Subband analysis structures and filters for still image compression

James Andrew
University of Wollongong

1994

Recommended Citation
NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
SUBBAND ANALYSIS STRUCTURES AND FILTERS
FOR STILL IMAGE COMPRESSION

A thesis submitted in fulfilment of the requirements for the award of the degree

DOCTOR OF PHILOSOPHY

from

UNIVERSITY OF WOLLONGONG

by

JAMES ANDREW, B.E. (HONS 1)

ELECTRICAL AND COMPUTER ENGINEERING

MARCH 1994
DECLARATION

This is to certify that the work reported in this thesis was done by the author, unless specified otherwise, and that no part of it has been submitted in a thesis to any other university or similar institution.

[Signature]

James Andrew
ACKNOWLEDGEMENTS

I would like to thank my supervisors, Dr Philip Ogunbona and Associate Professor Frank Paoloni of the University of Wollongong, for their advice, support and encouragement. In particular I would like to thank them for their technical interaction and their effort in reading various papers and this thesis.

I would like to thank Associate Professor Greg Doherty, Department of Computer Science, University of Wollongong, for his help and ideas in the pursuit of the optimum eigenfilters, discussed in Chapter 4. In particular he derived some very interesting results that are mentioned in Section 4.4.1.

I would also like to thank Philip Secker, other postgraduate students in the Department of Electrical and Computer Engineering, and the global net news community, for their help and stimulating interaction. The staff, and all involved, in the Department of Electrical and Computer Engineering, deserve thanks for making my time at the University of Wollongong enjoyable.

To Tricia and my family, I thank you for your support, encouragement, and for being my friends.
ABSTRACT

Digital images are replacing analogue images such as photographs and x-rays in many different fields. Compression of these digital images is desirable for efficient storage and transmission. Subband coding has proved an effective method of image compression. This thesis investigates subband analysis structures and filters which are optimised for still image compression. Among other results it is shown that a high coding gain, based on a typical image model, and good spatial localisation are desirable filter bank characteristics for subband image coding.

Assuming a high bit rate it is well known that the Karhunen-Loeve transform (KLT) is the optimum orthogonal block transform in terms of a coding gain metric. It is shown further that the KLT is the optimum invertible block transform using the unified coding gain. The coding gain metric is examined under a rate constraint. It is shown that for highly correlated sources increased low frequency subband resolution is required for optimum performance at low rates as compared to high rates: a result that is corroborated using a practical subband coder.

Subband filters (CQF's) that globally maximise the coding gain for all two-band perfect reconstruction orthogonal filter banks are derived. Various characteristics of these filters are predicted using a new theorem on the zeros of an eigenvector of a symmetric Toeplitz matrix corresponding to the minimum (maximum) eigenvalue. These filters are shown to enjoy the three properties of the KLT: namely maximum coding gain, minimum basis restriction error, and subband decorrelation. It is also shown that there is some freedom to select different impulse responses. The design of maximum gain filters is extended to include filters constrained to certain subspaces. For example maximum gain wavelets may be designed.

A modified two-dimensional discrete wavelet transform (DWT) is proposed based on a typical image model. A generic subband quantisation and encoding method suitable for any subband structure is introduced. This method is essentially a generalisation of the JPEG quantisation and encoding method and has good spatial adaptation properties.

Using the generic subband quantisation and encoding method various subband analysis structures and filters are compared for still image compression. The best orthogonal
filters, Daubechies wavelets and maximum gain filters designed using an image source model, and the discrete cosine transform (DCT) perform in a similar manner. The filters with the minimum spatial (time) width perform better than other impulse responses in a mean square error sense and exhibit significantly less ringing. The performance of cosine modulated filter banks, with poorer spatial resolution, is slightly inferior to the DCT. Preliminary investigations show that biorthogonal filters, with a smaller spatial width and higher coding gain, can outperform the best orthogonal filters, especially at low rates. These biorthogonal filters also exhibit minimal ringing. Finally, the modified DWT is shown to be superior to the DWT for head and shoulders type images.
TABLE OF CONTENTS

Acknowledgements .................................................................................................................. i
Abstract ................................................................................................................................... ii
Table of Abbreviations ........................................................................................................... viii
List of Symbols .................................................................................................................... ix

CHAPTER 1: INTRODUCTION ................................................................................................. 1
1.1. Motivation and Applications of Subband Analysis / Synthesis .................................... 1
1.2. Subband Image Coding ................................................................................................. 3
   1.2.1. History of Subband Coding ................................................................................. 3
   1.2.2. Subband Image Codec ....................................................................................... 4
   1.2.3. Advantages and Advances of Subband Coding .................................................. 6
1.3. Overview of Thesis ....................................................................................................... 7
1.4. Contributions ............................................................................................................... 10
1.5. Publications Relating to Thesis .................................................................................... 12

CHAPTER 2: SUBBAND ANALYSIS / SYNTHESIS ................................................................. 13
2.1. Notation and Preliminaries .......................................................................................... 13
2.2. Two-band Analysis /Synthesis .................................................................................... 15
   2.2.1. Frequency Domain Characterisation of the Two-band A/S System ................ 16
   2.2.2. Time Domain Characterisation of the Two-band A/S System ......................... 19
2.3. Arbitrary Analysis/Synthesis Configurations ............................................................... 23
   2.3.1. M-Band or Uniform Analysis/Synthesis .............................................................. 24
   2.3.2. Octave-band Analysis/Synthesis ........................................................................ 26
   2.3.3. Arbitrary Analysis/Synthesis ............................................................................ 28
   2.3.4. Perfect Reconstruction and Signal Boundary Extensions .................................. 29
2.4. Orthogonal Analysis Matrix Properties ....................................................................... 31
   2.4.1. Mean Square Error / Energy Conservation ...................................................... 31
   2.4.2. Filter bank Response to a Constant Input .......................................................... 32
2.5. Two-Dimensional Analysis/Synthesis ......................................................................... 35
2.6. Filter Bank Time Widths ............................................................................................. 38
CHAPTER 3: SUBBAND CODING GAIN .................................................. 40

3.1. Introduction ............................................................................. 40
  3.1.1. Overview of Chapter 3 ..................................................... 41
  3.1.2. Definition of Coding Gain ................................................. 42
  3.1.3. Statistical Image Models .................................................... 43
  3.1.4. Power Spectral Density and Correlation ......................... 45
  3.1.5. Lowpass to Highpass Filter Transformations .................. 48

3.2. Coding Gain and Bit Allocation ............................................. 50
  3.2.1. Background and Assumptions ......................................... 51
  3.2.2. Bit Allocation for Minimum Overall Distortion ............... 53
  3.2.3. Coding Gain .................................................................. 55
  3.2.4. Optimum Block Transform: The Karhunen-Loeve Transform (KLT) .............................................................. 57
  3.2.5. Coding Gain Bounds for Arbitrary Analysis/Synthesis Schemes ...... 61

3.3. Rate Constrained Coding Gain ............................................. 63
  3.3.1. Asymptotic Performance of the Rate Constrained Coding Gain .... 65
  3.3.2. Coding Gain for Highly Correlated Sources at Low Rates ....... 68

3.4. Transform Frequency Domain Characteristics .................... 73
  3.4.1. Pairwise Symmetric Transforms and a Symmetric Coding Gain .... 73
  3.4.2. Frequency Domain Characterisation of Transforms ............... 78

3.5. Conclusion ............................................................................ 80

CHAPTER 4: MAXIMUM CODING GAIN TWO-BAND FILTERS .................................................. 83

4.1. Introduction ............................................................................ 83
  4.1.1. Overview of Chapter 4 ..................................................... 85

4.2. Properties of Correlation matrices ....................................... 87
  4.2.1. Zeros of a Symmetric Toeplitz Matrix Eigenvector, Corresponding to a Repeated Minimum Eigenvalue ..................... 89

4.3. Optimum Two-band Orthogonal (CQF) Filters .................... 90
  4.3.1. Length Four Optimum Filters ......................................... 92
  4.3.2. Arbitrary Length Optimum Filters ................................... 96
  4.3.3. Design Algorithm .......................................................... 98
  4.3.4. Optimum Filter Examples ................................................. 99
    4.3.4.1. Maximising average stopband attenuation ................... 99

4.4. Properties of Optimum Filters ............................................. 102
### TABLE OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(N)</td>
<td>Nth order autoregressive (source)</td>
</tr>
<tr>
<td>bpp</td>
<td>bits per pixel</td>
</tr>
<tr>
<td>CQF</td>
<td>conjugate quadrature filter</td>
</tr>
<tr>
<td>DCT</td>
<td>discrete cosine transform (type II as defined by Rao and Yip 1990, p11)</td>
</tr>
<tr>
<td>DFT</td>
<td>discrete Fourier transform</td>
</tr>
<tr>
<td>DST</td>
<td>discrete sine transform (type I as defined by Rao and Yip 1990, p20)</td>
</tr>
<tr>
<td>DPCM</td>
<td>differential pulse code modulation</td>
</tr>
<tr>
<td>DWT</td>
<td>discrete wavelet transform (octave-band filter bank)</td>
</tr>
<tr>
<td>ELT</td>
<td>extended lapped transform</td>
</tr>
<tr>
<td>FIR</td>
<td>finite impulse response</td>
</tr>
<tr>
<td>FFT</td>
<td>fast Fourier transform</td>
</tr>
<tr>
<td>ICASSP</td>
<td>International Conference on Acoustics Speech and Signal Processing</td>
</tr>
<tr>
<td>IIR</td>
<td>infinite impulse response</td>
</tr>
<tr>
<td>KLT</td>
<td>Karhunen-Loeve transform</td>
</tr>
<tr>
<td>LOT</td>
<td>lapped orthogonal transform</td>
</tr>
<tr>
<td>MDWT</td>
<td>modified discrete wavelet transform</td>
</tr>
<tr>
<td>MLT</td>
<td>modulated lapped transform</td>
</tr>
<tr>
<td>MSE</td>
<td>mean square error</td>
</tr>
<tr>
<td>PCM</td>
<td>pulse code modulation</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>PR</td>
<td>perfect reconstruction</td>
</tr>
<tr>
<td>PR1</td>
<td>perfect reconstruction condition 1 (unit energy)</td>
</tr>
<tr>
<td>PR2</td>
<td>perfect reconstruction condition 2 (overlapping orthogonality)</td>
</tr>
<tr>
<td>SNR</td>
<td>signal to noise ratio</td>
</tr>
<tr>
<td>QMF</td>
<td>quadrature mirror filter</td>
</tr>
<tr>
<td>PSNR</td>
<td>peak signal to noise ratio</td>
</tr>
<tr>
<td>UC</td>
<td>unit circle</td>
</tr>
<tr>
<td>VQ</td>
<td>vector quantisation</td>
</tr>
<tr>
<td>WHT</td>
<td>Walsh-Hadamard transform</td>
</tr>
<tr>
<td>WSS</td>
<td>wide sense stationary</td>
</tr>
</tbody>
</table>
**LIST OF SYMBOLS**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>modified correlation matrix (also an arbitrary matrix)</td>
</tr>
<tr>
<td>$a_{ij}$</td>
<td>$(i,j)^{th}$ entry of $A$ matrix</td>
</tr>
<tr>
<td>$b_k$</td>
<td>number of bits allocated to subband $k$</td>
</tr>
<tr>
<td>$d_k$</td>
<td>decimation factor of subband $k$</td>
</tr>
<tr>
<td>$e_k$</td>
<td>eigenvalue $k$</td>
</tr>
<tr>
<td>$E{x}$</td>
<td>expected value of random variable $x$</td>
</tr>
<tr>
<td>$g_k(n)$</td>
<td>impulse response of synthesis filter $k$</td>
</tr>
<tr>
<td>$G_{TC}$</td>
<td>transform coding gain (subband coding gain for block transforms)</td>
</tr>
<tr>
<td>$G_{SBC}$</td>
<td>subband coding gain</td>
</tr>
<tr>
<td>$\hat{G}_{SBC}$</td>
<td>approximate subband coding gain (applicable to low analysis level subband methods with a highly correlated input source at low rates)</td>
</tr>
<tr>
<td>$h_k(n)$</td>
<td>impulse response of analysis filter $k$</td>
</tr>
<tr>
<td>$h(n)$</td>
<td>vector of filter coefficients (usually the coefficients of $h_k(n)$)</td>
</tr>
<tr>
<td>$H(e^{j\omega})$</td>
<td>frequency response (DFT) of filter $h(n)$</td>
</tr>
<tr>
<td>$I$</td>
<td>identity matrix</td>
</tr>
<tr>
<td>$J$</td>
<td>counter-identity matrix (ones on the anti-diagonal)</td>
</tr>
<tr>
<td>$M$</td>
<td>number of one-dimensional subbands for a uniform subband decomposition (implies $M^2$ subbands in two-dimensions)</td>
</tr>
<tr>
<td>$N$</td>
<td>length of filter or vector</td>
</tr>
<tr>
<td>$P$</td>
<td>permutation matrix</td>
</tr>
<tr>
<td>$QF$</td>
<td>quantisation factor</td>
</tr>
<tr>
<td>$r_{xx}(m)$</td>
<td>$m$ step correlation of signal $x$ (assumed to be wide sense stationary)</td>
</tr>
<tr>
<td>$R$</td>
<td>average (overall) system rate (in bits per pixel)</td>
</tr>
<tr>
<td>$R_{xx}$</td>
<td>correlation matrix of signal $x$</td>
</tr>
<tr>
<td>$[R_{xx}]_{ij}$</td>
<td>$(i,j)^{th}$ entry of matrix $R_{xx}$</td>
</tr>
<tr>
<td>$S$</td>
<td>octave-band or DWT tree-depth</td>
</tr>
<tr>
<td>$S_q$</td>
<td>quadic analysis tree-depth</td>
</tr>
<tr>
<td>$S_k$</td>
<td>error variance gain factor for subband $k$</td>
</tr>
<tr>
<td>$S_{xx}(e^{j\omega})$</td>
<td>power spectral density of signal $x$ (DFT of correlation signal $r_{xx}(m)$)</td>
</tr>
<tr>
<td>$T_a$</td>
<td>analysis matrix</td>
</tr>
<tr>
<td>$T_s$</td>
<td>synthesis matrix</td>
</tr>
<tr>
<td>$W$</td>
<td>overlapping orthogonality matrix</td>
</tr>
<tr>
<td>$x(n)$</td>
<td>filter bank input signal</td>
</tr>
</tbody>
</table>
\(x(n,m)\) two dimensional filter bank input image signal
\(\hat{x}(n,m)\) two dimensional filter bank output (reconstructed) image signal
\(x\) filter bank input data vector (also used as an arbitrary vector)
\(x'\) filter bank output (reconstructed) vector without subband quantisation
\(\hat{x}\) filter bank output data vector with subband quantisation
\(y_k(n)\) subband \(k\) signal
\(y\) subband data vector (also used as an arbitrary vector)
\(\hat{y}\) quantised subband data vector
\(\varepsilon^2\) quantiser performance factor
\(\rho\) one step correlation for an AR(1) source
\(\sigma_r^2\) reconstruction error variance
\(\sigma_k^2\) variance of subband \(k\)
\(\sigma_i^2\) time (spatial) width of a filter
\(\sigma_x^2\) variance of input signal \(x(n)\)
\(\omega\) radian frequency
INTRODUCTION

CHAPTER 1:

INTRODUCTION

Data represented digitally has advantages over analogue representation in terms of storage, flexibility, and errorless transmission [Jain 1981]. Digital images are replacing analogue images such as photographs and x-rays in many different fields. For example, a proposed standard for high definition television (HDTV), MPEG-2, is based on a digital image format. Many multimedia applications employ images, which are most conveniently stored, transmitted and processed in a digital form.

Digital images require large amounts of storage space and transmission bandwidth. A typical monochrome image with resolution 512x512 pixels requires 262,144 bytes of storage space. The current trend is for digital images with higher spatial resolution, such as HDTV, which further increases the demands on storage or transmission facilities. Image data compression is obviously desirable for storage and transmission purposes. In this thesis subband analysis structures and filters for the purpose of still image compression are investigated.

1.1. MOTIVATION AND APPLICATIONS OF SUBBAND ANALYSIS / SYNTHESIS

The human visual system has been modelled as a bank of independent parallel detection mechanisms, called spatial filters, which are tuned to different spatial frequencies and orientations [Sakrison 1979, p37]. Further the bandwidth of each filter is proportional to its centre frequency, and hence the filters have an equal bandwidth on a logarithmic scale. Since nature usually employs efficient systems, a motivation for subband coding is that it is possible in some sense to mimic this spatial frequency decomposition. Further, from the perspective of the end user, by mimicking this decomposition subband image coding can be tailored to exploit psychovisual properties of the human visual system. By carefully controlling the introduction of quantisation errors into different subbands, a reconstructed image that is subjectively pleasing can be obtained. For example, greater or lesser distortion can be tolerated in various
subbands, commensurate with the contrast sensitivity of the human eye to various frequency (sinusoidal) bands. In a similar manner subband coding of speech has been used to exploit psychoacoustic properties of the human ear.

As shown in Chapter 2, subband analysis can be considered as a linear transform: a decomposition of a signal onto a set of basis vectors. Subband synthesis is the inverse transform or a return to the original basis set. As such, many operations that are usually performed on the original signal may be performed more efficiently and/or effectively in the subband domain. For example, adaptive filtering is generally more efficient in the subband domain as compared to the fullband domain [Shynk 1992] [Malvar 1992]. In another quite different example, Kundu and Chen (1992) reported that better image texture classification was achieved in the subband domain as compared to the original domain.

There are a plethora of other applications in which subband analysis/synthesis is of use. As the title of this thesis suggests, subband coding is an effective method for data compression. For example, subband coding of speech, audio, image, video, and medical data signals has been investigated. Subband analysis also finds use in spectral estimation, antenna systems, transmultiplexing, data (image and speech) restoration and enhancement, echo cancellation, equalisation, and various medical image applications. More detail of these applications can be found in [Vaidyanathan 1990], [Malvar 1992] and in several papers published in ICASSP-91 and ICASSP-92.

Last, but certainly not least, there has recently been significant interest in the theory and application of "wavelets". The discrete wavelet transform (DWT) is an octave-band filter bank or a class of subband analysis. The DWT is considered for image compression in this thesis. Other than the topics listed in the previous paragraph, wavelets find applications in many diverse areas, such as seismic signal processing. An overview of wavelets and their applications can be found in [Rioul and Vetterli 1991].

The main emphasis of this thesis is the investigation of subband analysis structures and filters for still image compression. Nevertheless, many of the theories and results presented in this thesis can be adapted to other applications where subband analysis/synthesis is used. For example, with a goal of efficient image compression, principle-component-like subband structures and filters are investigated. Further investigation of such subband methods for other types of data compression should
prove fruitful. Also, principle-component-like methods are useful in other areas such as adaptive filtering and various remote sensing applications.

1.2. **SUBBAND IMAGE CODING**

1.2.1. **History of Subband Coding**

Subband coding of images can be traced back to the work of Kretzmer in 1956 [Woods and O'Neil 1986]. In his approach the image was split into a low and a high frequency subband which were subsequently encoded using PCM. The decoded image was reconstructed using this information. Another two-band coder, called the synthetic highs system, was described by Schreiber et al (1959,1989). After this early work, subband coding was largely ignored for quite some period of time. In 1976 Crochiere et al described a subband speech coder. Spectral coding of speech has since become a popular method of speech compression.

Burt and Adelson (1983) introduced a Laplacian pyramid as a compact image code, which can be considered as a type of subband image coder. In their approach an image is represented by a series of images, each image an approximation of the previous image at a reduced resolution. The error image between the interpolated approximation image and the current image at each stage is also stored. This pyramid can be used to reconstruct the original image exactly, but at the expense of an increase in the number of data samples.

Crosier et al (1976) introduced a quadrature mirror filter (QMF) bank, that splits a signal into two subbands. The subbands are decimated so that the total number of subband samples is equal to the original number of input data samples. The signal is reconstructed in a such a manner that any aliasing distortion introduced by the decimation is cancelled. This QMF filter bank is described in more detail in Chapter 2. Vetterli (1984) demonstrated that the QMF bank could be extended in a simple manner to multidimensional subband decompositions. Using a two-dimensional QMF bank Woods and O'Neil (1986) described one of the first modern subband image coders. Since this introduction subband coding of still images has become very popular.

Transform coding of images has also been a widespread method of image compression. It is widely recognised, and demonstrated later in this thesis, that transform coding is a
specific form of subband coding. However, modern subband coding has developed largely independently from transform coding. Wintz (1972) was one of the first to describe transform coding of images. With the advent of the discrete cosine transform (DCT), introduced by Ahmed et al (1974), transform coding of images has become extremely popular. Much of the work in transform coding has culminated in the JPEG still image compression standard, which is a DCT based transform coder [Wallace 1991].

There are two factors that make data compression possible: namely the statistical structure of the source, and the requirements of the end user. From a statistical perspective it is usually more efficient to quantise a signal in the subband (or transform) domain than in the original domain. This topic is considered in some detail in Chapter 3 of this thesis.

1.2.2. Subband Image Codec

Figure 1.1 illustrates a subband image codec. Subband coding consists of two separate tasks: namely subband analysis or decomposition, and quantisation/encoding. Subband synthesis reverses the analysis process, while the decoding reverses, as much as possible, the quantisation/encoding process. These two tasks are described briefly in this subsection. A detailed description of subband analysis/synthesis is given in Chapter 2, while a detailed description of a particular quantisation/encoding method is given in Chapter 5.
An input image to a subband codec, $x(n,m)$, is decomposed into two-dimensional spectral subbands by the analysis filter bank. The $k^{th}$ subband is formed by passing the input image through the $k^{th}$ filter, $h_k(n,m)$, followed by decimation. Decimation, as indicated by $d_k^\downarrow$ in Figure 1.1, involves subsampling or discarding certain signal samples, and is required so that the size of the data set representing the image does not increase following the analysis.

Following the subband analysis or decomposition the subband data is quantised and encoded. The subband coder is a very general waveform coder since any quantisation and encoding method may be used on the subbands. This information is then stored or transmitted according to the application. To reconstruct the image the subband data is decoded and synthesised. Decoding reverses the encoding operation giving the quantised subband data. Synthesis involves upsampling, indicated by $d_k^\uparrow$, filtering each subband signal independently, and then adding the resulting signals. Upsampling replaces the signal samples discarded in the decimation process with zero values.

Two-dimensional subband analysis/synthesis (A/S) is indicated in Figure 1.1. This type of A/S is suitable for still image compression where the input signal is two-dimensional. In the case of video signals three-dimensional subband A/S may be employed.
1.2.3. Advantages and Advances of Subband Coding

Subband coding is a subset of a class of coding methods that utilise linear dependencies in the input signal to aid coding performance. Another subset of this class are differential pulse code modulation (DPCM) schemes. Theoretically the mean square error (MSE) performance for optimum DPCM and optimum transform coding, using PCM to quantise the transform coefficients, is the same [Jayant and Noll 1984, p511]. As shown in Chapter 3, this optimum bound applies to subband coding in general. In practice transform, and subband coding in general, is superior to DPCM, especially at low bit rates, in terms of matching the codec to the data statistics, effects of transmission errors and subjective quality [Jayant and Noll 1984, p511]. Woods and O'Neil (1986) demonstrated that optimum DPCM on the original fullband signal and optimum DPCM on the subband signals perform equally in terms of MSE. However, Pearlman (in Woods 1991, p32) noted that in practice optimum subband DPCM is likely to outperform optimum fullband DPCM in a MSE sense.

Although transform image coders perform efficiently there is an annoying blocking effect attendant on the reconstructed image. Subband coding in general will obviate this blocking. However, other distortions such as ringing or "mosquito" noise around edges can be a problem in this case. Using a subband or transform coder it is a simple task to adapt the quantisation to suit the human visual system, giving superior subjective results. Forchheimer and Kronander (1989, p2009) described a subband coder as the "ultimate" waveform coder. This impressive title merely reflects the fact that the subband coder can incorporate any coding scheme to encode the subbands.

Subband coding of images has attracted sufficient attention to warrant a book published in 1991, edited by Woods (1991). Much of the work in several topics associated with subband coding is summarised by various authors in each chapter. In Chapter 4 of this book Simoncelli and Adelson summarised the following points: strongly non-orthogonal subband analysis methods are undesirable for image coding; joint spatial and spatial frequency localisation is desirable; subbands of equal width on a logarithmic scale may be desirable (ie octave or wavelet structure); the Laplacian pyramid approach is generally inferior to a critically sampled subband analysis for coding purposes; the QMF (or CQF) subband analysis satisfies most of the desirable properties for image coding; non-separable filters designed on non-square sampling
lattices offer similar results to the square-lattice QMF, but are difficult to design and require extra computation to convert the square sampling lattice as required.

Kronander (1989a, 1989b) investigated several criteria pertinent to filter banks as applied to image and video coding. Kronander concluded that filter banks should have a good coding efficiency (for a typical image model) and introduce minimal aliasing, phase, and ringing distortions.

Several traditional subband coding schemes are discussed in the introduction of Chapter 5. The octave-band or wavelet subband structure has recently received much attention in the image coding literature. It has been argued that this type of structure models the human visual system [Sakrison 1979]. Antonini et al (1992) described an image compression scheme consisting of a wavelet transform (octave subband analysis) and subband vector quantisation. Their results in terms of peak signal to noise ratio (PSNR) are commensurate with those of Woods and O'Neil (1986) and Westerink et al (1989) (see Antonini et al 1992, p217), while using shorter subband filters. Further, the coding of the subbands was designed to minimise a distortion measure weighted with subjective considerations. It was suggested that with the incorporation of entropy constrained vector quantisation (ECVQ) a significant improvement in PSNR can be made. These points illustrate flexibility inherent in a subband coder.

1.3. OVERVIEW OF THESIS

There are many open questions remaining in the area of subband analysis as applied to image coding, some of which are addressed in this thesis. Chapter 2 provides an overview of subband analysis/synthesis (A/S) systems, while the main contributions of this thesis are given in the four following chapters. Detailed literature references are given in each chapter, where they are relevant to the topic at hand. A gain metric for measuring the coding performance of a subband coder is examined in Chapter 3. Using this metric some maximum gain two-band subband filters are developed in Chapter 4. Chapter 5 considers some two-dimensional subband A/S structures, and introduces a generic subband quantisation and encoding method suitable for any subband structure. Finally in Chapter 6, using this quantisation and encoding method, various subband structures and filters are compared for image compression using a practical subband image codec. These comparisons are also used to test the various themes developed in this thesis.
INTRODUCTION

Various one-dimensional A/S systems are described in the time and frequency domains in Chapter 2. Finite input A/S is considered as a linear system using an input-output matrix equation. Using this formulation some properties of orthogonal A/S are derived. Two-dimensional A/S is considered as a separable application of one-dimensional A/S on each dimension. Finally a measure of the time (or spatial) localisation of a subband is discussed.

In Chapter 3 subband schemes where the subbands are quantised independently are considered. Using a quantisation noise model the quantisation bit allocation among subbands, under an overall bit rate constraint, that minimises the mean square image reconstruction error is derived. A general subband coding gain metric, introduced by Katto and Yashuda (1991), is derived assuming this optimum bit allocation and a high overall bit rate. This gain is an estimate of the reduction of the reconstruction error variance of a subband analysis system, using PCM on the subbands, as compared to a PCM system operating on the fullband signal.

It is shown that, under various assumptions valid at high rates, the KLT is the globally optimum invertible transform in terms of coding gain. The derivation of the general coding gain also provides a simple bit allocation procedure for all perfect reconstruction subband coders, which is used in Chapter 6. A rate constrained coding gain metric is then examined, which considers the effect of a low or moderate overall bit rate. This metric suggests that increased analysis levels are required for highly correlated sources at low rates for optimum performance. This prediction is corroborated by the results in Chapter 6. Finally the frequency domain characteristics of various block transforms are investigated. An interesting theorem, pertaining to the performance of block transforms as the correlation ($p$) varies for an AR(1) source is given. The DCT in particular is investigated and the impressive performance of this transform, for image coding purposes, is explained from a frequency domain perspective.

The frequency domain characteristics of the DCT provide the motivation for the work presented in Chapter 4. This chapter addresses the problem of maximising the coding gain for an orthogonal two-band subband coder. The optimum orthogonal two-band subband system, defined by some optimum filters (CQF's), can then be used as the building block in an octave-band subband coder. Chapter 4 begins with an overview of
the properties of symmetric Toeplitz matrices and presents a new theorem pertaining to
an eigenvector associated with a minimum (or maximum) repeated eigenvalue. This
theorem is directly relevant to the design of the optimum subband filters, which are
termed eigenfilters. Some necessary conditions for optimality in this regard are
derived. Also a sufficient condition is given. The design method is demonstrated for a
simple example using 4-tap filters and then generalised to arbitrary length filters.

Some eigenfilter properties are derived using the necessary and sufficient conditions
for optimality. It is shown that there are $N/2$ zeros on the unit circle for $N$ length filters.
Further, for filters of 8-taps in length or greater, it is shown that there is some freedom
in the selection of the optimum filter's impulse response. This freedom allows different
phase responses while maintaining the same magnitude response and is used to design
minimum time width impulse response filters. The optimum filters are shown to exhibit
the same properties as the block transform KLT, and that they can be thought of as a
generalisation of this transform in the case of two subbands or basis functions. The
design of optimum zero constrained filters is also described, which is equivalent to an
optimum wavelet problem. Finally these optimum eigenfilters are compared to other
filters and transforms using the coding gain metric.

In Chapter 5 two-dimensional subband A/S structures are considered. A two-
dimensional power spectral density (PSD) of an image model is discussed. Three
distinguishing characteristics of a subband analysis method are identified: namely the
ideal subband structure that is approximated by the analysis, the degree of this
approximation, and the time (spatial) resolution of the subbands. Based on the PSD
image model some new subband analysis structures are suggested.

A generic subband codec is also introduced. A quantisation and encoding methodology,
suitable for any subband analysis structure, is described. In the case where a block
transform DCT is used the coding scheme is essentially the same as the baseline JPEG
still image compression standard.

The generic subband coder is then used to evaluate and compare various subband
analysis structures and filters in Chapter 6. Some $M$-band transforms are compared in
the first section. The performance of these transforms, at different compression ratios
and analysis levels ($M$), is evaluated. In the next section some octave-band (DWT)
subband coders are compared. Issues such as filter length, filter phase, and analysis
level are addressed. Eigenfilters designed for different source models are considered. Further, these eigenfilters, Daubechies wavelets, and some linear phase biorthogonal filters are compared. Finally the new subband structures, suggested in Chapter 5, are compared to the octave-band and M-band analysis structures using the optimum filters in each case.

This thesis is concluded in Chapter 7. The main contributions of the thesis are reviewed and future work and ideas are discussed.

1.4. CONTRIBUTIONS

The main original contributions of this thesis are listed in this section. The relevant section of the thesis and published work is also given.

1. In Chapter 2 a necessary and sufficient condition for an orthogonal filter bank to be zero-constrained is given. This condition, whereby all but one of the analysis filters has a zero at DC, is shown in Chapter 6 to be a desirable attribute for subband image coding. As a consequence of this condition a zero-constrained block transform must have a constant DC basis vector. (Also, mean square error and energy conservation for an orthogonal filter bank is demonstrated using a simple proof involving orthogonal matrices). See Section 2.4.2.

2. It is well known that the Karhunen-Loeve transform is the optimum orthogonal block transform in terms of coding gain. Using the unified coding gain it is shown further that the KLT is the optimum invertible block transform. Also, as a consequence it is shown that the coding gain of any perfect reconstruction subband A/S scheme is bounded by that of the infinite block size KLT (the inverse of the spectral flatness measure). See Section 3.2.4.

3. The usual coding gain assumes a high rate. A rate constrained coding gain is examined. It is shown that for highly correlated sources, a higher level of analysis is required for near optimum performance at low rates as compared to high rates. This result is corroborated using a practical subband coder in Chapter 6. See Section 3.3.

4. A theorem relating filter pairwise symmetry to a symmetric coding gain with respect to the sign of $\rho$ for an AR(1) source is given. Also the coding gain of various transforms is explained from a frequency domain perspective and in relation to this theorem. See Section 3.4.
5. A theorem (Theorem 4.1) relating the zeros of an eigenvector of a symmetric Toeplitz matrix corresponding to the minimum (maximum) eigenvalue to the multiplicity of this eigenvalue is given. See Section 4.2.1. and Andrew et al (1993a).

6. A necessary condition and a sufficient condition are given for the globally maximum coding gain orthogonal two-band filters (CQF's). These optimum filters are referred to as eigenfilters. A design procedure is outlined and some MATLAB¹ code used to implement the design listed. See section 4.3 and Andrew et al (1993c).

7. Theorem 4.1 is used to predict certain magnitude characteristics of these eigenfilters. Aliasing energy minimisation is shown to be a similar measure to maximum coding gain for highly correlated sources. KLT-like properties of the maximum gain CQF's are demonstrated. See Section 4.4. and Andrew et al (1993c).

8. The design of maximum gain CQF's is extended to include filters constrained to lie in certain subspaces. For example maximum gain wavelets may be designed. See Section 4.5.

9. It is shown that there is some freedom in the selection of the impulse response of the maximum gain CQF's (or any CQF). The number of different impulse responses in general is derived for a given filter length. This freedom is used to select minimum time width filters, a characteristic desirable for image coding. See Section 4.6. and Andrew et al (1993c, 1994).

10. Various subband schemes, including an octave-band filter bank (or discrete wavelet transform DWT) using the maximum gain CQF's are investigated in terms of coding gain. It is shown that the coding gain for Daubechies filters is nearly the same as that for the best eigenfilters. See Section 4.7. and Andrew et al (1993c).

11. A generic subband quantisation and encoding method is introduced in Chapter 5. See Section 5.3 and Andrew et al (1993d).

12. A modified two-dimensional DWT is proposed based on a typical image model. A comparison of various other subband structures is made in terms of coding gain. See Section 5.2 and Andrew et al (1993d).

13. Using the generic subband quantisation and encoding method various subband analysis structures and filters are compared for image compression. Since there are too many conclusions to list here the reader is referred to the conclusion of Chapter 6: Sections 6.5 and 6.6. Also see Andrew et al (1993b, 1993d, 1994).

¹MATLAB is a trademark of the Math Works, Inc.
1.5. PUBLICATIONS RELATING TO THESIS


Chapter 2: Subband Analysis / Synthesis

Subband analysis is the decomposition of a signal into spectral subband components. This analysis can be thought of as a preprocessing of the input signal. Among other properties subband analysis allows exploitation of linear dependencies between samples of the input signal in a simple fashion. As a consequence it is more efficient to quantise and compress many signals using subband analysis.

This chapter describes in some detail various analysis/synthesis methods. In particular the following sections discuss one-dimensional two-band, $M$-band, and octave-band analysis. Equations describing these systems are given from both a time domain and frequency domain perspective. These equations are used in subsequent chapters of this thesis. The time domain framework is used to illustrate that orthogonal analysis/synthesis systems are a generalisation of orthogonal block transforms. Two-dimensional separable analysis is considered using one-dimensional analysis on each dimension. A measure of the spread of a filter’s impulse response, the filter time width, is also discussed.

2.1. Notation and Preliminaries

A finite impulse response (FIR) filter is usually denoted by $h$, a vector of the filter coefficients. The coefficients are arranged in reverse order. For example,

$$h = [h(N-1), h(N-2), \ldots, h(0)]^T$$

where $^T$ denotes the vector/matrix transpose operator and $N$ is the length of the filter. On the other hand a vector can be considered as a FIR filter with coefficients denoted as above. Vectors of input and output samples are arranged in terms of increasing time. For example,

$$x = [\ldots, x(-1), x(0), x(1), x(2), \ldots]^T$$
By using a reverse arrangement of filter coefficients, one step of a filtering or convolution process can be represented as a vector inner product. Multiple steps can be represented using a matrix vector product.

The Z-transform of a vector is the Z-transform of the corresponding FIR filter,

\[ H(z) = \sum_{n=0}^{N-1} h(n)z^{-n} = h^T \begin{bmatrix} z^{-(N-1)} \\ \vdots \\ 1 \end{bmatrix} \]

This Z-transform can be factored into a product form,

\[ H(z) = A \prod_{k=1}^{N-1} (z^{-1} - z_k) \]

where \( A \) is a magnitude factor. The zeros of a vector are the \( z_k \)'s. The frequency response of a vector or filter is evaluated by setting \( z = e^{j\omega} \). A zero of a vector on the unit circle corresponds to a zero in the frequency response. For example,

\[ z_k = e^{-j\phi} \Rightarrow H(e^{j\omega}) \bigl|_{\omega=\phi} = 0 \]

For software reasons, filter magnitude responses are plotted against normalised frequency (cycles/sample) rather than \( \omega \) radians. The normalised frequency range of 0 to 0.5 cycles/sample corresponds to 0 to \( \pi \) radians.

An ideal filter is one that has infinite attenuation in the stopband and zero transition bandwidth [Oppenheim and Schafer 1989, p42]. In other words the magnitude response is rectangular. Obviously an ideal filter cannot be implemented in practice. However a ideal filter is a useful mathematical abstraction, and is used to simplify various mathematical analyses. An ideal filter bank is a filter bank (or subband analysis/synthesis system) which employs ideal filters.

A decimator and interpolator are illustrated in Figure 2.1. The \( D \)-fold decimator, denoted by \( D \downarrow \) in Figure 2.1, is characterised by the input-output relation,
\[ y(n) = x(Dn) \]

which reduces the sampling rate by a factor of \( D \).

\[ x(n) \quad y(n) \quad z(n) \]

Figure 2.1. \( D \)-fold Decimator and \( I \)-fold Interpolator.

The \( I \)-fold interpolator, denoted by \( I \uparrow \) in Figure 2.1, is characterised by the input-output relation,

\[ z(n) = \begin{cases} 
  y\left(\frac{n}{I}\right) & \text{if } n \text{ is an integer} \\
  0 & \text{else}
\end{cases} \]

The interpolator inserts \( I-1 \) zero valued samples between the adjacent samples of \( y(n) \) giving the signal \( z(n) \). The input-output relations for the decimator and interpolator in the Z-domain are [Vaidyanathan 1990],

\[
Y(z) = \frac{1}{D} \sum_{k=0}^{D-1} X \left( z^{D} e^{-2\pi jk/D} \right), \quad Z(z) = Y(z')
\]

respectively. For the decimator the term corresponding to \( k=0 \) is an \( D \)-fold stretched version of the input spectrum. The other terms, called aliasing terms, correspond to uniformly shifted versions of this stretched spectrum. The interpolator on the other hand forms an \( I \)-fold compressed version of the input signal. A good overview of multirate digital signal processing may be found in [Vaidyanathan 1990].

2.2. TWO-BAND ANALYSIS / SYNTHESIS

Figure 2.2 illustrates a generic one-dimensional two-band analysis/synthesis (A/S) subband system. A subband analysis or analysis/synthesis system is often referred to as a filter bank. The input signal, \( X \), is filtered into a lowpass and highpass signal by \( H_0 \).
and $H_i$ respectively. Each signal is decimated by two, as depicted by $2\downarrow$; that is every second sample is discarded. The motivation for decimation is to maintain (at point Y) the original sampling rate of $X$. The decimated signals are then processed according to the application. For example, quantised and encoded in a subband coder. The synthesis consists of upsampling by two, depicted as $2\uparrow$, which inserts a zero between every sample. Ignoring the quantisation process the cascade of the decimation and interpolation sets every second sample to zero. The signals are then filtered with lowpass and highpass filters, $G_0$ and $G_1$ respectively, and summed to give the reconstructed signal.

\[
\begin{align*}
\mathbf{X}(z) & \quad 2\downarrow \quad Y_0(z) \quad 2\uparrow \quad G_0(z) \quad \mathbf{X}'(z) \\
\mathbf{H}_0(z) & \quad 2\downarrow \quad Y_1(z) \quad 2\uparrow \quad G_1(z) \\
\mathbf{H}_1(z) &
\end{align*}
\]

Figure 2.2. Two-band analysis/synthesis subband system

A system is said to be a perfect reconstruction (PR) system if the output $X'$ is equal to a delayed version of the input $X$. There are four possible distortions introduced in this analysis/synthesis system: namely aliasing, imaging, magnitude and phase distortions. These distortions are described in the following subsection.

### 2.2.1. Frequency Domain Characterisation of the Two-band A/S System

Historically the two-band and more general A/S systems have been characterised mathematically in the frequency domain. The aim of this subsection is to give a brief overview of this characterisation. A detailed description may be found in the listed references. Further motivation is to give some equations that are used later in this thesis.
For the two-band A/S system it can be shown that the output is related to the input as [Smith and Barnwell, 1986],

\[
x'(e^{j\omega}) = \frac{1}{2} \left[ H_0(e^{j\omega})G_0(e^{j\omega}) + H_1(e^{j\omega})G_1(e^{j\omega}) \right] X(e^{j\omega}) \\
+ \frac{1}{2} \left[ H_0(-e^{j\omega})G_0(e^{j\omega}) + H_1(-e^{j\omega})G_1(e^{j\omega}) \right] X(-e^{j\omega})
\]  

(2.1)

The second term in (2.1) represents aliasing (and imaging) distortion. The problem of attaining PR or near PR has received much attention in the research community. The first solution was the quadrature mirror filter (QMF) bank [Croisier et al, 1976]. The QMF's remove aliasing by constraining the synthesis filters as,

\[
G_0(e^{j\omega}) = H_1(-e^{j\omega}), \quad G_1(e^{j\omega}) = -H_0(-e^{j\omega})
\]

(2.2)

which is referred to as the aliasing cancellation condition. The system transfer function is then,

\[
\frac{X'(e^{j\omega})}{X(e^{j\omega})} = \frac{1}{2} \left[ H_0(e^{j\omega})H_1(-e^{j\omega}) - H_1(e^{j\omega})H_0(-e^{j\omega}) \right]
\]

(2.3)

The filter bank phase and magnitude distortions are dependent on this transfer function. Perfect reconstruction (PR) is attained when the transfer function is a simple delay.

The QMF solution of Croisier et al (1976) specifies that the highpass analysis filter is, \(H_1(e^{j\omega}) = H_0(-e^{j\omega})\) which gives,

\[
\frac{X'(e^{j\omega})}{X(e^{j\omega})} = \frac{1}{2} \left[ H_0^2(e^{j\omega}) - H_0^2(-e^{j\omega}) \right]
\]

(2.4)

Unfortunately, in only two special cases does the above system function describe a delay giving PR. These cases are when the length of the analysis/synthesis filters is two or when they are infinitely long. The former case actually describes a two by two block transform such as the discrete Fourier transform (DFT) or discrete cosine transform (DCT).
The first PR filter solution was offered by Smith and Barnwell (1984, 1986). These filters were termed conjugate quadrature filters (CQF's). The highpass analysis filter is given by,

\[ H_i(e^{j\omega}) = -e^{-j\omega(N-1)}H_0(-e^{-j\omega}) \]  

(2.5)

where \( N \) is the length of the lowpass filter. The \( e^{-j\omega(N-1)} \) term is required so that the highpass filter is causal, and generates an overall system delay of \( N-1 \). In the time domain the filters are related as follows,

\[ g_0(n) = h_0(N-1-n), \quad g_1(n) = h_1(N-1-n), \quad h_2(n) = (-1)^n h_0(N-1-n) \]  

(2.6)

The CQF solution is almost the same as the QMF solution. The only new ingredient is a time-reversal: the highpass analysis filter (and hence the lowpass synthesis filter) is a time reversed version of the QMF solution. Using CQF's the resulting transfer function for real-valued filters is,

\[ X(e^{j\omega}) = \frac{1}{2} \left[ |H_0(e^{j\omega})|^2 + |H_0(e^{j(\pi+\omega)})|^2 \right] e^{-j\omega(N-1)}X(e^{j\omega}) \]  

(2.7)

It is possible to constrain the term in square brackets to be two, giving an output that is a delayed version of the input.

The CQF solution specifies that the lowpass synthesis filter is the time reversed lowpass analysis filter and similarly for the highpass filters. This time reversal means that any phase distortion introduced in the analysis is cancelled in the synthesis. Further the CQF's employ the aliasing cancellation condition. Hence the only constraint left is the magnitude distortion condition, the square bracket term in equation (2.7). The benefit of this characteristic of the CQF solution is illustrated beautifully in the time-domain as shown in the next subsection.

As discussed later in this chapter CQF's cannot have linear phase. However there are other PR filter banks with linear phase filters. Equation (2.3) can be written as,

\[ \frac{X(e^{j\omega})}{X(e^{j\omega})} = \frac{1}{2} [P(e^{j\omega}) - P(-e^{j\omega})], \quad P(e^{j\omega}) = H_0(e^{j\omega})H_i(-e^{j\omega}) \]
A necessary and sufficient condition for PR with FIR filters is that the aliasing cancellation condition (2.3) holds and,

\[ P(e^{j\omega}) - P(-e^{j\omega}) = 2e^{-j(2k+1)\omega} \]

where \( k \) is an arbitrary integer [Vetterli and Herley 1990]. In other words the above transfer function is an odd delay. In the time domain this means that,

\[ p(n) - (-1)^n p(n) = 2\delta(n - 2k - 1) \]

which implies that every odd coefficient of \( p(n) \), baring one, must be zero. Linear phase PR filters, \( h_0(n) \) and \( h_1(n) \), can be designed as spectral factorisations of such \( p(n) \) which has linear phase (ie is symmetric). For example see [Le Gall and Tabatabai, 1988] and [Vetterli and Herley, 1990].

### 2.2.2. Time Domain Characterisation of the Two-band A/S System

Additional insights may be gained by considering a time domain characterisation of a two-band A/S system. This has been done in the literature, although it has been a more recent approach.

The analysis can be described in the time domain using block-Toeplitz or block circulant matrices [Vetterli and Le Gall 1989, p1059]. Consider an infinite vector,

\[ x = \ldots x(-1), x(0), x(1), \ldots \]  

(2.8)

as the input to the two-band analysis filter bank. The output, a lowpass and a highpass decimated subband signal, can be represented as an infinite output vector,

\[ y = \ldots y_0(-1), y_1(-1), y_0(0), y_1(0), y_0(1), y_1(1), \ldots \]  

(2.9)

where the lowpass and highpass analysis signals are multiplexed in time. Note that \( y \) is the time domain equivalent of the two Z-domain signals, \( Y_0(z) \) and \( Y_1(z) \), indicated in Figure 2.2. The analysis can then be described as,

\[ y = T_y x \]  

(2.10)
where the analysis matrix,

\[
T_a = \begin{bmatrix}
... & 0 & 0 & \ldots & 0 & 0 \\
0 & h_0(N-1) & h_0(N-2) & h_0(N-3) & \ldots & h_0(0) & 0 & 0 \\
0 & h_1(N-1) & h_1(N-2) & h_1(N-3) & \ldots & h_1(0) & 0 & 0 \\
& \ldots & 0 & 0 & h_0(N-1) & \ldots & h_0(1) & h_0(0) \\
& 0 & 0 & h_1(N-1) & \ldots & h_1(1) & h_1(0) & \ldots \\
& \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots 
\end{bmatrix}
\] (2.11)

In the case of infinite input/output signals, the analysis matrix is infinite in size and block Toeplitz. For the two-band case the blocks are two by two. Note that moving down the matrix each block is shifted across by two, which effects the decimation by two. For finite input/output signals circular convolution (filtering) is usually employed to obviate any problems associated with the signal boundaries. In this case the filter coefficients wrap around either the last or first few columns of the analysis matrix (or both), giving an analysis matrix that is block circulant. Finite block circulant matrices are considered henceforth.

The synthesis may be described in the same manner using a matrix vector product as,

\[ x' = T_s y \]

where \( T_s \) is the synthesis matrix, defined in a similar manner to the analysis matrix, and \( x' \) is the reconstructed (output) vector. For PR, if the analysis matrix is orthogonal, the synthesis matrix is simply the transpose of the analysis matrix. This is in fact the case with the CQF solution, when PR is attained.

To clarify these ideas, example 2.1 describes the relevant matrices and vectors for an orthogonal two-band system using 4-tap filters and an 8-tap input vector.

\textit{Example 2.1}

The analysis, \( y = T_a x \), is expanded as,
and the synthesis, \( x' = T_s y \) where \( T_s \) is the synthesis matrix, is expanded as,

\[
\begin{bmatrix}
\gamma_0(0) \\
\gamma_1(0) \\
\gamma_0(1) \\
\gamma_1(1) \\
\gamma_0(2) \\
\gamma_1(2) \\
\gamma_0(3) \\
\gamma_1(3)
\end{bmatrix}
= \begin{bmatrix}
h_0(3) & h_0(2) & h_0(1) & h_0(0) & 0 & 0 & 0 & 0 \\
h_1(3) & h_1(2) & h_1(1) & h_1(0) & 0 & 0 & 0 & 0 \\
0 & 0 & h_0(3) & h_0(2) & h_0(1) & h_0(0) & 0 & 0 \\
0 & 0 & h_1(3) & h_1(2) & h_1(1) & h_1(0) & 0 & 0 \\
0 & 0 & 0 & 0 & h_0(3) & h_0(2) & h_0(1) & h_0(0) \\
0 & 0 & 0 & 0 & h_1(3) & h_1(2) & h_1(1) & h_1(0) \\
h_0(1) & h_0(0) & 0 & 0 & 0 & 0 & h_1(3) & h_1(2) \\
h_1(1) & h_1(0) & 0 & 0 & 0 & 0 & h_1(3) & h_1(2)
\end{bmatrix}
\begin{bmatrix}
x(0) \\
x(1) \\
x(2) \\
x(3) \\
x(4) \\
x(5) \\
x(6) \\
x(7)
\end{bmatrix}
\]

noting that \( T_s = T_a^T \). The analysis matrix structure implies that the filtering begins four (in general \( N \) samples from the beginning of the input vector.

It is simple to verify that the analysis matrix in example 2.1, \( T_a \), is orthogonal if and only if (iff),

\[
\begin{align*}
h_0^T h_0 &= 1, h_1^T h_1 = 1, h_0^T h_0 = 0 \\
h_0(3)h_0(1) + h_0(2)h_0(0) &= 0 \\
h_1(3)h_1(1) + h_1(2)h_1(0) &= 0 \\
h_0(3)h_1(1) + h_0(2)h_1(0) &= 0 \\
h_1(3)h_0(1) + h_1(2)h_0(0) &= 0
\end{align*}
\]

or in matrix form,

\[
H^T H = I, \quad H^T WH = 0
\]

(2.12)
where $I$ is a four by four identity matrix, $0$ is a four by four zero matrix and,

$$
H = \begin{bmatrix}
h_0(3) & h_1(3) \\
h_0(2) & h_1(2) \\
h_0(1) & h_1(1) \\
h_0(0) & h_1(0)
\end{bmatrix},
W = \begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
$$

(2.13)

The two matrix equations in (2.12) actually only specify three independent equations each, giving six independent equations in total. Coupled with eight free parameters (2x4 filter coefficients) there are two degrees of freedom remaining. If these equations are satisfied then $T_x = T_a^T = T_a^{-1}$ which gives $x' = x$.

