.1.4.

The bit and power constrained problems considered in this chapter are equivalent when $M = 1$ and $\boldsymbol{C}(z) = \boldsymbol{I}$, since under these conditions the block MSE is equal in both problems, and the bit and power constraint are equivalent.

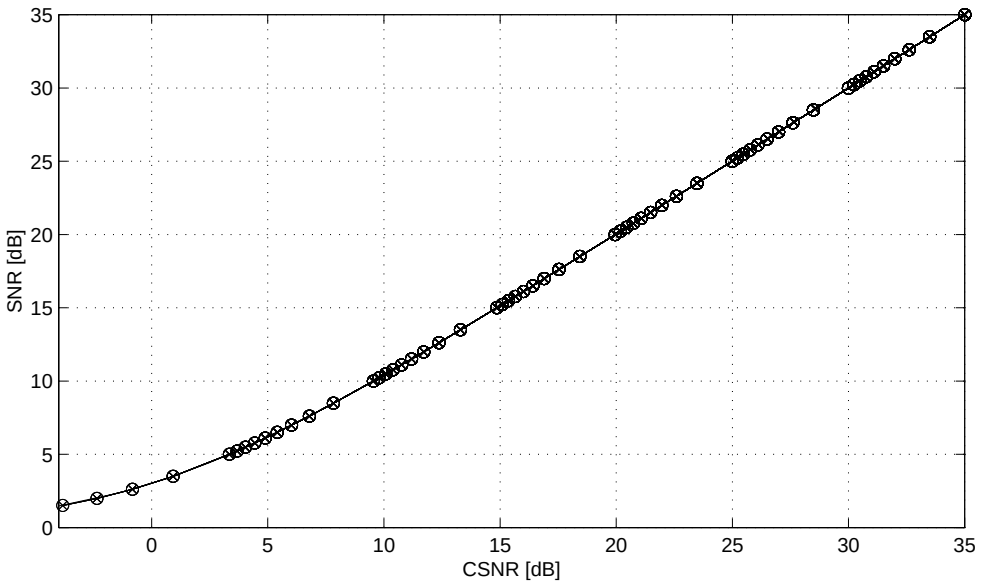

Figure 4.5 SNR vs. CSNR performance of proposed power constrained FIR filter banks are shown by the ×-marks while the system proposed in [Honig et al. 1992] is shown by the circles. Both the input time series and the additive channel noise are white and Gaussian, $N = M = 2$, $m = l = 2$, $o = 0$, $d_v = 1$, $d_s = 0$, and $\boldsymbol{C}(z) = \boldsymbol{I}$. 6_6 filter banks are used.

4.2.3 Power Constrained FIR Filter Bank Results

In this subsection, results in the FIR power constrained case are included, and comparisons are made to the methods proposed in [Honig et al. 1992] and [Malvar 1986]. In order to compare the results found in [Malvar 1986], the theory will be extended to include additive signal independent noise on the original signal.

4.2.3.1 Comparison to a Method Proposed by Honig et al.

In [Honig et al. 1992], a formula is given for finding the optimal FIR transmitter matrix for a given receiver matrix $\boldsymbol{R}(z)$ and Lagrange multiplier μ. This is Equation (4.14) in [Honig et al. 1992]. By two examples, it will be shown that this formula gives correct results for an uncorrelated source, but it results in incorrect results when a correlated input source is used.

If the performance of the proposed system is compared to the results found by the formulas in [Honig et al. 1992], the same results are obtained if the

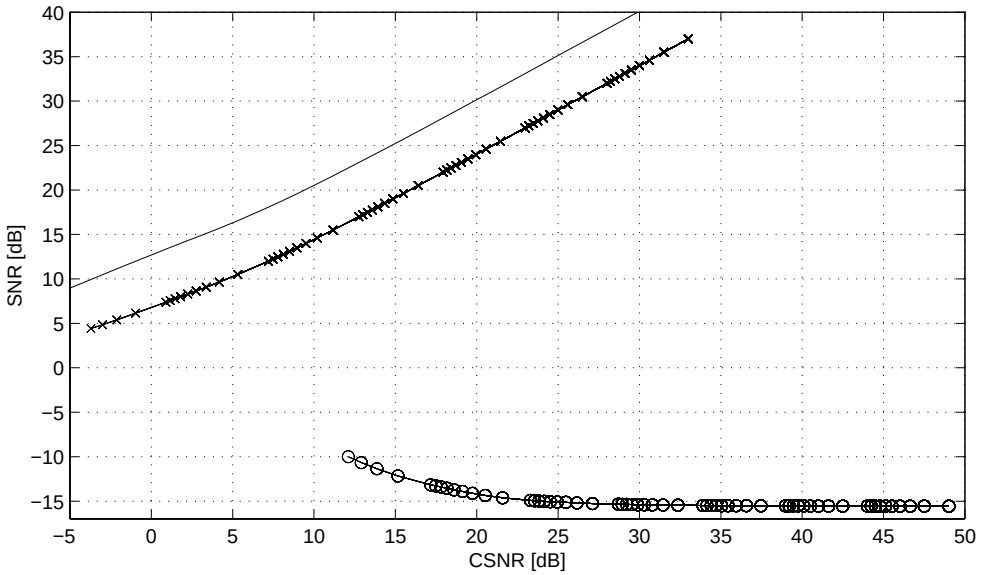

Figure 4.6 SNR vs. CSNR performance of proposed power constrained FIR filter banks are shown by the ×-marks while the system proposed in [Honig et al. 1992] is shown by the circles. The input time series is a Gaussian AR(1) time series with correlation coefficient 0.95, Gaussian white noise is added on the channel, with: $N = M = 2$, $m = l = 2$, $o = 0$, $d_v = 1$, $d_s = 0$, and $\boldsymbol{C}(z) = \boldsymbol{I}$. 5_3 filter banks are used. The upper curve is OPTA.

input signal is uncorrelated. This is shown in Figure 4.5 where the identity channel matrix is used, the input is uncorrelated, $N = M = 2$, $m = l = 2$, and $o = 0$. 6_6 filter banks are used in both systems.

From Figure 4.5, it is seen that the same performance is achieved by the system proposed in this section and the system in [Honig et al. 1992]. If the OPTA curve is compared to the results obtained in Figure 4.5, it is equal to the performance of the two systems. It can be shown that this is always the case if the input signal is white Gaussian, $N = M$, $\boldsymbol{C}(z) = \boldsymbol{I}$, and the channel noise is signal independent, white and Gaussian [Berger 1971].

In Figure 4.6, the performance of the proposed system using 5_3 filter banks when $\boldsymbol{C}(z) = \boldsymbol{I}$, and $N = M = 2$ is shown by the ×-marks. The OPTA curve is the upper curve in the figure. The curve with the circles is obtained by using the synthesis filters of the proposed system with the analysis filter bank found by Equation (4.12) in [Honig et al. 1992]. The rightmost circle and ×-mark correspond to the same value of Lagrange multiplier μ.

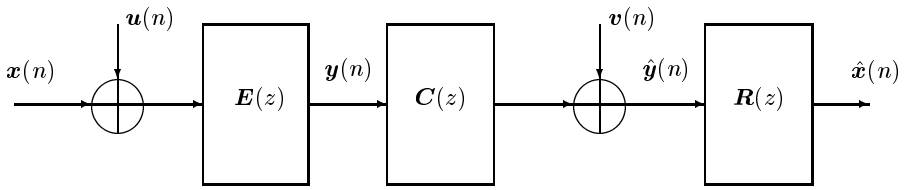

Figure 4.7 The power constrained MIMO block model with noise added to the original signal.

From Figure 4.6, it is seen that the performance obtained by Equation (4.14) in [Honig et al. 1992] gives negative values of SNR. Therefore, this formula must be wrong for the correlated input source used in the example. The correct formula is given by Equation (4.27). Figure 4.5 can also be obtained by using the synthesis filter bank from the proposed system in this section and Equation (4.14) in [Honig et al. 1992]. Equation (4.14) in [Honig et al. 1992] is therefore correct for the uncorrelated source used in Figure 4.5.

In [Crespo, Honig & Steiglitz 1989], which is an earlier article by the same authors as [Honig et al. 1992], it was assumed that the input signal was uncorrelated, and it was mentioned that the results were easy to generalize to correlated sources. This is what has been attempted in [Honig et al. 1992], but the equation for an optimal transmitter filter bank for a given synthesis filter bank is wrong for the correlated source used in the example shown in Figure 4.6.

4.2.3.2 Comparison to a Method Proposed by Malvar

In [Malvar 1986], jointly optimal FIR analysis and synthesis filter banks with linear phase were proposed for $M = 1$. Only one filter was used in both the analysis and synthesis filter bank, and it was assumed that signal independent noise was added both at the original signal and at the channel. In order to make a fair comparison to the result in [Malvar 1986], the theory developed in this section will be extended to include additive noise on the original signal.

Figure 4.7 shows the MIMO block system when noise is added to the original signal. The additive noise vector has dimension $N \times 1$, and it will be denoted $\boldsymbol{u}(n)$. It is assumed that this noise vector has zero mean and that it is uncorrelated with the original signal vector $\boldsymbol{x}(n)$ and the additive channel noise vector $\boldsymbol{v}(n)$ for all lags.

It is straightforward to extend the theory developed earlier in this section to include the additive noise on the input signal. Expressions for the block

MSE and the power used by the system must be found. The block MSE can be found in a similar manner as was done when deriving Equation (4.22), and the result is:

$$\begin{aligned}\mathcal{E}_{N,M}(d_v,d_s) = \mathrm{Tr}\Big\{ & \boldsymbol{R}_{\text{-}}(\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\left(\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}+\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m+o+l,N)}\right)\left((\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\right)^H\boldsymbol{R}_{\text{-}}^H \\ & -\boldsymbol{R}_{\text{-}}(\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s) \\ & -\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s)\right)^H\left((\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\right)^H\boldsymbol{R}_{\text{-}}^H \\ & +\boldsymbol{\Phi}_{\boldsymbol{x}}^{(0,N)}+\boldsymbol{R}_{\text{-}}\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}\boldsymbol{R}_{\text{-}}^H\Big\},\end{aligned} \tag{4.28}$$

where the matrix $\boldsymbol{\Phi}_{\boldsymbol{u}}^{(p,N)}$ is an $(p+1)N\times(p+1)N$ autocorrelation matrix for the $(p+1)M\times 1$ additive noise vector $\boldsymbol{u}(n)_{\text{ı}}$, and $\boldsymbol{\Phi}_{\boldsymbol{u}}^{(p,N)}$ is defined in a similar way as in Equation (4.8).

The power used by the system will now increase due to the additive noise vector, and it can be shown that the power used by the vector $\boldsymbol{y}(n)$ in Figure 4.7, is given by

$$\mathrm{Tr}\left\{\boldsymbol{E}_{\text{-}}\left(\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}+\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m,N)}\right)\boldsymbol{E}_{\text{-}}^H\right\}=P. \tag{4.29}$$

To optimize the system with additive input noise, necessary conditions for the optimal solution are needed. These equations can be found in a similar manner as used in Subsection 4.2.2. For a given transmitter filter bank, the optimized receiver filter bank can be found by solving the following equation

$$\begin{aligned}\boldsymbol{R}_{\text{-}} = & \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s)\right)^H\left((\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\right)^H \\ & \cdot\left((\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\left(\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}+\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m+o+l,N)}\right)\left((\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\right)^H+\boldsymbol{\Phi}_{\boldsymbol{v}}^{(l,M)}\right)^{-1}.\end{aligned} \tag{4.30}$$

The equations for optimizing the transmitter filter bank for a given receiver filter bank are given by:

$$\begin{aligned}\boldsymbol{C}_{\text{ı}}^H\mathcal{T}_2\Big\{\boldsymbol{R}_{\text{ı}}^H\mathcal{T}_1\Big\{ & \boldsymbol{R}_{\text{-}}(\boldsymbol{C}_{\text{-}}\boldsymbol{E}_{\text{⌐}})_{\diagdown}\left(\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}+\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m+o+l,N)}\right) \\ & -\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v,d_s)\right)^H\Big\}\Big\} = -\mu\boldsymbol{E}_{\text{-}}\left(\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}+\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m,N)}\right).\end{aligned} \tag{4.31}$$

If Equations (4.30) and (4.31) are compared to Equations (4.26) and (4.27), respectively, it is seen that if the matrices $\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m+o+l,N)}$ and $\boldsymbol{\Phi}_{\boldsymbol{u}}^{(m,N)}$ are set to zero, the corresponding equations are equal.

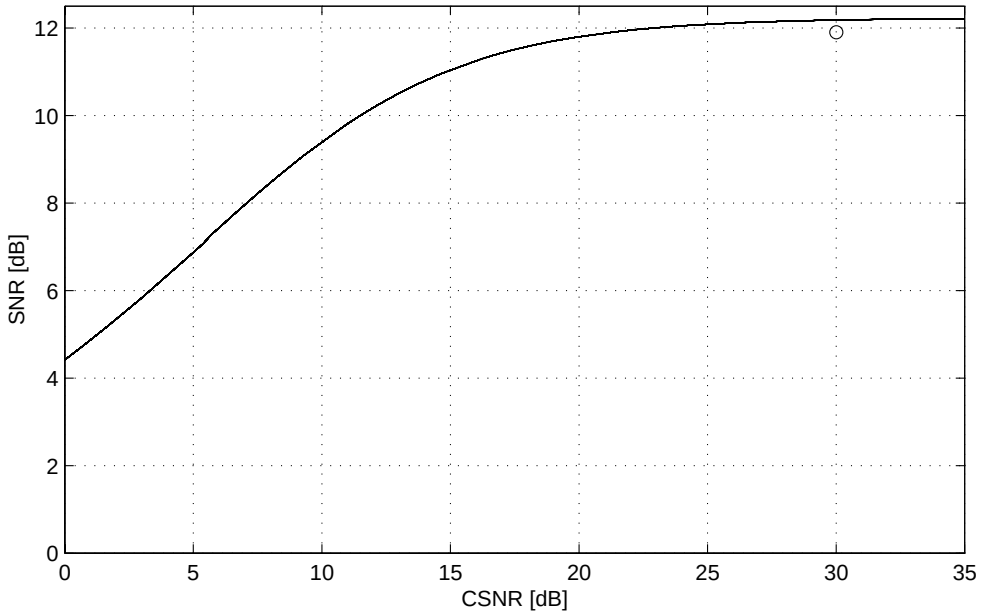

Figure 4.8 The solid curve shows the SNR vs. CSNR performance using the proposed theory with the following parameters: $N = 3$, $M = 1$, $m = l = 4$, $d_v = 3$, $d_s = 1$, and the input PSD is an AR(1) source with correlation coefficient 0.9. The circle shows the performance of the system proposed in [Malvar 1986]. ISNR = 30.0 dB in both systems.

In Section 3.4 in [Malvar 1986], an example is given using the following parameters: $N = 3$, $M = 1$, $m = l = 4$, $d_v = 3$, $d_s = 1$, and the input PSD is an AR(1) source with correlation coefficient 0.9. Furthermore, it was assumed that the subband samples were uncorrelated with the white additive channel noise. Linear phase is assumed in [Malvar 1986], and the filter lengths are 13 in both the analysis and synthesis filter. The same parameters were used in the proposed theory. Figure 4.8 shows the results achieved with the proposed method.

Let the input signal to noise ratio (ISNR) be defined as the ratio between the input signal variance and the input noise variance σ_x^2/σ_u^2, where σ_u^2 is the variance of the additive input noise. In this definition, it is assumed that all the vector components of $\boldsymbol{u}(n)$ have zero mean and the same variance σ_u^2. In Figure 4.8, ISNR = 30.0 dB.

From Figure 4.8, it is seen that the performance of the proposed system reaches a certain limit when the CSNR values are increased. The reason for

this is that $M = 1$, and with $M < N$ it is impossible to achieve PR. In the example in [Malvar 1986], the noise level in both the input signal and on the channel was 30.0 dB. Then the overall performance of the system was reported to have an SNR = 11.9 dB in [Malvar 1986]. This is 0.29 dB worse than the results obtained by the proposed system.

In [Malvar 1986], linear phase impulse responses were imposed in the optimization of the system. In the proposed system, no such restriction was made, and the resulting filter banks do not have linear phase. This shows that linear phase might lead to suboptimal solutions because the number of free parameters is reduced.

If M is increased, the performance of the system will improve at the expense of increasing the bandwidth used on the channel.

4.3 Summary

An optimization algorithm was proposed for finding FIR signal-adaptive jointly optimized analysis and synthesis filter banks with arbitrary given filter lengths and delay through the filter banks, under a bit constraint. The results show that the filter banks have neither linear phase nor PR in general. The proposed filter banks are source and rate dependent, which is not the case for PR systems. In Table 4.2, it was shown that the results for the proposed method are better than other well known filter banks found in the literature with the same filter lengths. For short filter lengths, the improvement is low, but for longer filters a significant improvement was achieved.

Equations for finding jointly optimal power constrained FIR analysis and synthesis filter bank were proposed. The equations for the synthesis filter bank are a generalization of the FIR Wiener filter bank equations found in the literature, to include arbitrary given filter lengths and arbitrary scalar and vector delay.

Both the bit and power constrained filter banks proposed in this chapter can be considered to be a generalization of the method proposed in [Malvar 1986], where an algorithm for finding jointly optimal analysis and synthesis filter banks having one analysis and one synthesis filter, i.e., $M = 1$, was given. It was assumed that the channel was power constrained. However, since only one channel was used, the power constraint is equivalent to the bit constraint, see Equation (4.13). The power constrained case was extended to include additive noise on the original signal, and this can also be done for the bit constrained problem.

Chapter 5

Connection between BPAM and Power Constrained MIMO

In this chapter, a communication system is studied where a continuous amplitude, discrete time, stationary source signal is transmitted over a power constrained, discrete time, continuous amplitude channel. The channel is assumed to have an identity transfer function and is contaminated by additive white noise. For a given channel quality, which can be measured by the CSNR, the objective is to design a communication system which minimizes the distortion between the original and the received signals at a given channel bandwidth.

Different system solutions exist, some of which are considered in [Fuldseth & Ramstad 1997, Vaishampayan 1989, Lee & Petersen 1976, Hjørungnes & Ramstad 1997]. Here, only *linear* solutions are studied. It is shown in this chapter that there exists a connection between the performances of the BPAM system considered in Section 3.2 and the power constrained MIMO system of infinite filter lengths presented in Section 2.2. It is shown that when the dimensions of the BPAM system approach infinity, the performance of the BPAM system approaches the performance of the unconstrained length MIMO system when a pre- and postprocessor containing modulation is used in the latter system.

The current chapter is organized as follows: In Section 5.1, the problem is formulated. The two system solutions to be compared, the BPAM and unconstrained length MIMO systems, are presented in Sections 5.2 and 5.3, respectively. The actual comparison of the two system solutions is given in Section 5.4, and a numerical example illustrating the main results in this chapter is presented in Section 5.5. Section 5.6 briefly summarizes this chapter.

This chapter is partly based on [Hjørungnes & Ramstad 1998*c*].

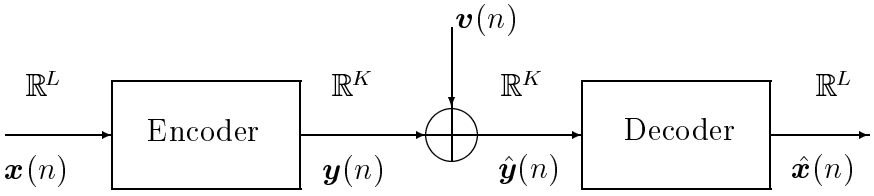

Figure 5.1 Block diagram of the communication system.

5.1 Problem Formulation

The transmission system considered in this chapter is shown in Figure 5.1. It is a block based system, where one output vector is produced for each input vector. In this BPAM system, L is used for the number of source samples in the source vector $\boldsymbol{x}(n)$ and K is used for the number of channel samples in the channel input vector $\boldsymbol{y}(n)$. The corresponding symbols in the MIMO system are N and M. The reason for using different symbols is to make the comparison of the system easier.

The time series $x(n)$, representing the source signal, is characterized by its PSD $S_x(f)$. The encoder is a device which maps L source samples into K channel samples using only a given amount of transmitter power. The output of the encoder is transmitted over a memoryless channel with additive white noise $v(n)$ which has variance σ_v^2. Since the BPAM system is developed for an identity channel transfer matrix, it is assumed that $\boldsymbol{C}(z) = \boldsymbol{I}$ in this chapter. The decoder is designed to reconstruct the L source samples from the K channel samples with as little distortion as possible. Let the $L \times 1$ source vector $\boldsymbol{x}(n)$ and the $K \times 1$ channel vector $\boldsymbol{y}(n)$ be given as shown by Figure 5.1. Furthermore, let $\boldsymbol{v}(n)$ be a $K \times 1$ vector containing the additive channel noise. The received vector is thus $\hat{\boldsymbol{y}}(n) = \boldsymbol{y}(n) + \boldsymbol{v}(n)$. The receiver is an estimator of $\boldsymbol{x}(n)$ producing the output $\hat{\boldsymbol{x}}(n)$. All signals are assumed to have zero mean.

The number of channel samples used per source sample is an arbitrary non-negative rational number, thus both bandwidth compression and expansion are possible.

Given L, K, and the statistical descriptions of the source and noise signals, the problem is to find an encoder-decoder pair that minimizes the expected value of the block MSE between the source vector $\boldsymbol{x}(n)$ and the decoded vector $\hat{\boldsymbol{x}}(n)$, i.e., min $E\left[\|\boldsymbol{x}(n) - \hat{\boldsymbol{x}}(n)\|^2\right]$, subject to the power constraint $E\left[\|\boldsymbol{y}(n)\|^2\right] = P$, where P is a given positive constant.

5.2 The BPAM System

In [Lee & Petersen 1976], the BPAM system was developed. It is a linear block based solution for communication of vectors over a *memoryless* vector channel. The linear vector system can therefore be used to solve the problem described in Section 5.1. This system groups the time series into blocks of L source samples, and the $L \times 1$ source vector is given by $\boldsymbol{x}(n) = [x(nL), x(nL-1), \ldots, x(nL-(L-1))]^T$. Each source vector is transformed into K channel samples by a $K \times L$ memoryless matrix. Then, the vector of K channel samples is transmitted over the channel, and an estimate of the L samples is recovered by an $L \times K$ matrix. BPAM treats each source vector independently and does not utilize inter-vector correlation. An alternative derivation of the BPAM system is given in Section 3.2.

Autocovariance matrices are used to describe the statistics of both the source and noise vectors in the BPAM system. The autocovariance matrix of the input signal is an $L \times L$ matrix and has L eigenvalues, which are not necessarily distinct. The eigenvalues are ordered in descending order, i.e., according to Equation (3.7). Since the channel is assumed to be a memoryless vector channel with white noise components all having variance σ_v^2, all K eigenvalues of the $K \times K$ noise autocovariance matrix $\boldsymbol{K_v}(0)$ at lag zero are equal to σ_v^2. Therefore, $\kappa_i^{(K)} = \sigma_v^2$ for all $i \in \{0, 1, \ldots, K-1\}$.

From Equations (3.38) and (3.42), the power used per *source sample* by the BPAM system $\mathcal{P}_{L,K}^{\text{BPAM}}(\mu)$ is given by

$$\mathcal{P}_{L,K}^{\text{BPAM}}(\mu) = \frac{1}{L} \sum_{i=0}^{\min(K,L)-1} \max\left\{0, \sqrt{\frac{\sigma_v^2 \lambda_i^{(L)}}{\mu}} - \sigma_v^2\right\}, \tag{5.1}$$

where μ is a Lagrange multiplier, and $\lambda_i^{(L)}$ is eigenvalue number i of the $L \times L$ autocovariance matrix $\boldsymbol{K_x}(0)$ given in Equation (3.1). Equations (3.1) and (3.7) must be used with the correct dimensions, which means that L eigenvalues exist with the notation introduced for the BPAM system.

The MSE per source sample is found from Equations (3.38) and (3.41) and is given by

$$\varepsilon_{L,K}^{\text{BPAM}}(\mu) = \begin{cases} \dfrac{1}{L}\left(\displaystyle\sum_{i=0}^{K-1} \min\left\{\lambda_i^{(L)}, \sqrt{\sigma_v^2 \mu \lambda_i^{(L)}}\right\} + \sum_{i=K}^{L-1} \lambda_i^{(L)}\right), & \text{if } K < L, \\ \dfrac{1}{L}\displaystyle\sum_{i=0}^{L-1} \min\left\{\lambda_i^{(L)}, \sqrt{\sigma_v^2 \mu \lambda_i^{(L)}}\right\}, & \text{if } K \geq L. \end{cases} \tag{5.2}$$

If Equations (5.1) and (5.2) are compared to Equations (B.2) and (B.3) on page 141 in [Vaishampayan 1989], it is observed that the expressions are different, and the reason is the incorrect use of the parameter θ in [Vaishampayan 1989]. The parameter θ in [Vaishampayan 1989] depends on an index, but is incorrectly treated as a constant. In [Vaishampayan 1989], the expressions given for the performance of the BPAM system and the corresponding graphs are inconsistent. The same mistake was made in [Vaishampayan & Farvardin 1992].

5.3 The MIMO System

In Section 2.2, jointly optimal transmitter and receiver filter banks were found for transmitting a vector time series over a vector channel with memory. The filters are unconstrained, i.e., they are non-causal and are allowed to have infinite length impulse responses. The MIMO system can therefore utilize first and second order correlation between vectors, and each vector is *not* treated independently, as it is in the BPAM system. For N source samples, M channel samples are used. With a suitable pre- and postprocessor, this system can be used to solve the problem in Section 5.1. In Subsection 5.4.2, a system with pre- and postprocessor containing modulation and linear filters is proposed.

If the MIMO system is used to solve the problem described in Section 5.1, the vector channel is assumed to be memoryless with a transfer matrix given by the identity matrix, and the noise in each sub-channel is modeled as additive white noise with variance σ_v^2. If these assumptions are inserted in Equations (2.75) and (2.78), the following expression is found for the power used per *source sample*

$$\mathcal{P}_{N,M}^{\text{MIMO}}(\mu) = \sum_{i=0}^{\min(M,N)-1} \int_{-\frac{1}{2N}}^{\frac{1}{2N}} \max\left\{0, \sqrt{\frac{\sigma_v^2 \lambda_i^{(N)}(fN)}{\mu}} - \sigma_v^2\right\} df, \tag{5.3}$$

where μ is a Lagrange multiplier, and $\lambda_i^{(N)}(f)$ is eigenvalue number i of the $N \times N$ PSD matrix $\boldsymbol{S_x}(f)$ modeling the source statistics. The PSD matrix is defined in Equation (1.8) and its eigenvalues are ordered according to Equation (2.7).

The MSE per source sample is found from Equations (2.75) and (2.77):

$$\varepsilon_{N,M}^{\mathrm{MIMO}}(\mu) = \begin{cases} \displaystyle\sum_{i=0}^{M-1} \int_{-\frac{1}{2N}}^{\frac{1}{2N}} \min\left\{\lambda_i^{(N)}(fN), \sqrt{\sigma_v^2 \mu \lambda_i^{(N)}(fN)}\right\} df \\ \qquad + \displaystyle\sum_{i=M}^{N-1} \int_{-\frac{1}{2N}}^{\frac{1}{2N}} \lambda_i^{(N)}(fN)\, df, & \text{if } M < N, \\ \displaystyle\sum_{i=0}^{N-1} \int_{-\frac{1}{2N}}^{\frac{1}{2N}} \min\left\{\lambda_i^{(N)}(fN), \sqrt{\sigma_v^2 \mu \lambda_i^{(N)}(fN)}\right\} df, & \\ & \text{if } M \geq N. \end{cases} \tag{5.4}$$

5.4 Comparison of the BPAM and Modulated MIMO Systems

In this section, alternative expressions are found for the performance of the MIMO system with a pre- and postprocessor containing linear filters and modulators. In addition, expressions are derived for the performance of the BPAM system when the dimensions of the matrices tend to infinity. These expressions are used to determine a connection between the performances of the two systems in this limiting case.

5.4.1 BPAM with Dimensions Approaching Infinity

Two cases are treated separately: Case 1: $K \geq L$ and Case 2: $K < L$.

5.4.1.1 Case 1: $K \geq L$

Letting $L \to \infty$ in Equation (5.2), and then using the Toeplitz distribution theorem [Grenander & Szegö 1958], results in

$$\begin{aligned} \lim_{L\to\infty} \varepsilon_{L,K}^{\mathrm{BPAM}}(\mu) &= \lim_{L\to\infty} \varepsilon_{L,L}^{\mathrm{BPAM}}(\mu) \\ &= \lim_{L\to\infty} \frac{1}{L} \sum_{i=0}^{L-1} \min\left\{\lambda_i^{(L)}, \sqrt{\sigma_v^2 \mu \lambda_i^{(L)}}\right\} \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \min\left\{S_x(f), \sqrt{\sigma_v^2 \mu S_x(f)}\right\} df. \end{aligned} \tag{5.5}$$

The same reasoning can be used for the power used per source sample given in Equation (5.1). The result is given by

$$\begin{aligned}\lim_{L\to\infty} \mathcal{P}_{L,K}^{\mathrm{BPAM}}(\mu) &= \lim_{L\to\infty} \mathcal{P}_{L,L}^{\mathrm{BPAM}}(\mu) \\ &= \lim_{L\to\infty} \frac{1}{L} \sum_{i=0}^{L-1} \max\left\{0, \sqrt{\frac{\sigma_v^2 \lambda_i^{(L)}}{\mu}} - \sigma_v^2\right\} \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \max\left\{0, \sqrt{\frac{\sigma_v^2 S_x(f)}{\mu}} - \sigma_v^2\right\} df. \end{aligned} \tag{5.6}$$

5.4.1.2 Case 2: $K < L$

When $K < L$, one cannot directly apply the Toeplitz distribution theorem because the following limit should be determined:

$$\lim_{L\to\infty} \varepsilon_{L,K}^{\mathrm{BPAM}}(\mu) = \lim_{L\to\infty} \frac{1}{L} \left(\sum_{i=0}^{K-1} \min\left\{\lambda_i^{(L)}, \sqrt{\sigma_v^2 \mu \lambda_i^{(L)}}\right\} + \sum_{i=K}^{L-1} \lambda_i^{(L)} \right). \tag{5.7}$$

The sum in Equation (5.7) consists of L addends, but all the addends are not of the same mathematical formula, and therefore, the Toeplitz distribution theorem cannot be directly applied.

It is known that the two sets

$$\left\{\lambda_i^{(L)}\right\} \text{ and } \left\{S_x\left(-\frac{1}{2} + \frac{i+1}{L+1}\right)\right\}, \tag{5.8}$$

are equally distributed [Grenander & Szegö 1958], where $i \in \{0, 1, \ldots, L-1\}$ and $L \to \infty$. By using this result in Equation (5.7), it is seen that the two sums are Riemann sums [Edwards & Penney 1986] with a *regular partition*. Furthermore, assume that $S_x(f)$ is Riemann integrable.

