Musical instrument sound source separation

Author:
Gunawan, David Oon Tao

Publication Date:
2009

DOI:
https://doi.org/10.26190/unsworks/21232

Musical Instrument Sound Source Separation

By

David Oon Tao Gunawan

A thesis submitted for the degree of

Doctor of Philosophy

THE UNIVERSITY OF
NEW SOUTH WALES

SYDNEY · AUSTRALIA

School of Electrical Engineering and Telecommunications

The University of New South Wales

May 2009

Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed:__________________________ Date:__________________________

Copyright Statement

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed:____________________ Date:____________________

Authenticity Statement

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed:____________________ Date:____________________

Abstract

The structured arrangement of sounds in musical pieces, results in the unique creation of complex acoustic mixtures. The analysis of these mixtures, with the objective of estimating the individual sounds which constitute them, is known as musical instrument sound source separation, and has applications in audio coding, audio restoration, music production, music information retrieval and music education.

This thesis principally addresses the issues related to the separation of harmonic musical instrument sound sources in single-channel mixtures. The contributions presented in this work include novel separation methods which exploit the characteristic structure and inherent correlations of pitched sound sources; as well as an exploration of the musical timbre space, for the development of an objective distortion metric to evaluate the perceptual quality of separated sources.

The separation methods presented in this work address the concordant nature of musical mixtures using a model-based paradigm. Model parameters are estimated for each source, beginning with a novel, computationally efficient algorithm for the refinement of frequency estimates of the detected harmonics. Harmonic tracks are formed, and overlapping components are resolved by exploiting spectro-temporal intra-instrument dependencies, integrating the spectral and temporal approaches which are currently employed in a mutually exclusive manner in existing systems. Subsequent to the harmonic magnitude extraction using this method, a unique, closed-loop approach to source synthesis is presented, separating sources by iteratively minimizing the aggregate error of the sources, constraining the minimization to a set of estimated parameters. The proposed methods are evaluated independently, and then are placed within the context of a source separation system, which is evaluated using objective and subjective measures.

The evaluation of music source separation systems is presently limited by the simplicity of objective measures, and the extensive effort required to conduct subjective evaluations. To contribute to the development of perceptually relevant evaluations, three psychoacoustic experiments are also presented, exploring the perceptual sensitivity of timbre for the development of an objective distortion metric for timbre. The experiments investigate spectral envelope sensitivity, spectral envelope morphing and noise sensitivity.

Keywords: source separation, music, sinusoidal estimation, instrument modelling, phase estimation, timbre, spectral envelope

Acknowledgements

I wish to acknowledge and thank my supervisor Dr. D. Sen, for his guidance, support, invaluable insights and sharing his wealth of expertise throughout this research.

I would also like to appreciate the support of Associate Professor David Taubman, Professor Eliathamby Ambikairajah and the rest of the staff and students in the Signal Processing Group at the University of New South Wales.

I am grateful for the financial support of the UNSW Faculty of Engineering and the School of Electrical Engineering and Telecommunications.

To my parents, Samuel and Aileen Gunawan, and my brother Jonathan, thank you for your love, prayers and encouragement. I truly appreciate the manner in which you have challenged me to rise to higher heights, and supported me along the way. Thank you also to my family and friends, for your friendship and support. You've made this journey truly enjoyable.

I am eternally grateful to my Lord Jesus, who has blessed me beyond my imagination. Thank you for the opportunity to explore the wonders of Your creation.

List of publications

Journal Articles

1. Gunawan, D. & Sen, D., “Separation of Harmonic Musical Instrument Notes using Spectro-Temporal Modeling of Harmonic Magnitudes and Multiple Input Spectrogram Inversion,” EURASIP Journal on Audio, Speech, and Music Processing. In Review.
2. Gunawan, D. & Sen, D., “Iterative Phase Estimation for the Synthesis of Single-Channel Separated Sources,” IEEE Signal Processing Letters. In Review.
3. Gunawan, D. & Sen, D., “Spectro-Temporal Modelling of Musical Instrument Harmonic Magnitude Trajectories for Source Separation,” IEEE Signal Processing Letters. In Review.
4. Gunawan, D. & Sen, D., “Spectral envelope sensitivity of musical instrument sounds,” Journal of the Acoustical Society of America, no. 123, vol. 1, pg 500–506, 2008.

Conference Papers

1. Gunawan, D. & Sen, D., “Spectro-Temporal Modelling of Harmonic Magnitude Tracks for Music Source Separation,” International Workshop on Multimedia Signal Processing (MMSP), Rio de Janeiro, 2009. In review.
2. Gunawan, D. & Sen, D., “Music Source Separation Synthesis using Multiple Input Spectrogram Inversion,” International Workshop on Multimedia Signal Processing (MMSP), Rio de Janeiro, 2009. In review.
3. Gunawan, D. & Sen, D., “Sensitivity to musical instrument noise in harmonics plus noise modelling,” International Conference on Music Communication Science, Sydney, December 2007.
4. Gunawan, D. & Sen, D., “Identification of partials in polyphonic mixtures based on temporal envelope similarity,” 123rd Audio Engineering Society (AES) Convention, New York, October 2007.

5. Gunawan, D. & Sen, D., “Musical instrument spectral envelope sensitivity,” Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hawaii, April 2007.

6. Gunawan, D. & Sen, D., “An exploration of the spectral envelope space of musical instruments using envelope morphing permutation strategies,” Association of Research in Otolaryngology (ARO) Convention, Denver, February, 2007.

7. Gunawan, D. & Sen, D., “Spectral envelope sensitivity of musical instruments,” Proc. Australasian International Conference on Speech Science and Technology, Auckland, December 2006.

8. Gunawan, D. & Sen, D., “Sinusoidal frequency estimation based on the time derivative of the STFT phase response,” Proc. International Conference on Information, Communications and Signal Processing, Bangkok, December 2005.

