Blind Audio Source Separation

by

Vincent Yan Fu Tan (SID)

Fourth-year undergraduate project in Group F, 2004/2005

Signal Processing Laboratory
Department of Engineering
University of Cambridge

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

May 2005

Abstract

Blind Audio Source Separation

Vincent Y.F. Tan (SID)
Signal Processing Laboratory,
Cambridge University Engineering Department,
Cambridge, CB2 1PZ, UK

In this project, the problem of blind separation of underdetermined mixtures of audio sources is considered. The sources have to be sparsely represented on a given basis or dictionary. They are assumed to follow a Student t distribution in the transformed domain. A Bayesian approach using Gibbs Sampling is employed to estimate the sources in the transformed domain. The performance resulting from the use of various transform methods is compared. A central theme of this project is *sparsity* and how this relates to separation quality and noise reduction.

The orthonormal bases that will be considered include the Discrete Cosine Transform, the Modified Discrete Cosine Transform, two Discrete Wavelet Transforms and a basis obtained using the Wavelet Packet Best Basis Algorithm. The overcomplete dictionaries that will be used include the Hybrid Transform and the Short-Time Discrete Cosine Transform. These transforms were chosen because of their varying abilities to provide sparse representations of different classes of audio signals.

The numerical results show that if the signal is sparse in the transformed domain, the separation results are better. Sparsity is measured using the normalized $\mathcal{L}_1$ norm. Various measures of separation quality, including the Source to Distortion Ratio will be used. In general, the Modified Discrete Cosine Transform is a good representation for audio signals with more tonals than transients. The Discrete Wavelets provide a sparser representation for audio signals with more transients than tonals. Due to its adaptive nature, the Wavelet Packet Best Basis Algorithm finds a basis that provides a sparse representation for all the classes of signals that were tested. Finally, contrary to our initial expectations, the overcomplete dictionaries failed to provide a significant improvement in separation quality. We propose an explanation for this observation.

Acknowledgements

I could have never achieved something like that without my parents. They have always supported me and my decisions, both financially and morally. They have given me valuable advice, and treated me like a mature adult. Mom, Dad, thank you for everything you have done for me.

Though writing this final report is a solitary act, my research this year was most certainly collaborative in nature. I have worked significantly with excellent researchers from the Signal Processing Laboratory here in Cambridge, UK and they have provided me with good advice. I owe the most and greatest thanks to my supervisor and friend Dr. Cédric Févotte for his many suggestions and constant support during this research. Your teaching opened me to the fascinating world of Blind Source Separation and Independent Component Analysis, giving me a solid background to my research.

I wish to thank the Public Service Commission (PSC) of Singapore for contributing significant financial support to my studies at the University of Cambridge and Massachusetts Institute of Technology (MIT). I could not have completed this major piece of work without the firm reassurance of financial aid throughout the course of my undergraduate education.

This past year has consisted of more than just research and studies. Further thanks to my friends from Singapore, Christopher Gordon, Gareth Tang, Melissa Tan and Xingfang Su for making my undergraduate years enjoyable and successful. Thanks for the unwavering support and good advice. Thanks to my friends in Cambridge and MIT for exposing me to a host of new ideas beyond the shores of tiny Singapore. I am also grateful to Sidney Sussex College, Cambridge for the excellent pastoral care and supervision through the course of my undergraduate education.

Finally, I would like to thank my girlfriend, Huili for her unconditional love so that I may have the strength to successfully complete my research. Her companionship has ensured that my time in Cambridge has not entirely been consumed by my studies.

To my parents,

Contents

Acknowledgements **ii**

List of Figures **vii**

List of Tables **ix**

Glossary **x**

List of Symbols **xi**

1 Introduction and Preview **1**

1.1 Literature review 2

1.2 Applications 4

1.2.1 Biomedicine 4

1.2.2 Image Denoising and Compression 4

1.2.3 Financial Time Series 5

1.2.4 Telecommunications 5

1.3 Structure of report 5

2 Blind Source Separation and Orthonormal Bases **6**

2.1 Model and Assumptions 6

2.1.1 BSS Model 7

2.1.2 Transform 7

2.1.3 Assumptions for BSS 8

2.2 Method 9

2.2.1 Gibbs Sampling 10

2.2.2 Reconstruction 11

2.2.3 Sparsity 11

2.3 Orthonormal Bases 13

2.3.1 Discrete Cosine Transform (DCT) 13

2.3.2 Modified Discrete Cosine Transform (MDCT) 14

2.3.3 Wavelet Transforms: Vaidyanathan (WT-Vai) 15
2.3.4 Wavelet Transforms: Symmlets (WT-Sym) 16
2.3.5 Wavelet Packet Best Basis (WPBB) 16
2.3.6 No Transform (NT) 17
2.3.7 Summary of Orthonormal Bases 18
2.4 Performance Measures 18
2.4.1 Source to Distortion Ratio (SDR) 19
2.4.2 Source to Interference Ratio (SIR) 19
2.4.3 Source to Artifacts Ratio (SAR) 20
2.4.4 Source to Noise Ratio (SNR) 20
2.5 Conclusions 20

3 Overcomplete Dictionaries 21
3.1 Analysis and Synthesis 21
3.2 Overcomplete Dictionaries: Examples 22
3.2.1 Short-Time Discrete Cosine Transform (STDCT) 23
3.2.2 Hybrid Transforms (HT) 25
3.3 Conclusions 26

4 Results: Orthonormal Bases 27
4.1 Introduction and Initialization 27
4.1.1 Signals and Estimation 27
4.1.2 Representation of transforms and performance measures 28
4.1.3 Implementation Details 28
4.1.4 Evolution of parameters 29
4.1.5 Basis Invariance 29
4.2 Speech Signals 29
4.2.1 Sparsity 30
4.2.2 Results 30
4.2.3 Discussion 31
4.3 Musical Signals 31
4.3.1 Sparsity 32
4.3.2 Results 33
4.3.3 Discussion 33
4.4 Percussion Signals 34
4.4.1 Sparsity 34
4.4.2 Results 34
4.4.3 Discussion 35
4.5 Combination of Signals 35

4.5.1 Sparsity . . . 36
4.5.2 Results . . . 36
4.5.3 Discussion . . . 36
4.6 Conclusions . . . 37
4.6.1 Relation between sparsity and performance metrics . . . 37
4.6.2 Performance of Transforms . . . 38
4.6.3 Effect of the Sources . . . 39

5 Results: Overcomplete Dictionaries 40
5.1 Introduction and Initialization . . . 40
5.2 Speech Signals . . . 40
5.3 Musical Signals . . . 41
5.4 Percussion Signals . . . 42
5.5 Combination of Signals . . . 42
5.6 Conclusions . . . 44
5.6.1 No Significant Improvement . . . 44
5.6.2 Proposed Explanation . . . 44
5.6.3 Further Work . . . 45

6 Conclusion 46
6.1 Summary . . . 46
6.1.1 The MDCT is an excellent basis for BSS . . . 46
6.1.2 Best Basis Algorithm finds an excellent basis for BSS . . . 47
6.1.3 Sparsity is very important . . . 47
6.1.4 Overcomplete dictionaries provide marginal improvement . . . 47
6.1.5 Computational Complexities . . . 47
6.2 Further Work . . . 47
6.2.1 Alternative Gibbs Sampling . . . 48
6.2.2 Alternative Prior Distributions . . . 48
6.2.3 Convolutive BSS . . . 48
6.2.4 BSS with unknown number of sources n . . . 48

A Probability Density Functions 49
A.1 Standard Distributions . . . 49
A.2 Conditional Densities . . . 49

Bibliography 51

List of Figures

2.1 Comparison of Student t distributions with different values of α 9

2.2 Block Diagram of the algorithm 11

2.3 Statistical properties of audio signals 13

2.4 Smooth windows used in the MDCT 15

2.5 Filter bank/Multiscale representation of a signal 16

3.1 Window functions used in the STDCT. 23

3.2 Lapped Block Diagonal Structure of $\mathbf{\Psi}$; Note that $K > N$ 24

4.1 Evolution of $\mathbf{r}_2$, $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ and σ^2 29

4.2 Speech Signals $\mathbf{s}_{11}, \mathbf{s}_{12}$ and $\mathbf{s}_{13}$ 30

4.3 Sparsity Indices for Signals using Orthonormal Bases a) Speech Signals, b) Musical Signals, c) Percussion Signals, d) Combination of Signals; Transform method: 1-DCT, 2-MDCT, 3-WT-Vai, 4-WT-Sym, 5-WPBB, 6-NT, 7-WPBB on sources (optimal) 31

4.4 Performances of the various transforms on Speech Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{12}$, brown-$\mathbf{s}_{13}$ 32

4.5 Musical Signals $\mathbf{s}_{21}, \mathbf{s}_{22}$ and $\mathbf{s}_{23}$ 32

4.6 Performances of the various transforms on Musical Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{21}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{23}$ 33

4.7 Percussion Signals $\mathbf{s}_{31}, \mathbf{s}_{32}$ and $\mathbf{s}_{33}$ 34

4.8 Performances of the various transforms on Percussion Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{31}$, green-$\mathbf{s}_{32}$, brown-$\mathbf{s}_{33}$ 35

4.9 Performances of the various transforms on Combination of Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{33}$ 36

4.10 Correlation between mean of sparsity indices ξ and mean of the Performance Indices for Orthonormal Bases 38

5.1 Sparsity Indices for Signals using Overcomplete Dictionaries a) Speech Signals, b) Musical Signals, c) Percussion Signals, d) Combination of Signals; Transform method: 1-STDCT, 2-HT 41

5.2 Performances of the two Overcomplete Dictionaries on Speech Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{12}$, brown-$\mathbf{s}_{13}$ 42

5.3 Performances of the two Overcomplete Dictionaries on Musical Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{21}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{23}$ 43

5.4 Performances of the two Overcomplete Dictionaries on Percussion Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{31}$, green-$\mathbf{s}_{32}$, brown-$\mathbf{s}_{33}$ 43

5.5 Performances of the two Overcomplete Dictionaries on the Combination of Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{33}$ 44

List of Tables

2.1 Comparison of computational complexities for the various transforms . . . 18

3.1 Comparison of computational complexities for the Overcomplete Dictionaries 26

4.1 Initialization of the parameters to be estimated 28
4.2 Correspondence between Transform Number and the orthonormal transform used 28
4.3 Correspondence between Performance Number and the performance index used 28
4.4 Ranking of the Orthonormal Bases for Speech Signals 31
4.5 Ranking of the Orthonormal Bases for Musical Signals 34
4.6 Ranking of the Orthonormal Bases for Percussion Signals 35
4.7 Ranking of the Orthonormal Bases for Combination of Signals 37
4.8 Transient and tonal indices for the audio signals 39
4.9 Transient and tonal indices for the mixtures and the preferred transform . 39

5.1 Parameters used in the implementation of the STDCT 40

Glossary

AR	Autoregressive
BSS	Blind Source Separation
CTFT	Continuous-Time Fourier Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
EM	Expectation-Maximization
FFT	Fast Fourier Transform
FWT	Fast Wavelet Transform
HT	Hybrid Transform
ICA	Independent Component Analysis
i.i.d.	independent, identically distributed
KLT	Karhunen-Loève Transform
MCMC	Markov Chain Monte Carlo
MDCT	Modified Discrete Cosine Transform
ML	Maximum Likelihood
MMSE	Minimum Mean Squared Error
NT	No Transform
pdf	probability density function
PR	Perfect Reconstruction
SAR	Source to Artifacts Ratio
SDR	Source to Distortion Ratio
SIR	Source to Interference Ratio
SNR	Source to Noise Ratio
STDCT	Short-Time Discrete Cosine Transform
STFT	Short-Time Fourier Transform
WPBB	Wavelet Packet Best Basis
WT-Vai	Vaidyanathan Discrete Wavelet Transform
WT-Sym	Symmlets of order 8 Discrete Wavelet Transform

List of Symbols

Scalar	Lower case, e.g. $s_{i,t}$, $\tilde{s}_{i,k}$
Vector	Bold lower case e.g. $\mathbf{s}_t$, $\mathbf{s}_i$ or underline for some Greek variables e.g. $\underline{\psi}_i$
Matrix	Bold upper case, e.g. $\mathbf{S}$, $\mathbf{\Psi}$
$p(\cdot)$	Probability density function
$\theta \sim p(\theta)$	θ is a random sample from $p(\theta)$
m	Number of observations
n	Number of sources
$\mathbf{X}$	$m \times N$ matrix of observations
$\mathbf{A}$	$m \times n$ mixing matrix
$\mathbf{S}$	$n \times N$ matrix of sources
$\mathbf{N}$	$n \times N$ matrix of i.i.d. Gaussian noise samples
$\mathbf{\Psi}$	Analysis Operator
$\mathbf{\Phi}$	Synthesis Operator
$\mathbf{I}_N$	$N \times N$ Identity matrix
$\mathrm{diag}(\mathbf{v})$	Diagonal matrix with elements of $\mathbf{v}$ along the diagonal
σ	Noise standard deviation
I_{tr}	Transient Index
I_{ton}	Tonal Index
$s(t)$	Continuous-Time Signal
s_t	Discrete-Time Signal
$\langle \mathbf{x}, \mathbf{y} \rangle$	Inner product i.e. $\mathbf{x}^T\mathbf{y} = \sum_t x_t y_t$
$\\|\mathbf{s}_i\\|_p$	$\mathcal{L}_p$ norm i.e. $(\sum_t \lvert s_{i,t}\rvert^p)^{1/p}$
$x = \mathcal{O}(y)$	Order of
$\triangleq$	Defined as
$\square$	'Quot erat demonstratum', denotes the end of a proof
$\mathbb{Z}$	Integers
$\mathbb{C}$, $\mathbb{C}^N$	Complex Numbers, Complex Signal/Vector of size N
$\mathbb{R}$, $\mathbb{R}^N$	Real Numbers, Real Signal/Vector of size N
$\mathbf{l}^2(\mathbb{Z})$	Finite energy discrete signals $\sum_t \lvert s_t\rvert^2 < +\infty$
$\mathcal{L}^2(\mathbb{R})$	Finite energy functions $\int \lvert s(t)\rvert^2 dt < +\infty$

Chapter 1

Introduction and Preview

Imagine you are in a party-room where 2 people are speaking simultaneously. There are 2 receivers, which are in different locations and can record the time signals generated by the 2 people. Assuming that the each of the recorded time signals $x_1(t)$ and $x_2(t)$ is a linear combination of the speeches $s_1(t)$ and $s_2(t)$ so that this can be expressed as

$$\begin{aligned} x_1(t) &= a_{1,1}s_1(t) + a_{1,2}s_2(t), \\ x_2(t) &= a_{2,1}s_1(t) + a_{2,2}s_2(t). \end{aligned}$$

for each time index t. It would be useful in several applications if we can estimate the signals $s_1(t)$ and $s_2(t)$ from the mixtures $x_1(t)$ and $x_2(t)$ for the time frame of interest. This is the canonical, determined 'cocktail party problem' and the chief motivation behind Blind Source Separation (BSS) and Independent Component Analysis (ICA).

This project focuses on various transform methods used in BSS [42, 85]. The goal of BSS and ICA [58, 69] is to determine the original sources given mixtures of those sources. Examples include speech separation in the 'cocktail party problem', processing of arrays of sonar signals, blind separation of electroencephalographic (EEG) [74, 76], electromyographic (EMG) [37] data and separation of artifacts in magnetoencephalographic (MEG) [58, 75] data in biomedicine. The adjective 'blind' emphasizes that the sources signals are not observed and no information is available with regard to the mixture process.

There are two key principles in BSS and ICA. Firstly, the key to estimating the ICA model is *nongaussianity.* To further elaborate, the independent components must be nongaussian for ICA to be possible because if the sources are Gaussian, the directions of the columns of the mixing matrix $\mathbf{A}$ cannot be inferred. Nongaussianity can be quantified via using measures such as kurtosis, mutual information and negentropy. Similarly, it is required to further assume that the source signals are *independent.* This is a statistically strong but physically plausible assumption. In fact, Hyvärinen [57, 58] argues that independent components can be recovered by maximizing the nongaussianity of the components. To maximize the nongaussianity for a demixing system $\mathbf{Y} = \mathbf{WX}$, where $\mathbf{X}$ contains the observations, we typically optimize over the demixing matrix $\mathbf{W}$ such that the kurtoses of the components in $\mathbf{Y}$ is maximized.

A vast number of fast and effective methods exist for solving the determined problem (see for example [18] and [57]). However, in this project, we focus on the more difficult underdetermined, noisy case. Fortunately, a linear instantaneous model is employed. A summary of the algorithms used for *convolutive* mixtures can be found in [16] and [84]. In the underdetermined case, the number of sources n is greater than the number of mixtures m and the problem is degenerate [118] as traditional matrix inversion demixing cannot be applied. In this case, the separation of the underdetermined mixtures requires prior [42, 86] information on the sources to allow for their reconstruction. Estimating the mixing system is not sufficient for reconstructing the sources, since for $m < n$, the mixing matrix is not invertible. As the system of linear equations is underdetermined, a *model* of the sources is required. In particular, we employ the use of the Student t model as the prior distribution of the sparse sources in the transformed domain. Given the observations, one seeks to estimate the sources and the mixing matrix and possibly the variance of the noise. The sources are sparsely represented by orthonormal bases and overcomplete dictionaries [1, 31]. This assumption means that only a few coefficients of the decomposition are significantly different from zero. We employ a Bayesian approach using Gibbs Sampling to estimate the sources in the transformed domain before reconstructing the sources. This Bayesian approach is not limited to underdetermined mixtures but applies to the (over)determined case as well [42]. It is particularly useful when applied to the general linear, instantaneous and noisy model, which we consider in this project.

To achieve the sparsest representation of the audio signal, wavelet packet best basis [114] algorithms are used to minimize some form of entropy measure. Instead of restricting ourselves to this, we will consider a series of orthonormal and overcomplete transforms. A sparsity measure will be compared to the separation quality. Standardized measures [40] for assessing audio sources separation algorithms will be used. We will show that sparsity of the sources in the transformed domain is closely related to the separation quality.

1.1 Literature review

BSS was first suggested by the French researchers Hérault, Jutten, and Ans [55] in the mid 1980s. The theory was further developed in the 1990s where researchers found ways to solve the (over)determined problem using, for instance, natural gradient methods [2, 3], maximum likelihood estimation [95], higher order statistics [18, 58], information maximization [69, 109], mutual information [110] and non-stationarity [7]. Adaptive BSS and ICA are discussed in detail in [21] and the concept of equivariance in [19]. Indeed, these well-established methods have performed well for the (over)determined BSS problem, for example in image coding [6]. Several books [21, 58] on ICA have been written as well.

The underdetermined problem has been of interest to researchers once solutions to the (over)determined problem had been established. In [119], the authors also exploited the use of sparse sources. A two stage approach was proposed, that is, the mixing matrix was estimated using a clustering algorithm. Following that, the source matrix is estimated. A related letter [73] examines the effect of sparsity using the algorithm in [119]. The authors focused mainly factorization of the data matrix by estimating the overcomplete basis matrix using the K-means method to find the best basis. Furthermore, in [64] and [118], the authors used a relatively new technique based on the Degenerate Unmixing and Estimation Technique (DUET). DUET can be applied to a class of signals known as W-disjoint orthogonal signals [101]. In essence, W-disjoint orthogonal signals have disjoint support for their time-frequency representation. Similarly, the authors in [41] and [43] contributed to BSS by using time-frequency representations. Assigning a prior to the hitherto unknown sources is also a popular way of performing underdetermined blind separation. Several authors have modelled the sources using Laplacian distributions [72, 86] or Gaussian Mixture Models (GMM) [31] and estimating the sources and mixing matrix using Expectation-Maximization (EM) [11]. As mentioned in [42], although using EM is computationally efficient, the algorithm is susceptible to convergence at local maxima. Using the successful results in [116], where audio restoration was modelled using Student t distributions, we will model the sources in the transformed domain using Student t distributions. Furthermore, inference will be performed using a Markov Chain Monte Carlo (MCMC) method based on the the Gibbs Sampler [47]. The Gibbs Sampler is, in fact, a stochastic analog of EM.

As mentioned, the concept of sparsity will be exploited in this project. Finding a sparse representation of a signal or an image is desired in many applications, for instance in JPEG image compression [10, 48, 59]. Researchers have developed many means based on optimization principles to express a signal in the sparsest possible representation [20, 32, 36, 50, 77]. Under mild conditions, minimizing the $\mathcal{L}_1$ norm is equivalent to minimizing the $\mathcal{L}_0$ norm [33], the number of nonzero entries in the sequence. Besides, minimizing the $\mathcal{L}_1$ norm is a relatively simple task, which in general, reduces to a a linear program [20, 73]. Thus, we will use the $\mathcal{L}_1$ norm [32, 104] to measure sparsity.

Numerous transform methods for representing a signal exist. These range from the traditional Discrete Fourier Transform (DFT) to more exotic representations using wavelets. The Fast Fourier Transform (FFT) attributed to Cooley and Tukey [26] revolutionized signal processing and sparked a range of research topics in frequency analysis. Time-frequency analysis and multi-resolution signal processing [27, 97, 111] soon followed. Most of these transforms have the effect of energy compaction such that in the transformed domain, only a small number of coefficients are significantly different from zero i.e. *sparse*. A subset of these methods will be considered in this project. Transforms have traditionally been used as a pre-processing step in BSS. This has the effect of compacting the energy of the sources. For instance, in [31], the authors present results with speech signals decomposed on a Modified Discrete Cosine Transform (MDCT) basis [28]. The authors of [66] use a multi-scale framework for performing blind separation instead. They used various versions of the Discrete Wavelet Transform (DWT) [27, 79, 112] as well as Wavelet Packets (WP) [114] to decompose the sources. We will consider these transforms among others in this project. Lastly, overcomplete representations such as the hybrid transforms mentioned in [29] and [88] will be considered.