The *fundamental period* is defined as the frequency interval $\left(-\frac{1}{2}, \frac{1}{2}\right]$. Define $\underline{\Gamma}(\gamma)$ to be the set of frequencies in the fundamental period of total length γ giving the smallest values of $S_x(f)$, and define $\overline{\Gamma}(\gamma)$ to be the set of frequencies of the fundamental period of total length γ giving the largest values of $S_x(f)$. If $S_x(f)$ is constant in some frequency regions, the sets $\underline{\Gamma}(\gamma)$ and $\overline{\Gamma}(\gamma)$ might not be uniquely defined, but in this case the final result is independent of the choice of regions. The expression $\lim_{L\to\infty} \varepsilon_{L,K}^{\mathrm{BPAM}}(\mu)$ can then be written as the

following integral:

$$\lim_{L\to\infty} \varepsilon_{L,K}^{\mathrm{BPAM}}(\mu) = \lim_{L\to\infty} \frac{1}{L}\left(\sum_{i=0}^{K-1} \min\left\{\lambda_i^{(L)}, \sqrt{\sigma_v^2\mu\lambda_i^{(L)}}\right\} + \sum_{i=K}^{L-1} \lambda_i^{(L)}\right)$$
$$= \int_{\overline{\Gamma}(\frac{K}{L})} \min\left\{S_x(f), \sqrt{\sigma_v^2\mu S_x(f)}\right\} df + \int_{\underline{\Gamma}(1-\frac{K}{L})} S_x(f)\, df. \tag{5.9}$$

Here, the ordering given in Equation (3.7) has been used to determine the region of integration.

The same derivation can be applied to $\lim_{L\to\infty} \mathcal{P}_{L,K}^{\mathrm{BPAM}}(\mu)$. The result will simply be stated here as:

$$\lim_{L\to\infty} \mathcal{P}_{L,K}^{\mathrm{BPAM}}(\mu) = \int_{\overline{\Gamma}(\frac{K}{L})} \max\left\{0, \sqrt{\frac{\sigma_v^2 S_x(f)}{\mu}} - \sigma_v^2\right\} df. \tag{5.10}$$

5.4.2 Modulated MIMO System

The MIMO system presented in Section 5.3 can be used for any preprocessing scheme which produces a vector as input to the MIMO system. A delay chain with decimation, which is usually used before the analysis polyphase filter bank could be used to produce the vector input. However, such a system will not perform well when fewer channel samples than source samples are used, i.e., when $K < L$ with the notation used in Figure 5.1. The reason why the MIMO system with decimation does not perform very well is, that in order to have alias cancellation through the whole system, the decimation process puts severe constraints on the frequency regions of the filters, see Chapter 2 and Appendix B. However, when the number of channel samples are greater than or equal to the number of source samples, i.e., when $K \geq L$ with the symbols used in Figure 5.1, the performance of the BPAM system will approach the performance of the MIMO system as the dimensions of the BPAM system tend to infinity, even if a delay chain with decimators is used as the preprocessing unit.

In order to get results that are valid for any choice of source and channel samples, a pre- and postprocessing scheme is introduced. This is based on ideal linear filters and linear modulation. The preprocessing unit is shown in Figure 5.2, and the MIMO system using this preprocessor and the corresponding postprocessor will be called *modulated MIMO system*.

The following definition is needed: Let $\Gamma_i^{(N)}$ be the frequency region of length $\frac{1}{N}$ of the fundamental period where $S_x(f)$ has the ith largest values,

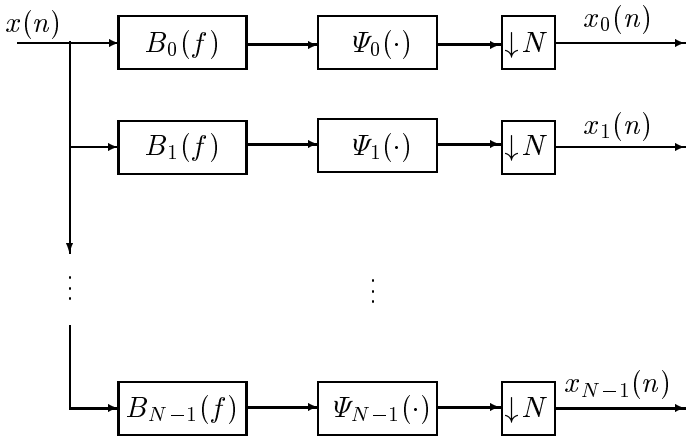

Figure 5.2 Preprocessing used in the modulated MIMO system.

where $i \in \{0, 1, \ldots, N-1\}$. An example of such a region is shown in Figure 5.3 (b), where the set of frequencies where the PSD is greater than zero is equal to $\Gamma_0^{(N)}$, with $N = 3$. The way these frequency regions are chosen is related to reverse water-filling [Berger 1971, Jayant & Noll 1984, Cover & Thomas 1991]. If the PSD function $S_x(f)$ is piecewise constant, the sets $\Gamma_i^{(N)}$ might not be uniquely defined, but this does not affect the final result.

In Figure 5.2, $B_i(f)$ is a digital filter, and it is therefore periodic with period 1. Within the fundamental period, the filter $B_i(f)$ is defined as:

$$B_i(f) = \begin{cases} \sqrt{N}, & \text{if } f \in \Gamma_i^{(N)}, \\ 0, & \text{if } f \notin \Gamma_i^{(N)}. \end{cases} \tag{5.11}$$

These filters are related to optimal unitary filter banks for source coding, but the frequency region for each filter is in general different [Vaidyanathan 1998].

After the filters, linear modulation is used. The modulation after filter $B_i(f)$ is performed by the operator $\Psi_i(\cdot)$. This modulator takes the output from the filter, which has non-zero frequency components of bandwidth $\frac{1}{N}$, and modulates this signal such that the output of the modulator has non-zero frequency components in the interval $\left(-\frac{1}{2N}, \frac{1}{2N}\right]$ and zero frequency components in the rest of the fundamental period. The modulator hence preserves the bandwidth of the signal, and the output of the modulator is periodic in the frequency domain with period equal to 1.

After the modulators, decimators with factor N are used to keep the number of samples into the preprocessor equal to the number out of the preprocessor. Since ideal filters are used and the signal into the decimators is non-zero

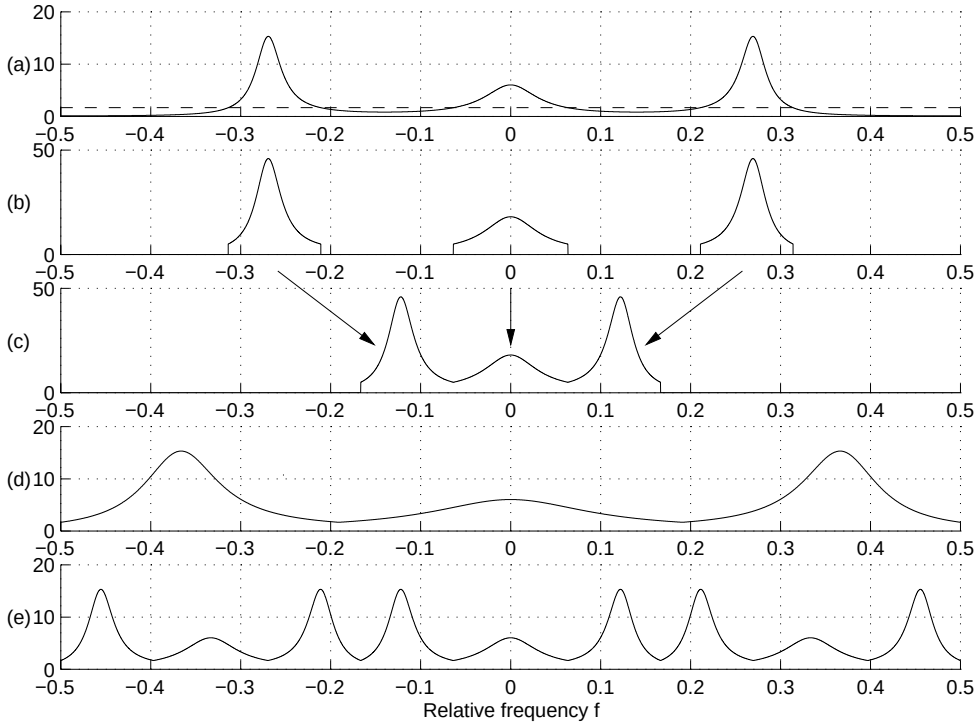

Figure 5.3 Illustration of PSDs in branch number 0 in the preprocessor when an AR(3) input source is used with $N = 3$. (a) PSD of the input signal $S_x(f)$. (b) PSD of the output of the filter $B_0(f)$. (c) PSD of the output of the modulator $\Psi_0(\cdot)$. (d) PSD of the output of the decimator in branch number 0, which is equal to $\lambda_0^{(N)}(f)$. (e) $\lambda_0^{(N)}(fN)$.

only in the frequency interval $\left(-\frac{1}{2N}, \frac{1}{2N}\right]$ of the fundamental period, there is no aliasing error introduced by the decimators.

To show how the preprocessor works, an example showing different PSDs in the preprocessor is shown in Figure 5.3. In part (a) of the figure, the PSD of the input source is shown. The input signal is an AR(3) signal, with poles at 0.8, $0.9e^{j2\pi 0.27}$, and $0.9e^{-j2\pi 0.27}$. The parts of the PSD that are below the dashed line in part (a) will be removed by $B_0(f)$. Part (b) shows the PSD of the output of $B_0(f)$. This spectrum is calculated according to $|B_0(f)|^2 S_x(f)$. The modulator $\Psi_0(\cdot)$ operates on the signal such that the output of the modulator is non-zero in the interval $\left(-\frac{1}{2N}, \frac{1}{2N}\right]$ and has zero frequency components in the rest of the fundamental period. This is shown in part (c) of the figure, and the arrows show how the spectrum has changed through the modulator.

In part (d), the PSD of the output of the decimator is given, and this is equal to $\lambda_0^{(N)}(f)$. Finally, part (e) shows $\lambda_0^{(N)}(fN)$, which is needed in some of the upcoming calculations.

On the receiver side, a postprocessor is used, consisting of expansion with factor N, inverse modulation, and filtering with the filters $B_i(f)$.

To be able to compare the performance expressions for the MIMO system to the BPAM system, the performance expressions for the MIMO system need to be reformulated. In order to do so, expressions for the PSD of the components $x_i(n)$ in Figure 5.2 are needed. Since the filters in Equation (5.11) are non-overlapping, the signals $x_i(n)$ are uncorrelated with each other [Ramstad, Aase & Husøy 1995]. Therefore, the matrix $\boldsymbol{S_x}(f)$ is diagonal, with diagonal element number i given by the PSD of $x_i(n)$. Due to the way the preprocessor in Figure 5.2 is designed, the PSD of $x_i(n)$ is equal to $\lambda_i^{(N)}(f)$.

Again, two cases are treated separately: Case 1: $M \geq N$ and Case 2 $M < N$.

5.4.2.1 Case 1: $M \geq N$

From Equation (5.4), it is seen that $\varepsilon_{N,M}^{\text{MIMO}}(\mu) = \varepsilon_{N,N}^{\text{MIMO}}(\mu)$ when $M \geq N$.

First, it is shown that $\varepsilon_{N,N}^{\text{MIMO}}(\mu)$ is independent of N. Equation (5.4) gives

$$\varepsilon_{N,N}^{\text{MIMO}}(\mu) = \sum_{i=0}^{N-1} \int_{-\frac{1}{2N}}^{\frac{1}{2N}} \min\left\{\lambda_i^{(N)}(fN), \sqrt{\sigma_v^2 \mu \lambda_i^{(N)}(fN)}\right\} df. \tag{5.12}$$

In Equation (5.12), the integrand is periodic with period $\frac{1}{N}$, and the length of each integration region is $\frac{1}{N}$. Figure 5.3 shows how $\lambda_0^{(N)}(fN)$ is found from $S_x(f)$ when $N = 3$. From the figure, it is seen that integral number 0 in Equation (5.12) can be calculated by replacing $\lambda_0^{(N)}(fN)$ with $S_x(f)$ and integrating over the interval of length $\frac{1}{N}$ where $S_x(f)$ is largest, i.e., the interval $\Gamma_0^{(N)}$. By using the same reasoning as in Figure 5.3, $\lambda_i^{(N)}(fN)$ can be found for all $i \in \{0, 1, \ldots, N-1\}$, and integral number i in Equation (5.12) can be obtained by substituting $\lambda_i^{(N)}(fN)$ with $S_x(f)$ and integrating over the interval $\Gamma_i^{(N)}$. Using this information, it is seen that the sum of all the integrals in Equation (5.12) can be written as one integral:

$$\varepsilon_{N,N}^{\text{MIMO}}(\mu) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \min\left\{S_x(f), \sqrt{\sigma_v^2 \mu S_x(f)}\right\} df. \tag{5.13}$$

From Equation (5.13), it is seen that $\varepsilon_{N,N}^{\text{MIMO}}(\mu)$ is independent of N, so $\varepsilon_{N,N}^{\text{MIMO}}(\mu) = \varepsilon_{1,1}^{\text{MIMO}}(\mu)$ for all values of N.

Using these results one obtains

$$\begin{aligned}\varepsilon_{N,M}^{\mathrm{MIMO}}(\mu) &= \varepsilon_{N,N}^{\mathrm{MIMO}}(\mu) \\ &= \varepsilon_{1,1}^{\mathrm{MIMO}}(\mu) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \min\left\{S_x(f), \sqrt{\sigma_v^2 \mu S_x(f)}\right\} df. \end{aligned} \tag{5.14}$$

The same derivation can also be applied for $\mathcal{P}_{N,M}^{\mathrm{MIMO}}(\mu)$, giving

$$\begin{aligned}\mathcal{P}_{N,M}^{\mathrm{MIMO}}(\mu) &= \mathcal{P}_{N,N}^{\mathrm{MIMO}}(\mu) \\ &= \mathcal{P}_{1,1}^{\mathrm{MIMO}}(\mu) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \max\left\{0, \sqrt{\frac{\sigma_v^2 S_x(f)}{\mu}} - \sigma_v^2\right\} df. \end{aligned} \tag{5.15}$$

5.4.2.2 Case 2: $M < N$

Equation (5.3) gives:

$$\mathcal{P}_{N,M}^{\mathrm{MIMO}}(\mu) = \sum_{i=0}^{M-1} \int_{-\frac{1}{2N}}^{\frac{1}{2N}} \max\left\{0, \sqrt{\frac{\sigma_v^2 \lambda_i^{(N)}(fN)}{\mu}} - \sigma_v^2\right\} df. \tag{5.16}$$

The integrand in Equation (5.16) is periodic with period $\frac{1}{N}$, and the integral is calculated over one period. Figure 5.3 illustrates how $\lambda_0^{(N)}(fN)$ can be found from $S_x(f)$.

For $i = 0$ in Equation (5.16), the integrand is a function of $\lambda_0^{(N)}(fN)$. From Figure 5.3, it is seen that an alternative way to calculate this integral is to replace $\lambda_0^{(N)}(fN)$ with $S_x(f)$ and integrating over the interval of length $\frac{1}{N}$ where $S_x(f)$ is largest, i.e., the interval $\Gamma_0^{(N)}$. By setting $i = 1$ in Equation (5.16) and using the same argument, $\lambda_1^{(N)}(fN)$ can be replaced by $S_x(f)$ with the integral taken over a length of $\frac{1}{N}$ where $S_x(f)$ is second largest, i.e., the interval $\Gamma_1^{(N)}$. By doing so for all the M addends in Equation (5.16), the result is obtained by replacing $\lambda_i^{(N)}(fN)$ by $S_x(f)$ and integrating over the set of frequencies of the fundamental period of total length $\frac{M}{N}$ containing the frequency values where $S_x(f)$ is largest. This interval is $\overline{\Gamma}\left(\frac{M}{N}\right)$. Thus, Equation (5.16) can be written

$$\mathcal{P}_{N,M}^{\mathrm{MIMO}}(\mu) = \int_{\overline{\Gamma}\left(\frac{M}{N}\right)} \max\left\{0, \sqrt{\frac{\sigma_v^2 S_x(f)}{\mu}} - \sigma_v^2\right\} df. \tag{5.17}$$

By performing the same calculations for the MSE as for the power, the following result is obtained

$$\varepsilon_{N,M}^{\text{MIMO}}(\mu) = \int_{\overline{\Gamma}(\frac{M}{N})} \min\left\{S_x(f), \sqrt{\sigma_v^2 \mu S_x(f)}\right\} df + \int_{\underline{\Gamma}(1-\frac{M}{N})} S_x(f)\, df. \quad (5.18)$$

5.4.3 Comparison

By comparing Equation (5.5) to Equation (5.14) and Equation (5.9) to Equation (5.18), it is seen that

$$\lim_{L\to\infty} \varepsilon_{L,K}^{\text{BPAM}}(\mu) = \varepsilon_{N,M}^{\text{MIMO}}(\mu), \text{ when } \frac{K}{L} = \frac{M}{N}. \quad (5.19)$$

The same comparison can be made for the power measure. By comparing Equation (5.6) to Equation (5.15) and Equation (5.10) to Equation (5.17), it is seen that

$$\lim_{L\to\infty} \mathcal{P}_{L,K}^{\text{BPAM}}(\mu) = \mathcal{P}_{N,M}^{\text{MIMO}}(\mu), \text{ when } \frac{K}{L} = \frac{M}{N}. \quad (5.20)$$

5.5 Numerical Example

Figure 5.4 shows a numerical example of the performance of decimated MIMO, the modulated MIMO system, and the BPAM system with different choices of L and K. The first axis shows the CSNR in dB, and this can be expressed as $10\log_{10}\frac{\mathcal{P}_{N,M}^{\text{system}}(\mu)}{\sigma_v^2}$. The second axis shows the SNR. This can be expressed as $10\log_{10}\frac{\sigma_x^2}{\varepsilon_{N,M}^{\text{system}}(\mu)}$, where σ_x^2 is the power of the input time series.

The equality $\frac{N}{M} = \frac{L}{K}$ holds for all the BPAM results in Figure 5.4. The results for the decimated MIMO are found from the theory developed in Section 2.2.

From Figure 5.4, it is seen that as the values of L and K are increased, the performance of the BPAM system approaches the performance of the modulated MIMO system. The decimated MIMO system has poorer performance compared to the modulated MIMO system. The reason for this is the different frequency partitioning of the two MIMO systems. This example illustrates the main results of this chapter.

5.6 Summary

From Equations (5.19) and (5.20), it was concluded that the performance of the BPAM system approaches the performance of the modulated MIMO system

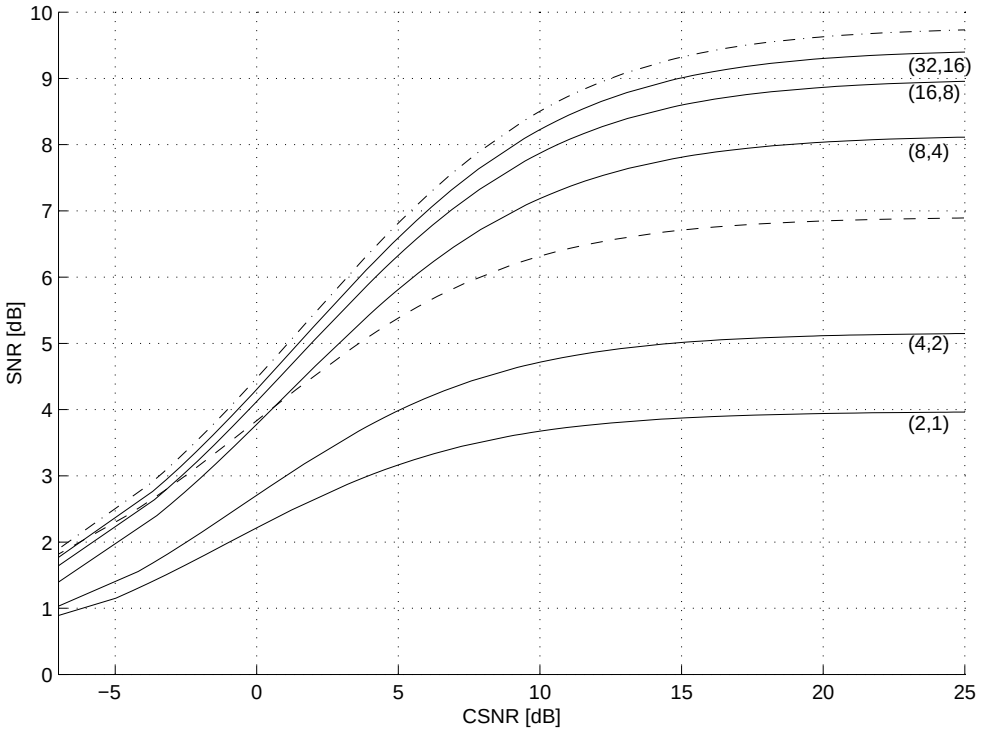

Figure 5.4 SNR vs. CSNR performance for decimated MIMO is shown by the dashed curves, modulated MIMO by the dash-dotted curves, and BPAM system by the solid curves. For both the MIMO systems the following parameters have been used: $(N, M) = (2, 1)$ and $(L, K) \in \{(2, 1), (4, 2), (8, 4), (16, 8), (32, 16)\}$ for the BPAM system. In the results for the BPAM system (solid) the values for L and K are increasing when going from bottom to top. An AR(3) source with the spectrum shown in Figure 5.3 (a) is coded.

when the ratio between K and L is the same as the ratio between M and N and when the size of K and L approaches infinity. This result is valid when the channel is memoryless and the noise on the channel is independent and identically distributed (i.i.d.).

It is also possible to use power per *channel sample* as a power measure. In that case, the result found here for the power has to be scaled by a constant.

If the number of channel samples is greater than or equal to the number of source samples, the same result holds for a MIMO system with only a delay chain and decimators in front of it, i.e., without ideal bandpass filters and modulators. However, when fewer channel samples than source samples are used,

the results in Subsection 5.4.3 do not hold for the decimated MIMO system. The reason is that in order to have alias cancellation through the MIMO system, the decimation process puts severe constraints on the frequency regions of the filters. This can be deduced from the ordering functions introduced in Chapter 2 and Appendix B.

Chapter 6

Practical Simulations for Bit Constrained FIR Filter Banks

Since a high rate model was used for modeling the coding of the subband signals, it is expected that there exists a mismatch between the results obtained by a practical coder and the results predicted by the theory at low rates. In this and the next chapter this mismatch will be examined. For the power constrained problem there exist *practical channels* which closely match the assumptions made for channels in this dissertation [Lee & Messerschmitt 1994, Bingham 1990], and therefore, this problem will not be treated in more detail.

In this chapter, a practical coding system is introduced. The performance of the practical coding system is compared to the performance obtained by the theory developed in Section 4.1, using a high rate model for the subband coding. The deviation between the theoretical and practical results is analyzed carefully. Theory for improved modeling for coding of the subband signals is proposed in the next chapter.

This chapter is organized as follows: In Section 6.1, the practical subband coder is introduced, and it is explained how the coding of the subband signals is performed. Section 6.2 contains distortion rate results which show the mismatch between the theoretical and practical performances. The reasons for the mismatch are analyzed in Section 6.3, and finally, a short summary is given in Section 6.4.

6.1 Practical Coding System

In this section, the structure of the practical coding system is explained.

Figure 1.1 shows the theoretical model which has been used to model the subband coder. In a practical coder, the analysis and synthesis filter banks are

equal to the FIR filters found in Section 4.1, but the coding of the subband signals have to be performed by a practical coding system. The performance of the practical coder will be evaluated by calculating average values by Monte Carlo simulations instead of finding theoretical expected values, as was done earlier in this dissertation.

In a practical coder, the coding of the subbands, which is represented by additive noise $q_i(n)$ in Figure 1.1, has to be performed by a practical source coding system. Midtread uniform threshold quantizers are used, and it is assumed that the quantizers have an infinite number of representation levels. The decision levels in uniform threshold quantizer number i are given by $\frac{2k+1}{2}\Delta_i$, where k is an integer and Δ_i is the quantizer step size in quantizer number i. The representation levels in a uniform threshold quantizer could be for example the midpoints of the decision intervals, the centroids of the pdf in the decision intervals, or some other function of the decision interval and the pdf of the subband signal. If the centroids are used as representation levels in the uniform threshold quantizers, these quantizers are close to the optimal entropy constrained scalar quantization [Farvardin & Modestino 1984, Sullivan 1996], for all rates and for a wide variety of memoryless sources. Entropy constrained scalar quantizers have a good distortion rate performance compared to other methods [Fischer & Wang 1992], and since uniform threshold quantizers are simple to implement and analyze, they will be used in this work. In this chapter, centroids will be used as representation levels in the uniform threshold quantizers.

It is also possible to use pdf optimized scalar quantizers with bit allocation in the coding of the subband signals. In the theoretical model for the coding of the subband signals, it is assumed that any number of bits could be used. However, the practical performance results obtained by pdf optimized scalar quantizers would not be very close to the theoretical results, because not every bit rate in each quantizer is possible with pdf optimized scalar quantizers. If L levels are used in a fix rate scalar quantizer, the number of bits used by the quantizer are $\log_2 L$, and for low values of L, this is a small subset of all real values. Therefore, a mismatch exists between the assumption of using any rate in each subband and the number of bits used in the practical coder for each subband. Another reason for not using pdf optimized scalar quantizers with bit allocation is that the distortion rate performance is worse than for entropy constrained uniform scalar quantizers [Farvardin & Modestino 1984].

The contribution that is given in this dissertation is mainly the filter bank part of the coder. Therefore, to simplify the implementation of the practical subband coder, the entropy coders are assumed to be ideal entropy coders. This means that the entropy coders are not actually implemented, and the

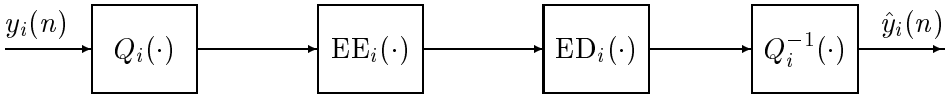

Figure 6.1 Subband coding of subband number i.

performance of the entropy coders is estimated by calculating the entropy of the quantization indices. Even though the subband signals are not uncorrelated, the zeroth order entropy is used as an estimate of the rate used when coding subband number i:

$$b_i = -\sum_{k=-\infty}^{\infty} P_k^{(i)} \log_2 P_k^{(i)}, \tag{6.1}$$

where $P_k^{(i)}$ is the estimated probability of index number k to occur in the ith uniform threshold quantizer in the practical subband coder.

The filter bank tries to remove correlation between different subbands, but inside one subband the filter bank approximately half-whitens the signal, see Subsection 2.1.3. Therefore, some correlation is left within the subband signals. If this correlation were utilized, better coding performance could have been achieved.

In a practical coder, the entropy coder could for instance be an arithmetic coder. In [Popat 1990], it is shown that the performance achieved by an arithmetic coder is very close to the zeroth order entropy of the quantization indices. Therefore, the practical results found here should be very close to the results found in a realization of a complete coder. The entropy is a lower bound for the rate in the quantizer if the source is memoryless [Blahut 1987].

The coding of subband signal number i is shown in Figure 6.1, and this figure replaces the additive noise in Figure 1.1. Figure 6.1 shows how the subband signal $y_i(n)$ is first quantized by the quantizer $Q_i(\cdot)$, and that the quantization indices are returned. These indices are sent into the ideal entropy encoder $\mathrm{EE}_i(\cdot)$. The output of the entropy encoder is the compressed signal. At the decoder, the compressed signal is first sent to the ideal entropy decoder $\mathrm{ED}_i(\cdot)$. The output of the entropy decoder is equal to the quantization indices, since the total entropy encoding/decoding process is lossless. The quantization indices are fed into the inverse quantization operator $Q_i^{-1}(\cdot)$, which outputs a representation level depending on the input index.

The midtread uniform threshold quantizer characteristics is shown in Figure 6.2 for quantizer number i. It is assumed that the quantizers have an infinite dynamic range. For practical purposes, this means that the signal is

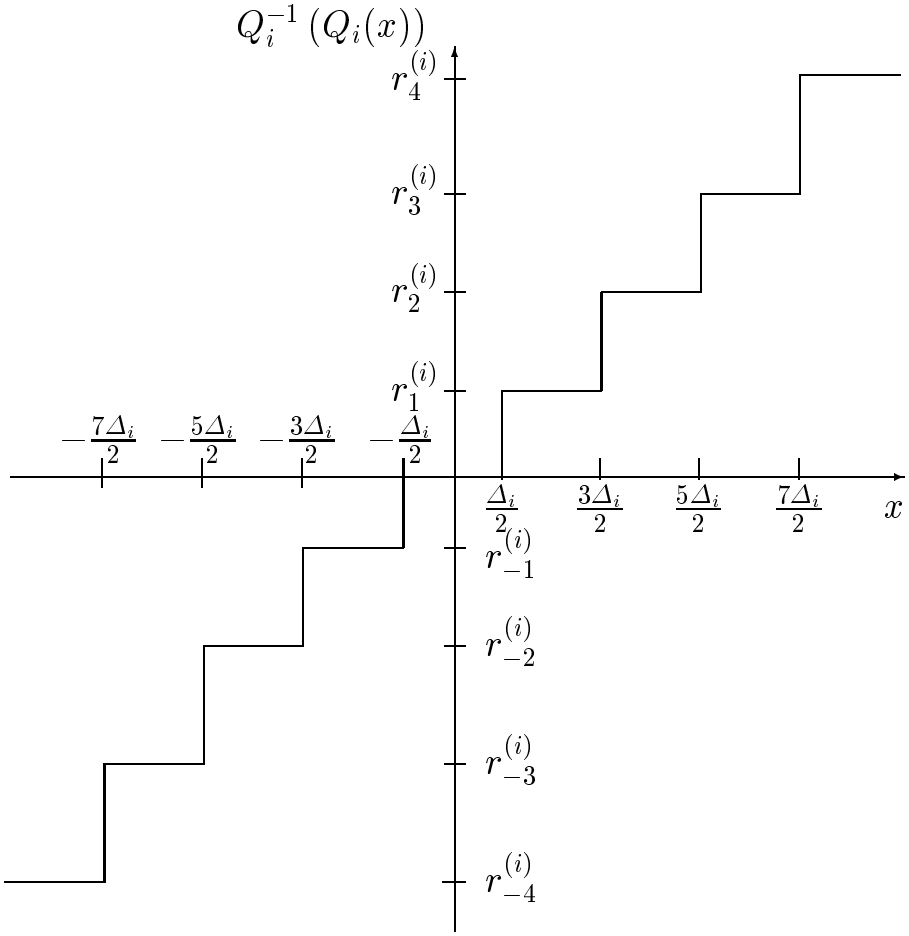

Figure 6.2 Characteristics of midtread uniform threshold quantizer number i. The quantizer step size is Δ_i and the number of representation levels are infinite.

never clipped by saturation of the quantizer. From Figure 6.2, it is seen that the *distance* between the decision levels are constant, and equal to Δ_i. The decision levels are given by $\frac{2k+1}{2}\Delta_i$, where $k \in \mathbb{Z}$.