Acronyms and Abbreviations

2AFC	Two-Alternative Forced Choice
AMT	Automatic Music Transcription
AWGN	Additive White Gaussian Noise
BA	Band Attenuation
CBR	Critical Band Rate
DFT	Discrete Fourier Transform
EL	Error Level
ERB	Equivalent Rectangular Bandwidth
F0	Fundamental Frequency
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
ICA	Independent Component Analysis
MFCC	Mel-Frequency Cepstral Coefficients
MISI	Multiple Input Spectrogram Inversion
MSE	Mean Square Error
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NMF	Non-negative Matrix Factorisation
ODG	Objective Difference Grade
PBM	Phase Binary Masking
PDFFT	Phase Derivative Fast Fourier Transform
PEAQ	Perceptual Evaluation of Audio Quality
QIFFT	Quadratically Interpolated Fast Fourier Transform
RLS	Regularised Least Squares
RMSE	Root Mean Square Error
SAR	Source -to-Artifacts Ratio
SD	Spectral Distortion
SDR	Source-to-Distortion Ratio
SEEVOC	Spectral Envelope Estimation VOCode
SIR	Source-to-Interference Ratio

SMR	Signal-to-Mask Ratio
SNR	Signal-to-Noise Ratio
STFT	Short Time Fourier Transform
STFTM	Short Time Fourier Transform Magnitude
WPT	Wavelet Packet Transform
WT	Wavelet Transform

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Scholars have long been intrigued by the enigmatic nature of music. Since ancient times, the mathematical relationships of music signals have been studied, and have been instrumental in fashioning music to be what it is today. The foundations of Western music theory in particular, are accredited to Pythagoras (ca. 570-497 B.C.) [29], who investigated the numerical relationships regarding the consonance of musical intervals and the relationships between pitch and the physics of strings.

In modern times, interest in the relationship between mathematics and music have escalated to new heights particularly with the advent of computer systems. The digitisation of music has led to a rapid change in the manner in which people create, listen and interact with music. With music stored in digital formats, music is now remarkably portable and accessible. People have the flexibility to manage the music they listen to and with the interconnections made possible through the Internet, instantaneous music choice is more extensive than ever.

The music production process has also become largely digital, with many recordings being created using a predominantly digital signal processing chain. Sounds are recorded and converted into the digital domain, where a plethora of hardware and software processors are available for signal modification. Processors are used for adding creative effects such as modulation and distortion, spatialisation effects such as reverb, and removing undesired noises such as clicks and pops. Much of this processing operates on the entire signal and if the signal contains multiple sources, then all the signal components are affected.

For this reason, a number of modern music recordings employ a multi-track approach, where each instrument is recorded to a separate channel, allowing each source to be individually modified, before being artificially summed together with other sources. However, for recordings involving a large number of instruments (such as orchestral recordings), recording each individual instrument to a separate channel is simply not practical, and a single channel usually contains a combination of a number of sources. In these cases, individual instruments cannot be modified without first separating the sources contained within the mixture.

This thesis investigates the source separation of music signals from single-channel mixtures, focussing on the development of a separation system that is applicable to a general class of music signals (i.e. not trained for a specific source). The system is built upon a generalised parametric model that facilitates separation. Separating musical sources is a non-trivial task, primarily due to its highly structured nature. The organisation of sounds in music is highly concordant, both spectrally and temporally, resulting in large statistical dependencies between sources. Employing a deterministic separation strategy based on a sinusoidal model framework, these issues are systematically addressed using refined parameter estimation, exploitation of temporal correlations, and iterative magnitude-constrained phase estimation. The separation system is then evaluated using objective and subjective measures.

The optimisation of separation systems requires an extensive understanding of the perceptual sensitivity to modifications made to the tone quality or *timbre* of a sound. In an ideal separation system, sources separated from a mixture should retain all the attributes belonging to the particular source, thereby retaining the timbrel integrity of each sound. While the reality of an ideal source separation system seems quite distant (particularly for mixtures containing many sources), recent systems are beginning to achieve perceptually acceptable separation for a limited number of sources. This has necessitated the development of a perceptually-relevant objective distortion metric, to evaluate the separation quality of separation systems. Through the results of a series of psychoacoustic experiments, insight into the salient parameters of such a distortion metric are discussed. The experiments investigate perceptual sensitivity to the spectral envelope, spectral envelope morphing strategies, and sensitivity to modifications made to the non-harmonic component of pitched sounds.

1.1 Approaches to Source Separation

As humans, we have an amazing ability to perceptually organise the sounds that we hear. When we listen to a piece of music, the acoustic pressure waves are analysed by our auditory system and we perceive the sensation of sound. Of greater intrigue however, is the ability to segregate the sound into separate components. It is not unusual for a listener to hear a piece of music played with numerous instruments, and within an instant, sift through the sounds to focus their attention on the melody of the song. Amongst the diverse sounds of percussive drum hits, distorted guitars, low-frequency bass and rich vocal harmonies, we as listeners are able to not only identify melodies, but we can then go on to make inferences about the nature of source generating the melody. We are able to identify if it is a singer or an instrument and if it is a singer, we think it to be a trivial exercise to identify the singer's gender. Musically trained listeners can even go so far as to identify the instruments being played simultaneously, and their corresponding pitches within a musical mixture.

This section is an exploration of the various approaches that have been employed for source separation. Much of the research in this area was initially motivated by the work in experimental psychoacoustics, and for this reason, we begin with an overview of some of the psychoacoustically motivated approaches in Section 1.1.1. This is followed by an overview of the information theoretic approaches adopted by unsupervised learning methods in Section 1.1.2, and in Section 1.1.3 we review some of the signal modelling approaches. Finally in Section 1.1.4, we look briefly at the related area of automatic music transcription.