For a set of CQF's one filter specifies the other three. Hence using the CQF solution leaves only four free parameters, which are the four coefficients of one of the filters. Although there are fewer parameters there are also fewer PR equations, which are,

$$
h^T h = 1, h^T W h = 0
$$

(2.14)

leaving, again two degrees of freedom. In this thesis the filter $h$ usually refers to the highpass analysis filter, although for these PR conditions it can refer to any one of the four analysis or synthesis filters. Note that due to the symmetry of the latter equation a symmetric version of the $W$ matrix may be used as,

$$
W = \begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
$$

The beauty of the CQF relationship is that it reduces the parameter space while maintaining the same degrees of freedom. For CQF filters of length $2n (n = 2,3,...)$ there are $n$ constraints and $n$ degrees of freedom. Daubechies (1988) used this CQF relationship to derive filters with $n$ zeros at $\pi$ radians for the lowpass filter giving her famous wavelets.
2.3. ARBITRARY ANALYSIS/SYNTHESIS CONFIGURATIONS

The purpose of this section is to give a time domain characterisation of an arbitrary A/S structure. This characterisation is subsequently used to derive general results for orthogonal A/S systems. Also, this structure is used to derive results pertinent to all subband A/S structures later in this thesis.

Figure 2.3 illustrates an arbitrary subband analysis/synthesis system. The filters and signals are indicated in the time domain, commensurate with the description to be given in this section. The description of this figure is similar to that of the two-band case. The input signal is filtered into $M$ subbands of arbitrary size. Each subband is decimated as depicted. The symbol $d \downarrow$ signifies decimation by $d$: that is only every $d^{th}$ sample is retained. The signals $\{ y_0, y_1, \ldots, y_{M-1} \}$ are then processed according to the particular application. These processed signals are then upsampled and filtered, which is termed interpolation, and combined to give the output. In the absence of processing it is desirable that the output signal $x'$ is equal to, or equal within a delay to, the input signal $x$.

A so called critically sampled system is one where,

$$\sum_{k=0}^{M-1} \frac{1}{d_k} = 1$$  \hspace{1cm} (2.15)
This implies that for every \( L \) input samples there are \( L \) subband samples in total. In this thesis only critically sampled systems are considered.

An arbitrary subband analysis can be described in the time domain using a matrix vector product as in the two-band case. In the following subsections some common analysis structures are considered.

### 2.3.1. \( M \)-Band or Uniform Analysis/Synthesis

\( M \)-band analysis is the case where there are \( M \) subbands of roughly equal bandwidth, and each decimation factor is \( M \). This type of analysis system is often referred to as a uniform filter bank. A block transform, such as the discrete cosine or discrete Fourier transform, can be considered as an \( M \)-band analysis structure. This view is also adopted by Simoncelli and Adelson (in Woods 1991) among others. One purpose of this subsection is to demonstrate this concept.

Vaidyanathan (1987) offered the first \( M \)-band PR solution which used paraunitary analysis building blocks. Vetterli and Le Gall (1989 p1061) showed that the paraunitary property is equivalent to subband system having an orthogonal analysis matrix.

For the \( M \)-band case, following Vetterli and Le Gall (1989), the analysis matrix is,

\[
T_a = \begin{bmatrix}
0 & A_0 & A_1 & A_{K-1} & 0 & \ldots \\
0 & 0 & A_0 & A_1 & A_{K-1} & 0 \\
& \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & A_0 & A_{K-1} & A_{K-1} 0 \\
& & & & \ddots & & \\
& & & & & \ddots &
\end{bmatrix}
\]  \hspace{1cm} (2.16)

where the output vector is again formed by multiplexing the \( M \) subbands,

\[
y = T_a x
\]

\[
y = \begin{bmatrix}
\ldots & [y_0(-M), y_1(-M), \ldots, y_{M-1}(-M)], [y_0(0), y_1(0), \ldots, y_{M-1}(0)]^T \\
\end{bmatrix}
\] \hspace{1cm} (2.17)
Each A matrix is of size $M$ by $M$. If $T_a$ is an infinite matrix then it is block-Toeplitz; if it is finite then it is block circulant, effecting a circular convolution on the input vector. Moving down the matrix each block is shifted across $M$ columns, which effects the decimation by $M$.

Non-overlapping or block transforms are a special case of $M$-band analysis. In this case the sub-matrices $A_k$ in (2.16) are zero matrices for $k \neq 0$. In example 2.1, if $h_0(1) = h_0(0) = h_1(1) = h_1(0) = 0$, then the analysis is equivalent to a two by two block transform. In this case the orthogonal analysis matrix is,

$$T_a = \frac{1}{\sqrt{2}} \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & \ldots \\
1 & -1 & 0 & 0 & 0 & \\
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & -1
\end{bmatrix}$$

Each block of this matrix performs a two by two transform on adjacent blocks of the input vector of length two. For $M$ by $M$ transforms the blocks are of size $M$ by $M$. As another example the common 8x8 discrete cosine transform (DCT) can be described using a block diagonal analysis matrix with blocks of size 8x8.

Rao and Yip (1993, p11 and p20) define a whole class of different DCT's and discrete sine transforms (DST's). The DCT referred to in this thesis is defined there as the DCT-II (DCT type two). Similarly the DST is defined as DST-I. The other DCT's and DST's used in this thesis are referred to using the Rao and Yip nomenclature.

Princen and Bradley (1986, 1987) introduced some PR cosine modulated $M$-band transforms (analysis/synthesis systems) where each analysis and synthesis filter is a cosine modulated version of some lowpass filter prototype and is of length $L=2M$. This idea was extended by various authors to include PR cosine modulated filter banks where the filter lengths are $L=2KM$ for positive integer $K$. Malvar and Staelin (1989) described a PR lapped orthogonal transform (LOT), which is an $M$-band transform with filter lengths $L=2M$. Malvar (1992) gave fast implementations for the LOT, and two cosine modulated filter banks: a modulated lapped transform (MLT) and an
extended lapped transform (ELT). The LOT referred to in this thesis is the fast LOT type II [Malvar 1992, p168], while the MLT is described in [Malvar 1992, p178]. The ELT is a cosine modulated filter bank \((K=2)\) using a lowpass filter prototype that is described by Malvar (1992, pp184-185) with parameter \(\gamma = 0.5\). Malvar (1992, p216) suggested this parameter setting for the encoding of highly correlated sources.

### 2.3.2. Octave-band Analysis/Synthesis

A tree-structured decomposition can be used to analyse a signal. The synthesis then is the reverse tree. For example, the lowpass signal of a two-band analysis may be repeatedly analysed as illustrated in Figure 2.4.

![Figure 2.4. Octave-band or DWT analysis: tree-depth two](image)

The tree-depth is the maximum number of levels of analysis used to decompose the signal. For example, in Figure 2.4 the tree-depth is two. The lowpass signal can be further analysed giving a tree-depth of three, four and so on. An octave-band analysis, often referred to as a dyadic analysis or discrete wavelet transform (DWT), is such an approach, with an arbitrary tree-depth.

The octave-band analysis (and synthesis) can be concisely described using block circulant matrices. In this case it is convenient to store the subbands sequentially in the analysis vector rather than in a multiplexed fashion. Consider example 2.1 using a rearranged analysis vector. The first stage of the octave-band analysis is \(y_1 = T_{a1}x\), which is expanded as,
where the subscript \( a1 \) is used to indicate stage one of the octave-band analysis. Note that the analysis vector \( y_1 \) consists of the lowpass and highpass subband data stored sequentially, as opposed to the multiplexed fashion in example 2.1. This difference is reflected by a different analysis matrix. The original analysis matrix \( T_a \) is related to the new analysis matrix by,

\[
T_{a1} = PT_a
\]

(2.18)

where \( P \) is a permutation matrix. Note that the same PR conditions apply since if \( T_a \) is orthogonal then so is \( T_{a1} \) (\( P^TP = I \), being a permutation matrix). The new synthesis matrix, the transpose of the new analysis matrix, is related to the old synthesis matrix of example 2.1 through the transposed permutation.

The lowpass subband can now be analysed using a second stage of the octave-band analysis as,

\[
y_2 = T_{a2}y_1
\]
I is a four by four identity matrix and 0 is a four by four zero matrix. If the original analysis matrix $T_{a1}$ is orthogonal it is easy to verify that $T_{a2}$ is orthogonal. The analysis and synthesis can now be described as,

$$y = y_2 = T_{a2}T_{a1}x = T_{a}x$$

$$x' = T_{al}^{T}T_{a2}^{T}y = T_{a}^{T}y = T_{a}^{T}T_{a}x = x$$

The overall analysis is described by the analysis matrix $T_{a}$. If orthogonal analysis matrices are used at each stage of the tree then $T_{a}$ is orthogonal. Assuming that the synthesis matrix is the transpose of the analysis matrix, it follows that the overall analysis/synthesis system is perfectly reconstructing and orthogonal.

For longer input vectors it is possible to analyse the lowpass signal further; again PR orthogonal analysis/synthesis is attained if at each stage orthogonal analysis is used. The tree-structure is generally more efficient for implementation purposes, whereas the analysis matrix cascade, giving one orthogonal matrix, is useful for mathematical analysis.

### 2.3.3. Arbitrary Analysis/Synthesis

The term level of analysis is used in this thesis to refer to the frequency resolution of the low frequency subbands for a general subband analysis structure. For example, for a DWT the level of analysis refers to the tree-depth while for $M$-band structures it refers to $M$. A high level of analysis means that the low frequency subbands have a narrow bandwidth or high frequency resolution.

Returning to Figure 2.3, an arbitrary analysis may be characterised with a matrix vector product. Assuming that the length of the input vector, $N_x$, is a common multiple of the decimation factors, the length of the analysis vector is the same as the input vector and the analysis matrix is square. This represents critically sampling, since for $L$ input samples there are $L$ output samples. The analysis matrix can be written as,
The first \( \frac{N_x}{d_0} \) rows of the analysis matrix analyse the lowpass subband, the next \( \frac{N_x}{d_1} \) analyse the next subband and so on. For the lowpass block each row is equivalent to the adjacent row above following a circular shift of \( d_0 \) columns. This circular shift effects the decimation by \( d_0 \). The subbands could be stored in the analysis vector in a multiplexed fashion, in which case the analysis matrix would be block circulant. Thus the above matrix is a permutation of a block circulant matrix. The A/S is orthogonal if the analysis matrix is orthogonal. From a frequency domain perspective, the system would be a paraunitary A/S system.

### 2.3.4. Perfect Reconstruction and Signal Boundary Extensions

For general A/S perfect reconstruction is attained when the synthesis matrix is given by the inverse of the analysis matrix \( (T_a = T_s^{-1}) \), assuming such an inverse exists. Since the inverse of a matrix is unique, for perfect reconstruction, the analysis matrix specifies the synthesis matrix and visa-versa. As discussed in the next subsection if the analysis matrix is orthogonal, then \( T_a = T_s^T \). This is referred to as orthogonal A/S. More generally, if \( T_s = T_s^{-1} \) the A/S is referred to as biorthogonal since \( T_s T_s = I \).
Note that A/S formulation presented here is not restricted to time invariant filters. Unless the analysis (and synthesis) matrix is block circulant, or a permutation of a block circulant matrix, the filters are effectively time varying.

The synthesis reconstructs the signal as a weighted sum of basis vectors. The weights are the subband coefficients, and the basis vectors lie in the columns of $T_r$. For time invariant systems, due to the upsampling operation, these basis vectors are shifted versions of the synthesis filters. The shift for synthesis filter $k$ is $d_k$, the interpolation factor for subband $k$. Hence one observes that subband analysis is a decomposition of a signal onto the synthesis basis vectors. For orthogonal A/S each analysis filter, which is used to calculate the weight for each synthesis basis vector, is the same as the corresponding synthesis basis vector.

Circular convolution, as effected by circulant analysis matrices, is equivalent to replicating the finite length input signal in time, making an infinite periodic signal, and analysing this infinite signal. Since the input signal is now periodic and infinite, the output subband signals are infinite and periodic. The analysis vector simply contains the subband signals for one of these periods. The synthesis, employing circular convolution, effectively replicates the finite subband signals, giving the above infinite length subband signals, and synthesises these infinite subbands. The reconstructed signal is simply one period of this output signal, which for PR is the same as finite input signal.

An even periodic, as opposed to a circular, data extension may be used with some filter banks with linear phase filters [Smith and Eddins 1987, 1990]. The finite input signal is first reflected about one of its boundaries, and then this new finite signal is replicated in time. The analysis and synthesis operate on a finite portion of this infinite signal. The filters are positioned in the same place in the analysis/synthesis matrices as with the circular extension method except at the boundaries of the matrix. Instead of wrapping around the columns, the filters are reflected about the last and first column. For critical sampling, where the length of the analysis vector is equal to that of the original input vector, only some linear phase filters may be used for perfect reconstruction with an even periodic extension. [Smith and Eddins 1987, 1990].
2.4. ORTHOGONAL ANALYSIS MATRIX PROPERTIES

If the analysis matrix is orthogonal then the A/S system is said to be orthogonal. If the analysis matrix is orthogonal then for PR the synthesis matrix is the transpose of the analysis matrix and is orthogonal. Under an assumption of PR this definition of an orthogonal filter bank is equivalent to that of Soman and Vaidyanathan (1993, p1832). Vetterli and Le Gall (1989) show that orthogonality and the paraunitary property are equivalent.

Orthogonal analysis/synthesis enjoys several properties analogous to properties enjoyed by orthogonal block transforms. The following subsections illustrate these properties and prove some new results which are useful in the design of orthogonal analysis/synthesis systems.

2.4.1. Mean Square Error / Energy Conservation

Quantisation in the subband domain involves the approximation of the analysis vector \( y \) by some "quantised" version \( \hat{y} \). Synthesising this approximation gives an output vector \( \hat{x} \). The mean square error (MSE) is defined as the average square error between the components of the input to the A/S system, \( x \), and the components of the output, \( \hat{x} \). Considering these vectors as wide sense stationary random vectors gives [Papoulis 1991, p329],

\[
MSE = E \left\{ \frac{1}{N_x} (x - \hat{x})^T (x - \hat{x}) \right\}
\]

If the analysis matrix is orthogonal then,

\[
MSE = E \left\{ \frac{1}{N_x} (T_a^T y - T_a^T \hat{y})^T (T_a^T y - T_a^T \hat{y}) \right\}
\]

\[
= E \left\{ \frac{1}{N_x} (y - \hat{y})^T T_a T_a^T (y - \hat{y}) \right\}
\]

\[
= E \left\{ \frac{1}{N_x} (y - \hat{y})^T (y - \hat{y}) \right\}
\]
Hence the system MSE is equal to the MSE in the subband domain. It has implicitly been assumed that the quantisation error vector $\mathbf{y} - \hat{\mathbf{y}}$ is wide sense stationary. This is a well known result that applies to orthogonal block transforms [Jayant and Noll 1984, p525]. The only difference here is that a block transform has been used to describe a finite input arbitrary (orthogonal) analysis/synthesis system. This idea is considered in more detail in Chapter 3.

Setting the approximation, or quantisation vector to zero gives,

$$E\{x^T x\} = E\{y^T y\}$$  \hspace{1cm} (2.19)$$

and hence there is energy conservation between the subband (analysis) and original domains. Soman and Vaidyanathan (1993) derive these results using a paraunitary A/S system assumption which is equivalent to an orthogonal A/S assumption. The above results obviously hold in a deterministic environment, which is the environment under which the MSE is usually calculated for a particular instance of an encoding of an image.

2.4.2. Filter bank Response to a Constant Input

The response of a filter to a constant input of unit magnitude is the frequency response at zero frequency. This is termed the DC component of a filter (or vector) and is denoted by,

$$\text{DC}(\mathbf{h}) = \sum_{n=0}^{N-1} h(n) = H(e^{j\omega})\bigg|_{\omega = 0}$$

where $N$ is the length of the vector $\mathbf{h}$ of filter coefficients. For a set of analysis (or synthesis) filters the DC subband or DC basis vector is defined as the subband or basis vector with the largest DC component. This is usually the lowpass subband.

Kronander (1989a, 1989b) and several other authors have suggested that for image coding every analysis filter, except the lowpass, should have a zero at zero frequency (DC). The reasons why this is so are discussed later in Section 6.3.1 of Chapter 6. An analysis filter set with this property is referred to as a zero-DC constrained or zero-constrained filter bank.
Using a zero-constrained filter bank means that a constant signal is represented wholly by the lowpass signal, or equivalently can be reconstructed from the lowpass signal only. Clarke (1983a) showed that a sufficient condition for a block or non-overlapping transform, to be zero-constrained is that the DC basis vector is constant. In fact it can be shown that this is a necessary condition. First, Clarke's result is generalised to arbitrary orthogonal analysis. The following theorem has been given by Daubechies (1988) in a different form for the two-band case. Note that it is assumed that the filter bank is perfectly reconstructing.

Theorem 2.1: Consider an orthogonal filter bank with DC subband filter $h_0$. A necessary and sufficient condition for the filter bank to be zero-constrained is that $\text{DC}(h_0) = \sqrt{d_0}$, where $d_0$ is the decimation factor of the DC subband.

Proof: Consider the output of such an orthogonal filter bank given a constant, unit magnitude input. This output is given by,

$$y = \left[\sqrt{d_0}, \sqrt{d_0}, \ldots, \sqrt{d_0}, \text{DC}(h_1), \ldots, \text{DC}(h_1), \ldots, \ldots, \ldots, \ldots, \text{DC}(h_{M-1}), \ldots, \text{DC}(h_{M-1})\right]^T$$

Energy conservation (2.19) gives,

$$y^Ty = x^Tx = N_x$$

$$= \sum_{n=0}^{N_x-1} \left(\sqrt{d_0}\right)^2 + \sum_{n=0}^{N_x-1} \left(\text{DC}(h_1)\right)^2 + \ldots + \sum_{n=0}^{N_x-1} \left(\text{DC}(h_{M-1})\right)^2$$

$$= N_x + \frac{N_x}{d_1}(\text{DC}(h_1))^2 + \ldots + \frac{N_x}{d_{M-1}}(\text{DC}(h_{M-1}))^2$$

Hence the DC component of $h_1, \ldots, h_{M-1}$ must be zero. In other words,

$$\text{DC}(h_1) = \ldots = \text{DC}(h_{M-1}) = 0$$

This shows that the above property is sufficient to enforce zeros at DC for non-DC filters. To show the necessity, if the DC component of $h_1, \ldots, h_{M-1}$ is zero then,
\[ y^T y = x^T x = N_x \]
\[ = \frac{N_x}{d_0} (\text{DC}(h_0))^2 \]
\[ \Rightarrow \text{DC}(h_0) = \sqrt{d_0} \]

This theorem places an interesting constraint on the DC basis function of a zero-constrained block transform as,

**Corollary 2.1**: For the block, or non-overlapping transform, a necessary and sufficient condition for zero constraint is that the DC basis vector is constant.

**Proof**: Consider maximising the DC component of the DC basis vector given a unit energy constraint, (required by an orthogonal transform). Forming a Lagrangian gives,

\[ L(h_0, \lambda) = h_0^T 1 - \lambda (h_0^T h_0 - 1) \]

where 1 is a vector of ones and \( h_0 \) denotes the DC basis vector. Differentiating and setting the gradient to zero gives,

\[ \frac{\partial L}{\partial h} = 1 - 2\lambda h_0 = 0 \Rightarrow h_0 = \frac{1}{2\lambda} [1 \ 1 \ \ldots \ 1]^T \]

which is a necessary condition for maximising the DC component of \( h_0 \). Using the unit energy constraint gives \( \frac{1}{2\lambda} = \frac{1}{\sqrt{d_0}} \), noting that for a block transform the length of the DC basis vectors is \( d_0 \) (and is the length of all the basis vectors). The maximum DC component is thus,

\[ \text{DC}(h_0) = d_0 \frac{1}{\sqrt{d_0}} = \sqrt{d_0} \]

It follows that if the DC basis vector is not constant then the DC component will be less than \( \sqrt{d_0} \), violating the necessary condition for a zero-constrained filter bank. If the DC basis vector is constant then for unit energy, as required by orthogonality, \( \frac{1}{2\lambda} = \frac{1}{\sqrt{d_0}} \).
As noted above, Clarke (1983a) showed the sufficiency of this corollary. Some examples that pertain to Corollary 2.1 are the discrete sine transform (DST) and the Karhunen-Loeve Transform (KLT) for an AR(1) source of correlation \(-1<\rho<1\). Both these transforms have a non constant DC basis vector, and hence both have non DC basis vectors with non-zero DC components. The discrete cosine transform (DCT) has a constant DC basis vector and hence all non-DC basis vectors have a zero DC component.

2.5. TWO-DIMENSIONAL ANALYSIS/SYNTHESIS

Two-dimensional analysis/synthesis is required for subband image coding since an image is represented by a two-dimensional data array. A four-band two-dimensional analysis/synthesis system (filter bank) is illustrated in Figure 2.5. The description of a two-dimensional filter bank is essentially the same as that for a one-dimensional filter bank. The input signal is analysed into various two-dimensional subbands which are then decimated. Decimation involves subsampling the subbands by a specified amount in each dimension. It is not necessary that the subsampling is the same in both dimensions, although it is often the case. In Figure 2.5 there are four subbands. The synthesis is the same reverse operation as for a one-dimensional filter bank.
The two-dimensional filter bank can also be described in either the time or frequency domain. From this description the necessary filter constraints for PR can be derived. By far the simplest analytical method of attaining PR is using separable two-dimensional filters that are the product of one-dimensional PR filters [Vetterli 1984]. For example consider a PR one-dimensional two-band filter bank using analysis filters $h_0$ and $h_1$ and synthesis filters $g_0$ and $g_1$. Setting,

$$
\begin{align*}
    h_{00}(n,m) &= h_0(n)h_0(m), & g_{00}(n,m) &= g_0(n)g_0(m) \\
    h_{01}(n,m) &= h_0(n)h_1(m), & g_{01}(n,m) &= g_0(n)g_1(m) \\
    h_{10}(n,m) &= h_1(n)h_0(m), & g_{10}(n,m) &= g_1(n)g_0(m) \\
    h_{11}(n,m) &= h_1(n)h_1(m), & g_{11}(n,m) &= g_1(n)g_1(m)
\end{align*}
$$

gives a two-dimensional four-band PR filter bank using the structure of Figure 2.5, where the decimation for each subband is effected by subsampling by two in both dimensions. The four subbands associated with these filters are referred to as LL, LH, HL and HH respectively where L refers to lowpass and H refers to highpass.

Using separable filters it is possible to analyse the input data rows first and then analyse the resulting data columns. Each row is analysed independently with a two-band one-dimensional filter bank. This results in two, two-dimensional subbands. The columns of these subbands are then analysed similarly to give four subbands in total. In
this form PR is obviously attained if the synthesis is the reverse cascade of these one-
dimensional analysis/synthesis operations.

Consider two one-dimensional filters, $h_{0}(n)$ and $h_{1}(n)$, with unit energy or norm so that,

$$\sum_{n}|h_{0}(n)|^{2} = \sum_{n}|h_{1}(n)|^{2} = 1$$

The energy of a two-dimensional filter, $h(n,m)$, generated using these two one-
dimensional filters is,

$$\sum_{n,m}|h(n,m)|^{2} = \sum_{n,m}|h_{0}(n)|^{2}|h_{1}(m)|^{2} = 1$$

Hence the two-dimensional filter is also of unit energy. The separable filters employed
in this thesis are usually generated from unit energy one-dimensional filters.

A filter bank employing separable filters is termed a separable filter bank. Bamberger
(1992) and Bamberger and Smith (1992) have investigated the use of non-separable
two-dimensional filter banks for various applications. They concluded that for subband
image coding that no significant subjective or objective gain achieved by using these
non-separable filters. Also Simoncelli and Adelson (1990) discuss QMF type filter
banks based on a hexagonal sampling lattice which offer superior orientation
selectivity. A drawback of these filters is the extra computation required to convert
existing rectangular sampling grids to hexagonal ones. There are other open questions
in this area of non-separable filter banks, which however are not the topics of this
thesis.

In this thesis only two-dimensional separable filter banks based on a rectangular image
sampling grid are considered. Although the separable filter banks are a subset of
general filter banks they offer certain advantages: namely an efficient implementation
via separate row/column analysis, simpler filter design, and simpler system analysis.
Also as concluded by Bamberger and Smith, at present, in terms of image coding a
non-separable filter bank does not appear to offer any advantages over a separable filter
bank.
One-dimensional filter properties are considered in the time and frequency domains in this thesis. Since separable filters are used for two-dimensional analysis/synthesis, these filters are characterised by the properties of the two corresponding one-dimensional filters. The one-dimensional time and frequency domains obviously correspond to the two-dimensional spatial and spatial frequency domains when applied to the two-dimensional filtering of images.

2.6. FILTER BANK TIME WIDTHS

Simoncelli and Adelson (in Woods 1991, Chapter 4 p182) conclude that for image compression purposes the basis functions of a filter bank should be localised in both the spatial and spatial-frequency domains. The time width of a filter is an attempt to quantify its time resolution (or spatial resolution as applied to images). The time width of a FIR filter of length $N$ is defined as [Marple 1987],

$$
\sigma_r^2 = \frac{\sum_{n=0}^{N-1} (n - \bar{n})^2 |h(n)|^2}{\sum_{n=0}^{N-1} |h(n)|^2}
$$

where $\bar{n}$ is the mean time,

$$
\bar{n} = \frac{\sum_{n=0}^{N-1} n|h(n)|^2}{\sum_{n=0}^{N-1} |h(n)|^2}
$$

Usually $h(n)$ is normalised to unit energy, in which case the time width is the variance of a probability mass distribution defined by $|h(n)|^2$.

A filter time width is a measure of the spread of its impulse response. Obviously for image coding purposes, this time width is really a one-dimensional spatial width. Since one-dimensional filter properties are generally considered in this thesis, and the actual label attached to a dimension, such as space or time, is largely irrelevant for purposes of this discussion, the impulse response spread is referred to as time width.

A frequency width, sometimes referred to as bandwidth, can be defined in a similar manner to the time width as,
\[ \sigma_\omega^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \omega^2 |H(e^{j\omega})|^2 d\omega \]

Note that the bandwidth interpretation in this case applies to lowpass filters only. The uncertainty principle states that [Marple 1987],

\[ \sigma_t^2 \sigma_\omega^2 \geq \frac{1}{4} \]

The time widths of the filters employed in a subband analysis method are dependent somewhat on the subband analysis structure. For example, given a particular subband bandwidth, the time width of the corresponding filter is lower bound by the uncertainty principle. As a general rule of thumb, as the frequency resolution increases the time resolution decreases. However it is important to realise that this is a general trend only and filters with the same bandwidth may have widely varying time widths.
CHAPTER 3:

SUBBAND CODING GAIN

3.1. INTRODUCTION

The coding gain metric, $G_{sbc}$ (or $G_{tc}$), is theoretical measure of the coding performance of a subband analysis/synthesis system for a given source model. This metric is commonly used to compare various block transforms and subband methods. Jayant and Noll (1984, p527) and Clarke (1985, p179) defined and derived a coding gain equation for orthogonal block transforms. Jayant and Noll (1984, p491) also defined and derived a subband coding gain equation for an $M$-band ideal filter bank. Although several assumptions were made in the derivation, it is generally perceived as being useful: to quote Malvar (1992, p241), "Although $G_{tc}$ may not predict directly the performance of a transform coder, differences in $G_{tc}$ from one transform to another generally correlate well with differences in practical performance."

Pearlman [in Woods 1991 pp27-29] formalised the derivation of the coding gain for ideal subband analysis structures using rate distortion theory. It was assumed that the input is a stationary Gaussian signal and that the rate is high. The subbands may be of arbitrary bandwidth but must be formed using ideal brick-wall filters. Soman and Vaidyanathan (1993) extended the block transform coding gain to include any orthogonal or paraunitary subband analysis. It must be noted that several authors had been using this extension for some time prior to this publication. Katto and Yashuda (1991) generalised the coding gain to a unified coding gain, suitable for any perfect reconstruction (PR) filter bank, including DPCM systems. Although the unified coding gain is very general it is worth noting (and shown later) that the assumptions required for biorthogonal analysis methods are stricter than those required for orthogonal methods. However, for sufficiently high rates all assumptions made are reasonable.

Akansu and Haddad (1990) showed that for some transforms the coding gain is independent of the sign of $p$ using an AR(1) source, and noted that this is not the case for the DCT. Malvar (1992) and Akansu and Wadas (1992) used the coding gain to
compare the DCT, LOT and ELT using an AR(1) model of high correlation, typically $p=0.95$, which was used as a basic model of still imagery. It was demonstrated that the performance hierarchy is ELT, LOT and DCT commensurate with the frequency resolving hierarchy of these subband methods. The coding gain of an ideal dyadic analysis or discrete wavelet transform was investigated by de Queiroz and Malvar (1992). They showed that the wavelet coding gain is asymptotic toward a lower bound than that of most block transforms (the asymptotic performance refers to the gain performance as the analysis level increases). However, one observes from their results that for highly correlated AR(1) sources the difference is small.

3.1.1. Overview of Chapter 3

Following this overview, the remainder of this introduction is concerned with background material. The coding gain for an arbitrary subband system is defined in Section 3.1.2. Some statistical image modes are then discussed. In particular correlation models and their relation to power spectral densities (PSD's) are considered. The correlation and PSD of the subband signals are derived given a correlation (or PSD) model of the input signal to a filter bank. Finally a lowpass to highpass filter transformation is considered.

In Section 3.2 a simple coding scheme which employs a PCM quantiser for each subband is considered. Using a rough quantisation noise model, the bit allocation among subbands required to minimise the overall system mean square error (MSE) is derived. This bit allocation is then used to derive the unified coding gain of Katto and Yashuda (1991). The motivation for giving the derivation here is several fold. The assumptions required to simplify the unified coding gain equation are given and it is shown that in the orthogonal case some of these may be relaxed. The relaxed assumptions are then in agreement with those given by Soman and Vaidyanathan (1993) for paraunitary or orthogonal subband coders. The unified coding gain is used to demonstrate that the KLT is the optimum transform among all transforms, including non-orthogonal ones. Finally it is shown that by normalisation of the synthesis filters to unit energy a simple optimum bit allocation scheme for biorthogonal (or general) subband coders may be used.

The coding gain considered previously in the literature assumes a high rate, and as such is independent of the rate. In Section 3.3 a rate constrained coding gain is considered.
In this case the assumptions used are generally valid only for orthogonal subband analysis methods. The rate constrained coding gain is demonstrated to be asymptotic to a lower level than the unconstrained case. Also, for highly correlated sources operating at very low rates it is shown that the rate constrained coding gain is almost negligible until quite high analysis levels. The minimum level of analysis required for a near optimum coding gain increases as the rate decreases for such sources. This characteristic has important implications in the design of subband coders for high definition television (HDTV) and other high resolution image systems.

In Section 3.4 it is shown that transforms or subband methods with pairwise symmetric filters have a symmetrical coding gain about \( \rho = 0 \) for an AR(1) source. This symmetrical response is dependent on the magnitude response of the subband filters. The asymmetric coding gain of the DCT is related to the fact that it performs so well for highly correlated sources. For negatively correlated sources the DCT performs poorly. The magnitude response of the DCT and other transforms is investigated in relation to this coding performance. It is demonstrated that the improved subband resolution of overlapping transforms means that a good coding performance for all values of correlation (\( \rho \)) can be attained. The disadvantage of the overlapping transforms is an increased computational expense and less time resolution or localisation.

### 3.1.2. Definition of Coding Gain

The signal to noise ratio of a coding/decoding system (codec) is defined as,

\[
\text{SNR(dB)} = 10 \log_{10} \left( \frac{\sigma_x^2}{\sigma_r^2} \right)
\]

where \( \sigma_x^2 \) is the input signal variance and \( \sigma_r^2 \) is the variance of the error between the decoded signal and the input signal [Jayant and Noll 1984, p6].

Jayant and Noll (1984, p491 and p528) define the coding gain for a subband codec as,

\[
G_{\text{SBC}} = \frac{\sigma_{r,\text{PCM}}^2}{\sigma_{r,\text{SBC}}^2}
\] (3.1)
where $\sigma^2_{r,PCM}$ is the variance of the reconstruction error of a PCM system and $\sigma^2_{r,SBC}$ is the variance of the reconstruction error of a subband coding system. For orthogonal block transforms Jayant and Noll use the familiar $G_{TC}$, where the subscript TC refers to transform coding. In this thesis the subscript SBC is used to refer to general perfect reconstruction subband coding systems, of which orthogonal transforms are a subset.

This metric is termed the subband coding gain since,

$$\text{SNR}_{SBC} (\text{dB}) = \text{SNR}_{PCM} (\text{dB}) + 10 \log_{10} G_{SBC}$$

(3.2)

### 3.1.3. Statistical Image Models

The normalised autocorrelation function of a wide sense stationary (WSS) discrete random process $x(n)$ is by definition,

$$r_{xx}(m) \equiv \frac{E\{x(n+m)x^*(n)\}}{E\{|x(0)|^2\}}$$

where $E\{x\}$ is the expected value of $x$. The denominator is simply a normalisation factor, so that the zero lag autocorrelation is unity. This normalised autocorrelation is usually referred to as correlation.

An $N$th order autoregressive process (AR($N$)) is generated by passing white noise through an all pole filter which has memory of only the $N$ preceding outputs [Jayant and Noll 1984, p62]. A Gauss-Markov source is an AR(1) source where input source is white Gaussian noise.

The correlation of an AR(1) source is,

$$r_{xx}(m) = e^{-\alpha|m|} = \rho^{|m|}$$

(3.3)

where $\alpha$ or $\rho$ is a correlation parameter. In this thesis the correlation of an AR(1) source refers to $\rho$. This model, with $\rho=0.95$, has been used as a basic model for the one-dimensional correlation properties of image data. For example see [Clarke 1985] or [Jayant and Noll 1984]. A more general model, is of the form,
Clarke (1985, p24) stated that with optimisation of \( \alpha \) and \( \gamma \) this model can represent actual measured values of one-dimensional correlation for a "well behaved" test image. However it was also stated that this model is not a great deal of use in the general case.

The normalised autocorrelation, or simply correlation, of a two-dimensional wide sense stationary random process is defined in a similar manner to the one-dimensional case. An obvious candidate for a two-dimensional model for image data is a separable form of an AR(1) process. In this case the correlation is,

\[
 r_{xx}(n, m) = e^{-\alpha_1 |n| e^{-\alpha_2 |m|}}
\]

The loci of constant correlation for this model are straight lines (for positive \( n \) and \( m \)). However, in terms of image data, the diagonal correlation falls off too rapidly with distance [Clarke 1985, p29]. This suggests the use of an isotropic model, whose loci of constant correlations are circles. However now the diagonal correlation is over estimated. Therefore a generalised correlation model has been devised to take into account these observations as [Mauersberger, 1979] [Clarke 1985, p268],

\[
 r_{xx}(n, m) = e^{-\alpha_a |n| + \alpha_h |m|} \rho^{(|n| + |m|)}
\]

where \( \alpha_a \) and \( \alpha_h \) are vertical and horizontal correlation parameters. As indicated in (3.4) it is usually assumed that these factors are equal. The separable and isotropic models can be generated using this model by appropriate selection of \( \beta \) and \( \gamma \). Clarke (1985, p30) noted that optimising the parameters for a particular image results in a model that is a surprisingly good fit for the image in question. Mauersberger demonstrated that independent of the correlation \( \rho \) of the image, a parameter setting of \( \beta = 1.75 \) and \( \gamma = 0.75 \) is valid for most images. The generalised model with these parameters is henceforth termed 2DG. For future reference the correlation function for the 2DG model is,

\[
 r_{xx}(n, m) = \rho^{(|n|^{.75} + |m|^{.75})^{.75}}
\]
3.1.4. Power Spectral Density and Correlation

The power spectral density (PSD) of a stationary source is defined as the Fourier transform of the correlation function, which for a discrete source is denoted as \( S_{xx}(e^{j\omega}) \). Thus,

\[
S_{xx}(e^{j\omega}) = \sum_{n=-\infty}^{\infty} r_{xx}(n)e^{-j\omega n}
\]

\[
= r_{xx}(0) + 2\sum_{n=1}^{\infty} r_{xx}(n) \cos \omega n
\]

since \( r_{xx}(n) \) is an even function and hence the PSD is a real function. A correlation function is by definition positive semidefinite, which means that,

\[
S_{xx}(e^{j\omega}) \geq 0
\]

Positive definite sources only are considered in this thesis which means that the PSD is strictly greater than zero.

Consider a linear filter and decimator, with discrete random signal \( x(n) \) input, as depicted in Figure 3.1.

It is straightforward to show that the correlation, \( r_{yy}(m) \), of signal \( y(n) \) is [Papoulis 1977, p321],

\[
r_{yy}(m) = r_{xx}(m) * h(m) * h(-m)
\]
where \( r_{xx}(m) \) is the correlation of the input signal \( x(n) \) and \( * \) denotes the convolution operator. It follows that the PSD of \( y \) is,

\[
S_{yy}(e^{j\omega}) = |H(e^{j\omega})|^2 S_{xx}(e^{j\omega})
\]

where \( S_{xx}(e^{j\omega}) \) is the PSD of \( x \). The correlation of the signal \( z \) is [Mintzer and Liu 1973, Appendix A],

\[
r_z(n) = r_{yy}(Dn)
\]

Assuming zero mean, the zero lag correlation of \( y \) is its variance and is also the variance of \( z \). In this case the output variance is,

\[
\sigma_z^2 = \sigma_y^2 = r_{yy}(0) = h^\top R_{xx} h
\]

where \( h \) is an \( N \) length vector of the filter coefficients as described in Chapter 2, and \( R_{xx} \) is the input correlation matrix of dimension \( N \times N \) whose \( (n,m) \)th entry is,

\[
[R_{xx}]_{nm} = r_{xx}(n-m)
\]

The PSD/correlation Fourier transform relationship gives,

\[
\sigma_z^2 = \sigma_y^2 = h^\top R_{xx} h = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{yy}(e^{j\omega}) e^{-j\omega 0} d\omega = \frac{1}{2\pi} \int_{0}^{\pi} |H(e^{j\omega})|^2 S_{xx}(e^{j\omega}) d\omega.
\]

This equation shows that the output variance is dependent on the magnitude of the filter transfer function only, and is independent of the phase.

The PSD of \( z \) is given by [Mintzer and Liu 1973, Appendix A],

\[
S_{zz}(e^{j\omega}) = \frac{1}{D} \sum_{k=0}^{D-1} S_{yy}(e^{j(\omega+2\pi k)/D}) = \frac{1}{D} \sum_{k=0}^{D-1} |H(e^{j(\omega+2\pi k)/D})|^2 S_{xx}(e^{j(\omega+2\pi k)/D})
\]

In effect it is an expanded version of the PSD of \( y \) (corresponding to \( k=0 \)) plus various aliased spectra of \( y \) (corresponding to \( k=1, 2, \ldots, D-1 \)).
The PSD for an AR(1) source is given by the Discrete Fourier Transform of equation (3.3) as [Jayant and Noll, 1984 p64],

\[ S_{xx}(e^{j\omega}) = \frac{1-p^2}{1+p^2-2p\cos\omega} \]  

(3.9)

where the signal variance, \( \sigma_x^2 \), is normalised to unity.

For a unit variance, zero mean AR(1) source input, if \( h(n) \) is an ideal filter then the output variance, either that of \( y \) or \( z \) in Figure 3.1, is given by [de Queiroz and Malvar 1992],

\[
\sigma_i^2 = 2 \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} K^2 \frac{1-p^2}{1+p^2-2p\cos\omega} \, d\omega \\
= \frac{2K^2}{\pi} \tan^{-1}\left(\frac{1+p}{1-p}\frac{\omega}{2}\right) \bigg|_{\omega=\omega_2}^{\omega=\omega_1}  
\]  

(3.10)

where \( \omega_1, \omega_2 \) are the passband (or stopband) edges and \( K \) is the passband gain. This equation is subsequently used to calculate the subband variances for an ideal filter bank.

The linear filter and decimator depicted in Figure 3.1 represents the filtering and decimation operation of one channel of an arbitrary subband analysis. Equations (3.7) and (3.8) therefore give the subband variance and PSD respectively. For example, consider the DC subbands of a \( M=8 \) DCT, \( M=8 \) ELT, and \( M=8 \) ideal \( M \)-band filter bank. The filter \( h(n) \) in Figure 3.1 is then the DC basic vector for the DCT, ELT or ideal filter bank respectively. The decimation factor is \( D=8 \) in all cases. Figure 3.2 illustrates the PSD of these DC subbands for a unit variance AR(1) input source of correlation \( \rho=0.95 \). It is evident that the PSD's of these three signals are very similar. Note that the energy, or variance, is maximum for the ideal case and minimum for the DCT as expected. The energy packing efficiency is (generally) better for subband schemes with longer filters.
This figure demonstrates the benefit of considering a block transform as a filter bank. The DC subband of the DCT is obviously very similar to that of the ideal and ELT filter banks. For the other subbands, especially the high frequency subbands, the difference between PSD's is greater. However, from a magnitude perspective at least, since these subbands contain very little energy, the increased difference is not particularly significant.

3.1.5. Lowpass to Highpass Filter Transformations

The lowpass to highpass transformation of a filter \( h(n) \) to a filter \( g(n) \) is defined as,

\[
g(n) = (-1)^n h(n) \quad \text{or} \quad g(n) = (-1)^n h(N-1-n)
\]

(3.11)

where in the latter case it is assumed that \( h(n) \) is a causal FIR filter of length \( N \). It is shown in Appendix B that for real \( h(n) \) the frequency response of \( g(n) \) is related to that of \( h(n) \) by,

\[
|G(e^{j\omega})| = |H(e^{j(\pi-\omega)})|
\]

(3.12)
The magnitude spectrum of $g$ is that of $h$ reflected in $\pi/2$. If $h$ is a lowpass filter then $g$ will be a highpass filter and visa-versa.

A source may be transformed using a lowpass to highpass transformation by simply considering $h(n)$ in the first equation in (3.11) as a source. It is demonstrated in Appendix B that in this case the correlation sequence undergoes a lowpass to highpass transformation and hence the PSD is reflected in $\pi/2$ radians. For example the PSD of an AR(1) source with correlation $\rho = -0.95$ is equal to the PSD of an AR(1) source of correlation $\rho = 0.95$ reflected in $\pi/2$ radians (or 0.25 cycles/sample normalised frequency). It follows that as a positive value of correlation represents a dominance of low frequencies, a negative correlation represents a dominance of high frequencies. Most real world signals exhibit a positive correlation, although difference signals, such as those generated by image sequence frame differences may exhibit a negative correlation [Akansu and Haddad 1990]. The relationship between the PSD's of an AR(1) source of correlation $\rho = -0.95$ and $\rho = 0.95$ is illustrated in Figure 3.3. Note that rho refers to the correlation $\rho$.

![Figure 3.3. PSD of an AR(1) source for $\rho(rho)=0.95$, $\rho=0.7$, $\rho=0.0$, $\rho=-0.95$](image)

Also shown in Figure 3.3 are the PSD's of AR(1) sources with correlation $\rho = 0.7$ and $\rho = 0.0$. It is evident that for increasing positive correlation the PSD is increasingly dominant at low frequencies. Due to the lowpass to highpass relationship between negative and positive correlation it holds that for negative correlation decreasing
towards -1 the PSD is increasingly dominant at high frequencies. The PSD for ρ=0.95 exhibits a steep PSD at low frequencies, which are also dominant, with an increasing flat PSD for higher frequencies.

3.2. CODING GAIN AND BIT ALLOCATION

One task associated with the design of a subband coder is the selection of an appropriate method to quantise or encode the subbands. The simplest solution is to treat the subbands as independent signals. For example it is possible to use a separate PCM quantiser for each subband. For subband analysis/synthesis structures that generate subbands with low correlation, there is little gain in using more sophisticated scalar quantisation techniques such as DPCM. If the subbands are considered independently then one has to decide how to allocate quantisation bits amongst the subbands. Usually this problem is posed as minimising the distortion of the reconstructed (synthesised) image, for a fixed overall bit rate. The overall bit rate (or simply rate) is defined as the average number of bits per pixel (bpp) required to transmit an image or sequence of images.

In this section, using a general subband structure, the bit allocation among subbands that minimises the overall distortion for a fixed bit rate is considered. It is assumed that the subbands are encoded independently using simple PCM quantisation. As an extension of this problem the unified coding gain metric of Katto and Yashuda (1991) is derived. This metric is a measure of the improvement of a subband coder using subband PCM over PCM quantisation of the original (fullband) signal. The assumptions under which it is possible to simplify the bit allocation and the coding gain equations are given in the derivation. It is shown that for orthogonal subband schemes some of these assumptions may be relaxed and that the resulting metric is in agreement with that given by Soman and Vaidyanathan (1993) for paraunitary or orthogonal subband coders under the same (relaxed) assumptions.

It is well known that the KLT is the optimum orthogonal transform in terms of coding gain. Using the unified coding gain, this result is extended by demonstrating that the KLT is optimum over all perfect reconstruction block transforms. Finally it is shown that a simple bit allocation scheme used for orthogonal subband schemes may be applied to biorthogonal subband schemes if the synthesis filters are normalised to unit energy.
3.2.1. Background and Assumptions

Figure 3.4 illustrates a general subband coding system which was described in Chapter 1. A general subband structure is used so that the following results may be applied to any critically sampled subband analysis/synthesis system. It is assumed in the ensuing analysis for simplicity that each source has a zero mean. Although one-dimensional filters are indicated, it is shown later that the following analysis is applicable to higher dimension filter banks.

Following Figure 3.4 the reconstruction error is by definition,

\[ r(n) \equiv x(n) - \hat{x}(n) \]  \hspace{1cm} (3.13)

and the \( k \)th subband quantisation error is,

\[ q_k(n) \equiv y_k(n) - u_k(n) \]

Following Katto and Yashuda (1991) let the variables \( A_k \) and \( S_k \) be defined by,
\[ \sigma_k^2 = A_k \sigma_x^2 \quad (3.14) \]

and

\[ \overline{\sigma}_r^2 = \sum_{k=0}^{M-1} S_k \sigma_{q_k}^2 \quad (3.15) \]

where \( \sigma_k^2 \equiv \sigma_{q_k}^2 \) is the variance of the \( k \)th subband, \( \sigma_{q_k}^2 \) is the quantisation error variance of the \( k \)th subband, and \( \overline{\sigma}_r^2 \) is the average reconstruction error variance. Strictly speaking the reconstruction error signal is nonstationary. However it is shown in Appendix B that for wide sense stationary (WSS) \( x(n) \) and jointly WSS \( q_k(n) \), it is cyclostationary. Therefore the reconstruction error is defined as the average reconstruction variance over one period. Normalising the input variance to unity gives \( A_k = \sigma_k^2 \) and hence from equation (3.6),

\[ A_k = h_k^T R_x h_k \]

In Appendix B it is shown that,

\[ S_k = \frac{1}{d_k} \sum_s |g_k(s)|^2 \quad (3.16) \]

where \( d_k \) and \( g_k(n) \) is the decimation factor and synthesis filter respectively of the \( k \)th subband, if the following assumptions hold,

*The input signal and quantisation noise are zero mean WSS signals.*

*The subband quantisation error noise is white and uncorrelated between subbands.*

The first assumption defines a class of signals for which (3.16) is valid, while the second assumption is valid only for high bit rates. Jayant and Noll [1984, p158] state that a white quantisation noise approximation is valid for rates \( R \geq 2 \) bits per pixel (bpp) if the lag one autocorrelation is not too close to +1. The latter criterion will generally be met by subband signals, since decorrelation, or spectral whitening, can be considered as one of the goals of subband analysis. Woods and Naveen (1992) arrived at equation (3.16), using the same assumptions, for the case of a (tree-structured) two-band subband analysis/synthesis.
For an orthogonal filter bank it is shown in Appendix B that the latter assumption can be relaxed to encompass jointly WSS subband quantisation error signals. In this case the second assumption is independent of the rate.

The synthesis filters can be normalised to unit energy (unit Euclidean norm) without loss of generality. Hence setting,

$$\sum_s |g_k(s)|^2 = 1$$

gives, under the assumptions listed above,

$$S_k = \frac{1}{d_k}$$  \hspace{1cm} (3.17)

Assuming PCM quantisers for each subband, the quantisation error variance can be modelled as,

$$\sigma_{q_k}^2 = \epsilon_{q_k}^2 \sigma_k^2 \approx \epsilon_{q_k}^2 2^{-2b_k} \sigma_k^2$$  \hspace{1cm} (3.18)

where $\epsilon_{q_k}^2$ is a quantisation performance factor and $b_k$ is the number of bits per pixel allocated to subband $k$ [Jayant and Noll 1984, p525]. The factor $\epsilon_{q_k}^2$ depends on the probability density function (pdf) of the $k$th subband, the corresponding quantiser characteristics and the number of bits allocated to that subband. It must be noted that the approximation, $\epsilon_{q_k}^2 = \epsilon_{q_k}^2 2^{-2R}$ for all $k$, is quite coarse for low values of $b_k$. Nevertheless it is subsequently used in the following bit allocation and coding gain sections.

For a Gaussian input signal $x(n)$, the subband signals are all Gaussian and it follows that the subband quantiser performance factors are similar. The degree of similarity depends on the rate allocation to the subbands. A PCM quantiser operating on this Gaussian input signal $x(n)$ also has a similar performance factor.

### 3.2.2. Bit Allocation for Minimum Overall Distortion

Since the quantisation error of each subband is dependent on the number of quantisation bits, the reconstruction error is dependent on the way the bits are allocated
to each subband. The bit allocation problem can be stated as that which minimises the average reconstruction error variance for given a fixed average overall (bit) rate. This is written mathematically as,

$$\min(\sigma^2) = \min\left(\sum_{k=0}^{M-1} S_k \sigma_q^2\right) = \min\left(\sum_{k=0}^{M-1} S_k \varepsilon_k^2 2^{-2b_k} \sigma_k^2\right)$$  \hspace{1cm} (3.19)

given,

$$\sum_{k=0}^{M-1} b_k = R$$  \hspace{1cm} (3.20)

where $R$ is the desired overall average rate in bits per pixel. Using a Lagrange multiplier technique it can be shown that the optimum allocation is [Katto and Yashuda 1991],

$$b_k = R + \frac{1}{2} \log_2 \frac{d_k S_k A_k}{\prod_{k=0}^{M-1} (d_k S_k A_k)^{1/2}}$$  \hspace{1cm} (3.21)

Using (3.17) and assuming a unit variance input, gives

$$b_k = R + \frac{1}{2} \log_2 \frac{\sigma_k^2}{\prod_{k=0}^{M-1} \sigma_k^2}$$  \hspace{1cm} (3.22)

noting that (3.17) is dependent on the quantisation error signal assumptions listed previously.