1.2 Applications

The classical application of BSS and ICA is trying to understand how the humans select the voice of a particular speaker from an ensemble of different voices corrupted by music and noise in the background. As described, this is the quintessential 'cocktail party problem' [58]. Other applications, besides those already mentioned in the Introduction, have also sprung to interest over the past decade or so. These include:

1.2.1 Biomedicine

Atrial fibrillation is one of the most prevalent abnormal heart rhythms. As the signal strength of the atrial component of the body surface electrocardiogram (ECG) is small, it has proven to be difficult to detect using non-invasive techniques. It would be thus, very helpful to clinicians if there are established algorithms to separate the atrial component from the stronger ventricular component in an ECG. This is discussed in the paper by Raine [98]. In EEG analysis, different artifacts such as eye-blinking deteriorate its quality. Identification of the various sources from the independent components is thus integral for clinal analysis. An innovative method [113] combining the use of standard BSS techniques and Support Vector Machines (SVM) [34] has been proposed to solve this problem. Advanced implementations of ICA as applied to neurophysiologic signals in the form of electromagnetic brain signals data are demonstrated in [60].

1.2.2 Image Denoising and Compression

In close relation to ICA is a field known as Sparse Code Shrinkage [56], which can be used for denoising of natural images. Briefly, the model used is

$$\mathbf{X} = \mathbf{S} + \mathbf{N}. \tag{1.1}$$

In Sparse Code Shrinkage, the density of the observations **S** are modelled by ICA. If the noise **N** is assumed to be Gaussian, then finding the Maximum Likelihood (ML) estimate for **S** given the mixtures **X** basically encapsulates the denoising process. The authors in [38, 39] exploited ICA to obtain transform-based compression schemes for images and similarly in [68], ICA was used to compute features extracted from natural images.

1.2.3 Financial Time Series

In econometrics, finding the common underlying factors [67] in perturbations of financial data, such as currency exchange rates or daily returns of stock is integral. This appears to be an appealing way to apply ICA on financial data. In a recent study of stock portfolio by Back and Weigend [5], ICA was found to be a complementary tool to Principal Component Analysis (PCA) [34]. Furthermore, in [117], ICA was used to forecast the U.S. real output and inflation variables. ICA is a powerful method for studying the underlying structure of the financial data and driving mechanisms in financial time series.

1.2.4 Telecommunications

In Code-Division Multiple Access (CDMA) and mobile communications, blind separation techniques [103] are used to separate the desired signals from other users' signals. Similar to the approach that we will be adopting later, the sources are assigned a prior distribution [102] to enhance the performance of the separation process.

1.3 Structure of report

This report will be structured as follows. We discuss the general method employed in Chapter 2. Here, the model and assumptions will be presented and a sparsity measure will be formally introduced. Various transform methods will also be discussed before the performance metrics to evaluate different BSS algorithms are defined. An extension to overcomplete dictionaries is provided in Chapter 3. We then turn our discussion to numerical results in Chapter 4. This chapter deals primarily with the effect of various orthonormal bases and consequently, sparsity on separation quality. It will be shown empirically, that there is a close correlation between sparsity and sound quality regardless of the nature of the sources. In Chapter 5, the effects of using the Short-Time Discrete Cosine Transform (STDCT) and Hybrid Transform (HT) are quantified. The importance of sparsity will be reinforced. We will see that the performance improves only marginally and we propose an explanation for this. Conclusions and perspectives are presented in Chapter 6.

Chapter 2

Blind Source Separation and Orthonormal Bases

In this chapter, the Blind Source Separation (BSS) model will be presented. A set of notation will be established for instantaneous linear mixtures [42]. Following that, the three main assumptions in BSS [18, 58] will be discussed. We will then proceed to summarize the method used for separating the sources in an underdetermined mixture. This involves using an orthonormal transform to achieve sparsity of the sources in the transformed domain. A Bayesian approach based on Gibbs Sampling [45, 46, 54] will be adopted to estimate the sources and the mixing matrix. The reconstruction step completes the algorithm. A selection of *orthonormal bases* will be presented. The bases were selected based on their different abilities to compress of various forms of audio data. For instance, it is well known that the Modified Discrete Cosine Transform (MDCT) [28, 79] is well suited for audio compression. Yet, there has been significant research devoted to using Wavelet Coefficients [30, 88] to model the transients in audio signals. The Wavelet Packet Best Basis Algorithm [20, 66] has also been used widely to achieve sparsity in BSS and audio and image [79] coding. Finally, we will discuss four different performance measures [40, 51] to evaluate the performance of source separation on the audio signals. Some other popular methods employed in audio source separation are given by Mitianoudis and Davies [85] and in the PhD thesis by Mitianoudis [83].

2.1 Model and Assumptions

The notation that we adopt is in line with that presented in [42] and [43]. We will present a discussion on the assumptions made on the sources and the noise. We will also briefly discuss about BSS indeterminacies here.

2.1.1 BSS Model

We consider the following linear instantaneous model that is commonly used in BSS.

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t, \qquad 0 \leq t \leq N-1 \tag{2.1}$$

where $\mathbf{x}_t = (x_{1,t}, \ldots, x_{m,t})^T$ is the vector of size m containing the observations, $\mathbf{s}_t = (s_{1,t}, \ldots, s_{n,t})^T$ is the vector of size n containing the non-stationary stochastic sources. $\mathbf{A}$ is the $m \times n$ unknown full rank mixing matrix (with possibly $m < n$). Variables without time index t denote whole sequences or samples, for example, $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_{N-1})$. As a consequence, (2.1) can be rewritten more succinctly as

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}. \tag{2.2}$$

The goal is to estimate the sources $\mathbf{S}$ and the mixing matrix $\mathbf{A}$ up to the standard BSS indeterminacies [18] of gain and order. Hence, the task is to compute $\hat{\mathbf{S}}$ and $\hat{\mathbf{A}}$ such that

$$\hat{\mathbf{A}} = \mathbf{APD}, \tag{2.3}$$

$$\hat{\mathbf{S}} = \mathbf{P}^T\mathbf{D}^{-1}\mathbf{S}, \tag{2.4}$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ a permutation matrix.

2.1.2 Transform

In the following discussion, $\boldsymbol{\Psi} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{N\times N}$ are orthonormal matrices. $\boldsymbol{\Psi}$ is the analysis operator such that the decomposition of the sources $\tilde{\mathbf{S}} = \mathbf{S}\boldsymbol{\Psi}$ results in a sparse representation. Note that the overhead tilde '$\sim$' is used to denote a vector in the transformed domain. Since it is sparse, the matrix $\tilde{\mathbf{S}}$ has few coefficients that are significantly different from zero. $\boldsymbol{\Phi}$ is the synthesis operator such that $\boldsymbol{\Phi} = \boldsymbol{\Psi}^{-1} = \boldsymbol{\Psi}^T$. Equation (2.2) can then be rewritten in the transformed domain as

$$\mathbf{X}\boldsymbol{\Psi} = \mathbf{A}\mathbf{S}\boldsymbol{\Psi} + \mathbf{N}\boldsymbol{\Psi}, \tag{2.5}$$

$$\tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{S}} + \tilde{\mathbf{N}}. \tag{2.6}$$

As $\boldsymbol{\Psi}$ is an orthonormal matrix, equations (2.1) and (2.6) are equivalent. Estimating the sources in the transformed domain is equivalent to estimating the sources in the time domain, except for an orthonormal transform. We will primarily work in the transformed domain because sparsity is essential for achieving good separation quality as we will see in Chapter 4.

2.1.3 Assumptions for BSS

We emphasize that for BSS to be performed, the sources have to be sparsely represented on a given basis or dictionary. Sparsity is well modelled by the Student t distribution. Furthermore, the sources have to be statistically independent and we assume that the noise is i.i.d. Gaussian.

Student t distribution

Although several authors have modelled the sources using Laplacian distributions [31, 72, 86], the successful results of Student t modelling for audio restoration [116] has inspired us to model the sources in the transformed domain using the Student t distribution. This leads to sparse modelling when the "degrees of freedom" is low. Hence, we assume that each source sequence $\tilde{\mathbf{s}}_i$ in the transformed domain is i.i.d. with Student t distribution $t(\alpha_i, \lambda_i)$

$$p(\tilde{s}_{i,k}|\alpha_i, \lambda_i) = \frac{\Gamma(\frac{\alpha_i+1}{2})}{\lambda_i\sqrt{\alpha_i\pi}\Gamma(\frac{\alpha_i}{2})}\left(1 + \frac{1}{\alpha_i}\left(\frac{\tilde{s}_{i,k}}{\lambda_i}\right)^2\right)^{-\frac{\alpha_i+1}{2}}, \tag{2.7}$$

where α_i is the "degrees of freedom" and λ_i a scale parameter. It is known [4] that the Student t distribution can be expressed as a Scaled Mixture of Gaussians (SMoG).

$$p(\tilde{s}_{i,k}|\alpha_i, \lambda_i) = \int_0^\infty \mathcal{N}(\tilde{s}_{i,k}|0, v_{i,k})\mathcal{IG}\left(v_{i,k}\middle|\frac{\alpha_i}{2}, \frac{2}{\alpha_i\lambda_i^2}\right) dv_{i,k}. \tag{2.8}$$

This property will be very useful when deriving the posterior distributions of the parameters when we implement the Gibbs Sampler. In comparison to the Laplacian prior, the Student t prior has the advantage of providing a supplementary hyperparameter α_i which controls the sharpness of the distribution. We denote $\mathbf{v}_t = (v_{1,t}, \ldots, v_{n,t})^T$, $\mathbf{V} = (\mathbf{v}_0, \ldots, \mathbf{v}_{N-1})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. The property in equation (2.8) implies that a random variable $y \sim t(\alpha, \lambda)$ can be sampled by

$$v \sim \mathcal{IG}\left(v\middle|\frac{\alpha}{2}, \frac{2}{\alpha\lambda^2}\right), \tag{2.9}$$

$$y \sim \mathcal{N}(y|0, v), \tag{2.10}$$

where $\mathcal{IG}(x|\gamma, \beta)$ and $\mathcal{N}(x|\mu, \sigma^2)$ are the Inverted-Gamma[1] and Normal distributions respectively. The analytical forms of the distributions can be found in Appendix A.1. A plot of several Student t distributions with different values of α is shown in Figure 2.1. The densities were normalized so that they have the same value at the mode. It was noted that small values of α tend to give peaky distributions. As α tends to large numbers, the

[1] The Inverted-Gamma distribution is the distribution of $1/X$ when X is Gamma distributed.

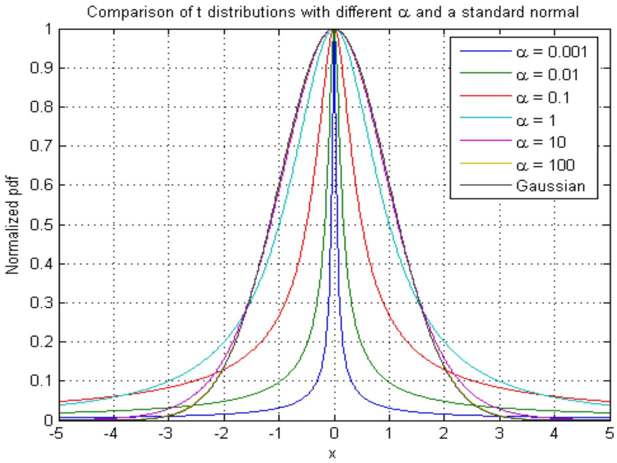

Figure 2.1: Comparison of Student t distributions with different values of α

density function tends to a standard normal $\mathcal{N}(0,1) = (2\pi)^{-1/2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$.

Independence of Sources

We will assume that the sources in the transformed domain are statistically and mutually independent, such that

$$p(\tilde{\mathbf{S}}) = \prod_{i=1}^{n} p(\tilde{\mathbf{s}}_i). \tag{2.11}$$

Since the next step is to decompose the sources on a given basis, it would be more relevant to consider that the *coefficients* of the sources on the basis are mutually independent [119].

Independent, Identically Distributed Gaussian Noise

Finally, in (2.1), $\mathbf{n}_t$ is an i.i.d noise vector independent of sources with covariance $\sigma^2\mathbf{I}_m$ and σ unknown. Since the transform is orthonormal, $\tilde{\mathbf{n}}_t$ is similarly an i.i.d noise vector independent of sources with covariance $\sigma^2\mathbf{I}_m$.

2.2 Method

The method employed is adapted from [42]. It involves the use of a transform as mentioned in section 2.1.2 to ensure that the sources are sparse before Gibbs Sampling is applied. Reconstruction then follows before the algorithms are evaluated using standard metrics [40, 51]. A block diagram of the method that will be described in detail is shown in Figure 2.2.

2.2.1 Gibbs Sampling

The tremendous improvements in computational power have led to a dramatic increase in interest in Markov Chain Monte Carlo (MCMC) methods. MCMC in the form of the Gibbs Sampler and the Metropolis-Hastings algorithm [47] allows any distribution to be simulated on a finite-dimensional state space specified by any conditional density. In particular, the Gibbs Sampler was first studied by the statistical physics community [82] and then later in the statistics community [46, 54]. The basis behind Gibbs Sampling is the Hammersley-Clifford [53] theorem which states that given the data $\mathbf{d}$, the conditional densities $p_i(a_i|a_{i\neq j}, \mathbf{d})$ contain sufficient information to produce samples from the joint density $p(a_1, a_2, \ldots, a_k|\mathbf{d})$. Gibbs Sampling has been used extensively and successfully in image [46] and audio restoration [47] and interpolation of missing samples [91]. The Gibbs Sampler is presented here to estimate $\{\tilde{\mathbf{S}}, \mathbf{A}, \sigma\}$ together with $\mathbf{V}$ and the hyperparameters $\{\boldsymbol{\alpha}, \boldsymbol{\lambda}\}$. As is common with most of the literature on the Gibbs Sampler, we define $\boldsymbol{\theta} = \{\tilde{\mathbf{S}}, \mathbf{A}, \sigma, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\}$, and $\boldsymbol{\theta}_{-\mathbf{y}}$ denotes the set of parameters except $\mathbf{y}$. Also, a Jeffrey's uninformative prior [62, 63] is assigned to the standard deviation of the noise such that $p(\sigma) = \kappa/\sigma$ and κ is a constant. The Gibbs Sampling algorithm is as follows:

```
Initialize $\boldsymbol{\theta}^{(0)} = \{\tilde{\mathbf{S}}^{(0)}, \mathbf{A}^{(0)}, \sigma^{(0)}, \mathbf{V}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\lambda}^{(0)}\}$.
for $k = 1 : K$ do
```

$$\begin{aligned}
\tilde{\mathbf{S}}^{(k)} &\sim p(\tilde{\mathbf{S}}|\mathbf{A}^{(k-1)}, \sigma^{(k-1)}, \mathbf{V}^{(k-1)}, \tilde{\mathbf{X}}) && (2.12)\\
\sigma^{(k)} &\sim p(\sigma|\tilde{\mathbf{S}}^{(k)}, \tilde{\mathbf{X}}) && (2.13)\\
\mathbf{A}^{(k)} &\sim p(\mathbf{A}|\tilde{\mathbf{S}}^{(k)}, \sigma^{(k)}, \tilde{\mathbf{X}}) && (2.14)\\
\mathbf{V}^{(k)} &\sim p(\mathbf{V}|\tilde{\mathbf{S}}^{(k)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}) && (2.15)\\
\boldsymbol{\alpha}^{(k)} &\sim p(\boldsymbol{\alpha}|\mathbf{V}^{(k)}, \boldsymbol{\lambda}^{(k-1)}) && (2.16)\\
\boldsymbol{\lambda}^{(k)} &\sim p(\boldsymbol{\lambda}|\mathbf{V}^{(k)}, \boldsymbol{\alpha}^{(k)}) && (2.17)
\end{aligned}$$

```
end for
```

where the superscript (k) indicates the value of a random vector/matrix at the k^{th} iteration and the $\sim$ denotes sampling from the conditional density on the right. The conditional densities from equations (2.12) to (2.17) can be found in Appendix A.2. K is the total number of iterations of the Gibbs Sampler and K_b is the burn-in length, the number of iterations before the Markov Chain reaches its stationary distribution. For the sake of brevity, we refer the interested reader to [42] for the derivation of the analytical forms of the various conditional densities $p(\boldsymbol{\theta}_{\mathbf{y}}|\boldsymbol{\theta}_{-\mathbf{y}}, \tilde{\mathbf{X}})$.

After the burn-in period, the Markov Chain reaches equilibrium and one will observe

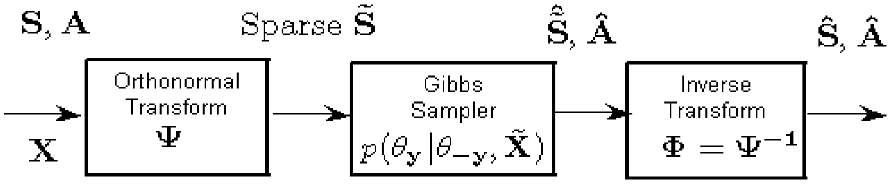
$\mathbf{S}, \mathbf{A}$
Sparse $\tilde{\mathbf{S}}$
$\hat{\tilde{\mathbf{S}}}, \hat{\mathbf{A}}$
$\hat{\mathbf{S}}, \hat{\mathbf{A}}$
Orthonormal Transform Ψ
Gibbs Sampler $p(\theta_{\mathbf{y}}|\theta_{-\mathbf{y}}, \tilde{\mathbf{X}})$
Inverse Transform $\Phi = \Psi^{-1}$
$\mathbf{X}$

Figure 2.2: Block Diagram of the algorithm

that the Gibbs Sampler draws random samples from the underlying posterior distributions $p(\boldsymbol{\theta}|\tilde{\mathbf{X}})$. Consequently, the Minimum Mean Squared Error (MMSE) Estimate [11, 34, 65] can be written as

$$\hat{\boldsymbol{\theta}}_{MMSE} = \frac{1}{K - K_b}\left(\sum_{k=K_b+1}^{K} \boldsymbol{\theta}^{(k)}\right). \tag{2.18}$$

2.2.2 Reconstruction

After the *burn-in* period, the Gibbs Sampler provides estimates of the mixing matrix $\mathbf{A}$, the sources $\tilde{\mathbf{S}}$ and the noise standard deviation σ. Equation (2.6) can then be rewritten as

$$\tilde{\mathbf{X}} = \hat{\mathbf{A}}\hat{\tilde{\mathbf{S}}} + \hat{\tilde{\mathbf{N}}}, \tag{2.19}$$

where the estimates are denoted $\hat{\mathbf{A}}$, $\hat{\tilde{\mathbf{S}}}$ and $\hat{\tilde{\mathbf{N}}}$. We perform the inverse transform on (2.19) by decomposing $\hat{\tilde{\mathbf{S}}}$ on the basis $\boldsymbol{\Phi}$ from which we obtain $\hat{\mathbf{S}}$.

$$\hat{\mathbf{S}} = \hat{\tilde{\mathbf{S}}}\boldsymbol{\Psi}^{-1} = \hat{\tilde{\mathbf{S}}}\boldsymbol{\Psi}^{T} = \hat{\tilde{\mathbf{S}}}\boldsymbol{\Phi}. \tag{2.20}$$

This concludes our discussion of the algorithm. A block diagram of the process is shown in Figure 2.2.

2.2.3 Sparsity

As shown in Figure 2.2, the representation of the sources is assumed to be sparse before the Gibbs Sampler is applied. Also, finding a sparse representation of a signal or an image is desired in many applications. It can be used in compression [48, 59, 79], regularization in inverse problems and feature extraction. In fact, the success of the JPEG2000 can be attributed to the sparsity of the wavelet coefficients of the image [10]. Furthermore, one of the most natural and effective priors in Bayesian theory for signal estimation is the existence of a sparse representation over a suitable basis or dictionary. Combining sparsity and overcompleteness, which will be discussed in Chapter 3, has been successfully used as a technique in dynamic range compression in images [44].

Sparsity Index

A useful metric for sparsity is the sparsity index [32, 104] of the signal $\tilde{\mathbf{s}}_i$ is defined as

$$\xi \triangleq \frac{\|\tilde{\mathbf{s}}_i\|_1}{\|\tilde{\mathbf{s}}_i\|_2}. \tag{2.21}$$

The smaller ξ is, the sparser the signal. The division by $\|\tilde{\mathbf{s}}_i\|_2$ is simply a normalization. $\|\tilde{\mathbf{s}}_i\|_2$ is equal for different orthonormal bases, so we are effectively comparing the $\mathcal{L}_1$ norms of $\tilde{\mathbf{s}}_i$. Other metrics based on the $\mathcal{L}_p$ norm exist. These would be expressed as $\xi_p \triangleq \|\tilde{\mathbf{s}}_i\|_p/\|\tilde{\mathbf{s}}_i\|_2$ for $0 < p < 2$. See for example [78] or [104] for more details. Besides, one can try to fit a Student t pdf to the histogram of the sources, estimate the corresponding value of α, which governs the concentration of coefficients near zero. This value of α would also be an alternative measure of sparsity. However, we will persist with the use of ξ as defined in equation (2.21) as our measure of sparsity. This sparsity measure was also used in [43].

Illustration

To illustrate the use ξ, in Figure 2.3a, a plot of a length $N = 65536$ musical signal is shown. It is known that the MDCT [28] produces a sparse representation of audio signals, and in particular musical signals, and this is shown in Figure 2.3b. The Orthogonal, Periodized[2], Discrete Wavelet Transform (using Symmlets of order 8) is also taken and shown in Figure 2.3c. Keeping in mind that the dyadic length of the signal is $J = \log_2(N) = 16$ and the length of the filter $M = 16$, the coarsest scale is chosen to be $L = 5$. Typically [17], $L \ll J$ and $M < 2^{L+1}$.

The normalized histogram of the original signal is plotted in Figure 2.3d and the corresponding sparsity index $\xi = 195.7$. The normalized histogram of the MDCT transformed signal is shown in Figure 2.3e. The histograms are normalized so their areas integrate to unity. One immediately observes that most of the coefficients are near zero implying that the signal is sparse in the MDCT domain. Reassuringly, the corresponding sparsity index $\xi = 50.1$. Finally, the value of the sparsity index for signal in the DWT domain is $\xi = 104.7$, indicating that the DWT produces coefficients that are not as sparse as the MDCT. The normalized histogram for the coefficients of the signal in the DWT domain is shown in Figure 2.3f.