In this chapter, the representation levels of the uniform threshold quantizer are chosen to be equal to the centroids of the pdf in the decision intervals. These values of the representation levels give the minimum MSE in each of the scalar quantizers [Gersho & Gray 1992]. The kth representation level in the

ith uniform threshold quantizer is given by

$$r_k^{(i)} = \frac{\int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i} y f_{y_i}(y)dy}{\int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i} f_{y_i}(y)dy}, \tag{6.2}$$

where f_{y_i} is the pdf of subband signal number i, which is assumed to be Gaussian with variance $\sigma_{y_i}^2$ and zero mean.

In a coder operating on real world signals, the statistics of the source is varying, and an adaptive system is preferable. This can be done by estimating statistical properties for segments of the signal and then designing the filter banks according to the estimated statistics. The filter coefficients of the synthesis filter bank or the signal statistics can be transmitted as side information. The coding of the subband signals can also be implemented in an adaptive fashion, by modeling the subband signals as infinite Gaussian mixture distributions. The subband samples can then be classified into a finite number of classes according to their estimated variance and coded by a uniform threshold quantizer with centroid representation levels and an entropy coder optimized for the class. Details on this method can be found in [Hjørungnes, Lervik & Ramstad 1996, Hjørungnes & Lervik 1997]. However, in the practical coder which is studied, the input signal is assumed to be stationary, and the statistics of the input signal are assumed to be known exactly. These assumptions are made in order to simplify the analysis of the differences between the theoretical and practical subband coder.

6.2 Comparison of Theoretical and Practical Results

Figure 6.3 shows the theoretical performance of the optimized 5_3 filter banks from Section 4.1. $N = 2$ channels are used. The input signal to the filter bank is a Gaussian AR(1) source with correlation coefficient 0.95. The performance of the practical source coder, using the same optimized filter banks and the coding system described in Section 6.1, is also shown in the figure. The distortion rate function is found from [Berger 1971].

In the simulations, time series of 300 000 samples are used. The length of the 95 % confidence interval [Hines & Montgomery 1990] for the MSE per source sample is less than 0.05 dB.

From Figure 6.3, it is seen that there is a mismatch between the performance of the practical coder and the theoretical coder. For high rates, the two curves are not far from each other, because a high rate model is used to

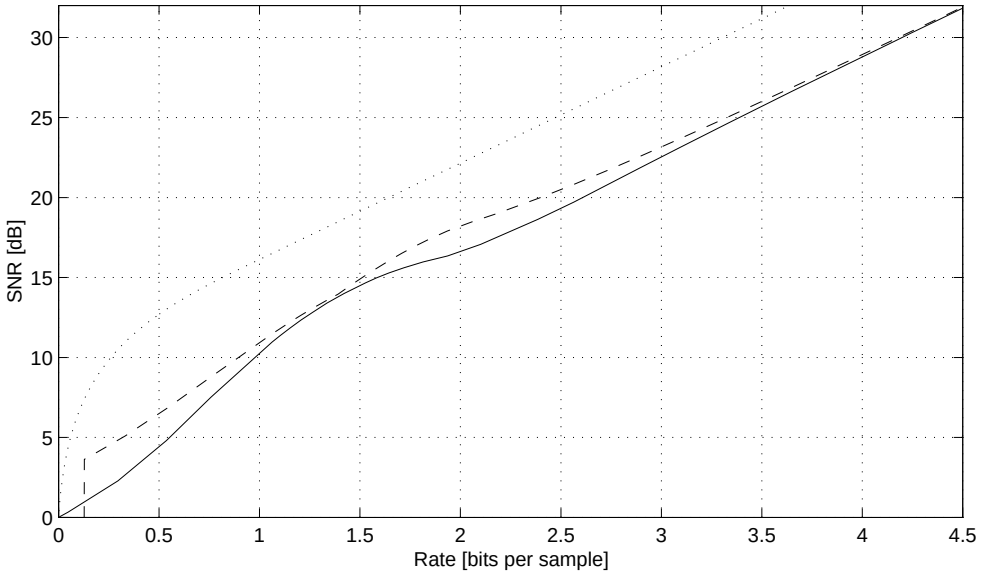

Figure 6.3 The distortion rate performance of the optimized 5_3 filter banks from Section 4.1. The performance of the practical coder is shown by the solid curve, while the dashed curve shows the theoretical performance found from the theory developed in Section 4.1. The dotted curve gives the distortion rate function for the input signal, which is a Gaussian AR(1) signal with correlation factor 0.95. $N = 2$ subbands are used.

model the coding of the subband signals. The curves are also close for bit rates around 1.3 bits per sample. In this region, only one quantizer receives bits, and the number of bits allocated to this quantizer is relatively high.

Figure 6.4 shows the mismatch between the practical performance and the performance predicted by the theory in Section 4.1 using the 9_7 PR filter bank proposed in [Antonini et al. 1992]. From the figure, it is seen that the theoretical and practical performances match much better for PR filter banks than for the optimized non-PR filter banks, like the filter banks that was used in Figure 6.3. For very low rates, there is a mismatch for PR filter banks as well, because the quantizer model is inaccurate for low bit rates. This is also the case around 2.0 bits per sample, since very few bits are used for the subband with the lowest variance.

In the next section, the reasons for the larger mismatch between practical and theoretical results using non-PR filter banks compared to PR filter banks are analyzed. In Chapter 7, an improved quantizer model is developed and

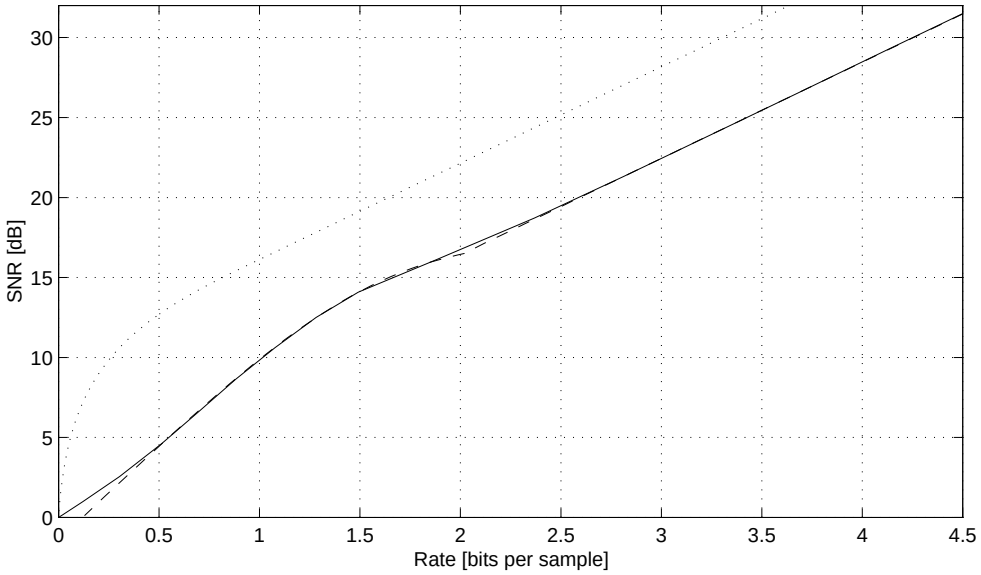

Figure 6.4 The distortion rate performance when using the 9_7 PR filter bank in [Antonini et al. 1992]. The performance of the practical coder is shown by the solid curve, while the dashed curve shows the theoretical performance given by the theory developed in Section 4.1. The dotted curve gives the distortion rate function for the input signal, which is a Gaussian AR(1) signal with correlation factor 0.95. $N = 2$ subbands are used.

used in a practical coder.

6.3 Analysis of the Reasons for the Mismatch

The performance is measured in distortion versus rate. Therefore, in this section, the estimation of both the rate and the distortion is compared in the practical and theoretical cases.

6.3.1 Bit Rate Estimation

In this subsection, the validity of the quantization model of Equation (1.13) is checked.

In the theoretical subband coding model in Equation (1.13), the coding coefficient c_i has to be specified. Since the entropy coding of Gaussian signals are used, this coefficient is chosen to $c_i = \frac{\pi e}{6}$ [Jayant & Noll 1984, Ramstad

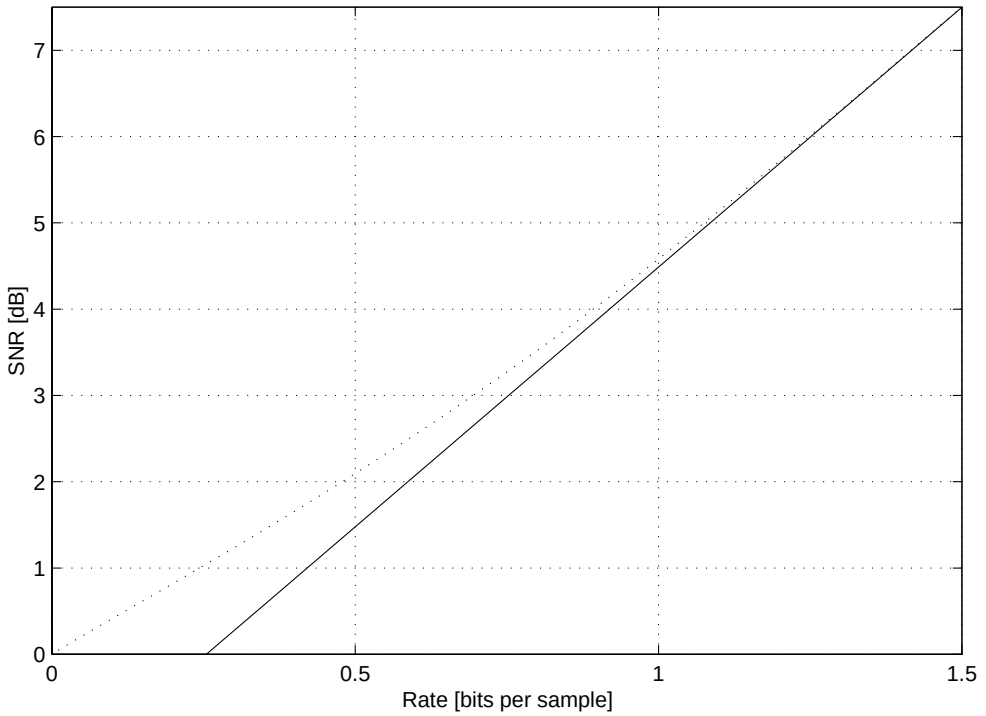

Figure 6.5 The solid curve shows the distortion rate performance of the theoretical quantization model when coding one subband signal, and the dotted curve shows the practical coding performance when coding the same Gaussian subband signal. The coding coefficient is $\frac{\pi e}{6}$ in the theoretical model.

et al. 1995].

Figure 6.5 shows the distortion rate performance of the quantization model in Equation (1.13) and the performance of the practical coder using a scalar uniform threshold quantizer and an ideal entropy coding of the quantization indices, see Figure 6.1. It is seen that the match between the quantization model and the practical coding is very good for high bit rates, but poor for low bit rates. The main reason for the deviation at low rates is that the coding coefficient c_i is the same for all rates. From Equation (1.13), it is seen that for low rates, the SNR in dB for the theoretical model is negative, but if c_i had approached 1 from above as the rate decreases, the SNR would have stayed positive for all rates. In the next chapter, the rate in the practical subband coder will be estimated in such a manner that the coding coefficients c_i are rate dependent.

6.3.2 Block MSE Estimation

Let the *signal block MSE* $\mathcal{E}_{N,M}^{(\boldsymbol{x})}(d_v, d_s)$ be defined as the block MSE that is achieved in the absence of quantization noise. Using the theory introduced in Section 4.1, it can be seen from Equation (4.12) that the theoretical signal block MSE is given by:

$$\begin{aligned}\mathcal{E}_{N,M}^{(\boldsymbol{x})}(d_v, d_s) = \mathrm{Tr}\Big\{ & \boldsymbol{R}_{\_}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{\_}^{H} - \boldsymbol{R}_{\_}\boldsymbol{E}_{\Gamma}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s) \\ & - \Big(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s)\Big)^{H}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{\_}^{H} + \boldsymbol{\Phi}_{\boldsymbol{x}}^{(0,N)}\Big\}\end{aligned} \tag{6.3}$$

The *quantization block MSE* $\mathcal{E}_{N,M}^{(\boldsymbol{q})}(d_v, d_s)$ is the block MSE that is due to the additive quantization noise. By using the quantization model introduced in Section 4.1, it is seen from Equation (4.12) that the theoretical quantization block MSE can be expressed as

$$\mathcal{E}_{N,M}^{(\boldsymbol{q})}(d_v, d_s) = \mathrm{Tr}\Big\{\boldsymbol{R}_{\_}\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}\boldsymbol{R}_{\_}^{H}\Big\}. \tag{6.4}$$

The part of the MSE that exists because of cross-correlation between the input signal and the additive quantization noise will here be denoted the *crossterm block MSE contribution* $\mathcal{E}_{N,M}^{(\boldsymbol{x},\boldsymbol{q})}(d_v, d_s)$. For the white signal independent quantization noise model used in Section 4.1, this is given by

$$\mathcal{E}_{N,M}^{(\boldsymbol{x},\boldsymbol{q})}(d_v, d_s) = \boldsymbol{0}, \tag{6.5}$$

since the input to the filter bank and the additive quantization noise are assumed to be uncorrelated with zero mean.

In the practical system the total, signal, quantization, and crossterm MSE contributions have to be estimated. Below, it will be shown how the estimation of these MSE contributions can be obtained in the practical coding system.

The output signal from the filter bank is given by the sum of the contributions from the signal and the quantization noise. Mathematically this can be expressed as

$$\hat{x}(n) = \hat{x}_{\mathrm{sig}}(n) + \hat{x}_{\mathrm{quant}}(n), \tag{6.6}$$

where $\hat{x}_{\mathrm{sig}}(n)$ is the output of the synthesis filter bank in the absence of quantizers, and $\hat{x}_{\mathrm{quant}}(n)$ is the output of the filter bank that is caused by the additive quantization noise.

The total block MSE in the practical coder can be estimated as:

$$\begin{aligned}\mathcal{E}_{N,M}(d_v,d_s) &= \frac{N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}|\hat{x}(n)-x(n-(d_vN+d_s+N-1))|^2\\ &= \frac{N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}|\hat{x}_{\text{sig}}(n)-x(n-(d_vN+d_s+N-1))|^2\\ &+ \frac{N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}|\hat{x}_{\text{quant}}(n)|^2\\ &+ \frac{2N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}\text{Re}\left\{\left(\hat{x}_{\text{sig}}(n)-x(n-(d_vN+d_s+N-1))\right)\hat{x}^*_{\text{quant}}(n)\right\},\end{aligned} \tag{6.7}$$

where the index set $\mathcal{I}$ is chosen according to the length of the input signal and the delay of the practical subband coder. $|\mathcal{I}|$ denotes the cardinality [Truss 1991] of the index set $\mathcal{I}$, Re is the real part of the argument, and the superscript * denotes complex conjugation.

The signal block MSE in the practical system is given by the MSE between the input and output of the filter banks when the quantization noise is equal to zero. The signal block MSE can be estimated as

$$\mathcal{E}^{(\boldsymbol{x})}_{N,M}(d_v,d_s) = \frac{N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}|\hat{x}_{\text{sig}}(n)-x(n-(d_vN+d_s+N-1))|^2. \tag{6.8}$$

The quantization block MSE can be estimated as

$$\mathcal{E}^{(\boldsymbol{q})}_{N,M}(d_v,d_s) = \frac{N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}|\hat{x}_{\text{quant}}(n)|^2, \tag{6.9}$$

while the crossterm block MSE contribution can be found by

$$\mathcal{E}^{(\boldsymbol{x},\boldsymbol{q})}_{N,M}(d_v,d_s) = \frac{2N}{|\mathcal{I}|}\sum_{n\in\mathcal{I}}\text{Re}\left\{\left(\hat{x}_{\text{sig}}(n)-x(n-(d_vN+d_s+N-1))\right)\hat{x}^*_{\text{quant}}(n)\right\}. \tag{6.10}$$

By comparing all the terms in Equations (6.8), (6.9), and (6.10) to Equation (6.7), it is seen that

$$\mathcal{E}_{N,M}(d_v,d_s) = \mathcal{E}^{(\boldsymbol{x})}_{N,M}(d_v,d_s) + \mathcal{E}^{(\boldsymbol{q})}_{N,M}(d_v,d_s) + \mathcal{E}^{(\boldsymbol{x},\boldsymbol{q})}_{N,M}(d_v,d_s). \tag{6.11}$$

Figure 6.6 shows the total, signal, quantization, and crossterm MSE contributions for both the practical and the theoretical coding system when coding a

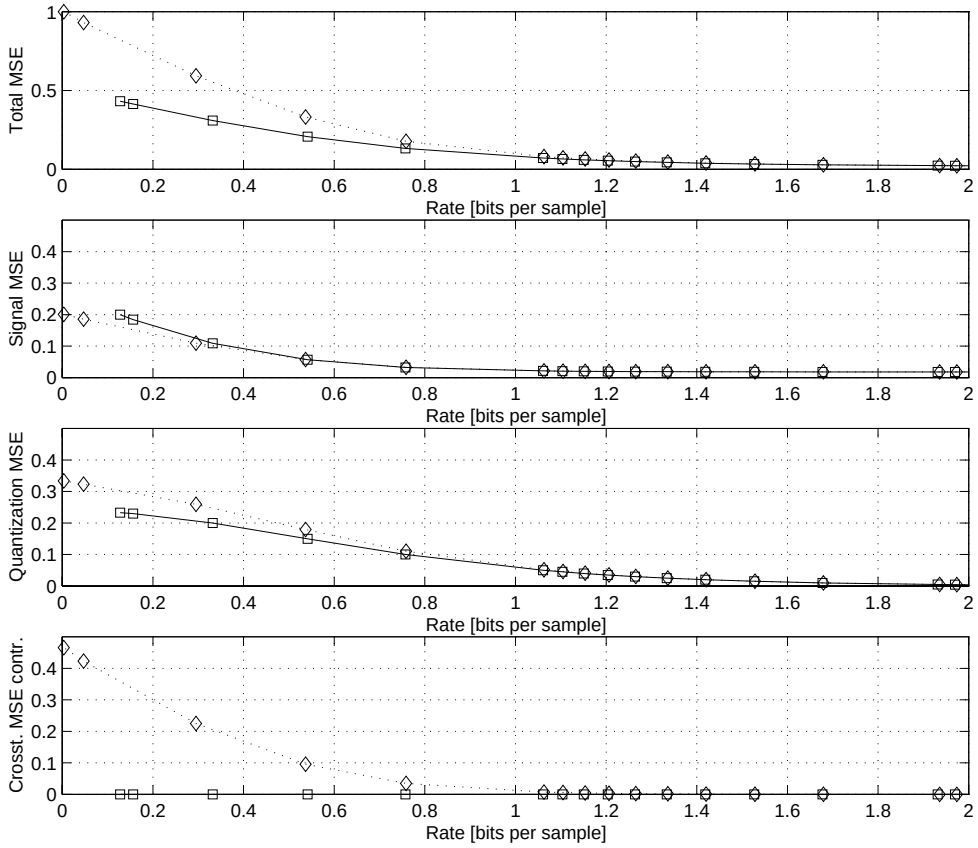

Figure 6.6 Different MSE contributions per source sample as a function of coding rate with $N = 2$ and $M = 1$. The optimized 5_3 filter banks from Section 4.1 are used. The dotted curves with diamonds show the MSE contributions for the practical coding system, and the solid curves with squares show the MSE contributions for the theoretical coding system. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95.

unit variance Gaussian AR(1) model with correlation coefficient 0.95, using the 5_3 filter bank optimized according to the theory developed in Section 4.1. In the figure, the Lagrange multipliers μ are the same in the theoretical and the practical results. The results obtained with the same Lagrange multiplier μ can be found by counting the squares and diamonds starting from high rates. The same procedure can be used in all the practical results presented in this dissertation. From Figure 6.6, it is seen that for high rates, the theoretical and practical results correspond well in both the MSE contributions and rate, but

this correspondence becomes worse as the rate decreases.

From the figure, it is seen that the main reason for the difference in the total MSE in the practical and theoretical case is the non-zero crossterm MSE contribution in the practical coder. In the theoretical case, this crossterm MSE contribution is zero, see Equation (6.5). From the size of the crossterm MSE contribution, it is seen that for low rates, the signal independent noise assumption does not hold. In Chapter 7, different ways of improving the model of the subband signal coding are proposed.

For a PR filter bank, $\hat{x}_{\text{sig}}(n) = x(n-(d_v N + d_s + N - 1))$, so the crossterm MSE contribution is equal to zero in the practical case, see Equation (6.10), even though the input time series and the additive quantization noise are correlated. Therefore, the white signal independent quantization model is a more suitable model for PR filter bank coders than for non-PR filter bank coders.

For low rates, it is also observed from the figure that there is a mismatch in the quantization MSE. The reason is that the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ is not diagonal in the practical coding system, because for low rates, very coarse quantization is used. This means that the quantization noise is approximately equal to the subband signal with the opposite sign. Since the subband signals are correlated, the quantization signal will be correlated too, and therefore, the off-diagonal terms in the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ are non-zero in the practical system. In Chapter 7, a new model for the quantization noise will be introduced, which includes this correlation and all other correlations that exist in the subband coder.

From Figure 6.6, it is observed that there is a mismatch in signal MSE curves. The size of the signal MSE is the same for the same Lagrage multiplier, but the rate is different. The reason for this is that the coding coefficients c_i are assumed to be constant in the theory. In Chapter 7, this will be treated more detailed. For a given Lagrange multiplier, both the theoretical and practical systems give the same signal MSE, it is the rate estimate that differs.

6.4 Summary

In this chapter, a mismatch between the theoretical and practical performance was found. One reason for the mismatch is that, in the theoretical quantization model, the coding coefficient is assumed to be the same for all rates. This can be improved by using a theoretical model in which the entropy of the quantization indices of a quantized Gaussian subband signal is used as an estimate of the bit rate used for coding the subband signal, see Chapter 7.

In the theoretical model, the quantization noise is assumed to be white. This is not correct if very low bit rates are employed. In Chapter 7, a signal

dependent colored quantization noise model will be introduced.

The most important reason for the mismatch is that the input signal and the additive quantization noise signal are assumed to be white and signal independent. In Chapter 7, improved models for the coding of the subband signals are proposed.

Chapter 7

Improvements of the Correspondence between Theory and Practice

In Chapter 6, it was shown that there exists a crossterm MSE contribution in a practical coder due to the cross-correlation between the subband signals and the additive quantization noise. In this chapter, three more accurate models of the coding of the subbands are proposed to account for this effect and thus obtain better correspondence between the theoretical model and the simulation results.

The first way to avoid the crossterm MSE contribution is to design scalar quantizers such that the input and the quantization noise are uncorrelated. The second method is to apply a subtractive dithering technique to make the additive coding noise uncorrelated to the subband signals. Both these methods match the assumption made in Section 4.1, which stated that the subband signals should be uncorrelated to the additive quantization noise generated by the coding of the subband signals. Both methods will be implemented and tested in a practical coder, where the filter banks optimized by the theory of Section 4.1 are used.

In the third method, the cross-correlation term between the input time series and the quantization noise and the correlation that exists within the quantization noise will be included in the theory for optimizing the filter banks. It will be shown that this more advanced signal dependent colored quantization noise model gives a good correspondence between performance results in practical simulations and theory. Furthermore, this improves the practical results.

In this chapter, the rate is theoretically estimated as the entropy obtained by a uniform threshold quantizer operating on a Gaussian time series. In the

theoretical model, the rate will be found by

$$b_i = -\sum_{k=-\infty}^{\infty} \int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i} f_{y_i}(y)dy \cdot \log_2 \left(\int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i} f_{y_i}(y)dy \right), \qquad (7.1)$$

where $f_{y_i}(y)$ is the Gaussian pdf of the subband samples in the ith subband, having zero mean and variance $\sigma_{y_i}^2$. By using this estimate of the rate in the theoretical model, it will be seen later that the theoretical results match the practical simulation results better than by the white signal independent noise model estimate used so far, see Equation (1.14). When subtractive dithering is used, the input pdf to the quantizer will change, and this must be taken into consideration when estimating the rate by Equation (7.1).

This chapter is organized as follows: In Section 7.1, two systems having uncorrelated subband signals and additive coding noise will be treated, thus, the coding method of the subband signals will be changed and the filter bank optimization theory developed in Section 4.1 is used. In Section 7.2, a signal dependent colored quantization noise model is introduced, and this model is used to find new equations for optimizing the filter banks. The equation for the FIR Wiener polyphase matrix is also derived in Section 7.2. Conditions for optimality of an FIR PR filter bank are derived in Section 7.3. Section 7.4 contains both practical and theoretical results achieved by the methods proposed in this chapter. Also, a comparison to results obtained by filter banks found in the literature will be given. Finally, a summary is given in Section 7.5.

7.1 Uncorrelated Subband Signals and Coding Noise

In this section, two coding systems where the subband signals and the additive coding noise are uncorrelated, will be studied. In the first system, which is presented in Subsection 7.1.1, the representation levels of the quantizer will be found such that the input to the quantizer and the additive quantization noise are uncorrelated. Subtractive dithering is described in Subsection 7.1.2 as another method where the subband signals and the coding noise are uncorrelated. The filter banks used with these two systems are the filter banks found with the theory developed in Section 4.1, except that the constraints in Equation (2.4) do not need to be satisfied with the two systems proposed in this section. The reason for this is that with the two coding methods in this section, the variance of the additive coding noise can be larger than the subband variance.

7.1.1 Redesigned Scalar Quantizers

In this subsection, scalar quantizers which have uncorrelated input and quantization noise will be designed. These quantizers will be used in the coding of the subband samples in the subband coder. The decision levels in quantizer number i is assumed to be given by $\frac{2k+1}{2}\Delta_i$, where k is an integer and Δ_i is the quantizer step size in quantizer number i.

The main reason why the theoretical and practical results in the previous chapter have a mismatch is that, in the theory, it is assumed that the additive quantization noise is uncorrelated with the quantizer input. However, by using centroids as representation levels in each of the scalar quantizers, the correlation of the quantizer input and the additive quantization noise is given by:

$$E\left[y_i(n)q_i(n)\right] = -\sigma_{q_i}^2, \tag{7.2}$$

since $E\left[\hat{y}_i(n)q_i(n)\right] = 0$ when centroids are used as representation levels [Gersho & Gray 1992].

In the following, the decision levels will be kept unchanged, that is, a uniform threshold quantizer is still used. However, the theory presented is straightforward to generalize for arbitrary decision levels.

It is assumed that the new representation levels $\hat{r}_k^{(i)}$ will be odd functions of the index k. This is justified by the assumptions that the pdf of the quantizer input signal is an even function, that is, $f_{y_i}(-y) = f_{y_i}(y)$, and that the decision levels are symmetric about zero. It can be verified that the condition $E\left[y_i(n)q_i(n)\right] = 0$ is equivalent to

$$\sum_{k=1}^{\infty} \hat{r}_k^{(i)} \int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i} y f_{y_i}(y)dy = \frac{\sigma_{y_i}^2}{2}. \tag{7.3}$$

If the condition in Equation (7.3) is satisfied, it can be shown that minimizing the MSE through the quantizer is equivalent to minimizing the following expression:

$$\sum_{k=1}^{\infty} \left(\hat{r}_k^{(i)}\right)^2 \int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i} f_{y_i}(y)dy. \tag{7.4}$$

The optimization problem is then to minimize the expression in Equation (7.4), subject to the constraint in Equation (7.3). By using the Lagrange multiplier method [Edwards & Penney 1986], the optimal representation levels

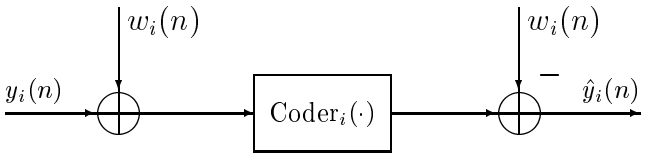

Figure 7.1 Subtractive dithering in subband number i.

can be derived as

$$\hat{r}_k^{(i)} = \frac{\sigma_{y_i}^2}{2}\left[\sum_{p=1}^{\infty}\frac{\left(\int_{\frac{2p-1}{2}\Delta_i}^{\frac{2p+1}{2}\Delta_i} y f_{y_i}(y)dy\right)^2}{\int_{\frac{2p-1}{2}\Delta_i}^{\frac{2p+1}{2}\Delta_i} f_{y_i}(y)dy}\right]^{-1} r_k^{(i)}, \tag{7.5}$$

where $r_k^{(i)}$ is the centroid representation level given by Equation (6.2). Equation (7.5) shows that the new representation levels are given by scaled versions of the centroid representation levels. Ideal entropy coding is also used in connection with the redesigned scalar quantizers introduced in this subsection. It can be shown that $\hat{r}_k^{(i)} = r_k^{(i)}$ only if infinitely high rates are used. Therefore, this is not possible in a practical source coder using a finite rate.

By using that $\sigma_{q_i}^2 \geq 0$ for a quantizer that is using centroids as representation levels, it is possible to show that the scaling factors in Equation (7.5) are greater or equal to 1, i.e.,

$$\frac{\sigma_{y_i}^2}{2}\left[\sum_{p=1}^{\infty}\frac{\left(\int_{\frac{2p-1}{2}\Delta_i}^{\frac{2p+1}{2}\Delta_i} y f_{y_i}(y)dy\right)^2}{\int_{\frac{2p-1}{2}\Delta_i}^{\frac{2p+1}{2}\Delta_i} f_{y_i}(y)dy}\right]^{-1} \geq 1, \tag{7.6}$$

for all $i \in \{0, 1, \ldots, M-1\}$.

7.1.2 Subtractive Dithering

An excellent survey on quantization and subtractive dithering is published in [Lipshitz, Wannamaker & Vanderkooy 1992]. When using subtractive dithering, the subbands $y_i(n)$ are uncorrelated with the additive coding noise $\hat{y}_i(n) - y_i(n)$ [Jayant & Noll 1984]. In this subsection, subtractive dithering will be introduced as a coding method of the subband signals.

Figure 7.1 shows the coding of the ith subband signal when using subtractive dithering. In the figure, the block denoted $\text{Coder}_i(\cdot)$ is given by Figure 6.1, which shows a uniform threshold quantizer with step size Δ_i followed by an ideal entropy encoder/decoder operating on the quantizer indices and an inverse quantizer. In Figure 7.1, the signal $w_i(n)$ is a pseudo-random uniformly distributed sequence within the interval $\left(-\frac{\Delta_i}{2}, \frac{\Delta_i}{2}\right)$. Note that it is the *same* pseudo-random sequence that is added and subtracted before and after the $\text{Coder}_i(\cdot)$ block in Figure 7.1. The signal $w_i(n)$ is called dither [Goodall 1951, Lipshitz et al. 1992], and it is statistically independent of the subband signal $y_i(n)$.