In the derivation of this optimum bit allocation it is assumed that the average rate, $R$, is sufficiently high so that $b_k \geq 0$. In fact, strictly speaking it is assumed that the $b_k$ are sufficiently high so that the quantisation noise from each subband is white. For highly correlated sources, such as still image data, even at moderate rates there may be negative bit allocations. In this case, again using (3.17) and a normalised input variance, the optimum bit allocation is given by,
Substituting the ideal bit allocation into the average reconstruction error variance equation, (3.19), gives,

$$\min(\tilde{\sigma}_e^2) = e_e^2 2^{-2R} \tilde{\sigma}_e^2 \prod_{k=0}^{M-1} (d_k A_k)^{1/2}$$

(3.24)
The coding gain, as defined in equation (3.1), is the ratio of the reconstruction error variances of fullband PCM to subband PCM. In this case it is assumed that $\sigma_{r,SBC}^2 = \min(\bar{\sigma}_i^2)$: that is the subband PCM quantisation bits have been distributed according to the ideal bit allocation. The fullband PCM quantisation error variance is modelled using the same equation as that used for subbands, equation (3.18), noting that the total bit rate is given as $R$ bpp. Therefore,

$$\sigma_{r,PCM}^2 = \epsilon^2 2^{-2R} \sigma_x^2$$

and hence,

$$G_{SBC} = \frac{\sigma_{r,PCM}^2}{\sigma_{r,SBC}^2} = \frac{\epsilon^2 2^{-2R} \sigma_x^2}{\epsilon^2 2^{-2R} \sigma_x^2 \prod_{k=0}^{M-1} (d_k S_k A_k)^{\frac{1}{d_k}}}$$

$$= \frac{1}{\prod_{k=0}^{M-1} (d_k S_k A_k)^{\frac{1}{d_k}}}$$

(3.25)

Using (3.17) and assuming a unit variance input gives,

$$G_{SBC} = \frac{1}{\prod_{k=0}^{M-1} (\bar{\sigma}_k^2)^{\frac{1}{d_k}}} = \frac{1}{\sigma_{WGM}^2}$$

(3.26)

The denominator is the weighted geometric mean (WGM) of the subband variances, where the weights are the inverse of the decimation factors. For orthogonal subband schemes, with a unit variance input, this WGM is upper bound by unity and hence in this case the coding gain is lower bound by unity.

In this derivation it is assumed that the subband quantisers and the corresponding fullband quantiser have the same quantiser performance factor. For Gaussian inputs and a reasonable rate allocation for each subband this is a good approximation. Using scalar quantisers and first order entropy coding it is possible at high rates to approach the rate distortion bound to within about 0.25 bits per pixel (sample) [Jayant and Noll 1984, p155]. This implies that for sufficiently high rates that $\epsilon^2 = 1.4$ for all the subbands and the fullband signal is a valid approximation. For other input pdf's a correction factor is possibly required [Jayant and Noll 1984, p529]. However, since the
coding gain is used primarily to compare various subband analysis/synthesis schemes, these considerations are not particularly significant.

In the case where orthogonal filters are used the only assumption that is made about the quantisation error signals is that they are jointly WSS with zero mean. (It is also assumed that the input is zero mean WSS). Soman and Vaidyanathan (1993, equation (4.6)) arrived at the same equation under the same assumptions for a paraunitary filter bank. It is worthwhile noting that a simple derivation is possible using the orthogonal analysis/synthesis matrices discussed in Chapter 2.

In the derivation of the optimum bit allocation (3.21) and coding gain (3.25) no reference to the filter bank dimension is made. In the rate constraint equation, (3.20), the decimation ratio $d_k$ refers to the ratio of output to input samples for the $k^{th}$ decimator. Therefore the coding gain and bit allocation equations are applicable to two-dimensional filter banks where the $d_k$ are defined as such. Under the same quantisation error signal assumptions listed above, for the two-dimensional case it is shown in Appendix B that,

$$ S_k = \frac{1}{d_k} \sum_{s,t} |g_k(s,t)|^2 $$

where $g_k(n,m)$ is the two-dimensional impulse response of the $k^{th}$ synthesis filter. Usually this filter is normalised to unit energy so that,

$$ \sum_{s,t} |g_k(s,t)|^2 = 1 $$

In this case equations (3.22) and (3.26) are applicable to two-dimensional filter banks. Note that as shown in Chapter 2, a two-dimensional separable filter has unit energy if both the corresponding one-dimensional filters have unit energy.

### 3.2.4. Optimum Block Transform: The Karhunen-Loève Transform (KLT)

The KLT is the optimum orthogonal block transform from several perspectives: namely for a given size it,
**K1** maximises the coding gain.

**K2** minimises the basis restriction error (or maximises the energy compaction).

**K3** decorrelates the input data.

A more detailed description can be found in Clarke (1985, p91). The KLT is defined by any one of these properties.

For a block transform, where the decimation factors are all $M$, the coding gain is,

$$G_{sbc} = \frac{1}{\left( \prod_{k=0}^{M-1} \sigma_k^2 \right)^{1/M}}$$

(3.27)

Note that this equation is the same as the familiar orthogonal transform coding gain metric, $G_{tc}$, assuming a unit variance input signal. (See [Jayant and Noll, 1984 p528]). The unified coding gain, denoted by $G_{sbc}$ generalises the orthogonal transform coding gain, $G_{tc}$, to all perfect reconstruction subband analysis/synthesis systems.

Equation (3.27) gives the maximum gain for a given transform, and assumes the optimum bit allocation. The transform with the maximum coding gain is referred to as the optimum transform. Maximising the coding gain, (3.27), requires minimisation of the denominator, the geometric mean of the transform variances.

Consider an input vector source, $x$, of length $N$ and a $N \times N$ transform operating on this source. Let the rows of the matrix $A$ consist of the transform basis vectors or filters. In the notation of Chapter 2 the subband analysis matrix is then $T_a = A$. The output vector of this transform is thus $y = Ax$. The transform correlation matrix is,

$$R_{yy} = E[yy^T] = E[Ax(Ax)^T] = AR_{xx}A^T$$

(3.28)

where $R_{xx}$ is the input vector source correlation matrix. Assuming a zero mean input source the diagonal entries of the transform correlation matrix are the transform variances. From matrix theory the product of diagonal elements of a positive definite
matrix is lower bounded by the determinant [Bellman 1960, p126]. Hence for an arbitrary positive definite matrix \( B \),

\[
\prod_{k=0}^{M-1} b_{kk} \geq |B| \tag{3.29}
\]

where \( |B| \) is the determinant of \( B \) and \( b_{kk} \) is \((k,k)\)th entry of \( B \). Equality holds if and only if \( B \) is diagonal. Assuming an invertible transform, if the input correlation matrix is positive definite then so is the transform correlation matrix. Applying (3.29) to the transform correlation matrix gives,

\[
\prod_{k=0}^{M-1} \sigma_k^2 \geq |R_{yy}| = |A R_{xx} A^T| = |A| |R_{xx}| |A^T| \tag{3.30}
\]

The inverse transform is given by \( x' = G y \), where \( x' \) is the output vector and \( G \) the inverse transform matrix. For perfect reconstruction \( G = (A)^{-1} \). Without loss of generality it is assumed that the synthesis basis vectors, the columns of \( G \), have been normalised to unit energy (unit Euclidean norm). In this case (3.29) gives,

\[
1 \geq |G^T G| = |G^T| |G| = |G|^2.
\]

This inequality holds since \( G^T G \) is positive definite for invertible \( G \). Also equality holds if and only if \( G^T G \) is diagonal, implying that \( G \) is orthogonal. Further since \( GA = I \),

\[
|A| = \frac{1}{|G|}
\]

Therefore \( |A| \geq 1 \) and,

\[
\prod_{k=0}^{M-1} \sigma_k^2 \geq |A| |R_{xx}| |A^T| \geq |R_{xx}|
\]

Substituting into (3.27) gives,
The KLT is defined by a transform matrix $A$ whose rows consist of the eigenvectors of $R_{xx}$. In this case $A$ is orthogonal, since $R_{xx}$ is symmetric, and

$$|R_{xx}| = |R_{yy}| = \prod_{k=0}^{M-1} \sigma_k^2$$

The latter equality holds because $R_{yy}$ is diagonal. Therefore the KLT coding gain is given by equality in (3.31), the maximum possible coding gain. For other transforms the gain is strictly less than this bound. The KLT totally decorrelates the transform coefficients: that is the transform correlation matrix, $R_{yy}$, has zero entries off the main diagonal. This implies that $\sigma_k^2 = e_k$ where $e_k$ is the $k$th eigenvalue of $R_{xx}$. The KLT is also optimum in that it minimises the basis restriction error, or equivalently maximises the energy compaction for orthogonal transforms.

The optimality of the KLT, in terms of coding gain, amongst orthogonal transforms is well known. Using the unified coding gain it has been shown further that the KLT is optimum over all invertible transforms. This result is based on an assumption of jointly WSS white subband quantisation noise. At high rates this is a reasonable assumption. The subband signals are generally less correlated than the fullband signal, resulting in less correlated intra-subband quantisation noise. Actually it is only required that the synthesised quantisation noise be uncorrelated. The synthesis will typically reduce any inter-subband correlation, especially so if it is near orthogonal.

The fact that the KLT has the maximum coding gain is commensurate with the rate distortion bound for encoding Gaussian variables in $N$-blocks. For small distortions (high rates) the distortion bound for the encoding of such variables in $N$-blocks, assuming that the blocks are independent, at an average rate of $R$ bits per pixel (or sample), is [Jayant and Noll 1984, p646],

$$D = 2^{-2R} \left( \prod_{k=0}^{N-1} e_k \right)^{1/N}$$  \hspace{1cm} (3.32)
where $e_k$ is the $k^{th}$ eigenvalue of the $NxN$ source covariance matrix, $R_s$. A transform with a coding gain larger than the KLT implies, conceptually at least, that the transform reconstruction distortion could be less than this rate distortion bound, which is obviously impossible.

Although the gain is of the KLT is expected to be maximum from this rate distortion analysis, the new result presented here demonstrates that, using the unified coding gain, the gain of the KLT of size $N$ bounds that of all transforms of size $N$. This result is extended in the following section.

### 3.2.5. Coding Gain Bounds for Arbitrary Analysis/Synthesis Schemes

In this chapter subband analysis has been characterised as a single input multiple output discrete linear system, where the sum of the subband sampling rates is equal to that of the input signal. The synthesis on the other hand is characterised as a multiple input single output linear system. It has been assumed that the signals are of infinite duration: in other words transient effects have been ignored. In Chapter 2, on the other hand, the analysis and synthesis were described as a $N$ input $N$ output linear systems using $NxN$ analysis and synthesis matrices. In this latter approach finite signals (or finite blocks of infinite signals) of length $N$ were assumed.

Since finite duration signals, or finite blocks of signals, are encountered in still image compression there must be some way of handling the signal boundaries. The matrix representation of Chapter 2 was used to illustrate that circular convolution is a simple method of ensuring perfect reconstruction when the input signal is of finite duration. An alternative that can be used for some linear phase analysis filters is an even periodic data extension [Smith and Eddins, 1987]. In either case these methods are concisely described using A/S matrices.

In coding gain derivation it was assumed that the input signal is of infinite duration. For a finite signal the coding gain must take into account the way in which the boundaries are handled. However, ignoring this problem for the moment, the coding gain for a general PR subband analysis technique is the equal to that of a PR block transform, with transform matrix equal to the subband analysis matrix. For the transform, there is one "subband" or coefficient for each input sample, where the number of input samples is the length of the input vector. However each "subband"
belongs to one of \( M \) groups, where there are \( M \) subbands in the original subband A/S system. Within each group the subbands have filters that are equivalent to within a delay. This was described in Chapter 2, where the delayed or (circularly) shifted subband filter coefficients form the rows of the analysis matrix. Since these "subbands" have equivalent filters to within a delay the variances are the same (a subband variance is dependent only on the magnitude response of the subband filter). Now the coding gain for an \( N \times N \) block transform is given by,

\[
G_{TC} = \frac{1}{\left( \prod_{n=0}^{N-1} \sigma_n^2 \right)^{\frac{1}{N}}}
\]

where it is assumed that the synthesis transform basis vectors have unit norm. For a subband analysis matrix system there are only \( M \) distinct variances. Hence,

\[
G_{TC} = \frac{1}{\prod_{k=0}^{M-1} (\sigma_k^2)^{N_k / N}} = \frac{1}{\prod_{k=0}^{M-1} (\sigma_k^2)^{1/d_k}}
\]

where the \( k^{th} \) variance is of multiplicity \( N_k \). Note that the decimation factor of the \( k^{th} \) subband in the original subband A/S system is by definition \( d_k = N / N_k \). Comparing with equation (3.26) one sees that indeed the transform coding gain is the same as that of the original subband A/S system.

The signal boundaries are automatically considered using the block transform approach to calculate the coding gain. As a consequence the gain figure attained using the block transform approach will generally be lower than the standard subband coding gain. However as the block size increases, the boundary effects become less significant and this difference will decrease, tending to zero as the block size tends to infinity.

As was shown in Section 3.2.4. the coding gain of the KLT is an upper bound for all block transforms. Therefore the coding gain of any finite subband scheme is bounded by the KLT of equivalent dimension to that of the subband analysis matrix. As the KLT block size tends to infinity the coding gain tends towards the inverse of the spectral flatness measure [Jayant and Noll 1984, p543]. This limit obviously bounds the gain of any finite block size KLT. It follows that the inverse of the spectral flatness measure is an upper bound for any perfect reconstruction subband A/S scheme. As
before this result is expected from the rate-distortion bound for the encoding of Gaussian variables in $N$-blocks.

Jayant and Noll (1984, p253 and p543) noted that the inverse of the spectral flatness measure is an upper bound for the gain of linear predictive DPCM schemes and orthogonal transforms. Pearlman (in Woods 1991, p32) and Woods and O'Neil (1986) showed that the gain of a subband A/S system, using ideal filters and optimum DPCM on the subbands, over optimum DPCM on the fullband signal is unity. In other words theoretically there is no advantage in using subband analysis prior to DPCM. (It is worth noting, that it was pointed out that in practice there is likely to be an advantage). Since the distortion introduced by a PCM is lower bound by that of optimum DPCM this also implies that the coding gain of an ideal filter bank is bounded by the inverse of the spectral flatness measure. Using the unified coding gain, it has been shown here that the coding gain of all perfect reconstruction schemes is bounded by the inverse of the spectral flatness measure.

The coding gain of the KLT bounds that of all PR subband analysis schemes of a given input vector dimension. However this does not mean that the coding gain of the KLT of dimension $N$ bounds that of a subband A/S scheme (ignoring boundary effects) whose filters are all less than or equal to $N$-taps in length. For example it has been observed that an 8-band LOT, where the A/S filters are 16-taps in length, has a higher coding gain for an AR(1) source of correlation around $p=0.7$, than the associated KLT of dimension 16x16.

The gain associated with the ideal filter bank bounds that of all orthogonal two-band filter banks [De Queiroz and Malvar, 1992]. However, it has also been observed that some two-band biorthogonal A/S systems have a higher gain than the ideal two-band filter bank for a highly correlated AR(1) source.

3.3. RATE CONSTRAINED CODING GAIN

In the previous section, the unified coding gain was derived assuming that the desired overall bit rate was sufficiently high so that an ideal bit allocation (3.22) is possible. In this section the coding gain is examined under a rate constraint. Under the assumptions made the following analysis is applicable only to orthogonal analysis/synthesis schemes. Nevertheless, the results indicate general trends that are also applicable to
biorthogonal schemes. In Section 3.3.1 it is shown that the asymptote of the rate constrained coding gain decreases as the rate decreases. The rate constrained coding gain for highly correlated sources is examined in Section 3.3.2. This work indicates that quite high levels of subband analysis may be required for near a optimum coding gain at low rates. This characteristic is an important consideration in the design of subband schemes with highly correlated input signals, such as HDTV resolution images.

The ideal coding gain assumes an ideal bit allocation. A more general form assumes a practical optimum bit allocation as given by (3.23). As in the ideal case it is assumed that equation (3.16) is valid: that is for orthogonal subbands schemes the quantisation noise is jointly WSS. Also it is assumed, without loss of generality, that the synthesis filters are normalised to unit energy. Using equations (3.19) and (3.16) the rate constrained coding gain becomes,

$$G_{sbc} = \frac{\sigma^2_{\epsilon,PCM}}{\sigma^2_{r,sbc}} = \frac{\epsilon^2_2 2^{-2R} \sigma^2_s}{\sum_{k=0}^{N_p-1} \frac{1}{d_k} \epsilon^2_2 2^{-2k} \sigma^2_k + \sum_{k=N_p}^{M-1} \frac{1}{d_k} \sigma^2_k}$$

(3.33)

where it is assumed that the subbands are indexed so that the first $N_p$ are allocated non-zero bits while the remainder are allocated zero bits. This gain equation is not valid for biorthogonal analysis, since some of the assumptions made in the derivation are not valid when the rate assigned to any subband is low or zero (see Appendix B).

For sufficiently high rates the coding gain is accurate for ideal analysis/synthesis filter banks with a Gaussian input signal [Pearlman, 1991]. However, in other cases the approximations made in the derivation of the (rate constrained) coding gain must be considered. For example PCM quantisers require integer bit allocations, or at least an integer number of quantisation levels. Further, the approximation that the parameter $\epsilon^2_2$ is independent of both the rate and subband is quite coarse, especially at low rates. Finally at rates below 1 bpp it is impossible to consider PCM quantisation of the fullband signal.

In this chapter the number of bits allocated to each subband is assumed to be a continuous variable (greater than or equal to zero). While in practice subband bit allocations may not be exactly as prescribed by the ideal allocation, the discrepancy is usually small. In other cases the continuous allocation can be considered as an
extrapolated or average result. It is illustrated later in this section that the substitution of quantisation error models that are more accurate for a given source and rate into the gain equation may give different absolute results but the relative performance of various subband schemes is largely unchanged. It follows that the approximations referred above are largely irrelevant when using the rate constrained coding gain to compare different subband schemes. For a fixed rate below 1 bpp the coding gain can be considered as the MSE of a subband scheme as compared to the fixed term $e^2 2^{-2R} \sigma_i^2$. It is then possible to compare different transforms and different analysis levels against this benchmark.

3.3.1. Asymptotic Performance of the Rate Constrained Coding Gain

As discussed in Section 3.2.5., the ideal coding gain for any subband scheme is bounded above by the inverse of the spectral flatness measure. The coding gain of the KLT is asymptotic towards this bound as the block size tends to infinity. Yemeni and Pearl (1979) showed that the coding gain of the DCT, DFT and DST is asymptotic towards this maximum bound. Other common M-band transforms such as the LOT, ELT and ideal filter bank also appear to asymptote toward this upper bound. This is expected since at any analysis level $M$, the coding gain of the LOT and ELT is greater than that of the DCT. De Queiroz and Malvar (1992) demonstrated that the ideal octave-band filter bank is asymptotic towards a lower bound. This sub-optimum asymptotic performance is expected since only the low frequency subbands are repeatedly analysed. Nevertheless, for a highly correlated AR(1) source the octave-band bound is quite close to the absolute upper bound.

The rate constrained coding gain at a rate of zero is unity. At zero rate it is only possible to quantise the subbands with zeros, giving an average reconstruction error variance equal to the signal variance. The equivalent error results in the case of "zero bit" PCM of the fullband signal. Hence the rate constrained coding gain is commensurate with the gain relative to a PCM system at zero rate. For non-zero mean signals the fullband and subband signals are quantised by their mean values giving, as above, a unit gain. This unit gain is independent of the level of analysis employed.

Figure 3.5 shows some gain versus $M$ curves for the DCT assuming an ideal bit allocation and using a practical allocation at rates of 1.0, 0.5, 0.2, 0.1 and 0.05 bpp. These curves were generated using the rate constrained coding gain equation, (3.33),
and the optimum practical bit allocation equation, (3.23). It is evident that the rate constrained coding gain decreases as the rate decreases. Further the curves are asymptotic toward decreasing levels as $R$ decreases. In the limit as $R$ tends to zero the gain is one (zero dB) for all $M$.

![Diagram of rate constrained coding gain versus $M$: DCT, AR(1) source $\rho = 0.95$](image)

Figure 3.5. Rate constrained coding gain versus $M$: DCT, AR(1) source $\rho = 0.95$

The optimum level of analysis is the minimum level of analysis that has a gain figure which is close to the asymptotic gain. Lower levels of analysis are desirable since they require less computation and offer advantages in the encoding of nonstationary sources (for example see Chapter 6 where it is demonstrated that high analysis levels may be inferior to medium analysis levels). This definition of optimum analysis level is obviously dependent on how one interprets "close": as a rough measure a figure of 90% is used.

The asymptotic nature of the rate constrained coding gain is expected. Obviously the rate constrained gain is upper bound by that of the ideal gain. While the rate $R$ is sufficiently large so that no $b_k$ are zero, then the rate constrained gain is that same as the ideal gain, and hence has the same asymptote. When $R$ is no longer sufficiently large, the coding gain decreases as the bit allocation deviates from the ideal optimum bit allocation. Since the practical bit allocation (3.23) is continuous in $R$, the rate constrained coding gain is continuous in $R$. It follows the asymptotic gain decreases in a continuous fashion towards unity as $R$ decreases.
SUBBAND CODING GAIN

Consider the octave-band (dyadic) analysis of an AR(1) source with positive correlation where the subbands are indexed in terms of decreasing frequency and the lowest subband is labelled separately. For example subband 0 and 1 are the highest and second highest frequency subband respectively. The reason for such nomenclature is that each subband label is independent of the tree-depth, S. In Appendix B it is shown that at the maximum rate where the first (highest) \( N+1 \) subbands are allocated zero bits, the coding gain is given by,

\[
G_{SBC} = \frac{\prod_{k=0}^{N} (\sigma_k^2)^{a_k}}{\sum_{k=0}^{N} \alpha_k \sigma_k^2} G_{SBC,\text{ideal}}
\]  

(3.34)

where,

\[
\sum_{k=0}^{N} \alpha_k = 1, \quad \alpha_k = \begin{cases} 
\frac{1}{d_k} & k = 0, 1, \ldots, N-1 \\
1 - \sum_{k=0}^{N-1} \frac{1}{d_k} & k = N
\end{cases}
\]

and \( G_{SBC,\text{ideal}} \) is the ideal or high rate coding gain. This rate constrained gain is given as a fraction of the ideal gain. This fraction is upper bound by unity since it is the ratio of a weighted geometric mean to arithmetic mean. As the rate decreases, more subbands are allocated zero bits, \( N \) increases and this fraction decreases. In Appendix B (Section B.3.1) it is illustrated that for a highly correlated source this fraction decreases quite rapidly for increasing \( N \). The asymptotic rate constrained gain, simply the limit of the gain as the tree-depth \( S \) tends to infinity, decreases according to this fraction. As the rate tends to zero, the asymptotic gain tends to one since the fraction tends to the inverse of the ideal coding gain.

An interesting characteristic of the rate constrained coding gain, evident in Figure 3.5, is that as the rate decreases the knee of the gain curves occur at increasing \( M \). The minimum level of analysis required for near optimum performance at high rates may be quite sub-optimum at low rates. This characteristic is the topic of the following section.
3.3.2. Coding Gain for Highly Correlated Sources at Low Rates

For low rates and highly correlated input sources it is possible that the DC subband variance dominates to an extent where it is the main source of quantisation error. Under these conditions the following approximation can be made,

\[ \varepsilon^2 2^{-2d_0R} \frac{\sigma^2_0}{d_0} \gg \sum_{k=1}^{M-1} \frac{\sigma^2_k}{d_k} \]  

(3.35)

where the subscript 0 refers to the DC subband. This approximation implies that the source energy is predominantly contained in the DC subband, and that a large percentage of the quantisation error derives from this subband. In this case the optimum DC subband bit allocation is \( b_0 = d_0R \), and the rate constrained coding gain is,

\[ G_{sbc} = \frac{\varepsilon^2 2^{-2d_0R} \sigma^2_0}{\varepsilon^2 2^{-2d_0R} \sigma^2_0 + \sum_{k=1}^{M-1} \sigma^2_k} \]  

(3.36)

At such low rates this equation is valid only for orthogonal filter banks. Also, as stated previously, for \( R < 1 \) the coding gain is relative to \( \varepsilon^2 2^{-2R} \sigma^2_x \) rather than relative to a PCM system. Substituting the relationship,

\[ \sigma^2_x = \sum_{k=0}^{M-1} \sigma^2_k = \frac{\sigma^2_0}{d_0} + \sum_{k=1}^{M-1} \frac{\sigma^2_k}{d_k} \]

applicable to orthogonal filter banks, into (3.36) gives,

\[ G_{sbc} = \frac{\varepsilon^2 2^{-2d_0R} \sigma^2_0}{\varepsilon^2 2^{-2d_0R} \sigma^2_0 + \sum_{k=1}^{M-1} \sigma^2_k} + \frac{\varepsilon^2 2^{-2R} \sum_{k=1}^{M-1} \sigma^2_k}{\varepsilon^2 2^{-2d_0R} \sigma^2_0 + \sum_{k=1}^{M-1} \sigma^2_k} \]

Now, given approximation (3.35) the second fraction is negligible compared to the first, as is the latter term in the denominator of either fraction, which gives,
This approximate coding gain, \( \hat{G}_{\text{SBC}} \), is derived using approximation (3.35). Hence for a fixed rate \( R \), while approximation (3.35) is valid, the coding gain is exponential in \( d_0 \). Note that usually \( d_0 \), the decimation factor of the DC subband, is inversely proportional to the DC subband bandwidth. It is interesting to note that this approximate coding gain is dependent only on \( d_0 \) and \( R \). Also, it can be shown that for \( \varepsilon_2^2 = 1 \), or sufficiently high \( d_0 \), this gain upper bounds the actual gain. Note that for the purposes of calculating the rate constrained coding gain it has been assumed that \( \varepsilon_2^2 = 1 \). In terms of comparing results, using different values of \( \varepsilon_2^2 \) has little effect.

The approximation, \( \hat{G}_{\text{SBC}} \), can be formulated more precisely in the case of an ideal octave-band or DWT analysis. Consider such an octave-band analysis of a sequence of Gauss-Markov sources, with correlation \( \rho \) tending to unity, at a high rate. In the limit as \( \rho \) tends to unity, approximation (3.35) is valid for all \( d_0 \), and hence the coding gain is given by (3.37). As a consequence, for finite \( d_0 \) the gain is finite, independent of \( \rho \), and comparatively small for small values of \( d_0 \). In comparison the ideal coding gain tends to infinity for all \( d_0 > 2 \) as \( \rho \) tends to unity.

For these Gauss-Markov sources approximation (3.35) becomes invalid for some value of \( d_0 \), and the exponential coding gain rise is curtailed. After this point the gain asymptotes with increasing \( d_0 \) to some upper bound as demonstrated in the previous subsection. As the source correlation increases this breakdown point occurs at increasing \( d_0 \). It follows that the optimum level of analysis increases with increasing correlation and/or decreasing rate for highly correlated sources at low rates. This is evident to some extent in Figure 3.5, where as \( R \) decreases the gain curves are pushed out or compressed along the \( M \) axis.

This compression is more pronounced as the correlation increases. Figure 3.6 illustrates the gain curves for the DCT, lapped orthogonal transform (LOT) and a DWT using Daubechies 12-tap filters (Daubechies 1988). An AR(1) source of correlation \( \rho = 0.99 \) at rate \( R = 0.05 \) bpp has been used. Also illustrated in the figure is the curve of (3.37) at this rate. It is evident that each curve is almost identical to that predicted by (3.37) for block sizes up to \( M = 16 \) or tree depth \( S = 4 \). Note that approximation (3.37) applies to all
orthogonal analysis/synthesis schemes. Figure 3.6 illustrates that the optimum level of analysis at low rates occurs for quite large analysis levels \((M,S)\). It follows that high levels of analysis are desired in the compression of highly correlated sources at low rates.

![Graph of Figure 3.6](image)

Figure 3.6. Rate constrained coding gain versus \(\log_2 M\) or \(S\): AR(1) source, \(\rho=0.99\), \(R=0.05\) bpp. (——) Graph of (3.37); (•••) LOT; (••••) Wavelet (DWT)

The same characteristics are observed using more accurate quantisation noise models. Segall (1976, example 2.2) described a quantiser noise model for Lloyd-Max quantisers for Gaussian sources operating at an arbitrary bit rate. Based on this model the optimum bit allocation between subbands was also given. The model that has been used previously in this chapter is referred to as the old model while the Segall model is referred to as the Lloyd-Max model. Figure 3.7 illustrates the gain predicted using the old and Lloyd-Max models for a DCT analysis and an ideal octave-band analysis of an AR(1) source assuming a bit rate of \(R=0.05\) bpp. Note that the old model results are the same as those shown in Figure 3.6.
Figure 3.7. Rate constrained coding gain versus $\log_2 M$ or $S$: AR(1) source $\rho=0.99$, $R=0.05$ bpp: (----) wavelet scheme using old model; (-----) wavelet scheme using Lloyd-Max model; (•••) DCT using old model; (−−−) DCT using Lloyd-Max model.

The Lloyd-Max model predicts a gain slightly lower than that of the old model. The reason for this discrepancy is that in the Lloyd-Max model the effective quantiser performance factor ($e^2$) increases at low rates, while in the old model it is fixed at all rates. However it is evident in Figure 3.7 that the same relative performance between the DCT and wavelet schemes is attained for each model. Also, in terms of comparing different analysis levels the results are largely the same. The coding gain is initially exponential and increases significantly until quite high analysis levels. A similar approximation to $\hat{G}_{\text{sbc}}$, tracking this initial exponential rise, can be derived using the Lloyd-Max model.

The ideal gain, which assumes a high rate, using the Lloyd-Max model is the same as the old model since the two noise models converge for an increasing rate. Figure 3.8 illustrates the ideal gain and rate constrained gain using the Lloyd-Max quantiser model for the same subband schemes and sources as Figure 3.7.
Figure 3.8 illustrates that the optimum level of analysis is much higher when operating at low rates as compared to high rates. At high rates the wavelet scheme is near optimum at a level $S=3$ or $S=4$, while at a rate of $R=0.05$ bpp a level of at least $S=6$ is required.

The conclusion drawn from these results is that any subband coding scheme, with highly correlated input data, requires higher levels of analysis for optimum performance at low rates as compared to the analysis level required at high rates. This conclusion is corroborated in Chapter 6 using a practical image compression algorithm. The performance of the wavelet scheme suggests that it is necessary to increase the analysis levels of low frequencies only, rather than across the whole spectrum. It is worthwhile noting that as image sampling densities increase toward HDTV resolutions and higher, the first order correlation of typical images will increase. Hence subband schemes used in high compression applications, involving these images, will require higher analysis levels than those used currently.

The results also suggest that in the comparison of different subband schemes over a range of rates, it is important to consider the relevant levels of analysis. For example the results of [Ebrahimi and Kunt 1992] suggest that the octave-band or DWT analysis is superior to the DCT at low rates for image compression. However, since the
comparison is between an 8x8 DCT and S=4 wavelet analysis, the difference may be attributed to the lower analysis level in the case of the DCT. A 16x16 DCT would be required for a fairer comparison at low rates. In the coders presented in Chapter 6, an 8x8 DCT coder is inferior to an S=4 wavelet scheme at low rates, while the 16x16 DCT coder performs in a similar manner. However, the wavelet scheme has other advantages over the DCT, which are discussed in Chapter 6.

3.4. TRANSFORM FREQUENCY DOMAIN CHARACTERISTICS

Clarke (1981) showed that the KLT of an AR(1) source is asymptotically equivalent to the DCT as \( p \) tends to unity. Similarly the KLT is asymptotically equivalent to the DST as \( p \) tends to zero. It is well know that the DST is inferior to the DCT for the encoding of highly correlated sources, such as still images [Clarke 1983a]. Hence it is of interest to compare the performance of various transforms and filter banks as the correlation coefficient varies for an AR(1) source. This is the topic of the subsection 3.4.1., where a theorem relating transform basis vector pairwise symmetry to a symmetric coding gain is given.

The coding gain is dependent on the transform coefficient variances, which in turn depend only on the magnitude response of the transform basis vectors and the PSD of the given source. Therefore, it is also of interest to examine the magnitude response of various transform basis vectors in light of their relative performance for a given source. This topic is considered in subsection 3.4.2., and provides the motivation for the design of optimum octave-band analysis filters, which is the subject of Chapter 4.

3.4.1. Pairwise Symmetric Transforms and a Symmetric Coding Gain

It has been observed that the coding gain of some transforms for an AR(1) source is symmetric about \( p=0 \): that is the coding gain is independent of the sign of \( p \).

Akansu and Haddad (1990) showed that transform variances for a block transform may be written in vector form as,

\[
\sigma^2 = V^T r_{xx}, \quad \nu_{pk} = (2 - \delta_p) \sum_n h_k(n)h_k(n+p)
\]
where $r_{xx}$ is a vector of the input source autocorrelation sequence values, $\sigma^2$ is a vector of the transform variances, and $v_{pk}$ is element $(p,k)$ of matrix $V$. The transform basis vectors are denoted as filters, $h_k(n)$, commensurate with the notation introduced in Chapter 2. The derivation is given in Appendix B and extended to include general $M$-band subband analysis methods. Akansu and Haddad showed that if $V$ has pairwise symmetric columns, then the transform coding gain is independent of the sign of $p$ for an AR(1) source. This follows from the fact that the variances are unchanged, other than a permutation in the vector $\sigma^2$, for such $V$ matrices given a change in the sign of $p$. Any objective measure that is a function of the variances, where the ordering is immaterial, is independent of the sign of $p$. This characteristic is referred to as a symmetric coding gain. The symmetric gain of the DST, modified Hermite transform, and discrete Hadamard transform (DHT) and the asymmetric performance of the DCT was illustrated by Akansu and Haddad for an 8x8 block size transforms.

Following Vaidyanathan and Nguyen et al (1989) pairwise symmetric (PWS) analysis filters are characterised as,

$$h_k(n) = (-1)^k h_{M-k}(n) \quad \text{or} \quad h_k(n) = (-1)^k h_{M-k}(N-1-n), \quad k = 0, 1, \ldots, \frac{M}{2} - 1$$

where $M$ is even. The actual labelling of the subbands is not important, just that there are pairs of filters that are related through the lowpass to highpass transformation. In the case of odd $M$, for a filter set to be PWS there must be one filter that is invariant under a lowpass to highpass transformation. This means every second sample of this filter must be zero.

A transform with basis filters obeying either of these equations is termed a PWS transform. If necessary, in order to differentiate between the equations, time-reversed PWS is used to refer to the latter equation. Note that the word transform is used to refer to any $M$-band or uniform subband analysis method. The following lemma is applicable to all PWS transforms,

**Lemma 3.1**: The coding gain associated with a PWS transform is symmetric for an AR(1) source.

**Proof**: If a transform is PWS then,
\[ v_{p,k} = \sum_{n} \left( 2 - \delta_{p} \right) h_{M-1-k}(n) h_{M-1-k}(n+p) \]
\[ = \sum_{n} \left( 2 - \delta_{p} \right) (-1)^{n} h_{k}(n) (-1)^{n+p} h_{k}(n+p) \]
\[ = (-1)^{p} \sum_{n} \left( 2 - \delta_{p} \right) h_{k}(n) h_{k}(n+p) \]
\[ = (-1)^{p} v_{p,k} \]

which shows that \( V \) has PWS columns and thus the transform has a symmetric coding gain.

\[ \nabla \nabla \nabla \]

In Appendix B the time-reversed PWS and complex transform cases are derived. If the transform matrix is also symmetric and the basis vectors are either symmetric or skew symmetric then one has,

**Corollary 3.1:** An orthogonal non-overlapping transform, where the block size \( M \) is even, with symmetric or skew-symmetric basis vectors and a symmetric transform matrix is a PWS transform, and hence has a symmetric coding gain.

**Proof:** Consider a matrix \( A \) whose rows consist of the transform basis vectors. Since there are \( M \) basis vectors of length \( M \) for a block transform, \( A \) is of dimension \( M \times M \). \( A \) is orthogonal since the basis vectors are orthogonal. It then follows, given that the basis vectors are symmetric, that there are \( M/2 \) symmetric basis vectors and \( M/2 \) skew symmetric basis vectors [Cantoni and Butler, 1976]. Ordering the basis vectors in the rows of \( A \) as alternately symmetric and skew-symmetric, gives a matrix with PWS columns. It follows that \( A^T \) defines a PWS transform since the basis filters, consisting of the rows of \( A^T \) are PWS. If \( A \) is also symmetric then \( A = A^T \) and hence the transform defined by \( A \) is PWS. Note that any reordering of the rows of \( A \) is irrelevant as far as the coding gain is concerned.

\[ \nabla \nabla \nabla \]

Some typical examples of orthogonal non-overlapping transforms that are symmetric and have (skew) symmetric basis vectors are the discrete sine transform (DST), and the Walsh-Hadamard transform (WHT). Jain (1976a) described a whole family of sinusoidal transforms that are asymptotically equivalent to the KLT for various source models and values of correlation. The DST and DCT referred to above are members of
The discrete Fourier transform (DFT) has PWS basis vectors and hence has a symmetric coding gain. In the case of overlapping transforms, the cosine modulated transforms are PWS and thus have a symmetrical coding gain. Similarly perfect reconstruction filter banks designed by Vaidyanathan et al (1989) enforce a pairwise symmetry, and therefore also have a symmetrical coding gain.

The discrete cosine transform (DCT) has (skew) symmetric basis vectors and thus the transform described by the transpose of the DCT is PWS. However, the DCT transform matrix is not symmetric and has a definite asymmetric coding gain with respect to the sign of $\rho$ for an AR(1) source. This is illustrated in Figure 3.9 where the coding gain for various transforms using an AR(1) source is plotted against the correlation $\rho$.

![Figure 3.9. Coding gain versus correlation coefficient: AR(1) source. (——) DCT; (— —) WHT; (• • •) ELT; (— — —) ideal filter bank. $M = 8$](image-url)
This figure shows a symmetrical coding gain for the ELT (a cosine modulated transform), and the WHT. The DCT performance is impressive for high positive correlation but in contrast is remarkably inferior for negative correlation. Also indicated in this figure is the gain associated with an ideal filter bank: that is one that employs brick wall filters. It is generally believed, although to the best of the authors knowledge not proven for $M>2$, that the gain for an orthogonal $M$-band analysis/synthesis system using a non-decreasing or non-increasing source is bounded by that of the ideal filter bank (for example see [Malvar 1992, p212]). The response of this filter bank is obviously symmetric, and in fact can be considered to have PWS (infinite length) basis vectors.

It is demonstrated in subsection 3.4.2 that the poor performance of the DCT for negative correlated sources is a consequence of its basis functions being optimum for highly positively correlated sources. Ogunbona et al (1993) used the WHT for the encoding of subbands in a still image compression scheme. Not only is the WHT more computationally efficient than the DCT, it has a superior coding gain for negative correlated sources, as shown in Figure 3.9. Since the WHT is orthogonal and has symmetric basis vectors it has a symmetric coding gain according to corollary 3.1. Also, partly as a consequence of the symmetric gain, it has a reasonable coding performance over a wide range of $\rho$. For a subband analysis of an image some (non DC) subbands typically have negative correlation, while others exhibit positive correlation [Tanabe and Favardin 1992], [Ogunbona et al 1993]. By using the WHT it is possible to attain a reasonable coding performance for all these subbands. It must be noted that the biggest advantage is in computational cost rather than a significant coding gain as compared to the DCT [Ogunbona et al 1993].

It is not surprising that PWS transforms have a symmetric coding gain. Changing the sign of $\rho$ for an AR(1) source is the equivalent to a lowpass to highpass source transformation. Hence the source PSD is reflected about $\pi/2$ radians (or 0.25 cycles/sample). If the basis vectors of any transform are also lowpass to highpass transformed then the variances corresponding to each basis vector will be the same as before (assuming the source has also been transformed). If follows that the coding gain will also be the same. A lowpass to highpass transformation of a PWS transform simply maps the transform to itself, since each filter in a PWS pair is mapped to its partner.
3.4.2. Frequency Domain Characterisation of Transforms

The coding gain of the DCT is close to that of the KLT for highly correlated sources. Figure 3.10 illustrates the magnitude response of the DC (first) and highest frequency DCT basis vectors for a transform of size 8. The magnitude response curves were generated using a 1024 point zero-padded FFT of the DCT basis vectors. The sidelobes of the DC magnitude response exhibit a shallow attenuation across the whole (stopband) spectrum. In contrast, the AC basis vector magnitude response exhibits sidelobes that have increased attenuation at lower frequencies. This is characteristic of all the high frequency basis vectors.

![Figure 3.10. Magnitude response of DC and highest frequency M=8 DCT basis vectors](image)

The variance of the $k^{th}$ transform component output is given by equation (3.7) as,

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

where $H_k(e^{j\omega})$ is the frequency response of the $k^{th}$ basis vector.

The PSD of a highly correlated AR(1) source, illustrated in Figure 3.3 Section 3.1.5, is characterised by a dominance of low frequencies. The high sidelobe attenuation at low frequencies for the DCT AC basis vectors means that for highly correlated sources
leakage from the dominant low frequencies is minimised. In other words, where the
PSD $S_x(e^{j\omega})$ is dominant the filter magnitude response $|H_k(e^{j\omega})|$ has a large
attenuating effect, thus minimising the high frequency subband variances. The DCT is
approximating, as much as possible for the short filter lengths, the ideal (brick-wall)
filter variances given a highly correlated source. It is shown in Chapter 4 that this type
of asymmetric sidelobe attenuation is a desirable characteristic for filters employed in a
system where the input is highly correlated. In fact this characteristic of the DCT was
the motivation for the filters investigated in Chapter 4.

The poor DC sidelobe attenuation at high frequencies is of little significance. In fact it
is conjectured that this poor performance is traded for a large low frequency sidelobe
attenuation. If the input source had a negative correlation close to -1, the DCT would
perform very poorly. In this case the low DC sidelobe attenuation would allow
significant leakage into the DC (and low frequency) components from the dominant
high frequencies. However, transforming the DCT basis vectors using the low-pass to
highpass transformation, would give a transform that would be asymptotically
equivalent to the KLT as $\rho$ tends to minus one.

Jain (1976) showed that the KLT of an AR(1) source is asymptotically equivalent to
the DST as $\rho$ tends to zero. Figure 3.11 illustrates the frequency response of the DC
(first) and highest frequency DST basis vectors for a transform of size 8.

![Figure 3.11. Magnitude Response of DC and highest frequency DST basis vectors](image-url)
In comparison to the DCT, the DST sidelobes exhibit a more even attenuation across the spectrum. This even attenuation minimises leakage for a source that has a relatively flat PSD. Further, due to the pairwise symmetry of the DST, mirroring the magnitude response of one basis filter about \(\pi/2\) radians (or 0.25 cycles/sample) gives the magnitude response of another basis filter. The DST performs poorly for image coding for two reasons: the first is that it is not zero-constrained, and the second is the relatively even nature of the stopband attenuation, which does not sufficiently suppress the dominant low frequencies.

Figure 3.9 illustrated that the gain of the ELT is very near the upper bound for an orthogonal filter bank, the gain of the ideal filter bank, for all correlation (\(\rho\)). The longer basis vectors employed in the ELT, offer sufficient attenuation of all stopband frequencies over the whole range of correlation. The ELT provides a good approximation to the ideal filter bank for all \(\rho\). This is in contrast to the non-overlapping transforms, such as the DCT, where optimum performance for high positive correlation means a sub-optimum performance for high negative correlation. The performance of the LOT, employing basis vectors of length greater than the DCT but shorter than the ELT, is somewhere in between. The disadvantage of transforms with long basis vectors is increased computational expense and less time resolution or localisation, which, as illustrated in Chapter 6, can be detrimental to coding performance. If some knowledge of a source is available, then transforms with a high gain and relatively short basis vectors can be designed and used.

3.5. CONCLUSION

In this chapter a subband coding gain metric has been defined. This metric provides a measure of the performance of a subband analysis/synthesis (A/S) scheme for a given input source model. It is a useful tool for comparing different A/S schemes. Some image correlation and power spectral density models were discussed in the Introduction.

Following the subband analysis of a signal, the next task is to quantise and encode the subband information efficiently. Using separate PCM quantisers for each subband requires a distribution of quantisation bits among subbands. Usually the problem is posed as the bit allocation that minimises the overall system mean square error (MSE)
for a given average bit rate. In Section 3.2 such a method was derived under certain assumptions that are valid at high rates. Using this bit allocation, the unified coding gain of Katto and Yashuda (1991) was derived. In the case of orthogonal A/S it was shown that some of the assumptions may be relaxed and that the resulting gain metric is equivalent to that given by Soman and Vaidyanathan (1993) for paraunitary or orthogonal subband coders. It was shown that using unit energy synthesis filters a simple bit allocation scheme may be used which is suitable for any subband analysis/synthesis system.

In Section 3.2.4 the KLT was shown to be the optimum transform, in terms of unified coding gain, extending a previously known result that the KLT is the optimum orthogonal block transform. It was demonstrated in Section 3.4.5 that the coding gain of an arbitrary subband A/S system can be evaluated using an equivalent block transform. It follows that the KLT coding gain bound applies to any PR subband system, and that the inverse of the spectral flatness measure, which bounds the gain of the KLT, is an upper bound for all PR subband systems.

The coding gain derived in the former portion of this chapter and previous work assumed a high rate. In Section 3.3 a rate constrained coding gain was considered for orthogonal subband analysis methods. The rate constrained coding gain was demonstrated to be asymptotic to a lower level than the unconstrained case. Also for highly correlated sources operating at very low rates it was shown that the rate constrained coding gain is almost negligible until quite high analysis levels. The minimum level of analysis required for a near optimum coding gain increases as the rate decreases for such sources. This characteristic has important implications in the design of subband coders for HDTV and other high resolution image systems.

A theorem relating transforms with pairwise symmetric basis filters and a coding gain that is independent of the sign of $\rho$ for an AR(1) source was given in Section 3.4. This means that the coding gain of such transforms is symmetric about $\rho=0$. The magnitude response of the DCT and other transforms was investigated in relation to this coding performance. The asymmetric coding gain of the DCT is related to the fact that it performs so well for highly correlated sources. For negatively correlated sources the DCT performs poorly. It was observed that the improved subband resolution of overlapping transforms means that a good coding performance for all values of correlation $\rho$ can be attained. The disadvantage of the overlapping transforms is an
increased computational expense and less time resolution or localisation. From another perspective, transforms with short basis vectors can achieve a high coding gain, if optimised for a particular source. This observation leads to the development of optimum two-band subband filters, which is the topic of Chapter 4.
CHAPTER 4:

MAXIMUM CODING GAIN TWO-BAND FILTERS

4.1. INTRODUCTION

The purpose of subband analysis is to separate various frequency bands for independent processing. It follows that, when no knowledge of the source is available, subband filters should have a high stopband attenuation, small transition bandwidth and a flat passband. Early subband filter designers aimed for filters with such characteristics under a constraint of perfect reconstruction (PR) or near PR.

As discussed in Chapter 2, for a two-band filter bank using QMF's it is possible to cancel aliasing and phase distortions introduced by the filter bank. However in this case there is some associated magnitude distortion that cannot be eliminated. Johnston (1980) tabulated different sets of QMF's, designed to minimise the filter bank magnitude distortion and filter stopband attenuation, for a given transition bandwidth. These filters were subsequently used by many authors for subband image coding. Jain and Crochiere (1984) designed QMF's using the same criteria in the time domain.

With the advent of PR filter banks, such as the two-band CQF solution of Smith and Barnwell (1986), filter bank magnitude distortion was automatically cancelled. This extra freedom allowed Smith and Barnwell to design CQF's with superior characteristics to their QMF counterparts. They used an optimal Chebyshev equiripple design, giving filters with a relatively high stopband attenuation and low transition bandwidth. Vaidyanathan (1987) offered the first M-band PR filter bank solution. A design procedure was given that maximised the filters average stopband attenuation for a given transition bandwidth. In subsequent papers by Vaidyanathan and co-authors several improved design procedures using the same criteria were given.

The nature of a typical input source to a filter bank was not considered by subband filter designers until recently. Kronander (1989b) noted that good coding efficiency or a high coding gain using a typical image model is a desirable filter bank characteristic.
Tabatabai (1989) investigated the design of subband filters for a given input source and quantisation noise model that minimise the average reconstructed signal error. No filter designs were given and the solution was posed as a difficult non-linear optimisation problem. The approach given in this chapter is another approach to essentially the same problem as considered by Tabatabai.

Gurski et al (1992) described a design procedure for an over-sampled subband pyramid scheme that minimises the reconstruction error when only the lowband is used to synthesize the signal. The filters can be designed for a given data set, such as a database of typical images, or using a stochastic model. It is worth noting however that the over-sampled pyramid decomposition is considered inferior to a critically sampled system for image compression [Simoncelli and Adelson, in Woods 1991, p163].

Desarte et al (1992) designed filters for an octave-band filter bank (DWT) that attempted to minimise the highpass signal variance, hence maximising the coding gain. The filters were numerically optimised for each image, and required a small transmission overhead. The authors reported better subjective results compared to the JPEG method. It is interesting to note that an improvement over static filters was reported only for some images depending on texture content.

Caglar et al (1991) considered the design of CQF's that attempted to optimise certain parameters such as energy compaction, aliasing minimisation, step response deviation, uncorrelated subbands, and linearity of phase. The design was formulated as a non-linear optimisation problem with non-linear constraints. Vandendorpe (1992) considered a similar problem, and attempted to maximise the coding gain for a two-band filter bank. Vandendorpe noted that the maximum coding gain also leads to the maximum energy compaction as with the block transform KLT.

Daubechies (1988) derived some CQF's that are subsequently referred to as Daubechies wavelets or filters. She considered continuous wavelets derived using a DWT of infinite tree-depth. The Daubechies CQF's were designed so that the resulting continuous wavelet is maximally regular or smooth. As a consequence, these filters generate lowpass filters at each stage of the DWT analysis tree that are reasonably smooth. A highly correlated source, a basic model of images, is relatively smooth in nature. As such the design of regular wavelets can in some sense be considered as a filter design based on a typical image model. This topic is discussed in more detail in
Section 4.7. Since Daubechies' paper there has been significant interest in wavelets and a detailed literature survey is beyond the scope of this chapter. Rioul (1993) presented results evaluating the effect of regularity for DWT image coding. He concluded that some measure of regularity is desirable and that this can be achieved with relatively short filters.

Most of the subband filter design procedures to date have been concerned with two-band filter banks. This follows from the fact that the two-band filter bank is easily cascaded to give various other subband configurations. Further, the recent interest in wavelets and the DWT, which are based on a tree-structured two-band filter bank, has fuelled interest in this area. This chapter is primarily concerned with the design of subband filters for a two-band filter bank. These filters may then be used in tree-structured filter banks, such as the DWT, to implement various subband decompositions.