[2] Periodized to ensure that the length of the transform is equal to the length of the signal.

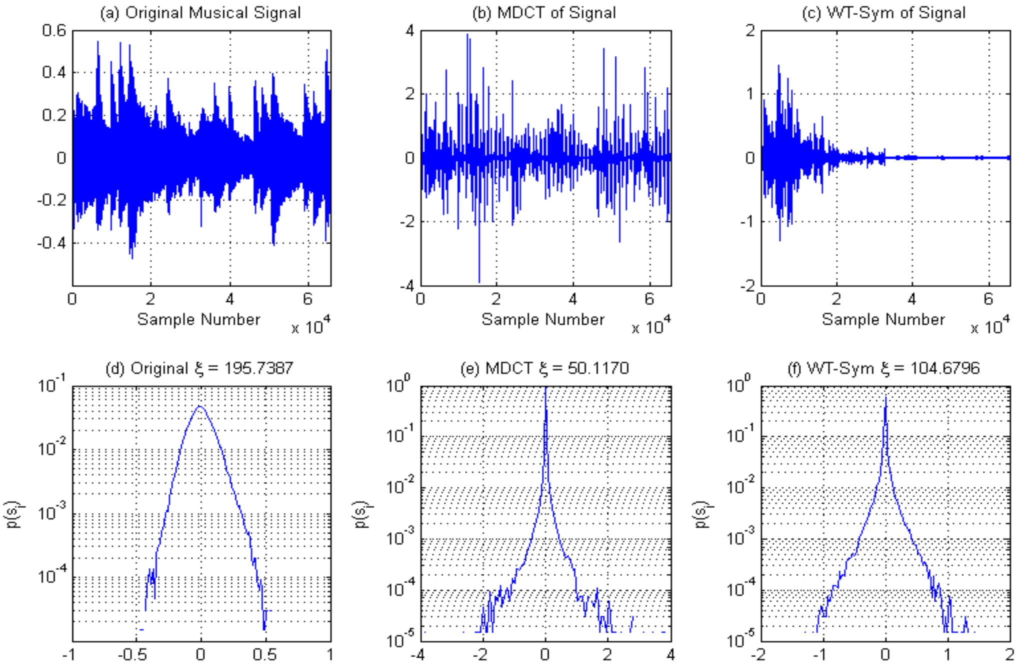

Figure 2.3: Statistical properties of audio signals

2.3 Orthonormal Bases

In the following discussion the row vector $\mathbf{s}_i$, for $1 \leq i \leq n$, is the i^{th} length N source sequence and its orthonormal transform is $\tilde{\mathbf{s}}_i$. $s_{i,t}$ is the t^{th} element of the source vector $\mathbf{s}_i$ while $\tilde{s}_{i,k}$ is the k^{th} element of transformed vector $\tilde{\mathbf{s}}_i$. Different bases $\mathbf{\Phi} = \mathbf{\Psi}^T$ will be introduced and used to transform the vector $\mathbf{s}_i$ to attain sparsity before the Gibbs Sampler is applied. Later, we will show that the degree of sparsity is closely related to separation quality. Transforms are discussed in greater detail in [49] where quantization effects are also considered. An orthonormal basis [107], which is a collection $\{\boldsymbol{\phi}_p : 1 \leq p \leq N\}$ has to satisfy the following two conditions.

- Orthogonality: If $1 \leq p, q \leq N$ and $p \neq q$, then $\langle \boldsymbol{\phi}_p, \boldsymbol{\phi}_q \rangle = \boldsymbol{\phi}_p^T \boldsymbol{\phi}_q = 0$;
- Normalization: For each p, $\|\boldsymbol{\phi}_p\| = 1$.

2.3.1 Discrete Cosine Transform (DCT)

The DCT [100] is related to the ubiquitous Discrete Fourier Transform (DFT). Martucci [81] proved that there are eight ways in total to extend a finite length signal $\mathbf{s}$ to make it symmetric. This leads to different versions of the Discrete Cosine Transform (DCT) and Discrete Sine Transform (DST). We consider here the Discrete Cosine IV basis [59, 79, 92]. A signal $\mathbf{s}_i$ of N samples is extended to a signal $\mathbf{s}_i^{(IV)}$ of period $4N$ with respect to $-1/2$

and antisymmetric with respect to $N-1/2$ and $-N+1/2$. It can be shown [79] that since $s_{i,t}^{(IV)} = s_{i,t}$ for $0 \leq t < N$, $\mathbf{s}_i$ can be written as a linear sum of odd frequency cosines. Hence, the family

$$\left\{ \sqrt{\frac{2}{N}} \cos\left[\frac{\pi}{N}\left(k+\frac{1}{2}\right)\left(t+\frac{1}{2}\right)\right]\right\}_{0\leq k<N} \tag{2.22}$$

is an orthonormal basis of $\mathbb{C}^N$. This is the discrete cosine transform IV (DCT-IV). The DCT-IV of a signal of size N is related to Discrete Fourier Transform (DFT) of a complex signal of size $N/2$ [35]. Hence, we need $\mathcal{O}(N\log_2 N)$ operations to calculate the DCT-IV using a method similar to the Fast Fourier Transform (FFT) [26].

The DCT has excellent energy compaction properties [92] and hence the major application of the DCT is in signal and image compression [61, 99], where it is a key part of many standardized algorithms, including JPEG[3] [48, 59]. This is because it is a good substitute for the optimal Karhunen-Loève transform (KLT) [34, 65] as the DCT functions approximate the eigenvectors of the autocorrelation matrix $\mathbf{R_{xx}}$ of a first order Gauss-Markov process [24].

2.3.2 Modified Discrete Cosine Transform (MDCT)

Malvar discovered that one can create orthogonal bases with smooth windows modulated by a cosine IV basis [80]. The support of a window g_p is $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$, with $l_p = a_{p+1} - a_p$ as seen in Figure 2.4. The design of the windows g_p implies symmetry and quadrature properties on overlapping intervals such that

$$g_p(t) = g_{p+1}(2a_{p+1} - t), \tag{2.23}$$

$$g_p^2(t) + g_{p+1}^2(t) = 1, \tag{2.24}$$

for all $t \in [a_{p+1} - \eta_{p+1}, a_{p+1} + \eta_{p+1}]$. This leads directly to the local cosine transform [79, 112], also called the modified discrete cosine transform (MDCT) [28] or lapped orthogonal transform (LOT) [80]. This overlapping, in addition to the energy-compaction properties of the DCT, makes the MDCT especially suited for several audio compression applications [96], since it helps to avoid artifacts resulting from the discontinuities at the block boundaries. Lapped orthogonal bases are discretized by replacing the orthogonal basis in $\mathcal{L}^2(\mathbb{R})$ with a discrete basis on $\mathbb{C}^N$ and uniformly sampling the windows $g_p(t)$ to obtain the smooth and discretized windows $g_{p,t}$. Also let N_f be the total number of frames or equivalently, windows. We restrict p to fall in the range $1 \leq p \leq N_f$ such that $\sum_{p=1}^{N_f} l_p = N$. Discrete local cosine bases are then derived with cosine-IV bases as

[3] However, JPEG2000 [10] uses wavelets instead of the DCT.

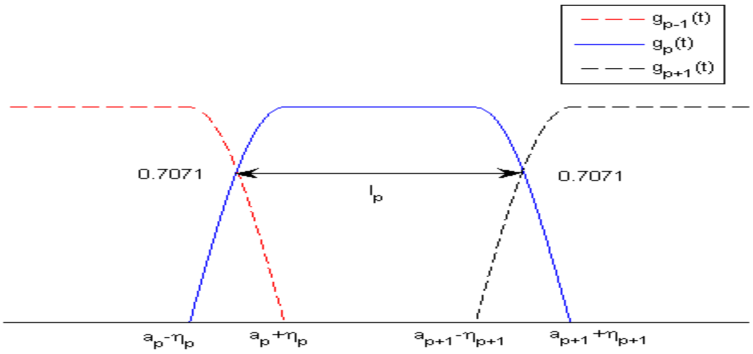

Figure 2.4: Smooth windows used in the MDCT

discussed in the previous section. Hence, the family of local cosine functions

$$\left\{ g_{p,k,t} = g_{p,t}\sqrt{\frac{2}{l_p}}\cos\left[\pi\left(k+\frac{1}{2}\right)\frac{t-a_p}{l_p}\right]\right\}_{0\le k<l_p, 1\le p\le N_f} \tag{2.25}$$

is an orthonormal basis for $\mathbb{C}^N$. Using a folding procedure, Malvar [80] devised a fast algorithm to compute the MDCT in $\mathcal{O}\left(\sum_{p=1}^{N_f} l_p \log_2 l_p\right)$ operations where l_p and N_f are defined above.

2.3.3 Wavelet Transforms: Vaidyanathan (WT-Vai)

The topic of wavelets is certainly the most trendy in signal processing and applied mathematics today. The first wavelet and the only example for a long time was found by Haar [52]. Over time, more wavelets as well as faster algorithms to compute them were devised. For an introduction to wavelets, see [27, 108, 112]. Wavelets represent a signal at different scales (See Figure 2.5) and the DWT is the collection of subband signals. Wavelets are used in various fields from image compression [48] to mechanical vibrations [90]. In this project, 2 different discrete wavelets are used. They were implemented using the Fast Wavelet Transform (FWT) [79, 108] by considering the filter bank approach in Figure 2.5. $H_0(z) = \sum_{t=-\infty}^{\infty} h_{0,t}z^{-t}$ is the z-transform [93] of the impulse response h_0 and $h_{0,t}$ is the t^{th} element of h_0. h_0 and h_1 are typically low-pass and high-pass filters respectively. The reconstruction filters can be derived from the synthesis filters and it is easy to show [59] that they have to satisfy the perfect reconstruction (PR) conditions.

$$H_0(-z)G_0(z) + H_1(-z)G_1(z) = 0, \tag{2.26}$$

$$H_0(z)G_0(z) + H_1(z)G_1(z) = 2, \tag{2.27}$$

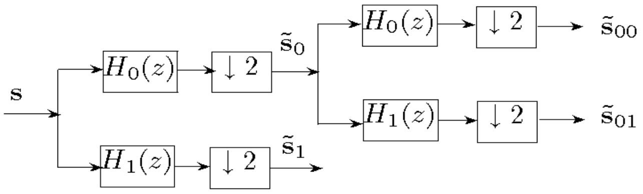

Figure 2.5: Filter bank/Multiscale representation of a signal

where $G_0(z)$ and $G_1(z)$ are the z-transforms of the reconstruction filters g_0 and g_1 respectively. The first wavelet we will use was designed by Vaidyanathan and Hoang [111]. This wavelet is used primarily because the constraint is important in optimizing the transform code of audio and speech signals [79, 111]. The phase is highly non-linear, which results in audible and visible coding artifacts [48, 112] in image processing. As a result, linear phase filters are always preferred. The Vaidyanathan filter is of length $M_{Vai} = 24$ and this wavelet will henceforth be abbreviated by WT-Vai.

2.3.4 Wavelet Transforms: Symmlets (WT-Sym)

Unfortunately, the ubiquitous Daubechies wavelets [27] are very asymmetric and hence they have non-linear phase and significant sidelobes [79]. To obtain a symmetric wavelet, the filter h_0 must be symmetric with respect to its center of support. Symmlet filters [79] are optimized such that they are as symmetric as possible whilst still satisfying the perfect reconstruction equations (2.26) and (2.27) and attaining an almost linear phase. The length of the Symmlet of order $p = 8$ is $M_{Sym} = 2p = 16$ and this wavelet will henceforth be abbreviated by WT-Sym.

A fast orthogonal wavelet transform was devised by Mallat [79, 108] and the computation of the DWT for both WT-Vai and WT-Sym requires at most $2MN$ operations where M is the length of the filter. This method is in fact a classical scheme known in signal processing as a *two-channel subband coder* or *filter banks* (See Figure 2.5).

2.3.5 Wavelet Packet Best Basis (WPBB)

In wavelet analysis, a signal is split into an approximation (low-pass component) and a detail (high-pass component). The approximation is then itself split into a second-level approximation and detail, and the process is repeated. For an n-level decomposition, there are $n + 1$ possible ways to decompose or encode the signal. Wavelet Packet analysis involves splitting not only the approximation but also the detail. This provides more than $2^{2^{n-1}}$ ways to represent a signal. Certainly, there exists ways that are, in some sense, better than others.

It turns out that it is possible to seek the best one by some criterion. If the algorithm is sufficiently cheap, then it is possible to assign to a particular signal its own adapted basis, or *basis of adapted waveforms*. Coifman and Wickerhauser [25, 114] have proposed a fast method based on dynamic programming to adaptively choose a single orthonormal basis that is the 'best basis'. If we define $\mathcal{E}$ as an entropy operator such that $\mathcal{E}(\tilde{\mathbf{s}}_i) = \sum_k e(\tilde{s}_{i,k})$, where $e(y)$ is a scalar function of a scalar argument, the optimization program can be stated as

$$\boldsymbol{\Phi}^* = \arg\min_{\boldsymbol{\Phi}} \mathcal{E}(\mathbf{s}_i \boldsymbol{\Phi}^{-1}) \qquad \text{subject to} \qquad \boldsymbol{\Phi}\boldsymbol{\Phi}^T = \mathbf{I}_N. \tag{2.28}$$

In this project, $e(y)$ was chosen to be $|y|$ because the sparsity index is defined in terms of the $\mathcal{L}_1$ norm. Another choice of $e(y)$ is the entropy $e(y) = -y \log_e y$ introduced by Shannon [106]. The algorithm in most cases delivers near-optimal sparsity representations [20]. In particular, when the signal has a sparse representation in an orthonormal basis taken from the library, the algorithm will work well. Using a brute force approach to find the best basis requires a total of $N2^{N/2}$ operations, which is computationally implausible. The fast dynamic programming algorithm that Coifman and Wickerhauser [25, 114] finds the best wavelet packet basis with $\mathcal{O}(N \log_2 N)$ operations, by taking advantage of the tree structure of the wavelet decomposition. Decomposing the observations on this basis requires another $2MN$ operations where M is the length of the filter. An similar formulation is presented in [21], where Cichocki and Amari discuss means to perform BSS and ICA adaptively.

Note that since we have no knowledge of the sources *a priori*, the optimization in (2.28) is done on $\mathbf{x}_1$, the first of the mixtures. This produces an optimal orthonormal basis $\boldsymbol{\Phi}_1 = \boldsymbol{\Psi}_1^T$ based on $\mathbf{x}_1$ that is, in general, not guaranteed to be the optimal one for any of the n original sources. Though sub-optimal, we hope that it will produce a sparse representation of the original sources. One can alternatively choose to perform the optimization in (2.28) for the sequence $\mathbf{x}_2$ and obtain another orthonormal basis $\boldsymbol{\Phi}_2 = \boldsymbol{\Psi}_2^T$.

2.3.6 No Transform (NT)

For the sake of comparison, the Gibbs Sampler was also applied to the sources directly, without transform. Hence, $\boldsymbol{\Psi} = \boldsymbol{\Phi} = \mathbf{I}_N$ and

$$\tilde{\mathbf{s}}_i = \mathbf{s}_i \mathbf{I}_N = \mathbf{s}_i, \tag{2.29}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. This is known in the literature [114] as the *standard orthonormal basis*.

Transform	DCT	MDCT	WT-Vai	WT-Sym	WPBB
Complexity $\mathcal{O}(\cdot)$	$N\log_2 N$	$\sum_{p=1}^{N_f} l_p \log_2 l_p$	NM_{Vai}	NM_{Sym}	$N\log_2 N$

Table 2.1: Comparison of computational complexities for the various transforms

2.3.7 Summary of Orthonormal Bases

A summary of the computational complexity of each transform is given in Table 2.1. Note that $M_{Vai} = 24$ is the length of the Vaidyanathan filter and the $M_{Sym} = 16$ is the length of the Symmlet of order 8. This completes our discussion on the various transforms that will be used. During the course of the project, some other transforms were also considered but not implemented. These included

- Gabor Frames [115, 116] or Short-Time Fourier Transform (STFT) [92]
  This was not implemented because it generates complex sequences from real sequences. However, in Chapter 3, we will be looking at the Short-Time Discrete Cosine Transform (STDCT), a close relative of the Gabor transform. The STDCT has the advantage over the STFT because it generates overcomplete *real* sequences.

- Discrete Sine Transform (DST) [59, 92]
  Except for a phase shift[4], this transform is very similar to the DCT as described section 2.3.1 and hence it was not implemented.

2.4 Performance Measures

Any algorithm has to be evaluated. In this section, we focus on the results of [40] and [51] where the authors addressed issues related to the evaluation of performance of BSS algorithms. The authors factored into their consideration that BSS algorithms will only be able to recover the sources up to a permutation and a gain as explained in section 2.1.1. The i^{th} length N signal will be denoted $\mathbf{s}_i$, for $1 \le i \le n$, and its estimate is $\hat{\mathbf{s}}_i$.

Let us assume that the source signals $\mathbf{s}_i$ are mutually orthogonal and the noise signals are always assumed to be mutually orthogonal to each other and to the sources. Then, the estimated source $\hat{\mathbf{s}}_i$ has an orthogonal decomposition

$$\hat{\mathbf{s}}_i = \left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} + \mathbf{e}_{interf} + \mathbf{e}_{noise} + \mathbf{e}_{artif}, \tag{2.30}$$

where $\langle \hat{\mathbf{s}}_i, \mathbf{s}_i/\|\mathbf{s}_i\| \rangle \mathbf{s}_i/\|\mathbf{s}_i\|$ is the contribution to the true source, $\mathbf{e}_{interf}$ is the error term due to interference of the other sources and $\mathbf{e}_{noise}$ is the error term due to additive noise and $\mathbf{e}_{artif} \overset{\triangle}{=} \hat{\mathbf{s}}_i - \langle \hat{\mathbf{s}}_i, \mathbf{s}_i/\|\mathbf{s}_i\| \rangle \mathbf{s}_i/\|\mathbf{s}_i\| - \mathbf{e}_{interf} - \mathbf{e}_{noise}$ is the error term attributed to the numerical

[4] However, in image coding, this effect of this phase shift results in visible differences when compared to the DCT [59].

artifacts of the separation algorithm. In general, the sources may be correlated but still linearly independent. We define $P_{\mathbf{s}}(\hat{\mathbf{s}}_i) \triangleq \sum_{l=1}^{n} \langle \hat{\mathbf{s}}_i, \mathbf{s}_l/\|\mathbf{s}_l\|\rangle \, \mathbf{s}_l/\|\mathbf{s}_l\|$ as the orthogonal projector onto the sources' span and $P_{\mathbf{s},\mathbf{n}}(\hat{\mathbf{s}}_i) \triangleq P_{\mathbf{s}}(\hat{\mathbf{s}}_i) + \sum_{l=1}^{m} \langle \hat{\mathbf{s}}_i, \mathbf{n}_l/\|\mathbf{n}_l\|\rangle \, \mathbf{n}_l/\|\mathbf{n}_l\|$ as the orthogonal projector onto the span of both the sources and the noise signals. The decomposition (2.30) still holds with

$$\begin{aligned} \mathbf{e}_{interf} &\triangleq P_{\mathbf{s}}(\hat{\mathbf{s}}_i) - \left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|}, && (2.31) \\ \mathbf{e}_{noise} &\triangleq P_{\mathbf{s},\mathbf{n}}(\hat{\mathbf{s}}_i) - P_{\mathbf{s}}(\hat{\mathbf{s}}_i), && (2.32) \\ \mathbf{e}_{artif} &\triangleq \hat{\mathbf{s}}_i - P_{\mathbf{s},\mathbf{n}}(\hat{\mathbf{s}}_i). && (2.33) \end{aligned}$$

Equipped with these definitions, we are now ready to define the *Source to Distortion Ratio* (SDR), the *Source to Interference Ratio* (SIR), the *Source to Artifacts Ratio* (SAR) and the *Source to Noise Ratio* (SNR).

2.4.1 Source to Distortion Ratio (SDR)

The total relative distortion is defined as

$$D_{total} \triangleq \frac{\|\hat{\mathbf{s}}_i\|^2 - \left|\left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle\right|^2}{\left|\left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle\right|^2}. \tag{2.34}$$

where $\|\cdot\|$ is the $\mathcal{L}_2$ norm. Note that when the estimate source is orthogonal to the original source i.e. $|\langle \hat{\mathbf{s}}_i, \mathbf{s}_i\rangle| \to 0$ then $D_{total} \to +\infty$. Thus, the SDR (dB) is defined as

$$\mathrm{SDR} \triangleq 10 \log_{10} D_{total}^{-1}. \tag{2.35}$$

The definition of D_{total} corresponds to the ratio of the energy of the two terms in the decomposition $\hat{\mathbf{s}}_i = \langle \hat{\mathbf{s}}_i, \mathbf{s}_i/\|\mathbf{s}_i\|\rangle \, \mathbf{s}_i/\|\mathbf{s}_i\| + \mathbf{e}_{total}$ where $\mathbf{e}_{total} = \mathbf{e}_{interf} + \mathbf{e}_{noise} + \mathbf{e}_{artif}$ is the error term, which is orthogonal (uncorrelated) to the contribution of the true source. In fact, by the Pythagorean theorem, $\|\mathbf{e}_{total}\|^2 = \|\hat{\mathbf{s}}_i\|^2 - |\langle \hat{\mathbf{s}}_i, \mathbf{s}_i/\|\mathbf{s}_i\|\rangle|^2$. The SDR provides an overall separation performance criterion. It is a *global* measure of distortion.