1.1.1 Psychoacoustically Motivated Approaches

This perceptual organisation of sounds has been studied in the psychoacoustics literature under the title *auditory scene analysis* and the work of experimental psychologist Albert S. Bregman [14], has played a particularly influential role. Studies in auditory scene analysis involve conducting experiments that measure the perceptual responses to combinations of simple stimuli such as sine-tones and bursts of white noise. The results of these experiments have resulted in Gestalt-like principles, which describe the salient cues related to the organisation and grouping of elementary components.

The principles particularly relevant to source separation involve the definition of the salient grouping cues for organising simultaneous spectral features. These important cues include:

1. Synchronicity cues -
    (a) Common amplitude and frequency modulation - Components whose amplitudes and frequencies exhibit similar variations are generally perceived as one object. These cues are primarily associated with the nature of the physical production systems affecting all the components. For example, when a violinist plays a note with vibrato, the frequency and amplitude modulations due to the variations in string length, manifest themselves in all of the harmonics.
    (b) Common onsets and offsets - Components that have similar onsets and offsets also tend to be grouped as a perceptual object. These can also be viewed as a specific case of amplitude modulation, where the amplitudes of the components are characterised by a sharp rise (for onsets) or fall (in the case of offsets).
2. Spectral cues -
    (a) Harmonicity - Components that are related in integer relationships of a fundamental frequency are also found to be fused together as a perceptual object.
    (b) Spectral proximity - Components are found to fuse with other components closer in spectral proximity. For example, combining the fundamental frequency with the first harmonic resulted in a stronger perceptual unit than the combination of the first harmonic and the 10th harmonic.
3. Spatial cues - With the human auditory system observing two channels - one

> from each ear, localisation of sources is made possible through the timing disparities between the signals from each ear. Components originating from similar spatial locations are more readily fused as a perceptual object.

Many systems were developed based on these experimental findings of audition. These included systems by Duda et. al. [30], Mellinger [83], Cooke [22], Brown [15], Kashino [60, 59], Ellis [32], and Virtanen [125].

While the performance of these systems were adequate for the limited applications they were designed for, there are concerns regarding the underlying principles of the psychoacoustically motivated approach. The systems adopt the grouping principles as the foundational elements of their computational infrastructure, however their validity as the governing principles of audition has been questioned [108]. The psychoacoustic experiments from which the principles were derived, targeted isolated principles and used simplistic, highly constrained stimuli. This provides information about the low-level groupings of synthetic stimuli such as sine-tones and gated white noise, but the information cannot be extrapolated to make definitive inferences about the nature of complex sounds. The descriptions of these principles also lacks the mathematical rigour required for accurate computational implementation. The principles are described using verbal descriptors such as “parallel” and “similar”, and their ambiguous nature makes it difficult to construct robust formulations.

1.1.2 Unsupervised learning

In stark contrast to the psychoacoustically motivated approaches, unsupervised learning methods make little inference about the nature of the sources *a priori*, opting to learn the characteristics of the sources from the data. Separation is achieved by applying information-theoretic principles to the mixtures, exploiting principles such as statistical independence. Unsupervised learning methods can be divided into 3 different classes of algorithms: independent component analysis (ICA), non-negative matrix factorisation (NMF) and sparse coding.

ICA has been used successfully in several blind source separation applications including biomedical signal processing [77], image processing [12] and speech and audio processing [116, 108, 127]. Using a linear signal model, ICA attempts to separate mixtures by identifying sources that are maximally independent. Mathematically, the objective of ICA is expressed as

$$y = Ax \tag{1.1}$$

where y are the observed mixed signals, A is the mixing matrix and x contains the vectors that are maximally independent. ICA algorithms must estimate both A and x, given the observations y, and they achieve this objective by finding solutions that approach a certain definition of statistical independence such as mutual information minimisation [20], info-max [11] and non-Gaussianity maximisation [54]. In the context of single-channel source separation, ICA cannot be directly applied to time-domain signals, as the number of sources extracted must be equal to the number of mixtures observed. For single-channel mixtures, the application of information-theoretic principles has been applied to the magnitude spectrogram in non-negative matrix factorisation [71, 109] and sparse coding has been used for polyphonic music transcription [3] and source separation [123].

While those pursuing source separation using unsupervised learning methods and psychoacoustically motivated methods have generally pursued common objectives from very different approaches, Smaragdis [108] has proposed a unification of both approaches by drawing upon the Barlowian theory[1] of redundancy reduction as a sensory processing mechanism. In his dissertation, he argues that the foundations of audition are better modelled by redundancy reduction, highlighting through a series of simple experiments how perceptual grouping principles relate to information theory.

1.1.3 Signal Models

Source separation systems that are based on signal models, exploit prior information about the sources to extract them from a mixture. Parametric models are derived for the sources and parameters are estimated either, deterministically with heuristics

---

[1] A good review is presented in [10].

[87, 125, 34], or using a Bayesian framework [25, 39, 128].

The predominant signal model employed in music processing is the sinusoidal model. Initially proposed by McAulay and Quatieri [81] as a parametric representation of speech, sinusoidal modelling soon found applications within music signal processing after its introduction by Serra and Smith [101, 111]. The signal model is particularly suited to the harmonic structure of pitched instruments and this will be discussed in greater detail in Chapter 2.

1.1.4 Automatic Music Transcription

Automatic music transcription (AMT) is closely related to music source separation. It is primarily concerned with music signal analysis, with the objective of finding symbolic representations of music signals, in a format commonly used as a performance instruction for musicians. AMT is much like the music analog of automatic speech recognition, and rather than transcribing speech signals into words and sentences, music signals are transcribed into notes of specified durations. Once the music has been transcribed, the information stored within the transcription is not only beneficial to musicians for performances, but can be also used for reproduction using synthesised sounds. AMT involves the estimation of several parameters, some of which pertain directly to source separation. These include the estimation of fundamental frequencies, and the estimation of note onsets and offsets.