In this chapter optimum CQF's are designed using an eigenvector decomposition. It is well known that the absolute minimum of a quadratic matrix function, under a unit energy constraint is given by the eigenvector corresponding to the minimum eigenvalue. This property can be used to solve optimisation problems that can be formulated as a quadratic equation. Several authors have used this method for filter design. For example Jain and Crochiere (1984) designed QMF's in the time domain based on an iterative procedure that used this method. Also, Vaidyanathan and Nguyen (1987) designed linear phase filters with good passband and stopband characteristics using this method. As stated by Vaidyanathan and Nguyen, the Kaiser window is a closed form approximation to an eigenvector solution. Eigenvectors are also used in other problems. The most obvious is the principal component decomposition or KLT, discussed in Chapter 3, which is the optimum block transform from the perspective of coding gain, energy compaction, and data decorrelation.

4.1.1. Overview of Chapter 4

As was illustrated in Chapter 3, the DCT performs remarkably well for highly correlated sources due to its asymmetric frequency response. This fact, coupled with the DCT's impressive performance for image coding, suggests that a filter design procedure based on a highly correlated source will produce effective filters for image coding. This chapter describes the design of CQF's that maximise the coding gain for a
given stationary source input to the two-band filter bank. The solution filter is an eigenvector of a stationary correlation matrix. In the design examples an AR(1) source is used as a basic model and unless otherwise indicated a correlation of $\rho = 0.95$ is assumed.

In Section 4.2, as background to the design method, various properties of the eigenvectors of stationary correlation matrices are discussed. A new theorem (4.1) pertaining to the eigenvectors corresponding to a repeated minimum eigenvalue is given. This theorem is used to determine certain properties of the maximum gain filters derived in this chapter.

The maximum coding gain CQF problem is formulated in Section 4.3., as a quadratic cost function with quadratic equality constraints. Using a Lagrangian the solution is derived whereby the optimum filter is the eigenvector of the modified source correlation matrix. The simple case of 4-tap filters is described and generalised to arbitrary even length filters. A description of the practical design algorithm is given and design examples are presented and discussed.

In Section 4.4 various properties of the optimum filters are discussed. The scope of this design procedure is considered in the first subsection. Using Theorem 4.1 some magnitude properties of these filters are discussed. Also some KLT-like properties are investigated in Subsection 4.4.3.

A constraint on the number of DC zeros of the highpass filter is considered in Section 4.5. The problem is formulated using a subspace constraint and is equivalent to the design of "optimum" wavelets. By constraining the filter in this way the optimum solution can be obtained using the previous eigenvector method.

In Section 4.6 it is demonstrated that the optimum solution is not unique. There are several optimum filters with the same magnitude response and differing phase (and impulse) responses. A design example is given to illustrate this point.

The coding gains of the optimum filters are evaluated in Section 4.7. Issues such as filter length and robustness to differing source models are considered. Finally this chapter is concluded in Section 4.8.
4.2. PROPERTIES OF CORRELATION MATRICES

This section summarises some properties of the eigenvectors of stationary correlation matrices. These properties are used in the derivation of the maximum gain CQF's. A new property of the eigenvectors of these matrices is also given as Theorem 4.1. This theorem is used later to determine various properties of the optimum CQF's.

From Chapter 3, the \((i,j)^{th}\) entry of the correlation matrix of a stationary scalar source is given by,

\[
[R_{xx}]_{i,j} = r_{xx}(i - j)
\]

so that \(R_{xx}\) is symmetric (Hermitian in the complex case) and Toeplitz.

The spectral theorem of linear algebra [Strang 1988, p309] states that every real symmetric matrix has a complete set of orthogonal eigenvectors. Consider a symmetric matrix with an eigenvalue \(\lambda\) of multiplicity \(k\), and \(k\) associated linearly independent eigenvectors, \(x_1, \ldots, x_k\) given by the spectral theorem. Any linear combination of these eigenvectors is itself an eigenvector corresponding to \(\lambda\). That is,

\[
A(\alpha_1x_1 + \ldots + \alpha_kx_k) = \lambda(\alpha_1x_1 + \ldots + \alpha_kx_k)
\]

Thus, for a symmetric matrix, an eigenvalue of multiplicity \(k\) has a subspace of corresponding eigenvectors of dimension \(k\). This subspace is referred to as a \(k^{th}\) dimensional eigenspace.

A matrix \(A\) is centrosymmetric if,

\[
a_{ij} = a_{N-1-i,N-1-j} .
\]

Since a stationary correlation matrix is symmetric and Toeplitz, it is also centrosymmetric [Cantoni and Butler, 1976a]. A symmetric centrosymmetric (SC) matrix is symmetric about both the main and the anti-diagonal and is sometimes referred to as doubly symmetric or per-symmetric.
Cantoni and Butler (1976a) gave several properties of SC matrices. Some of these properties are repeated here for convenience. Following Cantoni and Butler's conventions, let $V_N$ denote the set of SC matrices of dimension $N \times N$, and let $Q$ be a member of this set. When $N$ is even $Q$ can be partitioned as,

$$Q = \begin{bmatrix} A & C^T \\ C & JA \end{bmatrix}$$

where $A$ and $C$ are of dimension $N/2 \times N/2$, and $J$ is an anti-diagonal matrix of ones. A similar partitioning is possible in the case where $N$ is odd. However since the problem at hand requires even $N$ the odd case is not considered here. For $Q, Q_1 \in V_N$ the following properties hold,

(a) $JQJ = Q$
(b) $QQ_1 \in V_N$
(c) $Q^{-1} \in V_N$
(d) $Q^T \in V_N$

A symmetric matrix that obeys equation (a) is centrosymmetric and hence is SC. In fact Cantoni and Butler define $V_N$ to be the set of symmetric matrices that obey equation (a). Also, $Q$ is orthogonally similar to the matrix,

$$\begin{bmatrix} A - JC & 0 \\ 0 & A + JC \end{bmatrix}$$

It follows that the eigenvalues of $A \pm JC$, given by the equations,

$$\begin{align*}
(A - JC)U &= UA_u \\
(A + JC)Y &= YA_v
\end{align*}$$

are the eigenvalues of $Q$. The matrix of eigenvectors of $Q$ is given by,

$$\frac{1}{\sqrt{2}} \begin{bmatrix} U & Y \\ -JU & JY \end{bmatrix}$$

and hence there are $N/2$ skew symmetric eigenvectors corresponding to $A - JC$ and $N/2$ symmetric eigenvectors corresponding to $A + JC$. It is possible to form an
asymmetric eigenvector if and only if there is a repeated eigenvalue and that associated with this eigenvalue there is a symmetric eigenvector and a skew symmetric eigenvector. In other words $A - JC$ and $A + JC$ have an eigenvalue in common.

To summarise, symmetric Toeplitz matrices, as a subset of SC matrices, possess the following properties,

$P1$ A complete set of orthogonal eigenvectors (applies to any symmetric matrix)

$P2$ There exists $N/2$ symmetric eigenvectors and $N/2$ skew symmetric eigenvectors. For multiple eigenvalues it may be possible to select otherwise.

$P3$ If the minimum (maximum) eigenvalue is distinct, the corresponding eigenvector has all its zeros on the unit circle.

A proof of $P3$ is given by Robinson (1967, p271) or Makhoul (1981), and is included as a special case of Theorem 4.1, given in the following subsection.

4.2.1. Zeros of a Symmetric Toeplitz Matrix Eigenvector, Corresponding to a Repeated Minimum Eigenvalue

The last property ($P3$), can be generalised to the case of a repeated minimum eigenvalue. Although only the eigenvectors corresponding to the minimum eigenvalue are considered here, the analysis is also applicable to eigenvectors corresponding to the maximum eigenvalue. The generalisation of $P3$ is stated as,

**Theorem 4.1.** An eigenvector of a symmetric Toeplitz matrix $R$, corresponding to the minimum eigenvalue of multiplicity $k$, has at least $N-k$ zeros on the unit circle. Further every such eigenvector has $N-k$ zeros on the unit circle in common; and any vector with these common zeros is such an eigenvector.

Obviously in the case that $k=1$, this theorem is equivalent to $P3$. The complete proof of this theorem is given in Appendix A, which is a copy of a technical report. The outline of the proof, as a sequence of statements, is given as follows;

1. Any vector that minimises $x^TRx$ given $x^Tx=1$, is an eigenvector corresponding to the minimum eigenvalue (minimum eigenvector).
2. Inverting zeros (real or conjugate pairs) of a minimum eigenvector gives another minimum eigenvector.

3. Let the family of $x$ denote the set of vectors that can be generated through zero inversions and linear combinations of the resulting vectors. Every vector in the family of a minimum eigenvector is a minimum eigenvector (from 2).

4. If a vector has $N-k$ zeros on the unit circle and $k-1$ other real zeros then its family spans $k$ dimensional space.

5. If a vector has $N-k$ zeros on the unit circle and $k-1$ other arbitrary zeros, there is a vector in its family that has these same $N-k$ unit circle zeros and $k-1$ other real zeros. Hence from 5 its family spans (at least) $k$ dimensional space.

6. Therefore every minimum eigenvector must have at least $N-k$ zeros on the unit circle else its family of minimum eigenvectors would span greater than $k$ dimensional space.

7. Every minimum eigenvector has the same $N-k$ zeros on the unit circle since it must be in the space spanned by $k$ linearly independent vectors with $N-k$ zeros on the unit circle in common.

8. Any vector with these same $N-k$ zeros on the unit circle and arbitrary other zeros (giving a real vector) is a minimum eigenvector since is lies in the space spanned by the above linearly independent eigenvectors.

Statement 1 is known as Rayleigh's principle [Strang 1988, p349]. Also, Statement 2 is known under a different guise using the Fejér factorisation of a filters autocorrelation function [Robinson 1967, p335 and p269-272].

4.3. OPTIMUM TWO-BAND ORTHOGONAL (CQF) FILTERS

In Chapter 3 it was shown that the Karhunen-Loeve Transform (KLT) maximises the coding gain for a given source model over all block transforms. The KLT also maximises the energy compaction and totally decorrelates the input source. In this section the maximisation of the coding gain of an orthogonal two-band filter bank is considered. The problem is formulated in a similar manner to that of the KLT. However in this case there are additional constraints imposed by an overlapping orthogonality requirement. In effect the two-band KLT is extended to the case of two-band overlapping orthogonal filter banks. The filters that achieve the maximum gain
are referred to as optimum eigenfilters. Note that Vaidyanathan and Nguyen (1987) used a similar name for some linear phase filters that are eigenvectors of certain matrices. In this thesis the term eigenfilter refers to the maximum gain two-band orthogonal filters.

After the following preliminaries the design of optimum 4-tap filters is considered. This method is then generalised to arbitrary even length filters. The design algorithm, as implemented in MATLAB\(^1\) is described and some filter design examples given. Using an appropriate PSD it is also shown that maximisation of the average stopband attenuation can also be achieved using this method.

From Chapter 3, the coding gain for an orthogonal two-band filter bank is,

\[
G_{SB} = \frac{1}{2} \frac{\sigma_0^2 + \sigma_i^2}{\sqrt{\sigma_0^2 \sigma_i^2}}
\]

where \(\sigma_0^2\), \(\sigma_i^2\), and \(\sigma_i^2\) are the variances of the lowpass, highpass and input signals respectively. Since,

\[
\frac{1}{2}(\sigma_0^2 + \sigma_i^2) = \sigma_i^2
\]

to maximise the coding gain it is sufficient to minimise \(\sigma_i^2\). This idea has been used by Vandendorpe (1992) and Caglar et al (1991) to design optimum CQF's. Vandendorpe (1992) illustrated that the two-band filter bank with the maximum coding gain also has the maximum energy compaction (or minimum basis restriction error) and drew the obvious parallel with the block transform KLT. The independent approach given here offers several further insights.

From Chapter 3, the highpass signal variance is given by,

\[
\sigma_i^2 = h^T \mathbf{R}_x h
\]

(4.2)

where \(h\) is a vector of the highpass filter coefficients.

\(^1\)MATLAB is a trademark of The Math Works, Inc.
4.3.1. Length Four Optimum Filters

In order to clarify the optimum filter design approach, the simplest case of 4-tap filter design is described in this subsection. In order to have a PR CQF filter set, where each filter is of length four, the highpass filter, $h$, is constrained by,

$$h^T h = 1, \text{ and } h^T W_2 h = 0$$  \hspace{1cm} (4.3)

where,

$$W_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$  \hspace{1cm} (4.4)

as described in Chapter 2. These two constraints are referred to as PR1 and PR2 respectively. The optimum filter minimises the variance or cost, $\sigma_i^2 = h^T R_{xx} h$, while satisfying the PR constraints. Forming a Lagrangian gives,

$$L(h, \lambda) = h^T R_{xx} h - \lambda_2 h^T W_2 h - \lambda_1 (h^T h - 1)$$

The first term on the right hand side is the cost function, and the second and third terms associated with the Lagrange multipliers $\lambda_2$ and $\lambda_1$ are the two PR constraints. Note that removing the overlapping orthogonality constraint term associated with $\lambda_2$ leaves the cost function and constraint associated with the two-band KLT optimisation problem. Thus the problem formulation presented here can be thought of as a generalisation of the two-band KLT.

A necessary condition for a minimum of the cost function under the constraints is that the gradient of the Lagrangian is zero [Bertsekas 1982]. Differentiating with respect to $h$ and setting the resulting gradient to zero, gives,

$$\frac{\partial L(h, \lambda)}{\partial h} = 2R_{xx} h - 2\lambda_2 W_2 h - 2\lambda_1 h = 0$$  \hspace{1cm} (4.5)

$$\Rightarrow [R_{xx} - \lambda_2 W_2] h = \lambda_1 h$$
Hence the optimum filter is the eigenvector of a modified correlation matrix. Vandendorpe (1992) and Caglar et al (1991) characterised the equation as,

\[
[R_{xx} - \lambda_2 W_2 - \lambda_1 I]h = 0
\]  

(4.6)

and solved this equation numerically using an augmented Lagrangian method which is a difficult task. No further conclusions were drawn from their analysis. Writing the zero gradient equation as in (4.5) one sees explicitly that the optimum \( h \) is an eigenvector of the modified correlation matrix,

\[
A \equiv R_{xx} - \lambda_2 W_2
\]

and \( \lambda_1 \) is the associated eigenvalue. A feasible solution vector is one where the constraint equations are met. For a feasible \( h \),

\[
\sigma_i^2 = L(h, \lambda) = h^T R_{xx} h - \lambda_2 h^T W_2 h - \lambda_1 (h^T h - 1)
\]

\[
= h^T [R_{xx} - \lambda_2 W_2] h - 0
\]

\[
= h^T \lambda_1 h
\]

\[
= \lambda_1
\]

That is, for a feasible vector the cost is given by the eigenvalue \( \lambda_1 \). From property P2 an eigenvector, corresponding to a distinct eigenvalue, is either symmetric or skew symmetric. Henceforth skew symmetric is subsumed under symmetric unless otherwise indicated. Vetterli and Le Gall (1989, p1064) showed that it is not possible to have symmetric filters in a two-band orthogonal filter bank except for trivial cases. This is readily shown by considering the orthogonal matrices discussed in Chapter 2. In the case of 4-tap filters the only symmetric feasible vectors are of the form,

\[
h = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 0 & \pm 1 \\ 0 & 1 & \pm 1 & 0 \\ \end{bmatrix}^T
\]

or \( \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 & \pm 1 & 0 \\ \end{bmatrix}^T \)

The latter vector is simply the orthogonal 2x2 block transform solution. Obviously for most correlation matrices these symmetric solutions are not optimum. In order to avoid the symmetric solutions it is necessary that the eigenvalue \( \lambda_1 \) be of multiplicity two. The two corresponding eigenvectors can be linearly combined to give an asymmetric eigenvector that satisfies the PR equations.
Let $\lambda_1$ be the minimum eigenvalue of $A$. Then the cost associated with any feasible vector $x$ is,

$$\text{cost} = x' R_x x$$
$$= x' \left[ R_x - \lambda_1 W \right] x \quad \text{(for feasible } x)$$
$$\geq \lambda_1$$

with equality if and only if (iff) $x$ is an eigenvector corresponding to the minimum eigenvalue. Therefore if the solution corresponds to the minimum eigenvalue of $A$ it is the *global* minimum.

The correlation matrix of dimension 4x4 for an AR(1) source is,

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

The four eigenvalues of the modified correlation matrix $A$, derived in Appendix C (Section C.2.) using the symmetric/centrosymmetric matrix partitioning of Cantoni and Butler, are,

$$e_a = \left(2 + \rho + \rho^3 + \sqrt{8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 - 8\rho \lambda_2 - 8\rho^2 \lambda_2} \right) / 2$$
$$e_b = \left(2 + \rho + \rho^3 - \sqrt{8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 - 8\rho \lambda_2 - 8\rho^2 \lambda_2} \right) / 2$$
$$e_c = \left(2 - \rho - \rho^3 + \sqrt{-8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 + 8\rho \lambda_2 - 8\rho^2 \lambda_2} \right) / 2$$
$$e_d = \left(2 - \rho - \rho^3 - \sqrt{-8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 + 8\rho \lambda_2 - 8\rho^2 \lambda_2} \right) / 2$$

It is also shown in Appendix C (Section C.2.) that for real $\lambda_2$ and $|\rho| \neq 0$ that there are only two possible solutions for a repeated eigenvalue: namely $e_a = e_c$ corresponding to the maximum eigenvalue and $e_b = e_d$ corresponding to the minimum eigenvalue. For both solutions there is a symmetric and a skew symmetric eigenvector corresponding to the repeated eigenvalue. It follows that there are other eigenvectors corresponding to
this repeated eigenvalue that are asymmetric. Since further analytical progress has proved elusive, the solution proceeds numerically.

The value \( \lambda_2 \) can be determined numerically by substituting a specific value of \( \rho \) into the equation \( e_b = e_d \), which gives a double minimum eigenvalue of \( A \). Corresponding to this eigenvalue is a symmetric eigenvector and a skew symmetric eigenvector, which are labelled \( x_1 \) and \( x_2 \) respectively. Consider a linear combination of these vectors as,

\[
h = x_1 + \alpha x_2
\]

where \( \alpha \) is such that \( h \) obeys PR2. This condition implies,

\[
(x_1 + \alpha x_2)^T W_2 (x_1 + \alpha x_2) = 0
\]

\[
\Rightarrow \alpha^2 x_2^T W_2 x_2 + 2\alpha x_2^T W_2 x_1 + x_1^T W_2 x_1 = 0
\]

Solving this quadratic gives the appropriate value of \( \alpha \). By normalising the vector \( h \) to unit energy, PR1 is also satisfied. The resulting vector is an eigenvector of \( A \) corresponding to the minimum eigenvalue that obeys PR2 and PR1, which are given in equation (4.3). Hence it is the globally optimum solution.

The filter coefficients (and zeros) using \( \rho = 0.95 \) are given in Table 4.1

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.8848518e-01</td>
<td>9.8773731e-01 + j 1.5612496e-01</td>
</tr>
<tr>
<td>-8.3221878e-01</td>
<td>9.8773731e-01 - j 1.5612496e-01</td>
</tr>
<tr>
<td>2.2619883e-01</td>
<td>-2.7180212e-01 + j 0.0000000e+00</td>
</tr>
<tr>
<td>1.3277131e-01</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Optimum filter coefficients: AR(1) source, \( \rho =0.95 \) N=4

At the solution point the modified correlation matrix is positive definite since the minimum eigenvalue is the minimum cost which is greater than zero. In fact, the minimum cost is greater than or equal to the minimum eigenvalue of the original correlation matrix \( R_x \) since this latter eigenvalue is the minimum cost without the
overlapping orthogonality constraint. Since the modified correlation matrix at the solution point is positive definite and Toeplitz it is a valid correlation matrix describing the autocorrelation of some real discrete random process.

The optimum filter is an eigenvector of a modified correlation matrix corresponding to the minimum eigenvalue of multiplicity two. It follows from Theorem 4.1 that the filter has at least two \((N=4, k=2\) and \(N-k=2\)) zeros on the unit circle. These two unit circle zeros are the conjugate pair of zeros in Table 4.1.

For an AR(1) source as \(\rho\) tends to unity it has been observed that the two unit circle zeros migrate toward DC. In the limit as \(\rho\) tends to unity the highpass filter has two zeros at DC and is equivalent to Daubechies 4-tap highpass filter. However, as discussed later, the same is not the case for longer filters.

### 4.3.2. Arbitrary Length Optimum Filters

The above approach is generalised to arbitrary even length \((N)\) filters in this subsection. For arbitrary length filters the PR constraints are,

\[
\begin{align*}
    h^T h &= 1, \quad h^T W_2 h = 0, \quad h^T W_3 h = 0, \ldots, \quad h^T W_{N/2} h = 0 \\
    \end{align*}
\]

where

\[
W_m = W^{m-1} + (W^{m-1})^T
\]

and is of dimension \(NxN\). Forming a Lagrangian where the highpass variance is the cost function and the PR conditions are equality constraints gives,

\[
L(h, \lambda) = h^T R_x h - \sum_{m=2}^{N/2} \lambda_m h^T W_m h - \lambda_1 (h^T h - 1)
\]
Note that $\lambda_1$ is associated with the unit energy term and the ensuing $\lambda_i$ with the overlapping orthogonal terms. Differentiating with respect to $h$ and setting the resulting gradient to zero gives,

$$\frac{\partial L(h, \lambda)}{\partial h} = 2R_{xx}h - \sum_{m=2}^{N/2} 2\lambda_m W_m h - 2\lambda_i h = 0$$

$$ \Rightarrow \begin{bmatrix} R_{xx} - & \sum_{m=2}^{N/2} \lambda_m W_m \end{bmatrix} h = \lambda_i h$$

(4.7)

As in the simple 4-tap case the optimum solution is an eigenvector of a Toeplitz modified correlation matrix, which at the solution point must be positive definite. From property $P_2$, such an eigenvector, if corresponding to a distinct eigenvalue, must be symmetric. However, as stated previously, it is not possible to have a symmetric CQF vector except in trivial cases [Vetterli and Le Gall 1989, p1064]. Hence a multiple eigenvalue is needed. Since there are $N/2$ constraints it is expected that a multiplicity of $N/2$ eigenvalues is required. In this way the resulting $N/2$ dimensional eigenspace can be searched for a feasible eigenvector.

The associated cost (highpass subband variance) for a feasible $h$ is,

$$\text{cost} = L(h, \lambda) = h^T R_{xx} h - \sum_{m=2,4}^{N-2} \lambda_m h^T W_m h - \lambda_i (h^T h - 1)$$

$$= h^T \left[ R_{xx} - \sum_{m=2}^{N/2} \lambda_m W_m \right] h - 0$$

$$= h^T \lambda_i h$$

$$= \lambda_i$$

Further if the solution exists at the point where $\lambda_1$ is the minimum eigenvalue then it is the global minimum. In other words consider the cost associated with any feasible vector $x$, 

cost = \( x^T R_{xx} x \)

\[
= x^T \left[ R_{xx} - \sum_{m=2}^{N} \lambda_m W_m \right] x 
\]

\[ \geq \lambda_1 \]

with equality iff \( x \) is an eigenvector corresponding to the minimum eigenvalue.

### 4.3.3. Design Algorithm

The design method was implemented in MATLAB (version 3.5), an interactive numerical matrix based mathematical package, on a 386 personal computer.

The algorithm is outlined as follows:

**Step 1.** Construct the correlation matrix \(( R_{xx} )\), and the overlapping orthogonal matrices \(( W_m )\).

**Step 2.** Numerically determine the Lagrange multipliers \(( \lambda_i )\), that give a multiplicity of \( N/2 \) minimum eigenvalues of, \( A = R_{xx} - \sum_{m=2}^{N/2} \lambda_m W_m \).

**Step 3.** Determine the eigenvectors of \( A \) corresponding to these \( N/2 \) minimum eigenvalues. Denote these as \( \mathbf{v}_i \).

**Step 4.** Consider a vector \( \mathbf{h} \) given by, \( \mathbf{h} = \sum_{i=0}^{N/2-1} w_i \mathbf{v}_i \). Determine the \( N/2 \) weights \( w_i \), so that \( \mathbf{h} \) is a feasible (eigen)vector satisfying the \( N/2 \) PR constraints.

**Step 5.** Using the weights of step 4, construct the optimum solution vector.

The solutions to the non-linear equations in steps 2 and 4 are obtained using the "fsolve" MATLAB function. Although these steps are computationally intensive, filters up to 40-taps in length have been designed. It is possible to design longer filters, but the design time is quite lengthy and several different starting points may be required before a solution is found.

The MATLAB code and commands used to implement the algorithm is given in Appendix C (Section C.4.). Note that this code is used for a more general algorithm, where the CQF highpass filter is constrained to lie in some predetermined subspace.
The generalised algorithm is described in Section 4.5. (To implement the algorithm described here set $g_B = I$ (the identity matrix) when using these M-files - see the Appendix for details).

4.3.4. Optimum Filter Examples

Figure 4.1 illustrates the frequency response of length $N=30$ tap filters optimised for an AR(1) source of correlation $p=0.01$ and $p=0.98$. It is evident that the stopband attenuation increases towards zero frequency for the filter designed using the highly correlated AR(1) source ($p=0.98$). A similar characteristic was observed in Chapter 3 for the frequency responses of the DCT high order basis vectors. The aim of the optimisation is to minimise leakage from the dominant low frequencies, characteristic of a highly correlated AR(1) source. The filter designed for low correlated AR(1) source ($p=0.01$) on the other hand exhibits a much more even attenuation across the whole stopband spectrum.

![Frequency response plot](image)

Figure 4.1. $N=30$ Optimum highpass CQF magnitude response: (—) filter designed using an AR(1) source $p = 0.01$, (— —) filter designed using an AR(1) source $p = 0.98$.

4.3.4.1. Maximising average stopband attenuation

From Chapter 3 the highpass subband variance (cost function) is given by,
Maximising the average stopband attenuation of a filter for a given transition bandwidth can be achieved by selection of an appropriate power spectral density (PSD). For example, consider a PSD that is unity from \(-\omega_s\) to \(\omega_s\) and zero elsewhere. Note that real sources only are considered. The variance is then,

\[
\sigma_i^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{xx}(e^{j\omega}) d\omega
\] (4.8)

where \(\sigma_i^2\) is the classical average stopband attenuation measure, and the stopband is from \(-\omega_s\) to \(\omega_s\). This equation can be written in a quadratic form as in equation (4.2), where the matrix \(R_{xx}\) is the correlation matrix of a source with PSD as described above. Therefore, the optimum design method can be used to obtain a filter that globally maximises the average stopband attenuation over all possible PR orthogonal two-band highpass filters.

The highpass subband variance, given by equation (4.8), can be considered as a generalisation of the classical average stopband attenuation. The minimum stopband attenuation, stopband spectrum shape and transition bandwidth relative to the given source are implicitly considered simultaneously in the optimisation of this function.

Usually with filter design it is necessary to also consider the passband characteristics when constructing a cost function. However, as noted by Vaidyanathan (1987), in this case the passband is implicitly considered via the PR conditions. The PR condition in the frequency domain, is stated here again for convenience as,

\[
|H(e^{j\omega})|^2 + |H(e^{j(\pi+\omega)})|^2 = 2
\] (4.9)

Consider the magnitude response at \(\omega_1\), a point in the stopband, \(\omega_1 \in [0, \omega_s]\). For a reasonable stopband attenuation, the term \(|H(e^{j\omega_1})|^2\) is negligible, and therefore from equation (4.9),
$$|H(e^{j(\pi+\omega_1)})|^2 = 2$$

This corresponds to the frequency $\pi+\omega_1$ radians (or $\pi-\omega_1$ for real filters), which is in the passband. It follows that the better the stopband attenuation the flatter the passband spectrum. The transition region is from $\omega_s$ to $\pi-\omega_s$. For real filters it is readily seen from (4.9) that,

$$|H(e^{j\pi/2})|^2 = 1$$

That is $\omega=\pi/2$ is the 3dB point since the passband gain is 2. Therefore the 3dB bandwidth of any CQF is $\pi/2$ radians (excluding those with an exotic frequency response). Equation (4.9) guarantees that the sum of the magnitude response at any frequency and that of its reflection about this 3dB point is 2, a constant. This illustrates that this PR condition is simply a mirror image constraint.

This eigenfilter method is suitable for the design of a lowpass or highpass filters with a 3dB bandwidth of $\pi/2$ radians. The stopband attenuation can be increased or decreased as a trade-off against the transition bandwidth. Figure 4.2 illustrates the magnitude response of 30-tap filters designed using such PSD's where $\omega_s=0.27\pi$ and $\omega_s=0.225\pi$. The filter designed for the former source exhibits significantly more stopband attenuation at the expense of a larger transition bandwidth.
Figure 4.2. Magnitude response of 30-tap filters maximising average stopband attenuation: (—) $\omega_s=0.225\pi$ (0.225 cycles/sample), (— —) $\omega_s=0.2\pi$ (0.2 cycles/sample)

The stopband attenuation characteristic of the filter designed with $\omega_s=0.225\pi$ is very similar to that of the filter designed using a low correlated (AR(1) $\rho=0.01$) source. The attenuation is fairly even across the whole of the stopband spectrum. However there is a slight increase in attenuation away from the passband. This is in contrast to the equiripple Chebyshev filters where the stopband attenuation is even across the stopband spectrum. The design method presented here is a minimum mean square error problem. Although the equiripple Chebyshev and minimum mean square error criteria are similar they are not equivalent.

### 4.4. PROPERTIES OF OPTIMUM FILTERS

In this section various properties of the optimum CQF's are discussed. The scope of the proposed solution is considered in the first subsection. Using Theorem 4.1 some magnitude properties of the optimum filters are derived. Finally it is shown that the optimum CQF's share three properties with the block transform KLT: namely maximum coding gain, maximum energy compaction, and data decorrelation.
4.4.1. Scope of the Optimum Solution

The overlapping orthogonal PR conditions apply to the highpass (or any) filter of any two-band orthogonal PR filter bank. Hence the maximum gain CQF's derived here, offer the maximum gain over all two-band orthogonal PR filter banks.

The question of the existence of the solution at the minimum eigenvalue point is obviously relevant. Although the above method has been used to derive the optimum filters for many different correlation models a proof of the general case has been elusive. This topic is an area of further research. Although it remains unproven, the solution point of the problem appears to be the point where the minimum eigenvalue of the A matrix is maximised. That is the desired Lagrange multipliers, $\lambda_2 \ldots \lambda_{N/2}$ are such that $\lambda_1$ is maximised. For most correlation matrices this point corresponds to a repeated minimum eigenvalue of multiplicity $N/2$. However it is possible to construct a "pathological" correlation matrix where the optimum vector is a trivial symmetric eigenvector corresponding to a unique minimum eigenvalue. It is worth noting that this "pathological" solution point corresponds to the maximum of the minimum eigenvalue as conjectured above.

4.4.2. Magnitude Response Properties

From Theorem 4.1, since the solution vector is an eigenvector corresponding to a minimum eigenvalue of multiplicity $N/2$, it will have at least $N/2$ zeros on the unit circle. This is observed in Figures 4.1 and 4.2 where there are seven zeros in the stopband spectrum, each comprising a complex pair of zeros, and one zero at DC. This gives fifteen zeros on the unit circle for these 30-tap filters. In general it has been observed that there are only $N/2$ zeros on the unit circle. Also, as has been observed, it is reasonable to expect that the unit circle zeros are in the stopband while the remaining $N/2-1$ zeros, and the overall filter scaling factor, are such that the $N/2$ PR equations are satisfied.

Assuming that there are $N/2$ zeros on the unit circle, if $N/2$ is odd, then there will be an odd number of zeros on the unit circle. Hence at least one unit circle zero must be real and resides at $\pm 1$ (0 or $\pi$ radians). For an AR(1) source of positive correlation this zero will reside at +1 (DC) since it will have more of an attenuating effect than at -1 ($\pi$ radians). Therefore if $N/2$ is odd there is a zero at DC. However, if $N/2$ is even this
is not the case in general. In Section 4.5 a method is discussed for placing an arbitrary number of zeros at DC and optimising the vector in the resulting subspace.

As discussed previously, filters designed using a highly correlated source exhibit an increasing attenuation toward DC. The unit circle zeros are located so as to produce such a magnitude response.

4.4.3. Karhunen-Loeve properties of optimum filters

The similarities between the KLT block transform and the optimum CQF's are obvious. Vandendorpe (1992) illustrated that the optimum CQF's share two properties with the KLT: namely the maximum coding gain \( (K_1) \) and optimum basis restriction error \( (K_2) \). In general these two criteria are closely linked and for the two-band case are equivalent.

As mentioned in the introduction of this chapter, the optimum CQF solution is a generalisation of the two-band KLT. The PR requirements require a modification of the correlation matrix so that a feasible eigenvector exists: this is a type of generalised eigenvector problem. As previously discussed, using a sufficient condition for the maximum coding gain filter set, it is straightforward to verify that the CQF's obtained achieve the globally maximum gain. It is also shown below that the optimum CQF decorrelates the source as in the case of the KLT.

Consider the output cross-subband correlation,

\[
    r_{y_h y_i} = h_0^T R_{xx} h_i = h_i^T R_{xx} h_0 = r_{y_i y_0}
\]

where \( h_0 \) and \( h_i \) are the lowpass and highpass filters respectively. If one has an optimum CQF then,

\[
    r_{y_h y_i} = h_0^T R_{xx} h_i = h_0^T \left( R_{xx} - \sum_{m=2}^{N} \lambda_m W_m \right) h_i = h_0^T \lambda_i h_i = 0
\]
since the filters are orthogonal. This relationship shows that the subbands are decorrelated. This is equivalent to the property $K3$ of the KLT: data decorrelation.

Clarke (1981) showed that as $\rho$ tends to unity for an AR(1) source, the KLT tends towards the DCT. A similar observation has been made about the optimum CQF's. As $\rho$ tends to unity it has been observed that the $N/2$ unit circle zeros of the optimum highpass CQF tend toward the first $N/2$ zeros of the highest frequency DCT basis vector of length $N$ (if $N/2$ is even then the zeros correspond to the first $N/2$ zeros of the second highest DCT basis vector). In the case $N=4$ this gives two zeros at DC which is equivalent to Daubechies 4-tap filter, as mentioned previously. However for longer $N$ this is not the case: there is one or two zeros at DC, but there are other unit circle zeros not at DC according to the distribution of DCT basis vector zeros. The highpass Daubechies filter of $N$-taps has $N/2$ zeros at DC.

### 4.4.4. Minimisation of Aliasing Energy

Even for perfect reconstruction filter banks, in the presence of quantisation noise, aliasing introduced by decimation is not cancelled in the synthesis stage. Therefore it is also desirable to minimise the aliasing introduced by the decimation stage of the analysis. It is worth noting that Kronander (1989b) observed that uncancelled aliasing can cause significant image degradation from a subjective perspective.

Consider the lowpass subband in a two-band filter bank. Aliasing is generated by the decimation process when the lowpass filter has a non-zero magnitude response between $\pi/2$ and $3\pi/2$ radians. The unwanted energy from the lowpass subband that is aliased in the decimation process is,

$$\frac{1}{2\pi} \int_{\pi/2}^{3\pi/2} |H_0(e^{j\omega})|^2 S_{xx}(e^{j\omega}) d\omega$$

The above integrand is the PSD of the lowpass filtered subband prior to decimation. If $H_0$ is an ideal lowpass filter then this cost is zero. Similarly for the highpass subband the unwanted energy is given by,
The total aliasing energy is defined as the sum of the lowpass and highpass aliasing energy. It is straightforward to show that the total aliasing energy is then given by,

\[
\frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} |H_1(e^{j\omega})|^2 S_{xx}(e^{j\omega}) d\omega
\]

Equation (4.10) can be interpreted as the highpass subband variance with an input source PSD given by equation (4.11). As such this aliasing cost can be written in a quadratic matrix form. Therefore a highpass filter can be designed using the previous eigenvector method that minimises this total aliasing under the PR constraints.

It is interesting to note that for a highly correlated source, the aliasing cost associated with a filter designed to maximise the coding gain for such a source, is very close to the absolute minimum aliasing cost. This is not surprising, since the PSD given by (4.11), is similar to the input PSD, for a highly correlated source. Therefore maximising the coding gain for a two-band filter bank also minimises the aliasing for a highly correlated source.

4.5. ZERO CONSTRAINED FILTERS: OPTIMUM WAVELETS

As mentioned in Section 2.4.2 of Chapter 2, for subband image coding applications it is generally believed that the highpass analysis filters should have a zero at DC. For CQF's, the zeros of the highpass filter at DC are mapped to zeros at π radians for the lowpass filter. Hence, for CQF's, this DC constraint leads to an admissible wavelet filter, whereby the lowpass filter has at least one zero at π radians [Daubechies, 1988]. The regularity of the lowpass filter increases with an increasing number of zeros at π radians. The problem of minimising the variance of the highpass CQF, with added DC constraints, can then be considered as an optimum wavelet problem.
In this section the design of maximum coding gain CQF's with a constraint on the zeros of the highpass filter is investigated. By constraining the solution vector to be in a desired subspace it is shown that an extension of the previous design method can be used. In this way globally optimum filters are obtained.

An arbitrary vector, \( h \), with a zero at DC can be written in the form,

\[
h = Bx
\]  

(4.12)

where,

\[
B = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
-1 & 1 & & & \\
0 & -1 & \cdots & \cdots & \\
0 & 0 & \cdots & \cdots & 1 \\
\vdots & \cdots & \vdots & \vdots & \\
0 & \cdots & 0 & -1
\end{bmatrix}
\]

and is of dimension \( N \times (N-1) \). The vector \( x \) is of length \( N-1 \) and the matrix \( B \) has rank \( N-1 \). The column space of \( B \) spans the space of zero DC component vectors. Alternatively one can show that a constant vector spans the left null space of \( B \). Hence every vector \( h \) with a zero at DC can be written in this form and any vector written in this form has a zero at DC. It is shown in Appendix C (Section C.3.) that if \( x \) is symmetric then \( h \) is skew symmetric and if \( x \) is skew symmetric then \( h \) is symmetric.

Constructing a Lagrangian equation using the subspace constrained vector gives,

\[
L_B(Bx, \lambda) = h^T R_x h - \sum_{m=2}^{N} \lambda_2 h^T W_m h - \lambda_1 (h^T h - 1)
\]

\[
= x^T B^T R_x B x - \sum_{m=2}^{N} \lambda_2 x^T B^T W_m B x - \lambda_1 (x^T B^T B x - 1)
\]

Differentiating and setting the gradient to zero gives,
\[ \frac{\partial L_B(Bx, \lambda)}{\partial x} = 2B^TR_xBx - \sum_{m=2}^{N} 2\lambda_mB^TW_mBx - 2\lambda_1B^TBx = 0 \]

\[ \Rightarrow (B^TB)^{-1}B^T \left( R_{xx} - \sum_{m=2}^{N} \lambda_mW_m \right) B = \lambda_1x \]

or \[ (B^TB)^{-1}B^TAX = \lambda_1x \quad (4.13) \]

Note that \( B^TB \) is invertible since \( B \) has linearly independent columns (or rank \( N-1 \)). This equation is similar to the unconstrained case, except that modified correlation matrix, \( A \), is further modified by projecting its columns onto the subspace of zero DC constrained filters (See Appendix C, Section C.3.). In Appendix C (Section C.3.) it is shown that \((B^TB)^{-1}B^TAX\) is symmetric and centrosymmetric (SC). If the eigenvalues of this matrix are distinct, then \( x \) is symmetric, giving (skew) symmetric \( h \). Hence, as with the unconstrained design method, it is necessary to enforce a multiplicity of eigenvalues. Similarly, since there are still \( N/2 \) PR constraints, it is expected that a multiplicity of \( N/2 \) eigenvalues is required, generating a \( N/2 \) dimension eigenspace. This subspace is then searched for a feasible \( x \). Finally the optimum filter is given from equation (4.12).

The matrix product \( Bx \) represents a convolution of a 2-tap filter with coefficients \{-1,1\} with the filter \( x \). Therefore the zeros of \( h \) are those of \( x \) and an extra zero at DC (the zero from the filter \{-1,1\}). Since the optimum feasible vector \( x \) will generally be an eigenvector corresponding to a repeated eigenvalue of multiplicity \( N/2 \), it will have \( N-1-N/2 \) zeros on the unit circle, noting that the length of \( x \) is \( N-1 \). Therefore \( h \) will have these unit circle zeros and another zero on the unit circle at DC giving \( N/2 \) unit circle zeros in total as in the unconstrained case.

Figure 4.3 shows the magnitude response of an \( N=12 \) tap filter designed for an AR(1) source with correlation coefficient \( \rho=0.98 \), under a zero DC constraint. As in the unconstrained case the attenuation increases toward DC, counteracting the dominant frequencies of a highly correlated AR(1) source. Note that there are two zeros at DC and a complex pair in the low frequency spectrum giving \( N/2 = 6 \) zeros on the unit circle as expected.
This method is not restricted to a single zero constraint. It is possible to use any \( B \) which is constructed from a single symmetric or skew symmetric vector \( b \) of the form,

\[
B = \begin{bmatrix}
  b(0) & 0 & 0 \\
  b(1) & b(0) & \ldots \\
  \ldots & b(1) & \ldots \\
  b(K-1) & \ldots & \ldots & 0 \\
  0 & b(K-1) & \ldots & b(0) \\
  \ldots & \ldots & \ldots & \ldots \\
  0 & 0 & \ldots & b(K-1)
\end{bmatrix}
\]

In Appendix C (Section C.3.) it is shown that such a \( B \) matrix leads to a SC \((B^TB)^{-1}B'AB\) modified correlation matrix, and that (skew) symmetric \( x \) leads to symmetric or skew symmetric \( h \). Therefore, as with the unconstrained design method it is necessary to enforce a multiplicity of eigenvalues. In general it is necessary to enforce a multiplicity of \( N/2 \) repeated eigenvalues giving an \( N/2 \) dimensional eigenspace in which to search for a vector that satisfies the \( N/2 \) PR conditions.
For a symmetric or skew symmetric vector \( b \) the matrix product \( Bx \) is simply a convolution (and multiplication by -1 for skew-symmetric \( b \)) of the filters \( b \) and \( x \). It follows that the zeros of \( h \) consist of those of \( b \) and \( x \). Since a multiplicity of \( N/2 \) repeated eigenvalues are required in general, from Theorem 4.1 \( x \) will have \( N-(K-1)-N/2 \) zeros on the unit circle, noting that \( x \) is \( N-(K-1) \) taps long. Therefore, if the zeros of \( b \) are all on the unit circle, then \( h \) will have \( K-1 \) unit circle zeros from \( b \) and \( N-(K-1)-N/2 \) unit circle zeros from \( x \), giving a total of \( N/2 \) unit circle zeros.

The skew symmetric binomial filter of length \( k \) has \( k-1 \) zeros at DC. Using this binomial filter as a basis vector it is possible to construct a basis matrix where the column space spans the space of filters with \( k-1 \) zeros at DC. For example in the case of \( k=4 \),

\[
B = \begin{bmatrix}
1 & 0 & 0 \\
-3 & 1 \\
3 & -3 & \ldots \\
-1 & 3 & \ldots & 0 \\
0 & -1 & \ldots & 1 \\
0 & 0 & \ldots & -3 \\
\ldots & \ldots & \ldots & 3 \\
0 & \ldots & \ldots & -1
\end{bmatrix}
\]

where \( B \) is of size \( N \times (N-3) \). Proceeding as before gives the optimum filter, which in this case will have at least three zeros at DC.

If the basis were to have the maximum number of zeros at DC \( (N/2) \), the solution would be that given by Daubechies (1988). In this case there is no optimisation since the whole remaining subspace is required in order to find a feasible vector. It is interesting to note that at this point it has been observed that the modified correlation matrix is a multiple of the identity matrix. The minimum eigenvalue of multiplicity \( N/2 \) is that on the diagonal of the matrix.

In Appendix C (Section C.3.) it is shown that the same method can be used with the substitution \( h=Boy \) where \( Bo \) has orthonormal columns. The optimum filter is then given by solving,

\[
(Bo^TABo)y = \lambda y
\]  

(4.14)
where $h = B_0 y$ satisfies the PR constraints. This is achieved by forcing a minimum eigenvalue ($\lambda_e$) of multiplicity $N/2$.

Equation (4.14) can be seen as a generalisation of the corresponding equation for the unconstrained case, (4.7). For $B_0 = I$ (the identity matrix), equation (4.14) is the same as that for the unconstrained case. The design algorithm for the optimum CQF's given in Section 4.3.3 is easily generalised to this subspace constrained problem. In steps 2 and 3 of the original algorithm $A$ is substituted for $B_0^T A B_0$. Now, consider a matrix $V$ whose columns consist of the $N/2$ eigenvectors of $B_0^T A B_0$ obtained in step 3. The solution vector is given by $h_e = B_0 y_e$, where $y_e$ is a linear combination of the columns of $V$ that satisfies constraints imposed by the PR constraints of $h$. Making the substitution $g_V = B_0 V$ means that $h_e$ can be determined directly from a linear combination of the columns of $g_V$. This substitution is made in step 3 of the generalised algorithm so that the ensuing steps 4 and 5 are identical to those of the previous algorithm.

The MATLAB M-files used to implement this generalised algorithm are given in Appendix C (Section C.4.). By making the substitution $g_B = I$ (the identity matrix), the M-files can be used to implement the design of the unconstrained optimum CQF's. Some zero-constrained optimum CQF's, designed using an AR(1) source of correlation $\rho = 0.98$, are listed in Appendix D.

4.6. IMPULSE RESPONSE SELECTION

The optimum filter is not unique. In the first place a time-reversed version of the optimum filter has the same minimum cost and obeys the PR constraints. For filters of length greater than 6 there are also other possibilities. The amount of freedom and the possible use thereof is the topic of this section.

Consider a filter in the Z-transform domain,

$$H(z) = A(1 + a_1 z^{-1} + \ldots + a_n z^{-n})$$

$$= A(1 - r_1 z^{-1}) \ldots (1 - r_n z^{-1})$$

and let,
\[H_i(z) = A_i(1-r_1z^{-1})(1-r_{k-i}z^{-1})(1-r_kz^{-1})(1-r_{k+i}z^{-1})(1-r_{k+1}z^{-1})\]

where the zero \(r_k\) is real. In other words \(h_i\) has the same zeros as \(h\) except that \(r_k\) is inverted. In the proof of Theorem 4.1, Statement 2 (see Appendix A), it is demonstrated that the magnitude response of \(h_i\) (\(H_i\)) and \(h\) (\(H\)) are identical assuming both are normalised to unit energy. Similarly inverting a complex conjugate pair of zeros leaves that magnitude response unchanged (following energy normalisation).

From Chapter 2, PR is dependent on the frequency domain condition,

\[|H(e^{j\omega})|^2 + |H(e^{j(\pi+\omega)})|^2 = 2\]  

(4.15)

which illustrates that PR is dependent only on the magnitude response of \(h\). Further the cost or variance equation is given by,

\[\sigma_i^2 = h'R_{xx}h = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 S_{xx}(e^{j\omega}) d\omega\]

which illustrates that the variance or cost is also dependent solely on the magnitude response of \(h\).

It follows that, given an optimum filter, it is possible to generate other optimum filters by inverting real zeros or complex conjugate pairs of zeros. This freedom allows the selection of differing impulse responses while maintaining the same magnitude response. The time reversed solution referred to previously is simply the case where all zeros are inverted.

In the design of various filters the following observations have been made,

- There are exactly \(N/2\) zeros on the unit circle. (Theorem 4.1 only indicates that there are at least \(N/2\) zeros on the unit circle).
- The non unit circle zeros are all complex conjugate pairs barring one real zero if \(N/2\) is even (there are \(N/2-1\) zeros not on the unit circle).
Inverting a complex zero on the unit circle (UC) maps the zero to its conjugate. Hence inverting a complex conjugate pair of zeros on the UC achieves nothing, since the zeros are unchanged. Similarly inversion of a real zero on the UC (at ±1) achieves nothing since the zero is mapped to itself. Hence only the inversion of zeros not on the UC generates a new impulse response.

Consider the case where \( N/2 \) is odd. There are \((N/2-1)/2\) non-UC complex conjugate zero pairs. An optimum filter can be generated as the product of the \( N/2 \) fixed UC zeros and \( N/2-1 \) non-UC zeros. For each complex conjugate non-UC zero pair one has two choices: inverted or non-inverted zero? The \((N/2-1)/2\) binary choices give,

\[
2^{(N/2-1)/2}
\]
different possible optimum filters. In the case where \( N/2-1 \) is odd, there are \((N/2-2)/2\) complex conjugate and one real non-UC zeros which gives,

\[
2^{(N/2)/2}
\]
different optimum filters. The above equations can be combined for arbitrary even \( N \) giving the number of optimum filters as,

\[
2^\left\lfloor \frac{N}{4} \right\rfloor
\]
where \( \lfloor x \rfloor \) is the largest integer less than \( x \). There are twice this number of filters if one considers negating the sign of the coefficients of each filter. In the design of optimum CQF's for image coding purposes time-reversal is largely irrelevant. In this case there are,

\[
2^\left\lfloor \frac{N}{4} \right\rfloor - 1
\]
(4.16)
different useful optimum filters. The freedom to select a different impulse response has been realised to some extent by Smith and Barnwell (1986). In their design of CQF’s they used a spectral factorisation of a half-band filter. In the process of the spectral factorisation one has the freedom to choose between zeros outside the unit circle and their (inversions) reflections inside the UC. They suggested a selection of zeros that
leads to a filter with approximately linear phase. This freedom has also been recognised in the case of Daubechies filters [Daubechies 1988].

The different possible impulse responses for a lowpass 12-tap optimum filter designed using an AR(1) source of correlation coefficient $\rho=0.98$ are illustrated in Figure 4.4.

![Figure 4.4. Impulse Response Freedom of Optimum 12-tap filters: AR(1) source, $\rho=0.98$](image)

There are four different impulse responses shown corresponding to equation (4.16). Time-reversing the above impulse responses will give another four optimum filters, giving a total of eight different filters. The impulse response freedom can be utilised for optimum coding performance. This topic is discussed further in Chapter 6.

### 4.7. CODING GAIN EVALUATION

The aim of this filter design procedure has been to maximise the coding gain of a two-band orthogonal filter bank for filters of a given length. The coding gain of a filter bank is dependent on the input source. Unless otherwise indicated the coding gain is calculated assuming an AR(1) input source of correlation $\rho=0.95$. Note that the filter design procedure involves maximising the coding gain for a particular source, which is not necessarily the same source as used to calculate the coding gain. This default source
for calculating the coding gain is used to avoid confusion. The coding gain for various eigenfilters is evaluated in this section. In particular the effects of filter length, differing source statistics, and performance as compared to other filters and filter banks are investigated.

Figure 4.5 illustrates the coding gain for an octave band filter bank (DWT) versus tree-depth $S$, using different length filters designed to maximise the coding gain for an AR(1) $\rho=0.95$ source (the same source used to calculate the coding gain). Also shown on the figure is the gain of the ideal (brick-wall) filter bank, which upper bounds the gain of all orthogonal two-band filter banks [De Queiroz and Malvar 1992]. It is evident that the coding gain increases with increasing filter length, as expected, and quickly approaches this upper bound. The gain of the 10-tap filter is quite close to this bound, while that of the 30-tap filter is nearly identical.