2.4.2 Source to Interference Ratio (SIR)

The SIR measures the level of interferences from the other sources in each source estimate. One can define the relative distortion due to interferences D_{interf} as

$$D_{interf} \triangleq \frac{\|\mathbf{e}_{interf}\|^2}{\left|\left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle\right|^2}, \tag{2.36}$$

and consequently the SIR as

$$\mathrm{SIR} \triangleq 10 \log_{10} D_{interf}^{-1}. \tag{2.37}$$

2.4.3 Source to Artifacts Ratio (SAR)

The SAR measures the level of numerical artifacts in each source estimate. The relative distortion due to algorithmic effects D_{artif} is

$$D_{artif} \triangleq \frac{\|\mathbf{e}_{artif}\|^2}{\left\| \left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} + \mathbf{e}_{interf} + \mathbf{e}_{noise} \right\|^2}, \tag{2.38}$$

Hence, the SAR can be defined as

$$\mathrm{SAR} \triangleq 10 \log_{10} D_{artif}^{-1}. \tag{2.39}$$

2.4.4 Source to Noise Ratio (SNR)

Finally, the relative distortion due to additive noise D_{noise} is

$$D_{noise} \triangleq \frac{\|\mathbf{e}_{noise}\|^2}{\left\| \left\langle \hat{\mathbf{s}}_i, \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} \right\rangle \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|} + \mathbf{e}_{interf} \right\|^2}. \tag{2.40}$$

Hence, the Source to Noise Ratio (SNR) is

$$\mathrm{SNR} \triangleq 10 \log_{10} D_{noise}^{-1}. \tag{2.41}$$

The SNR measures the error due to the additive noise on the sensors.

2.5 Conclusions

This chapter has considered the BSS method that will be used. Firstly, the model was introduced and BSS assumptions were stated. The Gibbs Sampler was also presented. The central theme of this chapter is signal expansion in a transformed domain. The sources have to be *sparse* on a given basis and so a series of transforms were introduced and their computational complexities were compared. As we will see, using a transform that minimizes some form of entropy measure will result in significant advantages when the Gibbs Sampler is applied. Finally, we introduced four different performance metrics to evaluate the performances of various BSS algorithms.

Chapter 3

Overcomplete Dictionaries

In recent years, there has been a growing interest in representing signals using overcomplete dictionaries to achieve a sparse representation. For more details, see the papers by Aharon [1], which discusses the K-SVD algorithm and Mitianoudis [86] and Lee [70], which discuss the use of overcomplete dictionaries in underdetermined BSS. Besides, Malioutov [78] demonstrates how to obtain optimal sparse representations using overcompleteness. Finally, the processes to learn these dictionaries is presented by Lewicki [72].

Using an overcomplete dictionary that leads to a sparse representation can either be chosen as a pre-specified linear transform, or designing its content to fit a given set of signal examples i.e. *adaptive*. The former yields dictionaries with respect to which representations can be readily found. This is the case with overcomplete wavelet and Short-Time Fourier Transform (STFT). Tight frames [79, 112] are usually preferred as the inversion is done using a pseudo-inverse [94] operation. The success of these dictionaries depends on how suitable these sparsely describe the signals in question. We will be considering two pre-specified overcomplete dictionaries in this report. In particular, we will be examining the effectiveness of the Short-Time Discrete Cosine Transform (STDCT) [92] and the Hybrid Transform (HT) [29, 30, 36, 87, 88]. The HT is a union of two bases, namely the MDCT basis and the WT-Vai basis. Their performances will be compared to the canonical orthonormal bases discussed in section 2.3.

3.1 Analysis and Synthesis

The analysis and synthesis operations are central to the development of overcomplete representations. As usual, we will be adopting the notation $\boldsymbol{\Psi}$ for the analysis operator. $\boldsymbol{\Psi}$ has N rows and K columns and $K > N$. The analysis operation can be written as

$$\tilde{\mathbf{S}} = \mathbf{S}\boldsymbol{\Psi} \tag{3.1}$$

and $\boldsymbol{\Phi}$ is the synthesis operator such that

$$\mathbf{S} = \tilde{\mathbf{S}}\boldsymbol{\Phi}. \tag{3.2}$$

Since neither $\boldsymbol{\Psi} \in \mathbb{R}^{N \times K}$ nor $\boldsymbol{\Phi} \in \mathbb{R}^{K \times N}$ is a square matrix, they cannot be inverted. It is well known [20] that many natural signals can be sparsely represented in a proper signal dictionary $\boldsymbol{\Phi}$, which is the synthesis operator. The pseudo-inverse of this dictionary is the analysis operator

$$\boldsymbol{\Psi} \triangleq \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T. \tag{3.3}$$

In linear algebra [107], this is also known as the *left inverse* such that $\boldsymbol{\Psi}\boldsymbol{\Phi} = \mathbf{I}_N$. $\boldsymbol{\Psi}$ can be post-multiplied to (2.2) to give (2.6) as usual. In the following, we assume that $\tilde{\mathbf{S}}$ is a sparse matrix. For the Gibbs Sampler to estimate the sources and the mixing matrix efficiently, we require the sources in the transformed domain to be sparse. It is thus necessary to carry out the analysis operation. The synthesis operator is related to the analysis operator via the following:

$$\boldsymbol{\Phi} \triangleq \boldsymbol{\Psi}^T\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right)^{-1}. \tag{3.4}$$

This is known as the *right inverse* such that $\boldsymbol{\Psi}\boldsymbol{\Phi} = \mathbf{I}_N$ again. The reconstruction of the sources from the estimated sources in the transformed domain requires the use of the synthesis operator. If we post-multiply $\boldsymbol{\Phi}$ to (2.19) we obtain

$$\tilde{\mathbf{X}}\boldsymbol{\Phi} = \hat{\mathbf{A}}(\hat{\tilde{\mathbf{S}}}\boldsymbol{\Phi}) + \hat{\tilde{\mathbf{N}}}\boldsymbol{\Phi}. \tag{3.5}$$

Noting that $\tilde{\mathbf{X}}\boldsymbol{\Phi} = \mathbf{X}\boldsymbol{\Psi}\boldsymbol{\Phi} = \mathbf{X}\mathbf{I}_N = \mathbf{X}$, we can equivalently express (3.5) as

$$\mathbf{X} = \hat{\mathbf{A}}\hat{\mathbf{S}} + \hat{\mathbf{N}}, \tag{3.6}$$

where $\hat{\mathbf{S}} \triangleq \hat{\tilde{\mathbf{S}}}\boldsymbol{\Phi}$ and $\hat{\mathbf{N}} \triangleq \hat{\tilde{\mathbf{N}}}\boldsymbol{\Phi}$. Note that all we have done here is verify that the operations result in perfect reconstruction (PR).

3.2 Overcomplete Dictionaries: Examples

The STDCT and the HT were chosen. The STDCT results in an overcomplete representation of the real signal of length N. There are K real coefficients in the transformed domain. We contrast this to using a pair of bases [36], a MDCT and a wavelet basis. This results in $K = 2N$ coefficients in the transformed domain. The advantage of this transform as we will see in Chapter 4 and 5 is that it tries to capture both the tonals and

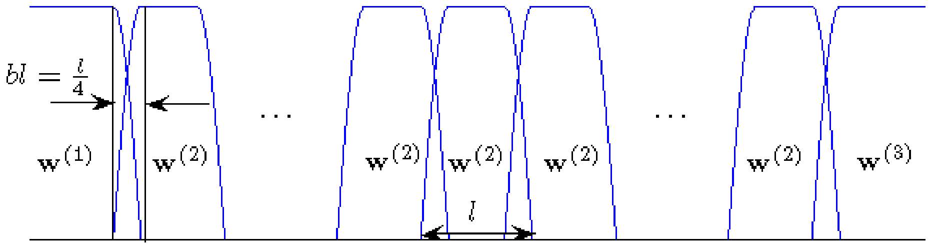

Figure 3.1: Window functions used in the STDCT.

the transients [29] of the audio signal and plausibly leads to sparser representations.

3.2.1 Short-Time Discrete Cosine Transform (STDCT)

The STDCT transforms a length N signal $\mathbf{s}_i$ to a length $K = l \times N_f$ signal $\tilde{\mathbf{s}}_i$. l and N_f are defined as the window length and number of frames respectively. This transform is overcomplete because $K > N$. Before the DCT-IV was taken, a tapering window was first applied to smooth abrupt edges. We will see that this operation results in a sparser representation. The windows are very similar to the one used in the MDCT. These are shown in Figure 3.1. The minor changes are:

- The first and the last windows have a value of unity at the edges. This is to ensure that $\mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{I}_N$ as will be detailed in Lemma 1. The reconstruction or synthesis step, in equation (3.4), will be trivial if this constraint is enforced.

- The overlap between adjacent windows is $b = 1/4 = 25\%$ and $b = l_{overlap}/l$ is the overlap fraction. The decomposition is overcomplete by about 25% and the length of the overlapping segments is $bl = l/4$.

The windows here satisfy the symmetry (2.23) and quadrature (2.24) properties. Let the $l \times l$ window matrices be

$$\mathbf{W}^{(k)} = \mathrm{diag}\left(\mathbf{w}^{(k)}\right), \qquad k = 1, 2, 3, \tag{3.7}$$

where $\mathbf{w}^{(k)}$ is the length l vector that contains the window coefficients. k takes on the values 1, 2 and 3 which represent the first window, the middle windows and the last window respectively as shown in Figure 3.1. The middle windows are all the same and they differ from the first and the last windows. The analysis operator $\mathbf{\Psi}$ is a lapped block diagonal matrix which has the structure as shown in Figure 3.2. Notice the $b = 25\%$ overlap between adjacent blocks. $\boldsymbol{\psi}$ is a $l \times l$, symmetric, orthonormal, DCT-IV matrix that is applied to each of the windowed segments after the windowing has been performed.

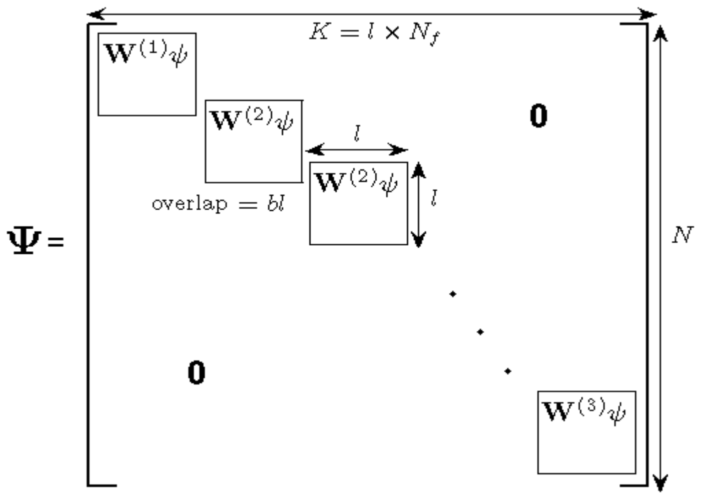

Figure 3.2: Lapped Block Diagonal Structure of $\boldsymbol{\Psi}$; Note that $K > N$

Lemma 1. *If the windows satisfy the symmetry* (2.23) *and quadrature properties* (2.24), *the window matrices are defined as in* (3.7) *and consequently, the analysis operator* $\boldsymbol{\Psi}$ *takes the form as shown in Figure* 3.2 *and* $b \in (0, 1/2]$, *then*

$$\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right)^{-1} = \mathbf{I}_N. \tag{3.8}$$

Proof. Computing $\boldsymbol{\Psi}\boldsymbol{\Psi}^T$ directly, we obtain a diagonal matrix

$$\boldsymbol{\Psi}\boldsymbol{\Psi}^T = \operatorname{diag}\begin{bmatrix} \left(w_1^{(1)}\|\underline{\psi}_1\|\right)^2 \\ \vdots \\ \left(w_{(1-b)l}^{(1)}\|\underline{\psi}_{(1-b)l}\|\right)^2 \\ \left(w_{(1-b)l+1}^{(1)}\|\underline{\psi}_{(1-b)l+1}\|\right)^2 + \left(w_1^{(2)}\|\underline{\psi}_1\|\right)^2 \\ \vdots \\ \left(w_l^{(1)}\|\underline{\psi}_l\|\right)^2 + \left(w_{bl}^{(2)}\|\underline{\psi}_{bl}\|\right)^2 \\ \left(w_{bl+1}^{(2)}\|\underline{\psi}_{bl+1}\|\right)^2 \\ \vdots \\ \left(w_{(1-b)l}^{(2)}\|\underline{\psi}_{(1-b)l}\|\right)^2 \\ \vdots \\ \left(w_{bl+1}^{(3)}\|\underline{\psi}_{bl+1}\|\right)^2 \\ \vdots \\ \left(w_l^{(3)}\|\underline{\psi}_l\|\right)^2 \end{bmatrix} = \operatorname{diag}\begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbf{I}_N \tag{3.9}$$

where $w_n^{(k)} \in \mathbb{R}$ is the n^{th} element of the vector $\mathbf{w}^{(k)} \in \mathbb{R}^l$ and the vector $\underline{\psi}_i \in \mathbb{R}^l$ is the i^{th} row of the DCT matrix $\boldsymbol{\psi} \in \mathbb{R}^{l \times l}$. This is because, by the orthogonality property of the matrix $\boldsymbol{\psi}$, $\underline{\psi}_i^T \underline{\psi}_j = 0$ if $i \neq j$. Note also that all the rows and columns are normalized, $\|\underline{\psi}_i\|^2 = 1$ for $1 \leq i \leq l$ and the windows satisfy the properties

$$\begin{aligned} w_n^{(1)} &= 1, \qquad 1 \leq n \leq (1-b)\,l, && (3.10)\\ w_n^{(2)} &= 1, \qquad bl + 1 \leq n \leq (1-b)\,l, && (3.11)\\ w_n^{(3)} &= 1, \qquad bl + 1 \leq n \leq l. && (3.12) \end{aligned}$$

It remains to observe that remaining elements are also unity by the quadrature property.

$$\begin{aligned} \left(w_{(1-b)l+n}^{(1)}\right)^2 + \left(w_n^{(2)}\right)^2 &= 1, && (3.13)\\ \left(w_{(1-b)l+n}^{(2)}\right)^2 + \left(w_n^{(2)}\right)^2 &= 1, && (3.14)\\ \left(w_{(1-b)l+n}^{(2)}\right)^2 + \left(w_n^{(3)}\right)^2 &= 1, \qquad 1 \leq n \leq bl. && (3.15) \end{aligned}$$

Hence *all* the diagonal entries of the matrix $\boldsymbol{\Psi}\boldsymbol{\Psi}^T$ equal to unity. The off-diagonal elements are equal to zero. This completes our proof of (3.8) and Lemma 1. □

Equipped with Lemma 1, the reconstruction step is now simple. Referring to equation (3.4), the synthesis operator $\boldsymbol{\Phi}$ takes the form

$$\boldsymbol{\Phi} = \boldsymbol{\Psi}^T. \tag{3.16}$$

The computational complexity of the STDCT is of the order $\mathcal{O}(K \log_2 l)$.

3.2.2 Hybrid Transforms (HT)

In [29], the authors were concerned with hybrid signal models that included components of different kinds. More specifically, they limited their investigations to additive models of the form

$$s(t) = s_{ton}(t) + s_{tr}(t) + s_r(t), \tag{3.17}$$

where $s_{ton}(t)$ is the tonal component, $s_{tr}(t)$ is the transient component and $s_r(t)$ is the stochastic or residual component. Such models have been considered for modelling and transformation [105], or for encoding or compression [71]. It turns out that the tonal component $s_{ton}(t)$ can be represented well using a MDCT (local cosine) basis. The transient components $s_{tr}(t)$ can be expanded using a wavelet basis. The coefficients of both expansions can be found by considering Hidden Markov Chains (HMCs) and Hidden Markov

Transform	STDCT	HT
Complexity $\mathcal{O}(\cdot)$	$K \log_2 l$	$NM + \sum_{p=1}^{N_f} l_p \log_2 l_p$

Table 3.1: Comparison of computational complexities for the Overcomplete Dictionaries

Models (HMMs) [29]. The residual component $s_r(t)$ is typically taken to be zero.

With these in mind, we set the synthesis operator $\mathbf{\Phi} \in \mathbb{R}^{2N \times N}$ to be a concatenation of two $N \times N$ bases $\mathbf{\Phi}_{MDCT}$ and $\mathbf{\Phi}_{DWT}$ [36]. The DWT basis that we use is the WT-Vai as described in section 2.3.3. Hence,

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_{MDCT} \\ \mathbf{\Phi}_{DWT} \end{bmatrix}. \tag{3.18}$$

We need to find out what the analysis operator $\mathbf{\Psi}$ is. Using equation (3.3) and the orthonormality property of $\mathbf{\Phi}_{MDCT}$ and $\mathbf{\Phi}_{DWT}$, we get

$$\mathbf{\Psi} = \frac{1}{2} \begin{bmatrix} \mathbf{\Phi}_{MDCT}^T & \mathbf{\Phi}_{DWT}^T \end{bmatrix}. \tag{3.19}$$

This is just a straightforward application of the MDCT and the DWT. After the Gibbs Sampling has been performed, one obtains estimates of the sources in the transformed domain as in equation (2.19). $\hat{\hat{\mathbf{S}}}$ can then be used to reconstruct the sources using $\hat{\mathbf{S}} \triangleq \hat{\hat{\mathbf{S}}}\mathbf{\Phi}$. Again this is just a simple inverse MDCT and inverse DWT. Hence, if

$$\hat{\hat{\mathbf{S}}} = \begin{bmatrix} \hat{\hat{\mathbf{S}}}_1 & \hat{\hat{\mathbf{S}}}_2 \end{bmatrix}, \tag{3.20}$$

where $\hat{\hat{\mathbf{S}}}_1$ and $\hat{\hat{\mathbf{S}}}_2$ are $n \times N$ matrices, the reconstructed sources would simply be

$$\hat{\mathbf{S}} = \hat{\hat{\mathbf{S}}}_1 \mathbf{\Phi}_{MDCT} + \hat{\hat{\mathbf{S}}}_2 \mathbf{\Phi}_{DWT}. \tag{3.21}$$

The computational complexity of the HT is simply the sum of the individual complexities of the MDCT and DWT. This is compared to the STDCT in Table 3.1.

3.3 Conclusions

In this chapter, we have presented the motivation and theory behind overcomplete dictionaries. We have discussed two overcomplete dictionaries, namely the STDCT and the HT. In Chapter 5, we will be assessing their performance on the source separation algorithm as shown in Figure 2.2. Now, we shall turn our attention to the performance of the orthonormal transforms.

Chapter 4

Results: Orthonormal Bases

In this chapter, the results of the algorithm presented in Chapter 2 will be tested on different sets of signals. We will show that there exists a close correlation between sparsity of the sources in the transformed domain and the separation quality.

4.1 Introduction and Initialization

In line with the work done in [42], we study the mixing of $n = 3$ audio sources with $m = 2$ observations. The original mixing matrix was chosen to be

$$\mathbf{A} = \begin{bmatrix} 0.8\cos(-\pi/3) & 0.9\cos(-\pi/8) & 0.8\cos(\pi/4) \\ 0.8\sin(-\pi/3) & 0.9\sin(-\pi/8) & 0.8\sin(\pi/4) \end{bmatrix} \tag{4.1}$$

and hence the independent components are given by the principal directions $\vartheta_1 = -\pi/3$, $\vartheta_2 = -\pi/8$ and $\vartheta_3 = \pi/4$. We added i.i.d. Gaussian noise on each observation, which resulted in the value of SNR $\approx$ 16.5 dB. This corresponded to $\sigma = 0.015$. The signals were sampled at 8 kHz with length $N = 65536$ ($\approx 8s$).

4.1.1 Signals and Estimation

Different orthonormal transforms, as described in section 2.3, were taken on the 4 sets of signals. These signals were chosen because they represent the spectrum of audio signals of interest. They included speech, musical and percussion signals as well as a combination of the three. Each Gibbs Sampling process was run until convergence was observed. The MMSE estimates of $\mathbf{A}$ and $\tilde{\mathbf{S}}$ were then computed from the final 1000 samples. The different parameters were initialized with the values as shown in Table 4.1. Finally, the discrete values from which the degrees of freedom α_i are sampled from are chosen from a set of values linearly spaced between 0.05 and 5, with step size 0.05.

$\tilde{\mathbf{S}}$	$\mathbf{r}_2$	σ	$\mathbf{V}$	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}$
`ones(n × N)`	$\begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$	0.1	`ones(n × N)`	`ones(1,n)`	`0.01.*ones(1,n)`

Table 4.1: Initialization of the parameters to be estimated

Transform Number	1	2	3	4	5	6
Transform Abbrev.	DCT	MDCT	WT-Vai	WT-Sym	WPBB	NT

Table 4.2: Correspondence between Transform Number and the orthonormal transform used

4.1.2 Representation of transforms and performance measures

In the following discussion, for convenience, integers will be used to denote the type of transform and the performance metric used. These are summarized in Tables 4.2 and 4.3. For example, transform 1 is the Discrete Cosine Transform (DCT) while performance number 3 is the Source to Artifacts Ratio (SAR). In Figures 4.3a, b, c and d, we see that there is also a seventh transform method used. It represents the effect of applying the best basis algorithm as described in section 2.3.5, to the sources *directly*. Hence, this provides the optimal value of ξ for the sources. Recall that in transform 5 (WPBB), we apply the best basis algorithm to the first of the mixture components $\mathbf{x}_1$ as we do not know the nature of the sources *a priori*. We then obtain $\boldsymbol{\Psi}_1$ and use this analysis operator to decompose the sources. Although we calculated the sparsity indices based on the sources themselves, we did not and certainly, could not implement it because the orthonormal transform $\boldsymbol{\Psi}$ has to be the same for all the sources. In this case, the analysis operator $\boldsymbol{\Psi}$ is most certainly not guaranteed to be the same for the n sources.

4.1.3 Implementation Details

The MDCT was implemented with a sine bell analysis window [28] of length 64ms ($l_p =$ 512 samples). This was empirically found [42, 43] to be the best window length. When performing the DWTs, we used the coarsest decomposition scale L permissible and L is typically much smaller than J the dyadic length of the signal. L is also the smallest integer such that $L > \log_2(M) - 1$, where M is the length of the QMF. Most of the transforms were implemented using the functions provided in WaveLab [17].