The intent of fundamental frequency (F0) estimation is to adequately approximate the F0s within a musical mixture. Once the F0s have been estimated, they are then quantised to note values on the musical scale. While robust F0 estimation methods exist for monophonic signals[2], the estimation of F0s for polyphonic signals [3] (*multiple F0 estimation*) is still a major topic of investigation. State-of-the-art multiple F0 estimators [115, 64, 66] are capable of estimating up to 6 six simultaneous sources with quickly diminishing performance with increasing polyphony.

The estimation of note onsets and offsets provides valuable information regarding musical note durations. While early systems tried to find signal onsets using the

---

[2] A good review can be found in [65]

[3] Signals consisting of two or more sounds.

amplitude envelope of the entire signal [19], filtering the signal into various bands and integrating the results, was later found to be more robust [98]. Psychoacoustic knowledge was later applied by Klapuri [63] to improve performance.

Recent AMT systems have achieved good performance, particularly with musical mixtures with limited polyphony [65, 25, 115, 78]. However transcribing mixtures involving numerous sources, like that of an orchestra, are still somewhat a fantasy.

1.2 Applications

Single channel source separation of musical mixtures has a vast array of potential applications which are described in the following subsections.

Audio coding

If individual sources can be isolated into separate music streams, then efficient, source-optimised compression algorithms can then be applied to achieve high compression gains. This is evident in speech compression algorithms, which have utilised models of the speech production system to achieve very low bit rates.

Denoising and restoration

A plethora of denoising methods are available for the removal of unwanted noises such as hissing, clicks and pops, in musical mixtures. However most of these methods are applied to the entire signal, often adversely affecting other sources in the vicinity of the unwanted noise. Source separation would facilitate the isolation of unwanted noises producing denoising results with fewer artifacts.

Music production

In the music production process, there are many instances in which single microphones are used to record multiple instruments. In these cases, if modifications to a particular source are required, there is no way to rectify the situation except to re-record the piece. The ability to separate sources would provide a flexibility that would fuel a number of creative applications. Instrumentation would be more readily interchangeable, effects could be applied to specific instruments, automatic music transcription would be greatly simplified and the possibilities for resampling would reach new heights. With separated sources, music remixing would be able to experiment with various spatialisations, taking music from monophonic or stereophonic sound fields, into 3-dimensional sound fields.

Music information retrieval

In recent times there has been a large shift in the manner in which people manage and listen to their music. With vast selections of music available, there has been a need for music management systems that have superior interactivity and automated organisation. Music information retrieval has been a growing area of research for this reason, automating the process of acquiring information about music recordings and samples directly from the music signals. Melodic information has been used for novel search methods such as 'query-by-humming' systems [37], the analysis of structural attributes of music have led to genre classification systems [117], and rhythmic analysis has led to the development of automatic tempo and beat tracking systems [98]. These systems will improve the way in which people browse their music and reduce the manual data entry currently required for the organisation of music databases.

The implications of musically meaningful source separation are diverse in this regard, providing improved reliability for information retrieval systems. Systems that search for instrument-specific information will obviously benefit significantly, and a greater flexibility will be offered to systems searching for more general information.

Music education

The separation of sources from music recordings would provide considerable flexibility in the manner in which music is taught and practised. In recent times, there has been a growing number recordings released in 'split-tracks', primarily for providing musical accompaniment for singers. If the same idea could be applied to all instruments from commercially-released recordings, the ability to isolate different sounds would provide great versatility for music teachers. Students would also benefit by being able to play along with a large selection of recordings, having the option to attenuate certain sounds as necessary.

1.3 Scope and Outline of the Thesis

This thesis is concerned with the separation of sounds in musical mixtures. In the signal processing literature, this is referred to as *source separation.* One of the primary objectives of this research is the separation of the harmonic components of pitched sounds in polyphonic mixtures. It does not address the separation of percussive sounds (e.g. drums, tambourines,...etc.), but interested readers are referred to [40, 35].

The separation system is concerned with the perceptual quality of the separated sounds, with the aim of preserving the timbrel integrity of each sound, with respect the original sounds. To facilitate the exploration of the perceptual sensitivity to changes in musical timbre, experiments are conducted to aid in the development of a suitable objective distortion metric for timbre.

Chapter 2. Music Signal Modelling for Source Separation

This chapter describes some of the theoretical foundations concerning musical source separation. It explores the nature of music signals and the inherent problems associated with the separation of pitched sounds. This is followed by an overview of several representations and models for music signals with respect to source separation.

Chapter 3. Harmonic Signal Modelling for Musical Mixtures and Separation

This chapter presents novel methods for the analysis and separation of harmonic sources in polyphonic mixtures, beginning with a novel algorithm proposed for sinusoidal parameter estimation based on phase derivatives. The spectro-temporal nature of instrument harmonics are then analysed, and models are presented for the estimation of harmonics using linear combinations of adjacent harmonics. These models are then used to identify ambiguous harmonics in mixtures.

Chapter 4. Synthesis of Separated Sources

This chapter explores methods of re-synthesising sounds from their parametric formulation. After estimating the magnitude spectra of each source, an iterative algorithm is proposed for the estimation of the phase of each of the sources. The algorithm adapts spectrogram inversion to be applied concurrently to multiple sources, recursively estimating the phase of each source by minimising the error between the sum of the source estimates and the original mixture.

Chapter 5. Perceptual sensitivity of Timbre: Towards an objective distortion metric

Three psychoacoustic experiments are presented in this chapter, providing insights into the perceptual sensitivity of timbre. The first experiment explores the sensitivity of the spectral envelope by attenuating frequency bands of musical instrument spectra. The second experiment investigates the timbre space using permutations of linear-logarithmic morphing between the spectral envelopes of instruments. The final experiment explores the sensitivity of the noise component of pitched instruments as a function of frequency and bandwidth. The results of each experiment contribute to the understanding of the timbre space with the aim of developing an objective distortion metric for timbre.