Since the subband signal $y_i(n)$ and the dither signal $w_i(n)$ are statistically independent, the pdf of the input signal to the $\text{Coder}_i(\cdot)$ block is given by the convolution of the pdf's of the signals $y_i(n)$ and $w_i(n)$ [Papoulis 1991]. The signal $y_i(n)$ is Gaussian with zero mean and variance $\sigma_{y_i}^2$, while $w_i(n)$ is uniformly distributed over the interval $\left(-\frac{\Delta_i}{2}, \frac{\Delta_i}{2}\right)$. If the subband signal is to be uncorrelated with the additive coding noise introduced by the subtractive dithering, the midpoint must be used as the representation level in the uniform threshold quantizer [Lipshitz et al. 1992]. This means that $r_k^{(i)} = k\Delta_i$ when using subtractive dithering.

From Figure 7.1, it is seen that the noise generated by subtractive dithering is equal to the noise generated by the middle block in the figure, because the same signal $w_i(n)$ is subtracted and added before and after the middle block. Therefore, the performance of the coder for the ith subband, shown in Figure 7.1, is given by Equation (D.10), using the correct values for the quantizer input pdf and the representation levels in the uniform threshold quantizer.

7.2 Signal Dependent Colored Quantization Noise Model

In this section, a signal dependent colored quantization noise model will be introduced. Until now, it has been assumed that the quantizer noise is white and uncorrelated with the subband signals. Formulas for finding the correlation that exists between quantization noise samples and between the input time series and the quantization noise are quite complex, and to improve the readability of this chapter, these formulas are derived in Appendix D.

If Lloyd-Max quantizers are used, the quantizer model can be improved by assuming that it not only adds noise to the signal, but also modifies the strength of the signal [Park & Haddad 1993], such that the output of the

quantizer is uncorrelated with the additive quantization noise. More accurate results can be obtained by using the true performance curve of the coding method.

This section is organized as follows: In Subsection 7.2.1, the problem is formulated when the cross-correlation between the input time series and the quantization noise is assumed to be known. Equations for optimality are derived in Subsection 7.2.2. The Wiener filter is formulated for FIR filter banks in Subsection 7.2.3 since this will be used later to derive the conditions for optimality of PR in the FIR filter bank case. In Subsection 7.2.4, a numerical optimization algorithm is proposed, describing the optimization of the FIR filter banks when using the signal dependent colored quantization noise model.

7.2.1 Problem Formulation

The $(m+l+1)N \times (l+1)M$ matrix

$$\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)} = E\left[\boldsymbol{x}(n)_{\mid}\boldsymbol{q}^H(n)_{\mid}\right], \tag{7.7}$$

and the $(l+1)M \times N$ matrix

$$\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v,d_s) = E\left[\boldsymbol{q}(n)_{\mid}\boldsymbol{x}_{d_s}^H(n-d_v)\right], \tag{7.8}$$

are used to include the cross-correlation between the input signal and the additive quantization noise signal. The vector $\boldsymbol{x}_{d_s}(n)$ is defined in Equation (4.5). The cross-correlation matrix $\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v,d_s)$ can be found from the matrix $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$, because the matrix $\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v,d_s)$ is equal to the Hermitian of the block matrix of $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ consisting of row number d_s+d_vN through row number d_s+d_vN+N-1. The numbering of the rows and columns starts at zero. In Appendix D, it is shown how the elements of the matrices in Equations (7.7) and (7.8) can be found. From the definition in Equation (7.7), it can be seen that the matrix $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ is a block Toeplitz matrix [Lütkepohl 1996], but the matrix is in general *not* a Toeplitz matrix.

The block MSE, which includes the cross-correlation between the input signal and the additive quantization noise, is derived in Appendix A. If the correlation matrices defined in Equations (4.8) through (4.10) are substituted

into Equation (A.8), the following result is obtained:

$$\begin{aligned}\mathcal{E}_{N,M}(d_v,d_s) = \mathrm{Tr}\Big\{ & \boldsymbol{R}_{-}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{-}^{H} \\ & -\boldsymbol{R}_{-}\boldsymbol{E}_{\Gamma}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s) - \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^{H}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{-}^{H} \\ & +\boldsymbol{\Phi}_{\boldsymbol{x}}^{(0,N)} + \boldsymbol{R}_{-}\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}\boldsymbol{R}_{-}^{H} \\ & +\boldsymbol{R}_{-}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}\boldsymbol{R}_{-}^{H} + \boldsymbol{R}_{-}\left(\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}\right)^{H}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{-}^{H} \\ & -\boldsymbol{R}_{-}\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v,d_s) - \left(\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v,d_s)\right)^{H}\boldsymbol{R}_{-}^{H}\Big\}, \end{aligned} \quad (7.9)$$

where the cross-correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ defined in Equation (7.7) and the cross-correlation matrix $\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v,d_s)$ given in Equation (7.8) have been used. In Equation (7.9), the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ has dimensions $(l+1)M \times (l+1)M$, and it is a block Toeplitz matrix, which is not necessarily diagonal in this chapter. The matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ is not a Toeplitz matrix, but it is Hermitian. For low rates, it has non-zero off-diagonal terms. Formulas for the elements of the quantization autocovariance matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ are found in Appendix D.

The bit constraint is assumed to be the same as before, that is, it is still given by Equation (4.13). The right hand side of Equation (4.13) is assumed to be a constant in the problem formulation, meaning that c_i is treated as a constant in the optimization. This is not accurate in practice, but it is done to simplify the optimization. In the theoretical evaluation of the results, the estimation of the rate, see Equation (7.1), is not an explicit function of c_i, and this is equivalent to letting c_i be rate dependent. The coding coefficients c_i can be calculated for a Gaussian time series as follows: Let $\sigma_{y_i}^2$ be the variance of the Gaussian time series, and let Δ_i be the quantizer step size resulting in $\sigma_{q_i}^2 = 1$. Then, the values of Δ_i and $\sigma_{y_i}^2$ can be used to find the entropy of the quantization indices when quantizing the subband samples using a uniform threshold quantizer with step size Δ_i. The entropy is assumed to be equal to b_i, see Equation (7.1). If the values of $\sigma_{y_i}^2$, Δ_i, $\sigma_{q_i}^2$, and b_i are inserted into Equation (1.13), the value of c_i can be found. The coding coefficient is theoretically calculated this way, and plotted in Figure 7.2. The result in the figure is independent of any correlation within the Gaussian time series, because zeroth order entropy is used to estimate the theoretical bit rate in the quantizers.

From Figure 7.2, it is seen that, for high rates, the coding coefficients are almost independent of rate, but for rates below approximately 1.4 bits per sample this assumption does not hold. After optimization, the theoretical performance of the optimized system can be found. Then, it is assumed that

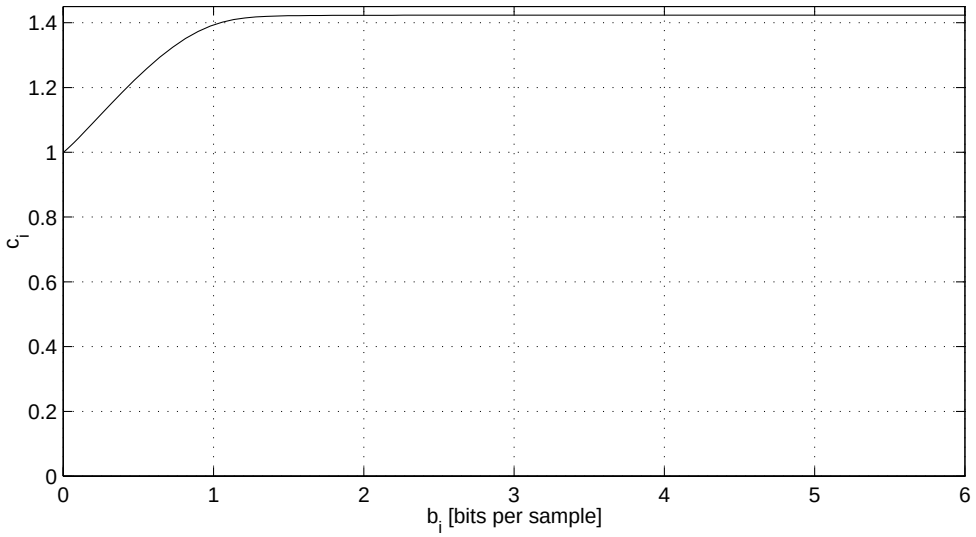

Figure 7.2 The coding coefficient c_i, as a function of rate b_i, when coding a Gaussian time series with a uniform threshold quantizer using centroids as representation levels.

the entropy of the quantization indices is used as an evaluation criterion. This means that in the theoretical evaluation of the system, the coding coefficients c_i are rate dependent.

The simplifying assumption described above might lead to results that are not optimal if the rate dependent coding coefficients had been taken into consideration. However, the results obtained by this simplification are very good, as will be shown later, and the theoretical and practical results match very well.

With the signal dependent colored quantization noise model, it is possible to classify the total block MSE in Equation (7.9) into signal, quantization, and crossterm MSE contributions, as was done in Subsection 6.3.2. The signal block MSE $\mathcal{E}_{N,M}^{(\boldsymbol{x})}(d_v, d_s)$ is unchanged, so it is again given by Equation (6.3), while the quantization block MSE $\mathcal{E}_{N,M}^{(\boldsymbol{q})}(d_v, d_s)$ is given by Equation (6.4), but now the quantization matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ has changed. The elements of this matrix are given by the autocorrelation function $R_{q_i(n),q_k}(m)$, given in Appendix D. The diagonal elements are given by Equation (D.10), while the off-diagonal elements are given by Equation (D.13).

With the signal dependent colored quantization noise model, the crossterm

block MSE contribution $\mathcal{E}_{N,M}^{(\boldsymbol{x},\boldsymbol{q})}(d_v, d_s)$ is in general non-zero, and given by:

$$\mathcal{E}_{N,M}^{(\boldsymbol{x},\boldsymbol{q})}(d_v, d_s) = \mathrm{Tr}\left\{ \boldsymbol{R}_{-}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}\boldsymbol{R}_{-}^{H} + \boldsymbol{R}_{-}\left(\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}\right)^{H}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{-}^{H} \right.$$
$$\left. -\boldsymbol{R}_{-}\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s) - \left(\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s)\right)^{H}\boldsymbol{R}_{-}^{H} \right\}. \quad (7.10)$$

The elements of the matrices $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ and $\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s)$ can be found by Equation (D.6).

It can also be shown theoretically that, in all quantizers using the centroids as representation levels, the variance of the additive quantization noise is always less than or equal to the variance of the subband signal. Therefore, when using the signal dependent colored quantization noise model, the constraints in Equation (2.4) are not needed any more because they will always be satisfied. The following choice has been made as before: $\sigma_{q_i}^2 = \sigma_q^2 = 1$ for all $i \in \{0, 1, \ldots, N-1\}$.

The problem that has to be solved when using the signal dependent colored quantization noise model can now be stated. The total block MSE in Equation (7.9) should be minimized with respect to the analysis and synthesis polyphase matrices $\boldsymbol{E}(z)$ and $\boldsymbol{R}(z)$, subject to the bit constraint in Equation (4.13).

7.2.1.1 PR Expressions when $N = M$

For a system having PR with scalar delay d_s, vector delay d_v, and where $N = M$, it can be shown that the following relations hold:

$$\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{-}^{H} = \boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s) \quad (7.11)$$

$$\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s)\right)^{H}\boldsymbol{E}_{\Gamma}^{H}\boldsymbol{R}_{-}^{H} = \boldsymbol{\Phi}_{\boldsymbol{x}}^{(0,N)} \quad (7.12)$$

$$\boldsymbol{R}_{-}\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,N)} = \left(\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,N)}(d_v, d_s)\right)^{H}. \quad (7.13)$$

Equations (7.11) and (7.12) can be used to show that for PR filter banks, the signal block MSE in Equation (6.3) is equal to zero, when $N = M$. The last relation above can be used to show that for PR filter banks, the crossterm block MSE contribution in Equation (7.10) is equal to zero, when $N = M$. This means that when using a PR filter bank, the total block MSE is given by

$$\mathcal{E}_{N,N}(d_v, d_s) = \mathrm{Tr}\left\{\boldsymbol{R}_{-}\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,N)}\boldsymbol{R}_{-}^{H}\right\}, \quad (7.14)$$

which is the same total PR block MSE obtained by the theory developed in Section 4.1, except that the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,N)}$ is *not* diagonal any more, see Appendix D.

7.2.2 Equations for Optimality

By means of the Lagrange multiplier method, the constrained optimization problem is converted to an unconstrained optimization problem. The unconstrained objective function for the signal dependent colored quantization noise model is given by:

$$\mathcal{E}_{N,M}\left(d_v, d_s\right) + \mu \sum_{k=0}^{M-1} \ln \sigma_{y_k}^2, \tag{7.15}$$

where μ is the Lagrange multiplier for the bit constraint in Equation (4.13).

Matrix differentiation of Equation (7.15) with respect to the matrix $\boldsymbol{E}_{-}$ gives the following equations for finding the optimal analysis filter bank $\boldsymbol{E}_{-}$ for a given synthesis filter bank $\boldsymbol{R}_{-}$

$$\begin{aligned}\boldsymbol{R}_{\mathsf{I}}^{H}\mathcal{T}\Bigg\{\boldsymbol{R}_{-}\left(\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)} + \left(\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}\right)^{H}\right) \\ - \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s)\right)^{H}\Bigg\} = -\mu\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{E}_{-}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)},\end{aligned} \tag{7.16}$$

where the matrix differentiation formulas found in Appendix C have been used. The operator $\mathcal{T}$ is defined in Equation (4.17).

Finding the equations for the optimal synthesis filter bank $\boldsymbol{R}_{-}$, for a given analysis filter bank $\boldsymbol{E}_{-}$, can be done by matrix differentiation of Equation (7.15) with respect to the matrix $\boldsymbol{R}_{-}$. The formulas found in Appendix C can be used to obtain the following equations

$$\begin{aligned}\boldsymbol{R}_{-} = &\left(\boldsymbol{E}_{\Gamma}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v, d_s) + \boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s)\right)^{H} \cdot \\ &\left[\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^{H} + \boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)} + \boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)} + \left(\boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}\right)^{H}\right]^{-1}.\end{aligned} \tag{7.17}$$

In Section 4.1, it was assumed that the additive quantization noise and the input subband signals were uncorrelated, and that the constraints in Equation (2.4) were imposed. With the white signal independent noise model, the correlation matrices in Equations (7.7) and (7.8) are equal to zero matrices, and the matrix $\boldsymbol{\Theta}$ contains the Kuhn-Tucker parameter θ_i for the inequality given in Equation (2.4). If these differences are taken into consideration, it is seen that the results from Equation (7.16) and (7.17) are consistent with those found in Equations (4.16) and (4.15), respectively.

7.2.3 FIR Wiener Synthesis Filter Bank

The FIR Wiener synthesis filter bank is used in the derivation of the conditions for optimality of PR FIR filter banks in Section 7.3, and is derived in this subsection.

Let the $N \times (l+1)M$ matrix $\boldsymbol{\phi}_{\boldsymbol{x},\hat{\boldsymbol{y}}}^{(l,N,M)}(d_v, d_s)$ be defined as

$$\boldsymbol{\phi}_{\boldsymbol{x},\hat{\boldsymbol{y}}}^{(l,N,M)}(d_v, d_s) = E\left[\boldsymbol{x}_{d_s}(n-d_v)\hat{\boldsymbol{y}}^H(n)_{\text{I}}\right], \tag{7.18}$$

and let the $(l+1)M \times (l+1)M$ matrix $\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,M)}$ be defined as

$$\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,M)} = E\left[\hat{\boldsymbol{y}}(n)_{\text{I}}\hat{\boldsymbol{y}}^H(n)_{\text{I}}\right]. \tag{7.19}$$

By calculating the matrices in Equations (7.18) and (7.19) using the signal dependent colored noise model, and comparing the result to Equation (7.17), it is seen that Equation (7.17), which gives the optimal synthesis polyphase matrix, can be written as

$$\boldsymbol{R}_{-} = \boldsymbol{\phi}_{\boldsymbol{x},\hat{\boldsymbol{y}}}^{(l,N,M)}(d_v, d_s)\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,M)}\right)^{-1}. \tag{7.20}$$

By means of the orthogonality principle [Therrien 1992], the FIR Wiener filter bank will now be derived. The error vector for the FIR filter bank system is given by $\hat{\boldsymbol{x}}(n) - \boldsymbol{x}_{d_s}(n-d_v)$. According to the orthogonality principle, the error vector has to be orthogonal to all the available observations at the input of the FIR synthesis filter bank. This can be expressed mathematically as

$$E\left[\left(\hat{\boldsymbol{x}}(n) - \boldsymbol{x}_{d_s}(n-d_v)\right)\hat{\boldsymbol{y}}^H(n-p)\right] = \boldsymbol{0}, \quad \forall p \in \{0, 1, \ldots, l\}. \tag{7.21}$$

By using that $\hat{\boldsymbol{x}}(n) = \sum_{k=0}^{l} \boldsymbol{r}(k)\hat{\boldsymbol{y}}(n-k)$, Equation (7.21) can be rewritten as

$$E\left[\boldsymbol{x}_{d_s}(n-d_v)\hat{\boldsymbol{y}}^H(n-p)\right] = \sum_{k=0}^{l} \boldsymbol{r}(k)\boldsymbol{K}_{\hat{\boldsymbol{y}}}(p-k), \quad \forall p \in \{0, 1, \ldots, l\}, \tag{7.22}$$

where the $M \times M$ matrix $\boldsymbol{K}_{\hat{\boldsymbol{y}}}(n)$ is defined as $\boldsymbol{K}_{\hat{\boldsymbol{y}}}(n) = E\left[\hat{\boldsymbol{y}}(n+k)\hat{\boldsymbol{y}}^H(k)\right]$. By putting the $l+1$ matrix equations in Equation (7.22) together, it is possible to rewrite these equations as the matrix equation in Equation (7.20). Therefore, the equation for the Wiener synthesis polyphase matrix is given by Equation (7.20). This shows that the FIR Wiener synthesis filter bank can be derived by both the orthogonality principle and by matrix differentiation.

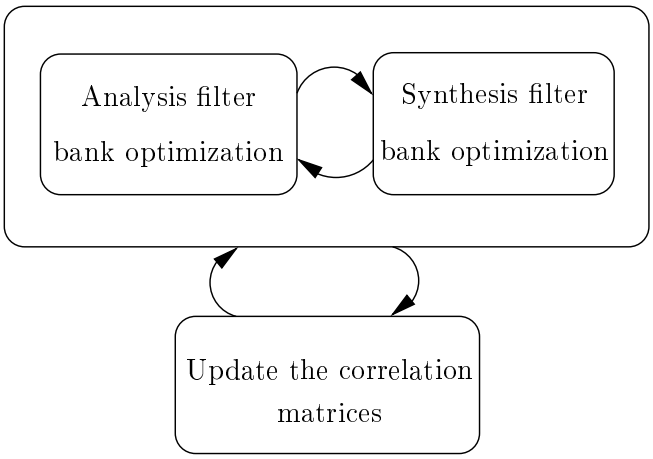

Figure 7.3 Illustration of the iterative numerical algorithm for optimizing the FIR filter banks when the cross-correlation between the input signal and the additive quantization noise is included.

When the optimal FIR synthesis filter bank is expressed as in Equation (7.20), it has a form related to the Wiener synthesis filter bank in the unconstrained case given in Equation (2.17) and the Wiener synthesis transform given in Equation (3.13). If arbitrary given filter lengths are used in the Wiener FIR synthesis filter bank, the procedure developed in Subsection 4.1.4 can be used.

By using the white signal independent quantization noise model used in Section 4.1 and calculating the Wiener polyphase matrix in Equation (7.20), it can be shown that the Wiener synthesis polyphase matrix in Equation (4.15) is found.

7.2.4 Numerical Optimization Algorithm

In Figure 7.3, the iterative numerical algorithm for optimization of the FIR filter bank is illustrated. As initial values, the filter banks found for the white signal independent noise model in Section 4.1 are used. Then the correlation matrices $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$, $\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s)$, and $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ are found by using the formulas in Appendix D. For the given values of the correlation matrices, Equations (7.16) and (7.17) are iteratively solved until convergence is reached. Then the correlation matrices are updated, until the whole algorithm stops when the correlation matrices have converged, see Figure 7.3.

7.3 Conditions for Optimality of PR in the FIR Case

In this section, the conditions for when PR FIR filter banks are optimal, for a given invertible FIR analysis polyphase filter bank, are derived by means of Wiener filter theory. This result is an extension of the result found in [Vaidyanathan & Chen 1994], where the same conditions were derived for unconstrained length filter banks and transforms.

In [Vaidyanathan & Chen 1994], the transform case and the unconstrained length filter bank case were treated for $d_v = d_s = 0$. The FIR case has to be treated in a slightly different manner, because in [Vaidyanathan & Chen 1994], it was assumed that the synthesis filter bank consists of the PR synthesis filter bank followed by a Wiener polyphase filter bank. Conditions for when the Wiener polyphase matrix is equal to the identity matrix were given for the transform case and the unconstrained length case. In both these cases, it can be assumed that the Wiener filter bank has the same order as the first part of the synthesis filter bank, which is either unconstrained or zero. In the FIR case, this is not possible because, if the first part of the synthesis filter bank is the FIR inverse of the analysis filter bank, the following Wiener polyphase filter bank can only be a memoryless matrix if the order of the total synthesis filter bank is to be kept constant. By using this method, the optimization does not include a search over all polyphase matrices of a given order. Therefore, a different method must be used.

Assume that the analysis FIR filter bank $\boldsymbol{E}(z)$ is FIR invertible, and let $\hat{\boldsymbol{R}}(z)$ be the FIR synthesis filter bank for which the filter banks possess the PR property, that is:

$$\hat{\boldsymbol{R}}\_\boldsymbol{E}_{\Gamma} = [\boldsymbol{0} \dots \boldsymbol{0}\boldsymbol{I}\boldsymbol{0} \dots \boldsymbol{0}], \tag{7.23}$$

where the right hand side of the equation is an $N \times (m+l+1)N$ matrix, and where the non-zero element in the first row and column of the $N \times N$ identity matrix $\boldsymbol{I}$ is placed in column number $Nd_v + d_s$. The numbering of the columns starts with 0.

In order to ensure that PR is possible, let $M = N$. The FIR PR system is shown in Figure 7.4. Since the system is a PR system, the output of the system is equal to $\hat{\boldsymbol{x}}(n) = \boldsymbol{x}_{d_s}(n - d_v) + \hat{\boldsymbol{R}}\_\boldsymbol{q}(n)_{\downarrow}$, where $\boldsymbol{x}_{d_s}(n - d_v)$ is the signal going through the filter bank with vector delay d_v and scalar delay d_s. $\hat{\boldsymbol{R}}\_\boldsymbol{q}(n)_{\downarrow}$ is the additive quantization noise filtered through the FIR synthesis PR filter bank $\hat{\boldsymbol{R}}(z)$.

From Figure 7.4, it can be seen that

$$\boldsymbol{x}_{d_s}(n - d_v) = \hat{\boldsymbol{R}}\_\hat{\boldsymbol{y}}(n)_{\downarrow} - \hat{\boldsymbol{R}}\_\boldsymbol{q}(n)_{\downarrow}. \tag{7.24}$$

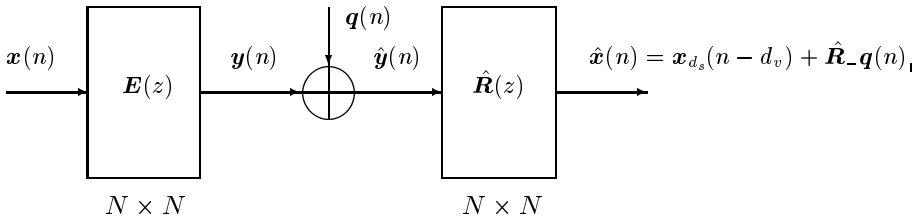

Figure 7.4 FIR PR filter bank model.

Now, conditions for when FIR PR filter banks are optimal will be derived, given invertible FIR analysis filter bank $\boldsymbol{E}(z)$. The total row-expanded polyphase matrix $\boldsymbol{W}_{-}$ through a system which has a row-expanded Wiener polyphase matrix $\boldsymbol{R}_{-}$ can be expressed as

$$\begin{aligned}
\boldsymbol{W}_{-} &= \boldsymbol{R}_{-}\boldsymbol{E}_{\urcorner} \\
&= \boldsymbol{\phi}_{\boldsymbol{x},\hat{\boldsymbol{y}}}^{(l,N,N)}(d_v, d_s)\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\right)^{-1}\boldsymbol{E}_{\urcorner} \\
&= E\left[\left(\hat{\boldsymbol{R}}_{-}\hat{\boldsymbol{y}}(n)_{\downarrow} - \hat{\boldsymbol{R}}_{-}\boldsymbol{q}(n)_{\downarrow}\right)\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\right)^{-1}\boldsymbol{E}_{\urcorner} \\
&= \hat{\boldsymbol{R}}_{-}\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\right)^{-1}\boldsymbol{E}_{\urcorner} - \hat{\boldsymbol{R}}_{-}E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\right)^{-1}\boldsymbol{E}_{\urcorner} \\
&= \hat{\boldsymbol{R}}_{-}\boldsymbol{E}_{\urcorner} - \hat{\boldsymbol{R}}_{-}E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\right)^{-1}\boldsymbol{E}_{\urcorner}, \qquad (7.25)
\end{aligned}$$

where $M = N$ and the results from Equations (7.18), (7.19), and (7.24) have been used. The system has the PR property with appropriate delay if, and only if, the total transfer matrix $\boldsymbol{W}_{-}$ is equal to $\hat{\boldsymbol{R}}_{-}\boldsymbol{E}_{\urcorner}$ given in Equation (7.23). From Equation (7.25), it is seen that this is the case if, and only if, the last term in the equation is equal to zero. By using the assumption that the matrix $\boldsymbol{E}(z)$ has a unique FIR inverse when studying Equation (7.23), it can be shown that:

$$\operatorname{rank}\left(\boldsymbol{E}_{\urcorner}\right) = (l+1)N. \qquad (7.26)$$

Using this when studying the dimension of the left nullspace [Strang 1988] of the matrix $\boldsymbol{E}_{\urcorner}$, it follows that the last term of Equation (7.25) is zero if, and only if, the matrix $\hat{\boldsymbol{R}}_{-}E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]\left(\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}\right)^{-1}$ is equal to the zero matrix. Since the matrix $\boldsymbol{\Phi}_{\hat{\boldsymbol{y}}}^{(l,N)}$ is assumed to be invertible, PR is optimal if, and only if, the following holds:

$$\hat{\boldsymbol{R}}_{-}E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right] = \boldsymbol{0}, \quad \forall n, \qquad (7.27)$$

where the zero matrix on the right hand side is of dimension $N \times (l+1)N$.

Given the conditions for PR being optimal in the unconstrained length case, see Equation (2.18), and in the transform case, see Equation (3.20), it is intuitively surprising that the matrix $\hat{\boldsymbol{R}}_{-}$ is part of the conditions in the FIR PR case. It might be more intuitive to guess that the condition in the FIR PR case would be that the matrix $E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]$ should be the zero matrix. However, the conditions in the FIR case are *not* that strict because, in the following example, it will be shown that the matrix $E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]$ is non-zero, but the conditions in Equation (7.27) are satisfied.

Let $N = M = 2$, $m = l = 1$, $d_s = 0$, $d_v = 1$, $H_0(z) = 1$, and $H_1(z) = z^{-3}$, where $H_i(z)$ is the transfer function of analysis filter number i. In order to ensure PR with the appropriate delay, the synthesis filters are given by $F_0(z) = z^{-3}$ and $F_1(z) = 1$. The matrix $\hat{\boldsymbol{R}}_{-}$ is now given by

$$\hat{\boldsymbol{R}}_{-} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{7.28}$$

With this choice the following non-zero block Toeplitz matrix will satisfy the condition in Equation (7.27):

$$E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right] = \begin{bmatrix} 0 & 0 & x_2 & x_3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ x_0 & x_1 & 0 & 0 \end{bmatrix}, \tag{7.29}$$

which is different from the zero matrix, because x_i is an arbitrary real number for all $i \in \{0, 1, 2, 3\}$.

With the quantization model used in Section 4.1, the additive quantization noise and the subband signals were assumed to be uncorrelated. In this case, the cross-correlation matrix in Equation (7.27) is given by $E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right] = \boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$, which is an invertible matrix for finite rates. The condition in Equation (7.27) cannot be satisfied because the only solution is $\hat{\boldsymbol{R}}_{-} = \boldsymbol{0}$, which is certainly not a PR filter bank. Therefore, with the quantization model used in Section 4.1, it is never optimal to use a PR filter bank for finite rates. However, for an infinite rate, the quantization noise will approach zero, and the cross-correlation matrix $E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]$ will approach zero as well. For infinitely high rates, the condition will be satisfied, and the PR filter bank is optimal.

With the quantization model used in Section 7.2, the *diagonal* elements of the matrix $E\left[\boldsymbol{q}(n)_{\downarrow}\hat{\boldsymbol{y}}^H(n)_{\downarrow}\right]$ will be zero, since centroids are used as the representation level in the quantizers. With this model, PR filter banks are optimal if $N = M = 1$ and $m = l = 0$. In all optimized cases other than $N = M = 1$ and $m = l = 0$, that has been considered, the condition in

Equation (7.27) is never satisfied. The reason for this, is that for all other values for N, M, m, and l, the off-diagonal elements of the cross-correlation matrix $E\left[\boldsymbol{q}(n)_{\mathsf{l}}\hat{\boldsymbol{y}}^H(n)_{\mathsf{l}}\right]$ are non-zero.

If the analysis filter bank is invertible and no quantization noise is present, it follows from the conditions in Equation (7.27) that PR is optimal. In this case, PR filter banks achieve zero MSE. Since MSE is always non-negative, this is optimal.

7.4 Results and Comparisons

To evaluate the proposed methods and make comparisons to existing PR filter banks, the following two channel FIR filter banks have been optimized: 5_3, 6_6, 9_7, and 10_18. The input source is a Gaussian AR(1) source with correlation coefficient 0.95.

This section is organized as follows: In Subsection 7.4.1, results obtained with the redesigned scalar quantizers are presented, and in Subsection 7.4.2 performance results with subtractive dithering are presented. Results for the model with signal dependent colored quantization noise are presented in Subsection 7.4.3. Finally, in Subsection 7.4.4, practical coding results obtained with all the proposed methods are compared to practical results obtained with filter banks found in the literature.

7.4.1 Redesigned Scalar Quantizer

For the quantizers redesigned as shown in Subsection 7.1.1, the different MSE contributions for the practical system are shown in Figure 7.5 together with the MSE contributions for the practical system using centroids as representation levels in the quantizers. The 9_7 filter banks were found by the optimization algorithm in Section 4.1, with $M = 1$.

From Figure 7.5, it is seen that the total MSE is reduced when using the redesigned quantizers. It is also seen that the crossterm MSE contribution vanishes with the new quantizers, and that is the main reason for the reduction of the total MSE.