Performance Number	1	2	3	4
Performance Abbrev.	SDR	SIR	SAR	SNR

Table 4.3: Correspondence between Performance Number and the performance index used

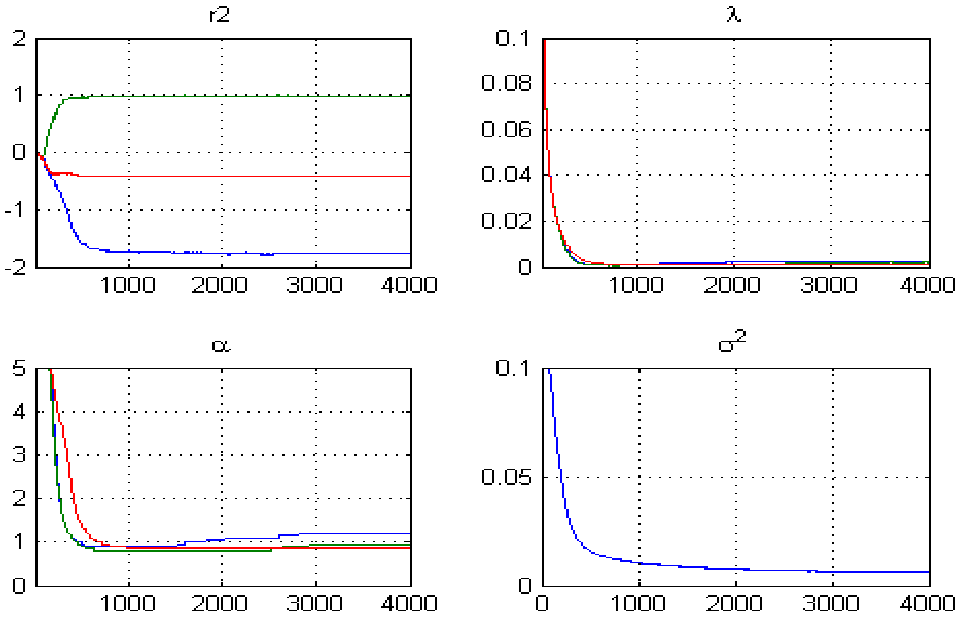

Figure 4.1: Evolution of $\mathbf{r}_2$, $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ and σ^2

4.1.4 Evolution of parameters

A typical plot of the evolution of the parameters is given in Figure 4.1. In view of the BSS indeterminacies as described in section 2.1.1, the first row of the mixing matrix $\mathbf{r}_1$ was clamped to $\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$. Then, $\mathbf{r}_2$ denotes the corresponding second row of the mixing matrix. The true values of $\mathbf{r}_2$ are the tangent of the principal directions ϑ_1, ϑ_2 and ϑ_3. Thus,

$$\mathbf{r}_2 = \begin{pmatrix} \tan(-\pi/3) & \tan(-\pi/8) & \tan(\pi/4) \end{pmatrix} = \begin{pmatrix} -\sqrt{3} & 1-\sqrt{2} & 1 \end{pmatrix}. \tag{4.2}$$

4.1.5 Basis Invariance

As mentioned in [42], the performance criteria are invariant to a change of basis, so that the SDR, SIR, SAR and SNR can be computed based on the time sequences ($\hat{\mathbf{S}}$ compared to $\mathbf{S}$) or transform coefficients ($\hat{\tilde{\mathbf{S}}}$ compared to $\tilde{\mathbf{S}}$). Finally, the reader is encouraged to visit the author's BSS homepage at `http://www2.eng.cam.ac.uk/~yfvt2/bss_demo.html` to listen to all the sound samples, including the original sources, mixtures and reconstructed signals.

4.2 Speech Signals

As mentioned in Chapter 1, the chief motivation behind BSS and ICA is to understand the 'cocktail party problem' and the separation of speech signals from their mixtures. Thus it

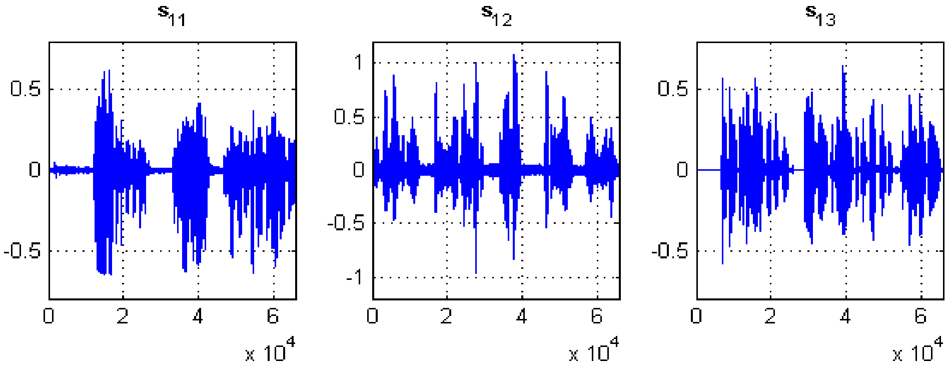

Figure 4.2: Speech Signals $\mathbf{s}_{11}, \mathbf{s}_{12}$ and $\mathbf{s}_{13}$

seems appropriate to start with the analysis of speech signals here. We will first examine the capability of each transform to compress the sources. We seek the best possible (sparsest) decomposition. Then, we will examine the results before a brief discussion is presented. The three speech signals, as shown in Figure 4.2, are each of length $N = 65536$ and they were normalized such that $\sigma_s = 0.1$, so that each source has the same standard deviation. Note that the first, second and third speech signals are labelled $\mathbf{s}_{11}$, $\mathbf{s}_{12}$ and $\mathbf{s}_{13}$ respectively. Notice that there are segments in which the signals are silent. For instance in $\mathbf{s}_{11}$, the first trace of an audible sound is after ≈ 1 second (8000 samples).

4.2.1 Sparsity

The sparsity indices ξ are plotted against the transforms for each of the signals $\mathbf{s}_{11}, \mathbf{s}_{12}$ and $\mathbf{s}_{13}$ in Figure 4.3a. Clearly, the MDCT and the WPBB provide the sparsest representation of the three speech sources. We see that they are also very close to optimal. This can be observed by comparing with transform 7. The DCT performs poorly in this case. This shows that windowing and smoothing are very important, especially since there are significant transient effects [30] in the speech signals. This explains why wavelets [30] are able to provide a sparse representation of speech signals. Wavelets are able to model the transients in an audio signal. For a more detailed analysis of wavelet analysis of transients, the reader is encouraged to refer to [29, 30, 87]. We also perfom an analysis to quantify the amount of tonals and transients in an audio signal in section 4.6.3.

4.2.2 Results

The complete set of numerical results are shown in Figure 4.4. As expected, the MDCT provided the best separation quality. It gave the highest mean values for the SDR, SIR and SAR. The WPBB outperformed the MDCT in terms of noise reduction. Very interestingly, even if we were to apply the Gibbs Sampler directly to the sources without any transform, the separation quality would be better than the DCT. The performances of the

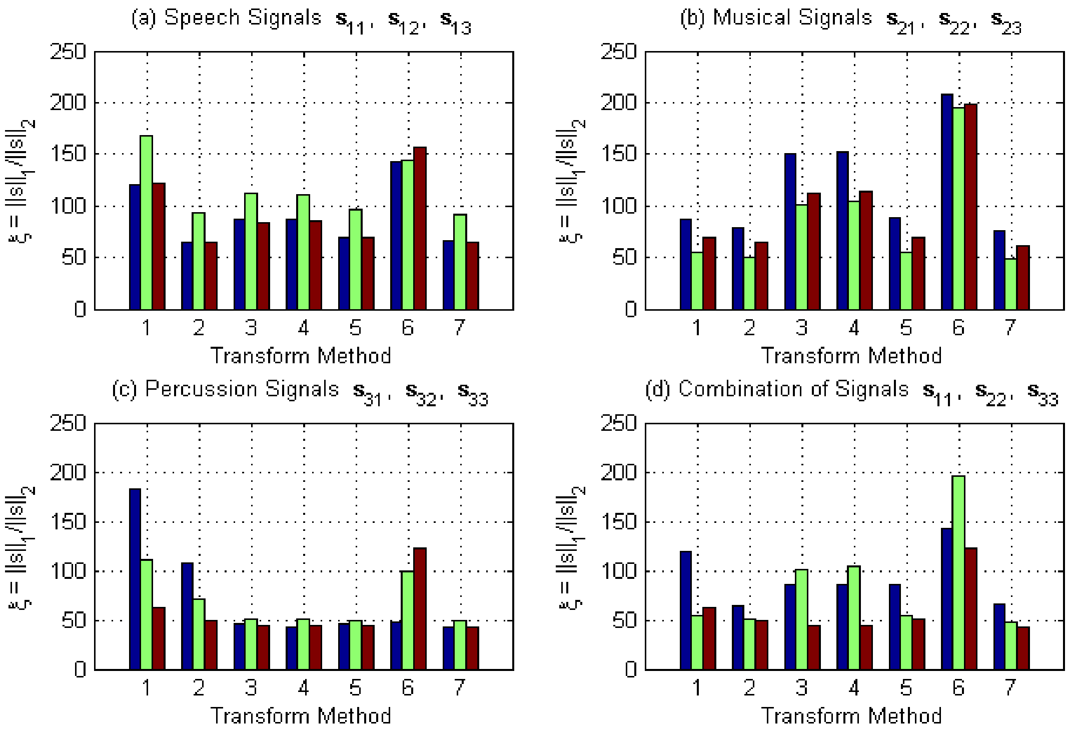

Figure 4.3: Sparsity Indices for Signals using Orthonormal Bases a) Speech Signals, b) Musical Signals, c) Percussion Signals, d) Combination of Signals; Transform method: 1-DCT, 2-MDCT, 3-WT-Vai, 4-WT-Sym, 5-WPBB, 6-NT, 7-WPBB on sources (optimal)

Rank	1	2	3	4	5	6
Transform	MDCT	WPBB	WT-Vai	WT-Sym	NT	DCT

Table 4.4: Ranking of the Orthonormal Bases for Speech Signals

transforms are ranked in Table 4.4. To assess the sound quality, the reader may listen to the reconstructed speech signals at `http://www2.eng.cam.ac.uk/~yfvt2/Speech.html`.

4.2.3 Discussion

There exists a strong correlation between sparsity (Figure 4.3a) and the separation quality (Figure 4.4). The MDCT performed the best followed closely by the WPBB when applied to one of the observations. The DCT performed the worst. Hence, to solve the 'cocktail party problem', one should decompose the observations on a MDCT basis.

4.3 Musical Signals

The three musical signals are plotted in Figure 4.5. The first, second, and third musical signals are strap guitar ($\mathbf{s}_{21}$), piano ($\mathbf{s}_{22}$) and guitar ($\mathbf{s}_{23}$) signals respectively. Unlike

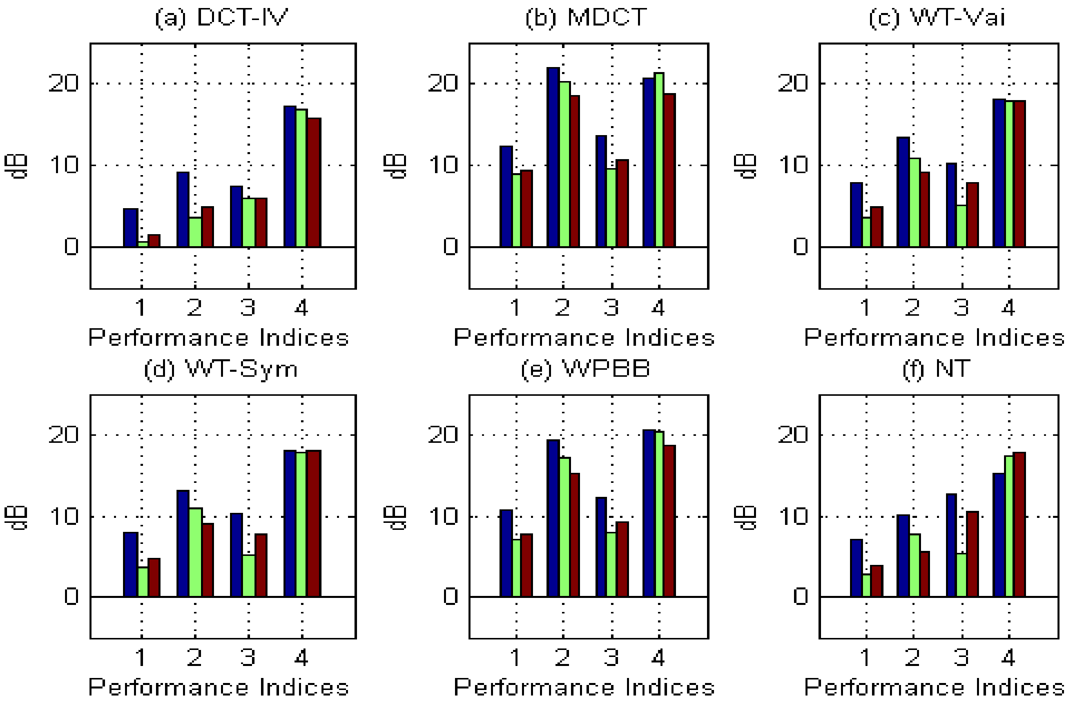

Figure 4.4: Performances of the various transforms on Speech Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{12}$, brown-$\mathbf{s}_{13}$

speech, the musical signals exhibit greater tonals than transients. As a result, we expect that the MDCT would perform very well here. This is substantiated in section 4.6.3.

4.3.1 Sparsity

The sparsity indices for each transform are plotted in Figure 4.3b. In this case, the DCT, MDCT and WPBB provided very sparse representations. They were close to optimal. Thus, the MDCT and DCT can model tonals accurately and wavelets are not as effective here. However, if one chooses to use a wavelet packet best basis algorithm to find a sub-optimal basis based on $\mathbf{x}_1$, the results would be just as good as the MDCT.

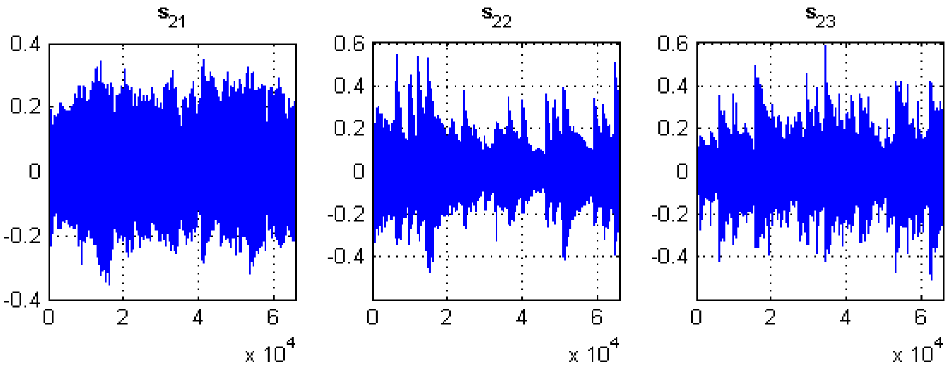

Figure 4.5: Musical Signals $\mathbf{s}_{21}, \mathbf{s}_{22}$ and $\mathbf{s}_{23}$

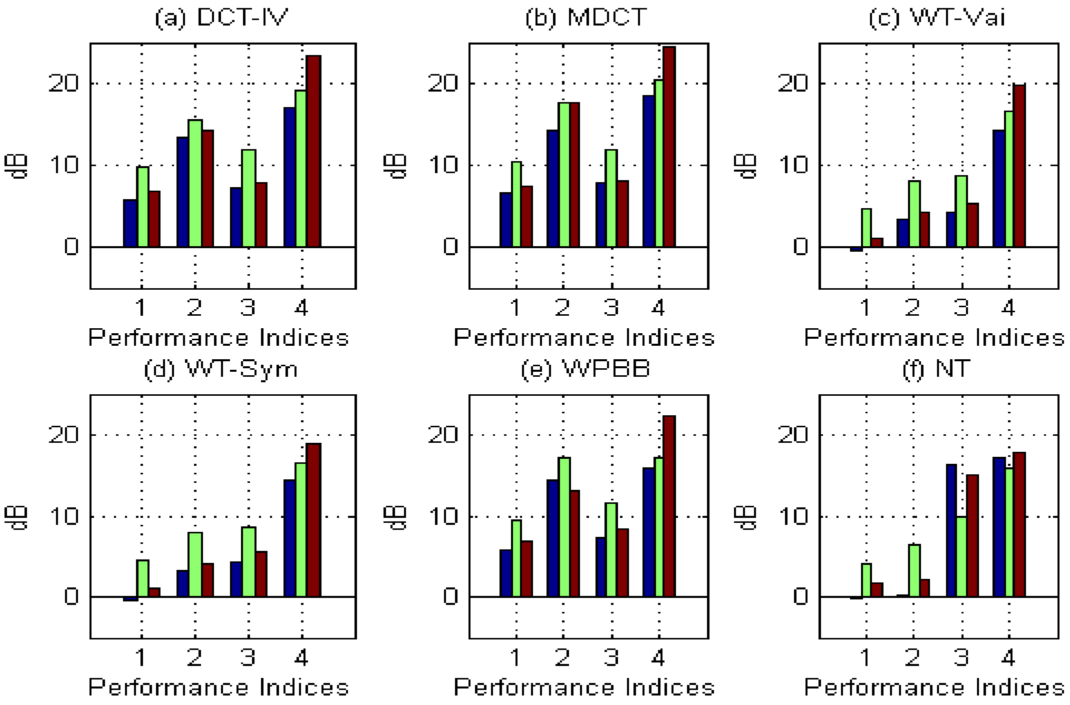

Figure 4.6: Performances of the various transforms on Musical Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{21}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{23}$

4.3.2 Results

The appalling performance resulting from the use of wavelets is demonstrated in Figure 4.6. Using either the WT-Vai or the WT-Sym to decompose the observations resulted in negative SDRs for some reconstructed sources. The reconstructed sources were unnatural. This was the case even after the Gibbs Sampler had converged correctly, emphasizing the fact that wavelets perform poorly for musical signals. The estimated second row of the $\mathbf{A}$ matrix for the WT-Vai transform was $\hat{\mathbf{r}}_2 = \begin{pmatrix} -1.7763 & -0.3800 & 1.0782 \end{pmatrix}$ and the maximum deviation of the elements[1] is within $d_{max} \approx 9\%$ of the true value in equation (4.2). The MDCT outperformed the other transforms in all the performance criteria.

4.3.3 Discussion

If we have prior knowledge that the original sources are musical signals, we should decompose the observations using the MDCT, DCT or WPBB. This would ensure that the separation quality is high. The performances of the transforms are ranked in Table 4.5 and the interested reader may listen to the reconstructed musical signals at `http://www2.eng.cam.ac.uk/~yfvt2/Musical.html`.

[1] If the elements of $\mathbf{r}_2$ are denoted $r_2^{(i)}$ for $i = 1, 2, 3$ and the elements of $\hat{\mathbf{r}}_2$ are denoted $\hat{r}_2^{(i)}$ this corresponds to $d_{max} = \max_i \left| \frac{r_2^{(i)} - \hat{r}_2^{(i)}}{r_2^{(i)}} \right|$.

Rank	1	2	3	4	5	6
Transform	MDCT	WPBB	DCT	NT	WT-Vai	WT-Sym

Table 4.5: Ranking of the Orthonormal Bases for Musical Signals

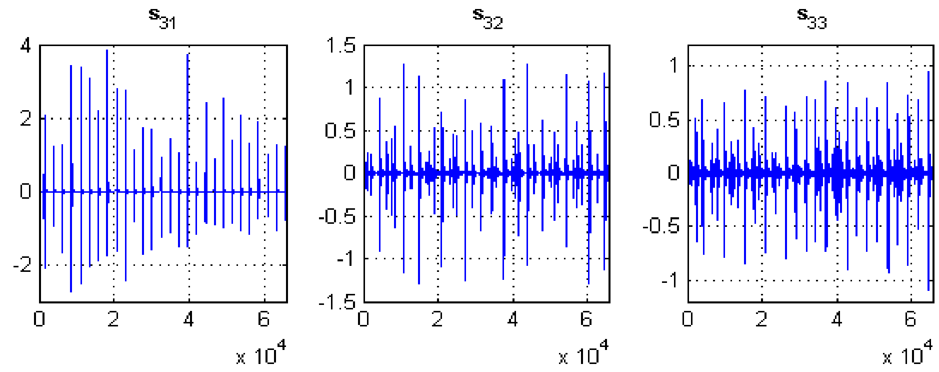

Figure 4.7: Percussion Signals $\mathbf{s}_{31}, \mathbf{s}_{32}$ and $\mathbf{s}_{33}$

4.4 Percussion Signals

We now turn our attention to percussion signals in this section. This analysis is particular useful and relevant because percussion signals are widespread in biomedicine, for example in the form of patients' heartbeats. Here, the $n = 3$ percussion signals are plotted in Figure 4.7. Note that the first, second and third percussion signals are labelled $\mathbf{s}_{31}$, $\mathbf{s}_{32}$ and $\mathbf{s}_{33}$ respectively. It is obvious that these signals contain a lot of transients and few tonals. This will be further quantified in section 4.6.3. Consequently, and as we will see shortly, wavelets would perform well. They would be able to provide a sparse representation of a percussion signal in the DWT domain.

4.4.1 Sparsity

We observe from Figure 4.3c that using either the DWT or the WPBB resulted in the sparsest transform coefficients. The MDCT and the DCT do not model the transients in the percussion signal as well as the DWT and the WPBB.

4.4.2 Results

Not surprisingly, the MDCT was not the best transform for percussion signals as shown in Figure 4.8. The reconstructed signals were not as natural as the original ones and clearly the transient parts are not as well defined. Vaidyanathan wavelets, as discussed in section 2.3.3 are optimized for speech and audio coding and thus result in excellent separation quality. The other wavelets also performed very well, resulting in high SDRs (> 10 dB). From a subjective point of view, the 'beats' and rhythm of the drums can be heard very clearly if either the DWT or the WPBB is used to decompose the sources.