Chapter 7. Separation of Harmonic Musical Instrument Notes using Spectro-Temporal Modelling of Harmonic Magnitudes and Multiple Input Spectrogram Inversion

Novel methods for the resolution of overlapping harmonics and source synthesis are presented in this chapter. Drawing on the insights discussed in Chapters 2 to 4, the proposed methods are evaluated individually as well as in the context of a source separation system. The system is described and evaluated using both objective and subjective methods, and this is followed by a discussion regarding the implications of the results.

Chapter 7. Conclusions

In the final chapter, the results of the thesis are summarised, and an outline of the potential areas of future research are presented.

1.4 Major Contributions

The original contributions of this thesis relate to the source separation of musical mixtures. The main contributions are summarised below:

- A computationally efficient algorithm for estimating the frequency of a sinusoid from the Short Time Fourier Transform (STFT) is proposed. Upon obtaining initial coarse estimates from the FFT of a given frame, the Phase Derivitave Fast Fourier Transform (PDFFT) makes further refinement to the frequency estimate using only the time derivative of the phase response. The algorithm is derived and is shown to require only 4 multiplications per peak. Single frequencies in the presence of noise are resolved well, outperforming the commonly used Quadratically Interpolated FFT (QIFFT) method even with zero-padding. The algorithm is then used to separate two sinusoids of close frequency proximity that appear as a single peak in the magnitude spectrum.

- In musical instrument sound source separation, the temporal envelopes of the harmonics are correlated due to the nature of the instruments. A quantitative investigation of the correlation between the temporal envelopes of harmonics is conducted on a large database of instrument samples and intra-instrument weighting functions are developed to model the similarities. A harmonic identification algorithm based on these models is then proposed and evaluated in polyphonic mixtures. The algorithm is shown to successfully discriminate between the harmonics of different sources.

- A novel algorithm is presented that concurrently synthesises multiple sources given the magnitude spectra of the sources. The multiple input spectrogram inversion (MISI) algorithm is evaluated for mixtures of up to 6 instruments, and given accurate estimates of the magnitude spectra, is shown to converge to the true phase spectra of each source. The iterative estimation procedure minimises the error between the sum of the estimated sources and the original mixture, and is shown to be capable of significantly reducing phase errors after a limited number of iterations.

- Using the intra-instrument correlation weighting functions derived in [48], a harmonic magnitude track prediction method is proposed for the resolution of overlapping harmonics. Using linear combinations of adjacent harmonic tracks, harmonics corrupted by interfering harmonics are successfully resolved. The spectro-temporal model is evaluated against existing spectral and temporal methods and is shown to provide superior estimates over a variety of musical instruments.

- The spectral envelope is well known to be a perceptually salient attribute in musical instrument timbre perception. A two-alternative forced choice (2AFC) experiment is presented, to observe perceptual sensitivity to modifications made on trumpet, clarinet and viola sounds. The experiment involves the attenuation of 14 frequency bands for each instrument in order to determine discrimination thresholds as a function of centre frequency and bandwidth. The results indicate that perceptual sensitivity is governed by the first few harmonics and sensitivity does not improve when extending the bandwidth any higher. However, sensitivity is found to decrease if changes are made only to the higher frequencies, continuing to decrease as the distorted bandwidth is widened. The analysis of the results is discussed with respect to two other spectral envelope discrimination studies in the literature as well as what is predicted from a psychoacoustic model.

- An experiment exploring the timbre space through novel morphing strategies is presented. The experiment is conducted using a 2AFC paradigm, using various linear-logarithmic permutations of the spectral envelopes of trumpet and clarinet sounds. Psychometric functions are approximated from the results and compared to Mel-Frequency Cepstral Coefficients (MFCC) and spectral centroid spectrum parametrisation as well as psychoacoustic masking models.

The results highlight the need for timbre space models to incorporate human-auditory related frequency resolution and masking models.

- An experiment investigating the perceptual sensitivity of the noise component of pitched instrument sounds is presented. A 2AFC paradigm is employed to explore the sensitivity to the noise component, by attenuating 7 frequency bands with different centre frequencies and bandwidths. The results show that the maximum sensitivity is around 6-11 kHz for sounds with $F0 = 311.1$ Hz, highlighting that low frequency harmonics are good maskers of low frequency noise. Sensitivity is also shown to vary for different instruments, and the sensitivity to noise is governed by broadband sensitivity.
- The PDFFT, harmonic magnitude track estimation method and the MISI algorithm are placed within a source separation architecture, and are evaluated using a variety of objective measures, as well as subjectively in a MUSHRA experiment. The results of the objective and subjective evaluations revealed that in conjunction with each other, the methods provided substantial improvements over existing approaches, over a wide variety of instruments and polyphonies.

Chapter 2

Music Signal Modelling for Source Separation

The quality of source separation that is achieved by a system, is largely dependent on the knowledge of the sources. Once a source can be modelled and parametrised, locating it in a mixture becomes a less arduous task. There are a number of ways to parametrise musical instruments, but it is necessary to select a model that facilitates source separation. This chapter explores the different representations and parametrisations that are appropriate for music signal source separation.

2.1 On the Nature of Musical Mixtures

In order to select an appropriate representation and model for source separation, it is important to understand the nature of music mixtures. Musical sounds can be broadly classified as being either pitched or non-pitched, and the nature and characteristics of such sounds are explored in Section 2.1.1. With an understanding of the general model of musical sounds, an investigation into the combination of these sounds is then discussed in Section 2.1.2.