In Appendix D, a formula for $E\left[x_i(n)q_k(p)\right]$ is given. This theory can be extended to include a formula for $E\left[y_i(n)q_k(p)\right]$ as well. With the quantizers introduced in Subsection 7.1.1, $E\left[y_i(n)q_i(n)\right] = 0$ for $i \in \{0, 1, \ldots, M-1\}$, as this is a design condition for these quantizers. If this result is utilized in the formulas for $E\left[y_i(n)q_k(p)\right]$ and $E\left[x_i(n)q_k(p)\right]$, it can be theoretically shown that the matrices $E\left[\boldsymbol{y}(n)_{\mathsf{l}}\boldsymbol{q}^H(n)_{\mathsf{l}}\right]$ and $E\left[\boldsymbol{x}(n)_{\mathsf{l}}\boldsymbol{q}^H(n)_{\mathsf{l}}\right]$ are equal to zero matrices when using the new quantizers. Therefore, the crossterm MSE contribution is

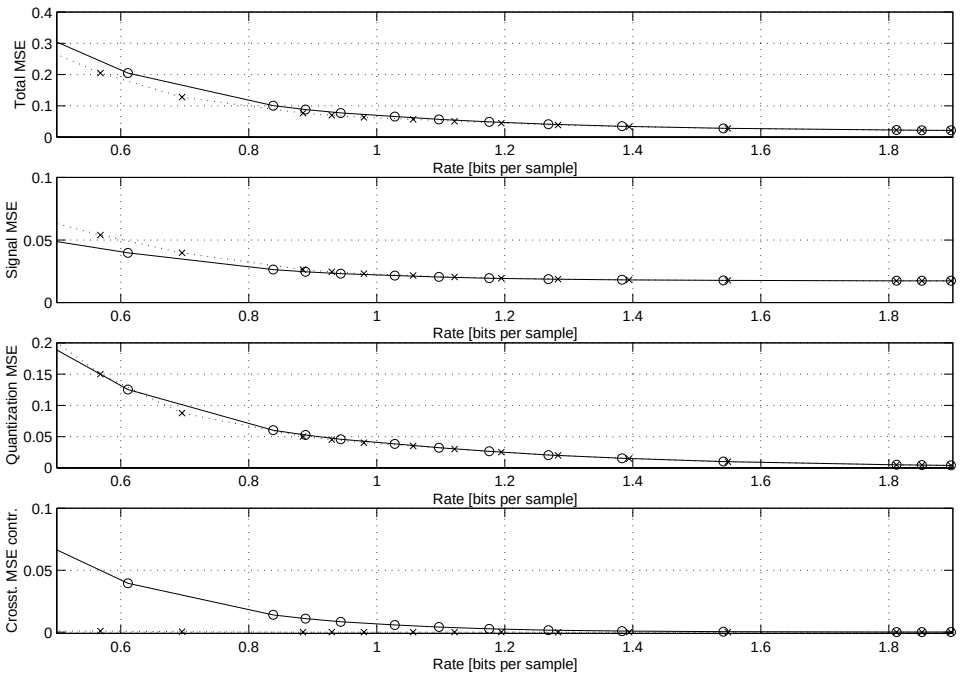

Figure 7.5 Different MSE contributions per source sample as a function of coding rate with $N = 2$ and $M = 1$. The dotted curves with $\times$-marks show the MSE contributions for the practical coding system with the redesigned quantizers, and the solid curves with circles show the MSE contributions for the practical coding system with centroids as representation levels in the quantizers. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95. The optimized 9_7 filter banks from Section 4.1 are used.

equal to zero when using the quantizers introduced in Subsection 7.1.1.

From Figure 7.5, it is seen that for a given Lagrange multiplier, the signal MSE is the same in the two systems, but the rate is different. The reason for this is that the output of the synthesis filter bank in the absence of quantizers $\hat{x}_{\text{sig}}(n)$ is equal in the two systems, and as will be shown later in Figure 7.8, for a given value of the additive quantization noise variance σ_q^2, more bits are required with the redesigned quantizers than with the quantizers using centroids as representation levels.

The quantization MSE vs. rate curves for the two systems seem to be quite close to each other. The reason for this is that for a given Lagrange multiplier, the rate increases using the redesigned quantizers compared to using

centroids as representation levels, but at the same time, the quantization noise using the redesigned quantizers will decrease since the off-diagonal terms in the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ are much closer to zero when using the redesigned quantizers compared to uniform quantizers with centroids as representation levels.

7.4.2 Subtractive Dithering

When using subtractive dithering, the input signal to the uniform threshold quantizer has variance $\sigma_{y_i}^2 + \frac{\Delta_i^2}{12}$, and the pdf of the input signal is given by the convolution of a rectangular and a Gaussian function. This will change the way the entropy is estimated since the pdf of the input signal of the uniform threshold quantizer will change. The way Δ_i is estimated has to be changed since the input pdf to the quantizer is changed. Midpoints must be used as representation levels in the uniform threshold quantizers in order to ensure that the subband signal and the quantization noise are uncorrelated.

The different MSE contributions when the changes described above are made in both the theoretical and practical coder, are shown in Figure 7.6. The 9_7 filter banks found by the optimization algorithm in Section 4.1 were used.

From Figure 7.6, it is observed that the the theoretical and practical results match very well for all the different MSE contributions. Here, the crossterm MSE contribution vanishes. The theoretical estimate of the MSE contributions fit very well to the corresponding practical values, because the assumptions made in the theoretical case are satisfied in the practical case. The theoretical rate estimate in Equation (7.1) gives a very good estimate of the rate in the practical subband coder.

7.4.3 Signal Dependent Colored Quantization Noise Model

Figure 7.7 shows different MSE contributions as a function of the coding rate for $N = 2$ and $M = 1$. The rate is estimated by calculating the entropy of the quantization indices from a uniform threshold quantizer having quantizer step size Δ_i and a Gaussian time series as input, with variance $\sigma_{y_i}^2$ and zero mean. The value of Δ_i is chosen such that $\sigma_{q_i}^2 = 1$. From the figure, it is seen that the crossterm MSE contribution is negative. This means that the crossterm MSE contribution is not an MSE itself, but just a part of the total MSE that might be negative. This is also the reason why the word *contribution* is used to describe this term. From the figure, it is also seen that all the theoretical MSE components match very well with the practical components. This shows that the signal dependent colored quantization noise model, introduced in Section 7.2, matches very well with the practical coder. Furthermore, observe from the figure that the quantization MSE is larger than,

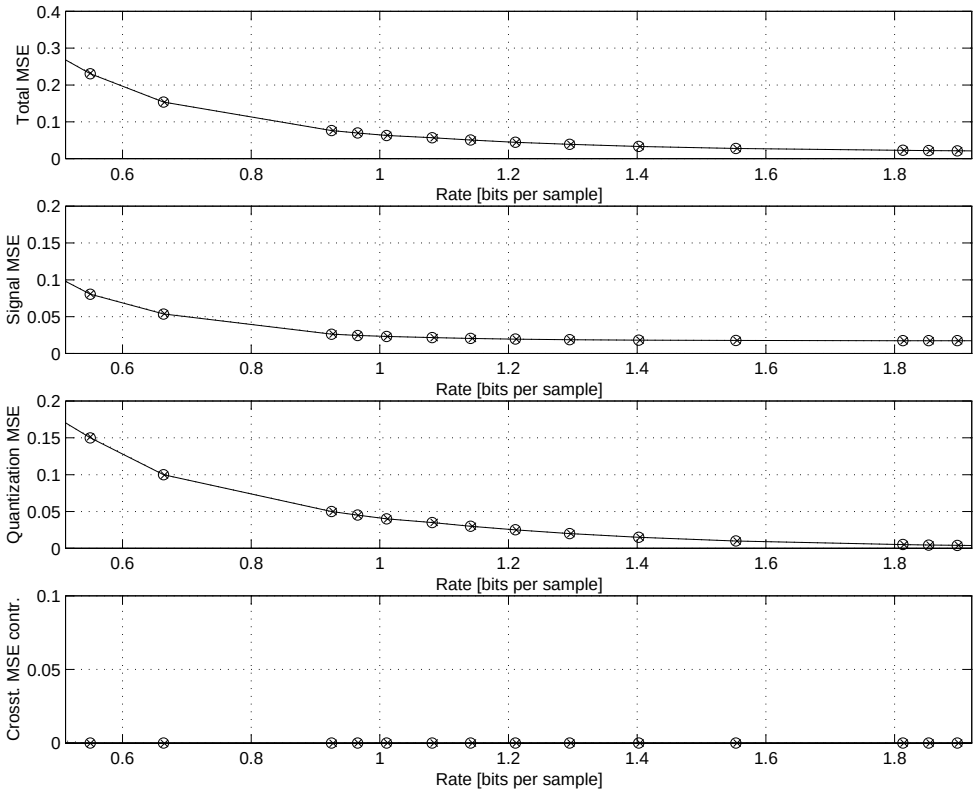

Figure 7.6 Different MSE contributions per source sample as a function of coding rate using subtractive dithering when $N = 2$ and $M = 1$. The dotted curves with $\times$-marks show the MSE contributions for the practical performance obtained with subtractive dithering, and the solid curves with circles show the theoretical MSE contributions. The filter banks used are the optimized 9_7 filter banks from Section 4.1. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95.

but almost equal to, the total MSE, while the signal MSE and the crossterm MSE contribution are small.

In Figure 7.7, the MSE components of the practical coder using the well known PR 9_7 filter bank from [Antonini et al. 1992] are included. From the figure, it is seen that the signal MSE and the crossterm MSE contribution are zero for the PR filter bank, and that the quantization MSE is equal to the total MSE. It is also observed that the performance of the proposed filter bank is significantly better than the PR filter bank at all the rates in Figure 7.7.

An interesting observation is that when the PR constraint is removed, as

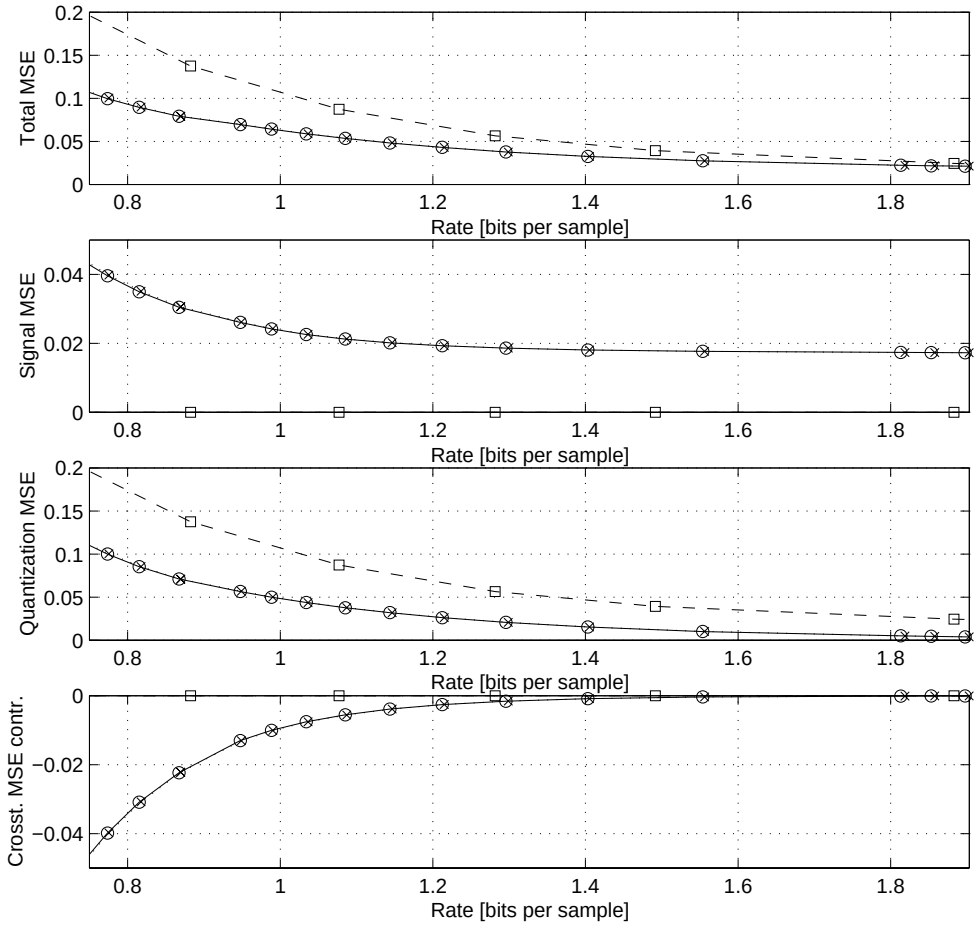

Figure 7.7 Different MSE contributions per source sample as a function of coding rate when $N = 2$ and $M = 1$. The dotted curves with $\times$-marks show the MSE contributions for the practical coding system, and the solid curves with circles show the MSE contributions for the theoretical coding system with the signal dependent colored quantization noise model. The dashed curves with squares show the MSE contributions for the 9_7 PR filter bank in [Antonini et al. 1992]. The input signal is unit variance Gaussian AR(1) with correlation coefficient 0.95. The filter banks used are the optimized 9_7 filter banks from Section 7.2.

it is in the proposed filter banks, the resulting optimized filter banks have a small signal MSE and a small crossterm MSE contribution, but the total MSE is significantly reduced compared to the PR filter bank used in the comparison,

Table 7.1 Practical distortion rate performances.

$N = 2$. The source is Gaussian AR(1) with correlation coefficient 0.95, d_v indicates the vector delay used in the system, and $d_s = 0$ in all cases. The case where convergence was not reached is marked by X.

Type of coding system	Bit rate [bits per sample]			
	0.50	1.00	2.00	3.00
	SNR [dB]			
Dist. rate func. [Berger 1971]	12.70	16.13	22.15	28.17
5_3 [Le Gall & Tabatabai 1988], $d_v = 1$	4.88	10.57	17.00	22.81
Biorth. 5_3 [Balasingham 1998], $d_v = 1$	4.90	10.59	17.03	**22.84**
5_3 [Balasingham 1998] Wiener, $d_v = 1$	**4.95**	**10.63**	17.03	**22.84**
Proposed 5_3 with cent., $d_v = 1$	4.43	10.26	16.63	22.54
Proposed 5_3 with new quant., $d_v = 1$	4.85	10.58	16.95	22.80
Proposed 5_3 with dith., $d_v = 1$	3.58	10.24	16.50	22.49
Proposed 5_3 with corr., $d_v = 1$	4.92	10.57	**17.05**	**22.84**
6_6 [Rodrigues et al. 1997], $d_v = 2$	4.68	10.18	16.77	22.57
Biorth. 6_6 [Balasingham 1998], $d_v = 2$	4.94	10.65	16.93	22.76
6_6 [Balasingham 1998] Wiener, $d_v = 2$	**5.82**	11.56	17.08	22.80
Proposed 6_6 with cent., $d_v = 2$	5.10	11.41	17.03	22.96
Proposed 6_6 with new quant., $d_v = 2$	5.39	11.85	17.14	23.23
Proposed 6_6 with dith., $d_v = 2$	5.22	11.57	16.76	22.93
Proposed 6_6 with corr., $d_v = 2$	5.81	**11.86**	**17.48**	**23.24**
9_7 [Antonini et al. 1992], $d_v = 5$	4.48	9.86	16.73	22.44
Biorth. 9_7 [Balasingham 1998], $d_v = 5$	5.26	11.17	17.26	23.07
9_7 [Balasingham 1998] Wiener, $d_v = 5$	**6.13**	11.87	17.42	23.13
Proposed 9_7 with cent., $d_v = 5$	5.35	11.56	17.35	23.40
Proposed 9_7 with new quant., $d_v = 5$	5.76	**12.04**	17.72	**23.69**
Proposed 9_7 with dith., $d_v = 5$	5.56	11.90	16.95	23.37
Proposed 9_7 with corr., $d_v = 5$	5.84	12.02	**17.83**	23.66
10_18 [Tsai et al. 1996], $d_v = 6$	4.72	10.30	17.03	22.80
10_18 [Tsai et al. 1996] Wiener, $d_v = 6$	6.42	11.68	17.21	22.86
Proposed 10_18 with cent., $d_v = 6$	5.92	12.28	17.90	24.06
Proposed 10_18 with new quant., $d_v = 6$	6.82	12.84	**18.41**	24.29
Proposed 10_18 with dith., $d_v = 6$	6.55	12.57	17.52	24.02
Proposed 10_18 with corr., $d_v = 6$	**7.47**	**12.87**	X	**24.32**

see Figure 7.7.

7.4.4 Practical Performance Comparisons

The 5_3, 6_6, 9_7, and 10_18 filter banks have been optimized, and the practical results are compared to some of the best performing filter banks with the same filter lengths found in the literature. The filter banks used in the comparisons are the following: 5_3 [Le Gall & Tabatabai 1988], 5_3 [Balasingham 1998], 6_6 [Balasingham 1998], 6_6 [Rodrigues et al. 1997], 9_7 [Antonini et al. 1992], 9_7 [Balasingham 1998], and 10_18 [Tsai et al. 1996]. The coding system described in Figure 6.1 is used with centroids as representation levels when coding the subband signals in the PR filter banks. The practical performance results are shown in Table 7.1. For the 10_18 filter bank with $b = 2.00$ bits per sample, simulation results have not been obtained when using the method proposed in Section 7.2. The reason for this is that the proposed signal dependent colored noise optimization method is very complex, and therefore, convergence might be difficult to reach. This happens when one of the subbands receives very few bits. In some cases using the signal dependent colored noise model, several different initial values of the filter banks had to be tried before convergence was reached.

In the table, results are also included for systems using a PR analysis filter bank and FIR Wiener synthesis filter bank with the same filter lengths as the PR FIR synthesis filter bank. In the 5_3, 6_6, and 9_7 cases, the analysis filter banks found in [Balasingham 1998] are used, while in the 10_18 case, the analysis filter bank in [Tsai et al. 1996] is used. The FIR Wiener filter bank is found from Equation (7.17), and the matrices $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$, $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$, and $\boldsymbol{\phi}_{\boldsymbol{q},\boldsymbol{x}}^{(l,N,M)}(d_v, d_s)$ are found by the formulas developed in Appendix D, with centroids as representation levels. The bit allocation used with the Wiener filter banks is the same as the bit allocation used with PR filter banks, i.e., the bits are distributed such that the product of the quantization variance and the squared norm of the *PR* synthesis filter is constant for each branch of the filter bank. It is seen from the table that the performance is improved by using the Wiener synthesis filter bank at low rates. For short filter lengths the gain is small, but for longer filter lengths and low rates much can be gained by using Wiener synthesis filter banks. A system using a PR analysis filter bank, Wiener synthesis filter bank, and bit allocation is proposed in [Gosse & Duhamel 1997]. An iterative algorithm is proposed where the bit allocation is optimized for a given Wiener synthesis filter bank, and vice versa. The Wiener result presented in Table 7.1 is therefore the result obtained by the theory proposed in [Gosse & Duhamel 1997] after one iteration of the synthesis filter bank optimization. Only one quantizer receives bits when using 0.50 bits/sample or 1.00 bits per sample in Table 7.1, and for these rates the bit allocation is uniquely given after

one iteration. Thus, for these rates, the Wiener filter bank results given in the table are the same as the results obtained by the method proposed in [Gosse & Duhamel 1997].

If the Wiener synthesis filter bank is estimated based on Equation (4.15), i.e., assuming that the input signal is uncorrelated to the additive quantization noise, and the autocovariance matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ is diagonal, the practical results are worse than the results obtained by the PR filter banks in most cases. This shows that the quantization noise model used in Section 4.1 is accurate enough.

When the practical results obtained using subtractive dithering are compared to the practical results of Chapter 6, where uniform threshold quantizers with centroid representation levels were used, the best results depend on the rate and filter length that is used. No definite trend can be found. The reason for this is that when using subtractive dithering, the assumption made in the signal independent white noise model in Section 4.1 is satisfied, but not when uniform threshold quantizers with centroid representation levels are used. However, as will be shown by Figure 7.8, the coding of the subbands is not as efficient when using subtractive dithering compared to uniform threshold quantizers with centroids as representation levels.

It is seen from the table that the results obtained with the redesigned scalar quantizer are better than the results obtained with subtractive dithering. This can be explained by Figure 7.8, which shows the theoretical distortion rate performance of three coding systems coding *one Gaussian subband signal*. The rate is found from Equation (7.1). From Figure 7.8, it is seen that the distortion rate performance of the redisigned scalar quantizers is better than the performance of the subtractive dithering when coding the subband signal. Two reasons for this are that the coding of the subbands in the subtractive dithering case is not as efficient because the midpoints are used as representation levels in the quantizers and the variance of the input to the quantizers is increased by $\frac{\Delta_i^2}{12}$, even though the variance of the subband signal is not changed. Since both the redesigned quantizers and subtractive dithering will produce coding noise which is uncorrelated with the subband samples, the redesigned quantizers will perform better than subtractive dithering when using the same filter bank.

The results using the redesigned quantizers are better than the results obtained by the quantizers using centroids as representation levels. This shows that it is very important to design the coding method in accordance with the assumptions made when designing the filter banks. Even though the distortion rate performance of the quantizers using centroid representation levels are better than the redesigned scalar quanitzers, see Figure 7.8, the overall subband coder which uses the redesigned quantizers performs better.

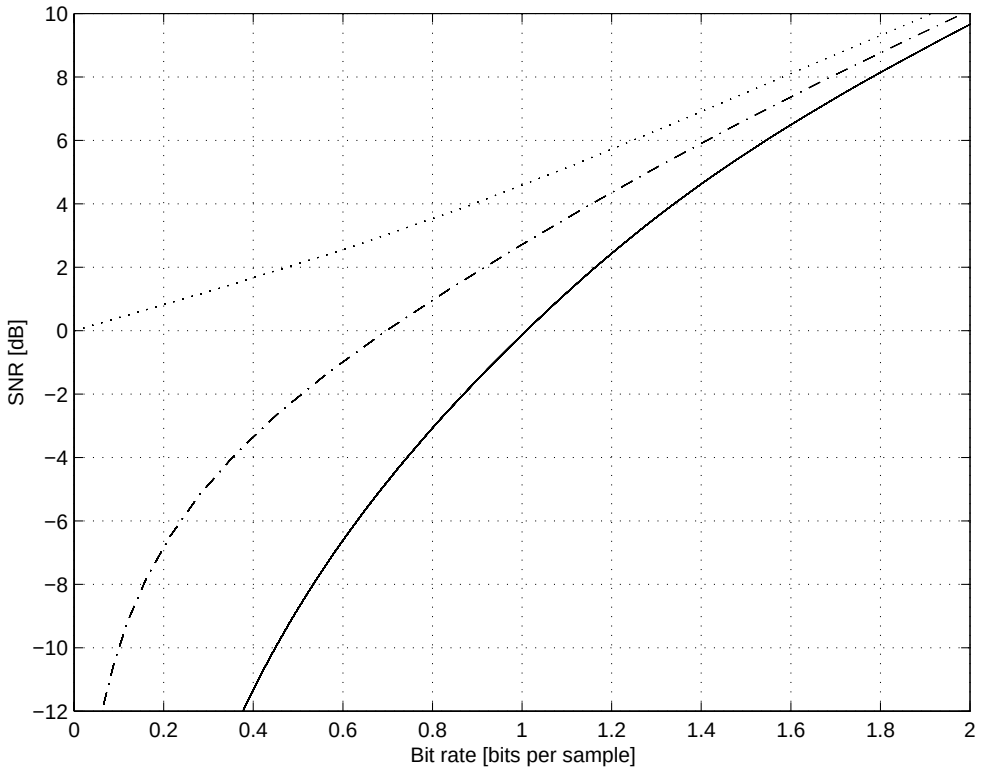

Figure 7.8 Theoretical distortion rate performance using uniform threshold quantizers having centroid representation levels (dotted), uniform threshold quantizers with scaled centroid representation levels (dash-dotted), and subtractive dithering (solid). A Gaussian time series is coded.

If the results obtained with the signal dependent colored quantization noise model are compared to the results of the redesigned scalar quantizers, it is seen that for low rates, better performance is achieved with the signal dependent colored quantization noise model in most cases. However, for higher rates, the performance of the two systems are very close. The reason why the signal dependent colored quantization noise is not optimal for all rates is a combination of the assumption that the coding coefficients c_i are rate independent and statistical deviation. In the simulations, time series of 300 000 samples are used, and the same seed is used when generating the input time series. The length of the 95 % confidence interval [Hines & Montgomery 1990] for the MSE per source sample is less than 0.05 dB, if the error time series is assumed to be Gaussian.

7.5 Summary

In this chapter, the modeling of the coding of the subband signals was improved. Three different techniques were proposed.

The two first methods studied were the redesigned scalar quantizers and subtractive dithering. In these methods, the subband signals and the additive coding noise are unorrelated. The practical simulations showed that the crossterm MSE contribution vanishes when using both these methods. The performance of the redesigned scalar quantizers is better than the performance of subtractive dithering. Therefore, if it is desired that the input is uncorrelated with the coding noise, the proposed new quantizers should be chosen instead of subtractive dithering if distortion rate performance is an important criterion. The redesigned quantizers gave better practical results than the quantizers using centroids as representation levels. When using subtractive dithering, very good correspondence was achieved between theoretical and practical results.

The third modeling technique was a signal dependent colored quantization noise model where the output of the quantizers were modeled to discrete values being equal to the representation levels in the quantizers. The description of the quantizers is given by the pdfs, providing a statistical quantization model. Very good correspondence was achieved between the theoretical and the practical simulations by this model. From Table 7.1, it was seen that the system using the redesigned quantizers will perform very close to the system using the signal dependent colored quantization noise model, and in some cases the performance was better. Since the complexity of the optimization of the redesigned quantizer system is lower than that using the signal dependent colored quantization noise model, the redesigned quantizers can be justified in many applications. However, the signal dependent colored quantization noise model gave the best results in most cases.

The conditions for when PR FIR filter banks are optimal were derived in this chapter. This is an extension of the results given in [Vaidyanathan & Chen 1994] to include FIR filter banks. The conditions are very strict, and for all the cases that were studied, PR is never optimal, except the case $N = M = 1$ and $m = l = 0$.

It should also be possible to extend the signal dependent colored quantization noise model to include a dead-zone [Sullivan 1996] around zero in the uniform threshold quantizers. Uniform threshold quantizers are used in JPEG [ISO 1991], and practical image coding has shown that improved distortion rate performance may be achieved by using such a quantizer.

Chapter 8

Conclusions

In this dissertation, the problems of optimizing filter banks and transforms under a bit or a power constraint have been investigated. The optimality criterion used is the minimization of the MSE between the output and the input signals. Three different classes of filter banks have been treated: Unconstrained length filter banks, transforms, and FIR filter banks with arbitrary given filter lengths. It has been shown that the synthesis filter bank and transform is equal to the Wiener filter bank or transform in all the cases that have been considered.

Conclusions from the Theoretical Results

In the three filter length cases treated, a high rate model has been used to model the quantizers performance, and this model was used for all rates. Better theoretical performance than PR filter banks was achieved for all cases considered. The reason for this is that in the proposed system, no constraint such as PR was employed. Thus, the set used in the optimizations in this dissertation includes the unitary and biorthogonal filter bank sets as proper subsets. This is illustrated in Figure 8.1. For the proposed unconstrained length filter bank system with a bit constraint, this means that it performs at least as well as the optimal unconstrained length unitary and biorthogonal filter banks at all rates and all sources. The results also showed that the proposed transform coder performs at least as well as the KLT for all rates and sources, and that the proposed FIR filter banks perform better than the FIR filter banks compared having the same filter lengths. This includes the case where an FIR Wiener filter bank was used on the synthesis side.

The theoretical results showed that the filter banks and transform matrices depend on the bit rate or channel quality used and the PSD of the signal which is compressed or transmitted. Consequently, side information is required in a

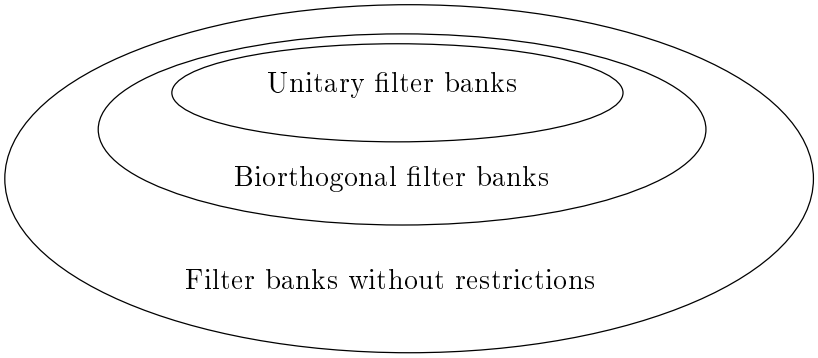

Figure 8.1 Venn-diagram of different sets of filter banks.

practical coder if the coding rate or channel quality and the input statistics are unknown. The traditional way of using bandpass filters with one contiguous passband in each subband is suboptimal for certain PSDs.

Three different filter length cases have been treated:

In the first case, the filters were allowed to be non-causal with infinite length impulse responses. The unconstrained length case is of theoretical interest only, since the filter banks found are not realizable. However, the solutions are fundamental for a thorough understanding of the problem, and they also provide an upper bound on the SNR vs. rate or CSNR performance of the transform and FIR filter banks cases. The unconstrained length filter bank results showed that the frequency response of one filter can have more than one passband. The resulting filters in the unconstrained length filter bank have ideal frequency separation between the subbands, and at high rates or very good channel qualities, the filters approximate half-whitening filters within each passband. If signal expansion is considered in the power constrained problem, i.e., $M > N$, $M - N$ of the filters will be set to zero. With bandwidth reduction, i.e., $M < N$, the performance of the linear system will not be very good for high quality channels because PR is impossible when not fully ranked polyphase matrices are used in filter banks.

In the second case, a transform coder was optimized. These results have immediate practical applications, since transforms are easily implementable. Analytical expressions have been derived for the jointly optimal analysis and synthesis transforms. For the bit constrained problem, analytical expressions for the bit allocation have been found. Differences between the proposed transform and the KLT have been pointed out. Formulas have been found for jointly optimal Wiener transform and bit allocation when a KLT analysis transform

with reduced rank was used. It was shown that the performance of this system is the same as the proposed bit constrained transform. The solution of the jointly optimal transmitter and receiver transforms under a power constraint was derived in [Lee & Petersen 1976], but an alternative derivation was presented in this work.

The third case studied was the optimization of an FIR filter bank structure. Analytical expressions for the optimal solution were not found in this case, but iterative numerical optimization algorithms were proposed for finding FIR signal-adaptive jointly optimized analysis and synthesis filter banks under a bit or a power constraint. These results are practically interesting, since FIR filters can be used in practical coders. The FIR filter banks should not be designed by approximating the unconstrained length solution, but the FIR constraints should be included in the problem formulation, and a procedure showing how this can be done was given. Global convergence is not guaranteed in the iterative algorithms proposed, but very good results were achieved.