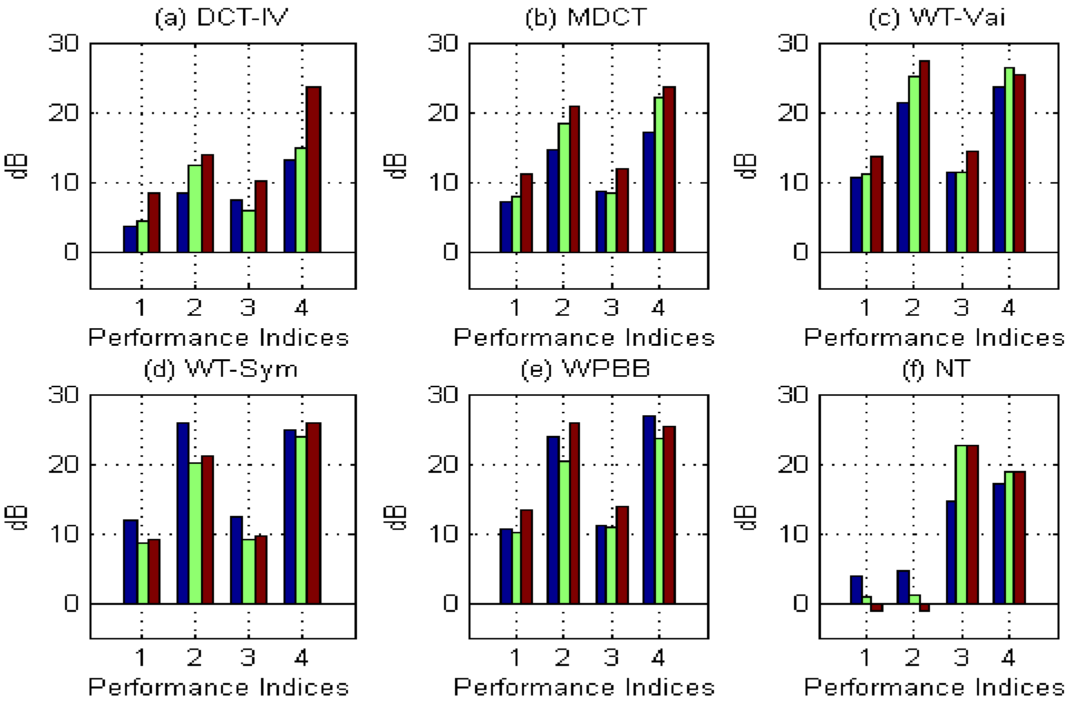

Figure 4.8: Performances of the various transforms on Percussion Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{31}$, green-$\mathbf{s}_{32}$, brown-$\mathbf{s}_{33}$

Rank	1	2	3	4	5	6
Transform	WT-Vai	WPBB	WT-Sym	MDCT	DCT	NT

Table 4.6: Ranking of the Orthonormal Bases for Percussion Signals

4.4.3 Discussion

Clearly, if one has prior knowledge that the underlying signals are percussive in nature, the DWT is the obvious transform to use. The transform produces very good results and its computational complexity is lower than the WPBB because the optimization step in equation (2.28) is not required. The performances of the transforms for the percussion signals are ranked in Table 4.6 and the sound samples can be found at `http://www2.eng.cam.ac.uk/~yfvt2/Per.html`.

4.5 Combination of Signals

To complete our discussion, a combination of the different types of signals was considered. We chose 1 speech signal, 1 musical signal and 1 percussion signal from the collection that has already been mentioned. The three signals used are $\mathbf{s}_{11}$, $\mathbf{s}_{22}$ and $\mathbf{s}_{33}$ and they are shown in Figures 4.2, 4.5 and 4.7 respectively. Hence, there exists a mixture of tonals and transients in the observations.

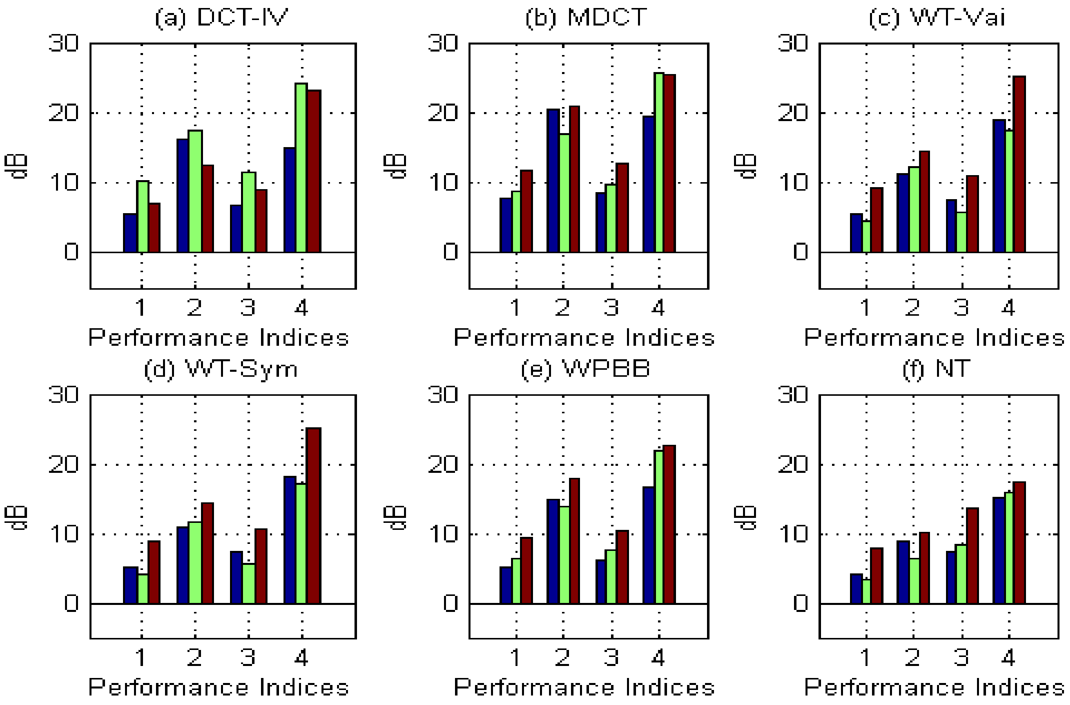

Figure 4.9: Performances of the various transforms on Combination of Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{33}$

4.5.1 Sparsity

The sparsity indices are plotted in Figure 4.3d. The MDCT clearly provided the sparsest representation but the WPBB came close. It was observed that a transform is required to achieve the necessary compression before the Gibbs Sampler was applied.

4.5.2 Results

The results are shown in Figure 4.9. Again, we observe that the MDCT performed the best. It was also the most superior when in terms of noise suppression. In this case, the overall performances of the DCT, the DWT and the WPBB were very similar but DCT outperformed the rest of the transforms for the synthesized musical signal $\mathbf{s}_{22}$.

4.5.3 Discussion

If one seeks to separate a mixture of signals that comprise an approximately equal amount of tonals and transients, the MDCT would be the best. This is demonstrated in the above experiment. The synthesized sound samples, which can be found at `http://www2.eng.cam.ac.uk/~yfvt2/Com.html`, substantiate this point. The performances of the transforms are ranked in Table 4.7.

Rank	1	2	3	4	5	6
Transform	MDCT	DCT	WPBB	WT-Vai	WT-Sym	NT

Table 4.7: Ranking of the Orthonormal Bases for Combination of Signals

4.6 Conclusions

This section concludes our discussion on the performances of the various orthonormal bases on different sets of signals. The preceding experiments emphasize the close relation between the degree of sparsity and the separation quality. This will be further substantiated shortly. The best transforms are, undoubtedly, the MDCT and the WPBB. However, the WPBB algorithm is fairly computationally expensive.

4.6.1 Relation between sparsity and performance metrics

Figure 4.10 shows the close relation between sparsity and performance. The data from the preceding 4 experiments were gathered. For each transform and each set of signals, the mean values of ξ, SDR, SIR, SAR and SNR were calculated. For the ease of presentation, these are given the same symbols ξ, SDR, SIR, SAR and SNR. Figure 4.10a is a plot of the SDR values against the sparsity indices. A linear least-squares fit and a quadratic fit are also shown in red and blue respectively. We observe that there is a negative correlation between the SDR and ξ. The equation of the quadratic is given as

$$\text{SDR} \approx -4.4057 \times 10^{-5}\xi^2 - 0.1061\xi + 14.7013, \tag{4.3}$$

and the correlation coefficient [9] is $\rho_{SDR} = -0.927$. Notice that the quadratic term is very small. We *conjecture* that the relation between the sparsity index ξ and the SDR is linear. Recall that the SDR is the overall separation performance criterion. This has a profound implication - that the sparser a signal, the better the overall separation and the relation is probably *linear*. The other relations are also almost linear.

$$\text{SIR} \approx 7.4670 \times 10^{-4}\xi^2 - 0.3517\xi + 36.6957, \qquad \rho_{SIR} = -0.912, \tag{4.4}$$

$$\text{SAR} \approx 2.1247 \times 10^{-4}\xi^2 - 0.0941\xi + 14.1323, \qquad \rho_{SAR} = -0.848, \tag{4.5}$$

$$\text{SNR} \approx 1.7826 \times 10^{-3}\xi^2 - 0.4139\xi + 40.3201, \qquad \rho_{SNR} = -0.892. \tag{4.6}$$

This can also be observed from Figures 4.10b, c and d. From equation (4.3), we note that the magnitude of the gradient of the least-squares line for the SDR is around 0.1. Hence, if one reduces the sparsity index by 10, the SDR will increase by ≈ 1 dB. In other words,

$$\frac{d(\text{SDR})}{d\xi} \approx -0.1 \text{ dB}. \tag{4.7}$$

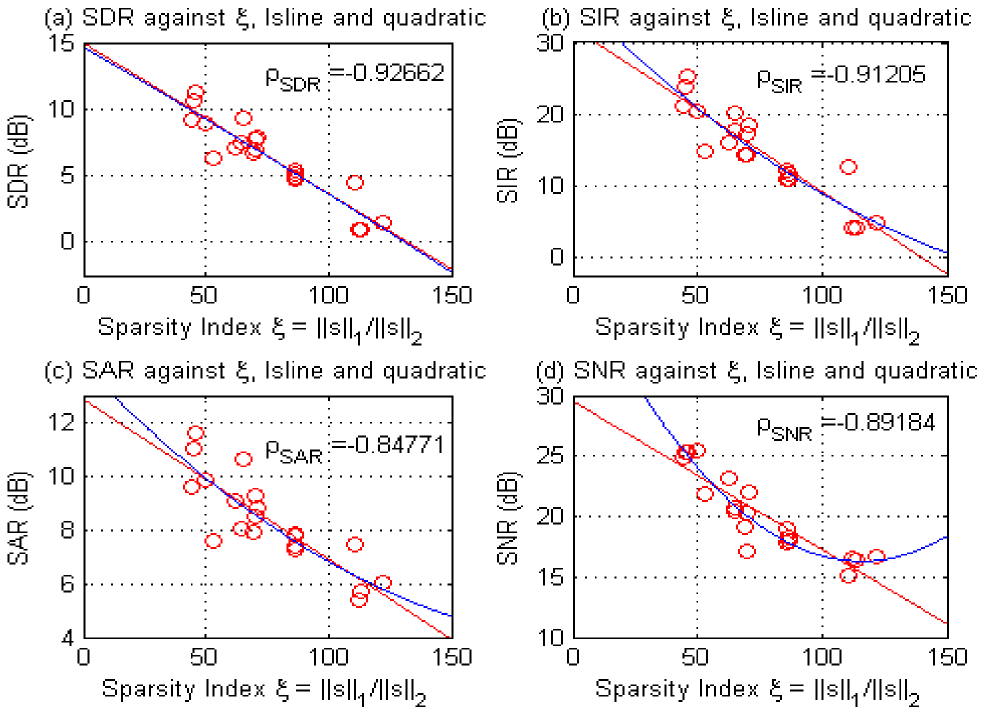

Figure 4.10: Correlation between mean of sparsity indices ξ and mean of the Performance Indices for Orthonormal Bases

In addition, if there is a lower limit on the value of ξ that one can achieve, this directly places a upper limit on the value of SDR. We emphasize that equation (4.7) is only a *conjecture* and further experiments are required to validate it.

Conjecture 1. *If the the audio sources are of the same length and are all normalized to the have same standard deviation*[2], *the* SDR, *in* dB, *and the sparsity index* ξ *are approximately linearly related.*

4.6.2 Performance of Transforms

Across the various sets of signals, it was observed that the MDCT and the WPBB performed the best. This was because, in general, they provided the best i.e. *sparsest* representation of the signals. The MDCT is particularly well suited for audio signals [28]. The MDCT is also used in audio coders for low-bit rates and in spectral estimation. As mentioned previously, wavelets can model the transient nature of speech and percussion signals. The WPBB algorithm gives the best basis based on one of the observations. We have seen that this gives good separation quality even though it was not applied on the sources directly. In fact, it is the most appropriate transform to use if one has no prior knowledge of the nature of the sources.

[2] As a reminder, in these experiments, the length of the signals is $N = 65536$ and the standard deviation is $\sigma_s = 0.1$.

Type	Speech			Musical			Percussion		
Sources	$\mathbf{s}_{11}$	$\mathbf{s}_{12}$	$\mathbf{s}_{13}$	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$	$\mathbf{s}_{31}$	$\mathbf{s}_{32}$	$\mathbf{s}_{33}$
I_{tr}	0.4154	0.4040	0.5615	0.0841	0.0208	0.0816	0.9203	0.7987	0.6607
I_{ton}	0.5846	0.5960	0.4385	0.9159	0.9792	0.9184	0.0797	0.2013	0.3393

Table 4.8: Transient and tonal indices for the audio signals

Type	Speech		Musical		Percussion		Combination	
Pref. Transform	MDCT		MDCT		DWT		MDCT	
Mixtures	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_1$	$\mathbf{x}_2$
I_{tr}	0.4216	0.4237	0.2865	0.2493	0.6632	0.6962	0.3595	0.4197
I_{ton}	0.5784	0.5763	0.7135	0.7507	0.3368	0.3038	0.6405	0.5803

Table 4.9: Transient and tonal indices for the mixtures and the preferred transform

4.6.3 Effect of the Sources

As we have seen, the transientsness of the signals is related to the performance of the transforms. We define the transient and tonal indices I_{tr} and I_{ton} of a signal vector $\mathbf{s}$ as

$$I_{tr} \triangleq \frac{\tau_M}{\tau_M + \tau_W}, \tag{4.8}$$

$$I_{ton} \triangleq \frac{\tau_W}{\tau_M + \tau_W} = 1 - I_{tr}, \tag{4.9}$$

where

$$\tau_M \triangleq \left(\prod_{i=1}^{N} \left| \left\langle \mathbf{s}, \boldsymbol{\psi}_i^{(M)} \right\rangle \right|^2 \right)^{1/N}, \qquad \tau_W \triangleq \left(\prod_{i=1}^{N} \left| \left\langle \mathbf{s}, \boldsymbol{\psi}_i^{(W)} \right\rangle \right|^2 \right)^{1/N} \tag{4.10}$$

are the geometric means of the MDCT and DWT coefficients respectively. $\boldsymbol{\psi}_i^{(M)}$ and $\boldsymbol{\psi}_i^{(W)}$ for $1 \leq i \leq N$, are the length N basis vectors of the MDCT and the DWT respectively. We refer the reader to [8, 29, 87] for more details. I_{tr} and I_{ton} were computed for the nine signals and tabulated in Table 4.8 and we confirm that percussion signals contain the most transients followed by speech signals. Musical signals contain the most tonals.

A plausible improvement to the algorithm is to compute I_{tr} for the mixtures $\mathbf{x}_1$ and $\mathbf{x}_2$, before the application of the transform. This was done and the results are displayed in Table 4.9. In this *pre-processing* step, if the mean value of the transient index for the mixtures, $\bar{I}_{tr}$ exceeds a threshold value, say $\bar{I}_{tr}^{thres} = 0.5$, the DWT (either WT-Vai or WT-Sym) should be used. Otherwise, the MDCT would be the preferred transform. We observe from Table 4.9 that this scheme concurs with our conclusions from the experiments performed, i.e. the DWT should be used for the set of percussion signals and the MDCT for the other set of signals. We emphasize that the chief aim is to achieve a very sparse representation for the sources $\tilde{\mathbf{S}}$ to maximize the various performance measures.

Chapter 5

Results: Overcomplete Dictionaries

In Chapter 3, the theory behind overcomplete dictionaries was presented. In particular, we introduced two overcomplete transforms, namely the Short-Time Discrete Cosine Transform (STDCT) and the Hybrid Transform (HT). We recall that the HT is a concatenation of a MDCT and a DWT (WT-Vai) basis. Both transforms were implemented successfully and the results are presented here. We will briefly compare the use of overcomplete dictionaries and orthonormal transforms. We are going to try to use overcomplete dictionaries to see if we can further improve sparsity of the sources and thus, hopefully, performance.

5.1 Introduction and Initialization

$\mathbf{A}$ and σ were given the *same* values as mentioned in section 4.1. In addition, the different parameters were also initialized with the *same* values as shown in Table 4.1. This was done to ensure that the results can be compared, without bias, to those obtained using the orthonormal transforms in Chapter 4. During the implementation of the STDCT, the parameters as described in Chapter 3, assumed the values in Table 5.1. As a reminder, b is the overlap fraction, l is the window length and N_f is the number of frames.

5.2 Speech Signals

The same set of speech signals, as shown in Figure 4.2, were tested using the two overcomplete transforms. The sparsity index ξ was computed and plotted in Figure 5.1a. As the sparsity index $\xi \triangleq \|\tilde{\mathbf{s}}_i\|_1/\|\tilde{\mathbf{s}}_i\|_2$ involves a normalization in the denominator, we

Parameter	b	l	bl	N_f	$K = l \times N_f$
Numerical Value	1/4	1024	256	85	87040

Table 5.1: Parameters used in the implementation of the STDCT

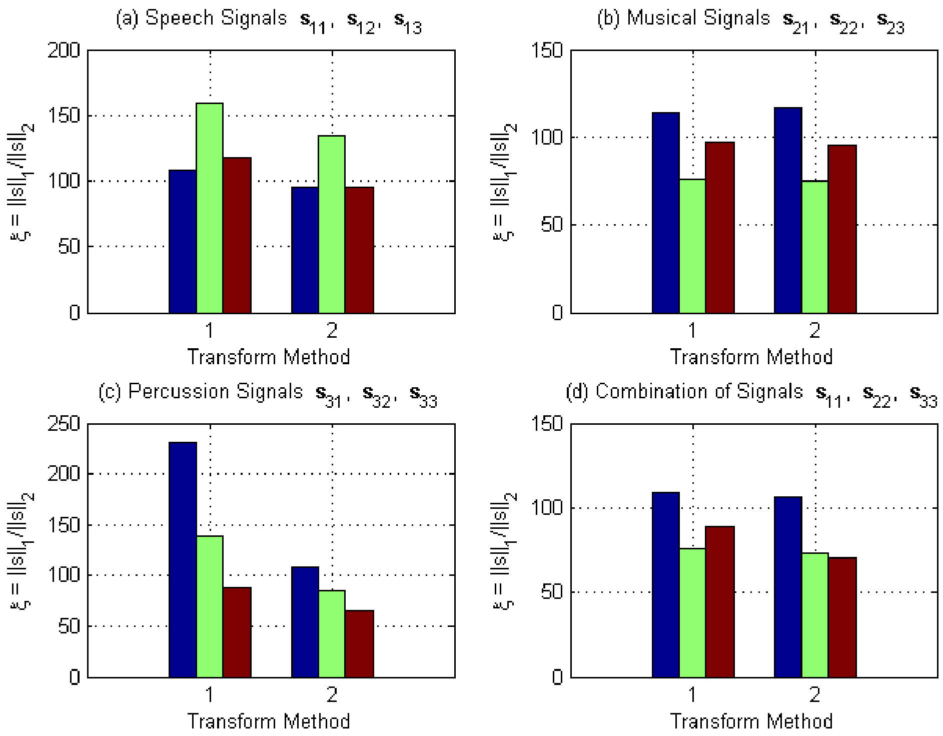

Figure 5.1: Sparsity Indices for Signals using Overcomplete Dictionaries a) Speech Signals, b) Musical Signals, c) Percussion Signals, d) Combination of Signals; Transform method: 1-STDCT, 2-HT

can compare it to the sparsity indices generated using the orthonormal transforms. The HT provided sparser representations of the signals as compared to the STDCT. However, referring to Figure 4.3a, we observe that some of the orthonormal bases performed much better than the overcomplete dictionaries. The results are presented in Figure 5.2. Not surprisingly, the HT outperformed the STDCT here. The separation quality was higher in all the performance criteria. Referring to Figure 4.4, we observe that as the MDCT provided sparser representations of the signals, the performance of the MDCT was marginally better than the HT. Speech signals contain almost equal amounts of transients and tonals (Table 4.8) and hence the Hybrid Transform, which models both tonals and transients, performs well. However, its performance was nonetheless slightly poorer than the MDCT. The computational expense was also doubled. For the overcomplete dictionaries, all the original signals, mixtures and synthesized sources can be found at `http://www2.eng.cam.ac.uk/~yfvt2/Overcom.html`.

5.3 Musical Signals

Similarly, the STDCT and the HT were tested on the musical signals. Since the musical signals contain more tonals than transients, we expect the STDCT to perform at least

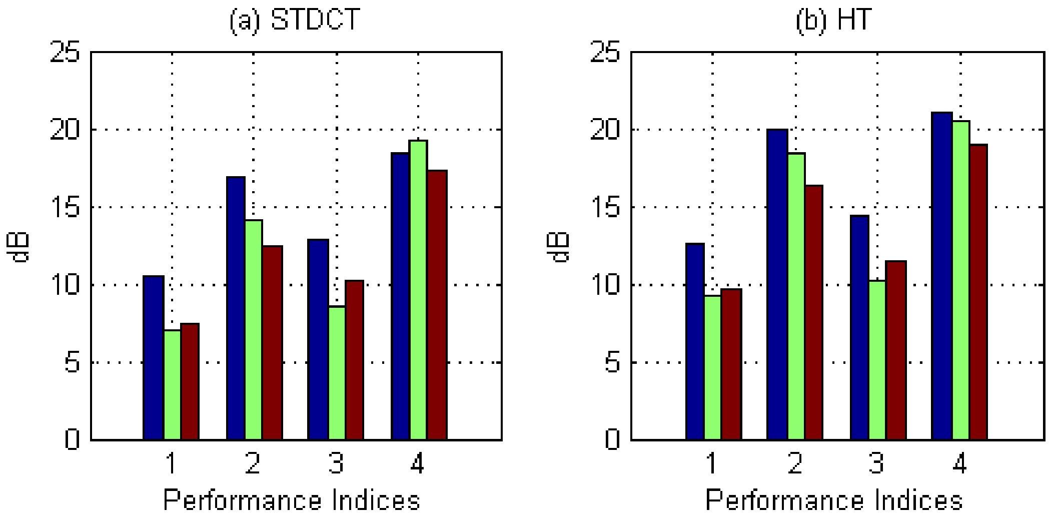

Figure 5.2: Performances of the two Overcomplete Dictionaries on Speech Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{12}$, brown-$\mathbf{s}_{13}$

as well as the HT. Referring to Figure 5.1b, we observe that indeed it does. The results are plotted in Figure 5.3. Both overcomplete representations performed similarly. In this case, the overcomplete dictionaries outperformed the best orthonormal transform, the MDCT. This is promising but not worthwhile given the increased computational expense.