2.1.1 Musical sounds

Pitch has been defined as “that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high” [9]. This definition, while vague, highlights that pitch is not a physical attribute, but a perceptual attribute. The fundamental frequency (F0) on the other hand, is a physical attribute defined for signals that exhibit periodicity, and the F0 is defined as the inverse of the period. Pitch is thus a psychophysical function of physical variables, primarily dependent on the F0, but also on intensity and duration. This relationship has been explored extensively through psychoacoustic experimentation [96].

A *pitched musical note* refers to a sound that has a pitch, onset time and finite duration. In Western music, notes are arranged in a 12 tone equal-tempered scale where the fundamental frequency in Hz, of a note m is given by

$$F(m) = 440 \times 2^{m/12} \tag{2.1}$$

where m is an integer in the range $-48 \leq m \leq 39$, for the notes of a standard piano. Each interval in the scale is known as a *semitone*, and a list of the fundamental frequencies corresponding to a musical octave ranging from C4[1] to C5 is given in Table 2.1.

Note	m	F0 (Hz)
C4	-9	261.63
C#4	-8	277.18
D4	-7	293.66
D#4	-6	311.13
E4	-5	329.63
F4	-4	349.23
F#4	-3	369.99
G4	-2	392.00
G#4	-1	415.30
A4	0	440.00
A#4	1	466.16
B4	2	493.88
C5	3	523.25

Table 2.1: Corresponding fundamental frequencies of the notes of the equal temperament scale from C4 to C5

Pitched musical instrument sounds are produced by a variety of physical systems, but their common denominator is that they all produce vibrations which are approx-

---

[1] In Western music, notes are denoted by letters and the number suffix denotes the octave, where C0 is the lowest note discernable to the trained ear.

imately periodic. First discovered by Helmholtz [51], these sounds are generally modelled as a harmonic series of sinusoids and knowledge of this has been instrumental to the development of music analysis and synthesis systems. The magnitude response of a flute playing a B4 note, clearly illustrating the harmonic properties of pitched instruments is illustrated in Figure 2.1.

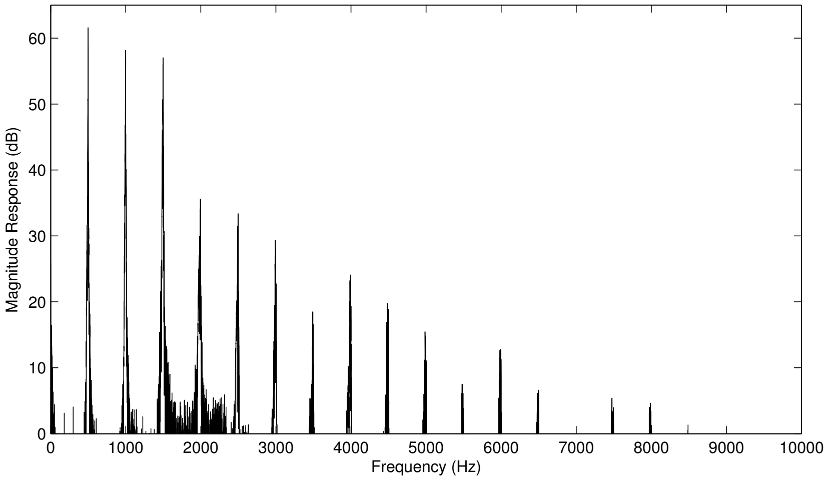

Figure 2.1: Magnitude response of a flute.

The temporal nature of musical notes is also relatively predictable even amongst the vast array of musical instruments that produce them. Since musical sounds are produced by physical systems, they have a time-onset and a time-offset and their duration is finite. In the music signal processing literature, the temporal evolution of the energy of a note is commonly segmented into the *attack*, *sustain* and *release* (Figure 2.2). During the attack, the energy of the signal rises to its maximum level. The energy is then sustained for a duration, after which the note is 'released', which is characterised by a sharp fall in energy.

Non-pitched sounds are another class of musical sounds that exhibit similar temporal structure to pitched notes, but have very different spectral structure. Non-pitched notes, as the name implies, refers to notes whose pitch is absent or ambiguous. The spectral content of such sounds are usually devoid of structure, and are typically produced by percussive instruments such as the drums.

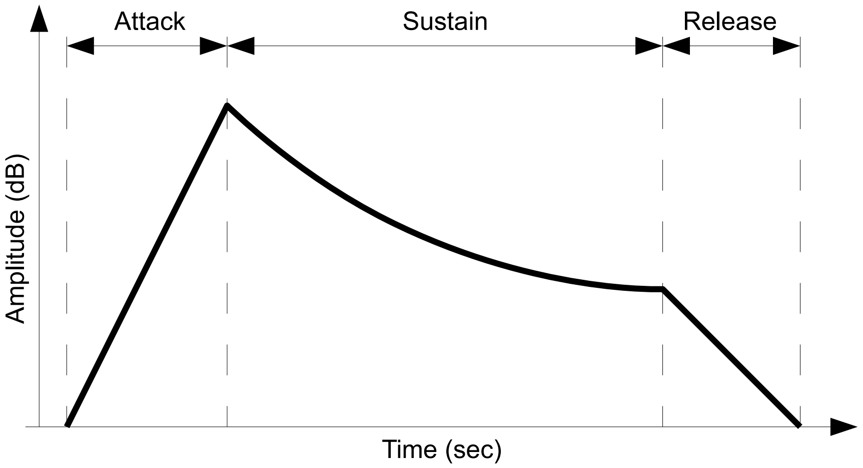

Figure 2.2: Temporal evolution of the energy of a note

2.1.2 Music mixtures

A music mixture will typically consist of a variety of musical instruments playing a series of musical notes organised sequentially and simultaneously. Notes are frequently played simultaneously by one or more instruments, and the combination of 2 or more notes is known as a *chord*. Chords can be consonant or dissonant [91] depending on the relationships between the pitches of the individual notes and the study of this is known as *harmony*.