Conclusions from the Practical Results

A practical source coder has been introduced, and it was shown that there is a mismatch between the theoretical results obtained by the white signal independent quantization noise model and the results found with the practical source coder. The reason for this has been analyzed, and it was concluded that the assumptions of white and signal independent quantization noise are the main reasons for the mismatch.

Three methods for improving the mismatch have been proposed. In the first two, the coding of the subband signals was changed and the filter banks were kept constant. In the third method, a signal dependent colored quantization noise model was introduced, and by means of this model the filter banks were re-optimized. The results show that the filter banks found with the signal dependent colored quantization noise model perform comparable or better than PR filter banks and filter banks using Wiener synthesis filter banks. For the largest filter lengths considered, the gain was highest, i.e., over 1.0 dB for rates below 3.0 bits/sample for the 10_18 filter bank. For smaller filter lengths, there were some cases where the filter banks with the synthesis Wiener filter performed slightly better than the proposed filter banks.

This can be explained by Figure 8.2. The upper left corner shows the real world problem, which in this case is a practical filter bank that is supposed to perform optimally. This problem was first modeled by the white signal independent quantization model, that was the first approximative mathematical model used for the practical problem. The solution was tried in a practical coder, and it was shown that this did not give very good correspondence be-

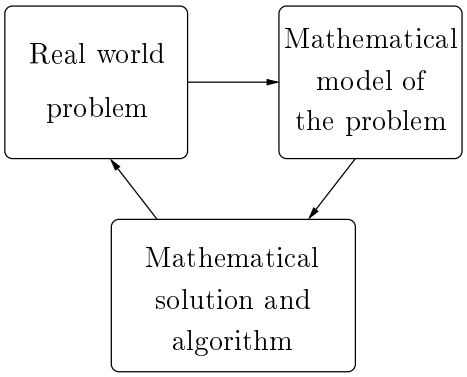

Figure 8.2 Basic concepts of research.

tween theoretical and practical results. This first attempt corresponds to the first iteration of Figure 8.2. More advanced quantization noise models were then introduced as the next mathematical model, and that lead to very good correspondence between the theoretical and practical results. Therefore, the next iterations of Figure 8.2 improved the results considerably, but even with these new models the results of the best performing filter banks were not always optimal in the practical system. This suggest that the mathematical model of the real world is still not complex enough to take into account all the effects in a practical subband coder. However, the results are significantly better than conventional filter bank results for the longest filters considered. The proposed theory has hopefully given some new insight into the problems of designing optimal filter banks. Some of the filter banks obtained may be applied in practical source coders and in communication systems.

Suggestions for Future Research

- Similar theory should be developed for nonuniform FIR filter banks as well, because this kind of filter banks is often used in practical systems.
- If the filter banks are to be used on multidimensional signals such as images, the theory should be adapted for these kind of signals, and non-separable filter banks should be found.
- The theory developed should be extended to include weighted MSE. If the filter banks are used in a practical system, the MSE optimization criterion does not match human perception very closely. This can be improved by using weighted MSE or other performance criteria.

- Theory can be found for other source models. In this dissertation, it has been assumed that the input signal is WSS, with a known PSD. Other source models, especially non-stationary source models, would be interesting in many practical applications.
- Formulas for finding the correlations that exist in a subband coder for other coding methods of the subband signals should be developed. Examples of other coding methods are trellis coded quantization, vector quantization, etc. These formulas can be used to optimize the filter banks when using these coding methods.

Appendix A

Derivation of Block MSE, Bit Constraint, and Power Constraint

In this appendix, expressions for the block MSE, bit constraint, and power constraint are derived for both unconstrained length and FIR filter banks. These expressions are used to formulate some of the optimization problems treated in the dissertation.

For simplicity, the summation limits will not be written explicitly in the unconstrained case, but all sums treating the unconstrained length case in this appendix go from $-\infty$ to ∞.

A.1 Block MSE Derivation

The derivation of the block MSE for unconstrained length and FIR filter banks are presented in this section. The derivation in the FIR case is *not* a special case of the unconstrained case since causal FIR filters are assumed, and therefore there exists a delay through the FIR filter banks which has to be taken into consideration.

A.1.1 Block MSE for Unconstrained Length Filter Banks

From Equations (1.9), (1.10), (1.11), and (1.12), the block MSE can be expressed as

$$\begin{aligned}
\mathcal{E}_{N,M} &= \mathrm{Tr}\left(E\left[\boldsymbol{\epsilon}(n)\boldsymbol{\epsilon}^H(n)\right]\right) \\
&= \mathrm{Tr}\left(E\left[(\hat{\boldsymbol{x}}(n)-\boldsymbol{x}(n))(\hat{\boldsymbol{x}}(n)-\boldsymbol{x}(n))^H\right]\right) \\
&= \mathrm{Tr}\left(E\left[\left(\sum_m \boldsymbol{w}(n-m)\boldsymbol{x}(m)-\boldsymbol{x}(n)+\sum_m \boldsymbol{r}(n-m)\boldsymbol{q}(m)\right)\cdot\right.\right. \\
&\qquad \left.\left.\left(\sum_k \boldsymbol{x}^H(k)\boldsymbol{w}^H(n-k)-\boldsymbol{x}^H(n)+\sum_k \boldsymbol{q}^H(k)\boldsymbol{r}^H(n-k)\right)\right]\right) \\
&= \mathrm{Tr}\left(\sum_m\sum_k \boldsymbol{w}(n-m)E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(k)\right]\boldsymbol{w}^H(n-k)+E\left[\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right]\right. \\
&\quad +\sum_m\sum_k \boldsymbol{r}(n-m)E\left[\boldsymbol{q}(m)\boldsymbol{q}^H(k)\right]\boldsymbol{r}^H(n-k) \\
&\quad \left.-\sum_m \boldsymbol{w}(n-m)E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(n)\right]-\sum_k E\left[\boldsymbol{x}(n)\boldsymbol{x}^H(k)\right]\boldsymbol{w}^H(n-k)\right),
\end{aligned} \tag{A.1}$$

where it is assumed that the input vector $\boldsymbol{x}(n)$ and the additive noise vector $\boldsymbol{q}(m)$ have zero mean, and that they are uncorrelated for all lags.

The last expression of Equation (A.1) contains five expressions, each of them will be further developed separately, below.

The first expression:

$$\begin{aligned}
&\mathrm{Tr}\left(\sum_m\sum_k \boldsymbol{w}(n-m)E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(k)\right]\boldsymbol{w}^H(n-k)\right) \\
&= \mathrm{Tr}\left(\sum_m\sum_k \boldsymbol{w}(n-m)\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S}_{\boldsymbol{x}}(f)e^{j2\pi f(m-k)}\,df\,\boldsymbol{w}^H(n-k)\right) \\
&= \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}}\sum_m \boldsymbol{w}(n-m)e^{-j2\pi f(n-m)}\boldsymbol{S}_{\boldsymbol{x}}(f)\left(\sum_k \boldsymbol{w}(n-k)e^{-j2\pi f(n-k)}\right)^H df\right) \\
&= \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}}\sum_k \boldsymbol{w}(k)e^{-j2\pi fk}\boldsymbol{S}_{\boldsymbol{x}}(f)\left(\sum_m \boldsymbol{w}(m)e^{-j2\pi fm}\right)^H df\right) \\
&= \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{W}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{W}^H(f)\,df\right),
\end{aligned} \tag{A.2}$$

where the matrix $E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(k)\right] = \boldsymbol{K}_{\boldsymbol{x}}(m-k)$ is found by calculating the inverse Fourier transform of the PSD matrix in Equation (1.8).

The second expression:

$$\mathrm{Tr}\left(E\left[\boldsymbol{x}(n)\boldsymbol{x}^H(n)\right]\right) = \mathrm{Tr}\left(\boldsymbol{K}_{\boldsymbol{x}}(0)\right) = \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S}_{\boldsymbol{x}}(f)\,df\right). \tag{A.3}$$

The third expression:

$$\begin{aligned}
&\mathrm{Tr}\left(\sum_m \sum_k \boldsymbol{r}(n-m)E\left[\boldsymbol{q}(m)\boldsymbol{q}^H(k)\right]\boldsymbol{r}^H(n-k)\right) \\
&= \mathrm{Tr}\left(\sum_m \sum_k \boldsymbol{r}(n-m)\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{\Lambda}_{\boldsymbol{q}} e^{j2\pi f(m-k)}\,df\,\boldsymbol{r}^H(n-k)\right) \\
&= \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_m \boldsymbol{r}(n-m)e^{-j2\pi f(n-m)}\boldsymbol{\Lambda}_{\boldsymbol{q}}\left(\sum_k \boldsymbol{r}(n-k)e^{-j2\pi f(n-k)}\right)^H df\right) \\
&= \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_k \boldsymbol{r}(k)e^{-j2\pi fk}\boldsymbol{\Lambda}_{\boldsymbol{q}}\left(\sum_m \boldsymbol{r}(m)e^{-j2\pi fm}\right)^H df\right) \\
&= \mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{R}(f)\boldsymbol{\Lambda}_{\boldsymbol{q}}\boldsymbol{R}^H(f)\,df\right),
\end{aligned} \tag{A.4}$$

where the quantization PSD matrix $\boldsymbol{\Lambda}_{\boldsymbol{q}}$ is given by Equation (1.17).

The fourth expression:

$$\begin{aligned}
&\mathrm{Tr}\left(-\sum_m \boldsymbol{w}(n-m)E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(n)\right]\right) \\
&= \mathrm{Tr}\left(-\sum_m \boldsymbol{w}(n-m)\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S}_{\boldsymbol{x}}(f)e^{-j2\pi f(n-m)}\,df\right) \\
&= \mathrm{Tr}\left(-\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_m \boldsymbol{w}(n-m)e^{-j2\pi f(n-m)}\boldsymbol{S}_{\boldsymbol{x}}(f)\,df\right) \\
&= \mathrm{Tr}\left(-\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_k \boldsymbol{w}(k)e^{-j2\pi fk}\boldsymbol{S}_{\boldsymbol{x}}(f)\,df\right) \\
&= \mathrm{Tr}\left(-\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{W}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\,df\right)
\end{aligned} \tag{A.5}$$

The fifth expression is given by the Hermitian of the fourth expression:

$$\begin{aligned}
&\mathrm{Tr}\left(-\sum_m E\left[\boldsymbol{x}(n)\boldsymbol{x}^H(m)\right]\boldsymbol{w}^H(n-m)\right)\\
&=\mathrm{Tr}\left(-\int_{-\frac{1}{2}}^{\frac{1}{2}}\left(\boldsymbol{W}(f)\boldsymbol{S}_{\boldsymbol{x}}(f)\right)^H\,df\right)\\
&=\mathrm{Tr}\left(-\int_{-\frac{1}{2}}^{\frac{1}{2}}\boldsymbol{S}_{\boldsymbol{x}}(f)\boldsymbol{W}^H(f)\,df\right),
\end{aligned}\tag{A.6}$$

where it has been used that the PSD matrix $\boldsymbol{S}_{\boldsymbol{x}}(f)$ is Hermitian.

Collecting all the expressions in Equations (A.2) to (A.6) and using that $\boldsymbol{W}(f)=\boldsymbol{R}(f)\boldsymbol{E}(f)$, the expression for the block MSE given in Equation (2.2) is obtained.

A.1.2 Block MSE for FIR Filter Banks

The block MSE for FIR filter banks for a given vector delay d_v and scalar delay d_s is defined by:

$$\begin{aligned}
\mathcal{E}_{N,M}(d_v,d_s)&=E\left[\|\hat{\boldsymbol{x}}(n)-\boldsymbol{x}_{d_s}(n-d_v)\|^2\right]\\
&=\mathrm{Tr}\left\{E\left[\left(\hat{\boldsymbol{x}}(n)-\boldsymbol{x}_{d_s}(n-d_v)\right)\left(\hat{\boldsymbol{x}}^H(n)-\boldsymbol{x}_{d_s}^H(n-d_v)\right)\right]\right\},
\end{aligned}\tag{A.7}$$

where N is the decimation factor used and M is the number of quantizers receiving a *positive* number of bits. The vector $\boldsymbol{x}_{d_s}(n)$ is defined in Equation (4.5). The output vector from the synthesis polyphase matrix is $\hat{\boldsymbol{x}}(n)$, see Figure 1.1, and it can be expressed by Equation (4.11). By inserting Equation (4.11) into

Equation (A.7), the block MSE can be expressed as:

$$\begin{aligned}\mathcal{E}_{N,M}(d_v,d_s) = \mathrm{Tr}\Big\{ & E\Big[\left(\boldsymbol{R}_{\text{⌊}}\boldsymbol{E}_{\text{⌐}}\boldsymbol{x}(n)_{\text{⌊}} + \boldsymbol{R}_{\text{⌊}}\boldsymbol{q}(n)_{\text{⌊}} - \boldsymbol{x}_{d_s}(n-d_v)\right)\cdot \\ & \left(\boldsymbol{x}^H(n)_{\text{⌊}}\boldsymbol{E}_{\text{⌐}}^H\boldsymbol{R}_{\text{⌊}}^H + \boldsymbol{q}^H(n)_{\text{⌊}}\boldsymbol{R}_{\text{⌊}}^H - \boldsymbol{x}_{d_s}^H(n-d_v)\right)\Big]\Big\} \\ = \mathrm{Tr}\Big\{ & \boldsymbol{R}_{\text{⌊}}\boldsymbol{E}_{\text{⌐}}E\left[\boldsymbol{x}(n)_{\text{⌊}}\boldsymbol{x}^H(n)_{\text{⌊}}\right]\boldsymbol{E}_{\text{⌐}}^H\boldsymbol{R}_{\text{⌊}}^H \\ & -\boldsymbol{R}_{\text{⌊}}\boldsymbol{E}_{\text{⌐}}E\left[\boldsymbol{x}(n)_{\text{⌊}}\boldsymbol{x}_{d_s}^H(n-d_v)\right] \\ & -\left(E\left[\boldsymbol{x}(n)_{\text{⌊}}\boldsymbol{x}_{d_s}^H(n-d_v)\right]\right)^H\boldsymbol{E}_{\text{⌐}}^H\boldsymbol{R}_{\text{⌊}}^H \\ & +E\left[\boldsymbol{x}_{d_s}(n-d_v)\boldsymbol{x}_{d_s}^H(n-d_v)\right] \\ & +\boldsymbol{R}_{\text{⌊}}E\left[\boldsymbol{q}(n)_{\text{⌊}}\boldsymbol{q}^H(n)_{\text{⌊}}\right]\boldsymbol{R}_{\text{⌊}}^H \\ & +\boldsymbol{R}_{\text{⌊}}\boldsymbol{E}_{\text{⌐}}E\left[\boldsymbol{x}(n)_{\text{⌊}}\boldsymbol{q}^H(n)_{\text{⌊}}\right]\boldsymbol{R}_{\text{⌊}}^H \\ & +\boldsymbol{R}_{\text{⌊}}\left(E\left[\boldsymbol{x}(n)_{\text{⌊}}\boldsymbol{q}^H(n)_{\text{⌊}}\right]\right)^H\boldsymbol{E}_{\text{⌐}}^H\boldsymbol{R}_{\text{⌊}}^H \\ & -\boldsymbol{R}_{\text{⌊}}E\left[\boldsymbol{q}(n)_{\text{⌊}}\boldsymbol{x}_{d_s}^H(n-d_v)\right] \\ & -\left(E\left[\boldsymbol{q}(n)_{\text{⌊}}\boldsymbol{x}_{d_s}^H(n-d_v)\right]\right)^H\boldsymbol{R}_{\text{⌊}}^H\Big\}. \end{aligned} \tag{A.8}$$

If the correlation matrices defined in Equations (4.8), (4.9), (4.10), (7.7), and (7.8) are substituted in the equation above, the block MSE in Equation (7.9) is obtained. If the white signal independent noise model in Section 4.1 is used, the last four terms of Equation (A.8) vanish, and the result of Equation (4.12) is obtained.

A.2 Bit Constraint Derivation

The bit constraint is derived for both the unconstrained length and FIR filter bank cases in this section.

A.2.1 Bit Constraint for Unconstrained Length Filter Banks

The bit constraint is expressed in Equation (1.15). By using the assumed quantization model in Equation (1.13), the bit constraint can be reformulated as

$$\prod_{i=0}^{M-1}\sigma_{y_i}^2 = 2^{2Nb}\cdot\prod_{i=0}^{M-1}\frac{\sigma_{q_i}^2}{c_i}. \tag{A.9}$$

The problem is to find an expression for the product on the left hand side of this equation. This can be done by defining the operator Pr as the product

of the diagonal elements on the main diagonal of the matrix to which the operator is applied. The matrix $E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]$ has $\sigma_{y_i}^2$ as the ith diagonal element. Therefore, by expressing $\boldsymbol{y}(n)$ as the convolution of the input vector time series and the polyphase analysis filter, the bit constraint can be expressed as

$$\Pr\left(E\left[\sum_m \boldsymbol{e}(n-m)\boldsymbol{x}(m)\sum_k \boldsymbol{x}^H(k)\boldsymbol{e}^H(n-k)\right]\right) = 2^{2Nb}\cdot\prod_{i=0}^{M-1}\frac{\sigma_{q_i}^2}{c_i}. \quad \text{(A.10)}$$

With some rearrangement, one obtains

$$\Pr\left(\sum_m\sum_k \boldsymbol{e}(n-m)E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(k)\right]\boldsymbol{e}^H(n-k)\right) = 2^{2Nb}\cdot\prod_{i=0}^{M-1}\frac{\sigma_{q_i}^2}{c_i}. \quad \text{(A.11)}$$

By taking the inverse Fourier transform of Equation (1.8) and inserting the result in Equation (A.11), one gets

$$\Pr\left(\sum_m\sum_k \boldsymbol{e}(n-m)\int_{-\frac{1}{2}}^{\frac{1}{2}}\boldsymbol{S}_{\boldsymbol{x}}(f)e^{j2\pi f(m-k)}\,df\,\boldsymbol{e}^H(n-k)\right) = 2^{2Nb}\cdot\prod_{i=0}^{M-1}\frac{\sigma_{q_i}^2}{c_i}. \quad \text{(A.12)}$$

Assuming convergence, one may rearrange the order of summation and integration as

$$\Pr\left(\int_{-\frac{1}{2}}^{\frac{1}{2}}\sum_m \boldsymbol{e}(n-m)e^{-j2\pi f(n-m)}\boldsymbol{S}_{\boldsymbol{x}}(f)\left(\sum_k \boldsymbol{e}(n-k)e^{-j2\pi f(n-k)}\right)^H df\right) = 2^{2Nb}\cdot\prod_{i=0}^{M-1}\frac{\sigma_{q_i}^2}{c_i}. \quad \text{(A.13)}$$

A change of summation variables and introduction of $\boldsymbol{E}(f)$, the Fourier transform of $\boldsymbol{e}(n)$, renders Equation (2.3), which is the desired result.

A.2.2 Bit Constraint for FIR Filter Banks

The problem of expressing the bit constraint in Equation (A.9) in the FIR case reduces to finding an expression for the product on the left hand side of the equation. This can be done by usage of the operator Pr defined in

Subsection A.2.1 and the matrix $E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]$ in the following way:

$$\begin{aligned}\prod_{i=0}^{M-1}\sigma_{y_i}^2 &= \Pr\left\{E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]\right\}\\ &= \Pr\left\{E\left[\boldsymbol{E}\_\boldsymbol{x}(n)_{\mid}\boldsymbol{x}^H(n)_{\mid}\boldsymbol{E}\_\right]\right\}\\ &= \Pr\left\{\boldsymbol{E}\_\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}\boldsymbol{E}_-^H\right\}. \end{aligned} \tag{A.14}$$

If this result is inserted in Equation (A.9), the bit constraint in the FIR case can be expressed as in Equation (4.13), where it has been assumed that the variances of the additive quantization noise are equal to σ_q^2. In Equation (A.14), the vector $\boldsymbol{x}(n)_{\mid}$ is an $(m+1)N \times 1$ vector.

A.3 Power Constraint Derivation

The power constraint is derived for both the unconstrained length and FIR filter bank case in this section.

A.3.1 Power Constraint for Unconstrained Length Filter Banks

In the unconstrained filter bank case, the power constraint in Equation (1.23) is rewritten by expressing the vector $\boldsymbol{y}(n)$ as the convolution of the input vector time series $\boldsymbol{x}(n)$ and the matrix impulse sequence $\boldsymbol{e}(n)$ of the transmitter filter bank:

$$\mathrm{Tr}\left(E\left[\sum_m \boldsymbol{e}(l-m)\boldsymbol{x}(m)\sum_p \boldsymbol{x}^H(p)\boldsymbol{e}^H(l-p)\right]\right) = P. \tag{A.15}$$

With some rearrangement, one obtains

$$\mathrm{Tr}\left(\sum_m\sum_p \boldsymbol{e}(l-m)E\left[\boldsymbol{x}(m)\boldsymbol{x}^H(p)\right]\boldsymbol{e}^H(l-p)\right) = P. \tag{A.16}$$

By taking the inverse Fourier transform of Equation (1.8) and putting the result in Equation (A.16), one gets

$$\mathrm{Tr}\left(\sum_m\sum_p \boldsymbol{e}(l-m)\int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{S}_{\boldsymbol{x}}(f)e^{j2\pi f(m-p)}\,df\,\boldsymbol{e}^H(l-p)\right) = P. \tag{A.17}$$

Another rearrangement in the order of summation and integration, assuming convergence, gives

$$\mathrm{Tr}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_m \boldsymbol{e}(l-m)e^{-j2\pi f(l-m)}\boldsymbol{S}_{\boldsymbol{x}}(f)\left(\sum_p \boldsymbol{e}(l-p)e^{-j2\pi f(l-p)}\right)^H df\right) = P. \tag{A.18}$$

A change of summation variables and the introduction of $\boldsymbol{E}(f)$, the Fourier transform of $\boldsymbol{e}(n)$, gives the power constraint in Equation (2.71).

A.3.2 Power Constraint for FIR Filter Banks

The problem of expressing the power constraint in Equation (1.23) in the FIR case reduces to finding an expression for the vector $\boldsymbol{y}(n)$ by means of the input vector time series. This can be done by $\boldsymbol{y}(n) = \boldsymbol{E}_{-}\boldsymbol{x}(n)_{|}$, where the vector $\boldsymbol{x}(n)_{|}$ is an $(m+1)N \times 1$ vector. If the vector $\boldsymbol{y}(n)$ is substituted in the left hand side of Equation (1.23) the following result is obtained:

$$\begin{aligned}\mathrm{Tr}\left\{E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]\right\} &= \mathrm{Tr}\left\{E\left[\boldsymbol{E}_{-}\boldsymbol{x}(n)_{|}\boldsymbol{x}^H(n)_{|}\boldsymbol{E}_{-}\right]\right\} \\ &= \mathrm{Tr}\left\{\boldsymbol{E}_{-}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}\boldsymbol{E}_{-}^H\right\}. \end{aligned} \tag{A.19}$$

If this result is inserted in Equation (1.23) the power constraint in the FIR case can be expressed as in Equation (4.24).

Appendix B

Ordering Functions and Eigenvalues of the PSD Matrix

In this appendix, the eigenvalues of the PSD matrix $\boldsymbol{S_x}(f)$ are derived and the ordering functions are introduced. Some properties of the ordering functions are also derived.

B.1 Eigenvalues of the PSD Matrix

Let $\lambda_i^{(N)}(f)$ be the ith largest eigenvalue of the $N \times N$ PSD matrix $\boldsymbol{S_x}(f)$, defined in Equation (1.8). In [Sathe & Vaidyanathan 1993], it is shown that $\boldsymbol{S_x}(f)$ is a pseudocirculant matrix when the time series $x(n)$ is WSS. Furthermore, in [Vaidyanathan & Mitra 1988], it is shown that the eigenvalues of a pseudocirculant matrix are given by

$$\lambda_i^{(N)}(f) = \sum_{k=0}^{N-1} \zeta_k^{(N)}(f)\ e^{-\frac{j2\pi k\left(f+l_i^{(N)}(f)\right)}{N}}, \tag{B.1}$$

where $\zeta_k^{(N)}(f)$ is the kth element in the first row of $\boldsymbol{S_x}(f)$ and the function $l_i^{(N)}(f)$ is the *ordering function*

$$l_i^{(N)} : \mathbb{R} \longrightarrow \mathbb{Z}_N, \quad i \in \mathbb{Z}_N, \tag{B.2}$$

where the set $\mathbb{Z}_N$ is defined as $\mathbb{Z}_N = \{0, 1, \ldots, N-1\}$. The ordering functions $l_i^{(N)}(f)$ were not used in [Vaidyanathan & Mitra 1988], but are introduced here to ensure that the ordering of the eigenvalues given in Equation (2.7) is maintained for all frequencies f.

The goal is to find $\lambda_i^{(N)}(f)$ expressed as a function of $S_x(f)$. From the definition of $\boldsymbol{S_x}(f)$, see Equation (1.8), one obtains

$$\zeta_k^{(N)}(f) = \sum_{m=-\infty}^{\infty} R_x(mN+k)\ e^{-j2\pi fm},\ \ k \in \mathbb{Z}_N, \tag{B.3}$$

where $R_x(n)$ is the autocorrelation function of the input time series $x(n)$. Thus, $\zeta_k^{(N)}(f)$ is the Fourier transform of a decimated version of the time series $R_x(m+k)$, which in turn has the Fourier transform $e^{j2\pi fk} \cdot S_x(f)$ [Proakis & Manolakis 1992]. Therefore, the Fourier transform of the decimated sequence $R_x(mN+k)$ is

$$\begin{aligned}\zeta_k^{(N)}(f) &= \frac{1}{N}\sum_{m=0}^{N-1} e^{j2\pi(\frac{f-m}{N})k} S_x\left(\frac{f-m}{N}\right) \\ &= e^{j2\pi f\frac{k}{N}}\frac{1}{N}\sum_{m=0}^{N-1} e^{-\frac{j2\pi mk}{N}} S_x\left(\frac{f-m}{N}\right).\end{aligned} \tag{B.4}$$

By inserting the result from Equation (B.4) into Equation (B.1), one gets

$$\begin{aligned}\lambda_i^{(N)}(f) &= \sum_{k=0}^{N-1} \zeta_k^{(N)}(f)\ \ e^{-\frac{j2\pi k\left(f+l_i^{(N)}(f)\right)}{N}} \\ &= \sum_{k=0}^{N-1} e^{j2\pi f\frac{k}{N}}\frac{1}{N}\sum_{m=0}^{N-1} e^{-\frac{j2\pi mk}{N}} S_x\left(\frac{f-m}{N}\right)\ \ e^{-\frac{j2\pi k\left(f+l_i^{(N)}(f)\right)}{N}} \\ &= \sum_{k=0}^{N-1}\frac{1}{N}\sum_{m=0}^{N-1} e^{-\frac{j2\pi mk}{N}} S_x\left(\frac{f-m}{N}\right)\ \ e^{-\frac{j2\pi kl_i^{(N)}(f)}{N}} \\ &= \sum_{m=0}^{N-1} S_x\left(\frac{f-m}{N}\right)\sum_{k=0}^{N-1}\frac{1}{N}\ \ e^{-\frac{j2\pi k\left(m+l_i^{(N)}(f)\right)}{N}} \\ &= \sum_{m=0}^{N-1} S_x\left(\frac{f-m}{N}\right)\delta\left(m+l_i^{(N)}(f)\right) \\ &= S_x\left(\frac{f+l_i^{(N)}(f)}{N}\right),\ \ i \in \mathbb{Z}_N,\end{aligned} \tag{B.5}$$

where $\delta(\cdot)$ is the Krönecker delta function. The values of the ordering functions can be found by Equation (B.5), while making sure that the ordering in Equation (2.7) is maintained.

B.2 Properties of the Ordering Functions

In this section, some of the properties of the ordering functions will be stated and proven.

For a fixed frequency, the N ordering functions will take each of the N numbers in the set $\mathbb{Z}_N$ only once. Therefore, the sum of all the ordering functions for a fixed frequency is given by:

$$\sum_{i=0}^{N-1} l_i^{(N)}(f) = \sum_{i=0}^{N-1} i = \frac{(N-1)N}{2} \quad \forall f. \tag{B.6}$$

From Equation (1.8), it is seen that $\boldsymbol{S_x}(f+1) = \boldsymbol{S_x}(f)$, i.e., the PSD matrix $\boldsymbol{S_x}(f)$ has a period equal to 1. Since $\lambda_i^{(N)}(f)$ is the eigenvalue of the PSD matrix $\boldsymbol{S_x}(f)$, the eigenvalue functions must also have period 1. $S_x(f)$ represents a discrete time signal, and therefore, $S_x(f+1) = S_x(f)$, which means that the period of $S_x(f)$ is also 1. By studying Equation (B.5) and imposing that both the functions $\lambda_i^{(N)}(f)$ and $S_x(f)$ should have period 1, the only possibility is

$$l_i^{(N)}(f+1) + 1 = l_i^{(N)}(f) + Nk, \quad k \in \mathbb{Z}. \tag{B.7}$$

This result leads to

$$l_i^{(N)}(f+p) + p = l_i^{(N)}(f) + Nk, \quad k, p \in \mathbb{Z}. \tag{B.8}$$

Since the function $l_i^{(N)}(f)$ returns values in $\mathbb{Z}_N$, it follows from Equation (B.8) that

$$l_i^{(N)}(f+N) = l_i^{(N)}(f), \tag{B.9}$$

which means that the ordering functions have period N.

Assume that the input time series is real. Then the PSD function $S_x(f)$ is symmetric, that is $S_x(-f) = S_x(f)$ [Proakis & Manolakis 1992]. It will now be shown that the following property holds:

$$l_i^{(N)}(f) + l_i^{(N)}(-f) = 0 \pmod N, \tag{B.10}$$

where $(\mathrm{mod}\, N)$ means that the arithmetic is modulo N [Judson 1994].