Strictly speaking the DWT with the maximum coding gain for a given filter length is one that employs filters optimised for each stage of the analysis tree-structure. Nevertheless it has been observed that there is virtually no increase in the coding gain by using such filters over using filters designed for the first stage only at all stages of the tree-structure (at least up to a tree-depth of $S=6$). A more general observation has
been that the coding gain for a given filter set is relatively insensitive to changes in the correlation $\rho$. For example, the coding gain of filters optimum for an AR(1) source of correlation $\rho=0.9$ is nearly identical to that of filters designed using $\rho=0.98$ using any AR(1) source of high correlation ($\rho>0.8$) to calculate the coding gain. A similar observation has been made about the block transform KLT [Clarke 1985] and "optimum" lapped orthogonal transform [Malvar and Staelin 1989]. It has been observed that if the highpass filter has a reasonable number of stopband zeros near the region where the input PSD is dominant, the coding gain using an AR(1) source is quite close to that for the optimum filter set. It is worthwhile noting that same is not necessarily true for other sources.

As the filter lengths increase, the optimum filter tends towards the ideal filter, regardless of the source, and hence a high gain can always be achieved using sufficiently long filters. However, for shorter filters the gain is significantly inferior for "mismatched" sources. For example Figure 4.6 illustrates the gain for a DWT using 8-tap filters optimised for AR(1) sources of correlation $\rho=0.01$, $\rho=0.95$ and $\rho=0.98$. The gain for the DWT using the $\rho=0.01$ optimum filters is significantly lower than that of the DWT using the $\rho=0.95$ and $\rho=0.98$ filters. As indicated previously the gain for the latter two filter sets is very similar.

Figure 4.6. Coding gain for DWT using 8-tap optimum filters.
Finally, Figure 4.7 illustrates the gain for a DWT using 12-tap Daubechies filters, AR(1) $\rho=0.95$ source optimum filters, and for the DCT. The gain of Daubechies filters is very close to the $\rho=0.95$ optimum filters. It has been observed that different length Daubechies filters all perform very close to the corresponding optimum filter for a highly correlated source. This is not surprising for two reasons. The first is that Daubechies highpass filters have many zeros at DC, which suppress the dominant low frequencies of a highly correlated source. The second reason is that, as mentioned previously, the coding gain is relatively insensitive to the location of the highpass filter zeros in the stopband, as long as they are near the dominant region of the input power spectral density.

The coding gain of the DCT is lower than that of the DWT until $\log_2 M=4,5$. After this point it is greater than that of the DWT, tending towards the absolute maximum for any filter bank; the inverse of the spectral flatness measure. For an AR(1) $\rho=0.95$ source this bound is at 10.11dB. De Queiroz and Malvar (1992) demonstrated that the ideal DWT tends to a lower bound than this absolute upper bound. However as illustrated here and by these authors the gain of the DWT for a highly correlated source is certainly close to this ideal bound for moderate $S$. It is also worth noting that since images are nonstationary in nature, the small increment in gain suggested by the coding gain for large $M$ and $S$ is not likely to be achieved in practice. For example the optimum DCT block size for image coding (of 256x256 or 512x512 resolution) is suggested to be around 8x8 to 32x32 [Clarke 1985]. Larger block sizes are not only computationally expensive, they are also often inferior. This phenomenon is also illustrated in Chapter 6.
Early subband filter designs attempted to minimise such criteria as average stopband attenuation, passband ripple, and aliasing. However, more recently there has been some effort toward designing filters based on a model of a typical source input to the filter bank. In this chapter the design of two-band orthogonal filters (CQF's) that maximise the coding gain metric have been considered. By maximising the coding gain the filter's stopband attenuation and transition bandwidth are in a sense optimised relative to a given source under the perfect reconstruction (PR) constraints.

The optimum filters, those that maximise the coding gain, are derived as the eigenvectors of a modified correlation matrix. In Section 4.2 various properties of correlation matrices were discussed and a new theorem (4.1) given. This theorem relates the number of zeros on the unit circle of an eigenvector to the multiplicity of the corresponding minimum eigenvalue.

In Section 4.3 the method of optimum CQF design was discussed. First the simple case of 4-tap filters was considered. This was then generalised to the case of arbitrary even length filters. The necessary conditions for a maximum coding gain were given.
Further, a sufficient condition for a global optimum was derived. Although it has not been shown that this sufficient condition is necessarily attainable in general, all filter designs to date have achieved this sufficient condition. The practical design algorithm, as implemented in MATLAB, was outlined and several design examples presented. It was shown that the minimum average stopband attenuation problem under the PR constraints, is a special case of this design method. In effect the maximum coding gain problem leads to filters with the best stopband attenuation relative to the given input source.

In Section 4.4 various properties of the optimum filters were considered. Theorem 4.1 was used to predict the number of zeros on the unit circle of these filters. It was demonstrated that the optimum filters share three properties with the block transform KLT: namely maximum coding gain, maximum energy compaction, and data decorrelation.

A constraint forcing a number of highpass filter zeros at DC was investigated in Section 4.5. Maximising the coding gain, under such a constraint, is an optimum wavelet problem. It was shown that constraining the solution vector to be in an appropriate subspace, an extension of the previous design method can be used. The resulting wavelets, globally optimise the coding gain for CQF's where the highpass filter has a given number of zeros at DC.

In Section 4.6 it was demonstrated that there are \(2^{|N/4|}\) different optimum filters with the same magnitude response and differing impulse responses. An example using \(N=12\)-tap filters was presented. This freedom can be used to enhance coding performance, and is used in the subband coders presented in Chapter 6.

Finally, in Section 4.7 the optimum filters were evaluated and compared against other filters using the coding gain metric. From this study the following conclusions were made about the optimum filters:

- Stopband attenuation increases at all stopband frequencies as the filter length increases, regardless of the source model. Hence a near maximum coding gain for any source can be achieved using long filters.
• However, a near maximum coding gain can be achieved with relatively short optimum filters (8-12 taps for example) if the filters are optimised for the given source.

• The coding gain for longer optimum filters is nearly identical to the absolute maximum for an orthogonal two-band filter bank.

• The optimum filters are relatively insensitive to the correlation for an AR(1) source.

• As a generalisation of the previous conclusion, CQF filter sets where the highpass filter has stopband zeros near the dominant region of the PSD, have a near optimum coding gain. As an example Daubechies filters have a near maximum coding gain for a highly correlated AR(1) source.
CHAPTER 5:

A GENERIC QUANTISATION AND ENCODING METHOD AND IMAGE SUBBAND STRUCTURES

5.1. INTRODUCTION

5.1.1. Background

Although a brief overview of a subband coding scheme was given in Chapter 1, a description is repeated here for convenience. A general subband image coding scheme is illustrated in Figure 5.1. An input image is decomposed by the analysis filter bank. The resulting subband signals are quantised and encoded. Quantisation is a lossy or irreversible entropy reducing process, while the encoding losslessly represents the quantised information. The encoded information is stored or transmitted depending on the application. Decoding simply reverses the encoding operation. For perfect reconstruction (PR) analysis/synthesis filter banks the quantisation is the only lossy component of the process: that is if the quantisation step is removed, the output signal is an exact replica, to within finite machine precision, of the input signal.

![Figure 5.1. Subband image codec](image)

Figure 5.1. Subband image codec
Subband analysis is a form of preprocessing. Efficient quantisation is possible by analysing the signal in such a manner. The coding gain material presented in Chapter 3 illustrates the theoretical advantage of PCM quantisation in the subband domain over PCM of the fullband original signal.

As discussed in Chapter 1, Woods and O'Neil (1986) presented the first (modern) results on subband image coding. A tree-structured subband analysis was used to partition the image into 16 equal width subbands. A bit allocation between the subbands, as described in chapter 3, was determined based on the subband variances. These subbands were then quantised using DPCM, and the resulting quantiser values Huffman coded. An adaptive technique was also proposed where the subbands were partitioned into blocks which were categorised as quiet, non-busy and busy. The available bits are then allocated between the subbands and these block categories. The results were compared favourably against the Chen and Smith (1977) adaptive DCT image coding scheme and various vector quantisation (VQ) techniques.

Gharavi and Tabatabai (1988) considered subband coding of monochrome and colour images using an octave-band decomposition. The image was decomposed into seven subbands using the octave tree-structure. The low-low subband was coded using DPCM and variable length codes, while for the other subbands it was concluded that PCM was sufficient. These non-zero PCM values were encoded using runlength positional information.

Westerink et al (1988) used VQ to quantise the subbands of a uniform 16-band (4x4) decomposition. The vectors were constructed using one pixel from each subband. It was concluded that this method offered a small improvement over PCM coding of the subbands but was inferior to the adaptive DPCM technique of Woods and O'Neil.

Since these initial subband image coders there have been various other subband hybrid proposals. For example, Fischer and Blain (1989) compared vector quantisation of block transform and subband data for image compression. It was concluded that (traditional) subband decompositions offered slightly better results than transform decompositions at low rates. Tanabe and Favardin (1992) reported a subband coding scheme that uses DPCM or DCT coding on the low (DC) subband, and PCM
quantisation on the other bands followed by entropy encoding. They reported superior results to the adaptive DPCM scheme of Woods and O'Neil.

As discussed previously in this thesis, a transform coder can be considered as a subband coder. Therefore many of the quantisation and encoding schemes proposed for transform coding can also be applied to general subband methods. For example, the adaptive Chen and Smith (1977) DCT coding scheme is readily adapted to general subband methods. Chen and Pratt (1984) presented a "Scene Adaptive Coder", which is DCT based codec. The Joint Photographies Experts Group (JPEG) standard is essentially based on this Chen and Pratt method [Wallace 1991]. The effectiveness of the JPEG method is generally acknowledged in the image coding community.

Ohta et al (1992) classified general subband coding into two broad groups: traditional subband coding and transform coding. In tradition subband coding coefficients are grouped into frequency blocks (subbands) and each block is encoded separately, while in transform coding coefficients are grouped into spatial blocks and encoded. Ohta et al argued that the transform type approach allows quantisation with better spatial adaptation, as compared to the traditional subband methods. This property is considered beneficial for the encoding of images. Using this classification, a transform type quantisation and encoding method, suitable for most subband structures, is introduced in this chapter.

5.1.2. Review and Overview

For nonstationary sources such as images, it is beneficial to allow the quantisation process to adapt to changing signal statistics. Traditional subband coders employ a fixed quantiser for each subband, with a bit allocation based on the variance of the subband. The bit allocation for each subband may be allowed to vary from image to image, but is fixed over the whole subband. It is more effective to allow (at least) the number of quantisation bits to vary within the subband, commensurate with spatially changing signal statistics within the subband.

The quantisation methods of Chen and Smith (1977) and Woods and O'Neil (1986) are an attempt to this end. As described above, using this type of method the input image is grouped not only by subband but by category such as quiet, non-busy and busy. The available quantisation bits are then allocated among the subbands and various
categories. Using different categories allows the quantisation bits to be allocated more effectively within each subband. However, the use of such categories allows only a limited degree of spatial adaptation. The Chen and Pratt (1984) scheme allows greater adaptation. Quantisation bits are implicitly allocated so that the energy of the quantisation error is independent of the energy of the subband coefficients. More bits are implicitly allocated to busy regions and less to quiet regions. This method is discussed in more detail later in this chapter.

A generic quantisation and encoding method that is suitable for any subband structure is proposed in this chapter. This method is essentially a generalisation of the Chen and Pratt (1984) method. As discussed above, this type of quantisation method has good spatial adaptation properties. Husoy (1991) used this type of scheme for an 8-band (64 in two-dimensions) uniform subband decomposition. Ohta et al (1992) also used a similar scheme in a dyadic or wavelet coding scheme. The generic subband quantisation and encoding method provides a platform with which to compare different subband analysis structures and filters for subband image coding. This comparison is the main topic of Chapter 6.

In Chapter 2 of this thesis, various one-dimensional subband analysis/synthesis systems were analysed, largely from a time domain perspective. The coding gain metric was introduced in Chapter 3 as a tool with which to evaluate these systems for a given source model. In this chapter various two-dimensional image subband structures are discussed. In Section 5.2, the power spectral density (PSD) of a typical image model is considered. Existing subband image structures are discussed in light of this model and the coding gain metric. In addition some new analysis structures are proposed. Finally some other considerations, such as filter or subband time localisation, are covered.

The generic subband quantisation and encoding method is described in Section 5.3. The baseline JPEG scheme is discussed as a basis for this method in Subsection 5.3.1. Two themes are then developed that describe how the quantisation and encoding method is to be applied to arbitrary subband decompositions. Various issues relating to these themes and a practical implementation are discussed.
5.2. SUBBAND ANALYSIS CHARACTERISTICS

Different subband analysis structures for subband image coding purposes are considered in this section. In Section 5.2.1 the power spectral densities of the generalised two-dimensional image correlation model (2DG) and a two-dimensional AR(1) model are examined.

In light of these models, in particular the 2DG model, three distinguishing characteristics of subband analysis are discussed in the following subsections. First, subband analysis methods are grouped according to some ideal partitions of the two-dimensional (spatial) frequency plane. These ideal partitions are referred to as subband structures. For example, the DCT LOT and ELT all belong to the class of uniform or $M$-band subband structures.

Secondly, the degree to which various analysis methods approximate the corresponding ideal structure is considered. In particular this approximation is considered relative to the PSD of a typical image. The coding gain metric implicitly measures these first two characteristics relative to a particular source model. Various subband methods are compared using the coding gain metric for the 2DG model, in order to predict their performance for image coding. A modified or extended wavelet analysis structure is proposed based on the 2DG model.

The third characteristic is the time resolution, or spatial resolution, of the filters employed in the subband analysis. From the perspective of the coding gain metric, time resolution is irrelevant for the coding of stationary sources. However, since still images are nonstationary, some measure of filter time resolution or localisation is desirable. In Section 5.2.3 time resolution properties are quantified and discussed in relation to coding performance.

5.2.1. Power Spectral Density of a typical Image and Ideal Subband Analysis Structures

Figure 5.2 illustrates a mesh plot of the power spectral density (PSD) of the two-dimensional AR(1) image correlation model. Similarly Figure 5.3 shows a mesh plot of the two-dimensional generalised (2DG) model. Both correlation models were introduced in Chapter 3. The DC point is in the middle of the mesh plane,
corresponding to the peak in each spectra. Both spectra are volcanic in nature, in that the PSD is increasingly dominant at lower frequencies. Also, both spectra have ridges running along the horizontal and vertical spatial frequency axes. However, it is evident that these ridges are more pronounced in the 2DAR(1) model. As stated in Chapter 3, in two dimensions the 2DG model is considered to be a much better image correlation model than the 2DAR(1) model. The results given later suggest that the large frequency axis ridges predicted by the 2DAR(1) model are the source of its inaccuracy.
Figure 5.2. PSD (dB) of 2DAR(1) source, $\rho=0.95$

Figure 5.3. PSD (dB) of 2DG source, $\rho=0.95$
The objective of signal analysis, for compression purposes, can be described as the production of subband signals with relatively flat spectra [Pearlman, in Woods 1991 p39]. Using a stationary source model the coding gain for orthogonal analysis is generally increasing for increasing analysis levels. However, this is not necessarily the case for still image compression. Since image data is nonstationary the coding performance at large analysis levels may be inferior to lower levels. Small improvements suggested by the coding gain metric for large analysis levels over moderate levels are unlikely to be attained in practice. For example, in the case of the DCT, coding experiments suggest that at moderate rates, no improvement is obtained using a block size larger than around 16x16 (See Clarke (1985) and Andrew et al (1993b, 1993d)). Hence it is prudent to ignore the small improvement suggested by the coding gain for large analysis levels.

A two-dimensional DCT divides the frequency plane roughly into subbands of equal bandwidth. For example, in the case of a 4x4 DCT image analysis the frequency plane is divided into 16 equal bandwidth subbands, as depicted in Figure 5.4.

![Figure 5.4. Frequency plane partitioning effected by 4x4 DCT](image)

Each square in Figure 5.4 represents an ideal subband encompassing all spatial frequencies delineated by the square boundaries. An ideal subband decomposition is one that is produced using ideal (brick-wall) filters. A 4x4 DCT certainly only approximates this ideal. An overlapping transform, such as the lapped orthogonal transform (LOT) or extended lapped transform (ELT) (see Chapter 2) with four one-dimensional basis filters, effects the same frequency plane partitioning. This type of analysis where the subbands are of equal bandwidth is termed uniform or $M$-band.
analysis, where $M$ refers to the number of one-dimensional subbands. Note that this gives $M^2$ subbands in two dimensions.

The octave-band or discrete wavelet transform (DWT) filter bank, on the other hand, does not partition the frequency plane into equal size subbands. Usually, as in the one-dimensional case, the two-dimensional DWT is implemented using a tree-structured analysis. At the first stage the two-dimensional frequency plane is partitioned into four subbands of roughly equal bandwidth. At subsequent stages, the low frequency subband only is partitioned in the same manner. The tree-depth, $S$, is the number of stages employed. For example a two-dimensional $S=3$ DWT analysis structure is illustrated in Figure 5.5a.

Usually this four band partition building block is implemented using two-dimensional separable filters. That is a two-band one-dimensional analysis is used on each dimension. The subband resulting from lowpass filtering in both dimensions is referred to as the LL (low-low) band, while the subband resulting from the lowpass vertical and highpass horizontal filtering is referred to as the LH (low-high) subband. The HL and HH subbands are named in a similar fashion, as described in Chapter 2.

From the PSD mesh plots in Figures 5.2 and 5.3 it is evident that the input (image) spectrum is relatively flat and of low energy in the high frequency region, while it becomes increasingly steep and contains more energy towards lower frequencies. The octave nature of the DWT analysis structure reflects this characteristic. Further analysis of high frequency regions is pointless from a coding gain perspective since the high frequency subbands have a relatively flat spectra. In addition, even if the image spectrum was uneven in these subbands, they are of such insignificant energy that any coding gain associated with these particular subbands would have little effect on the overall coding gain. In contrast, the increasing steepness and energy concentration of the PSD at low frequencies suggests a gain in analysing the low frequencies in an increasingly fine manner.

A modified discrete wavelet transform (MDWT) analysis structure is shown in Figure 5.5b. This structure is the same as the DWT except that at each stage the LH and HL bands are further analysed using a two-band one-dimensional analysis. The direction of this further analysis, as depicted in Figure 5.5b, is an attempt to utilise a gain afforded by the ridge like nature of the PSD models.
The DWT is a tree-structured analysis, where a two-band (four in two-dimensions) analysis is used at each stage of the tree. There is a whole family of such analysis structures, where an $M$-band analysis is used at each stage of the tree. For example, an analysis structure implemented using a 4-band analysis for a two-stage tree is illustrated in Figure 5.6. The structure is termed a quadic analysis, with a tree-depth $S_q$.

The one-dimensional bandwidth of the DC subband in the quadic case is given roughly by $\pi/4^{S_q}$ radians ($0.5/4^{S_q}$ cycles/sample), whereas in the DWT case the bandwidth is roughly $\pi/2^S$ radians ($0.5/2^S$ cycles/sample). As discussed in Chapter 2, the level of analysis refers to the bandwidth of the DC subband. For example, an $M$-band, DWT and MDWT have the same level of analysis for $M=4$, $S=4$ and $S_q=2$ respectively (in both one or two dimensions).
Westerink et al (1988) examined a range of different analysis structures, with equivalent analysis level up to that of the \( S=3 \) DWT, for subband image coding. They concluded that for 256x256 images the best subband splitting scheme was a uniform 4x4 (4-band) subband decomposition. However, one observes from their results that the performance of some other structures was close to that of the 4x4 decomposition. Their conclusion is dependent on the quantisation scheme employed and the filters used to implement the subband decomposition. Their scheme and results are discussed in more detail in the Discussion section of Chapter 6.

5.2.2. Approximation of Ideal Subband Analysis Structures

The analysis structures presented above are ideal subband decompositions. In practice it is only possible to approximate this ideal partitioning. Hence for coding purposes it is not only the analysis structure that is important, but also how well the structure is approximated using practical filters. For example, for the \( M \)-band transforms considered previously, the overlapping transforms such as the LOT and ELT provide a better approximation to the ideal \( M \)-band partitioning as compared to that effected by the DCT.

In terms of coding performance, the approximation of the ideal analysis structure is dependent on the frequency response of the filters relative to the input PSD. The coding gain metric is an attempt to measure the performance of a subband scheme for a given source model. In other words it implicitly accounts for the approximation of some ideal analysis structure for a given transform and source model. As illustrated in Chapter 3, in terms of compression purposes, the frequency responses of the DCT basis filters are optimised for a highly correlated source.

The coding gains for the DCT, KLT (optimum block transform), DWT, and modified DWT are illustrated in Figure 5.7. Figure 5.7a gives the coding gain using a 2DAR(1) model and Figure 5.7b gives the gain using the 2DG model, where the inter-element correlation, \( \rho=0.95 \) is the same for both models. The DWT and modified DWT structures are implemented using Daubechies 12-tap filters.
The KLT gain plot is not shown for the 2DAR(1) model since it is nearly identical to that of the DCT. For the 2DG model the KLT gain is shown up to an analysis level of $M=32$ (computational complexity is prohibitive beyond this point). This KLT has been determined by considering a one-dimensional source that is formed by column stacking the 2DG source, where the KLT of size $M$ corresponds to columns of length $M$. The resulting one-dimensional KLT has length $M \times M$. For both models the DCT performance is very close to that of the KLT.

The 2DAR(1) model suggests at least a 3dB improvement using the modified DWT or DCT as compared to the standard DWT. However, in the coding of typical imagery it has been found that the DCT and DWT perform in a similar manner in terms of PSNR (see Andrew et al (1993b) or Section 6.2 of Chapter 6). As mentioned in Chapter 3, the 2DAR(1) model is generally a poor model. The 2DG model on the other hand, being more accurate in general, suggests a much more even result for the three methods. Considering the limiting performance of the DCT at a block size of 8x8 or 16x16, then the DCT and DWT are similar while the modified DWT offers a small improvement. As stated previously the PSD for a 2DAR(1) source has more pronounced ridges than the 2DG source. This phenomenon explains the more dramatic gain improvement suggested by the 2DAR(1) source for the modified DWT over the DWT.

The coding gains for the DWT and the quadic analysis method are illustrated in Figure 5.8. Two quadic structures are indicated: one uses a 4-band MLT while the other a 4-band ELT. The DWT uses Daubechies 12-tap filters.
Figure 5.8. Coding gain for the DWT and quadic analysis structures. 2DG model

In Figure 5.8 the quadic gain is plotted against $2xS_q$ while the DWT gain is plotted against $S$. The quadic gain is given at the points $S_q=0,1,2,3$ corresponding to the points $2xS_q=0,2,4,6$ respectively. This convention is used so that the bandwidth of the DC subband, or level of analysis, is roughly the same at each point on the $x$ axis. This figure suggests that the quadic analysis structure will perform in a similar manner for image compression. The ELT based quadic structure has a slightly higher gain as compared to the MLT based structure. This is expected due to the superior frequency resolution of the ELT. However, taking into consideration the nonstationary nature of image data, it is difficult to ascertain whether the ELT based quadic structure will be superior to the MLT based quadic structure or DWT for image compression.

5.2.3. Time Width Properties of Filter Banks

The subband coding gain metric assumes a stationary source model. As demonstrated in Chapter 3 this gain metric is dependent only upon the magnitude response of the subband filters. As a consequence time resolution or localisation properties are not considered. However, since image data is nonstationary, some measure of time resolution is important. Simoncelli and Adelson (in Woods 1991, Chapter 4 p182)
concluded that, for image compression purposes, the filter bank basis functions should be localised in both the spatial (time) and spatial frequency domains.

The time width, a measure of the mean square deviation of a filter's impulse response about its mean, was introduced in Chapter 2. The uncertainty principle was also introduced whereby the time width / frequency width product is lower bound by a constant. In the case of continuous wavelets, constructed mathematically by iterating the DWT to an infinite number of decomposition levels, minimum time width Daubechies wavelets come close to this bound [Dorize and Villemoes 1991]. However for small number of decomposition levels it has been observed that the discrepancy is larger.

For filters with the same magnitude response their frequency widths are obviously the same. However filters with the same magnitude response but different phase responses have different time widths. In Chapter 4 it was illustrated that the optimum eigenfilter is not unique. There are several different eigenfilters with the same magnitude response but differing phase, and hence impulse responses. The minimum time width eigenfilter is the eigenfilter that has the smallest time width out of the group of eigenfilters with the same frequency response. It follows that this filter has the smallest time/bandwidth product out of all the filters in this group. Further, it has been observed that the minimum time width eigenfilter also has the minimum passband group delay deviation in this group.

It is proposed that the minimum time width filters are desirable for image coding for two reasons. On one count a minimum time width filter has the smallest "window" as it moves past an edge in the analysis process. This means that it is possible for the coding scheme to adapt quickly to changing image statistics. On the second count it has been observed that the step-edge response of the minimum time width filter has less ringing than filters with other phase responses. Although the step-edge rise time is slower, there is less associated ringing, which is generally perceived as visually annoying.

From the perspective of analysing images, the DWT is a good example of a trade-off between time and frequency localisation. The DWT analyses lower frequencies in an increasingly fine manner commensurate with the increasingly dominant PSD. High frequency subbands have good time localisation since the subband bandwidths are relatively large. There is little coding gain in analysing these subbands any further
since the image PSD is relatively flat and of insignificant energy in this region. Also, this extra analysis will be at the expense of high frequency subband time localisation. On the other hand low frequency subbands, where the spectrum is steep and dominant, have good frequency localisation as required for a near optimum coding gain. Since the coding gain of the DWT is close to that of the "optimum" transforms, such as the DCT, for a typical image model (see Figure 5.7b), it is predicted that the DWT has sufficient, or nearly sufficient, subband frequency resolution for image coding. The modified DWT is an attempt to increase the frequency resolution, in the regions where it is most useful, without significantly sacrificing the good time localisation properties of the DWT.

5.3. GENERIC SUBBAND IMAGE COMPRESSION METHOD

In this section a quantisation and encoding method suitable for any subband analysis structure is proposed. This coding strategy provides a platform with which to compare various subband analysis structures and filters. Due to its simplicity and effectiveness, as demonstrated by the results, the method offers a fair means of comparison. For example, for a given subband structure the same coder is used to compare different sets of filters that implement this structure. This coder does not require initialisation of such parameters as quantisation tables and entropy codes. Hence any differences in results can be attributed with a high degree of confidence to the differences between the filters.

This generic quantisation and encoding scheme is essentially a generalisation of the method proposed by Chen and Pratt (1984) as used in the JPEG baseline sequential coder [Wallace 1991]. The JPEG scheme is relatively simple and offers impressive results for still image compression.

In Subsection 5.3.1 a brief overview of the baseline sequential JPEG still image compression method is given. In Subsection 5.3.2 two principles are developed that extend this method to arbitrary subband structures. These principles involve grouping of spatially corresponding pixels from each subband, and a specified zigzag scan of the pixels in each group into a one-dimensional data stream. An alternative scanning procedure is considered and evaluated. Finally various issues relating to the estimation of the number of bits required to encode the data are discussed.
5.3.1. The Baseline Sequential JPEG Coder

A brief overview of the baseline sequential JPEG coder is described in this subsection. More detail can be found in Wallace (1991) and Pennebacker and Mitchell (1992). The outline does not follow the JPEG standard exactly but provides a guide to the method that is used. This description provides the basis for the explanation of the proposed quantisation and encoding method suitable for any subband A/S scheme in the following subsection.

The JPEG method is described as follows: An input image is transformed block by block with a two-dimensional DCT, which in the JPEG standard is an 8x8 DCT. Figure 5.9 illustrates this process for a 4x4 DCT on an image of dimension 8x8.

In Figure 5.9 each 4x4 block of pixels is transformed by a DCT to a 4x4 matrix of DCT coefficients.

The DCT coefficients are then scaled according to a quantisation factor and rounded. For example, if $Y_{k,l}(i,j)$ represents the $(k,l)^{th}$ coefficient in the $(i,j)^{th}$ DCT image block then the quantised value, $\hat{Y}_{k,l}(i,j)$, is,

$$\hat{Y}_{k,l}(i,j) = \text{round} \left\{ \frac{Y_{k,l}(i,j)}{QF} \right\}$$
where $QF$ is the quantisation factor. For fixed $(k,l)$, $Y_{k,l}(i,j)$ represents a subband of the image. On the other hand, for fixed $(i,j)$, $Y_{k,l}(i,j)$ represents the collection of DCT coefficients for the $(i,j)^{th}$ block in the original image.

In the JPEG standard it is possible to specify a quantisation factor ($QF$) that is different for different coefficients: that is $QF$ is a function of $(k,l)$. Improved subjective results can be obtained by varying $QF$ in this way.

The DC coefficient from each block is subtracted from that of the next block. This is written as,

$$
\hat{Y}_{0,0}(i,j) = \begin{cases} 
\hat{Y}_{0,0}(i,j) - \hat{Y}_{0,0}(i,j-1) & j \neq 0 \\
\hat{Y}_{0,0}(i,0) - \hat{Y}_{0,0}(i-1,0) & i \neq 0, j = 0 \\
\hat{Y}_{0,0}(0,0) - \text{DC\_estimate} & i = 0, j = 0 
\end{cases} 
$$

$\text{DC\_estimate}$ is an estimate of the average DC term. The subtraction is an attempt to exploit any correlation between DC coefficients in adjacent blocks. One can think of this process as DPCM with a first order predictor whose coefficient is unity. Note that this process is reversible since the input and output are integers.

Each DCT block is scanned in a zigzag manner into a one-dimensional data vector. The zigzag scan is illustrated for a $4 \times 4$ DCT block in Figure 5.10.

![Figure 5.10. Zigzag scan of one 4x4 DCT coefficient block.](image)

These data vectors are represented using zero run-length, amplitude and end of block (EOB) codes. This representation is best illustrated using an example: consider a one-dimensional scan block of coefficients (for a $4 \times 4$ DCT giving 16 coefficients) as,
This block is represented as,

- **Amplitude Stream**: 50, -10, 5, -1
- **Run-length Stream**: 3, 0, 1, EOB

This is an efficient representation assuming that the block contains many runs of zeros. The amplitude streams from each block are concatenated into a super amplitude stream, and similarly the run-length streams are concatenated into a super run-length stream. These super streams are then encoded using an arithmetic or Huffman coder to exploit their first order entropy. Strictly speaking the JPEG Huffman coder specifies slightly different amplitude and run-length symbols. However the basic idea is the same. The arithmetic coding method offers up to 10% improvement on the Huffman coding [Wallace 1991] but the particular implementation is covered by a patent.

This entropy coding process is also lossless or reversible. The only lossy part of the whole process is the (scaling and) rounding of the coefficients which is simply uniform quantisation. The larger the quantisation factor the coarser the quantisation.

The decoding of an encoded image is simply the reverse operation. The entropy coded streams are decoded to give the amplitude and run-length super streams. The two-dimensional DCT blocks are then reconstructed from these streams. Addition of adjacent DC coefficients reverses the DC subtraction or DPCM process. The reconstructed DCT coefficients, $U_{i,j}(k,l)$, are given by,

$$U_{i,j}(k,l) = QF \times \hat{Y}_{i,j}(k,l) = QF \times \text{round}\left\{\frac{Y_{i,j}(k,l)}{QF}\right\}$$

This reconstructed DCT image is synthesised using an inverse block DCT to give the decoded image. It is thus evident that the scaling and rounding is the only lossy part of the whole process (given that the DCT and inverse DCT are lossless).

The coefficients of the DCT transformed image are usually grouped in spatial blocks. That is, the different DCT coefficients from the same spatial block are grouped together to form the usual DCT blocks. Through a simple permutation this can be changed so that coefficients are grouped by subband. The number of subbands is given...
by the DCT block size, or equivalently the number of coefficients for each block. In mathematical terms, for fixed \((k,l)\), \(Y_{kl}(i,j)\) represents the \((k,l)\)th subband. With this conceptual grouping, the zigzag scan is seen as a scan through the subbands gathering spatially corresponding pixels. Each scan block contains coefficients with the same \((i,j)\) (spatial) index. In the next subsection a similar scan is outlined for a general subband analysis structure.

5.3.2. General Subband Analysis/Synthesis Quantisation and Encoding Method

The type of quantisation method described above, as applied to a general subband analysis scheme, is shown schematically in Figure 5.11.

The input to the quantisation and encoding process is an analysed image. Each analysed image coefficient is scaled by a constant scaling factor and rounded to the nearest integer. This scaling and rounding is simply uniform quantisation, which is described mathematically as,

\[
\hat{Y}_{kl}(i,j) = \text{round}\left(\frac{Y_{kl}(i,j)}{QF}\right)
\]

where \(Y_{kl}(i,j)\) is the \((i,j)\)th pixel in the \((k,l)\)th subband, \(\hat{Y}_{kl}(i,j)\) is the quantised version, and \(QF\) is a quantisation factor. The scaling and rounding is simply uniform quantisation. As such, the minimum mean square error bit allocation given in Chapter 3, assuming the synthesis filters are of unit energy, implicitly suggests a \(QF\) that is fixed for all subbands. Therefore, since the results in this thesis are usually evaluated in an objective manner, a constant scaling parameter is employed. Note that for biorthogonal subband schemes the analysis filters should be scaled so that the synthesis filters, or effective synthesis filters in a tree-structured approach, are of unit energy.
So far this is the same as the JPEG quantisation method described previously. These resulting coefficients are grouped into blocks which are sorted into a one-dimensional data stream using a zigzag scan through the subbands. Each block is delineated spatially. The details of this process are described later.

Each block of coefficients is represented using zero run-length, amplitude and end of block (EOB) symbols in the same manner as the JPEG method. The amplitude streams from each block are concatenated into a super amplitude stream, and similarly the runlength streams are concatenated into a super runlength stream. Each stream is then entropy encoded.

5.3.2.1. Zigzag Scan and Blocking Procedure for Arbitrary Subband Analysis

The only difference between this general method, and that described previously for the JPEG method is the grouping and scanning of the subband pixels. Nevertheless the basic principle is the same. There are two principles which are stated as,

1. Group spatially corresponding subband pixels

2. Sort the pixels within each group into a one-dimensional vector according to a zigzag path through the subbands

Since the groups in Principle 1. are delineated spatially they are referred to as spatial blocks. The blocking and scanning method for analysis schemes with the same (ideal) subband structure is the same.

Consider an $M$-band analysis of an image. There are $M^2$ subbands that are of roughly equal size. Each pixel in a subband corresponds roughly to a spatial region in the image of size $M \times M$ pixels. This correspondence is dependent on the synthesis basis functions, but is present nonetheless. There is a group of $M \times M$ pixels, with one pixel from each subband, that contain most of the information about this $M \times M$ spatial block. The critically sampled nature of the subband analysis means that there is this one-to-one block correspondence. These $M \times M$ spatial blocks are then scanned in a zigzag manner. For example consider a 4-band image analysis as depicted in Figure 5.12.
The first spatial block of pixels is formed by scanning through the subbands, as indicated in this figure, collecting the top left hand pixel from each subband. The next spatial block is formed similarly by collecting the pixel adjacent to the top left hand one in each subband and so on. In mathematical terms the analysed image pixels, the $Y_{k,l}(i,j)$, are grouped according to $(i,j)$, and each group is sorted in a zigzag manner through the subband indices $(k,l)$. Note that if an $MxM$ DCT was used as the image analysis, this would be the exact method described previously for the JPEG method.

In the case of general subband analysis each pixel in the smallest subband (that with the smallest bandwidth or largest decimation factor) corresponds to a spatial region in the original image whose size is roughly proportional to the corresponding decimation factor. Each spatial block of pixels therefore must correspond to a region in the original image of at least this size. For most subband analysis/synthesis methods the subband decimation factors are roughly inversely proportional to the subband bandwidths.

Assuming that the decimation factor of this smallest subband is an integer multiple of the other decimation factors, the number of spatial blocks is given by the number of pixels in the smallest subband. Each spatial block contains one pixel from the smallest subband and spatially corresponding pixels in other subbands sorted in a zigzag manner, commensurate with that used in the JPEG method. For example if a subband is twice as large in both the vertical and horizontal spatial frequency directions as the smallest subband, it contributes $2\times2$ pixels to each zigzag scan block. The ordering of these four pixels in the scan block has been done using a column stacking approach since this is simple to implement, and other methods are not believed to offer a significant improvement.
To illustrate the spatial blocks and the zigzag scan procedure consider Figure 5.13 showing the zigzag scan path for the DWT and modified DWT. The solid dots in this figure refer to subband pixels, which are delineated spatially within the subbands. Each scan block begins with one pixel in the DC subband (the top left hand subband) and collects spatially corresponding pixels in the other subbands by following the line indicated in this figure.

![Figure 5.13. (a) DWT zigzag scan path, (b) modified DWT scan path](image)

For the DWT, Figure 5.13a, the smallest subbands (the four top left hand subbands) contribute one pixel each, while the other subbands contribute four or sixteen pixels depending on their size. The same method applies to the modified DWT shown in Figure 5.13b. In this case, in some subbands rectangular groups of pixels are collected in the scan path reflecting the rectangular shape of those subbands.

In the $M$-band, DWT and modified DWT analysis methods the subband bandwidths are all integer multiples of the smallest subband. Hence each subband contributes the relevant integer multiple of pixels to each spatial block. In a more general case it may be necessary to group pixels from the smallest subband so that integer multiples of pixels in other subbands correspond to the same spatial block. When the number of pixels in the smallest subband required to do this is large, this method will not be particularly effective and some other method may be required. In this case a relaxed spatial grouping may be appropriate.
5.3.2.2. Variations on the Quantisation and Encoding Method

The aim of the zigzag scan is to maximise the length of zero runs and hence minimise the entropy of the run-length stream [Chen and Pratt 1984, p226]. This motivation suggests two variations of the quantisation and encoding method that are discussed in this subsection. The first variation discussed is an alternative zigzag scan procedure while the second is a simple variation of the quantisation method.

In order to maximise zero run-lengths an alternative scanning procedure has been investigated whereby the subbands are scanned following a path of decreasing subband variance. The same spatial blocks as the zigzag method are used (principle 1.), but the pixels within each spatial block are sorted in terms of decreasing subband variance. Consider an $S=3$ two-dimensional DWT, as shown in Figure 5.14.

![Figure 5.14. $S=3$ DWT two-dimensional analysis structure and variance scan path](image)

Using the 2DG model, correlation $\rho=0.95$, and Daubechies 12-tap filters, the variance of these subbands is given in Figure 5.15.
The subband labelling in Figure 5.15 extends the labelling given in Figure 5.14. Figure 5.15 illustrates that the subband variances are increasing according to the given labelling. It is worthwhile noting that the variance of every subband at a given level in the analysis tree (or resolution) is always greater than that of a subband at a lower level in the tree. Further, at each level the edge subbands have a higher variance than the diagonal subband. When the horizontal and vertical correlation are equal, as assumed in the 2DG model, the edge subband variances will be the same, as in Figure 5.15.

Based on this model a modified scan is proposed. This scan path, termed variance scan, is illustrated in Figure 5.14. The zigzag and variance scan method are compared in the Section 6.3.5 of Chapter 6.

Gharavi and Tabatabai (1988) observe that picture or camera noise manifests itself as a low level signal in the high frequency subbands of an image subband decomposition. Obviously it is not desirable to waste quantisation bits encoding this noise. In the design of subband quantisers for the high frequency subbands it is commonplace to employ quantisers with a dead zone around zero to eliminate such noise. The quantisation method suggested previously, of scaling and rounding, is easily adapted to include a dead zone. One method is to threshold the subband pixels before scaling and rounding. A similar method is suggested by Chen and Pratt (1984) for a DCT codec.
This can be achieved simply by truncating any subband pixels to zero that are less than a predetermined threshold following the scaling and rounding. In the results presented in Chapter 6, this technique is evaluated by truncating all quantised subband pixels with unit magnitude to zero before the entropy encoding process.

The quantisation technique could be improved with the use of more sophisticated quantisation methods. Nevertheless this method is very simple to implement, and is surprisingly effective, as the results demonstrate later. It is interesting to note that in the high rate encoding of memoryless Gaussian sources, the uniform quantiser has a codeword entropy that comes within 0.255 bits of the rate distortion bound [Jayant and Noll 1984, p155].

5.3.2.3. Entropy Rate Estimation versus Arithmetic Coding

The quantised subband data is represented as two streams; one of run-length symbols and the other of amplitude symbols, which are encoded separately. In order to calculate the number of bits required to transmit each stream of symbols the first order entropy of the symbols is calculated. The average rate for a stream is thus,

$$ R_{\text{stream}} = -\sum_i p_i \log_2 p_i $$

where $p_i$ is the probability of the $i^{th}$ symbol in the stream. The probabilities are determined by,

$$ p_i = \frac{N_i}{N} $$

where $N$ is the length of the stream and $N_i$ is the number of occurrences of the $i^{th}$ symbol.

Denoting the amplitude stream rate and length as $R_a$ and $N_a$ respectively, and similarly the run-length stream rate and length as $R_r$ and $N_r$ respectively, the total number of bits required to transmit the encoded image is,

$$ \text{bits} = N_a R_a + N_r R_r $$
In a practical codec the streams could be coded using an arithmetic or Huffman coder, as with the JPEG standard. The motivation for estimating the rate in terms of entropy is several fold. First an arithmetic coder can approach the bound determined by the data entropy. In fact an adaptive arithmetic coder, as described by Witten et al (1987), can surpass the first order "entropy" estimated using symbol frequencies by adapting to local changes in data. Secondly an arithmetic (or Huffman) coder generally requires initialisation or prior knowledge of symbol frequencies. The performance will be sub-optimum if the symbol frequencies are not initialised correctly. The coder described by Witten et al (1987) generally needs no initialisation since it is adaptive. For long sequences of symbols the inefficiency of the coder at the beginning of the sequence is relatively insignificant. However, for shorter sequences this is not necessarily the case, and initialisation of symbol frequencies is desirable. In terms of comparing subband schemes, the entropy technique obviates any problems of initialisation. This rate estimation by entropy calculation is often used. For example, much of the work in entropy constrained VQ uses this approach (see Chou et al, 1989).

To substantiate the claim that an arithmetic coder can perform near the entropy bound an arithmetic coder as described by Witten et al (1987) was implemented. The file size generated by the arithmetic coder with the run-length and amplitude streams as inputs (encoded separately) was compared to the file size calculated using the stream entropies. An experiment was performed where an image was encoded using an 8x8 DCT, and the generic subband quantised and encoding method. The size of the file generated by arithmetically encoding this data and the size estimated from the stream entropies is shown in Table 5.1 for various compression ratios.

| (a) Arithmetic Coded File Size (bytes x 1024) | (b) Entropy Estimated File Size (bytes x 1024) | % Difference $\frac{|b-a|}{b} \times 100$ |
|---------------------------------------------|---------------------------------------------|---------------------------------|
| 37.82                                       | 37.23                                       | 1.6%                            |
| 32.60                                       | 32.05                                       | 1.7%                            |
| 23.48                                       | 22.98                                       | 2.2%                            |
| 12.35                                       | 11.89                                       | 3.9%                            |

Table 5.1. File size for Arithmetic coder as compared to entropy estimated file size. 8x8 DCT subband image coder. Note original image file size was 256 (bytes x 1024)
Table 5.1 illustrates that the arithmetic coder achieves a rate very close to that predicted using the entropy estimate. If the arithmetic coder was initialised the discrepancy would be smaller. As the file size decreases, or stream lengths decrease, the discrepancy increases. However it is believed that proper initialisation would alleviate this increased discrepancy to a large extent.

5.4. SUMMARY

In this chapter three distinguishing characteristics of subband analysis structures were considered: namely, the ideal subband structure that is approximated by the analysis, the degree of this approximation, and the time resolution properties of the subbands. The power spectral density of a typical image was investigated using a correlation model of image data. A two-dimensional AR(1) and a generalised two-dimensional model (2DG) correlation model were considered. The characteristics of various subband image structures were discussed in light of these models. The DWT was shown to exhibit a high coding gain while attaining good time localisation properties. A modified two-dimensional DWT analysis structure was proposed that exhibits a coding gain over the DWT structure, while maintaining good time localisation properties. The two-dimensional AR(1) model predicts that this modified structure offers a large increase in coding gain over the DWT, while the 2DG model, generally a more accurate model, predicts only a small improvement. A quadic analysis structure, based on a 4-band one-dimensional analysis and a tree-structure analysis of the low band was also considered.

In Section 5.3 a generic subband image compression scheme, suitable for any subband analysis structure was proposed. This scheme is essentially a generalisation of that proposed by Chen and Pratt (1984) as used in the JPEG baseline sequential coder [Wallace 1991]. Quantisation is effected on the subband signals by scaling and rounding. The subsequent data is grouped into blocks which correspond to spatially distinct regions in the original image. These blocks are represented using zero run-length and amplitude symbols which are entropy encoded. The proposed scheme simply defines the method of grouping the subband pixels into spatial blocks. Various issues relating to this method were discussed.
The generic subband quantisation and encoding method provides a platform with which to compare different subband analysis structures and filters for subband image coding. This comparison is the main topic of Chapter 6.
CHAPTER 6:

SUBBAND CODEC RESULTS AND COMPARISONS

6.1. INTRODUCTION

One of the main themes of this thesis is the investigation of image "optimised" subband analysis structures and filters. By optimised it is meant that the subband structure, and filters used to implement the structure, are suited for a typical image. The aim of this chapter is to evaluate various subband analysis structures and filters using a practical still image compression technique. This is achieved by using the generic subband quantisation and encoding method described in Chapter 5. Not only are different methods evaluated, but in doing so the new theories and results given in Chapters 3, 4 and 5, which are derived from theoretical models, are tested on real images.

As discussed in the introduction of Chapter 5, the generic quantisation and encoding method has good spatial adaptation properties. This is in contrast to traditional subband coding methods that have a fixed bit allocation for each subband and hence are not spatially adaptive (or less so). Such an adaptive method has been selected so as to provide a fair comparison between different filter banks for image compression. Filter banks with good spatial resolution properties tend to confine edges, or regions associated with changing signal statistics, into a relatively narrow region within each subband. This confinement and separation of different regions can be exploited by a spatially adaptive quantiser. Quantisers that are not adaptive are unable to exploit this property so effectively. For filter banks with poor spatial resolution properties, spatial adaptation is less important, since edge boundaries, and the regions near the boundaries, are smeared over a larger spatial area. In the extreme case this smearing occurs across the whole subband, and spatial adaptation offers little, if any, improvement over static quantisation.
6.1.1. Preliminaries

The image codecs considered in this chapter consist of a subband image analysis followed by the quantisation and encoding method described in Chapter 5. The results are generally given as plots of peak signal to noise ratio (PSNR) versus compression ratio. Some observations of the subjective quality of the results are also given where appropriate. The PSNR is defined as,

$$\text{PSNR (dB)} = 10 \log_{10} \left( \frac{255^2}{\text{MSE}} \right)$$

where MSE is the mean square error between the reconstructed and original images: that is,

$$\text{MSE} = \frac{1}{N_x N_y} \sum_{i=0}^{N_x-1} \sum_{j=0}^{N_y-1} [x(i, j) - \hat{x}(i, j)]^2$$

where $N_x$ and $N_y$ are the number of columns and rows of the image respectively. This form of PSNR is commonly used in the image coding literature [Jain 1981].

The compression (ratio) is defined as the ratio of original image size to compressed file size. As discussed in Chapter 5, the compressed file size is estimated using the first order entropy of the quantised subband data symbols.

The test images are 512x512 pixels in spatial dimension with 8 bits per pixel grey scale. The image files baboon.g, jet, lena.g, peppers.g and urban were obtained from the ftp site eedsp.gatech.edu in the /database/images directory using an anonymous ftp login. The .g extension indicates that the image is actually the green field from an RGB image. The image files airplane, sailboat and tiffany were obtained from the ftp site ftp.ipl.rpi.edu in the /pub/image/still/usc/gray601 directory. These images are referred to by these names (without the .g reference). A head and shoulders image, of the same resolution, referred to as Newscaster is also used, which is available locally.

The word optimum is used to describe the best filter in terms of PSNR at a given compression level among a group of filters. There may be several filters that are optimum in this regard. In terms of analysis level, the smallest level that offers the best
results (PSNR) is usually referred to as the optimum analysis level, commensurate with the description used in Chapter 3.

6.1.2. Overview of Chapter 6

The generic subband and quantisation method, proposed in Chapter 5, was developed as a platform with which to compare various subband structures and filters for still image compression. As the results will attest, it is also an effective method for many subband structures.

The results for various $M$-band subband schemes are given in Section 6.2. The DCT, LOT and ELT are compared in terms of reconstructed image PSNR. The optimum analysis level ($M$) at different bit rates is evaluated for these transforms, in light of the rate constrained coding gain theory presented in Chapter 3. Finally, to illustrate the importance of selecting a good transform, the results for the DCT are compared against the Walsh-Hadamard transform (WHT) and the discrete sine transform (DST).

Section 6.3 investigates some DWT subband schemes. A large number of different filter characteristics such as length, phase, magnitude response, and time width are evaluated. As with the $M$-band transforms, the effect of varying the analysis level at various bit rates is considered. Some "source" optimised eigenfilters, introduced in Chapter 4, are compared. The effects of using different sources and filter lengths are considered. A DWT using the best eigenfilters is compared to a DWT using Daubechies filters, and to the $M$-band DCT. A comparison of the variations of the quantisation and encoding method, as discussed in Chapter 5, is made. Finally, a DWT using some biorthogonal filters is compared to a DWT using the best orthogonal filters.

In Section 6.4, the results for the other analysis structures presented in Chapter 5 are given. The modified DWT is compared to the DWT and $M$-band results. The results of the modified DWT in relation to different images is considered. Also, a quadic analysis structure using a 4-band MLT and a 4-band ELT is evaluated.

The results of this chapter are discussed in Section 6.5. Various issues relating to a subjective evaluation of the results are considered. Some results for 256x256 pixel resolution images are noted. The conclusions drawn from these results and the results of this chapter, are extrapolated to the case of HDTV resolution images. The results are
compared and contrasted against the related results of other authors. Finally, this
chapter is concluded in Section 6.6.

6.2. M-BAND ANALYSIS STRUCTURES

In this section the results for various $M$-band uniform subband codecs are presented
and discussed. The subband coders use a uniform $M$-band analysis followed by the
generic subband quantisation and encoding method. For subband schemes using the
same analysis level ($M$), exactly the same quantisation and encoding method is used.
The $M$-band transforms used are the DCT, LOT, ELT, WHT and DST. See Chapter 2
for a description of the exact transforms to which these abbreviations refer. The results
are presented in terms of PSNR versus compression ratio.