5.4 Percussion Signals

The percussion signals (Figure 4.7) contain more transients than tonals (Table 4.8). This is also reflected in Figure 5.1c where we see that the HT is able to produce much sparser representations of the percussion signals than the STDCT. Figure 5.4 emphasizes, once again, that sparsity is of paramount importance to separation quality. The HT outperformed the STDCT in all the criteria. Its performance was comparable to the Vaidyanathan Discrete Wavelets (WT-Vai) and the Wavelet Packet Best Basis (WPBB) algorithm.

5.5 Combination of Signals

Finally, we turn our attention to the results obtained when the overcomplete dictionaries were used to decompose a combination of signals. Recall that we chose 1 speech signal, 1 musical signal and 1 percussion signal from the collection that was presented. The sparsity results are shown in Figure 5.1d. We see that the HT provided a slightly sparser representation overall. However, the difference in the separation quality, as shown in Figure 5.5, was minimal. Both the overcomplete dictionaries performed almost as well as the MDCT, which was discussed in section 4.5.

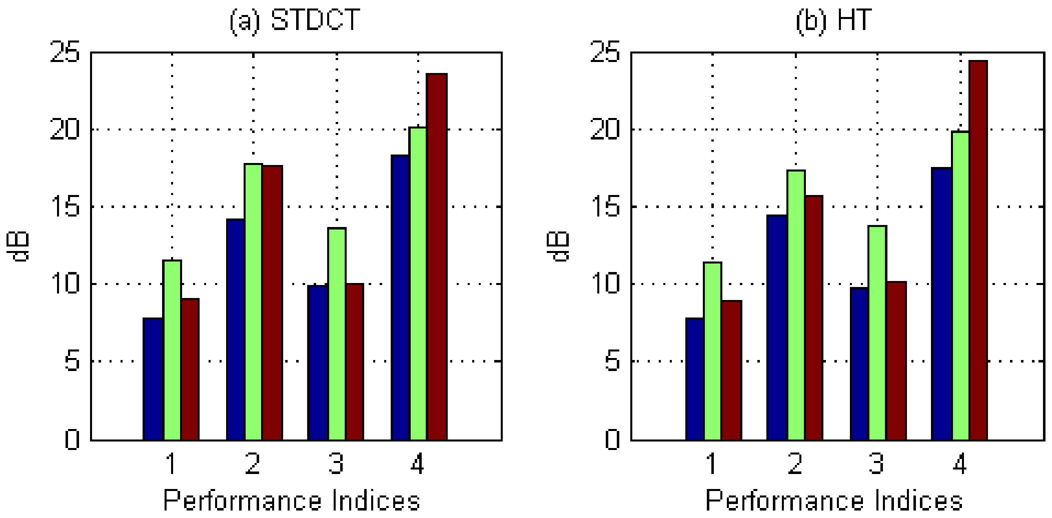

Figure 5.3: Performances of the two Overcomplete Dictionaries on Musical Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{21}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{23}$

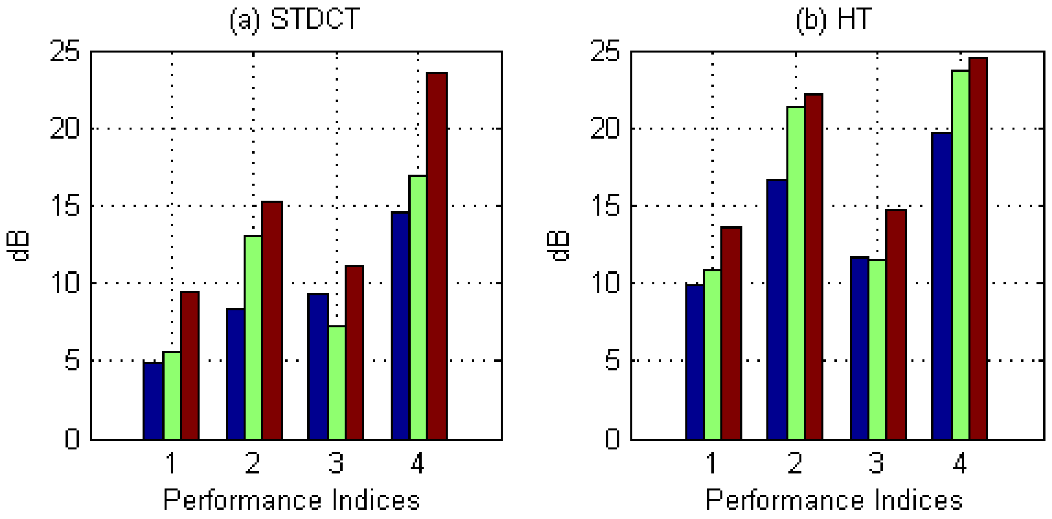

Figure 5.4: Performances of the two Overcomplete Dictionaries on Percussion Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{31}$, green-$\mathbf{s}_{32}$, brown-$\mathbf{s}_{33}$

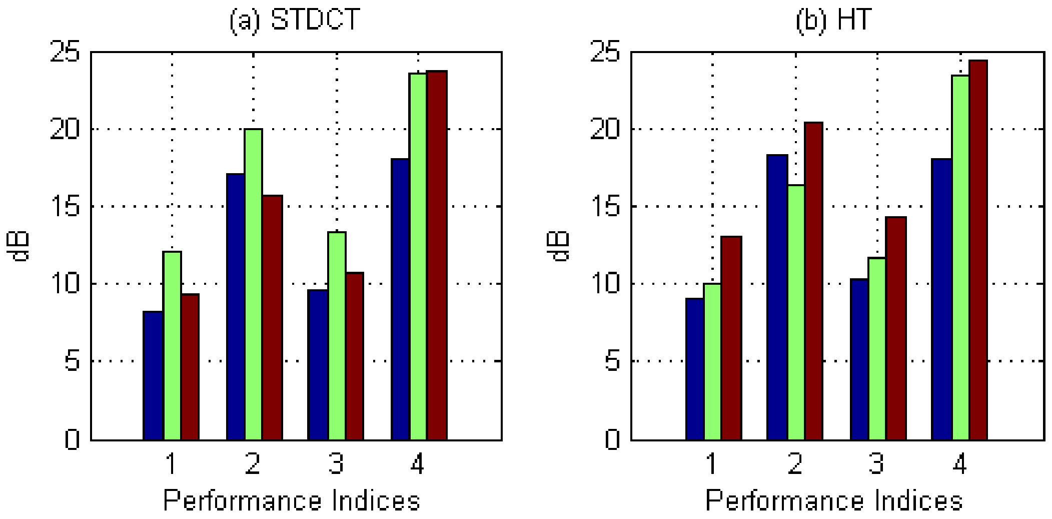

Figure 5.5: Performances of the two Overcomplete Dictionaries on the Combination of Signals; Performance Indices: 1-SDR, 2-SIR, 3-SAR, 4-SNR; Signals: blue-$\mathbf{s}_{11}$, green-$\mathbf{s}_{22}$, brown-$\mathbf{s}_{33}$

5.6 Conclusions

The overcomplete dictionaries provided good separation quality but the increased computational complexity was a major drawback. For instance, the HT took twice as long to run but produced results of similar quality compared to the MDCT.

5.6.1 No Significant Improvement

The overcomplete dictionaries neither improve the sparsity nor the separation quality. The orthonormal bases, in particular the MDCT and the WPBB, performed just as well. There is a also negative relation between the sparsity indices and the separation quality but this is not as straightforward as for the orthonormal transforms (See conjecture 1 on page 38). Referring to Figures 5.1 to 5.5, we conclude that the if the sources are sparsely represented, the separation quality is, in general, better.

5.6.2 Proposed Explanation

We perform an analysis to understand why, contrary to intuition, the overcomplete dictionaries performed relatively poorly. Let $\mathbf{s}$ be a length N signal, which has a sparse representation of length K, $\tilde{\mathbf{s}}_{sparse}$ using an overcomplete dictionary $\mathbf{\Phi} \in \mathbb{R}^{K \times N}$. Hence,

$$\mathbf{s} = \tilde{\mathbf{s}}_{sparse}\mathbf{\Phi}. \tag{5.1}$$

We analyzed the signal using the analysis operator $\mathbf{\Psi}$ to obtain a signal $\tilde{\mathbf{s}}_{ana}$.

$$\tilde{\mathbf{s}}_{ana} = \mathbf{s}\mathbf{\Psi}. \tag{5.2}$$

Using equations (3.3) and (5.1), we obtain

$$\tilde{\mathbf{s}}_{ana} = \tilde{\mathbf{s}}_{sparse}\left(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\right) = \tilde{\mathbf{s}}_{sparse}\widetilde{\mathbf{I}}_K. \tag{5.3}$$

where $\widetilde{\mathbf{I}}_K \triangleq \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ is the projection matrix onto the the column space of $\mathbf{\Phi}$ [107]. Note that $\widetilde{\mathbf{I}}_K \neq \mathbf{I}_K$, hence $\tilde{\mathbf{s}}_{ana} \neq \tilde{\mathbf{s}}_{sparse}$. This means that the signal in the transformed domain $\tilde{\mathbf{s}}_{ana}$ may not be sparse even though the signal $\mathbf{s}$ may have a sparse representation $\tilde{\mathbf{s}}_{sparse}$. Thus, the overcomplete dictionaries performed poorer than expected.

5.6.3 Further Work

To alleviate the above problem, we propose two methods to improve the sparsity and hence the results.

Optimize to find $\mathbf{\Phi}$

One can conduct a study on how to optimize $\widetilde{\mathbf{I}}_K$, or equivalently $\mathbf{\Phi}$, such that it is as close as possible to $\mathbf{I}_K$. In other words, one seeks to find a representation $\mathbf{\Phi}^*$ such that

$$\mathbf{\Phi}^* = \arg\min_{\mathbf{\Phi}} f(\mathbf{\Phi}) \triangleq \|\widetilde{\mathbf{I}}_K - \mathbf{I}_K\|_F, \tag{5.4}$$

where $\|\mathbf{C}\|_F = \sqrt{\sum_i \sum_j c_{ij}} = \sqrt{\text{Tr}(\mathbf{C}\mathbf{C}^H)}$ is the Frobenius[1] norm of the matrix $\mathbf{C}$.

Alternative Gibbs Sampling

Let us consider the following model [115, 116] instead.

$$\mathbf{X} = \mathbf{A}\tilde{\mathbf{S}}\mathbf{\Phi} + \mathbf{N} \tag{5.5}$$

Hence, given $\mathbf{X}$, we want to estimate the mixing matrix $\mathbf{A}$, the sparse sources contained in the matrix $\tilde{\mathbf{S}}$ and the standard deviation of the noise σ *directly.* All the conditional densities in (2.12) to (2.17) stay the same except for $p(\tilde{\mathbf{S}}|\mathbf{A}^{(k-1)}, \sigma^{(k-1)}, \mathbf{V}^{(k-1)}, \tilde{\mathbf{X}})$ which is replaced by $p(\tilde{\mathbf{S}}|\mathbf{A}^{(k-1)}, \sigma^{(k-1)}, \mathbf{V}^{(k-1)}, \mathbf{X})$. This has to be re-derived. We conjecture that this Gibbs Sampling scheme would produce better results because $\tilde{\mathbf{S}}$ is assumed to be sparse given the dictionary $\mathbf{\Phi}$ and we are estimating $\tilde{\mathbf{S}}$ *directly* from the observations $\mathbf{X}$. The analysis step $\tilde{\mathbf{S}} = \mathbf{S}\mathbf{\Psi}$, which does not guarantee sparsity, is avoided. As a consequence, the problem as described in section 5.6.2 may also be avoided.

[1] $\mathbf{C}^H$ is the conjugate transpose or Hermitian of the matrix $\mathbf{C}$.

Chapter 6

Conclusion

In this project, we have investigated the use of different bases and overcomplete dictionaries in *Blind Audio Source Separation*. A Bayesian approach based on Gibbs Sampling [42] was used to estimate the sources in the transformed domain $\tilde{\mathbf{S}}$, the mixing matrix $\mathbf{A}$ and the noise standard deviation σ. *Sparsity* of the sources was a central theme of the whole project and was found to be integral to the source separation algorithm's performance.

6.1 Summary

The chief aim in BSS is to recover the n sources from the m mixtures and we exploited the use of a Markov Chain Monte Carlo (MCMC) method in the form of the Gibbs Sampler. The algorithm is summarized in Figure 2.2. Gibbs Sampling allows us to consider the more difficult underdetermined ($m < n$) and noisy case. Here, various transforms were applied on different sets of signals, before the application of the Gibbs Sampler, and they yielded vastly diverse results. These transforms included orthonormal bases and overcomplete dictionaries. We will now summarize the key ideas developed in the project.

6.1.1 The MDCT is an excellent basis for BSS

It is well known that the MDCT [28] provides sparse representations of audio signals and, in particular, musical signals. The MDCT is used in most current audio coding protocols [12, 14] such as MPEG and Windows Media Player. We used it to decompose the source signals for the purpose of producing sparse representations before the application of the Gibbs Sampler. Chapter 4 substantiates that the MDCT is a good basis for BSS, except when the signals are percussive in nature. The transients [30] are not well modelled by the MDCT. Speech and musical signals, which comprise significant amounts of tonals, are well modelled by the MDCT and it produces very impressive results. The MDCT does not perform so well on percussion signals.

6.1.2 Best Basis Algorithm finds an excellent basis for BSS

An adaptive algorithm based on the observations also performs well. As the sources and the mixing process were unknown, the next best thing was to apply the algorithm on the observations and find a basis $\mathbf{\Phi}_1$ such that the $\mathcal{L}_1$ norm of the observation $\mathbf{x}_1$ in the transformed domain was minimized. This also produced remarkable results as demonstrated in Chapter 4. In fact, it is probably the best transform to use if we do not have prior knowledge of the nature of the sources. It models both tonals and transients well and can deal adequately with signals that are percussive in nature as well as otherwise. Compared to the MDCT, the audio quality appears to be better for the percussion and speech signals. The MDCT is more superior for the musical signals.

6.1.3 Sparsity is very important

An important observation was made from Figure 4.10. From the experiments that were conducted, we *conjecture* that the Source to Distortion Ratio (SDR) in decibels (dB) is a *linear* function of the sparsity index ξ as detailed in equation (4.7) and conjecture 1.

6.1.4 Overcomplete dictionaries provide marginal improvement

It is somewhat surprising that the overcomplete dictionaries did not perform much better than the orthonormal bases. They failed to provide sparse representations for the sources. A plausible explanation is provided in section 5.6.2. This is attributed to the nature of the Moore-Penrose pseudo-inverse [94]. Further research can be conducted to find ways of producing a dictionary $\mathbf{\Phi}$ such that $\widetilde{\mathbf{I}}_K$ is as close to the identity matrix $\mathbf{I}_K$ as possible.

6.1.5 Computational Complexities

MCMC methods are computer-intensive [47] and can only be done off-line and when the complexity of the problem warrants their use. Despite this, the Gibbs Sampling approach is robust to the initialization of the parameters as compared to EM [31, 42]. The sampling approach is feasible because it is easy to sample from univariate conditional distributions.

6.2 Further Work

To conclude the overview of the audio source separation problem, we will briefly outline some of the most important outstanding issues in blind audio source separation. Research in these issues might enhance the performance of current source separation systems. Finally, some interesting extensions using current technologies are also suggested here.

6.2.1 Alternative Gibbs Sampling

As mentioned in section 5.6.3, we can derive the density $p(\tilde{\mathbf{S}}|\mathbf{A}^{(k-1)}, \sigma^{(k-1)}, \mathbf{V}^{(k-1)}, \mathbf{X})$, implement it and compare the new Gibbs Sampling scheme to the results we have obtained in this project. We expect that this sampling scheme would improve the performance.

6.2.2 Alternative Prior Distributions

We can apply the Bayesian approach using priors for the elements in $\tilde{\mathbf{S}}$. The authors in [86] consider using the Laplacian distribution

$$p(\tilde{s}_{i,k}|\alpha_i) = \frac{1}{2\alpha_i} \exp\left(-\frac{|\tilde{s}_{i,k}|}{\alpha_i}\right), \qquad \alpha_i > 0 \quad \forall i. \tag{6.1}$$

Similar to the Student t distribution, α_i contains information about the width or variance of the distribution. Other authors [31, 89] also considered Gaussian Mixture Models

$$p(\tilde{s}_{i,k}|\boldsymbol{\omega}_i, \boldsymbol{\sigma}_i) = \sum_{j=1}^{c} \omega_{i,j} \mathcal{N}(\tilde{s}_{i,k}|0, \sigma_{i,j}^2), \qquad \sum_{j=1}^{c} \omega_{i,j} = 1. \tag{6.2}$$

The conditional densities $p(\boldsymbol{\theta}_{\mathbf{y}}|\boldsymbol{\theta}_{-\mathbf{y}}, \tilde{\mathbf{X}})$ will have to be derived and it will be interesting to observe whether these priors produce better results than the Student t prior.

6.2.3 Convolutive BSS

We have thus far only considered the linear, instantaneous and noisy model. In the convolutive BSS task [15, 16, 84], we can represent the mixing process as

$$\mathbf{x}_t = \sum_{l=0}^{L-1} \mathbf{A}_l \mathbf{s}_{t-l} + \mathbf{n}_t, \qquad 0 \leq t \leq N-1 \tag{6.3}$$

where $\mathbf{A}_l$ is a sequence of $m \times n$ matrices, which are the impulse responses of the acoustical environment. We may possibly use the Bayesian approach to solve this more realistic problem. Convolutive mixing can be rearranged [13] into an instantaneous mixing model.

6.2.4 BSS with unknown number of sources n

In our study thus far, we have assumed that the number of sources n is known. Certainly, in reality, one cannot make this assumption based on the observations. It would be useful and certainly, intriguing to study the possibility of estimating the number of sources before attempting to separate them. Cichocki uses neural networks [22, 23] but a Bayesian approach may produce equally promising results.

Appendix A

Probability Density Functions

A.1 Standard Distributions

The following probability density functions (pdfs) are commonly used in this report.

1. Univariate Gaussian Distribution $\mathcal{N}(x|\mu, \sigma^2)$:

$$\mathcal{N}(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{A.1}$$

2. Multivariate Gaussian Distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{A.2}$$

3. Gamma Distribution $\mathcal{G}(x|\gamma, \beta)$:

$$\mathcal{G}(x|\gamma, \beta) = \frac{x^{\gamma-1}}{\Gamma(\gamma)\beta^{\gamma}} \exp\left(-\frac{x}{\beta}\right) \mathbb{I}_{[0,+\infty)}(x) \tag{A.3}$$

4. Inverted-Gamma Distribution $\mathcal{IG}(x|\gamma, \beta)$:

$$\mathcal{IG}(x|\gamma, \beta) = \frac{x^{-(\gamma+1)}}{\Gamma(\gamma)\beta^{\gamma}} \exp\left(-\frac{1}{\beta x}\right) \mathbb{I}_{[0,+\infty)}(x) \tag{A.4}$$

A.2 Conditional Densities

The following conditional densities are used in the Gibbs Sampling algorithm as described in section 2.2.1. For the sake of brevity, we refer the interested reader to [42] for the detailed derivations.

1. Sampling $\tilde{\mathbf{S}}$:

$$p(\tilde{\mathbf{S}}|\boldsymbol{\theta}_{-\tilde{\mathbf{S}}}, \tilde{\mathbf{X}}) = \prod_{k=0}^{N-1} \mathcal{N}\left(\tilde{\mathbf{s}}_k|\boldsymbol{\mu}_{\tilde{\mathbf{s}}_k}, \boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_k}\right) \tag{A.5}$$

   with $\boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_k} = \left(\frac{1}{\sigma^2}\mathbf{A}^T\mathbf{A} + \mathrm{diag}(\mathbf{V}_k)^{-1}\right)^{-1}$ and $\boldsymbol{\mu}_{\tilde{\mathbf{s}}_k} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}_{\tilde{\mathbf{s}}_k}\mathbf{A}^T\tilde{\mathbf{X}}_k$.

2. Sampling σ and $\mathbf{A}$.

$$p(\sigma^2|\boldsymbol{\theta}_{-(\mathbf{A},\sigma)}, \tilde{\mathbf{X}}) = \mathcal{IG}(\alpha_\sigma, \beta_\sigma) \tag{A.6}$$

   with $\alpha_\sigma = \frac{(N-n)m}{2}$ and $\frac{2}{\beta_\sigma} = \sum_{i=1}^m \sum_{k=0}^{N-1} \tilde{x}_{i,k}^2 - \left(\sum_{k=0}^{N-1} \tilde{x}_{i,k}\tilde{\mathbf{s}}_k^T\right)\left(\sum_{k=0}^{N-1} \tilde{\mathbf{s}}_k\tilde{\mathbf{s}}_k^T\right)^{-1}$ $\times \left(\sum_{k=0}^{N-1} \tilde{x}_{i,k}\tilde{\mathbf{s}}_k\right)$. Let $\mathbf{r}_i$ be the transposed rows of $\mathbf{A}$ such that $\mathbf{A}^T = [\mathbf{r}_1 \ldots \mathbf{r}_m]$,

$$p(\mathbf{A}|\boldsymbol{\theta}_{-\mathbf{A}}, \tilde{\mathbf{X}}) = \prod_{i=1}^{m} \mathcal{N}\left(\mathbf{r}_i|\boldsymbol{\mu}_{\mathbf{r}_i}, \boldsymbol{\Sigma}_{\mathbf{r}}\right) \tag{A.7}$$

   with $\boldsymbol{\Sigma}_{\mathbf{r}} = \sigma^2\left(\sum_{k=0}^{N-1} \tilde{\mathbf{s}}_k\tilde{\mathbf{s}}_k^T\right)^{-1}$ and $\boldsymbol{\mu}_{\mathbf{r}_i} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}_{\mathbf{r}}\sum_{k=0}^{N-1} \tilde{x}_{i,k}\tilde{\mathbf{s}}_k$. As mentioned in section 4.1.4, $\mathbf{r}_1$ is clamped to $[1, 1, 1]$ to resolve the BSS indeterminacies.