In Western music where the fundamental frequencies of the notes are given by Equation 2.1, the logarithmic organisation of the F0s results in the F0s being approximately related in small whole number ratios [62]. Major thirds[2] for example, have fundamental frequencies that are related in a 5:4 ratio. This relationship combined with the harmonic nature of pitched instruments, implies that every 5th harmonic of the root note will overlap with every 4th harmonic of the other. Octaves have a 2:1 ratio, while fifths have a 3:2 ratio. Figure 2.3 illustrates the numerous overlaps which occur when 3 notes (C4, E4, G4) are combined. Music mixtures therefore contain many overlapping harmonics - an issue that presents a significant challenge in building music source separation systems. The frequency of overlapping harmonics is not as prominent in speech source separation systems, as the likelihood of multiple

---

[2] Major thirds are a common musical interval comprising of a note and a second note 4 semitones higher.

speakers having related fundamental frequencies is lower. Speech mixtures also tend to have less temporal synchronicity.

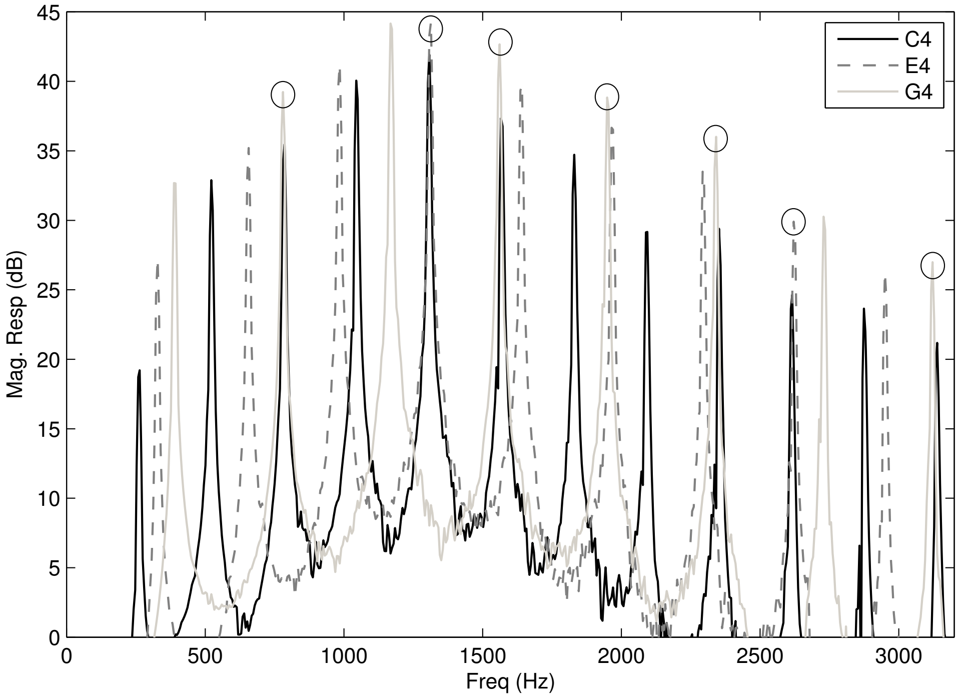

Figure 2.3: Magnitude response of a C major chord trumpet mixture. The regions of harmonic overlap are circled.

In music, the sounds are often intentionally arranged to be rhythmically concordant. The performance of musical pieces typically requires musicians to synchronise their playing with other musicians. This results in a large proportion of simultaneous note onsets and offsets and quite often musical expression such as loudness, vibrato and accents are also synchronised.

There is a high level of statistical dependency between the source sounds that combine to create musical mixtures. The harmonic concordance and temporal synchronicity of the notes, makes music source separation a non trivial task, and the design of a quality music source separation system requires the selection of an appropriate representation and signal model that addresses the issues presented in this section.

2.2 Music Signal Representation

Before a music mixture can be separated, it is necessary to transform the time-domain representation of the mixture into a representation that will facilitate source separation. There are a vast array of representations that are used in audio analysis systems including the Short-Time Fourier Transform [34, 64, 126, 127, 58], multi-resolution transforms such as the constant-Q transform [16] and the discrete wavelet transform [118], and physiologically motivated representations such as the correlogram [106, 30] and the weft [31]. There are many representations employed for audio content analysis, but not all are appropriate for music source separation. So what qualities should a representation for music source separation have?

- Firstly, the representation should be *invertible.* Ultimately, the aim is to re-synthesise the individual sources, so it is necessary to adopt a representation that not only facilitates the analysis of musical sounds, but also the reconstruction of them. An invertible representation should not add additional artifacts to the reconstructed sound, but should ideally offer perfect reconstruction so that for a transform F applied to a signal x, there exists an inverse transform F^{-1} such that

$$F^{-1}\left\{F\left\{x\right\}\right\} = x \tag{2.2}$$

- If the representation is viewed as a decomposition into components from which sources are created, then *component exclusivity* is also important. If the resolution of the components is insufficient to discriminate between the sources, then complete separation will not be attained. Thus the representation should decompose the mixture into components that are sufficiently fine for source reconstruction.

- Finally, it is also desirable that the representation be *linear.* Consider a mixture containing two sounds $x_1[n]$ and $x_2[n]$. If their transformed representations are denoted by $F\left\{x_1[n]\right\}$ and $F\left\{x_2[n]\right\}$, and the representation is linear, then

$$F\left\{\alpha_1 x_1[n] + \alpha_2 x_2[n]\right\} = \alpha_1 F\left\{x_1[n]\right\} + \alpha_2 F\left\{x_2[n]\right\} \tag{2.3}$$

where α_1 and α_2 are scalar constants. Being closed under addition and scalar multiplication implies that iterative subtraction is then made possible without any introduction of artifacts. For example, if $\alpha_1 F\{x_1[n]\}$ is found, then $\alpha_2 F\{x_2[n]\}$ can be obtained without any residual artifacts arising from the transformation itself, where

$$\alpha_2 F\{x_2[n]\} = F\{\alpha_1 x_1[n] + \alpha_2 x_2[n]\} - \alpha_1 F\{x_1[n]\} \tag{2.4}$$

The following sections review the short-time Fourier transform, multi-resolution transforms and physiologically motivated transforms, with respect to the desirable attributes of a music source separation representation.