6.2.1. DCT Image Codec

Figure 6.1 illustrates the results for the DCT codec using the Lena image. There are
five curves corresponding to $M=4, 8, 16, 32$, and 64.

![DCT image codec results for Lena image](image)

Figure 6.1. DCT codec results for Lena image $M=4$ as indicated, $(--)$ $M=8,$
$(•••)$ $M=16,$ $(--)$ $M=32,$ and $(--)$ $M=64.$
From Figure 6.1 it is evident that the DCT 4x4 block size is sub-optimum at all levels of compression, although by a small margin at low compression. At low compression levels the \( M=8 \) and \( M=16 \) DCT's are optimum. At higher compression the \( M=8 \) DCT is sub-optimum. The performance of the \( M=32 \) and \( M=64 \) DCT improve as the compression increases, becoming optimum at high compression.

The inferiority of the \( M=32 \) and \( M=64 \) DCT at low compression is due to the nonstationary nature of the image data. For blocks that contain edges, the energy of the high frequency DCT coefficients is significant, making these blocks difficult to encode. Any extra gain afforded by a longer memory (larger block sizes) is offset by the fact that larger blocks are more likely to contain edges.

The differences between different block sizes is largely the same for other images. For the Jet image the \( M=32,64 \) block size DCT's are inferior by up to 3 dB at low compression, and are still inferior at high compression. On the other hand for the Baboon image and DCT of block sizes \( M=8,16,32, \) and 64, the results are more even than those for Lena. Nevertheless the \( M=8 \) DCT is still inferior at high compression.

The rate constrained coding gain theory presented in Chapter 3 predicts that as the compression increases the optimum level of analysis increases. This is evident in Figure 6.1. For example the performance of the \( M=4 \) DCT decreases with respect to the optimum performance as the compression increases. The same is true of the \( M=8 \) DCT although to a lesser extent. Even the \( M=16 \) DCT is surpassed at high compression by the \( M=32 \) DCT, albeit by a small margin.

In order to verify the performance of the subband codecs, the \( M=8 \) DCT codec was compared to the JPEG standard. (The JPEG method uses an \( M=8 \) DCT). Table 6.1 lists the reconstructed PSNR for the \( M=8 \) DCT codec and the baseline JPEG codec for Lena at various compression ratios.
Table 6.1. PSNR for JPEG and DCT Coders at various compression ratios

<table>
<thead>
<tr>
<th>Compression ratio</th>
<th>JPEG PSNR (dB)</th>
<th>DCT coder PSNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.7</td>
<td>33.0</td>
<td>33.6</td>
</tr>
<tr>
<td>12.9</td>
<td>33.8</td>
<td>34.5</td>
</tr>
<tr>
<td>11.0</td>
<td>34.4</td>
<td>35.2</td>
</tr>
<tr>
<td>9.5</td>
<td>35.0</td>
<td>35.8</td>
</tr>
<tr>
<td>7.7</td>
<td>35.9</td>
<td>36.7</td>
</tr>
<tr>
<td>5.9</td>
<td>37.0</td>
<td>38.1</td>
</tr>
</tbody>
</table>

This table indicates that the DCT codec performs in a similar manner to the JPEG standard as expected. The slightly lower PSNR values for the JPEG coder are probably due to its use of sub-optimum variable-length and Huffman codes as compared to the entropy estimate. Also the JPEG method uses slightly different run-length and amplitude symbol information and includes some header information overhead.

6.2.2. LOT, ELT and DCT Image Codec Comparison

The conclusions drawn from the DCT results apply roughly to the results for the ELT, as illustrated in Figure 6.2. The inferiority of the large block size ELT is more evident, especially at low compression. This is expected since the ELT basis filters are four times as long as the DCT basis filters, offering less time resolution. The performance of the $M=4$ ELT at low compression is near optimum, in contrast to that of the $M=4$ DCT. It appears that the improved frequency resolution of the ELT means that optimum performance is attained for smaller $M$ as compared to the DCT. On the other hand, for larger analysis levels ($M$) the ELT degrades more noticeably. Again this is commensurate with its better frequency resolution. The ELT has a better frequency resolution at the expense of the time resolution as compared to the DCT. The LOT results, lying between those of the DCT and ELT, corroborate these conclusions.
Figure 6.2. ELT codec results for Lena image $M=4$ as indicated, (- - -) $M=8$, (• • •) $M=16$, (- - -) $M=32$, and (—) $M=64$.

Overall the $M=16$ DCT, $M=16$ LOT and $M=16$ ELT are the best in terms of PSNR. This analysis level, $M=16$, is thus used in subsequent comparisons between different $M$-band subband schemes.

Figure 6.3 illustrates some results for the DCT, LOT and ELT using $M=16$. Note that in the case of the DCT, $M$ refers to the block size. The most interesting feature in Figure 6.3 is that the DCT, LOT and ELT perform in a similar manner. The same is generally true for all the other images. Using a stationary source model one would expect that the ELT would perform the best, followed by the LOT and then the DCT, commensurate with the frequency resolving capabilities of these transforms. However, it appears that any decreased subband resolution in the case of the DCT is compensated by improved spatial adaptation capabilities due to its shorter basis filter lengths. A closer inspection of these figures reveals that the DCT is slightly better at low compression, while the LOT and ELT improve comparatively as the compression increases. The slight improvement of the LOT and ELT at high compression can be attributed to the increased subband resolution of these transforms.
Kovacevic et al (1989) compared the $M=8$ DCT with various $M=8$ MLTs for image compression using a JPEG type method. It was concluded that the MLT gave inferior results to the DCT. However, non-zero constrained MLT's were used and it was noted that this was the likely cause of the poorer results. The MLT achieved results commensurate, and possible slightly better, than the DCT when the DC component of each image block was subtracted prior to the transform and transmitted separately. These results are then commensurate with those presented here.

Malvar (1992, pp262-274) compared the DCT, LOT, MLT and ELT using $M=8$ for image compression. The Chen Smith (1977) method was used to quantise the subbands. It was concluded that the LOT outperformed the DCT by 0.4 dB on average, and that similar results were obtained with the MLT. Similar observations were made by Cassereau et al (1989) and Malvar (1989). In contrast the results presented here suggest that the DCT and LOT performance is more even. It is worth noting that 0.4 dB is not a particularly large increase in any case.

The Chen Smith quantisation method, used by Malvar, was limited to four classes, limiting the spatial adaptability of this quantisation method. The JPEG type method on the other hand, which is used here, allows a greater spatial adaptability, since each transform coefficient is quantised according to its magnitude only. Since the DCT has
better spatial (time) resolution than the LOT, it is expected that the performance of the DCT will improve in comparison to the LOT when adaptive quantisation schemes are employed rather than non-adaptive (or less adaptive) schemes. This in part explains the slight discrepancy between the results of Malvar and those presented here. Note that Cassereau et al (1989) used a similar partially adaptive technique to Malvar (1992).

Malvar's (1992) results showed that the objective performance of the ELT was slightly inferior to that of the LOT. Subjectively the LOT exhibited less blocking and slightly more ringing than the DCT, while the ELT exhibited a large amount of ringing. A similar observation is made here. Figure 6.4 illustrates the original and reconstructed images for an $M=16$ DCT, $M=16$ LOT and $M=16$ ELT codec operating at approximately 0.15 bpp using Lena. These images have been resampled for use with a display device that assumes a rectangular sampling grid (4:3 width to height ratio). This resampling has not affected any of the visible reconstructed image artefacts.

The DCT coded image exhibits the characteristic blocking and ringing or edge smearing distortions. The LOT image exhibits fainter blocking artefacts, but a significant increase in edge distortion. The edge distortion, or ringing, can be seen to propagate a significant distance away from the hat, and the edge of the mirror. The ELT image exhibits even more edge distortion. In some areas the ringing can be seen to propagate about twice the distance of that associated with the LOT. An appropriately weighted subband quantisation factor would improve the results from a subjective perspective. For example, the blocking artefacts could be reduced in the DCT image. However, since all codecs will improve to some extent, it is not believed that subjective quantisation will affect this comparison between these codecs to a large extent.
Figure 6.4. $M$-band codec results for Lena. (a) Original 8 bpp (b) $M=16$ DCT, 0.1515 bpp, PSNR = 28.86 dB (c) $M=16$ LOT, 0.1493 bpp, PSNR = 28.72 dB (d) $M=16$ ELT, 0.1497 bpp, PSNR = 28.51 dB.

6.2.3. DCT WHT and DST Image Codecs

From the preceding results it is tempting to conjecture that any subband scheme will perform in a similar manner. However, this is definitely not the case. Figure 6.5 illustrates some results for the DCT ($M=16$), DST ($M=64$) and WHT ($M=16$). The $M=64$ DST is shown since it is superior to the smaller block size DST's. As $M$ becomes
large the DST, DCT and other asymptotically optimum transforms tend to the KLT [Yemini and Pearl 1979]. Hence it is expected that as the block size increases the DST performance will approach that of the DCT. This explains in part the improved performance of the \( M=64 \) DST over smaller block size DST's.

Figure 6.5. Results for \( M=16 \) DCT, \( M=64 \) DST and \( M=16 \) WHT codec using Lena.

The WHT and especially the DST are quite sub-optimum as compared to the DCT. This sub-optimum performance applies to all the eight images. At low compression the difference is about 3 dB, increasing to about 5 dB at high compression.

As suggested by Clarke (1983a), a significant factor in the inferior performance of the DST is that it is not a zero-constrained transform. That is some basis filters, other than the DC filter, do not have a zero at DC. This results in significant leakage into the corresponding subband when encoding DC (or very low frequency) dominant sources. The results using an \( M=16 \) "zero-constrained DST", a zero-constrained KLT as \( \rho \) tends to zero for an AR(1) source, are observed to be less than 2 dB down on those of the \( M=16 \) DCT at all compression levels.
6.3. DWT SUBBAND CODEC RESULTS

In this section the results for various DWT codecs are presented and discussed. These codecs employ a DWT subband analysis followed by the generic subband quantisation and encoding method. The results are used to compare different filters and analysis levels. In particular the subsequent subsections address the following issues:

(i) The effect of zero-constrained filters as opposed to non-zero constrained filters
(ii) The optimum analysis level at various compression ratios
(iii) The effect of different filter phase for a given magnitude response
(iv) Comparison of "source" optimised eigenfilters, Daubechies filters, and the DCT
(v) Evaluation of variations on the quantisation and encoding method
(vi) Comparison of orthogonal and biorthogonal filters

6.3.1. Zero-constrained Filters

As mentioned in Section 2.4.2 of Chapter 2, for subband image coding purposes the highpass filter(s) should have a zero at DC, giving a zero-constrained filter bank. Kronander (1989a, 1989b) gave two reasons why this is so. The first is that zero-constraint is required for maximum coding efficiency (or coding gain). This argument has also been given by Clarke (1983a) and Caglar et al (1991). The second reason is that the large DC component of an image is aliased into a sinusoid at half the sampling frequency. In the presence of quantisation noise this aliasing is not perfectly cancelled, and unless the filter bank is zero-constrained this can cause an annoying ripple in the reconstructed image.

In order to verify this reasoning, a zero-constrained and a non-zero constrained eigenfilter were compared. Both filters were optimised for an AR(1) source of correlation, ρ=0.95. Figure 6.6 illustrates the results for a subband image codec using these filters for the Jet image.
Figure 6.6. Zero-constrained and non-zero constrained filters. Jet image.

At low compression the zero-constrained filter set outperforms the non zero-constrained filter set by about 0.8 dB. As the compression increases the difference reduces to about 0.2 dB. Similar results are obtained for Airplane, Tiffany and Urban. However for these other images the difference at high compression is very small. For Baboon, Lena, Peppers and Sailboat the filter sets perform nearly identically.

Although the difference does not appear to be large, it is important to note that the non zero-constrained filter set has been optimised for an AR(1) source of high correlation. This means that there will be significant attenuation near DC, even if there is not a zero at this point. For filters optimised for a flat PSD Rioul (1993) noted that there is up to 5 dB difference between filters with and without the zero DC constraint.

6.3.2. Optimum Analysis Level and Optimum Filter Length

Figure 6.7 illustrates the PSNR versus compression ratio results for the Lena image using a DWT image codec with different analysis levels (tree-depths, S). The filters used are 12-tap minimum time width Daubechies filters.
The PSNR increases at all levels of compression as $S$ increases from 1 to 3. At low compression, no improvement is made using an analysis level greater than $S=3$. Indeed, there is only a relatively small improvement attained using $S=3$ as opposed to $S=2$. At high compression the performance improves more dramatically as $S$ increases to $S=4$. This characteristic, whereby a larger analysis level ($S$) is desirable as the compression increases, is commensurate with the rate constrained coding theory presented in Chapter 3. A similar characteristic was noted for the DCT (and other $M$-band transforms) in the $M$-band results section.

The differencing of adjacent DC subband pixels, implemented by the generic quantisation and encoding method, is effectively a further analysis of the DC subband (DPCM in particular). If this differencing were absent, the PSNR would increase for higher $S$ than indicated above for all compression ratios.

At all compression levels shown in Figure 6.7, there is little gain achieved using a larger analysis level than $S=4$. However, it is worth noting that at any compression ratio, there is no performance degradation associated with increasing the analysis level. This is in contrast to the $M$-band transforms, whose performance at low compression can be quite sub-optimum if the analysis level $M$ is too large. These conclusions are generally applicable to all the other images and to DWT's using different length filters.
However, for filters that are short and "optimised" using a flat PSD model, the gain may increase for $S>4$ at high compression.

Since $S=4$ is the lowest analysis level near optimum at all levels of compression, it is generally used in subsequent comparisons. In order to estimate the optimum filter length a similar comparison to that above has been made. Using an $S=4$ DWT the results for different length minimum time width Daubechies filters are shown in Figure 6.8.

\[ \text{Figure 6.8. } S=4 \text{ DWT codec results for Lena using minimum time width Daubechies filters. 4-tap as indicated, (---) 6-tap, (—) 8-tap, (•••) 12-tap, and (---) 14-tap.} \]

Figure 6.8 illustrates that for a fixed compression ratio the PSNR increases as the filter lengths increase, up to a length of about 8-taps. Beyond this point there is very little improvement. In fact for longer filters the PSNR can even decrease. At higher analysis levels similar results are obtained. For some of the other images, the PSNR increases for filters up to about 12-taps. For the Jet image the 4-tap filter gives results equal to all the other length filters. These observations corroborate those of Rioul (1993) who concluded that 12-tap Daubechies filters are optimum in terms of length and notes that longer filters can give sub-optimum results.

Most of the results presented in this chapter use 12-tap filters since this is around the optimum length for all images. However it is worth noting that the 12-tap filters only
offer a small improvement over 8 or 10-tap filters for some images. The advantage of the shorter filters is a reduced computational cost required to analyse an image.

6.3.3. Effect of Filter Time Width or Phase

As discussed in Chapter 4, there are certain freedoms in selecting the phase response of the filters used in the two-band perfect reconstruction filter bank. Figure 6.9 illustrates the PSNR versus compression for the Peppers image using an $S=4$ DWT image analysis and Daubechies 12-tap filters. There are four curves which correspond to the four different wavelet filters with the same frequency response. Strictly speaking there are eight different filters. However, for every filter there is a corresponding time-reversed filter. Only one orientation per pair is shown since time-reversal has a negligible effect on the coding performance.

![Figure 6.9. $S=4$ DWT Peppers image results: Daubechies 12-tap filters.](image)

The different phase filters are labelled phase a, phase b, phase c and phase d. Phase a corresponds to the minimum (or maximum via time reversal) phase implementation, while phase c corresponds to the minimum time width impulse response. The point to note in Figure 6.9 is that the minimum phase (a), performs the worst while the minimum time width (phase c) performs the best. The Peppers image displays the
largest difference between filters. Nevertheless for the other images the same hierarchy of results is obtained.

Table 6.2 gives the impulse response time width and the group delay deviation (in the passband) for the four different filters.

<table>
<thead>
<tr>
<th>Daubechies 12-tap Filter: phase</th>
<th>time width, $\sigma_t^2$</th>
<th>Passband group delay deviation (samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.9577</td>
<td>1.88</td>
</tr>
<tr>
<td>b</td>
<td>0.5941</td>
<td>0.87</td>
</tr>
<tr>
<td>c</td>
<td>0.5316</td>
<td>0.49</td>
</tr>
<tr>
<td>d</td>
<td>0.8288</td>
<td>1.53</td>
</tr>
</tbody>
</table>

Table 6.2. Time width and group delay deviation (passband) for Daubechies 12 tap filters

The passband group delay deviation is a measure of the deviation from linear phase of the frequency response of a filter. It is interesting to note that the hierarchy of time widths is the same as that of the passband group delay deviation. Further, this hierarchy is the same as that of the filter performance for nearly all the images (minimum time width filters offering the best results). The only exception is the Urban image, which exhibits nearly identical results for each filter. Also worth noting is that the same hierarchy of results has been observed using various eigenfilters.

It is proposed that a minimum time width response is desirable for subband image coding filters for two reasons. First the results demonstrate that the minimum time width filters offer the best objective results (PSNR). Secondly the results are better from a subjective perspective. Although the step-edge rise time is slower, as compared to other phase responses, there is less associated ringing. Hence, as noted by Kronander (1989b), the reconstructed image suffers from less ringing or mosquito noise around edges. It is assumed henceforth that the filters used are minimum time width filters unless otherwise indicated.

The effect of phase is more dramatic for longer filters. The difference between minimum time width and the maximum time width filters can be well over 1 dB for longer filters. From a subjective perspective the difference between the minimum time
width and maximum time width 22-tap Daubechies filters is quite dramatic (Some results were presented in MCAT-93, see [Andrew et al 1993b]). Subjectively the former exhibits much less ringing and is, as a consequence, more visually pleasing. In Section 6.4 a reconstructed image using a DWT codec and maximum time width Daubechies 22-tap filters is shown in Figure 6.21 As discussed in this section, this reconstructed image exhibits a significant increase in ringing as compared to encoded images using smaller time width filters.

6.3.4. Source "Optimised" DWT Filters and the DCT

In this subsection different source models are evaluated for filter design purposes in terms of DWT image coding. The description, source model, is used fairly loosely. For example, as demonstrated in Chapter 4, minimising the classical average stopband attenuation can be considered as a maximum gain filter design problem using an appropriate source.

Firstly, eigenfilters optimised for different values of correlation $\rho$, for an AR(1) source, are compared. Figure 6.10 illustrates the results for a DWT of tree-depth $S=4$ using 8-tap eigenfilters for the Lena image. Note that the correlation $\rho=1.0$ refers to the limiting behaviour as $\rho$ tends to unity, although in practice a value of $\rho=0.9999$ has been used. The eigenfilters designed for a highly correlated source outperform those designed for a low correlated source by up to 6 dB. As discussed in Chapter 4, the highpass 8-tap eigenfilters generally will not have a zero at DC (unless $\rho$ tends to one). On the other hand, the highpass eigenfilters of length $N$ where $N/2$ is odd, will have a zero at DC. In this case the difference between eigenfilters designed for a high and low correlated source is much less than that indicated in Figure 6.10, with less than 2dB separating all the eigenfilters. This improvement is attributed to the large increase in attenuation at and near DC for the equivalent bandpass and highpass filters in the DWT. This observation further corroborates the proposition that a zero-constrained filter bank is desirable for image coding purposes.
The results using other images are similar, and suggest the same conclusion: using maximum gain eigenfilters, those designed for a highly correlated source are desirable. However, there is some variation in the difference between the eigenfilters designed for different correlation. For example, using the Baboon image, the difference is much less than for Lena. It was observed that zero constraint, although desirable, is not so critical for such images as Baboon.

The above results illustrate that filters designed using a "mismatched" source perform poorly. However, relatively short filters were used in this study. As discussed in Chapter 4 (Section 4.7), it is possible to design filters with a high coding gain for any correlation by using sufficiently long filters. A sufficiently long highpass eigenfilter will have a magnitude response that is a good approximation to the ideal (brick-wall) highpass filter, with bandwidth $\pi/2$ radians, for any (non-increasing) source. Using non zero-constrained eigenfilters designed for a low correlated source, the PSNR, for a given bit rate, increases for filter lengths increasing up to about 24-taps. The 24-tap filters offer results that are only between 0.5 to 1.0 dB down from those of the optimum 8-tap (or 12-tap) eigenfilters illustrated in Figure 6.10.

The results for zero-constrained eigenfilters optimised for a highly correlated source, and Daubechies filters are now compared. In Table 6.3 the eigenfilters are listed along
with their associated correlation coefficient for which they are optimised and the time width of the impulse response. The label wdl2 refers to 12-tap minimum time width Daubechies filters. The labels ez12p90, ez12p95, etc, refer to 12-tap minimum time width zero constrained eigenfilters designed for an AR(1) source of correlation $\rho = 0.90$, $\rho = 0.95$ etc.

<table>
<thead>
<tr>
<th>Filter</th>
<th>time width, $\sigma_t^2$</th>
<th>optimum correlation coefficient $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ez12p90</td>
<td>1.003</td>
<td>0.90</td>
</tr>
<tr>
<td>ez12p95</td>
<td>0.9953</td>
<td>0.95</td>
</tr>
<tr>
<td>ez12p98</td>
<td>0.9928</td>
<td>0.98</td>
</tr>
<tr>
<td>ez12p100</td>
<td>0.9923</td>
<td>(1.0) 0.9999</td>
</tr>
<tr>
<td>wdl2</td>
<td>0.5316</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3. Eigenfilter and Daubechies 12-tap filter time widths.

Figure 6.11 illustrates the results for an $S=4$ DWT using these 12-tap (minimum time width) eigenfilters and Daubechies 12-tap filters for the Lena image.

![Figure 6.11. S=4 DWT results for Lena using 12-tap Daubechies filters and 12-tap zero-constrained eigenfilters optimised for an AR(1) source of correlation $\rho = 0.90, 0.95, 1.0$](image)

Figure 6.11. $S=4$ DWT results for Lena using 12-tap Daubechies filters and 12-tap zero-constrained eigenfilters optimised for an AR(1) source of correlation $\rho = 0.90, 0.95, 1.0$
In Figure 6.11 no labelling is given since the curves are nearly identical. It is evident that for DWT image coding the different eigenfilters and the Daubechies filters perform in a very similar manner. This is not surprising, considering the coding gain evaluation given in Chapter 4, Section 4.7. The same is generally true for the other images and different analysis levels. In the case of the Jet and Peppers images, the Daubechies filters offer a very small improvement at high compression, especially using an analysis level of $S=6$. It is proposed that this improvement for these images is due to the smaller time width of Daubechies filters, as demonstrated in Table 6.3. Note that the Peppers (and Jet) image displayed the largest difference in results for Daubechies (or eigenfilters) with the same magnitude response and differing phase responses (or time widths). For other images such as Baboon the eigenfilters perform slightly better, suggesting that frequency resolution is more important for this image. Note that the eigenfilters have a better frequency resolution than Daubechies filters. However it is important to stress that for all images all the filters perform in a very similar manner.

As mentioned in Chapter 1, the (QMF) filters tabulated by Johnston (1980) have been used extensively for subband image coding. Johnston used an objective function that included a filter bank distortion measure and an average stopband attenuation measure. The filter bank distortion measure is required because these QMFs are not perfectly reconstructing. These filters have been compared against the 12-tap Daubechies filters using a DWT structure [Andrew et al. (1994)]. The 8-tap Johnston filters performed poorly. This is probably related to distortion introduced by the filter bank. The 32-tap Johnston filters offer slightly inferior results to Daubechies 12-tap filters, usually about 0.2 to 0.5 dB down. The 16-tap filters perform similarly to the 32-tap filters at high compression, while they are inferior at low compression. Again this characteristic is probably due to the distortion introduced by the filter bank.

In light of the previous discussion on filters designed for "mismatched" sources these results are not particularly surprising. Depending on the stopband edge frequency the average stopband attenuation measure is similar to the variance of the subband using a low correlated source. Hence Johnston filters, being optimised in such a manner, are expected to perform like eigenfilters designed for a low correlated source, noting that quite long filters are used. Johnston filters have an advantage in one regard, which is that the filters have a time width that is relatively small compared to the filter length.
Nevertheless, the time widths of the longer filters is significantly larger than that of the short eigenfilters designed for a highly correlated source.

A similar performance is expected, and observed, for the (CQF) filters designed by Smith and Barnwell (1986). Smith and Barnwell use a Chebyshev equiripple filter design method which is similar to using an average stopband attenuation objective function. In this case slightly better results are sometimes obtained since a relatively large transition bandwidth is allowed. This means that the corresponding source, using a subband variance interpretation of the average stopband attenuation, has a reasonable level of correlation. In [Andrew and Ogunbona 1991] it was observed that a high stopband attenuation, rather than a narrow transition bandwidth, is desirable for subband image coding in order to prevent leakage of dominant low-frequency energy into high frequency subbands. This is commensurate with the concept of subband filters optimised for a highly correlated source.

Finally the performance of the DCT is compared to that of the DWT. Figure 6.12 compares the results for a DWT codec using Daubechies 12-tap filters and a DCT codec for the Lena image. Also shown are the results for DWT codec using 8-tap Smith and Barnwell filters.

![Figure 6.12](image)

Figure 6.12. Subband image coder results for Lena. (—) $M=16$ DCT, (---) $S=4$ DWT using 12-tap Daubechies filters, (•••) $S=4$ DWT using 8-tap Smith and Barnwell CQFs.
The DWT using 12-tap Daubechies filters and the DCT perform in a very similar manner. This result is commensurate with those using other images, and the results given by Andrew et al (1993b) and Ohta et al (1992). For some images such as Jet, the DWT performs slightly better at high compression. As noted previously, the $M=64$ DCT can be sub-optimum at low compression. At high compression the $S=6$ DWT scheme, using 12-tap Daubechies wavelets, gives the best results. However, the improvement over the $M=16,64$ DCT and $S=4$ DWT scheme is only about 0.2 dB on average. Some reconstructed images using the DCT and DWT are presented and compared, from a subjective perspective, in Section 6.3.6.

Note that as expected, the performance of the 8-tap CQF's is inferior to Daubechies filters. This comparison generally holds for other images. However in the case of the Jet image the CQF's give PSNR values that are over 5dB down at low compression, and 2.5 dB at high compression. For Baboon on the other hand, the CQF's perform nearly identically to the DCT and Daubechies 12-tap filters at all compression levels.

6.3.5. Variations on the Quantisation and Encoding Method

The PSNR at various compression levels for two $S=4$ DWT coders using 12-tap Daubechies filters has been evaluated. The first coder uses the generic subband quantisation and encoding method, using a zigzag scan, while the second coder uses an identical method except that the subbands are scanned according the 2DG variance method. These different scanning methods were described in Section 5.3.2.2 in Chapter 5. The results are illustrated in Figure 6.13.
The two scanning methods perform in a very similar manner as illustrated in Figure 6.13. At low compression the methods are nearly identical. At high compression the 2DGVariance scan is superior by about 0.2 dB. However this difference is the largest for any of the eight other images at any compression level. At an analysis level of $S=6$ similar results are obtained.

From these similar results it is concluded that either the zigzag or 2DG variance scan is an effective technique. The method that lends itself to the simplest implementation is thus the recommended technique. The zigzag scan has been used in the other simulations presented in this chapter largely because it was the first method implemented.

The effectiveness of using a quantiser dead zone, as discussed in Chapter 5 (Section 5.3.2.2), is now considered. A modified quantisation method has been employed where all quantised subband pixels with unit magnitude are truncated to zero before the entropy encoding process. Otherwise the entire subband codec process is the same. Figure 6.14 illustrates the results for an $S=4$ DWT using the original quantisation method, and the same codec using the modified quantisation method, for the Peppers image.
Figure 6.14. $S=4$ DWT results for Peppers: (—) original quantisation method, (- - -) modified (dead zone) quantisation method

The results, in terms of differences between methods, are similar for other images. Also worth noting is that using this type of dead zone improves the results for other subband coders such as a DCT based codec. The incorporation of such a dead zone into the quantisation process offers an improvement of around 0.4dB at all levels of compression. This observation supports the proposition that the low magnitude pixels in the high frequency subbands represent picture noise. A reduction in bit rate, without significantly reducing the reconstructed image quality, can be achieved by eliminating this noisy component, which is difficult to encode efficiently. Depending on the quantisation factor ($QF$) or compression ratio, it may be advantageous to use a higher threshold than unity.

6.3.6. Linear Phase and Biorthogonal Wavelet Filters

As discussed in Chapter 4, an orthogonal PR two-band filter bank or DWT cannot have linear phase filters. If linear phase filters are required, it is necessary to employ a biorthogonal DWT. In this thesis generally only orthogonal wavelet filters have been considered. An open question is whether there are some biorthogonal filters that are better than the best orthogonal filters for DWT image coding. This issue is addressed
briefly in this section. The material presented suggests several avenues of future research into the design of "optimum" biorthogonal wavelet filters.

Orthogonality is generally believed to be a desirable attribute for subband image coding for several reasons. Not the least of which is that the maximum coding gain block transform is the orthogonal KLT. It is also believed that linear phase filters are desirable. As was pointed out in Chapter 4, some orthogonal filters have near linear phase. On the other hand, some PR biorthogonal filter banks with linear phase filters are near orthogonal. Simoncelli and Adelson [in Woods 1991, Chapter 4] concluded that near orthogonality is desirable for image coding. Another advantage of linear phase filters is that, as discussed in Section 2.4.6 of Chapter 2, for some filters an even periodic data extension, as opposed to a circular data extension, may be employed. Smith and Eddins (1990) demonstrated that the even periodic extension technique is slightly more efficient for coding purposes, both in an objective and subjective sense.

Le Gall and Tabatabai (1988) introduced some linear phase biorthogonal perfect reconstructing two-band filters known as symmetric short kernel filters (SSKF). These SSKF, denoted SSKF(53) to indicate that the lowpass is 5-taps and the highpass is 3-taps long, are very short and have linear phase (since they are symmetric). Further the highpass analysis filter coefficients follow a binomial expansion (ignoring signs) and hence it approximates a Gaussian, giving a near optimum time bandwidth product. Katto and Yashuda (1991) demonstrated that the coding gain of these filters using a DWT structure is similar to the DCT for an AR(1) source of high correlation. However this impressive performance is partly due to the fact that the highpass analysis filter approximates the optimum 2-sided predictor for an AR(1) source of high correlation. Therefore, for an AR(1) source, the highpass signal variance is (almost) minimised.

Ebrahimi and Kunt (1992) proposed some near perfect reconstruction biorthogonal linear phase filters that are designed to minimise the time bandwidth product. Further the filter coefficients are constrained to be powers of 2, so that an efficient computational implementation is possible. These filters are referred to as EK filters where the analysis filters are 10-taps long while the synthesis filters are 6-taps long.

Figure 6.15 illustrates the PSNR (dB) versus compression for an S=4 DWT image compression schemes based on Daubechies 12-tap filters, SSKF(53) and EK filters for the Urban image.
At low compression Daubechies filters offer a 0.4dB improvement over the SSKF(53) filters. However the converse is the case at high compression. The EK filters are sub-optimum at all compression. The results for other images are similar. Daubechies filters are optimum at low compression while SSKF(53) are optimum at high compression. Similar results are obtained for S=6 except that at high compression the superiority of the SSKF(53) filters is not so apparent.

Cheong et al (1992) investigated a family of SSKF's and compared them to some of Johnston's QMF's and some orthogonal wavelets. They concluded that the Le Gall and Tabatabai SSKF(53) were optimum or near optimum for all images. However, the MSE performance of the orthogonal wavelets was fairly close to the SSKF(53). It is worth noting that they used 34-tap Daubechies filters, whereas superior performance could have been obtained using shorter Daubechies filters, both in an objective and subjective sense. Cheong et al (1992) also noted that the shorter filters exhibited less ringing distortion.

The performance of the SSKF(53) filters is very impressive given their short length. Any lack of frequency resolution is compensated by their fine time resolution. The SSKF(53) filters outperform Daubechies filters at moderate to high compression for
images such as Jet where time resolution is important. The highpass analysis filter has 2 zeros at DC (this actually specifies the filter) so that the dominant low frequencies of typical image data are attenuated to reasonable degree even though the filter is only 3-taps long. The EK filters on the other hand do not perform so well. The filters have a small time width, being optimised in this respect, but the coding gain is not particularly high. It is proposed that this latter characteristic is the reason for the sub-optimum performance.

A possible improvement to the design method of Ebrahimi and Kunt is to minimise an inverse coding gain / time width product rather than the time / bandwidth product. In a rather ad hoc attempt to this end, a 6-tap highpass linear phase analysis filter has been designed. This filter has 3 zeros at DC which provides high attenuation of low frequencies and allows a small filter time width. The two other zeros are selected to be at $-5$ and $-1/5$ so that the bandwidth of the filter is roughly $\pi/2$ radians. This means that a corresponding two-band biorthogonal filter bank can be constructed that is "near orthogonal". The lowpass analysis filter is selected so as to form a perfect reconstruction filter bank, and have at least two zeros at $\pi$ radians, for some measure of regularity (see Rioul (1993)).

This filter set is referred to as lin6, and the filter coefficients are listed in Appendix D. Since the filters are of even length and linear phase, an even periodic extension may be employed in the two-band filter bank. The time width of the lowpass filter is $\sigma_1^2 = 0.4066$, while that of the highpass filter is $\sigma_2^2 = 0.3617$, which are both considerably less than the time widths of 12-tap eigenfilters, or 12-tap Daubechies filters. These latter time widths were given in Table 6.3.

Antonini et al (1993) compared three different biorthogonal wavelet filter sets for image compression. They concluded that the best filter set was a "most even length" spline variant, the coefficients of which are listed in Table II in [Antonini et al 1993, p209]. This filter set is referred to as anton7. It is worth noting that the coding gain of the anton7 and lin6 filters in a DWT (tree-depth of $S=4,5,6$) using an AR(1) source of correlation $\rho=0.95$, is nearly identical to that of a 12-tap eigenfilter optimum for this AR(1) source. The time width of the highpass analysis lin6 filter is slightly lower than that of the anton7 filter set, while that of the 12-tap Daubechies filter is considerably larger.
A biorthogonal DWT, and the generic quantisation and encoding method, has been used to evaluate the performance of the anton7 and lin6 filter sets. Figure 6.16 illustrates the results for an $S=4$ DWT using the anton7, lin6 and Daubechies 12-tap filters. At low compression the filters perform in a similar manner (the lin6 filters offer about 0.15dB advantage). However, at high compression the lin6 filters offer a 0.5dB improvement over the anton7 filters which in turn offer a 0.5dB improvement over the Daubechies filters. The Peppers image exhibits one of the largest differences among the images tested. Nevertheless the lin6 filters consistently exhibit a 0.5dB improvement over the 12-tap Daubechies filters at high compression, with a somewhat smaller improvement at low compression. This impressive performance is attributed to the high coding gain and low time widths of the lin6 filters.

A circular periodic extension was used for the anton7 filters, while an even periodic extension was used for the lin6 filters. Using the lin6 filters and a circular periodic extension it has been observed that the results are very similar to the anton7 filters. This suggests that the advantage of the lin6 filters over the anton7 filters is largely due to the ability to use an even periodic extension.

![Graph](image)

Figure 6.16. $S=4$ DWT results for the Peppers image. (—) anton7 filters, (---) 12-tap Daubechies filters, (•••) lin6 filters

The results using the lin6 filters are also impressive from a subjective perspective. As illustrated later, there is significantly less ringing in the reconstructed images than that
associated with 12-tap Daubechies filters. This follows from the minimal ringing
associated with the step response of the lowpass synthesis filter: a fact no doubt related
to its small time width. Some subjective results have been presented in Andrew et al
(1993), and more will be presented in Andrew et al (1994).

Figure 6.17 illustrates the reconstructed (Lena) images for a subband coder using an
$M=16$ DCT, $S=4$ DWT using Daubechies 12-tap filters, an $S=4$ DWT using ez12p98
filters, and an $S=4$ DWT using the linear phase lin6 filters. The compressed image bit
rate is approximately 0.14 bpp. The original image can be observed in Figure 6.4. As
before these images have been resampled for a display device that assumes a
rectangular sampling grid. The DCT image exhibits blocking and ringing distortions.
The DWT images are free from any blocking, but there is still some edge distortion.

A close inspection of this figure reveals that the image encoded using the lin6 filters
exhibits less ringing than that encoded using Daubechies 12-tap filters, which in turn
exhibits less ringing than the image encoded using the 12-tap eigenfilters (ez12p98). As
suggested previously, this is commensurate with the hierarchy of filter time widths.
The lin6 filters have the smallest time width, and exhibit less step-edge ringing than the
other filters. In the authors opinion all these DWT schemes exhibit less edge distortion
than the DCT scheme. This is in agreement with the observations of Ohta et al (1992),
who observed that the DWT compressed images, using orthogonal filters, exhibited
less mosquito (ringing) noise and blocking as compared to the DCT compressed
images.
Figure 6.17. Subband coded Lena images.  (a) $M=16$ DCT, 0.1410 bpp, PSNR = 28.54 dB.  (b) $S=4$ DWT using 12-tap Daubechies filters 0.1408 bpp, PSNR = 28.56 dB.  (c) $S=4$ DWT using ez12p98 filters, 0.1467 bpp, PSNR = 28.51 dB.  (d) $S=4$ DWT using lin6 filters, 0.1408 bpp, PSNR = 29.31 dB

6.4. OTHER TREE-STRUCTURED ANALYSIS SCHEMES

In this section the results for various subband structures, discussed in Chapter 5, are compared to the $M$-band and DWT analysis structures.
Figure 6.18 illustrates some results for the $M=16$ DCT, $S=4$ DWT and $S=4$ modified DWT codecs. At $S=4$ the low frequency DWT (and modified DWT) subband occupies roughly the same band in the frequency plane as does the DC subband in the DCT of block size $M=16$. This correspondence holds similarly for $S=6$ and $M=64$.

Figure 6.18. Results for Airplane and Lena. (—) $M=16$ DCT, ( — — ) $S=4$ DWT, and (•••) $S=4$ modified DWT.

The most prominent feature in Figure 6.18 is that the DCT, DWT and modified DWT subband schemes perform in a similar manner. The DWT and modified DWT schemes employ 12-tap minimum time width Daubechies filters. Figure 6.19 illustrates the results for the $S=4$ DWT and $S=4$ modified DWT analysis structures for the Newscaster (head and shoulders) image.
Figure 6.19. Results for Newscaster: (—) $S=4$ DWT, and (---) $S=4$ modified DWT

For head and shoulders images such as Newscaster, the modified DWT structure offers an improvement of 0.5-0.8 dB over the DWT. For other types of images the improvement is generally smaller. It is interesting to note that the 2DG model suggests a gain of around 0.3-0.4 dB as appears to be the case on average (see Chapter 5, Section 5.2). The modified DWT structure offers more improvement at higher compression, although this is somewhat image dependent. As with the DWT scheme the $S=6$ modified DWT method offers a small improvement of about 0.2-0.4 dB at high compression as compared to the $S=4$ method. Hence the modified DWT analysis method offers a small improvement over the DWT scheme at all levels of analysis.

Figure 6.20 illustrates the results for Lena using the standard $S=4$ DWT coder and the quadic coder using an MLT and ELT.
Figure 6.20. Lena results. (—) $S=4$ DWT, (— — —) $S_q=2$ MLT based quadic analysis, and (• • •) $S_q=2$ ELT based quadic analysis.

Figure 6.20 indicates that the quadic ELT performs in a similar manner to the standard DWT method while the quadic MLT is slightly inferior at high compression. For images where spatial resolution is more important, the MLT quadic scheme outperforms the ELT quadic scheme at low compression, while the converse is the case for all images at high compression. Overall the quadic schemes are slightly inferior to the optimum DWT scheme. At $S=4$ for the DWT scheme, and $S_q=2$ for the quadic scheme, the DC subbands have (roughly) the same bandwidth. (Similarly for $S=6$, $S_q=3$.) The quadic ELT at $S_q=3$ can be quite inferior as compared to the DWT and MLT scheme at low compression, suggesting that spatial resolution does affect the efficiency of quantisation in the low frequency subbands to some extent.

Figure 6.21 shows the original Newscaster image and reconstructed images using two DWT coders and a modified DWT coder. There first DWT coder uses minimum time width Daubechies 12-tap filters, while second uses maximum time width Daubechies 22-tap filters. The modified DWT coder uses minimum time width Daubechies 12-tap filters. The DWT using the maximum time width Daubechies 22-tap filters exhibits a large amount edge distortion or ringing. These filters were selected for their large time width, to illustrate the relation between ringing and filter time width. Similar observations are made with the other images.
The modified DWT image quality is roughly commensurate with the 12-tap Daubechies DWT image, yet it is coded at a lower bit rate. The modified DWT does increase ringing distortion slightly, but for this type of image can reduce the bit rate by above 10% for a given PSNR. This improvement is largely independent of the filters employed, although the improvement is slightly larger for orthogonal filters, such as Daubechies filters or the eigenfilters, as compared to linear phase filters, such as the lin6 filter set.
Figure 6.21. Reconstructed Newscaster images. (a) Original 8 bpp. (b) DWT using 12-tap Daubechies filters, 0.124 bpp, PSNR = 33.9 dB, (c) DWT maximum time-width 22-tap Daubechies filters, 0.120 bpp, PSNR = 33.1 dB (d) modified DWT using 12-tap Daubechies filters, 0.109 bpp, PSNR = 33.8 dB.

6.5. DISCUSSION

The performance of various subband filters and structures have been evaluated in this thesis in an objective manner. For image coding purposes a rigorous subjective
evaluation of the results is desirable. However this is a difficult task. The objective evaluations serve to highlight various characteristics of subband filters and structures for coding purposes, and are useful to determine what attributes are desirable. A rigorous subjective comparison can then be made of the various methods that perform in the optimum or near optimum fashion in an objective sense.

Various subjective observations, such as that relating ringing to filter time width, can be used in conjunction with objective measures to provide a better evaluation. Subjective observations have been made in this chapter, where appropriate. Further some of these observations have been given a physical explanation, such as the ringing associated with step-edge response of a filter.

In this chapter the subband image codecs have used 512x512 pixel resolution images. The conclusions drawn from these results are not necessarily applicable to other resolution images. However, a number of 256x256 pixel images have been tested in a similar manner, and the results, in terms of comparative performance between different subband filters and structures, are largely the same. For the 256x256 pixel images it appears that the time width of the filters is slightly more important than for the 512x512 pixel images. Also, there is less improvement at high compression using an $S=4$ (or greater) DWT over using an $S=3$ DWT. Both these observations are expected. For an image sampled at a lower rate (smaller number of pixels) there will be more edges relative to the total number of pixels, hence spatial resolution is more important. Further, the correlation of the image will be lower, and the latter observation above is then consistent with the rate constrained coding gain theory.

From these observations it is expected that the results for HDTV type resolution images will be similar. However, it is expected that for high compression quite high analysis levels ($M=32, 64$ or $S=5, 6$) are required for optimum performance, since the HDTV images generally exhibit a very high inter-pixel correlation. The DWT has an advantage over other subband schemes in that it is possible to increase the tree-depth or analysis level with relative computational ease.

The results of this chapter illustrate that good objective results can be obtained using a DWT with short filters. The best filters are those "optimised" in some sense for a highly correlated source. These filters exhibit a high coding gain and a low time width, but more importantly offer the best practical results. Long filters designed using
"mismatched" sources offer reasonable results, but are inferior to these other filters. It has also been observed that shorter filters offer better subjective results, with less edge distortion such as ringing or mosquito noise. A similar observation has been made by Fernandez and Ansari (1989). Lower edge distortion is attributed to less ringing associated with the step-edge response of a filter, which is related to its time width. Long filters, with large time widths, tend to introduce a large amount of ringing distortion at moderated to high compression, which can be quite objectionable to a human viewer.

Rioul (1993) compared the effects of regularity for orthogonal DWT image compression and concluded that some measure of regularity is desirable. The regularity of the lowpass filter is dependent on the number of zeros at DC of the corresponding highpass filter. Daubechies filters, having the maximum number of highpass zeros at DC, are the maximally regular orthogonal wavelet filters for a given length. It also appears that they also have the minimum time width among orthogonal CQFs for a given filter length. This fact coupled with a high coding gain suggests, as observed, an impressive coding performance. Longer filters, with more zeros at DC, are more regular, but have a larger time width. Rioul concluded that 12-tap Daubechies filters are optimum. A similar conclusion was given previously in this chapter: filters of length 8 to 12-taps are optimum. Longer filters, although more regular, have a larger time width and hence may perform in an inferior fashion.

Westerink et al (1988) examined several different subband analysis structures for subband image coding. Johnston's (1980) QMF's were used in a tree-structure to implement the subband decompositions. The DC subband was coded with DPCM, while the other bands were coded with PCM. Among the different structures evaluated were a uniform 4x4 decomposition, an S=2 DWT, and an S=2 modified DWT. Of the structures tested they concluded that for 256x256 images the best subband splitting scheme was a uniform 4x4 subband decomposition. However, the performance of the S=2 modified DWT was close to that of the 4x4 uniform decomposition. They observed that structures using a greater tree-depth (analysis levels) offered inferior results. However, as they noted, this is probably due to the fact that a non perfect reconstruction filter bank was used.

In contrast to the results of Westerink et al (1988), the results presented here suggest that coders using an S=3 DWT or MDWT outperform a coder using a uniform 4x4
SUBBAND CODEC RESULTS AND COMPARISONS

subband partition, since at most rates this latter partition does not sufficiently analyse the low frequency subbands. The difference in results is attributed to the use of different DWT filters and a different quantisation and encoding method. Using "optimum" DWT filters means that there is no degradation associated with increasing tree-depth, and that short filters with a high coding gain and low time width are employed. Using a spatially adaptive quantisation method, as opposed to the traditional non-adaptive subband quantisation method used by Westerink et al, allows exploitation of the good spatial resolution properties of these low time width filters.

Woods and Naveen (1992) compared various subband filters for image compression using a 4-band uniform subband decomposition, and a similar quantisation scheme to that of Westerink et al (1988) described above. The filters investigated were 12, 16 and 32-tap Johnston filters, 6 and 8-tap minimum phase Daubechies filters, SSKF(53) and two of Kronander's filters. The first observation made was that all the filters perform in a similar manner in a MSE sense, as is observed in this thesis. Upon closer inspection it was observed that the 16B Johnston QMF performed the best with Daubechies filters close behind. As with Westerink et al (1988) these results are somewhat different to the results presented in this thesis. Nevertheless, as above, a different subband decomposition and quantisation method were employed. It is proposed that the SSKF(53) filters did not perform as well as suggested in this thesis because the quantisation method was unable to exploit the excellent spatial resolution properties of these filters. As discussed in Section 6.1, using a uniform 4-band decomposition and a non-adaptive quantisation method tends to favour filters with better frequency resolution. Also, for the 8-tap Daubechies filters, using minimum time-width impulse responses may have given improved results. Nevertheless the subjective evaluation of Woods and Naveen is commensurate with subjective evaluations made in this thesis.

Ramchandran and Vetterli (1993) proposed a "best wavelet packet bases" approach to subband image analysis for compression purposes. This is actually an extension of some of the work by Coifman, Meyer, Quake, and Wickerhauser [See Ramchandran and Vetterli, 1993]. Using orthogonal two-band wavelet filters and a tree-structure the "best" subband structure for a particular image is determined in a rate distortion sense. Orthogonal filters are required so that an efficient search strategy is possible. Since the maximum coding gain eigenfilters are orthogonal they are suitable for such a method. Further, filters optimised for various parts of the tree-structure could be designed and employed. For a DWT (and modified DWT), as mentioned in Chapter 4, there is little
advantage in using different filters at different levels. However, this is not the case with other tree-structured approaches. Filters designed for tree branches that extend from a highpass subband may offer a potential improvement.

In this chapter Daubechies filters have been shown to perform in a similar manner to optimum eigenfilters. Although Daubechies filters exhibit a slightly inferior coding gain, or "weighted frequency response", this is compensated by better time localisation properties. The lin6 filters on the other hand, demonstrate that there exist (very) high coding gain biorthogonal filters, with excellent time localisation properties. Chung and Smith (1993) presented a time-varying $S=2$ DWT filter-bank using different IIR filters for "flat" image regions and edge or transition image regions. Filters with a good frequency response were used in the former regions while filters with a good step response were used in the latter regions. It is worth noting that the filters with a good step response have a frequency response similar to Daubechies filters and the lin6 filters (see Figure 3 in [Chung and Smith 1993]). The lin6 filters, while exhibiting good step response properties, also exhibit a very high coding gain for a highly correlated source. Therefore, it remains to be seen whether time-varying filters can outperform static filters with such properties.

6.6. CONCLUSION AND RECOMMENDATIONS

Using the generic subband quantisation and encoding method proposed in Chapter 5, various $M$-band transforms, DWT's, and other subband analysis structures have been compared for image compression. The results have been evaluated in terms of peak signal to noise ratio between the original and reconstructed image at various levels of compression. The results have illustrated that this generic quantisation and encoding method is a simple yet effective method of image compression for many subband analysis structures.

The observations and conclusions drawn from the $M$-band analysis scheme results are as follows: For the DCT, at low compression the PSNR increases for increasing $M$ (analysis level) up to $M=8$, while at $M=32$ or above the PSNR may decrease. At high compression the PSNR increases up to $M=16$ and does not decrease for $M=32$ and $M=64$. This observation is consistent with the rate-constrained coding theory presented in Chapter 3, which predicts that higher analysis levels ($M$) are required for optimum performance as the compression ratio increases. It is proposed that the sub-optimum
performance of the large $M$ DCT's at low compression is attributed to the nonstationary nature of image data. Some measure of time (spatial) localisation is desired when analysing a nonstationary source as opposed to a stationary source.

The LOT and ELT $M$-band schemes offered a similar performance to the DCT. The comparisons between the analysis levels ($M$) were similar to those of the DCT. However the performance of the large $M$ ELT was particularly inferior at low compression. This performance was attributed to the poor time localisation of the large $M$ ELT. Also the $M=8$ ELT was less sub-optimum at high compression as compared to the $M=8$ DCT. The $M=16$ analysis level was optimum at all compression ratios for the DCT, LOT and ELT. At this level these three transforms performed in a similar manner. It was proposed that any lack of frequency resolution on the part of the DCT is compensated by its fine time resolution (localisation). The DCT and the LOT generally outperformed the ELT by a small margin, suggesting that the ELT lacks sufficiently fine time resolution. The DCT was shown to outperform the DST and WHT by a large margin, illustrating that for short filters, a good frequency resolution relative to the input source is essential.

The following observations and conclusions were drawn from the results of the DWT schemes: The optimum filter length using Daubechies filters is from 8 to 12-taps. For any fewer number of taps the PSNR is less for a given compression ratio. Using filters longer than 12-taps the PSNR does not increase, and can even decrease for very long filters. These results corroborate those of Rioul (1993) who used a different quantisation and encoding method. It has been verified that using zero constrained filters is desirable in terms of PSNR. The optimum analysis level is $S=3$ at low compression and $S=4$ at high compression. It is worth noting that at high compression $S=5$ or greater offers a small improvement over $S=4$. Again these results are commensurate with the rate constrained coding gain theory. In contrast to the $M$-band schemes, the higher analysis levels are not sub-optimum at low compression. The minimum time width impulse response filters were observed to outperform the other phase response filters by a small margin.