3. Sampling $\mathbf{V}$:

$$p(\mathbf{V}|\boldsymbol{\theta}_{-\mathbf{V}}, \tilde{\mathbf{X}}) = \prod_{k=0}^{N-1}\prod_{i=1}^{n} \mathcal{IG}\left(\nu_{i,k}|\gamma_{\nu_i}, \beta_{\nu_{i,k}}\right) \tag{A.8}$$

   with $\gamma_{\nu_i} = \frac{\alpha_i+1}{2}$ and $\beta_{\nu_{i,k}} = \frac{2}{\tilde{s}_{i,k}^2+\alpha_i\lambda_i^2}$.

4. Sampling $\boldsymbol{\alpha}$:

$$p(\boldsymbol{\alpha}|\boldsymbol{\theta}_{-\boldsymbol{\alpha}}, \tilde{\mathbf{X}}) \propto \prod_{i=1}^{n} \frac{P_i^{-\left(\frac{\alpha_i}{2}+1\right)}}{\Gamma\left(\frac{\alpha_i}{2}\right)^N}\left(\frac{\alpha_i\lambda_i^2}{2}\right)^{\frac{\alpha_i N}{2}} \exp\left(-\frac{\alpha_i\lambda_i^2}{2}R_i\right)p(\alpha_i) \tag{A.9}$$

   with $R_i = \sum_{k=0}^{N-1}\frac{1}{\nu_{i,k}}$ and $P_i = \prod_{k=0}^{N-1}\nu_{i,k}$. It is difficult to sample from $p(\boldsymbol{\alpha}|\boldsymbol{\theta}_{-\boldsymbol{\alpha}}, \tilde{\mathbf{X}})$ [42, 43] but since the precise value of α_i is unlikely to be important provided it is within an appropriate small range, we sample $\boldsymbol{\alpha}$ from a uniform grid of discrete values with probability masses proportional to equation (A.9) with $p(\boldsymbol{\alpha}) \propto 1$ [63].

5. Sampling $\boldsymbol{\lambda}$:

$$p(\lambda_i^2|\boldsymbol{\theta}_{-\boldsymbol{\lambda}}, \tilde{\mathbf{X}}) = \mathcal{G}(\gamma_{\lambda_i}, \beta_{\lambda_i}) \tag{A.10}$$

   with $\gamma_{\lambda_i} = \frac{\alpha_i N}{2}$ and $\beta_{\lambda_i} = \frac{2}{\alpha_i R_i}$.

Bibliography

[1] Aharon, M., Elad, M. and Bruckstein, A. K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. On Image Processing*, 2005.

[2] Amari, S. Natural Gradient works efficiently in learning. *Neural Computation*, 10:251–276, Jan 1998.

[3] Amari S., Cichocki A. and Yang H.H. A New Learning Algorithm for Blind Signal Separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. The MIT press, 1996.

[4] Andrews D.F. and Mallows, C.L. Scale mixtures of normal distributions. *J.R. Statist. Soc.* **B**, B(36):99–102, 1974.

[5] Back, A.D. and Weigend A.S. A first application of Independent Component Analysis to extracting structure from stock returns. *Int. J. on Neural Systems*, 8(4):473–479, 1998.

[6] Bell A.J. and Sejnowski T.J. Edges are the 'independent components' of natural scenes. In *Advances in Neural Information Processing Systems*, volume 9, page 831. The MIT Press, 1997.

[7] Belouchrani, A., Abed-Meraim, K, Cardoso, J.F. and Moulines, E. A blind source separation technique based on second order statistics. *IEEE Trans. Signal Processing*, 45(2):1837–1848, 1997.

[8] Berger, J., Coifman, R. and Goldberg, M. Removing noise from music using local trigonometric bases and wavelet packets. *J. Audio Eng. Soc.*, 42(10):808–818, 1994.

[9] Bertsekas, D.P. and Tsitsiklis, J.N. *Introduction to Probability.* Athena Scientific, 2002.

[10] Bilgin, A, Marcellin, M.W., Gormish, M.J. and Boliek, M.P. An Overview of JPEG2000. *Data Compression Conference 2000*, pages 523–541, 2000.

[11] Bishop, C.M. *Neural Networks for Pattern Recognition.* Oxford University Press, 2nd edition, 1995.

[12] Bosi, M. High-quality multichannel audio coding: Trends and challenges. *J. Audio Eng. Soc*, 48(6), 2000.

[13] Bousbiah-Salah, H. Belouchrani, A. and Abed-Meraim, K. Jacobi-like algorithm for blind separation of convoultive mixtures. *Electronics Letters*, 37(16):1049–1050, Aug 2001.

[14] Brandenburg, K. MP3 and AAC explained. *In Proc. AES 17th Int. Conf. High Quality Audio Coding*, Sep 1999. Florence, Italy.

[15] Buchner, H., Aichner, R., and Kellermann, W. *Blind source separation for convolutive mixtures: A unified treatment.* Kluwer Academic Publishers, Boston,, April 2004.

[16] Buchner, H., Aichner, R and Kellermann, W. A generalization of blind source separation algorithms for convolutive mixtures based on second order statistics. *IEEE Trans. Speech and Audio Processing*, 13(1):120–134, Jan 2005.

[17] Buckheit, J.B. and Donoho, D.L. *WaveLab and Reproducible Research.* Stanford University, Stanford CA 94305, USA.

[18] Cardoso, J.F. Blind signal separation: Statistical principles. *IEEE Proc.*, 9(10):2009–2025, 1998.

[19] Cardoso J.F. and Laheld B. Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, 45:434–444, 1996.

[20] Chen, S.S., Donoho, D.L. and Saunders, M.A. Atomic Decomposition by basis pursuit. *SIAM J. Scientific Computing*, 20(1):33–61, 1998.

[21] Cichocki, A. and Amari, S. *Adaptive Blind Signal and Image Processing.* John Wiley and Sons, 1st edition, 2002.

[22] Cichocki, A. and Kasprzak, W. Local adaptive learning algorithms for blind separation of natural images. *Neural Network World*, 6:515–523, 1996.

[23] Cichocki, A., Karhunen, J., Kasprzak, W. and Vigario R. Neural networks for blind separation with unknown number of sources. *Neurocomputing*, 24:55–93, 1999.

[24] Clarke, R.J. Relation between the Karhunen-Loève and cosine transforms. *IEEE Proc.*, 128(F):359–360, 1981.

[25] Coifman, R.R. and Wickerhauser, M.V. Entropy-based algorithms for best-basis selection. *IEEE Trans. Inf. Theory*, 38:713–718, 1992.

[26] Cooley, J.W. and Tukey, J.W. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[27] Daubechies, I. *Ten Lectures on Wavelets.* Society for Industrial and Applied Math, Philadelphia, PA, 1992.

[28] Daudet, L. and Sandler, M. MDCT Analysis of Sinusoids: exact results and applications to coding artifacts reduction. *IEEE Trans. Speech and Audio Processing*, 12(3):302–312, 2004.

[29] Daudet, L. and Torrésani, B. Hybrid representations for audiophonic signal encoding. *Image and Video Coding Beyond Standards*, 82(11):1595–1617, 2002.

[30] Daudet, L., Molla, S. and Torrésani, B. Transient modeling and encoding using trees of wavelet coefficients. *In Proc. 18th Symposium GRETSI'01 on Signal and Image Processing, Toulouse*, Sep 2001.

[31] Davies, M. and Mitianoudis, N. A simple mixture model for sparse overcomplete ICA. *IEE Proc. on Vision, Image and Signal Processing*, 151(1):35–43, 2004.

[32] Donoho, D.L. and Elad, M. Maximal sparsity representation via l1 minimization. *the Proceedings National Academy of Science*, 100:2197–2202, March 2003.

[33] Donoho, D.L. and Huo, X. Uncertainty Principles and Ideal Atomic Decomposition. *IEEE Trans. Inf. Theory*, 47(4):2845–62, 2001.

[34] Duda, R.O., Hart, P.E. and Stork, D.G. *Pattern Classification.* Wiley-Interscience, 2nd edition, 2000.

[35] Duhamel, P., Mahieux, Y. and Petit, J. A fast algorithm for the implementation of filter banks based on time domain aliasing cancellation. *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, pages 2209–2212, May 1991.

[36] Elad, M. and Bruckstein, A.M. A Generalized Uncertainty Principle and Sparse Representation in Pairs of Bases. *IEEE Trans. Inf. Theory*, 48(9):2558–2567, Sep 2002.

[37] Farina, D., Févotte, C., Doncarli, C. and Merletti, R. Blind Separation of Linear Instantaneous Mixtures of Nonstationary Surface Myoelectric Signals. *IEEE Transactions on Biomedical Engineering*, 51(9), Sep 2004.

[38] Ferreira, A. Image Compression Using Independent Component Analysis. Master's thesis, Technical University of Lisbon, 2002.

[39] Ferreira A. and Figueiredo, M. Class-adapted image compression using independent component analysis. *IEEE International Conference on Image Processing - ICIP 2003*, Sep 2003. Barcelona, Spain.

[40] Févotte, C. *Approche temps-fréquence pour la séparation aveugle de sources non-stationnaire (Time-frequency approach for blind separation of nonstationary sources).* PhD thesis, l'École Centrale de Nantes et de l'Université de Nantes, 2003.

[41] Févotte, C. and Doncarli, C. Two contributions to blind source separation using time-frequency distributions. *IEEE Signal Processing Letters*, 11(3), Mar 2004.

[42] Févotte, C. and Godsill, S.J. A Bayesian Approach for Blind Separation of Sparse Sources. Technical report, Cambridge University Engineering Dept, Jan 2005. Accepted for publication in IEEE Trans. on Acoustics, Speech, and Signal Processing.

[43] Févotte, C. and Godsill, S.J. A Bayesian Approach to Time-Frequency Based Blind Source Separation. *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005. Oct 16-19, New Paltz, NY.

[44] Gastaud, R. and Starck, J.L. Dynamic Range Compression: A new method based on wavelet transform. *Astronomical Data Analysis Software and System Conference*, 2003. Strasbourg.

[45] Gelfand, A.E. and Smith, A.F.M. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85:389–409, 1990.

[46] Geman, S. and Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[47] Godsill, S.J. and Rayner, P.J.W. *Digital Audio Restoration: A Statistical Model Based Approach.* Springer-Verlag, 1998.

[48] Gonzalez, R.C. and Woods, R.E. *Digital Image Processing.* Prentice-Hall, Inc., 2002.

[49] Goyal, V.K. *Beyond Traditional Transform Coding.* PhD thesis, Electronics Research Laboratory, University of California Berkeley, 1999.

[50] Gribonval, R. and Nielsen, M. Sparse representation in unions of bases. Technical Report 1499, IRISA, Rennes, France, Nov 2003.

[51] Gribonval, R., Benoroya, L., Vincent, E. and Févotte, C. Performance Measures in Blind Audio Source Separation. *IEEE Trans. Speech and Audio Processing*, 2003. to appear.

[52] Haar, A. Zur Theorie der orthogonalen Funktionensysteme. *Math. Ann.*, 69:331–371, 1910.

[53] Hammersly, J.M. and Clifford, M.S. Markov fields on finite graphs and lattices. Unpublished, 1970.

[54] Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrica*, 57:97–109, 1970.

[55] Hérault, J., Jutten, C. and Ans, B. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétrique en apprentissage non supervisé. *Gretsi*, 2:1017–1020, Mai 1985. Nice, France.

[56] Hyvärinen, A. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768, 1999.

[57] Hyvärinen, A. and Oja, E. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, Oct 1997.

[58] Hyvärinen, A., Karhunen, J. and Oja, E. *Independent Component Analysis.* Wiley Interscience, 1st edition, 2001.

[59] Jain, A.K. *Fundamentals of Digital Image Processing.* Prentice Hall, 1988.

[60] James, C.J. and Hesse, C.W. Independent component analysis for biomedical signals. *Physiological Measurement*, 26:R15–R19, 2005.

[61] Jayant, N.S. and Noll, P. *Digital Coding of Waveforms.* Prentice Hall, Eaglewood Cliffs, NJ, 1984.

[62] Jaynes, E.T. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[63] Jeffreys, H. *Theory of Probability.* Oxford University Press, 1939.

[64] Jourjine, A., Rickard, S. and Yilmaz, Ö. Blind Separation of Disjoint Orthogonal Signals: Demixing N sources from 2 mixtures. *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (ICASSP2000)*, 5:2985–2988, Jun 2000. Istanbul, Turkey.

[65] Kay, S.M. *Fundamentals of Statistical Signal Processing.* Prentice Hall, 1998.

[66] Kisilev, P., Zibulevsky, P. and Zeevi, Y.Y. A Multiscale Framework for Blind Separation of Linearly Mixed Signals. *Journal of Machine Learning Research*, 4:1339–1364, Dec 2003.

[67] Kiviluoto, K. and Oja, E. Independent component analysis for parallel financial time series. *Proc. ICONIP'98*, 2:895–898, 1998.

[68] Le Borgne, H., Guérin-Dugué, A. and Antoniadis, A. Representation of images for classification with independent features. *Pattern Recogn. Lett.*, 25(2):141–154, 2004.

[69] Lee T.W., Girolami M., Bell A.J. and Sejnowski T.J. A unifying Information-theoretic framework for Independent Component Analysis . *International Journal on Mathematical and Computer Modeling*, 1998.

[70] Lee, T.W., Lewicki, M.S., Girolami, M. and Sejnowksi, T.J. Blind Source Separation of More Sources than Mixtures using Overcomplete Representations. *IEEE Signal Processing Letters*, 6(4), 1999.

[71] Levine, S. *Audio Representations for Data Compression and Compressed Domain Processing.* PhD thesis, Stanford University, 1998.

[72] Lewicki, M.S. and Sejnowski, T.J. Learning Overcomplete Representations. *Neural Computations*, 12:337–365, 2000.

[73] Li, Y., Cichocki, A. and Amari, S. Analysis of sparse representation and blind source separation. *Neural Computation*, 16(6):1193–1234, Jun 2004.

[74] Makeig S., Jung T-P., Bell A.J. and Sejnowski T.J. Independent Component Analysis of analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8, 1995.

[75] Makeig S., Jung T-P., Bell A.J., Ghahremani D. and Sejnowski T.J. Blind Separation of Event-related Brain Responses into Independent Components. *Proc. Natl. Acad. Sci. USA*, 1997.

[76] Makeig S., Westerfield, M., Jung T-P., Enghoff, S., Townsend, J, Courchesne, E. and Sejnowski, T.J. Dynamic brain sources of visual evoked responses. *Science*, 295:690–694, 2002.

[77] Malioutov, D.M. A sparse signal reconstruction perspective for source localization with sensor arrays. Master's thesis, MIT, EECS, July 2003.

[78] Malioutov, D.M., Çetin, M. and Willsky, A.S. Optimal Sparse Representations in General Overcomplete Bases. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2:793–796, 2004.

[79] Mallat, S. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, CA, 1999.

[80] Malvar, H.S. and Staelin, D.H. The LOT: Transform Coding Without Blocking Artifacts. *IEEE Trans. Acoustics, Speech and Signal Processing*, 37(4):553–559, 1989.

[81] Martucci, S.A. Symmetric Convolution and the Discrete Sine and Cosine Transforms. *IEEE Trans. Signal Processing*, 42(5):1083–1051, May 1994.

[82] Metropolis, N, Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, W. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[83] Mitianoudis, N. *Audio Source Separation using Independent Component Analysis.* PhD thesis, Queen Mary, University of London, April 2004.

[84] Mitianoudis, N. and Davies, M. Audio Source Separation of convolutive mixtures. *IEEE Trans. Speech and Audio Processing*, 151(5):489–497, Sep 2003.

[85] Mitianoudis, N. and Davies, M. Audio Source Separation: Solutions and Problems. *Int. J. Adapt. Control Signal Process.*, 18(3):299–314, Apr 2004.

[86] Mitianoudis, N. and Stathaki, T. Overcomplete source separation using Laplacian Mixture Models. *IEEE Signal Processing Letters*, 12(4), Apr 2005.

[87] Molla, S. and Torrésani, B. Determining the local transientness in audio signals. *IEEE Signal Processing Letters*, 11(7):625–628, July 2004.

[88] Molla, S. and Torrésani, B. An Hybrid Audio Scheme Using Hidden Markov Models Of Waveforms. *Applied and Computational Harmonic Analysis*, 2005.

[89] Moulines, E., Cardoso, J.F. and Gassiat, E. Maximum likelihood for blind signal separation and deconvolution of noisy signals using mixture models. *ICASSP-97*, April 1997. Munich, Germany.

[90] Newland, D.E. *Random Vibrations, Spectral and Wavelet Analysis.* John Wiley and Sons Inc., 3rd edition, 1993.

[91] Ó Ruanaidh, J.J.K. and Fitzgerald, W.J. Interpolation of missing samples for audio restoration. *Electronic Letters*, 30(8):622–623, 1994.

[92] Oppenheim, A.V., Schafer, R.W. and Buck, J.R. *Discrete-Time Signal Processing.* Prentice-Hall, Inc., 1999.

[93] Oppenheim, A.V., Willsky, A.S. and Nawab, S.H. *Signals and Systems.* Prentice-Hall, Inc., 2nd edition, 1996.

[94] Penrose, R. A Generalized Inverse for Matrices. *Proc. Cambridge Phil. Soc.*, 51:406–413, 1955.

[95] Pham D.T., Garrat P and Jutten C. Separation of a mixture of independent sources through a maximum likelihood approach. *Proc. EUSIPCO*, pages 771–774, 1992.

[96] Princen, J.P. and Bradley, A.B. Analysis/Synthesis filter bank design based on time domain aliasing cancellation. *IEEE Trans. Acous. Speech Signal Processing*, 34(5):1153–1161, 1986.

[97] Proakis, J.G. and Manolakis, D.G. *Digital Signal Processing: Principles, Algorithms and Applications.* Prentice-Hall, Inc., 3rd edition, 1996.

[98] Raine, D., Langley, P., Murray, A., Dunuwille, A and Bourke, J.P. Surface 12-lead electrocardiogram waveform analysis on patients with atrial fibrillation: a tool for evaluating the effects of intervention? *Eur. Heart*, 24, 2003.

[99] Rao, K.R. and Hwang, J.J. *Techniques and Standards for Image, Video and Audio Coding.* Prentice Hall, Upper Saddle River, NJ, 1996.

[100] Rao, K.R. and Yip, P. *Discrete Cosine Transform: Algorithm, Advantages, Applications.* Academic Press, 1990.

[101] Rickard, S. and Dietrich, F. DOA Estimation of many W-disjoint orthogonal sources from two mixtures using DUET. *10th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pages 311–314, Aug 2000. Pocono Manor, PA.

[102] Ristaniemi, T. and Jousesalo, J. On the performance of blind source separation in CDMA downlink. *Proc. Int. Workshp on Independent Component Analysis and Signal Separation (ICA'99)*, pages 437–441, 1999.

[103] Ristaniemi, T., Raju, K. and Karhunen, J. Jammer Mitigation in DS-CDMA Using Independent Component Analysis. *In Proc. of the IEEE Int. Conf. on Communications*, Apr 2002. New York, USA.

[104] Saito, N. The Generalized Spike Process, Sparsity and Statistical Independence. *Modern Signal Processing*, 46, 2004.

[105] Serra, X. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition.* PhD thesis, Stanford University, 1989.

[106] Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J*, 27:379–423, July 1948.

[107] Strang, G. *Introduction to Linear Algebra.* Wellesley Cambridge Press, 3rd edition, 2003.

[108] Strang, G. and Nguyen, T. *Wavelets and Filter Banks.* Wellesley College, 1996.

[109] Torkkola K. Blind separation of delayed sources based on information maximisation. *Proc. ICASSP*, 1996.

[110] Torkkola, K. Blind separation for audio signals – are we there yet. *Proc. Workshop on Independent Component Analysis and Blind Source Separation*, Jan 1999. Aussois, France.

[111] Vaidyanathan, P.P. and Hoang, P.Q. Lattice Structures for optimal design and robust implementation of two-channel perfect reconstruction filter banks. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(1):81–94, Jan 1988.

[112] Vetterli, M. and Kovačević, J. *Wavelets and Subband Coding.* Prentice Hall, 1st edition, 1995.

[113] Wang, W., Chambers, J.A. and Sanei, S. A Novel Hybrid Approach to the Permutation Problem of Frequency Domain Bind Source Separation. *ICA 2004*, pages 532–539, 2004.

[114] Wickerhauser, M.V. *Adapted Wavelet Analysis from Theory to Software.* AK Peters, Ltd., Wellesley, MA, 1994.

[115] Wolfe, P.J. and Godsill, S.J. Bayesian estimation of time-frequency coefficients for audio signal enhancement. In *Advances in Neural Information Processing Systems.* The MIT Press, 2003. Cambridge, MA.

[116] Wolfe, P.J., Godsill, S.J. and Ng, W.J. Bayesian variable selection and regularisation for time-frequency surface estimation. *J.R. Statist. Soc.* **B**, 2004.

[117] Yau, R. Macroeconomic Forecasting with Independent Component Analysis. Technical Report 741, Econometric Society, Aug 2004.

[118] Yilmaz, Ö. and Rickard, S. Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, Jul 2004.

[119] Zibulevsky, M., Pearlmutter, B.A., Boll, P. and Kisilev, P. Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neural Computation*, 13(4):863–882, 2001.