By using the definition $\boldsymbol{K_x}(m) = E\left[\boldsymbol{x}(n+m)\boldsymbol{x}^H(n)\right]$ and the assumption that the input is real, it can be shown that $\boldsymbol{K}_{\boldsymbol{x}}^T(-m) = \boldsymbol{K_x}(m)$. This can again be used to show that for real input signals:

$$\boldsymbol{S}_{\boldsymbol{x}}^T(-f) = \boldsymbol{S_x}(f). \tag{B.11}$$

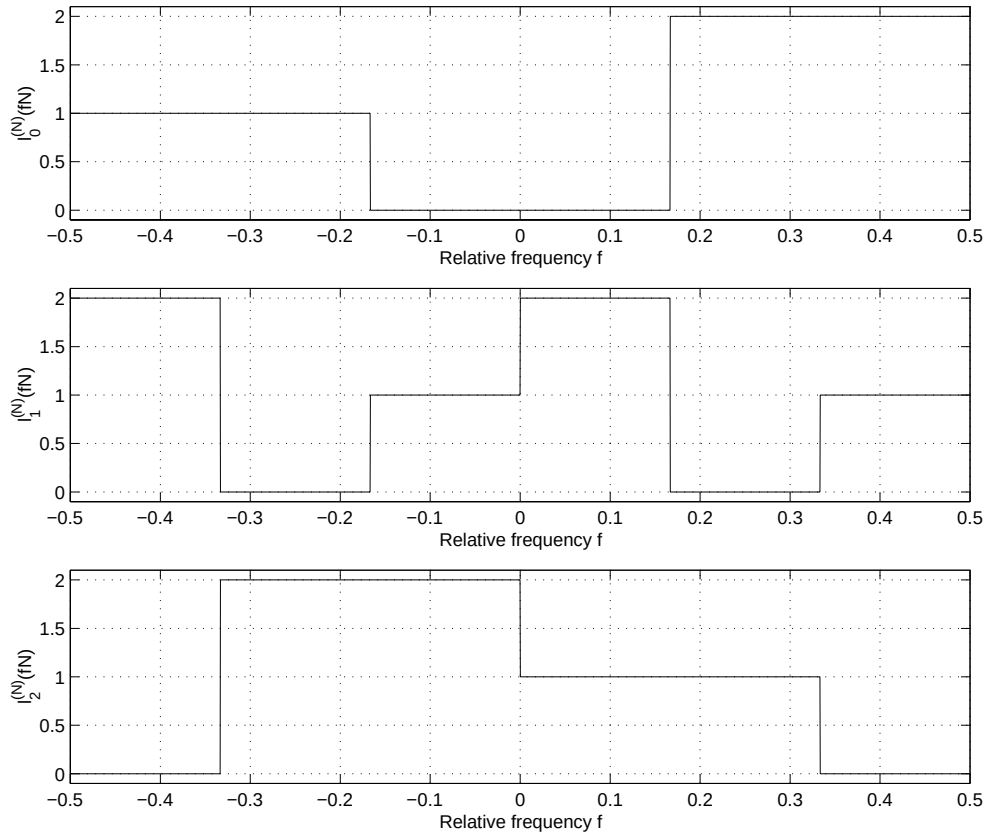

Figure B.1 Ordering functions $l_i(fN)$ as a function of frequency f for $N = 3$, where the input PSD $S_x(f)$ is an AR(1) with correlation coefficient 0.95.

Since the transpose of a matrix and the matrix itself have the same eigenvalues [Kreyszig 1988], it follows from the result in Equation (B.11) that

$$\lambda_i^{(N)}(-f) = \lambda_i^{(N)}(f) \ \ \forall f. \tag{B.12}$$

The result in Equation (B.12) combined with $S_x(-f) = S_x(f)$ leads to:

$$S_x\left(\frac{-f + l_i^{(N)}(-f)}{N}\right) = S_x\left(\frac{f + l_i^{(N)}(f)}{N}\right) = S_x\left(\frac{-f - l_i^{(N)}(f)}{N}\right). \tag{B.13}$$

Since the PSD function $S_x(f)$ has period 1, the only possibility for Equa-

tion (B.13) to be valid is that

$$\frac{-f + l_i^{(N)}(-f)}{N} = \frac{-f - l_i^{(N)}(f)}{N} + p, \quad p \in \mathbb{Z}, \tag{B.14}$$

and with some rearrangements, this leads to the desired result in Equation (B.10).

For real input signals, it is sufficient to decide the values of the ordering functions $l_i^{(N)}(f)$ in the frequency interval $\left(0, \frac{1}{2}\right]$. The reason can be seen from Equations (B.7), (B.9), and (B.10).

As an example of how the ordering function may look, the ordering functions $l_i^{(N)}(fN)$, which have period 1, are shown in Figure B.1. Here, $N = 3$ and the input PSD is an AR(1) source with correlation coefficient equal to 0.95. From the figure, it can be observed that the properties given in this appendix are satisfied by the ordering functions for all frequencies except for frequencies where the ordering functions are discontinuous. At these discontinuous points, the ordering functions are undefined. Since these functions are used to evaluate the SNR vs. rate or CSNR performance through the calculation of integrals of these functions, the performance is unaffected by the fact that the ordering functions are undefined in a countable number of frequencies. From Equations (2.61) and (2.62), it is seen that the frequency regions where the unconstrained length filters are different from zero are given by the frequency regions where the ordering functions $l_i^{(N)}(fN)$ are equal to zero.

The properties of the ordering functions shown here are in agreement with the results reported in [Unser 1993, Vaidyanathan 1998] through the way the passbands for optimal unconstrained length unitary filter banks should be chosen. The passbands of optimal unconstrained length biorthogonal filter banks should also be chosen according to the ordering function introduced in this appendix [Aas & Mullis 1996, Vaidyanathan & Kıraç 1998, Moulin et al. 2000]. The ordering functions introduced in this appendix are related to the integer valued functions in [Sathe & Vaidyanathan 1993], which are used to study pseudocirculant matrices.

Appendix C

Matrix Variational Calculus and Differentiation

In this appendix, some results of matrix variational calculus and differentiation will be included, which are used in solving some of the optimization problems in Chapters 2, 4, and 7.

C.1 Matrix Variational Calculus

In the unconstrained length filter bank case, the unknowns are *functions* of frequency. The optimization of these can be found through variational calculus with respect to the elements in the matrices.

The Gâteaux variation [Troutman 1996] is needed in the optimization of the unconstrained length filter banks. The use of this in the unconstrained length case for continuous time filters can be seen in [Amitay & Salz 1984, Salz 1985]. In [Yang & Roy 1994], a related result to what will be found in this section was given, but no derivation of the result was included.

The Lagrange function for the problem of minimizing the block MSE in Equation (2.8), subject to the equality constraint in Equation (2.9) and the

inequality constraints in Equation (2.4), can be expressed as:

$$\begin{aligned}\mathcal{L} = &\sum_{n=0}^{N-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{m=0}^{N-1} \left| \delta(n-m) - \sum_{k=0}^{M-1} T_{n,k}(f) G_{k,m}(f) \right|^2 \lambda_m^{(N)}(f)\, df \\ &+ \sum_{n=0}^{N-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{m=0}^{M-1} |T_{n,m}(f)|^2 \, \sigma_{q_m}^2 \, df \\ &+ \mu \sum_{m=0}^{M-1} \ln \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{n=0}^{N-1} |G_{m,n}(f)|^2 \, \lambda_n^{(N)}(f)\, df \right) \\ &- \sum_{m=0}^{M-1} \theta_m \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{n=0}^{N-1} |G_{m,n}(f)|^2 \, \lambda_n^{(N)}(f)\, df, \end{aligned} \tag{C.1}$$

where $T_{n,m}(f)$ is the element in row number n and column number m of the matrix $\boldsymbol{T}(f)$. $G_{m,n}(f)$ is the element in row number m and column number n of the matrix $\boldsymbol{G}(f)$ and $\lambda_m^{(N)}(f)$ is diagonal element number i of the diagonal matrix $\boldsymbol{\Lambda}_{\boldsymbol{x}}(f)$. In Equation (C.1), $\mu \in \mathbb{R}^+ = (0, \infty)$ is a Kuhn-Tucker parameter for the equality constraint (2.9), and $\theta_i \geq 0$ are the Kuhn-Tucker parameters for the inequality constraints in Equation (2.4).

If the Gâteaux variation [Troutman 1996] of $\mathcal{L}$ is calculated with respect to $T_{i,j}(f)$ in the direction of $V(f)$ and set equal to zero, the following result is obtained:

$$\begin{aligned}\delta_g \mathcal{L}\left(T_{i,j}(f); V(f)\right) = \int_{-\frac{1}{2}}^{\frac{1}{2}} 2\,\mathrm{Re} \Bigg\{ \Bigg[&\sum_{n=0}^{N-1} \lambda_n^{(N)}(f) G_{j,n}(f) \\ &\left(-\delta(i-n) + \sum_{m=0}^{M-1} T_{i,m}^*(f) G_{m,n}^*(f) \right) + T_{i,j}^*(f) \sigma_{q_j}^2 \Bigg] V(f) \Bigg\}\, df = 0, \quad \forall V(f), \end{aligned} \tag{C.2}$$

where δ_g is the Gâteaux variation operator. Operator Re returns the real value part, and the superscript * means complex conjugation. In order to find the optimal value, Equation (C.2) must be satisfied for all functions $V(f)$.

Since Equation (C.2) must hold for all values of $V(f)$, the only possibility is that

$$\sum_{n=0}^{N-1} \lambda_n^{(N)}(f) G_{j,n}(f) \left(-\delta(i-n) + \sum_{m=0}^{M-1} T_{i,m}^*(f) G_{m,n}^*(f) \right) + T_{i,j}^*(f) \sigma_{q_j}^2 = 0, \tag{C.3}$$

for all f and this equation must be valid for $i \in \{0, 1, \ldots, N-1\}$ and $j \in \{0, 1, \ldots, M-1\}$. If the equations are manipulated and written in matrix form, the following expression is obtained:

$$\boldsymbol{T}(f)\left(\boldsymbol{G}(f)\boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f) + \boldsymbol{\Lambda_q}\right) = \boldsymbol{\Lambda_x}(f)\boldsymbol{G}^H(f), \tag{C.4}$$

which is equivalent to Equation (2.11). The necessary conditions with respect to the synthesis polyphase matrix are now found, but the necessary conditions for the analysis matrix remain to be found. These can be obtained by finding the Gâteaux variation of $\mathcal{L}$ with respect to $G_{m,n}(f)$ in the direction of $V(f)$, and then setting the result equal to zero. If this is done, the following result is obtained:

$$\begin{aligned}
&\delta_g \mathcal{L}\left(G_{m,n}(f); V(f)\right) = \\
&\int_{-\frac{1}{2}}^{\frac{1}{2}} 2\,\mathrm{Re}\left\{\left[\sum_{k=0}^{N-1} \lambda_n^{(N)}(f) T_{k,m}(f)\left(-\delta(k-n) + \sum_{i=0}^{M-1} T_{k,i}^*(f) G_{i,n}^*(f)\right)\right.\right. \\
&\left.\left. + \mu\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{j=0}^{N-1} \left|G_{m,j}(f)\right|^2 \lambda_j^{(N)}(f)\, df\right)^{-1} G_{m,n}^*(f)\lambda_n^{(N)}(f)\right] V(f)\right\} df = 0,
\end{aligned} \tag{C.5}$$

for all $V(f)$. Since the equation must hold for all values of $V(f)$, the only possibility is that

$$\begin{aligned}
&\sum_{k=0}^{N-1} \lambda_n^{(N)}(f) T_{k,m}(f)\left(-\delta(k-n) + \sum_{i=0}^{M-1} T_{k,i}^*(f) G_{i,n}^*(f)\right) \\
&\quad + \mu\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{j=0}^{N-1} \left|G_{m,j}(f)\right|^2 \lambda_j^{(N)}(f)\, df\right)^{-1} G_{m,n}^*(f)\lambda_n^{(N)}(f) = 0,
\end{aligned} \tag{C.6}$$

for all f and this equation must be valid for $m \in \{0, 1, \ldots, M-1\}$ and $n \in \{0, 1, \ldots, N-1\}$. These equations can be rearranged in matrix form, which gives:

$$\boldsymbol{T}^H(f)\boldsymbol{T}(f)\boldsymbol{G}(f) + \mu\boldsymbol{\Sigma_y}^{-1}\boldsymbol{G}(f) - \boldsymbol{\Theta}\boldsymbol{G}(f) = \boldsymbol{T}^H(f) \tag{C.7}$$

where the matrix $\boldsymbol{\Sigma_y}$ is an $M \times M$ diagonal matrix with diagonal element number m given by Equation (2.13). The matrix $\boldsymbol{\Theta}$ is a diagonal matrix where diagonal element number i is θ_i. Equation (C.7) is equivalent to Equation (2.12).

C.2 Matrix Differentiation Results

In the FIR filter bank case, the unknown variables are the elements of the impulse response functions. These independent variables can be found through optimization in the ordinary way, i.e., through differentiation.

Differentiation of a scalar function v with respect to a matrix $\boldsymbol{A}$ can be defined [Graham 1981] as another matrix $\boldsymbol{B}$ with the same dimension as the matrix $\boldsymbol{A}$. The matrix element in row number i and column number j of the matrix $\boldsymbol{B}$ is given by the ordinary scalar differentiation of the scalar function v with respect to the element in row number i and column number j of the matrix $\boldsymbol{A}$.

Many matrix differentiation results can be found in [Magnus & Neudecker 1988, Lütkepohl 1996]. However, most of the results given in this section have *not* been found in the literature. The detailed derivation of the results will not be given here, only the final results will be shown.

In order to differentiate the objective function in both the bit and power constrained FIR cases with respect to the matrices $\boldsymbol{E}_{\_}$ and $\boldsymbol{R}_{\_}$, the results in this section are needed.

To differentiate the block MSE with respect to $\boldsymbol{E}_{\_}$, the following three results are needed:

$$\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\left\{\boldsymbol{R}_{\_}\boldsymbol{E}_{\sqcap}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\sqcap}^{H}\boldsymbol{R}_{\_}^{H}\right\} = 2\boldsymbol{R}_{|}^{H}\mathcal{T}\left\{\boldsymbol{R}_{\_}\boldsymbol{E}_{\sqcap}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\right\}, \tag{C.8}$$

where the operator $\mathcal{T}$ is defined in Equation (4.17),

$$\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\left\{\boldsymbol{R}_{\_}\boldsymbol{E}_{\sqcap}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right\} = \boldsymbol{R}_{|}^{H}\mathcal{T}\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^{H}\right\}, \tag{C.9}$$

and

$$\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^{H}\boldsymbol{E}_{\sqcap}^{H}\boldsymbol{R}_{\_}^{H}\right\} = \boldsymbol{R}_{|}^{H}\mathcal{T}\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^{H}\right\}. \tag{C.10}$$

These three results are found from the definition of the differentiation of a scalar with respect to a matrix, stated above.

To include the bit constraint in the optimization, the following result is needed:

$$\frac{\partial}{\partial \boldsymbol{E}_{\_}} \sum_{i=0}^{M-1} \ln \sigma_{y_i}^2 = 2\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\boldsymbol{E}_{\_}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}. \tag{C.11}$$

The result in Equation (C.11) is used to include the equality constraint given in Equation (4.13) in the optimization of the FIR analysis filter bank through a Lagrange multiplier.

To include the inequality constraints given in Equations (2.4) in the optimization of the analysis filter bank, the following result will be useful:

$$\frac{\partial}{\partial \boldsymbol{E}_{\_}} \sum_{i=0}^{M-1} \theta_i \sigma_{y_i}^2 = 2\boldsymbol{\Theta}\boldsymbol{E}_{\_}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}, \tag{C.12}$$

where the matrix $\boldsymbol{\Theta}$ is an $M \times M$ diagonal matrix with θ_i as diagonal element number i.

To find equations for an optimal synthesis filter bank for a given analysis filter bank, the following three results will be needed:

$$\frac{\partial}{\partial \boldsymbol{R}_{\_}} \operatorname{Tr}\left\{\boldsymbol{R}_{\_}\boldsymbol{E}_{\urcorner}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\urcorner}^H\boldsymbol{R}_{\_}^H\right\} = 2\boldsymbol{R}_{\_}\boldsymbol{E}_{\urcorner}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\urcorner}^H, \tag{C.13}$$

$$\frac{\partial}{\partial \boldsymbol{R}_{\_}} \operatorname{Tr}\left\{\boldsymbol{R}_{\_}\boldsymbol{E}_{\urcorner}\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right\} = \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^H \boldsymbol{E}_{\urcorner}^H, \tag{C.14}$$

and

$$\frac{\partial}{\partial \boldsymbol{R}_{\_}} \operatorname{Tr}\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^H \boldsymbol{E}_{\urcorner}^H\boldsymbol{R}_{\_}^H\right\} = \left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+l,N)}(d_v,d_s)\right)^H \boldsymbol{E}_{\urcorner}^H. \tag{C.15}$$

The results in Equations (C.13), (C.14), and (C.15) can be found in [Magnus & Neudecker 1988].

To optimize the transmitter polyphase matrix for power constrained FIR filter banks, the following result is needed:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\Big\{\boldsymbol{R}_{\_}(\boldsymbol{C}_{\_}\boldsymbol{E}_{\urcorner})_{\diagdown}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}\left((\boldsymbol{C}_{\_}\boldsymbol{E}_{\urcorner})_{\diagdown}\right)^H\boldsymbol{R}_{\_}^H\Big\}& \\ = 2\boldsymbol{C}_{\shortmid}^H\mathcal{T}_2\left\{\boldsymbol{R}_{\shortmid}^H\mathcal{T}_1\left\{\boldsymbol{R}_{\_}(\boldsymbol{C}_{\_}\boldsymbol{E}_{\urcorner})_{\diagdown}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+o+l,N)}\right\}\right\},&\end{aligned} \tag{C.16}$$

where the operator $\mathcal{T}_1 : \mathbb{R}^{N\times(m+o+l+1)N} \rightarrow \mathbb{R}^{(l+1)N\times(m+o+1)N}$ produces an $(l+1)N \times (m+o+1)N$ block Toeplitz matrix from an $N \times (m+o+l+1)N$ matrix in the same way as shown in Equation (4.17), but where m is replaced by $m+o$. The operator $\mathcal{T}_2 : \mathbb{R}^{M\times(m+o+1)N} \rightarrow \mathbb{R}^{(o+1)M\times(m+1)N}$ produces an $(o+1)M \times (m+1)N$ block Toeplitz matrix from an $M \times (m+o+1)N$ matrix in the following way: Let $\boldsymbol{W}_{\_}$ be an $M \times (m+o+1)N$ matrix where the ith

$M \times N$ block is given by $[\boldsymbol{W}_{\_}]_i$, $i \in \{0, 1, \ldots, m+o\}$. Then the operator $\mathcal{T}_2$ is defined as follows:

$$\mathcal{T}_2\{\boldsymbol{W}_{\_}\} = \begin{bmatrix} [\boldsymbol{W}_{\_}]_o & [\boldsymbol{W}_{\_}]_{o+1} & \cdots & [\boldsymbol{W}_{\_}]_{m+o} \\ \vdots & \vdots & \ddots & \vdots \\ [\boldsymbol{W}_{\_}]_1 & [\boldsymbol{W}_{\_}]_2 & \cdots & [\boldsymbol{W}_{\_}]_{m+1} \\ [\boldsymbol{W}_{\_}]_0 & [\boldsymbol{W}_{\_}]_1 & \cdots & [\boldsymbol{W}_{\_}]_m \end{bmatrix}. \tag{C.17}$$

The following two results are also needed to find the equations for an optimal transmitter filter bank in the power constrained FIR case for a given synthesis filter bank:

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\left\{\boldsymbol{R}_{\_}(\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\diagdown}\, \boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v, d_s)\right\} \\ &\qquad = \boldsymbol{C}_{\lfloor}^H \mathcal{T}_2\left\{\boldsymbol{R}_{\lfloor}^H \mathcal{T}_1\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v, d_s)\right)^H\right\}\right\} \end{aligned} \tag{C.18}$$

and

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v, d_s)\right)^H \left((\boldsymbol{C}_{\_}\boldsymbol{E}_{\lceil})_{\diagdown}\right)^H \boldsymbol{R}_{\_}^H\right\} \\ &\qquad = \boldsymbol{C}_{\lfloor}^H \mathcal{T}_2\left\{\boldsymbol{R}_{\lfloor}^H \mathcal{T}_1\left\{\left(\boldsymbol{\phi}_{\boldsymbol{x}}^{(m+o+l,N)}(d_v, d_s)\right)^H\right\}\right\}. \end{aligned} \tag{C.19}$$

These two results can be used to derive the result in Equation (C.16).

To include the power constraint in the optimization, the following result is needed:

$$\frac{\partial}{\partial \boldsymbol{E}_{\_}} \operatorname{Tr}\left\{\boldsymbol{E}_{\_}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}\boldsymbol{E}_{\_}^H\right\} = 2\boldsymbol{E}_{\_}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m,N)}. \tag{C.20}$$

The last result can be found in [Magnus & Neudecker 1988].

Appendix D

Correlation Matrix Elements

In this appendix, formulas for finding the matrix elements of the block Toeplitz matrices $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ and $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ will be given. These formulas are needed to include the correlation between the input signal and the additive quantization noise in Section 7.2. In [Korn 1965, Sripad & Snyder 1977, Gosse & Duhamel 1997], midpoints were used as the representation levels in the uniform threshold quantizers. The results developed are an extension of these results to include other types of representation levels in the uniform threshold quantizers. In the theory developed in this appendix, the representation levels can be chosen as an arbitrary even function.

It is assumed that the input time series is Gaussian. Since the analysis filter bank is a linear operator, the subband signals are also Gaussian. The results will be derived by first finding the pdf that can be used to find the autocorrelation or the cross-correlation terms, then the characteristic function will be found, and finally, the correlation function will be found by differentiation of the characteristic function. This is the same procedure used in the derivation of similar formulas in [Korn 1965, Sripad & Snyder 1977].

Let $a_k^{(i)}$ be defined as follows

$$a_k^{(i)} = r_k^{(i)} - k\Delta_i. \tag{D.1}$$

$k\Delta_i$ is the midpoint of the kth decision interval. Therefore, $a_k^{(i)}$ is the distance between the representation level $r_k^{(i)}$ and the midpoint $k\Delta_i$ in decision interval number k in the ith uniform threshold quantizer. For midpoints as representation levels, $r_k^{(i)} = k\Delta_i$, thus with midpoints as representation levels, $a_k^{(i)} = 0$.

D.1 Elements of the $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ Matrix

The matrix $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ is an $(m+l+1)N \times (l+1)M$ matrix, and to calculate the element in row number $nN+i$ and column number $pM+k$ of this matrix, where $i \in \{0,1,\ldots,N-1\}$, $k \in \{0,1,\ldots,M-1\}$, $n \in \{0,1,\ldots,m+l\}$, and $p \in \{0,1,\ldots,l\}$, the joint pdf of the stochastic variables $x_i(n)$ and $q_k(p)$ is needed. By generalizing the results in [Korn 1965, Sripad & Snyder 1977], it can be shown that the joint pdf of the stochastic variables $x_i(n)$ and $q_k(p)$ is given by:

$$f_{x_i(n),q_k(p)}(a,b) = \sum_{p_1=-\infty}^{\infty} f_{x_i(n),y_k(p)}\left(a, r_{p_1}^{(k)} - b\right) \operatorname{rect}\left(\frac{b - \left(r_{p_1}^{(k)} - p_1\Delta_k\right)}{\Delta_k}\right), \tag{D.2}$$

where the function $\operatorname{rect}\colon \mathbb{R} \to \mathbb{R}$ is the rectangular function [Haykin 1983] defined by

$$\operatorname{rect}(x) = \begin{cases} 1, & \text{if } |x| \leq \frac{1}{2}, \\ 0, & \text{if } |x| > \frac{1}{2}. \end{cases} \tag{D.3}$$

In Equation (D.2), the symbol $f_{x_i(n),y_k(p)}$ represents the joint Gaussian pdf of the stochastic variables $x_i(n)$ and $y_k(p)$.

The characteristic function $\chi_{x_i(n),q_k(p)}(a,b)$ can be found (except for a sign in the exponent) as the Fourier transform of the pdf [Papoulis 1991] in Equation (D.2), and it can be shown that it is given by:

$$\chi_{x_i(n),q_k(p)}(a,b) = \frac{\Delta_k}{2\pi} \int_{-\infty}^{\infty} \sum_{p_1=-\infty}^{\infty} \cos\left(b r_{p_1}^{(k)} - \nu p_1 \Delta_k\right) \chi_{x_i(n),y_k(p)}(a,\nu-b) \frac{\sin\frac{\nu\Delta_k}{2}}{\frac{\nu\Delta_k}{2}} d\nu, \tag{D.4}$$

where $\chi_{x_i(n),y_k(p)}$ is the characteristic function of the joint Gaussian variables $x_i(n)$ and $y_k(p)$, which can be found in standard references on statistics [Papoulis 1991].

The element in row number $nN+i$ and column number $pM+k$ in the matrix $\boldsymbol{\Phi}_{\boldsymbol{x},\boldsymbol{q}}^{(m,l,N,M)}$ can be found by the moment theorem [Papoulis 1991]:

$$E\left[x_i(n) q_k(p)\right] = \frac{1}{j^2} \frac{\partial^2}{\partial a \partial b} \chi_{x_i(n),q_k(p)}(a,b)\Big|_{a=0,b=0}. \tag{D.5}$$

By performing the calculations in Equation (D.5), it can be shown that the following result is obtained

$$E\left[x_i(n)q_k(p)\right] = 2R_{x_i,y_k}(n-p)\sum_{p_1=1}^{\infty}\left((-1)^{p_1}e^{-\frac{2\pi^2 {p_1}^2\sigma_{y_k}^2}{\Delta_k^2}}\right.$$
$$\left.+a_{p_1}^{(k)}\sqrt{\frac{2}{\pi\sigma_{y_k}^2}}e^{-\frac{\Delta_k^2\left(4{p_1}^2+1\right)}{8\sigma_{y_k}^2}}\sinh\left(\frac{\Delta_k^2 p_1}{2\sigma_{y_k}^2}\right)\right), \quad \text{(D.6)}$$

where $R_{x_i,y_k}(n-p) = E\left[x_i(n)y_k(p)\right]$ is the element in row number $nN+i$ and column number $pM+k$ of the matrix $E\left[\boldsymbol{x}(n)_{\mathsf{I}}\boldsymbol{y}^H(n)_{\mathsf{I}}\right] = \boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^H$.

If $a_p^{(k)} = 0$ and $i = k$ in Equation (D.6), the result corresponds to Equation (23) in [Gosse & Duhamel 1997]. Therefore, Equation (D.6) is a generalization of the result in [Gosse & Duhamel 1997] to include arbitrary symmetric representation levels and cross-correlation between *different* subbands.

D.2 Elements of the $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ Matrix

By assuming that the input to the quantizer shown in Figure 6.2 is Gaussian with zero mean and input variance $\sigma_{y_i}^2$, it is possible to calculate the pdf of the additive quantization noise [Papoulis 1991]:

$$f_{q_i}(a) = \sum_{p=-\infty}^{\infty} f_{y_i}\left(r_p^{(i)} - a\right)\operatorname{rect}\left(\frac{a-\left(r_p^{(i)}-p\Delta_i\right)}{\Delta_i}\right). \quad \text{(D.7)}$$

The characteristic function of q_i is found (except for a sign in the exponent) by Fourier transform of the pdf in Equation (D.7), and it can be shown that it is given by

$$\chi_{q_i}(a) = \frac{\Delta_i}{2\pi}\int_{-\infty}^{\infty}\sum_{p=-\infty}^{\infty}\cos\left(ar_p^{(i)} - \nu p\Delta_i\right)\chi_{y_i}(\nu-a)\frac{\sin\frac{\nu\Delta_i}{2}}{\frac{\nu\Delta_i}{2}}d\nu, \quad \text{(D.8)}$$

where χ_{y_i} is the characteristic function of the Gaussian variable y_i, and it is fully specified by the mean, which is assumed to be zero, and the variance of the stochastic variable $y_i(n)$, which is denoted $\sigma_{y_i}^2$. The characteristic function of a Gaussian variable can be found in [Papoulis 1991].

From the characteristic function $\chi_{q_i}(a)$, it is possible to find the variance of the additive quantization noise by the moment theorem. By performing the

calculations, it can be shown that it is given by:

$$\begin{aligned}\sigma_{q_i}^2 =& \frac{\Delta_i^2}{12} + \frac{\Delta_i^2}{12}\sum_{k=1}^{\infty}\frac{(-1)^k}{k^2}e^{-\frac{2\pi^2k^2\sigma_{y_i}^2}{\Delta_i^2}} \\ &- 4\sqrt{\frac{2\sigma_{y_i}^2}{\pi}}\sum_{k=1}^{\infty}a_k^{(i)}e^{-\frac{\Delta_i^2\left(4k^2+1\right)}{8\sigma_{y_i}^2}}\sinh\left(\frac{\Delta_i^2 k}{2\sigma_{y_i}^2}\right) \\ &+ \sum_{k=1}^{\infty}\left(2k\Delta_i a_k^{(i)} + \left(a_k^{(i)}\right)^2\right) \\ &\qquad \cdot\left(\operatorname{erf}\left(\frac{2k+1}{2}\frac{\Delta_i}{\sqrt{2\sigma_{y_i}^2}}\right) - \operatorname{erf}\left(\frac{2k-1}{2}\frac{\Delta_i}{\sqrt{2\sigma_{y_i}^2}}\right)\right),\end{aligned} \tag{D.9}$$

where erf is the error function [Abramowitz & Stegun 1972]. If $a_k^{(i)} = 0$ is substituted in the equation, only the first two parts of the right hand side remain, and this is the same result as Equation (20) in [Gosse & Duhamel 1997], where midpoints were used as representation levels in the quantizers.

Diagonal element number $pM + i$ in the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ is equal to the variance of the additive quantization noise in subband number i. As an alternative expression, the quantization noise variance of a uniform threshold quantizer can be found the ordinary way [Gersho & Gray 1992]

$$\sigma_{q_i}^2 = \sum_{k=-\infty}^{\infty}\int_{\frac{2k-1}{2}\Delta_i}^{\frac{2k+1}{2}\Delta_i}\left(y - r_k^{(i)}\right)^2 f_{y_i}(y)dy, \tag{D.10}$$

where $f_{y_i}(\cdot)$ is the pdf of the subband signal y_i and $r_k^{(i)}$ is representation level number k in uniform threshold quantizer number i. The value of $r_k^{(i)}$ giving the minimum value of the variance $\sigma_{q_i}^2$ is given by the centroid [Gersho & Gray 1992] of the function $f_{y_i}(y)$ in the kth decision interval $\left(\frac{2k-1}{2}\Delta_i, \frac{2k+1}{2}\Delta_i\right]$, see Equation (6.2).

The determination of the off-diagonal elements of the $(l+1)M \times (l+1)M$ matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ can be derived by first finding the joint pdf of the variables $q_i(n)$ and $q_k(p)$, where $i \in \{0, 1, \ldots, M-1\}$, $k \in \{0, 1, \ldots, M-1\}$, $n \in \{0, 1, \ldots, l\}$,

and $p \in \{0, 1, \ldots, l\}$. It can be shown that the joint pdf is given by

$$f_{q_i(n),q_k(p)}(a,b) = \sum_{p_1=-\infty}^{\infty} \sum_{p_2=-\infty}^{\infty} f_{y_i(n),y_k(p)}\left(r_{p_1}^{(i)} - a, r_{p_2}^{(k)} - b\right) \cdot \operatorname{rect}\left(\frac{a - \left(r_{p_1}^{(i)} - p_1\Delta_i\right)}{\Delta_i}\right) \operatorname{rect}\left(\frac{b - \left(r_{p_2}^{(k)} - p_2\Delta_k\right)}{\Delta_k}\right), \quad \text{(D.11)}$$

where $f_{y_i(n),y_k(p)}(\cdot,\cdot)$ is the joint Gaussian pdf of the stochastic variables $y_i(n)$ and $y_k(p)$. It is fully specified by the means of these variables, which are assumed to be zero, their variances, and the cross-correlation $R_{y_i,y_k}(n-p)$.