Some AR(1) source optimised eigenfilters were evaluated for image compression. It was shown that short eigenfilters designed for a highly correlated source significantly outperform eigenfilters designed for a low correlated source. Various eigenfilters designed for an AR(1) source of high correlation, with a zero constraint, were
compared. As with the block-transform KLT, the performance is very similar for $\rho > 0.9$ in terms of image coding. Further these filters perform in a similar manner to Daubechies filters. This is not surprising since both filter sets have a high coding gain and good time (spatial) localisation properties (see Chapter 4). The results of this chapter have illustrated that a high coding gain is an important characteristic for image coding purposes. Similarly good time localisation is important, although to a lesser extent. Nevertheless, to obtain the best results, both characteristics must be considered.

The DWT schemes employing Daubechies filters, or the best eigenfilters, perform in a similar manner to the DCT in terms of PSNR at a given compression ratio. Other DWT filters, such as the CQF's designed by Smith and Barnwell, are inferior. This result was explained by the fact that these latter filters were designed for maximum average stopband attenuation rather than considering a typical image source, and are not zero-constrained.

A preliminary investigation of the performance of a DWT using linear phase biorthogonal filters has been performed. Le Gall and Tabatabai (1989) filters were shown to perform remarkably well given their short length. Some linear phase filters were designed, using a somewhat ad hoc approach based on the conclusions of this chapter, that outperformed the best orthogonal DWT filters by up to 0.5 at low compression and 1.0 dB at high compression. Some other biorthogonal filters, used by Antonini et al (1992), performed in a similar but slightly sub-optimum manner. These investigations certainly illustrate that biorthogonal filters have the potential to outperform orthogonal filters for DWT image coding. It is possible to design biorthogonal filters with a high coding gain (using a DWT structure) and very good time localisation properties. Optimum biorthogonal DWT filters should prove to be a fruitful topic of future research.

The modified DWT scheme was shown to offer an improvement of up to 0.8 dB over the DWT scheme for some images. The average gain is more in the range of 0.2-0.4dB, commensurate with the gain suggested by a two-dimensional generalised correlation model. Optimum filters were used for both the DWT and the modified DWT. The quadic analysis schemes gave results slightly down from the DWT scheme.

The results for the coding of 256x256 pixel resolution images were noted. Similar conclusions were drawn to those given for the 512x512 pixel images considered in this
chapter. These conclusions were then extrapolated to HDTV type resolution images. Some subjective issues were discussed, and the results compared to relevant results of other authors in the Discussion section.

Based on the conclusions drawn from this chapter, the following recommendations are made in relation to the design of subband image codecs:

- Subband filters should exhibit a high coding gain and good time localisation. For example use 8 to 12-tap Daubechies filters, or similar length eigenfilters designed for a highly correlated source.
- Minimum time width DWT filters should be used for the best objective results. Further it has been observed that the minimum time width filters exhibit less ringing giving better subjective results.
- For low compression (less than 10 times) of 512x512 images, an analysis level of $S=3$ and $M=8$ for a DWT and an $M$-band transform respectively, is sufficient for optimum performance. At moderate to high compression (greater than 30 times) a level of $S=4$ or $M=16$ is required. Higher resolution images, such as HDTV, are likely to require greater analysis levels.
- From a mean square error (MSE) perspective, the DCT, LOT and ELT and DWT using the best filters perform in a similar manner. Hence any of these methods is appropriate for minimum MSE applications.
- Should the best possible performance be required, the modified DWT offers a small improvement over the DWT for some images, at the expense of a small increase in computational cost.
- Biorthogonal filters should be considered for DWT image coding, as they have the potential to outperform the best orthogonal filters. However for some applications orthogonal filters are required.
CONCLUSION

CHAPTER 7:

CONCLUSION

The main theme of this thesis has been the investigation of subband methods that are optimum for image compression purposes. To this end each chapter has pursued certain aspects of the theory and practicalities associated with subband image coding.

In Chapter 1, the Introduction, a brief overview and history of subband image coding was given. Chapter 2 provided the background mathematics for subband analysis and synthesis. In particular, various one-dimensional subband methods were described using linear matrix equations. Some properties of orthogonal analysis/synthesis were discussed and derived. Separable two-dimensional subband analysis was considered, as the product of two one-dimensional subband analyses.

In Chapter 3 subband coding schemes that use independent subband quantisation were considered. The quantisation bit allocation among subbands was derived using a simple quantisation model, under an overall bit rate constraint, to minimise the mean square image reconstruction error. A general subband coding gain metric, estimating the gain of a subband system over a PCM system, was derived assuming this optimum bit allocation. This subband coding gain was introduced by Katto and Yashuda (1991), and is applicable to any perfect reconstruction subband scheme under a high rate assumption. The coding gain was extended to a rate constrained coding gain metric. Various properties of this rate constrained coding gain metric were examined. Common block transforms, as subband structures, were examined from a frequency domain perspective. In particular the coding gain performance for different sources was interpreted.

Using the subband analysis/synthesis background of Chapter 2, and the coding gain theory of Chapter 3, some maximum gain two-band orthogonal filters were derived in Chapter 4. The problem was considered as the extension of the two-band optimum block transform, the Karhunen-Loeve transform (KLT). The highpass solution filter was derived as an eigenvector of a modified correlation matrix. Various properties of
correlation matrices were thus considered, and used to derive properties of the optimum filters. As an extension of this filter design method optimum wavelet filters were derived.

Various two-dimensional subband image structures were considered in relation to a typical image power spectral density in Chapter 5. Three distinguishing characteristics of a subband analysis method were identified: namely the ideal subband structure that is approximated by the analysis, the degree of this approximation, and the time (or spatial) resolution of the subbands. The coding gain metric implicitly measures the first two characteristics, while filter time widths and the subband structure can be used to evaluate the third. Also, a generic subband quantisation and encoding method, suitable for any subband structure, was proposed.

Using this quantisation and encoding method various subband structures and filters were evaluated and compared for still image compression in Chapter 6. Among other conclusions the results illustrated that this quantisation and encoding method is a simple yet effective method for image compression for many different subband structures.

7.1. MAJOR FINDINGS

In Chapter 2 a necessary and sufficient condition for an orthogonal filter bank to be zero-constrained was given. A zero-constrained filter bank is one where all but one of the filters have a zero at DC. This attribute is widely regarded, and shown in Chapter 6, to be desirable for subband image coding purposes.

In Chapter 3 it was demonstrated using the unified subband coding gain metric that the KLT of size $N$ has the maximum gain of any block transform of size $N$, orthogonal or not. Although this result is expected from the rate-distortion bound of encoding Gaussian variables in $N$-blocks, it provides an upper bound on the unified subband coding gain of all subband schemes. In particular it means that the subband coding gain is absolutely bounded by the inverse of the spectral flatness measure.

The validity of the subband coding gain metric is based on a high bit rate assumption. In Chapter 3 a rate constrained coding gain, applicable to orthogonal subband methods, was introduced. This constrained metric is useful for predicting the performance of
subband schemes at various bit rates. It was shown that the rate constrained coding gain decreases as the bit rate decreases. More importantly, it was shown that highly correlated sources require quite high analysis levels for optimum performance at low bit rates. This characteristic has important implications in the design of high compression applications of high resolution images such as HDTV.

A theorem relating filter pairwise symmetry to a symmetric coding gain with respect to the sign of the correlation coefficient $\rho$ for an AR(1) source was given in Chapter 3. Although this theorem is mainly of theoretical interest, coupled with the frequency domain interpretation of the performance of block transforms (also given in Chapter 3), it provides some interesting insights into desirable transform attributes for various applications. These insights led to the design of the filters given in Chapter 4.

In Chapter 4 a theorem relating the zeros of an eigenvector of a symmetric Toeplitz matrix corresponding to the minimum (maximum) eigenvalue to the multiplicity of this eigenvalue was given. A sufficient condition was given for the globally maximum coding gain orthogonal two-band filters, extending a previously known necessary condition. These filters are referred to as (optimum) eigenfilters.

The optimum eigenfilter design was formulated as a type of extended two-band KLT problem. This approach offers several insights. For example, the above theorem was used to predict certain characteristics of the optimum filters. Also, the optimum filter bank was shown to exhibit some characteristic properties of the KLT: namely globally maximum coding gain (over orthogonal two-band filter banks), and data decorrelation. Finally, it was shown that there are several different optimum filters with different impulse responses but the same magnitude response. This freedom can be used to select the minimum time width impulse response, which was shown in Chapter 6 to be a desirable attribute for subband image coding.

The design of maximum gain filters was extended to include filters constrained to lie in certain subspaces. For example, maximum gain wavelets were designed. Various properties relating to the unconstrained filters were shown to be applicable to the optimum constrained filters.

Some two-dimensional subband analysis schemes have been evaluated using the coding gain metric in Chapter 5. Using an octave-band filter bank (discrete wavelet transform,
DWT) it was shown that the coding gain of Daubechies (1988) filters is very close to that of the optimum eigenfilters. Daubechies filters were also shown to have a small time width in Chapter 3 (see also [Dorize and Villemoes 1991]), suggesting that they will perform well for DWT image compression. A modified DWT was proposed based on a typical image model. In Chapter 6 it was illustrated that for some images the modified DWT offers a small improvement over the DWT.

A generic subband quantisation and encoding method was introduced in Chapter 5 which is suitable for any subband structure. This method is a generalisation of the baseline JPEG method and provides a useful platform with which to compare various subband analysis schemes. Also, as the results of Chapter 6 attest, it is a simple yet effective subband quantisation and encoding method.

In Chapter 6, using this generic subband quantisation and encoding method, many different subband analysis structures and filters were compared for still image compression. There are many conclusions that have been drawn from these results. For example, the rate constrained coding gain prediction that higher analysis levels are required at high compression, as compared to low compression, was confirmed. A DWT (or modified DWT) using the optimum eigenfilters designed for a highly correlated source, or Daubechies filters, was shown to offer mean square error results commensurate with, or possibly slightly better than, those of the DCT.

Filter bank coding gain and time (spatial) localisation properties were shown to be important characteristics in the design of a subband image coder. A filter bank with high coding gain and good time localisation properties performs a principle-component-like analysis of an input image. It was shown that biorthogonal DWT filters have the ability to outperform orthogonal DWT filters in this regard. Also, it was observed that images coded by subband structures using minimum time width filters exhibit less ringing. Some further conclusions and recommendations were given in Chapter 6, Section 6.6.

7.2. FUTURE WORK

Several questions regarding subband analysis/synthesis systems for still image compression were addressed in this thesis. However, there are several avenues of future
research that should prove fruitful. Some of these areas are spin-offs from this work, while others are directly related to subband methods for image compression.

As discussed in Chapter 3, Section 3.4.1, the existence of the optimum eigenfilters at the minimum eigenvalue point has not been proven in the general case. Such a proof is desirable, although it has eluded the author and several people from whom help has been sought to date. Also, it may be possible to extend this method to the solution of general quadratic problems with quadratic and linear constraints. Certainly, at least Toeplitz quadratic forms should be investigated. The advantage of such a method is that it gives the globally optimum solution.

Another area worth pursuing, in the area of eigenfilter design, is the extension of the KLT like design formulation to the general $M$-band case as opposed to the 2-band case. This extension is not immediately obvious. Nevertheless it is straightforward to show using similar arguments to those used in the derivation of the KLT, that a necessary condition for the optimum filters is that the subband correlation matrix is diagonal: that is the subband data is decorrelated. However, unlike the block transform KLT, this is not a sufficient condition. Soman and Vaidyanathan (1993) offered an interesting approach to this problem. However it is uncertain as to whether their solution is globally optimum.

In Chapter 6 it was illustrated that biorthogonal filters have the potential to outperform orthogonal filters for image compression. A hybrid coding gain / inverse time width weighted measure for a two-band (or possibly the general $M$-band case) could be developed. Using this measure, high coding gain and low time width biorthogonal filters could be designed. Some further examination of the relative importance of filter time width as compared to coding gain may be required to this end.

The assumptions used in the optimum bit allocation among subbands, derived in Chapter 3, are valid only at high rates or for orthogonal schemes. Therefore the bit allocation for biorthogonal schemes needs to be investigated at low bit rates. A related problem was considered by Moulin (1993). However, the results of Chapter 6 demonstrate impressive results for the biorthogonal schemes at all bit rates. It is uncertain if a significant improvement can be made for the biorthogonal schemes considered in this thesis. If such an improvement is possible, then the results will certainly be impressive.
Finally, it is worth investigating whether the best fixed subband structure, such as a DWT or modified DWT using optimum filters, can be significantly improved upon using a "best basis" type approach [Ramchandran and Vetterli 1993]. Certainly there will be some images where this will be the case. However, for most images this is not obvious in the authors opinion. Another closely related open question, discussed in Section 6.5 of Chapter 6, is whether time-varying filters can outperform the best static filters.
REFERENCES


REFERENCES


APPENDIX A

TECHNICAL REPORT: SOME PROPERTIES OF THE EIGENVECTORS OF SYMMETRIC TOEPLITZ MATRICES

A.1 INTRODUCTION

This appendix contains a technical report that summarises some properties of the eigenvectors of symmetric Toeplitz matrices. Stationary correlation matrices are positive definite (symmetric) Toeplitz matrices. A generalisation of a previously known property is stated and proved as Theorem 1. This theorem corresponds to Theorem 4.1, stated in the main text of Chapter 4.
SOME PROPERTIES OF THE EIGENVECTORS OF SYMMETRIC TOEPLITZ MATRICES

Technical Report: Electrical and Computer Engineering
University of Wollongong
July 1993
J. P. Andrew, P. O. Ogunbona, and F. J. Paoloni

ABSTRACT

This technical report summarises various properties of the eigenvectors of symmetric Toeplitz matrices. These properties pertain to wide sense stationary source correlation matrices which are symmetric (positive definite) and Toeplitz. Further a new property of the eigenvectors of such matrices is also given as Theorem 1.

1. INTRODUCTION

In Section 1 of this report we summarise various properties of the eigenvectors of symmetric Toeplitz matrices. Symmetric Toeplitz matrices are of interest since they encompass wide sense stationary source correlation matrices. We also give a new property of the eigenvectors of such matrices, which is stated in Theorem 1.

Theorem 1. Any eigenvector of a symmetric Toeplitz matrix of dimension N x N corresponding to the minimum eigenvalue of multiplicity k, has at least N-k zeros on the unit circle. Further every such eigenvector has N-k zeros on the unit circle in common; and any vector with these common zeros is such an eigenvector.

In this report N refers to the dimension (N x N) of the symmetric Toeplitz matrix R, and hence to the length of any associated eigenvectors.

The proof of this theorem is given in Section 3, where we outline the steps of the proof as a series of propositions and definitions. Each proposition is proved in a later subsection.
2. SUMMARY OF EIGENVECTOR PROPERTIES

A positive definite matrix is assumed symmetric by definition. Symmetric Toeplitz matrices, as a subset of doubly symmetric or symmetric centro-symmetric (SC) matrices (see Cantoni and Butler 1976), possess the following properties,

A complete set of orthogonal eigenvectors (applies to any symmetric matrix).

There are \[ \frac{N}{2} \] symmetric eigenvectors and \[ \frac{N}{2} \] skew symmetric eigenvectors where the matrix is of dimension NxN.

If the minimum (maximum) eigenvalue is distinct the corresponding eigenvector has all its zeros on the unit circle.

The first property applies to any symmetric matrix. Cantoni and Butler (1976) demonstrated the second property as applied to SC matrices, and hence this property applies to symmetric Toeplitz matrices. If any eigenvalues are repeated it may be possible to select associated asymmetric eigenvectors. Nevertheless it is also possible to select a full complement of (skew) symmetric eigenvectors if one desires. Robinson (1967, p271) proves the third property in the case of positive definite Toeplitz matrices. Theorem 1 is a generalisation of this third property. Makhoul (1981) also summarises several properties of symmetric Toeplitz matrices.

3. PROOF OF THEOREM 1

The proof of Theorem 1 is outlined as a sequence of statements, also given in the main text of Chapter 4, as follows,

1. Any vector, \( x \), that minimises the quadratic sum, \( x'Rx \) given \( x'x=1 \), is a minimum eigenvector.

2. Inverting zeros (real or conjugate pairs) of a minimum eigenvector gives another minimum eigenvector.
3. The family of a vector, \( x \), denotes all vectors that can be generated through zero inversion and linear combinations of the resulting vectors. Every vector in the family of a minimum eigenvector is a minimum eigenvector (from 2).

4. If a vector has \( N-k \) zeros on the unit circle and \( k-1 \) other real zeros then its family spans \( k \) dimensional space.

5. If a vector has \( N-k \) zeros on the unit circle and \( k-1 \) other arbitrary zeros, there is a vector in its family that has these \( N-k \) same unit circle zeros and \( k-1 \) other real zeros. Hence its family spans \( k \) dimensional space (at least, and can be shown to span only \( k \) dimensional space).

6. From 5 every minimum eigenvector must have \( N-k \) zeros on the unit circle else we could generate a family of such eigenvectors that spans greater than \( k \) dimensional space.

7. Every minimum eigenvector has the same \( N-k \) zeros on the unit circle since it must lie in the space spanned by \( k \) linearly independent eigenvectors with \( N-k \) zeros on the unit circle in common.

8. (Corollary) Any vector with these same \( N-k \) zeros on the unit circle and arbitrary other zeros (giving a real vector) is a minimum eigenvector since it lies in the space spanned by the above linearly independent eigenvectors.

The proof of each statement, where required, is given in the following subsections. Statement 1 is known as Rayleigh's principle [Strang 1988, p349] and the proof is given here for clarity. Statement 2 is also known under a different guise as the Fejér factorisation of a filter's auto-correlation function [Robinson p335 and p269-272]. The remaining properties extend results previously determined for the case \( k=1 \).

### 3.1. Proof of Statement 1

*Any vector that minimises the quadratic sum, \( x'Rx \) given \( x'x = 1 \), is an eigenvector corresponding to the minimum eigenvalue.*

This theorem is known as Rayleigh's principle [Strang 1988, p349]. The minimum eigenvalue is labelled \( \lambda_1 \) and the corresponding eigenvector, referred to as the minimum eigenvector, is labelled \( x_1 \). Generally the labelling of eigenvalues and associated eigenvectors is such that \( \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \). Since \( R \) is symmetric there is a complete set of orthogonal eigenvectors [Strang p290 or p309]. Let \( S \) denote an orthogonal matrix whose columns consist of the eigenvectors of \( R \). The columns of \( S \)
are sorted corresponding to the above ordering of eigenvalues. i.e. column one contains a/the minimum eigenvalue while column $N$ contains a/the maximum eigenvector. Using this notation one has,

$$RS = S\Lambda$$

$$S'RS = S'S\Lambda = \Lambda$$

where $\Lambda$ is a diagonal matrix of eigenvalues. Also let,

$$y = S'x \text{ or } x = Sy$$

giving $y$ as the orthogonal eigen-transform of the vector $x$. The quadratic sum in $x$ can be manipulated thus;

$$\frac{x'Rx}{x'x} = \frac{y'S'RSy}{y'y}$$

$$= \frac{y'\Lambda y}{y'y}$$

$$= \frac{\lambda_1 y_1^2 + ... + \lambda_N y_N^2}{y_1^2 + ... + y_N^2}$$

Assuming the ordering of eigenvalues given above yields,

$$\frac{x'Rx}{x'x} \geq \frac{\lambda_1(y_1^2 + ... + y_N^2)}{y_1^2 + ... + y_N^2} = \lambda_1, \quad \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$$

Given $x'x=1$, then $y'y=1$ and the left hand side of the above equation is the quadratic, i.e.

$$x'Rx \geq \lambda_1$$

(3)

If $\lambda_1$ is of multiplicity $k, \lambda_1 = \lambda_2 = \ldots = \lambda_k < \lambda_{k+1} \leq \ldots \leq \lambda_N$, then

$$\frac{x'Rx}{x'x} = \frac{y'\Lambda y}{y'y}$$

$$= \frac{\lambda_1(y_1^2 + ... + y_k^2) + (\lambda_{k+1}y_{k+1}^2 + \ldots + \lambda_N y_N^2)}{y_1^2 + ... + y_N^2}$$

$$= \lambda_1 \left( \frac{y_1^2 + ... + y_k^2}{y_1^2 + ... + y_N^2} \right) + (\lambda_{k+1}y_{k+1}^2 + \ldots + \lambda_N y_N^2)$$

(4)
and hence,
\[
x'Rx \left\{ \begin{array}{l}
\lambda_1 (y_1^2 + \ldots + \lambda_N y_N^2) \\
y_1^2 + \ldots + y_N^2
\end{array} \right. \\
= \lambda_1
\]
for some \( y_{k+1}, \ldots, y_N \neq 0 \)
for all \( y_{k+1}, \ldots, y_N = 0 \)

Any vector that minimises the quadratic sum is a linear combination of the eigenvectors that correspond to the minimum eigenvalue since \( y_{k+1}, \ldots, y_N = 0 \). This vector is thus a minimum eigenvector because,

If \( Rx_1 = \lambda_1 x_1, Rx_2 = \lambda_2 x_2 = \lambda_1 x_2 \), then
\[
R(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 Rx_1 + \alpha_2 Rx_2
\]
\[
= \lambda_1 (\alpha_1 x_1 + \alpha_2 x_2)
\]

ie. any linear combination of eigenvectors corresponding to the same eigenvalue is another such eigenvector.

As noted by Strang (1988 p349) nowhere is it assumed in this proof that \( R \) is positive definite. This theorem holds for any symmetric \( R \).

3.2. Proof of Statement 2

*Inverting a zero, real or conjugate pair, of a minimum eigenvector gives another minimum eigenvector.*

Consider a real FIR filter of length \( N \) represented as a vector of coefficients, \( h \). The Z-transform representation gives \( N-1 \) zeros, which are either real or come in complex conjugate pairs. ie.,

\[
H(z) = A(1 + a_1 z^{-1} + \ldots + a_n z^{-n})
\]
\[
= A(1 - r_1 z^{-1})(1 - r_n z^{-1})
\]

where \( H(z) \) is the Z-transform of the filter \( h \). For simplicity the scaling factor \( A \) is set to one. The magnitude of the frequency response of \( H \) is,
Inverting a real zero, say $r_k$, and denoting the resulting filter by $H_i$ (or $h_i$) gives;

$$
|H_i(e^{j\omega})| = \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_k e^{-j\omega}\right) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_k e^{-j\omega}\right) \right|
$$

In other words its magnitude response is the same as $H(e^{j\omega})$ to within a scale factor. Similarly inverting a conjugate pair of zeros, say $r_k$ and $r_{k+1}$, and again denoting the resulting filter by $H_i$, gives;

$$
|H_i(e^{j\omega})| = \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

$$
= \left| H(e^{j\omega}) \right| \left| \left(1 - r_k e^{-j\omega}\right) \ldots \left(1 - r_{k-1} e^{-j\omega}\right) \right|
$$

Again the magnitude response is unchanged to within a scale factor.
The energy of the filter \( h \) is defined as,

\[
h'h = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{j\omega})|^2 \, d\omega
\]  
(10)

Assuming that \( h \) has unit energy we normalise \( h \) to unit energy by multiplying by \( |r_i|^2 \).

Hence when both the filters are normalised to unit energy, the magnitude responses of \( h \) and \( h_i \) are identical.

If \( \mathbf{R}_{xx} \) is a symmetric Toeplitz matrix, then,

\[
x'\mathbf{R}_{xx}x = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\omega})|^2 S_{xx}(e^{j\omega}) \, d\omega
\]  
(11)

where \( S_{xx}(e^{j\omega}) \) is the Fourier Transform of the sequence \( R(n) \) which is defined as,

\[ R(n) = \mathbf{R}_{xx}[i,j], \quad n = i - j \]

Since \( \mathbf{R}_{xx} \) is Toeplitz there is no ambiguity in this last equation. If \( \mathbf{R}_{xx} \) is positive (semi) definite then \( S_{xx}(e^{j\omega}) \) is the power spectral density of some source with autocorrelation sequence given by \( R(n) \). However equation (11) holds whether \( \mathbf{R}_{xx} \) is positive definite or not. The quadratic form \( x'\mathbf{R}_{xx}x \) makes sense only when \( \mathbf{R}_{xx} \) is symmetric, otherwise the integral is valid for all Toeplitz matrices.

Unless otherwise specified it is assumed henceforth that each vector has unit energy.

Since unit energy normalised \( x \) and \( x_i \) have the same magnitude response they have the same quadratic cost. It follows from Rayleigh's principle (Statement 1) that if \( x \) is an eigenvector of \( \mathbf{R}_{xx} \) corresponding to the minimum eigenvalue then so is \( x_i \). Further any linear combination of these two vectors similarly gives an eigenvector.

### 3.3. Vector families: Definition of Statement 3

Let \( x_i \) denote any vector that has zeros in a one to one correspondence with those of \( x \) or those of \( x \) inverted, noting that all filters considered are real. For example,
\[
X(z) = (z-c_1)(z-c_2^*)(z-r_1)
\]
\[
X_i(z) = (z-c_1)(z-c_2^*)\left(z - \frac{1}{r_1}\right)
\]
\[
\text{or } X_i(z) = \left(z - \frac{1}{c_1}\right) \left(z - \frac{1}{c_2^*}\right) (z-r_1)
\] (12)

For simplicity we use \(z\) rather than \(z^{-1}\). Note that if the inverted zero(s) is on the unit circle (UC), then \(x_i\) and \(x\) are identical: complex zeros on the UC map to their conjugates and a zero at \(\pm 1\) is mapped to itself under zero inversion.

The family of \(x\) is thus defined,

\[\text{The family of } x, \mathcal{F}_x, \text{ is the set of vectors consisting of } x, \text{ all } x_i, \text{ linear combinations thereof and ensuing zero inversions.}\]

Note that the family of any vector in \(\mathcal{F}_x\) is a subset of \(\mathcal{F}_x\). Every vector in the family of a minimum eigenvector, \(x_1\), of \(R\) is a minimum eigenvector. This follows directly from statement 2 since all such vectors have the same minimum quadratic cost.

3.4. Proof of Statement 4

\[\text{If a vector, } x, \text{ has } N-k \text{ zeros on the unit circle and } k-1 \text{ other real zeros then its family spans } k \text{ dimensional space.}\]

The proof proceeds by induction. Consider the case \(k=2\). (The case \(k=1\) is trivial and offers less insight). In this case \(x\) has one real zero off the unit circle which we denote \(r_1\). Let \(x_i\) have the same UC zeros as \(x\) and a real zero at \(\frac{1}{r_1}\). Since \(x_i \neq \alpha x\), these two vectors are linearly independent and span a two-dimensional space. Any vector in the family of \(x\), has the same \(N-2\) UC zeros as \(x\). There is no way to move these UC zeros through zero inversion and/or linear combinations of vectors with these zeros in common. Therefore the family of \(x\) can span at most a two-dimensional space. Alternatively one may show that the family of \(x\) is spanned by the set of vectors \(\{x, x_i\}\).

Now assume that the proposition is true for \(k\), and consider the case \(k+1\). Let
\[ X(z) = (z-u_1) \ldots (z-u_{N-k-1})(z-r_1)(z-r_2) \ldots (z-r_k) \]

\[ X_i(z) = (z-u_1) \ldots (z-u_{N-k-1}) \left( z - \frac{1}{r_i} \right)(z-r_2) \ldots (z-r_k) \] \hfill (13)

where \( u \) represents a unit circle zero and \( r \) represents a real zero not on the unit circle. Further, consider the linear combination,

\[ Y = \alpha X + (1-\alpha)X_i \]

\[ = (z-u_1) \ldots (z-u_{N-k-1})(z-r_2) \ldots (z-r_k) \left[ \alpha(z-r_1) + (1-\alpha) \left( z - \frac{1}{r_i} \right) \right] \] \hfill (14)

\[ = (z-u_1) \ldots (z-u_{N-k-1})(z-r_2) \ldots (z-r_k) \left( z - \frac{\alpha r_i^2 + (1-\alpha)}{r_i} \right) \]

noting that \( y \) is a member of the family of \( x \). It is possible to place a zero at 1 \((z=1)\) which is on the UC, by selecting,

\[ \frac{\alpha r_i^2 + (1-\alpha)}{r_i} = 1 \]

\[ \alpha(r_i^2 - 1) = r_i - 1 \] \hfill (15)

\[ \alpha = \frac{r_i - 1}{(r_i^2 - 1)} \]

noting that \( r_i \neq \pm 1 \). Hence \( y \) has \( N-k \) zeros on the UC and \( k-1 \) real other zeros. From the inductive assumption the family of \( y \) spans \( k \)-dimensional space. Further every vector in this space has (the same) \( N-k \) zeros on the UC. Obviously \( x \) is linearly independent from the family of \( y \) since it has only \( N-k-1 \) zeros on the UC. Hence the family of \( x \), which contains \( x \) and the family of \( y \), spans at least \( k+1 \) dimensional space. Reasoning as in the case \( k=2 \) the family of \( x \) can span at most \( k+1 \) dimensional space since every vector in this space has the same \( N-k-1 \) zeros on the UC.

To conclude: because the proposition is true for \( k=2 \), and true for \( k+1 \) given it is true for \( k \), it is true for all \( k \).
3.5. Proof of Statement 5

*If a vector has N-k zeros on the unit circle and arbitrary other zeros, there is a vector in its family that has the same N-k unit circle zeros and k-1 real zeros not on the unit circle. Hence its family spans k dimensional space (at least, but can be shown to span only k dimensional space).*

Consider as above,

\[ X(z) = (z-u_1) \cdots (z-u_{N-k})(z-z_1)(z-z_2) \cdots (z-z_{k-1}) \]

\[ X_i(z) = (z-u_1) \cdots (z-u_{N-k}) \left( z - \frac{1}{z_1} \right) \left( z - \frac{1}{z_2} \right) \cdots (z-z_{k-1}) \] (16)

where \( z \) represents an arbitrary zero, real or complex, not on the unit circle. Let \( z_1 \) be complex and \( z_2 = z_1^* \). If there are no complex zeros off the unit circle statement 5 is equivalent to statement 4. Consider the linear combination,

\[ Y = \alpha X + (1-\alpha)X_i \]

\[ = (z-u_1) \cdots (z-r_k) \left[ \alpha(z-z_1)(z-z_1^*) + (1-\alpha) \left( z - \frac{1}{z_1} \right) \left( z - \frac{1}{z_1^*} \right) \right] \] (17)

\( Y \) is the product of the common zeros of \( X \) and \( X_i \), and a quadratic \( Y_q \). The following analysis shows that with propitious choice of \( \alpha \) it is possible to generate a quadratic term with real roots not on the UC. Consider the quadratic,

\[ Q = z^2 + 2 \text{Re}(z_1) + |z_1|^2 \equiv (1, 2b, c) \] (18)

where \( b = \text{Re}(z_1) \) and \( c = |z_1|^2 \). Now let,

\[ Q_i = z^2 + 2 \text{Re} \left( \frac{1}{z_1} \right) + \left| \frac{1}{z_1} \right|^2 \]

\[ = z^2 + \frac{2 \text{Re}(z_1)}{|z_1|^2} + \left| \frac{1}{z_1} \right|^2 \]

\[ = \left( 1, \frac{2b}{c}, \frac{1}{c} \right) \]
Note that $b^2 < c$ since the roots are complex and that $c>0$ since it is the magnitude squared of the complex roots.

The quadratic $Y_q$ is a linear combination of $Q$ and $Q'$,

$$Y_q = \alpha(1,2b,c)+(1-\alpha)\left(1,\frac{2b}{c},\frac{1}{c}\right)$$

$$= \left(1,\alpha 2b+(1-\alpha)\frac{2b}{c},\alpha c+(\frac{1-\alpha}{c})\right)$$

$$= \left(1,\frac{\alpha b(c-1)+b}{c},\frac{\alpha(c^2-1)+1}{c}\right)$$

$$= (1,2b',c') \quad (19)$$

Now consider the case where $Y_q$ has double roots or a zero discriminant,

$$b'^2 - c' = 0$$

$$\left(\frac{\alpha b(c-1)+b}{c}\right)^2 - \frac{\alpha(c^2-1)+1}{c} = 0 \quad (20)$$

$$\frac{\alpha^2 b^2(c-1)^2 + 2\alpha b^2(c-1) + b^2 - \alpha c(c^2-1) - c}{c^2} = 0$$

Hence,

$$\alpha^2 b^2(c-1)^2 + \alpha(2b^2(c-1) - c(c^2-1)) + b^2 - c = 0$$

We can write this equation as a quadratic in $\alpha$ as,

$$a_\alpha \alpha^2 + b_\alpha \alpha + c_\alpha = 0$$

It is evident that $a_\alpha > 0$, and $c_\alpha < 0$ since $b^2 - c < 0$. Therefore for some real $\alpha$ the above quadratic in $\alpha$ is zero and the original discriminant is zero. Hence for some real $\alpha$, $Y_q$ has a real double root.

It is now shown that this double root can't lie at $\pm 1$, ie. it is not on the unit circle. For a double root at $\pm 1$ it is necessary that $-b' = \pm 1$. Then,
\[ \frac{-b \alpha (c-1) + b}{c} = \pm 1 \]

\[ b \alpha (c-1) = \mp c - b \]

\[ \alpha = \frac{\mp c - b}{b(c-1)} \]

Substituting this value of \( \alpha \) into the quadratic \( Y \) at \( \pm 1 \),

\[ Y_q(\pm 1) = 1 \pm 2 \frac{b(\frac{\mp c - b}{b(c-1)})(c-1) + b}{c} + \frac{(\frac{\mp c - b}{b(c-1)})(c^2 - 1) + 1}{c} \]

\[ = 1 \pm 2 \frac{\mp c - b + b}{c} + \frac{(\mp c - b)(c+1) + 1}{c} \]

and simplifying gives,

\[ Y_q(\pm 1) = \frac{bc \pm (\mp 2bc) \mp c^2 \mp c - bc - b + b}{bc} \]

\[ = \frac{-2bc \mp c^2 \mp c}{bc} \]

\[ = \frac{\mp c(c \pm 2b + 1)}{bc} \]

Equating to zero, assuming a root, gives,

\[ c \pm 2b + 1 = 0 \]

\[ b = \frac{\mp (c + 1)}{2} \]

which is necessary for a double root at \( \pm 1 \). In the case where this relationship between \( b \) and \( c \) holds the discriminant of the original quadratic \( Q \) is,

\[ b^2 - c = \left( \frac{\mp (c + 1)}{2} \right)^2 - c \]

\[ = \frac{c^2 + 2c + 1 - 3c}{4} \]

\[ = \frac{(c - 1)^2}{4} \geq 0 \]
This is contrary to the assertion of complex zeros, ie. a negative discriminant. Therefore there can't be a double root at ±1 if the original root $z_1$ is complex.

It is possible to construct a polynomial $Y$ which is the product of the common roots of $X$ and $X_i$ and two real non UC roots. Proceeding in a similar manner one can move all non UC conjugate zero pairs of $X$ to real non UC pairs. This generates a vector with $N-k$ UC zeros and $k-1$ real zeros off the UC. Since this process involves only zero inversion and linear combinations, this vector is in the family of $x$.

From proposition 4 the resulting vector has a family that spans $k$ dimensional space and hence the family of $x$ spans at least $k$ dimensional space. The family can't span greater than $k$ dimensional space since every vector in the family has the same $N-k$ UC zeros. Alternatively the family of $x$ can be generated with $k$ linearly independent vectors.

3.6. Proof of Statement 6

*Every minimum eigenvector, corresponding to an eigenvalue of multiplicity $k$, must have at least $N-k$ zeros on the unit circle.*

Consider a minimum eigenvector, corresponding to eigenvalue of multiplicity $k$, with $N-I$ zeros on the UC and $l-1$ arbitrary other zeros. From statement 5 the family of this vector spans $l$ dimensional space. Hence from Statement 3 there is a subset of minimum eigenvectors that span $l$ dimensional space. From the spectral theorem of linear algebra (Strang 1988, p296) there are $N-k$ eigenvectors orthogonal to this $l$ dimensional space. The dimension of the composite space consisting of these two spaces is bounded by $N$ since the spanning vectors are of length $N$. Hence one has $N-k+l \leq N$ which implies that $l \leq k$.

As mentioned in Section 2, Robinson (1967, p271) proves statement 6 for positive definite Toeplitz matrices where $k=1$. In the proof we have presented for case of arbitrary $k$, the positive definite requirement has been relaxed to include all symmetric Toeplitz matrices.
3.7. Proof of Statement 7

Every minimum eigenvector has the same $N-k$ zeros on the unit circle.

Corresponding to the minimum eigenvalue of multiplicity $k$ we have $k$ linearly independent (minimum) eigenvectors, $x_1, ..., x_k$. Consider the eigenvectors,

$$x_1 = (x_{11}, ..., x_{1N})$$
$$x_k = (x_{k1}, ..., x_{kN})$$

(25)

Another minimum eigenvector $y$ can be formed, so that the first $k-1$ components are zero, ie.,

$$y = (0, ..., 0, y_{k}, ..., y_{N})$$

For example consider,

$$\alpha_1 (x_1, ..., x_{i(k-1)}) + \ldots + \alpha_{k-1} (x_{(k-1)i}, ..., x_{(k-1)(k-1)}) - (x_{kl}, ..., x_{k(k-1)}) = (0, ..., 0)$$

(26)

Rearranging and writing in matrix form,

$$\begin{bmatrix}
  x_{11} & \ldots & x_{(k-1)i} \\
  \ldots & \ldots & \ldots \\
  x_{(k-1)i} & \ldots & x_{(k-1)(k-1)}
\end{bmatrix} \begin{bmatrix}
  \alpha_1 \\
  \ldots \\
  \alpha_{k-1}
\end{bmatrix} = \begin{bmatrix}
  x_{kl} \\
  \ldots \\
  x_{k(k-1)}
\end{bmatrix}$$

(27)

If the above matrix on the left hand side of the equation is non-singular then one can solve for $\alpha_1, ..., \alpha_{k-1}$. If the matrix is singular then there exists some $\alpha_1, ..., \alpha_{k-1} \neq 0$ such that the right hand side vector is identically zero. ie. we don't have to include $x_k$ in the sum to generate $k-1$ zero components of $y$. {We suspect that this matrix is non-singular in any case}.

From $y$ one can easily generate $k$ linearly independent minimum eigenvectors. For example consider,
\[ y_1 = (0, \ldots, 0, y_k, \ldots, y_N, 0) \]
\[ y_2 = (0, \ldots, 0, y_k, \ldots, y_N, 0, 0) \]
\[ \vdots \]
\[ y_{k-1} = (y_k, \ldots, y_N, 0, \ldots, 0) \]  

(28)

In effect one has \( y_n = z^n y \), that is one is adding zeros at zero, while maintaining the original zeros. Alternatively one may think of this as inverting real infinite zeros. Obviously the magnitude response of each \( y_n \) is identical to that of \( y \) and hence each is a minimum eigenvector. Further it is evident that \( y \) and the all the \( y_n \) form a linearly independent set, thus spanning the \( k \) dimensional space of minimum eigenvectors.

From proposition 6 \( y \) must have at least \( N-k \) zeros on the UC. In fact \( y \) really has only \( N-k \) zeros with \( k-1 \) zeros at infinity. Each \( y_n \) has these same zeros, with \( n \) other zeros at zero. Hence any vector in the space spanned by \( y \) and the \( y_n \) has these same \( N-k \) UC zeros. Since any minimum eigenvector is in this space it too must have such zeros.

3.8. **Proof of Statement 8 (Corollary)**

*Any vector with these same \( N-k \) zeros on the unit circle and arbitrary other zeros (giving a real vector) is a minimum eigenvector.*

Any vector with these \( N-k \) zeros on the unit circle can be written as a linear combination of \( y \) and the \( y_n \) and hence is a minimum eigenvector. For example consider,

\[
X(z) = (z-u_1)(z-u_{N-k})(z-z_1)(z-z_{k-1}) \\
= (z-u_1)(z-u_{N-k})[z^{k-1} + \alpha_{k-2}z^{k-2} + \ldots + \alpha_0] \tag{29}
\]

Now consider some linear combination of \( y \) and the \( y_n \),

\[
y_{k-1} + \beta_{k-2}y_{k-2} + \ldots + \beta_0y = (z-u_1)(z-u_{N-k})[z^{k-1} + \beta_{k-2}z^{k-2} + \ldots + \beta_0] \tag{30}
\]

Equating the \( \alpha_i \) and \( \beta_i \) one has the desired combination. If \( x \) has \( l \) zeros at infinity, \( l < k \), then \( y_{k-1}, \ldots, y_{k-l} \) are not required in the linear combination.
4. REFERENCES


APPENDIX B: Subband Coding Gain

B.1 Introduction

In this appendix some of the mathematical detail required for the material presented in Chapter 3 is given. The title of each section or subsection in this appendix is the same as the title of the corresponding section in Chapter 3.

B.1.1. Lowpass to highpass transformations

In this section it is shown that the lowpass to highpass transformation reflects the magnitude spectrum of a real filter or source about π/2 radians (0.25 cycles/sample).

If \( g(n) = (-1)^n h(n) \), as given by equation (3.11) in the main text of Chapter 3, then,

\[
G(e^{j\omega}) = \sum_n g(n)e^{-j\omega n} \\
= \sum_n (-1)^n h(n)e^{-j\omega n} \\
= \sum_n h(n)e^{-j(\omega + \pi)n} \\
= H(e^{j(\omega + \pi)})
\]

For real filters \( |H(e^{j\omega})| = |H(e^{j(2\pi - \omega)})| \) and hence

\[
|G(e^{j\omega})| = |H(e^{j(2\pi - (\omega + \pi))})| = |H(e^{j(\pi - \omega)})|
\]

This relationship between the magnitude response of \( g \) and that of \( h \) holds when \( g(n) = (-1)^n h(N-1-n) \) since time reversal does not effect the magnitude response of a filter.
Consider the lowpass to highpass transformation of a WSS source \( x(n) \) to a source \( y(n) \). The correlation of \( y \) is,

\[
R_{yy}(m) = E\{y(n+m)y^*(n)\} \\
= E\{(-1)^{m+n} x(n+m)(-1)^n x^*(n)\} \\
= (-1)^n R_{xx}(m)
\]

Thus the correlation sequence of \( y \) is the lowpass to highpass transform of the correlation sequence of \( x \). It follows that the PSD of \( y \) is that of \( x \) reflected about \( \pi/2 \) radians, when \( x \) is real.

### B.2 CODING GAIN AND BIT ALLOCATION

#### B.2.1. Background and Assumptions

In this section the average reconstruction error variance for an arbitrary perfect reconstruction subband codec is determined in terms of the subband quantisation error variances, the decimation/interpolation factors, and the synthesis filters. Various assumptions that are required to simplify the expression for the reconstruction error variance are listed. This expression is subsequently used in the main text of Chapter 3 to determine the optimum bit allocation among subbands. This optimum bit allocation is used in turn to determine a subband coding gain metric.

Following Figure 3.4 in Chapter 3 the output of each synthesis filters is,

\[
v_k(n) = I_k(u_k(n)) \\
= I_k(y_k(n)) - I_k(q_k(n))
\]

where \( I_k \) is a linear interpolation operator effecting the \( k^{th} \) upsampling and synthesis filtering operation. This operator is implicitly given by Malvar (1992, p91 equation 3.8) as,

\[
I_k(q_k(n)) = \sum_s g_k(n-sd_k)q_k(s)
\] (B.1)
For perfect reconstruction (PR), with no quantisation,

\[ x(n) = \hat{x}(n) = \sum_{k=0}^{M-1} y_k(n) = \sum_{k=0}^{M-1} I_k(y_k(n)) \]

which gives,

\[ r(n) \equiv x(n) - \hat{x}(n) = \sum_{k=0}^{M-1} I_k(y_k(n)) - \sum_{k=0}^{M-1} I_k(y_k(n)) - I_k(q_k(n)) \]
\[ = \sum_{k=0}^{M-1} I_k(q_k(n)) \]

In other words the reconstruction error is the synthesis of the subband quantisation error signals. Assuming a zero mean reconstruction error, the reconstruction error variance is,

\[ \sigma^2(n) \equiv E[|r(n)|^2] = E\left\{ \sum_{k=0}^{M-1} I_k(q_k(n)) \sum_{l=0}^{M-1} I^*_l(q_l(n)) \right\} \]
\[ = \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} E[I_k(q_k(n))I^*_l(q_l(n))] \]
\[ = \sum_{k=0}^{M-1} \left\{ E[I_k(q_k(n))]^2 \right\} + \sum_{k=0}^{M-1} \sum_{l \neq k} E[I_k(q_k(n))I^*_l(q_l(n))] \]

Using the (B.1) the synthesised quantisation error cross-correlation is,

\[ E[I_k(q_k(n))I^*_l(q_l(n))] = E\left\{ \sum_s g_k(n-d_k s)q_k(s) \sum_t g^*_l(n-d_l t)q^*_l(t) \right\} \]
\[ = \sum_s \sum_t g_k(n-d_k s)g^*_l(n-d_l t)E[q_k(s)q^*_l(t)] \]

(B.3)

where \( k \neq l \). Assuming the subband quantisation noise is uncorrelated between subbands, then,

\[ E[q_k(n)q_l(m)] = 0 \quad \forall k \neq l \]

and the cross-correlation given by (B.3) is zero. In this case the reconstruction variance from (B.2) is,
\[ \sigma_r^2(n) = \sum_{k=0}^{M-1} E \left[ |I_k(q_k(n))|^2 \right] \]  

(B.4)

Mintzer and Liu (1978, Appendix A) show that \( I_k(q_k(n)) \) is cyclo wide sense stationary (CWSS) with period \( d_k \) if \( q_k(n) \) is WSS. Setting \( L \) as the least common multiple of the decimation factors \( (d_k) \) gives,

\[ \sigma_r^2(n + L) = \sum_{k=0}^{M-1} E \left[ |I_k(q_k(n + L))|^2 \right] = \sum_{k=0}^{M-1} E \left[ |I_k(q_k(n))|^2 \right] \]

\[ = \sigma_r^2(n) \]

which shows that the reconstruction variance is periodic with period \( L \). In fact the reconstruction error signal is CWSS with period \( L \). The average reconstruction variance is thus defined as,

\[ \overline{\sigma_r^2} \equiv \frac{1}{L} \sum_{n=0}^{L-1} \sigma_r^2(n) = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} E \left[ |I_k(q_k(n))|^2 \right] \]  

(B.5)

Now from (B.3),

\[ E \left[ |I_k(q_k(n))|^2 \right] = \sum_s \sum_t g_k(n - d_k s) g_k^*(n - d_k t) E \{ q_k(s) q_k^*(t) \} \]

Assuming that the subband quantisation noise is white with zero mean, then,

\[ E \{ q_k(s) q_k^*(t) \} = \begin{cases} \sigma_{q_k}^2, & s = t \\ 0, & s \neq t \end{cases} \]

and,

\[ E \left[ |I_k(q_k(n))|^2 \right] = \sum_s g_k(n - d_k s) g_k^*(n - d_k s) \sigma_{q_k}^2 \]

Substituting this equation into (B.5) gives the average reconstruction variance as,
\[ \bar{\sigma}^2_r = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} E\left[ |I_k(q_k(n))|^2 \right] = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} \sum_s g_k(n-d_k s)^2 \sigma_{q_s}^2 \]
\[ = \frac{1}{L} \sum_{k=0}^{M-1} \sigma_{q_s}^2 \left( \sum_{n=0}^{L-1} \sum_s |g_k(n-d_k s)|^2 \right) \]

Setting \( m_k = \frac{L}{d_k} \) means that \( m_k \) is an integer, since \( L \) is a multiple of each of the \( d_k \).

The previous equation then reduces to,
\[ \bar{\sigma}^2_r = \frac{1}{L} \sum_{k=0}^{M-1} \sigma_{q_s}^2 \left( \sum_{n=0}^{L-1} \sum_s |g_k(n-d_k s)|^2 \right) = \frac{1}{L} \sum_{k=0}^{M-1} \sigma_{q_s}^2 \sum_{s} \left( \sum_{t=0}^{d_k-1} \sum_s |g_k(nd_k + t - d_k s)|^2 \right) \]
\[ = \sum_{k=0}^{M-1} \sigma_{q_s}^2 \frac{1}{L} \sum_{n=0}^{d_k-1} \left( \sum_s |g_k(s)|^2 \right) \]
\[ = \sum_{k=0}^{M-1} \sigma_{q_s}^2 \frac{1}{d_k} \sum_s |g_k(s)|^2 \]

Comparing with the definition of \( S_k \) in Chapter 3, equation (3.15), gives,
\[ S_k = \frac{1}{d_k} \sum_s |g_k(s)|^2 \]  
\[ \text{(B.6)} \]

That is \( S_k \) is given by the \( k^{th} \) synthesis filter energy weighted by the \( k^{th} \) decimation factor. The assumptions that have been made in this derivation are,

\[ \text{The input signal is a zero mean WSS signal} \]
\[ \text{The subband quantisation error is zero mean white noise and uncorrelated between subbands.} \]

\subsection*{B.2.2. Background and Assumptions: Orthogonal Filters}

In the previous subsection an expression for the reconstruction error variance was derived in terms of the subband quantisation error variances, the decimation/interpolation factors, and the synthesis filters. The above assumptions were used to simplify this expression. This subsection demonstrates that for an orthogonal \( M \)-band filter bank the latter assumption above can be relaxed to include jointly WSS subband quantisation noise signals. This result is then generalised to arbitrary (nonuniform) orthogonal subband analysis/synthesis systems.
To simplify the notation let \( w_k(n) \) be the response of the \( k^{th} \) upsampler and synthesis filter with input \( q_k(n) \). That is,

\[
w_k(n) = I_k(q_k(n))
\]

The average reconstruction error variance is given from equations (B.2) and (B.5) as,

\[
\bar{\sigma}_r^2 = \frac{1}{L} \sum_{n=0}^{L-1} \sigma_r^2(n) = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} E\{I_k(q_k(n))I_l^*(q_l(n))\}
\]

\[
= \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} \sigma_{w_k,w_l}(n) + \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} \sigma_{w_k,w_l}^2(n)
\]

where the synthesised quantisation error covariance,

\[
\sigma_{w_k,w_l}^2(n) \equiv E\{w_k(n)w_l^*(n)\} = E\{I_k(q_k(n))I_l^*(q_l(n))\}
\]

The average synthesised quantisation error covariance is denoted as,

\[
\bar{\sigma}_{w,\text{cross}}^2 = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} \sum_{l=0,l\neq k}^{M-1} \sigma_{w_k,w_l}^2(n)
\]

where the averaging is over time \( n \) and subbands. Note that this average is independent of the consecutive \( L \) samples of time over which the covariances are averaged. In the general case of the previous section, it was necessary to assume that the subband quantisation error cross correlation was zero to remove this term. However if the filters are orthogonal it is necessary only to assume that the subband quantisation errors are jointly wide sense stationary (WSS). To demonstrate this, substituting in (B.3) to the previous equation gives,

\[
\bar{\sigma}_{w,\text{cross}}^2 = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} \sum_{s} \sum_{t} g_k(n-d_k s)g_l^*(n-d_l t)E\{q_k(s)q_l^*(t)\}
\]

\[
= \sum_{k=0}^{M-1} \sum_{l=0,l\neq k}^{M-1} \left( \frac{1}{L} \sum_{n=0}^{L-1} \sum_{s} \sum_{t} g_k(n-d_k s)g_l^*(n-d_l t)E\{q_k(s)q_l^*(t)\} \right)
\]

\[
= \sum_{k=0}^{M-1} \sum_{l=0,l\neq k}^{M-1} \bar{\sigma}_{w_k,w_l}^2
\]
where $\sigma^2_{w_1, w_2}$ is the time averaged covariance of the $k^\text{th}$ and $l^\text{th}$ synthesised quantisation error and is given by,

$$
\sigma^2_{w_1, w_2} = \frac{1}{L} \sum_{n=0}^{L-1} \sum_{s} \sum_{t} g_k(n-d_k s)g_\ast_l(n-d_l t)E\left[q_k(s)q_\ast_l(t)\right]
$$

$$
= \frac{1}{L} \sum_{n=0}^{L-1} \sum_{s} \sum_{t} g_k(n-d_k s)g_\ast_l(n-d_l t)r_{q_k q_\ast_l}(s-t)
$$

which assumes that the $q_k(n)$ are jointly WSS. Setting $\alpha = s-t$ gives,

$$
\sigma^2_{w_1, w_2} = \frac{1}{L} \sum_{\alpha} r_{q_k q_\ast_l}(\alpha) \sum_{n=0}^{L-1} \sum_{s} g_k(n-d_k(\alpha+t))g_\ast_l(n-d_l t)
$$

Now consider the uniform filter bank. In this case the decimation factors are equal: that is $d_k = M$, the number of subbands. Also $L = M$ so that,

$$
\sigma^2_{w_1, w_2} = \frac{1}{M} \sum_{\alpha} r_{q_k q_\ast_l}(\alpha) \sum_{n=0}^{M-1} \sum_{s} g_k(n-dt-\alpha d)g_\ast_l(n-dt)
$$

$$
= \frac{1}{M} \sum_{\alpha} r_{q_k q_\ast_l}(\alpha) \sum_{s} g_k(s-\alpha d)g_\ast_l(s)
$$

$$
= \frac{1}{M} \sum_{\alpha} r_{q_k q_\ast_l}(\alpha) \delta(\alpha) \delta(k-l) \quad \text{for orthogonal filters}
$$

$$
= 0 \quad \forall k \neq l
$$

Orthogonal synthesis filters were defined in Chapter 2 and are also defined by Soman and Vaidyanathan (1993, equation (4.1)). Therefore the bracket term in (B.8), the time averaged covariance of the $k^\text{th}$ and $l^\text{th}$ synthesised quantisation error, is zero (since $k \neq l$) and hence the average cross correlation term is zero when $d_k = d$. Removing this term from (B.7) gives the average reconstruction error variance as,

$$
\sigma_r^2 = \frac{1}{M} \sum_{n=0}^{M-1} \sum_{k=0}^{M-1} \sigma^2_{w_1, w_2}(n) = \frac{1}{M} \sum_{n=0}^{M-1} \sum_{k=0}^{M-1} E\left[l_k(q_k(n))^2\right]
$$

$$
= \sum_{k=0}^{M-1} \left( \frac{1}{M} \sum_{n=0}^{M-1} \sum_{s} g_k(n-d_k s)g_\ast_k(n-d_k t)r_{q_k q_\ast_k}(s-t) \right)
$$

$$
= \sum_{M=0}^{M-1} \sigma^2_{w_1, w_2}
$$
Using (B.9) (for $k=t$) assuming an orthogonal filter bank gives,

$$
\hat{\sigma}_r^2 = \frac{1}{M} \sum_{k=0}^{M-1} \left( \sum_{\alpha} r_{q_1,q_2}(\alpha) \delta(\alpha) \delta(k-k) \right)
$$

$$
= \frac{1}{M} \sum_{k=0}^{M-1} r_{q_1,q_2}(0)
$$

(B.10)

Note that for an orthogonal filter bank the synthesis filters have unit energy.

Comparing with equation (3.15) in Chapter 3 gives for an orthogonal $M$-band (uniform) filter bank,

$$
S_k = \frac{1}{d_k} = \frac{1}{M}
$$

The only assumption required is that the subband quantisation error signals are jointly WSS.

An arbitrary nonuniform filter bank, with $M$ subbands, may be transformed into an equivalent uniform filter bank, with $L$ subbands, where $L$ is the lowest common multiple of the decimation factors ($d_k$). For example see Section 3.2.5 in Chapter 3 or Soman and Vaidyanathan (1993). The uniform subband filters can be arranged into $M$ groups each corresponding to a filter in the original nonuniform filter bank. There are $m_k = L/d_k$ filters in each group which are identical to within a delay. If the nonuniform filter bank is orthogonal then the corresponding uniform filter bank is orthogonal.

Assuming that the nonuniform filter bank subband quantisation noise is jointly WSS, the equivalent uniform filter bank subband quantisation noise is jointly WSS and hence the reconstruction error variance is given by equation (B.10). Also given this assumption, there will be groups of subbands with the same quantisation error variance. These groups correspond to the same subband groups discussed in the previous paragraph. Hence the reconstruction error variance can be rewritten as,

$$
\hat{\sigma}_r^2 = \frac{1}{L} \sum_{n=0}^{L-1} \sigma_{q_n}^2 = \sum_{k=0}^{M-1} \frac{m_k}{L} \sigma_{q_k}^2 = \sum_{k=0}^{M-1} \frac{1}{d_k} \sigma_{q_k}^2
$$
Finally comparing this with equation (3.15) in Chapter 3, one sees that for the nonuniform (general) orthogonal filter bank,

\[ S_k = \frac{1}{d_k} \]

Note that the only assumption that has been made in this case is that the subband quantisation noise is jointly WSS.

B.2.3. Background and Assumptions: Two-Dimensional Filter Banks

As noted in the main text of Chapter 3, the optimum bit allocation and coding gain is applicable to two-dimensional filter banks. However, if any assumptions are to be made concerning the \( S_k \) then special consideration of the two-dimensional nature is required. In this subsection an expression for the reconstruction error variance is determined as before in terms of the subband quantisation error variances, the decimation/interpolation factors, and the synthesis filters for two-dimensional filter banks.

The symbols used in the following analysis are the two-dimensional equivalent to those used in the first part of this appendix.

For a two-dimensional filter bank with \( M \) two-dimensional subbands the reconstruction error is,

\[ r(n,m) = \sum_{k=0}^{M-1} I_k(q_k(n,m)) \]

For zero mean sources the reconstruction variance is thus,
\[ \sigma^2_r(n,m) = E\left[ |r(n,m)|^2 \right] = E \left\{ \sum_{k=0}^{M-1} I_k(q_k(n,m)) \sum_{l=0}^{M-1} I_l^*(q_l(n,m)) \right\} \]

\[ = \sum_{k=0}^{M-1} \sum_{l=0}^{M-1} \sum_{s,t} g_k(n-d_{k_1},s,m-d_{k_2},t) q_k(s,t) \sum_{u,v} g_l^*(n-d_{l_1},u,m-d_{l_2},v) q_l^*(u,v) \]

where \( d_{k_1} \) and \( d_{k_2} \) are the decimation factors of the \( k^{th} \) subband in the vertical and horizontal directions (assumes separability). As in the one dimensional case it is assumed that the quantisation error noise is jointly WSS white noise: that is,

\[ E\{q_k(s,t)q_l(u,v)\} = \sigma^2_{q_k} \delta(k-l) \delta(s-u) \delta(t-v) \]

Under this assumption the reconstruction error is,

\[ \sigma^2_r(n,m) = \sum_{k=0}^{M-1} \sum_{s,t} g_k(n-d_{k_1},s,m-d_{k_2},t) g_l^*(n-d_{l_1},u,m-d_{l_2},v) \sigma^2_{q_k} \delta(k-l) \delta(s-u) \delta(t-v) \]

As in the one-dimensional case, if the subband quantisation noise is jointly WSS (as it is for white noise), \( r(n,m) \) is CWSS with period \( L_1 \) and \( L_2 \) in each dimension (\( L_1 \) is the least common multiple of the decimation factors in the vertical direction and similarly \( L_2 \) is the least common multiple of the decimation factors in the horizontal dimension). Therefore the average reconstruction variance is defined as,

\[ \overline{\sigma^2_r} = \frac{1}{L_1 L_2} \sum_{n=0}^{L_1-1} \sum_{m=0}^{L_2-1} \sigma^2_r(n,m) \]

and is independent of which \( (L_1,L_2) \) samples of \( \sigma^2_r(n,m) \) over which it is averaged. Substituting in (B.11) gives,
The $d_k$ used in the coding gain and bit allocation equations is,

$$d_k = d_k^A d_k^S$$

This is the ratio of input to output samples for the $k^{th}$ decimator. Hence normalising the synthesis filters to unit energy (unit Euclidean norm) gives,

$$S_k = \frac{1}{d_k}$$

as in the one dimensional case. To conclude: the bit allocation, ideal and practical, and coding gain equations, are applicable to two-dimensional filter banks under the assumption of jointly WSS white subband quantisation noise. As discussed in the main text of Chapter 3 this assumption is valid in general only for high rates. However, in the orthogonal case only jointly WSS quantisation error signals are required, which is a reasonable assumption regardless of the rate.

### B.3 RATE CONSTRAINED CODING GAIN

#### B.3.1. Asymptotic Performance of the Rate Constrained Coding Gain: Ideal Octave-band Analysis

De Queiroz and Malvar (1992) derived the ideal coding gain for an ideal octave-band or dyadic A/S system, and derived the asymptotic ideal coding gain as the tree-depth increases to infinity. In this section the rate constrained coding gain for an ideal dyadic analysis/synthesis system, using an AR(1) source model, is considered. The coding gain is shown to decrease with rate. For a fixed tree-depth, $S$, and for certain rates, the rate constrained coding gain is calculated as a fraction of the ideal gain. As a consequence the rate constrained coding gain is asymptotic to decreasing levels as the rate decreases. The value of the asymptote for certain rates can be derived using these results.
From Chapter 3, for a unit variance input source, the ideal gain for a dyadic analysis/synthesis is,

$$G_{DYAD} = \frac{1}{\sigma_{WGM}^2}$$  \hspace{1cm} (B.12)

where the weighted geometric mean of the subband variances is,

$$\sigma_{WGM}^2 = \left(\sigma_{\text{low band}}^2\right)^{\frac{1}{2^{N-1}}} \prod_{k=0}^{M-2} \left(\sigma_k^2\right)^{\frac{1}{2^{4-k}}}$$

and $\sigma_{\text{low band}}^2$ is the variance of the DC subband. The other subbands are indexed in terms of decreasing frequency bands. The motivation for such nomenclature, somewhat contrary to the nomenclature used in Chapter 3, is that the subbands labels are independent of the tree-depth. For example $\sigma_0^2$ always refers to the variance of the highpass subband.

Now, consider the ideal dyadic analysis of an AR(1) source with positive correlation. Since the PSD is decreasing with increasing frequency, the subband variances are increasing with increasing index (increasing index means decreasing frequency). Consider a rate, $R$, so that,

$$\alpha + \frac{1}{2} \log_2 \frac{\sigma_N^2}{\sigma_{WGM}^2} = 0$$  \hspace{1cm} (B.13)

noting that the practical optimum bit allocation is,

$$b_k = \max\left\{ \alpha + \frac{1}{2} \log_2 \frac{\sigma_k^2}{\sigma_{WGM}^2}, 0 \right\}$$

In this case subbands 0,1,...,$N$ have an optimum practical bit allocation of zero bits, while the other subbands have a positive allocation since $\sigma_k^2 < \sigma_N^2$ for $k < N$ and $\sigma_k^2 > \sigma_N^2$ for $k > N$.

For simplicity it is assumed that $\epsilon,^2 = 1$. Using equation (3.33) from Chapter 3, the reconstruction error variance is,
where subband $M-1$ refers to the lowband. Note that subbands 0 to $N$ are allocated zero bits, while subbands $N+1$ to $M-1$ are allocated a positive number of bits.

Now from (B.13),

$$2^{-2\alpha} = \frac{\sigma^2_N}{\sigma^2_{WGM}}$$

which gives,

$$\sigma^2_{r,DYAD} = \sum_{k=0}^{N} \frac{\sigma_k^2}{d_k} + \left(1 - \sum_{k=0}^{N} \frac{1}{d_k}\right) \sigma_N^2$$

To determine the rate required for (B.13) to hold, consider the rate constraint,

$$R = \sum_{k=0}^{M-1} b_k$$

$$= \sum_{k=N+1}^{M-1} \alpha + \frac{1}{2} \log_2 \frac{\sigma_k^2}{\sigma_{WGM}^2} d_k$$

$$= \left(1 - \sum_{k=0}^{N} \frac{1}{d_k}\right) \alpha + \frac{1}{2} \sum_{k=N+1}^{N} \log_2 \frac{\sigma_k^2}{\sigma_{WGM}^2} d_k$$

Coupled with (B.13) this gives,

$$R = -\frac{1}{2} \log_2 \frac{\sigma_N^2}{\sigma_{WGM}^2} \left(1 - \sum_{k=0}^{N} \frac{1}{d_k}\right) + \frac{1}{2} \sum_{k=N+1}^{M-1} \log_2 \frac{\sigma_k^2}{\sigma_{WGM}^2} d_k$$

(B.15)
Hence,
\[
2^{-2R} = \left( \frac{\sigma_N^2}{\sigma_{WGM}^2} \right)^{\left(1 - \sum_{k=0}^{N} \frac{d_k}{d_{k+1}} \right)} \prod_{k=N+1}^{M-1} \left( \frac{\sigma_{WGM}^2}{\sigma_k^2} \right)^{\frac{1}{d_k}} \\
= \frac{\sigma_N^2}{\sigma_{WGM}^2} \prod_{k=0}^{N} \left( \frac{\sigma_k^2}{\sigma_N^2} \right)^{\frac{1}{d_k}} \prod_{k=0}^{M-1} \left( \frac{\sigma_{WGM}^2}{\sigma_k^2} \right)^{\frac{1}{d_k}} \\
= \frac{\sigma_N^2}{\sigma_{WGM}^2} \prod_{k=0}^{N} \left( \frac{\sigma_k^2}{\sigma_N^2} \right)^{\frac{1}{d_k}}
\]

Finally combining this equation with (B.14) and setting the input variance to unity, the rate constrained coding gain is (remembering that \( \epsilon_k^2 = 1 \)),
\[
G_{SBC} = \frac{\sigma_{r,PCM}^2}{\sigma_{r,SBC}^2} = 2^{-2R} \frac{\sigma_x^2}{\sigma_{r,SBC}^2} \\
= \frac{\sigma_N^2}{\sigma_{WGM}^2} \prod_{k=0}^{N} \left( \frac{\sigma_k^2}{\sigma_N^2} \right)^{\frac{1}{d_k}} \\
= \frac{\sum_{k=0}^{N} \sigma_k^2 + \left(1 - \sum_{k=0}^{N} \frac{1}{d_k} \right) \sigma_N^2}{\sum_{k=0}^{N} d_k}
\]

This can be rearranged to,
\[
G_{SBC} = \frac{\prod_{k=0}^{N} (\sigma_k^2)^{\alpha_k}}{\sum_{k=0}^{N} \alpha_k \sigma_k^2} \frac{1}{\sigma_{WGM}^2} \tag{B.16}
\]

where,
\[
\alpha_k = \begin{cases} 
\frac{1}{d_k} & k = 0, 1, \ldots, N-1 \\
\left(1 - \sum_{k=0}^{N-1} \frac{1}{d_k} \right) & k = N
\end{cases}
\]

Finally note that,
since the subband analysis/synthesis is a critically sampled system.

The rate constrained coding gain is given as a fraction of the ideal coding gain. The fraction is determined by the ratio of weighted geometric mean of subband variances that are allocated zero bits to the weighted arithmetic mean. This ratio is upper bound by one, and equality occurs when all the variances are equal. As \( N \) tends to infinity, the denominator of this ratio tends to the signal variance (normalised to one) and the numerator tends to \( \sigma^2_{WGM} \). Hence as \( N \) tends to infinity the gain tends to unity. This is expected since as \( N \) tends to infinity the rate is implicitly tending to zero. For an AR(1) source of correlation \( \rho=0.95 \) the fraction of the ideal gain is 1, 0.8790, 0.6127, 0.3697, 0.2157, 0.1392, and 0.1018 for \( N=0,1,2,3,4,5,6 \) respectively. The gain decreases quite significantly even for low \( N \).

In the above derivation it was assumed that \( \varepsilon^2 = 1 \). In practice \( \varepsilon^2 \) will usually be greater than one. In this case these gain estimates will be lower than that attained in practice. Nevertheless these estimated gain curves serve as a guide to the general trends as the rate decreases. As stated above the gain curves are significantly lower than that of the ideal gain, when some subbands are allocated zero bits.

De Queiroz and Malvar (1992) derived the coding gain for an ideal dyadic A/S system where an ideal bit allocation is assumed, and derived the asymptotic ideal coding gain as the tree-depth increases to infinity (using an AR(1) source). The asymptotic rate constrained coding gain can be calculated as a fraction of the asymptotic ideal coding gain, using the results of section.

### B.4 TRANSFORM FREQUENCY DOMAIN CHARACTERISTICS

#### B.4.1. Pair-Wise Symmetric Transforms and Symmetric Coding Gain

In this section it is shown that the subband variances for an \( M \)-band analysis/synthesis system can be determined using a matrix vector product. Also it is shown that \( M \)-band A/S systems with (time-reversed) pairwise symmetric analysis filters have a coding
gain that is independent of the sign of the correlation coefficient, $\rho$, for an AR(1) source.

Consider a general parallel subband analysis with $M$ subbands. From Chapter 3 or Papoulis (1977 p321), for zero mean sources the variance of the $k^{th}$ subband is,

$$\sigma_k^2 = r_{y_k}(0) = r(n)*h(n)*h^*(-n)\big|_{n=0}$$

where $r(n)$ is the correlation sequence of the input source. This variance equation may be written as,

$$\sigma_k^2 = \sum_p r(p) \sum_n h_k(n)h_k^*(n+p) = \sum_p r(p)s_k(p)$$

Let $N$ denote the length of the longest analysis filter. Then $s_k(p)$ is zero for $|p| > N-1$ since $h_k(n)h_k^*(n+p) = 0$ for $|p| > N-1$. Also

$$s_k(-p) = \sum_n h_k(n)h_k^*(n-p)$$
$$= \sum_q h_k(q+p)h_k^*(q)$$
$$= s_k^*(p)$$

Therefore, since $r(p) = r(-p)$ for a real source,

$$\sigma_k^2 = r(0)s_k(0) + \sum_{p=1}^{N-1} r(p)[s_k(p) + s_k(-p)] = \sum_{p=0}^{N-1} r(p)v_k(p)$$

where,

$$v_k(p) = \begin{cases} 2\text{Re}\{s_k(p)\} = 2\text{Re}\left\{\sum_n h_k(n)h_k^*(n+p)\right\} & p \neq 0 \\ s_k(0) = \sum_n h_k(n)h_k^*(n) & p = 0 \end{cases}$$

Collecting the subband variances into a vector, $\sigma^2$, means that the variances can be determined using a matrix equation as,

$$\sigma^2 = V^T r_{xx}$$
where $r_{xx}$ is a vector of the correlation samples, $r_{xx} = [r(0) \ r(1) \ \ldots \ r(N-1)]'$ and the $V$ matrix entries are defined as,

$$v_{pk} = v_k(p) = (2 - \delta_{p,0}) \text{Re} \left\{ \sum_n h_k(n) h_k^*(n+p) \right\}$$

Consider a PWS analysis filter set where $h_k(n) = (-1)^n h_{M-1-k}(n)$. Then,

$$v_{p,M-1-k} = (2 - \delta_{p,0}) \text{Re} \left\{ \sum_n h_{M-1-k}(n) h_{M-1-k}^*(n+p) \right\}$$

$$= (2 - \delta_{p,0}) \text{Re} \left\{ \sum_n (-1)^n h_k(n)(-1)^{n+p} h_k^*(n+p) \right\}$$

$$= (-1)^p v_{p,k}$$

Thus $V$ has PWS columns. It follows that the coding gain of such a filter bank is independent of the sign of $p$ for an AR(1) source. If the analysis filters are time-reversed PWS then $h_k(n) = (-1)^n h_{M-1-k}(N-1-n)$ and,

$$v_{p,M-1-k} = (2 - \delta_{p,0}) \text{Re} \left\{ \sum_n h_{M-1-k}(n) h_{M-1-k}^*(n+p) \right\}$$

$$= (2 - \delta_{p,0}) \text{Re} \left\{ \sum_n (-1)^n h_k(N-1-n)(-1)^{n+p} h_k^*(N-1-n-p) \right\}$$

$$= (-1)^p (2 - \delta_{p,0}) \text{Re} \left\{ \sum_q h_k(q+p) h_k^*(q) \right\}$$

$$= (-1)^p v_{p,k}$$

Hence again $V$ has PWS columns and the filter bank coding gain is independent of the sign of $p$. 
C.1. INTRODUCTION

In this appendix some of the mathematical detail required for the material presented in Chapter 4 is given. The title of each section or subsection in this appendix corresponds to the section in Chapter 4 of the same title.

C.2. LENGTH FOUR OPTIMUM FILTERS

In this section the eigenvalues of the 4x4 modified correlation matrix, used in the design of the optimum 4-tap two-band filters, are derived analytically. Also it is shown that only the minimum or maximum eigenvalue can be forced to have multiplicity greater than one for real λ₂.

For the 4-tap optimum filter design example given in the main text of Chapter 4 (Section 4.3.1),

\[ A = R - \lambda_2 W = \begin{bmatrix} 1 & \rho & \rho^2 - \lambda_2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 - \lambda_2 \\ \rho^2 - \lambda_2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 - \lambda_2 & \rho & 1 \end{bmatrix} \]

Using equation (4.1) in the main text of Chapter 4, the eigenvalues of A are the eigenvalues of

\[ \begin{bmatrix} 1 - \rho^3 & \rho - \rho^2 + \lambda_2 \\ \rho - \rho^2 + \lambda_2 & 1 - \rho \end{bmatrix} \text{ and } \begin{bmatrix} 1 + \rho^3 & \rho + \rho^2 - \lambda_2 \\ \rho + \rho^2 - \lambda_2 & 1 + \rho \end{bmatrix} \]

which are calculated as,
\[ e_{a,b} = \frac{1}{2} \left( 2 + \rho + \rho^3 \pm \sqrt{8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 - 8\rho \lambda_2 - 8\rho^2 \lambda_2} \right) \]

\[ e_{c,d} = \frac{1}{2} \left( 2 - \rho - \rho^3 \pm \sqrt{-8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 + 8\rho \lambda_2 - 8\rho^2 \lambda_2} \right) \]

The eigenvalues denoted by \( e_a \) and \( e_b \) correspond to skew symmetric eigenvectors and those denoted by \( e_c \) and \( e_d \) correspond to symmetric eigenvectors. For real \( \lambda_2 \) these eigenvalues are real since \( A \) is a symmetric (and in fact SC) matrix.

Consider \( e_a = e_b \). This implies that the square root term is zero: that is,

\[ 8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6 + 4\lambda_2^2 - 8\rho \lambda_2 - 8\rho^2 \lambda_2 = 0 \]

This is a quadratic in \( \lambda_2 \), \( a\lambda_2^2 + b\lambda_2 + c = 0 \), where the discriminant,

\[ b^2 - 4ac = (8\rho + 8\rho^3)^2 - 16(8\rho^3 + 5\rho^2 + 2\rho^4 + \rho^6) \]

Simply algebraic manipulation reveals,

\[ b^2 - 4ac = -16\rho^2(\rho^2 - 1)^2 < 0, \quad (\rho \neq 0) \]  

(C.1)

meaning that the discriminant is less than zero. Therefore for real \( \lambda_2 \) it is impossible to force a zero square root term, and hence impossible to force \( e_a = e_b \). The same method can be used to show that it is impossible to enforce \( e_c = e_d \). Further it is immediately evident, noting that the eigenvalues are real, that it is impossible to enforce \( e_a = e_d \) and \( e_b = e_c \). Therefore to force a repeated eigenvalue, there are only two solutions: namely \( e_a = e_c \) giving a repeated maximum eigenvalue, and \( e_b = e_d \) giving a repeated minimum eigenvalue. In both cases the repeated eigenvalue will have a corresponding symmetric and skew symmetric eigenvector. This means that the corresponding eigenspace contains asymmetric vectors and possibly a vector that obeys the PR constraints.

C.3. ZERO CONSTRAINED FILTERS: OPTIMUM WAVELETS

In this section, using the general basis matrix \( B \) introduced in Chapter 4, it is shown that the matrix product \( B^TQB \) is a symmetric centrosymmetric (SC) matrix when \( Q \) is SC. Also, using the subspace constraint \( h = Bx \), it is shown that if \( x \) is symmetric then so is \( h \). Finally it is demonstrated that it is convenient to use a basis matrix \( B \) with orthonormal columns.
Consider a vector \( \mathbf{b} \) of length \( K \) and a basis matrix \( \mathbf{B} \) of dimension \( N \times (N-K+1) \) of the form,

\[
\mathbf{B} = \begin{bmatrix}
    b(0) & 0 & 0 \\
    b(1) & b(0) & \ddots \\
    \vdots & b(1) & \ddots \\
    b(K-1) & \ddots & \ddots & 0 \\
    0 & b(K-1) & \ddots & b(0) \\
    \vdots & \ddots & \ddots & \ddots \\
    0 & \ddots & \ddots & 0 \\
    0 & \ddots & \ddots & 0 \\
\end{bmatrix}
\]

\( \mathbf{B} \) has full rank \( (N-K+1) \) since the columns are linearly independent. Consider the row reflection effected by pre-multiplication with an anti-diagonal matrix of ones, \( \mathbf{J} \),

\[
\mathbf{J}_{(N,N)} \mathbf{B} = \begin{bmatrix}
    0 & 0 & 0 & b(K-1) \\
    \ddots & \ddots & \ddots & \ddots \\
    \ddots & 0 & \ddots & b(1) \\
    \vdots & 0 & \ddots & \ddots & b(0) \\
    b(K-1) & \ddots & \ddots & \ddots & 0 \\
    \ddots & b(1) & \ddots & \ddots & \ddots \\
    b(1) & b(0) & \ddots & \ddots & \ddots \\
    b(0) & 0 & \ddots & \ddots & \ddots \\
\end{bmatrix}
\]

and the column reflection effected by the post-multiplication by \( \mathbf{J} \),

\[
\mathbf{B} \mathbf{J}_{(N-K+1,N-K+1)} = \begin{bmatrix}
    0 & 0 & 0 & b(0) \\
    \ddots & \ddots & \ddots & b(1) \\
    \ddots & 0 & \ddots & \ddots \\
    0 & b(0) & \ddots & \ddots & b(K-1) \\
    b(0) & b(1) & \ddots & \ddots & 0 \\
    b(1) & \ddots & \ddots & \ddots & \ddots \\
    \ddots & b(K-1) & \ddots & \ddots & \ddots \\
    b(K-1) & 0 & \ddots & \ddots & \ddots \\
\end{bmatrix}
\]
The \( J \) matrix subscripts refer to the matrix dimension. Henceforth these subscripts are dropped since the appropriate dimension of \( J \) can be ascertained from the relevant equation. If \( b \) is symmetric then,

\[
JB = BJ, \quad JB^T = B^TJ
\]  
(C.2)

Alternatively if \( b \) is skew symmetric then,

\[
JB = -BJ, \quad JB^T = -B^TJ
\]  
(C.3)

Consider a SC square matrix \( Q \) and the product matrix \( B^TQB \). If \( B \) satisfies (C.2) or (C.3) one has,

\[
J(B^TQB)J = (B^TQ)J(B) = B^TQB
\]  
(C.4)

which shows that the product matrix is SC. By inspection this product matrix is also symmetric and hence is SC. Also consider,

\[
h = Bx
\]

If \( x \) and \( b \) are both symmetric or both skew symmetric then,

\[
h = BJx = JBx = Jh
\]

ie. \( h \) is symmetric. If one of \( b \) and \( x \) is symmetric and the other skew symmetric then \( h \) is skew symmetric.

In the main text of Chapter 4 it is shown that the optimum solution is given by the eigenvector equation,

\[
(B^TB)^{-1}B^TAX = \lambda, x
\]

From (C.4) the product matrix \( B^TQB \) is SC where \( Q \) is SC. Since \( A \) and \( I \), the identity matrix, are SC then,

\[
[(B^TIB)^{-1}]B^TAB = [(B^TB)^{-1}]B^TAB
\]
is SC, being the product of two SC matrices. (Noting that the inverse of a SC matrix is also SC) As a consequence, if the eigenvalues of this matrix are distinct then \( \mathbf{x} \) is symmetric. If follows that \( \mathbf{h} \) is symmetric, contrary to the theorem that symmetric \( \mathbf{h} \) can't satisfy the PR equations. Therefore, one concludes as in the unconstrained case, it is necessary to enforce a multiplicity of eigenvalues.

It is often convenient to use a basis matrix \( \mathbf{B}_o \) with orthonormal columns so that \( \mathbf{B}_o^T \mathbf{B}_o = \mathbf{I} \). Using the subspace constraint \( \mathbf{h}_o = \mathbf{B}_o \mathbf{y} \), the optimum eigen-equation becomes

\[
\mathbf{B}_o^T \mathbf{A} \mathbf{B}_o \mathbf{y} = \lambda \mathbf{y}
\]

Solving this equation gives the same solution as that obtained using the original \( \mathbf{B} \) matrix since,

\[
\lambda \mathbf{h} = \mathbf{B} \lambda \mathbf{x} = \mathbf{B} \left( \mathbf{B}^T \mathbf{B} \right)^{-1} \mathbf{B}^T \mathbf{A} \mathbf{B} \mathbf{x}
\]
\[
= \left[ \mathbf{B} \left( \mathbf{B}^T \mathbf{B} \right)^{-1} \mathbf{B}^T \right] \mathbf{A} \mathbf{h}
\]
\[
\lambda \mathbf{h}_o = \mathbf{B}_o \lambda \mathbf{y} = \mathbf{B}_o \mathbf{B}_o^T \mathbf{A} \mathbf{B}_o \mathbf{y}
\]
\[
= \left[ \mathbf{B}_o \mathbf{B}_o^T \right] \mathbf{A} \mathbf{h}_o
\]

In other words \( \mathbf{h} \) and \( \mathbf{h}_o \) are the eigenvectors of the projection of \( \mathbf{A} \) onto the subspace spanned by the columns of \( \mathbf{B} \) (See Strang 1988 p158). That is \( \mathbf{B}_o \mathbf{B}_o^T = \mathbf{B} \left( \mathbf{B}^T \mathbf{B} \right)^{-1} \mathbf{B}^T \) is the projection matrix for this subspace. The eigenvectors corresponding to the zero eigenvalues in the above eigen-equations, are the eigenvectors that are not in the subspace, and hence are not feasible solutions.

Using an orthogonal basis it is again necessary in general to enforce a minimum eigenvalue (of \( \mathbf{B}_o \mathbf{A} \mathbf{B}_o \)) of multiplicity \( N/2 \). To prove this consider a vector \( \mathbf{x}_* \) that is the globally optimum solution using the original subspace constrained method. In general \( \mathbf{x}_* \) will correspond to a minimum eigenvalue of \( \left( \mathbf{B}^T \mathbf{B} \right)^{-1} \mathbf{B}^T \mathbf{A} \mathbf{B} \) of multiplicity \( N/2 \). Label this minimum eigenvalue as \( \lambda_{\text{min}} \) which is also the global minimum of the cost function under the PR constraints. Now consider \( \mathbf{y}_* \) so that,

\[
\mathbf{B}_o^T \mathbf{A} \mathbf{B}_o \mathbf{y}_* = \lambda_{\text{min}} \mathbf{y}_*
\]
where $\lambda_{\text{min}}$ is the minimum eigenvalue of $B_o^T A B_o$. Since $B_o^T A B_o$ and $\left(B^T B\right)^{-1} B^T A B$ are similar matrices (see below) then $\lambda_{\text{min}} = \lambda_{\text{min}}$ and is of multiplicity $N/2$. Also $x_* = M^{-1} y_*$ where $M$ is the similarity transformation matrix.

The cost associated with the vector $h_o$ is given by,

$$y_*^T B_o^T A B_o y_* = \lambda_{\text{min}}$$

which is the globally minimum cost. Also $h_o = B_o y_* = B_o M x_* = B x_* = h$, so that if $x_*$ satisfies the PR constraints so does $y_*$. Therefore $y_*$ is the desired solution point, where it is the eigenvector of $B_o^T A B_o$ corresponding to the minimum eigenvalue of multiplicity $N/2$.

It remains to show that the matrices $B_o^T A B_o$ and $\left(B^T B\right)^{-1} B^T A B$ are similar. Since $B$ and $B_o$ have the same column space (same subspace constraint) one can write,

$$B = B_o M$$

where $B$ and $B_o$ are of dimension $N \times M$ and $M$ is of dimension $M \times M$, noting that $N > M$. From Strang (1988 p201),

$$r(B_o M) \leq r(M)$$

where $r(A)$ is the rank of matrix $A$. Since $B$ (and $B_o$) is of rank $M$ (being a basis matrix) it follows that,

$$r(B_o M) = r(B) = M \leq r(M)$$

Obviously $r(M)$ can't be greater than $M$ so that $r(M) = M$. Therefore $M$ is invertible. Now consider the substitution (C.5) into the matrix equation $\left(B^T B\right)^{-1} B^T A B$,

$$\left(B^T B\right)^{-1} B^T A B = \left(M^T B_o^T B_o M\right)^{-1} M^T B_o^T A B_o M$$

$$= M^{-1} \left(M^T\right)^{-1} M^T B_o^T A B_o M$$

$$= M^{-1} B_o^T A B_o M$$
Hence the matrices \( (B^T B)^{-1} B^T A B \) and \( B_0^T A B_0 \) are similar. Similar matrices share the same eigenvalues and the eigenvectors are related through a linear transformation by \( M \) [Stang 1988 p304].

### C.4. MATLAB M-FILES TO IMPLEMENT FILTER DESIGN

In this section the MATLAB M-files, version 3.5, used to implement the 5 steps of the maximum gain CQF design algorithm described in the main text of Chapter 4 are listed. Each subsection corresponds to one step of the algorithm, and gives the listing for the appropriate M-file. The last subsection lists various auxiliary function used in the following functions.

#### C.4.1 Step 1.

```matlab
%ewvN_i
%MAXIMUM CODING GAIN CQF DESIGN FUNCTION (STEP 1: INTIALISATION)
%Global parameter initialisation for maximum coding gain CQF design
%using subspace constraint(eigen-wavelet, length N initialisation)
%Default subspace is that of zero DC vectors. If no subspace constraint
%is required set g_B to be identity matrix. Note that g_B is assumed
%to have orthonormal columns. ie use g_B = orth(g_B);
%
%Usage:
%  ewvN_i user is prompted for filer length and correlation coefficient
%  Change g_B as required for different subspace solutions.
%(ie g_B = Identity matrix for no subspace constraint)
%
%Input:
%  g_N  filter length
%  g_rho AR(1) source correlation coefficient
%
%Output:
%  g_R  AR(1) source correlation matrix
%  g_W  Overlapping orthogonal matrix
%  g_B  Basis for subspace desired. Set to identity if want
%        unconstrained solution.
%  g_V  Initialise eigenvector matrix
%
%Note: the prefix g_ indicates a global variable
%  Set g_B (after running ewvnz_i) as required for an
%  arbitrary subspace constrained solution. Note that it is
%  assumed that g_B has orthonormal columns.
%
%To design the filter, the following 5 steps as outlined in Thesis are
%required.
%  >>ewvN_i;
%  >>l = fsolve('ewvnz_sr',l_initial_guess);
%  >>[g_V,d] = ewvnz_me(l);
```
```matlab
>> w = fsolve('ewvn-sw', w_initial_guess);
>> [h0,h1] = ewvn_smk(w);

% h0, h1 are the maximum and minimum coding gain CQF filters. Note that the
% length of the l (and w) vector is g_N/2 - 1. (Lambda 2 -> Lambda N/2).
% To display results of fsolve iterations use fsolve('fn', init_guess, 1);
%
% Recommendations:
% l_initial_guess = (rand(g_N/2-1,1)-1)/1024; % (default uniform rand)
% w_initial_guess = rand(g_N/2-1,1)-0.5;

g_N = input('Enter filter length, N >');
g_rho = input('Enter the correlation factor >');
g_W = wmake(2, g_N/2);
g_R = covarel(g_rho, g_N);
g_W = wmake(2, g_N/2);
g_B = dct(g_N);
g_B = g_B(:, 2:g_N); % Use orthogonal DCT (excluding DC) basis vectors
  % as a basis for DEFAULT zero DC vector subspace.

C.4.2 Step 2

function y = ewvNz_sl(l);
% function y = ewvNz_sl(l);
% MAXIMUM CODING GAIN CQF DESIGN FUNCTION (STEP 2)
% Use with fsolve to find Lagrange multipliers for maximum coding gain
% subspace constrained CQF filters. Use ewvnz_i to initialise design
% parameters. Set g_B (subspace) as required, noting that is assumed that
% the columns of this matrix are orthonormal. The default g_B is for a single
% highpass filter zero at DC, as initialised in ewvNz_i.m
%
% Determines the square deviation of the N/2 minimum eigenvalues of the
% subspace constrained modified correlation matrix g_B*A*g_B from the minimum
% eigenvalue. (Returns zero, when the minimum eigenvalue has multiplicity N/2)
% A is g_R-l(1)W2-l(2)W2 ..- l(N/2)WN/2, modified correlation matrix
%
% Usage:
%  % l_out = fsolve('ewvNz_sl', initial_guess);
%  %
%  % Input:
%  %   l   initial guess of Lagrange multipliers
%  %
%  % Output:
%  %   y   square error sum of difference between min eigenvalue (of g_B*A*g_B)
%  %       and the next g_N/2-1, minimum eigenvalues
%  %
%  % See ewvNz_i.m
```

**APPENDIX C: EIGENWAVELETS**

250


\%g_R, g_W, g_N, g_B are (or need to be !) global. (see ewvN_i.m)

L = g_N / 2 - 1;
y = zeros(L,1);
A = g_R;
W = g_W;
for n = 1:L
    A = A - l(n)*(W+W);
    W = W*g_W;
end
Rp = g_B*A*g_B;

e = eig(Rp);
e = sort(e);
a = e(2:length(e));
a = abs(a(1:L)-e(1:L)); \% force multiple at minimum eigenvalue
y = a(1:L)*1E5;
disp(y'*y);

C.4.3 Step 3

function [V,d,Rp] = ewvNz_me(l);
\%function [V,d,{Rp}] = ewvNz_me(l);
\%MAXIMUM CODING GAIN CQF DESIGN FUNCTION (STEP 3)
\%Given the Lagrange multipliers that force a multiplicity of N/2
\%minimum eigenvalues of the subspace constrained modified correlation matrix
\%(g_B'* A*g_B), this function returns the associated eigenvectors,
\% eigenvalues and matrix Rp = g_B'*A*g_B;
\%
\%Usage:
\%   [V,d,Rp] = ewvn_me(l);
\%
\%Input:
\% 1  vector of feasible Lagrange multipliers, so that the modified
\%  correlation matrix has a multiplicity of N/2 minimum eigenvalues
\%
\%Output:
\% V  matrix of associated minimum eigenvectors
\% d  vector of all eigenvectors of Asoln
\% Rp  (g_B*A*g_B) solution modified correlation matrix
\%
\%See ewvNz_i.m

\%g_R, g_W, g_N, g_B are (or need to be !) global. (see ewvNz_i.m)

L = g_N / 2 - 1;
[Mb,Lb] = size(g_B);
A = g_R;
W = g_W;
for n = 1:L
    A = A - l(n)*(W+W);
    W = W*g_W;
end
Rp = g_B*A*g_B;
\[
\begin{align*}
\text{[g_V, d]} &= \text{eig}(R_p); \\
\text{d} &= \text{diag}(d); \\
\text{find the } g_{N/2} \text{ eigenvectors corresponding to the minimum repeated eig.} \\
\text{d_temp} &= d; \\
V &= \text{zeros}(L_b, g_{N/2}); \% \text{N-1 D vectors due to zero DC constraint} \\
\text{for } n = 1:g_{N/2}; \\
\text{[min_d, index]} &= \text{min}(d_temp); \\
V(:, n) &= g_V(:, index); \\
d_temp(index) &= \text{inf}; \\
\text{end} \\
V &= g_B \ast V; \\
\end{align*}
\]

**C.4.4 Step 4**

function cost = ewvn_sw(w);
%
function cost = ewvn_sw(w);
%MAXIMUM CODING GAIN CQF DESIGN FUNCTION (STEP 4)
%Use with fsolve to find the weights associated with the g_{N/2} eigenvectors
%corresponding to the minimum eigenvalue of the modified correlation matrix (A)
%such that the weighted sum gives a vector that obeys the g_{N/2} PR
%constraints.
%
%An eigenvector h is determined as the weighted (weight vector w) sum of
%eigenvectors. The PR constraints should be h'Wnh == 0. This function
%determines the sum of square deviations from 0 of these h'Wnh PR equations
%
%Usage:
%  w_out = fsolve('ewvn_sw', initial_weight_guess); 

%Input:
%  w  weight vector
%
%Output:
%  cost  sum of square deviations from zero of PR constraints
%
%See ewvn_i.m
%
%g_V, g_W are global;
\[
L = g_{N/2} - 1; \\
\text{cost} &= \text{zeros}(L, 1); \\
h &= g_V(:, 1); \\
\text{for } n = 1:L \\
& \quad h = h + w(n) \ast g_V(:, n+1); \\
\text{end} \\
W &= g_W; \\
\text{for } n = 1:L \\
& \quad \text{cost}(n) = h' \ast W \ast h; \\
& \quad W = W \ast g_W; \\
\text{end} \\
\]
APPENDIX C: EIGENWAVELETS

\[ \text{cost} = \text{cost} \times 1E03/(h'*h); \]
\[ g\_weight = w; \]
\[ \text{disp} \left( \text{cost} \times \text{cost} \right); \]

### C.4.5 Step 5.

```matlab
function [h0,h1] = ewvn_smk(w);

%function [h0,h1] = ewvn_smk(w);
%MAXIMUM CODING GAIN CQF DESIGN FUNCTION (STEP 5)
%Combines \( g_N/2 \) eigenvectors corresponding to the minimum eigenvalue of
%multiplicity \( g_N/2 \) of the modified correlation matrix \( A \), using the weight
%vector \( w \) to give the maximum coding gain CQF highpass filter \( h_1 \).
%
%Input:
% \( w \)     weight vector for eigvectors
%
%Output:
% \( h_0 \)  lowpass maximum gain CQF
% \( h_1 \)  highpass maximum gain CQF
%
%See ewvn_ij.m

%g_V, g_W are global;

L = g_N/2 - 1;

h = g_V(:,1);
for n = 1:L
    h = h + w(n)*g_V(:,n+1);
end

h = h ./ sqrt(h'*h);

h0 = lp_hp(h(length(h):-l:l));
h1 = h*sign(sum(h0));
h0 = h0*sign(sum(h0));
```

### C.4.6 Auxiliary Functions.

```matlab
function Rxx = covarl(rho,N);

%function Rxx = covarl(rho,N);
% This function constructs the \( N \times N \) covariance matrix of an AR(1)
% source where \( p = \rho \) (correlation).

Rxx = zeros(N);

for i = 1:N
    for j = 1:N
        Rxx(i,j) = rho^abs(i-j);
    end;
end;

% End of covarl.m
```

```matlab
function y = dct(size);
% This function generates a dct (type II) coefficient matrix of size size!
```
Usage y = dct(size) where size is the required dimension of the matrix
Note that the transform basis vectors are return in the columns of y

```matlab
C = sqrt(2/size);
dc = sqrt(1/size);
D = 2*size;
y = zeros(size);
y(:,1) = ones(size,1)*dc;
disp(size); disp('iterations');
for i = 1:size;
    for j = 2:size;
        y(i,j) = C*cos( (2*(i-1)+1) * (j-1)*pi/D);
    end;
    disp(i)
end;
% end of dct.m function
```

```matlab
function W = wmake(M,P);
%function W = wmake(M,P); determines the 1 block shift of a P*M
%block matrix.

zv = zeros(M*(P-1),M);
zh = zeros(M,P*M);
I = eye(M*(P-1));
W = [zv I; zh];
% end of wmake function
```

```matlab
function h1 = lp_hp(h0);
%function h1 = lp+hp(h0); %h1 is a lowpass to highpass version of h0

N =length(h0);
n = (-1) .^((0:N-1).');
h1 = h0.*n;
%end of lp_hp.m function
```
In this appendix various filter coefficients are listed. In Tables D.1 and D.2 some lowpass eigenfilter coefficients are listed. These eigenfilters were designed using a single zero at DC constraint for the highpass filter, and optimised for an AR(1) source of correlation $\rho=0.98$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$h_0(n), N=12$</th>
<th>$h_0(n), N=14$</th>
<th>$h_0(n), N=16$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6.2468412e-02</td>
<td>6.7772530e-02</td>
<td>2.6629950e-02</td>
</tr>
<tr>
<td>1</td>
<td>7.4482149e-03</td>
<td>5.5079566e-02</td>
<td>-1.7806126e-03</td>
</tr>
<tr>
<td>2</td>
<td>-2.0183990e-01</td>
<td>-1.1960838e-01</td>
<td>-5.6831937e-02</td>
</tr>
<tr>
<td>3</td>
<td>-6.3915226e-02</td>
<td>-3.6947256e-02</td>
<td>1.1077425e-02</td>
</tr>
<tr>
<td>4</td>
<td>4.7897172e-01</td>
<td>4.4588517e-01</td>
<td>5.1713162e-02</td>
</tr>
<tr>
<td>5</td>
<td>7.4541323e-01</td>
<td>7.4795584e-01</td>
<td>-4.5825473e-02</td>
</tr>
<tr>
<td>6</td>
<td>3.9661933e-01</td>
<td>4.3768488e-01</td>
<td>2.2269941e-02</td>
</tr>
<tr>
<td>7</td>
<td>-1.9586013e-02</td>
<td>-4.7988003e-02</td>
<td>4.3702409e-01</td>
</tr>
<tr>
<td>8</td>
<td>-3.3713016e-02</td>
<td>-1.3527342e-01</td>
<td>7.3189656e-01</td>
</tr>
<tr>
<td>9</td>
<td>7.6328841e-02</td>
<td>2.7380012e-02</td>
<td>4.4354928e-01</td>
</tr>
<tr>
<td>10</td>
<td>4.6002297e-03</td>
<td>2.9420706e-02</td>
<td>-8.4132609e-02</td>
</tr>
<tr>
<td>11</td>
<td>-3.8582271e-02</td>
<td>-6.1474671e-02</td>
<td>-2.1287565e-01</td>
</tr>
<tr>
<td>12</td>
<td>-1.8774706e-02</td>
<td>1.8126958e-02</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.3101293e-02</td>
<td>1.1430222e-01</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>-2.5652430e-03</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>-3.8364489e-02</td>
<td></td>
</tr>
</tbody>
</table>

Table D.1. $N=12$, 14 and 16-tap zero constrained minimum time width lowpass eigenfilter coefficients: AR(1) source correlation $\rho=0.98$. Note $h_1(n) = (-1)^n h_0(N-1-n)$
### Table D.2. N=6, 8 and 10-tap zero constrained minimum time width lowpass eigenfilter coefficients: AR(1) source correlation $\rho=0.98$. Note $h_1(n) = (-1)^n h_0(N-1-n)$

<table>
<thead>
<tr>
<th>n</th>
<th>$h_0(n), N=6$</th>
<th>$h_0(n), N=8$</th>
<th>$h_0(n), N=10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.8219100e-01</td>
<td>-1.2772050e-01</td>
<td>-7.3682218e-02</td>
</tr>
<tr>
<td>1</td>
<td>7.9716521e-01</td>
<td>-3.5165245e-02</td>
<td>-5.7051927e-02</td>
</tr>
<tr>
<td>2</td>
<td>4.3020211e-01</td>
<td>5.1916674e-01</td>
<td>1.1188410e-03</td>
</tr>
<tr>
<td>3</td>
<td>-1.4053666e-01</td>
<td>7.6801888e-01</td>
<td>-3.1217235e-01</td>
</tr>
<tr>
<td>4</td>
<td>-1.0528632e-01</td>
<td>3.3302001e-01</td>
<td>-7.1713813e-01</td>
</tr>
<tr>
<td>5</td>
<td>5.0478225e-02</td>
<td>-8.8796585e-02</td>
<td>-5.4843315e-01</td>
</tr>
<tr>
<td>6</td>
<td>-1.7359462e-02</td>
<td>3.7829340e-02</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6.3049731e-02</td>
<td></td>
<td>2.6836485e-01</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td>4.4765384e-02</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>-5.7814209e-02</td>
</tr>
</tbody>
</table>

The lin6 filter coefficients are listed in Table D.3.

### Table D.3. Coefficients of lin6 linear phase lowpass and highpass analysis filters.

Note that $g_1(n) = (-1)^n h_0(n)$, $g_0(n) = (-1)^n h_0(n)$.

<table>
<thead>
<tr>
<th>n</th>
<th>$h_0(n), \sigma^2 = 0.4066$</th>
<th>$h_1(n), \sigma^2 = 0.3617$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.4986852e-03</td>
<td>-5.9676240e-02</td>
</tr>
<tr>
<td>1</td>
<td>1.6497107e-02</td>
<td>-1.3128773e-01</td>
</tr>
<tr>
<td>2</td>
<td>-1.1058717e-01</td>
<td>6.9224438e-01</td>
</tr>
<tr>
<td>3</td>
<td>3.5059426e-02</td>
<td>-6.9224438e-01</td>
</tr>
<tr>
<td>4</td>
<td>7.0610570e-01</td>
<td>1.3128773e-01</td>
</tr>
<tr>
<td>5</td>
<td>7.0610570e-01</td>
<td>5.9676240e-02</td>
</tr>
<tr>
<td>6</td>
<td>3.5059426e-02</td>
<td>-1.1058717e-01</td>
</tr>
<tr>
<td>7</td>
<td>-1.1058717e-01</td>
<td>1.6497107e-02</td>
</tr>
<tr>
<td>8</td>
<td>1.6497107e-02</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>7.4986852e-03</td>
<td></td>
</tr>
</tbody>
</table>