The joint characteristic function of the variables $q_i(n)$ and $q_k(p)$ can be found (except for a sign in the exponent) by the Fourier transform of the pdf in Equation (D.11). By performing the Fourier transformation, it can be shown that the joint characteristic function is

$$\chi_{q_i(n),q_k(p)}(a,b) = \frac{\Delta_k\Delta_i}{4\pi^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sum_{p_1=-\infty}^{\infty} \sum_{p_2=-\infty}^{\infty} \cos\left(ar_{p_1}^{(i)} - \nu_1 p_1 \Delta_i\right) \cdot \cos\left(br_{p_2}^{(k)} - \nu_2 p_2 \Delta_k\right) \chi_{y_i(n),y_k(p)}(\nu_1 - a, \nu_2 - b) \frac{\sin\frac{\nu_1\Delta_i}{2}}{\frac{\nu_1\Delta_i}{2}} \frac{\sin\frac{\nu_2\Delta_k}{2}}{\frac{\nu_2\Delta_k}{2}} d\nu_1 d\nu_2, \quad \text{(D.12)}$$

where $\chi_{y_i(n),y_k(p)}$ is the joint characteristic function of the joint Gaussian variables $y_i(n)$ and $y_k(p)$.

After finding the joint characteristic function of the stochastic variables $q_i(n)$ and $q_k(p)$, it is possible to find $E[q_i(n)q_k(p)]$ by using the moment theorem. $E[q_i(n)q_k(p)]$ is the off-diagonal element in the matrix $\boldsymbol{\Phi}_{\boldsymbol{q}}^{(l,M)}$ in row number $nN + i$ and column number $pM + k$, and it can be shown that it is

found by the following expression:

$$
\begin{aligned}
R_{q_i(n),q_k}(p) = E\left[q_i(n)q_k(p)\right] = \sum_{p_1=1}^{\infty}\sum_{p_2=1}^{\infty} \frac{(-1)^{p_1+p_2}\Delta_i\Delta_k}{\pi^2 p_1 p_2} e^{-\frac{2p_1^2\pi^2\sigma_{y_i}^2}{\Delta_i^2}-\frac{2p_2^2\pi^2\sigma_{y_k}^2}{\Delta_k^2}} \\
\cdot \sinh\frac{4\pi^2 p_1 p_2 R_{y_i,y_k}(n-p)}{\Delta_i\Delta_k} \\
+ \sum_{p_1=1}^{\infty}\sum_{p_2=1}^{\infty}\left\{\frac{\Delta_i\Delta_k}{\pi^2 p_1} a_{p_2}^{(k)}(-1)^{p_1+1}\int_{-\infty}^{\infty}\sin\left(v_2 p_2\Delta_k\right)\right. \\
\left. \cdot \operatorname{sinc}\left(\frac{v_2\Delta_k}{2\pi}\right) e^{-\frac{2p_1^2\pi^2\sigma_{y_i}^2}{\Delta_i^2}-\frac{1}{2}v_2^2\sigma_{y_k}^2-\frac{2p_1\pi v_2 R_{y_i,y_k}(n-p)}{\Delta_i}} dv_2\right\} \\
+ \sum_{p_1=1}^{\infty}\sum_{p_2=1}^{\infty}\left\{\frac{\Delta_i\Delta_k}{\pi^2 p_2} a_{p_1}^{(i)}(-1)^{p_2+1}\int_{-\infty}^{\infty}\sin\left(v_1 p_1\Delta_i\right)\right. \\
\left. \cdot \operatorname{sinc}\left(\frac{v_1\Delta_i}{2\pi}\right) e^{-\frac{1}{2}v_1^2\sigma_{y_i}^2-\frac{2p_2^2\pi^2\sigma_{y_k}^2}{\Delta_k^2}-\frac{2p_2\pi v_1 R_{y_i,y_k}(n-p)}{\Delta_k}} dv_1\right\} \\
- \sum_{p_1=1}^{\infty}\sum_{p_2=1}^{\infty}\left\{\frac{\Delta_i\Delta_k}{\pi^2} a_{p_1}^{(i)} a_{p_2}^{(k)}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\sin\left(v_1 p_1\Delta_i\right)\sin\left(v_2 p_2\Delta_k\right)\right. \\
\left. \cdot \operatorname{sinc}\left(\frac{v_1\Delta_i}{2\pi}\right)\operatorname{sinc}\left(\frac{v_2\Delta_k}{2\pi}\right) e^{-\frac{1}{2}v_1^2\sigma_{y_i}^2-\frac{1}{2}v_2^2\sigma_{y_k}^2-v_1 v_2 R_{y_i,y_k}(n-p)} dv_1 dv_2\right\}
\end{aligned}
\tag{D.13}
$$

where $R_{y_i,y_k}(n-p) = E\left[y_i(n)y_k(p)\right]$ is the element in row number $nN+i$ and column number $pM+k$ in the matrix $E\left[\boldsymbol{y}(n)_{\mathsf{I}}\boldsymbol{y}^H(n)_{\mathsf{I}}\right] = \boldsymbol{E}_{\Gamma}\boldsymbol{\Phi}_{\boldsymbol{x}}^{(m+l,N)}\boldsymbol{E}_{\Gamma}^H$. The function $\operatorname{sinc}\colon \mathbb{R}\to\mathbb{R}$ is defined [Haykin 1983] as follows:

$$
\operatorname{sinc}(x) = \begin{cases} 1, & \text{if } x = 0, \\ \frac{\sin(\pi x)}{\pi x}, & \text{if } x \neq 0. \end{cases} \tag{D.14}
$$

The first double-sum in Equation (D.13) represents the cross-correlation that exists when the midpoints are used as representation levels in the quantizers, and this is a generalization of Equation (21) in [Gosse & Duhamel 1997] to include cross-correlation between quantization noise in *different* subbands.

Equation (D.13) contains three integrals with infinite integration range, which might result in an unstable numerical solution. Therefore, these three integrals have been rewritten as finite integrals. This can be done by using Parseval's theorem for aperiodic continuous valued functions [Proakis &

Manolakis 1992]. The first integral in Equation (D.13) is equal to

$$
\begin{aligned}
e^{-\frac{2p_1^2\pi^2}{\Delta_i^2}\left(\sigma_{y_i}^2-\frac{\left(R_{y_i,y_k}(n-p)\right)^2}{\sigma_{y_k}^2}\right)} \int_0^{\frac{\Delta_k}{4\pi}} \frac{4\pi}{\Delta_k}\sqrt{\frac{2\pi}{\sigma_{y_k}^2}}e^{-\frac{2\pi^2}{\sigma_{y_k}^2}\left(f^2+\frac{p_2^2\Delta_k^2}{4\pi^2}\right)} \\
\left\{\cos\left(\frac{2\pi p_1 p_2 R_{y_i,y_k}(n-p)\Delta_k}{\Delta_i\sigma_{y_k}^2}\right)\sinh\left(\frac{2\pi p_2\Delta_k}{\sigma_{y_k}^2}f\right)\cdot\right. \\
\sin\left(\frac{4\pi^2 p_1 R_{y_i,y_k}(n-p)}{\Delta_i\sigma_{y_k}^2}f\right) \\
-\sin\left(\frac{2\pi p_1 p_2 R_{y_i,y_k}(n-p)\Delta_k}{\Delta_i\sigma_{y_k}^2}\right)\cosh\left(\frac{2\pi p_2\Delta_k}{\sigma_{y_k}^2}f\right)\cdot \\
\left.\cos\left(\frac{4\pi^2 p_1 R_{y_i,y_k}(n-p)}{\Delta_i\sigma_{y_k}^2}f\right)\right\}\,df. \qquad \text{(D.15)}
\end{aligned}
$$

The second integral in Equation (D.13) is equal to

$$
\begin{aligned}
e^{-\frac{2p_2^2\pi^2}{\Delta_k^2}\left(\sigma_{y_k}^2-\frac{\left(R_{y_i,y_k}(n-p)\right)^2}{\sigma_{y_i}^2}\right)} \int_0^{\frac{\Delta_i}{4\pi}} \frac{4\pi}{\Delta_i}\sqrt{\frac{2\pi}{\sigma_{y_i}^2}}e^{-\frac{2\pi^2}{\sigma_{y_i}^2}\left(f^2+\frac{p_1^2\Delta_i^2}{4\pi^2}\right)} \\
\left\{\cos\left(\frac{2\pi p_1 p_2 R_{y_i,y_k}(n-p)\Delta_i}{\Delta_k\sigma_{y_i}^2}\right)\sinh\left(\frac{2\pi p_1\Delta_i}{\sigma_{y_i}^2}f\right)\cdot\right. \\
\sin\left(\frac{4\pi^2 p_2 R_{y_i,y_k}(n-p)}{\Delta_k\sigma_{y_i}^2}f\right) \\
-\sin\left(\frac{2\pi p_1 p_2 R_{y_i,y_k}(n-p)\Delta_i}{\Delta_k\sigma_{y_i}^2}\right)\cosh\left(\frac{2\pi p_1\Delta_i}{\sigma_{y_i}^2}f\right)\cdot \\
\left.\cos\left(\frac{4\pi^2 p_2 R_{y_i,y_k}(n-p)}{\Delta_k\sigma_{y_i}^2}f\right)\right\}\,df, \qquad \text{(D.16)}
\end{aligned}
$$

whilst the third integral, which is a double integral, can be written as a one-dimensional integral

$$
\begin{aligned}
\int_0^{\frac{\Delta_i}{4\pi}} \frac{4\pi}{\Delta_i}\sqrt{\frac{2\pi}{\sigma_{y_i}^2}}e^{-\frac{2\pi^2}{\sigma_{y_i}^2}\left(f^2+\frac{p_1^2\Delta_i^2}{4\pi^2}\right)} \left\{\sinh\left(\frac{2\pi p_1\Delta_i}{\sigma_{y_i}^2}f\right)\cdot g\left(\Delta_k,a_1,a_2,a_3,a_4\right)\right. \\
\left.-\cosh\left(\frac{2\pi p_1\Delta_i}{\sigma_{y_i}^2}f\right)\cdot g\left(\Delta_k,a_1,a_3,a_2,a_4\right)\right\}\,df, \qquad \text{(D.17)}
\end{aligned}
$$

where $a_1 = p_2\Delta_k$, $a_2 = \frac{p_1\Delta_i R_{y_i,y_k}(n-p)}{\sigma_{y_i}^2}$, $a_3 = \frac{2\pi R_{y_i,y_k}(n-p)f}{\sigma_{y_i}^2}$, and $a_4 = \frac{1}{2}\left(\sigma_{y_k}^2 - \frac{\left(R_{y_i,y_k}(n-p)\right)^2}{\sigma_{y_i}^2}\right)$. The function $g(\Delta_k, a_1, a_2, a_3, a_4)$ is defined

as

$$
\begin{aligned}
g(\Delta_k, a_1, a_2, a_3, a_4) = -\frac{\pi}{4\Delta_k} \Bigg\{ & \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} + \frac{a_1+a_2+a_3}{2\pi}\right)\right) \\
& + \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} - \frac{a_1+a_2+a_3}{2\pi}\right)\right) - \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} + \frac{a_1-a_2-a_3}{2\pi}\right)\right) \\
& - \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} - \frac{a_1-a_2-a_3}{2\pi}\right)\right) - \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} + \frac{a_1+a_2-a_3}{2\pi}\right)\right) \\
& - \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} - \frac{a_1+a_2-a_3}{2\pi}\right)\right) + \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} + \frac{a_1-a_2+a_3}{2\pi}\right)\right) \\
& + \operatorname{erf}\left(\frac{\pi}{\sqrt{a_4}}\left(\frac{\Delta_k}{4\pi} - \frac{a_1-a_2+a_3}{2\pi}\right)\right)\Bigg\}. \qquad \text{(D.18)}
\end{aligned}
$$

References

Aas, K. C. & Mullis, C. T. [1996], 'Minimum mean-squared error transform coding and subband coding', *IEEE Trans. Inform. Theory*, vol. 42, no. 4, pp. 1179–1192, July 1996.

Aase, S. O. [1993], Image Subband Coding Artifacts: Analysis and Remedies, Ph.D. dissertation, The Norwegian Institute of Technology, Norway.

Aase, S. O. & Ramstad, T. A. [1995], 'On the optimality of nonunitary filter banks in subband coders', *IEEE Trans. Image Processing*, vol. 4, no. 12, pp. 1585–1591, December 1995.

Abramowitz, M. & Stegun, I. A. [1972], *Handbook of Mathematical Functions*, Dover Publications, Inc., New York, USA.

Ahmed, N., Natarajan, T. & Rao, K. [1974], 'Discrete cosine transform', *IEEE Trans. Computers*, vol. 23, no. 1, pp. 90–93, January 1974.

Amitay, N. & Salz, J. [1984], 'Linear equalization theory in digital data transmission over dually polarized fading radio channels', *AT & T Bell Laboratories Technical Journal*, vol. 63, no. 10, pp. 2215–2259, December 1984.

Antonini, M., Barlaud, M., Mathieu, P. & Daubechies, I. [1992], 'Image coding using wavelet transform', *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 205–220, April 1992.

Balasingham, I. [1998], On Optimal Perfect Reconstruction Filter Banks for Image Compression, Ph.D. dissertation, Norwegian University of Science and Technology (NTNU), Norway.

Bellanger, M., Bonnerot, G. & Coudreuse, M. [1976], 'Digital filtering by polyphase network: Application to sample rate alteration and filter banks', *IEEE ASSP Mag.*, vol. 24, pp. 109–114, 1976.

Berger, T. [1971], *Rate Distortion Theory*, Prentice-Hall, Inc, Englewood Cliffs, New Jersey, USA.

Berger, T. & Tufts, D. W. [1967], 'Optimum pulse amplitude modulation—Part I: Transmitter-receiver design and bounds from information theory', *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 196–208, April 1967.

Bingham, J. A. [1990], 'Multicarrier modulation for data transmission: An idea whose time has come', *IEEE Communications Mag.*, vol. 28, no. 5, pp. 5–14, May 1990.

Blahut, R. E. [1987], *Principles and Practice of Information Theory*, Addison Wesley, Reading, Massachusetts, USA.

CCITT Rec. T.81 [1992], *Information Technology – Digital Compression and Coding of Continuous-Tone Still Images – Requirements and Guidelines.*

Chen, B.-S., Lin, C.-W. & Chen, Y.-L. [1995], 'Optimal signal reconstruction in noisy filter bank systems: Multirate kalman synthesis filtering approach', *IEEE Trans. Signal Processing*, vol. 43, no. 11, pp. 2496–2504, November 1995.

Coleman, T., Branch, M. A. & Grace, A. [1999], *Optimization Toolbox: For Use with Matlab*, The Math Works Inc., USA.

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

Cover, T. M. & Thomas, J. A. [1991], *Elements of Information Theory*, Wiley, New York, USA.

Crespo, P., Honig, M. L. & Steiglitz, K. [1989], Optimization of pre- and post-filters in the presence of near and far-end crosstalk, *in* 'Proc. Int. Conf. on Communications', vol. 1, Boston, USA, June 1989, pp. 541–547.

Delopoulos, A. N. & Kollais, S. D. [1996], 'Optimal filter banks for signal reconstruction from noisy subband components', *IEEE Trans. Signal Processing*, vol. 44, no. 2, pp. 212–224, February 1996.

Donoho, D. L., Vetterli, M., DeVore, R. A. & Daubechies, I. [1998], 'Data compression and harmonic analysis', *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2435–2476, October 1998.

Edwards, C. H. & Penney, D. E. [1986], *Calculus and Analytic Geometry*, 2nd ed., Prentice-Hall, Inc, Englewood Cliffs, New Jersey, USA.

Farvardin, N. & Modestino, J. W. [1984], 'Optimum quantizer performance for a class of non-gaussian memoryless sources', *IEEE Trans. Inform. Theory*, vol. IT-30, no. 3, pp. 485–497, May 1984.

Fischer, T. R. & Wang, M. [1992], 'Entropy-constrained trellis-coded quantization', *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 415–426, March 1992.

Fuldseth, A. & Ramstad, T. A. [1997], Bandwidth compression for continuous amplitude channels based on vector approximation to a continuous subset of the source signal space, *in* 'Proc. Int. Conf. on Acoustics, Speech, and Signal Proc.', vol. 4, Munich, Germany, April 1997, pp. 3093–3096.

Gersho, A. & Gray, R. M. [1992], *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston, MA, USA.

Goodall, W. M. [1951], 'Television by pulse code modulation', *Bell Syst. Tech. J.*, pp. 33–49, January 1951.

Gosse, K. & Duhamel, P. [1997], 'Perfect reconstruction versus MMSE filter banks in source coding', *IEEE Trans. Signal Processing*, vol. 45, no. 9, pp. 2188–2202, September 1997.

Gosse, K., Pothier, O. & Duhamel, P. [1995], Optimizing the synthesis filter bank in audio coding for minimum distortion using a frequency weighted psychoacoustic criterion, *in* 'IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics', New Paltz, NY, October 1995, pp. 191–194.

Graham, A. [1981], *Krönecker Products and Matrix Calculus with Applications*, Ellis Horwood Limited, England.

Grenander, U. & Szegö, G. [1958], *Toeplitz Forms and Their Applications*, University of California Press, Berkely, California, USA.

Haykin, S. [1983], *Communication Systems*, 2nd ed., John Wiley & Sons, Inc., Singapore.

Hines, W. W. & Montgomery, D. C. [1990], *Probability and Statistics in Engineering and Management Science*, 3rd ed., John Wiley & Sons, Inc.

Hjørungnes, A., Coward, H. & Ramstad, T. A. [1999], Minimum mean square error FIR filter banks with arbitrary filter lengths, *in* 'Proc. Int. Conf. on Image Processing', vol. 1, Kobe, Japan, October 1999, pp. 619–623.

Hjørungnes, A. & Lervik, J. M. [1997], Jointly optimal classification and uniform threshold quantization in entropy constrained subband image coding, *in* 'Proc. Int. Conf. on Acoustics, Speech, and Signal Proc.', vol. 4, Munich, Germany, April 1997, IEEE, pp. 3109–3112.

Hjørungnes, A., Lervik, J. M. & Ramstad, T. A. [1996], Entropy coding of composite sources modeled by infinite gaussian mixture distributions, *in* 'Proc. DSP Workshop', Loen, Norway, September 1996, IEEE, pp. 235–238.

Hjørungnes, A. & Ramstad, T. A. [1997], Linear solution of the combined source-channel coding problem using joint optimal analysis and synthesis filter banks, *in* 'Proc. for Thirty-First Asilomar Conference on Signals, Systems and Computers', vol. 2, Naval Postgraduate School; San Jose, CA, USA, Maple Press, November 1997, pp. 990–994.

Hjørungnes, A. & Ramstad, T. A. [1998*a*], Jointly optimal analysis and synthesis filter banks for bit constrained source coding, *in* 'Proc. Int. Conf. on Acoustics, Speech, and Signal Proc.', vol. 3, Seattle, USA, May 1998, pp. 1337–1340.

Hjørungnes, A. & Ramstad, T. A. [1998*b*], Minimum mean square error transform coders, *in* 'Proc. The 6th IEEE Int. Workshop on Intelligent Signal Processing and Communication Systems', vol. 2, Melbourne, Australia, November 1998, IEEE, pp. 738–742.

Hjørungnes, A. & Ramstad, T. A. [1998*c*], On the performance of linear transmission systems over power constrained, continuous amplitude channels, *in* 'Proc. for the UCSB Workshop on Signal & Image Processing', Santa Barbara, USA, December 1998, pp. 31–35.

Hjørungnes, A. & Ramstad, T. A. [1999*a*], Algorithm for jointly optimized analysis and synthesis FIR filter banks, *in* 'Proc. for the 6th IEEE International Conference on Electronics, Circuits and Systems', vol. 1, Paphos, Cyprus, September 1999, pp. 369–372.

Hjørungnes, A. & Ramstad, T. A. [1999*b*], 'Minimum mean square error subband coding', *IEEE Trans. Inform. Theory*, 1999. Submitted.

Hjørungnes, A. & Ramstad, T. A. [1999*c*], Minimum mean square error transforms and filter banks for source coding, *in* 'Proc. Second Int. Workshop on Transforms and Filter Banks', Brandenburg, Germany, March 1999, pp. 249–277.

Honig, M. L., Crespo, P. & Steiglitz, K. [1992], 'Suppression of near- and far-end crosstalk by linear pre- and post-filtering', *IEEE J. Select. Areas Commun.*, vol. 10, no. 3, pp. 614–629, April 1992.

Hotelling, H. [1933], 'Analysis of a complex of statistical variables into principal components', *J. Educational Psychology*, vol. 24, pp. 417–441 and 498–520, 1933.

Huang, J. J. Y. & Schultheiss, P. M. [1963], 'Block quantization of correlated gaussian random variables', *IEEE Trans. Circuits, Syst.*, pp. 289–296, September 1963.

ISO [1991], *Digital Compression and Coding of Continuous-Tone Still Images.* (JPEG).

ISO/IEC IS 11172 [1995], *Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Up to about 1.5 Mbit/s.* (MPEG-1).

ISO/IEC IS 13818 [1998], *Information Technology – Generic Coding of Moving Pictures and Associated Audio Information.* (MPEG-2).

Jain, A. K. [1989], *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Jayant, N. S. & Noll, P. [1984], *Digital Coding of Waveforms, Principles and Applications to Speech and Video*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA.

Judson, T. W. [1994], *Abstract Algebra: Theory and Applications*, PWS Publishing Company, Boston, USA.

Kıraç, A. [1998], Optimal Orthonormal Subband Coding and Lattice Quantization with Vector Dithering, Ph.D. dissertation, California Institute of Technology, Pasadena, California, USA.

Korn, G. A. [1965], 'Hybrid-computer techniques for measuring statistics from quantized data', *Simulation*, vol. 4, pp. 229–239, April 1965.

Kreyszig, E. [1988], *Advanced Engineering Mathematics*, 6th ed., John Wiley & Sons, Inc., New York, USA.

Le Gall, D. & Tabatabai, A. [1988], Subband coding of digital images using symmetric short kernel filters and arithmetic coding techniques, *in* 'Proc. Int. Conf. on Acoustics, Speech, and Signal Proc.', April 1988, pp. 761–764.

Lee, E. A. & Messerschmitt, D. G. [1994], *Digital Communication*, 2nd ed., Klüwer Academic Publishers, 3300 AH Dordrecht, Netherlands.

Lee, K.-H. & Petersen, D. P. [1976], 'Optimal linear coding for vector channels', *IEEE Trans. Commun.*, vol. COM-24, no. 12, pp. 1283–1290, December 1976.

Lipshitz, S. P., Wannamaker, R. A. & Vanderkooy, J. [1992], 'Quantization and dither: A theoretical survey', *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.

Luenbereger, D. G. [1984], *Linear and Nonlinear Programming*, 2nd ed., Addison–Wesley Publishing Company, Reading, Massachusetts, USA.

Lütkepohl, H. [1996], *Handbook of Matrices*, John Wiley & Sons, Inc., USA.

Magnus, J. R. & Neudecker, H. [1988], *Matrix Differential Calculus with Application in Statistics and Econometrics*, John Wiley & Sons, Inc., Essex, England.

Makhoul, J. [1981], 'On the eigenvectors of symmetric toepliz matrices', *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 4, pp. 868–872, August 1981.

Malvar, H. [1992], *Signal Processing with Lapped Transforms*, Artech House.

Malvar, H. S. [1986], Optimal Pre- and Post-Filtering in Noisy Sampled-Data Systems, Ph.D. dissertation, Massachusetts Institute of Thechnology, Cambridge, MA, USA.

Malvar, H. S. & Staelin, D. H. [1988], 'Optimal FIR pre- and postfilters for decimation and interpolation of random signals', *IEEE Trans. Commun.*, vol. 36, no. 1, pp. 67–74, January 1988.

Moulin, P., Anitescu, M. & Ramchandran, K. [1998], Asymptotic performance of subband coders using constrained, signal-adapted fir filter banks, *in* 'Proc. for the 32nd Annual Conference on Information Science and Systems', vol. I, Princeton, New Jersey, USA, March 1998, pp. 341–346.

Moulin, P., Anitescu, M. & Ramchandran, K. [2000], 'Theory of rate–distortion–optimal, constrained filter banks — application to IIR and FIR biorthogonal designs', *IEEE Trans. Signal Processing*, vol. 48, no. 4, pp. 1120–1132, April 2000.

Nayebi, K., Barnwell, T. P. & Smith, M. J. T. [1992], 'Time-domain filter bank analysis: A new design theory', *IEEE Trans. Signal Processing*, vol. 40, no. 6, pp. 1412–1429, June 1992.

Papoulis, A. [1991], *Probability, Random Variables, and Stochastic Processes*, Electrical Engineering, 3nd ed., Mcgraw-Hill Book Company, Singapore.

Park, K. & Haddad, R. A. [1993], Optimum subband filter bank design and compensation in the presence of quantizers, *in* 'Proc. for Twenty-Seventh Asilomar Conference on, Signals, Systems and Computers', vol. 1, Naval Postgraduate School; San Jose State, Maple Press, San Jose, CA, USA, November 1993, pp. 80–84.

Popat, K. [1990], Scalar quantization with arithmetic coding, M.Sc. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Proakis, J. G. & Manolakis, D. M. [1992], *Digital Signal Processing: Principles, Algorithms, and Applications*, 2nd ed., Macmillan Publishing Company, New York.

Ramstad, T. A., Aase, S. O. & Husøy, J. H. [1995], *Subband Compression of Images – Principles and Examples*, Elsevier Science Publishers B.V., Netherlands.

Rodrigues, M. A. M., da Silva, E. A. B. & Diniz, P. S. R. [1997], 'Design of wavelets for image compression satisfying perceptual criteria', *Electronics Letters*, vol. 33, no. 1, pp. 40–41, January 1997.

Salz, J. [1985], 'Digital transmission over cross-coupled linear channels', *AT & T Technical Journal*, vol. 64, no. 6, pp. 1147–1159, July–August 1985.

Sathe, V. & Vaidyanathan, P. P. [1993], 'Effects of multirate systems on the statistical properties of random signals', *IEEE Trans. Signal Processing*, vol. 41, pp. 131–146, January 1993.

Scharf, L. L. & Tufts, D. W. [1987], 'Rank reduction for modeling stationary signals', *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 3, pp. 350–355, March 1987.

Segall, A. [1976], 'Bit allocation and encoding for vector sources', *IEEE Trans. Inform. Theory*, vol. 22, no. 2, pp. 162–169, March 1976.

Shannon, C. E. [1948], 'A mathematical theory of communication', *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, July 1948.

Shannon, C. E. [1959], 'Coding theorems for a discrete source with a fidelity criterion', *IRE Nat. Conv. Rec.*, pp. 142–163, March 1959.

Song, B.-G. & Ritcey, J. A. [1997], 'Joint pre and postfilter design for spatial diverity equalization', *IEEE Trans. Signal Processing*, vol. 45, no. 1, pp. 276–280, January 1997.

Sripad, A. B. & Snyder, D. L. [1977], 'A necessary and sufficient condition for quantization errors to be uniform and white', *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 5, pp. 442–448, October 1977.

Strang, G. [1988], *Linear Algebra and its Applications*, 3rd ed., Harcourt Brave Jovanovich, Inc., San Diego, USA.

Sullivan, G. J. [1996], 'Efficient scalar quantization of exponential and Laplacian random variables', *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1365–1374, September 1996.

Therrien, C. W. [1992], *Discrete Random Signals and Statistical Signal Processing*, Prentice – Hall Inc., Englewood Cliffs, New Jersey, USA.

Troutman, J. L. [1996], *Variational Calculus and Optimal Control Optimization with Elementary Convexity*, Springer-Verlag, New York, USA.

Truss, J. K. [1991], *Discrete Mathematics for Computer Scientists*, Addison-Wesley Publishing Company, Inc., Suffolk, England.

Tsai, M. J., Villasenor, J. D. & Chen, F. [1996], 'Stack-run image coding', *IEEE Trans. Circuits, Syst. for Video Technol.*, vol. 6, no. 5, pp. 519–521, October 1996.

Tsatsanis, M. K. & Giannakis, G. B. [1995], 'Principal component filter banks for optimal multiresolution analysis', *IEEE Trans. Signal Processing*, vol. 43, no. 8, pp. 1766–1777, August 1995.

Tufts, D. W. & Berger, T. [1967], 'Optimum pulse amplitude modulation—Part II: Inclusion of timing jitter', *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 209–216, April 1967.

Tuqan, J. & Vaidyanathan, P. [1997], 'Statistically optimum pre- and postfiltering in quantization', *IEEE Trans. Circuits, Syst. II: Analog and Digital Signal Processing*, vol. 44, no. 1, pp. 1015–1031, December 1997.

Unser, M. [1993], 'On the optimality of ideal filters for pyramid and wavelet signal approximation', *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3591–3596, December 1993.

Vaidyanathan, P. & Chen, T. [1994], Statistically optimal synthesis banks for subband coders, *in* 'Proc. for Twenty-Eighth Asilomar Conference on Signals, Systems and Computers', October–November 1994, pp. 986 –990.

Vaidyanathan, P. P. [1993], *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, New Jersey, USA.

Vaidyanathan, P. P. [1998], 'Theory of optimal orthnormal subband coders', *IEEE Trans. Signal Processing*, vol. 46, no. 6, pp. 1528–1543, June 1998.

Vaidyanathan, P. P. & Kıraç, A. [1998], 'Results on optimal biorthogonal filter banks', *IEEE Trans. Circuits, Syst. II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 932–947, August 1998.

Vaidyanathan, P. P. & Mitra, S. K. [1988], 'Polyphase networks, block digital filtering, LPTV systems, and alias-free QMF banks: A unified approach based on pseudocirculants', *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 3, pp. 381–391, March 1988.

Vaishampayan, V. A. [1989], Combined Source-Channel Coding for Bandlimited Waveform Channels, Ph.D. dissertation, University of Maryland, USA.

Vaishampayan, V. A. & Farvardin, N. [1992], 'Joint design of block source codes and modulation signal sets', *IEEE Trans. Inform. Theory*, vol. 38, pp. 1230–1248, July 1992.

Vandendorpe, L. [1991], 'Optimized quantization for image subband coding', *Signal Processing: Image Communication*, vol. 4, no. 1, pp. 65–80, November 1991.

Vembu, S., Verdú, S. & Steinberg, Y. [1995], 'The source-channel separation theorem revisited', *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 44–54, January 1995.

Vetterli, M. & Kovačević, J. [1995], *Wavelets and Subband Coding*, Prentice Hall, Englewood Cliffs, New Jersey, USA.

Yang, J. & Roy, S. [1994], 'On joint transmitter and receiver optimization for multiple-input-multiple-output (MIMO) transmission systems', *IEEE Trans. Commun.*, vol. 42, no. 12, pp. 3221–3231, December 1994.

Young, N. [1990], *An Introduction to Hilbert Space*, Press Syndicate of the University of Cambridge, The Pitt Building, Trumpington Street, Cambridge. Reprinted edition.