2.2.1 The Short-Time Fourier Transform (STFT)

Audio signals can be described as 'quasi-stationary', meaning that over sufficiently short periods of time, the statistical properties of the signals change negligibly. The Short-Time Fourier Transform (STFT) [6] provides a time-localised representation of the frequency domain behaviour of a signal, making it particularly suited to the analysis of audio signals. This coupled with computationally efficient implementations of the STFT has resulted in its extensive use in audio processing.

Mathematically, the continuous STFT of a signal $x(t)$ is given by

$$STFT(\tau, \omega) = \int_{-\infty}^{\infty} x(t) h^{'}(t - \tau) e^{-j\omega t} dt \tag{2.5}$$

where $h^{'}(t)$ is the analysis window function, such as the Hanning window.

In the discrete case, where $x[n]$ is a sampled version of $x(t)$, the discrete STFT of a one dimensional signal $x[n]$ is found by segmenting the time axis into frames (see Figure 2.4), which can be overlapping, and applying the DFT to each time frame. The result is a two-dimensional representation of the signal which provides time and frequency information. This is expressed mathematically as

$$X[k, r] = \sum_{n=-\infty}^{\infty} x[n] h[n - rN_{hop}] e^{-j\omega_k n} \tag{2.6}$$

where $h[n]$ is the discrete analysis window function, k is the frequency index, r is the frame index, N_{hop} is the frame hop size in samples and $\omega_k = \frac{2\pi k}{N}$ where N is the size of the discrete Fourier transform (DFT).

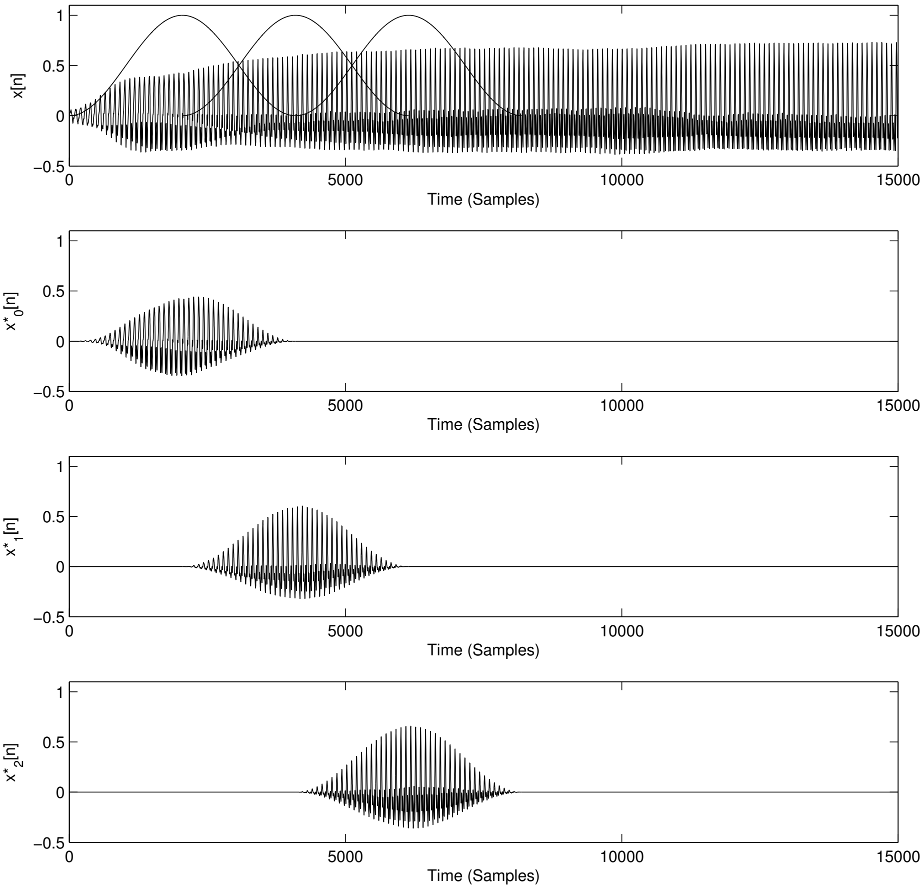

Figure 2.4: Segmentation of $x[n]$ into frames, where $x_r^*[n] = x[n]h[n - rN_{hop}]$

The STFT can be efficiently calculated by using the fast Fourier transform (FFT) to obtain the DFT. Using the FFT to compute a N-point DFT where N is a power of 2, the FFT has complexity of order $Nlog_2N$, which is significantly less than the direct calculation of the DFT which has order N^2.

In addition to its computational efficiency, the STFT is also linear and invertible, and perfect reconstruction is obtained by using the weighted overlap-add method [23]. The inverse STFT is simply found by overlap-adding the inverse DFT of each frame, multiplying each frame with the appropriate synthesis window.

2.2.2 Multi-resolution Transforms

For the analysis of quasi-stationary signals such as music, it is desirable to maximise the time resolution Δt and the frequency resolution Δf . Governed by the uncertainty principle, the product of the time resolution and frequency resolution has a lower bound, where the time bandwidth product [95] is given by

$$\Delta t \Delta f \leq \frac{1}{4\pi} \tag{2.7}$$

Thus, the time and frequency resolution cannot be arbitrarily small and one can only trade time resolution for frequency resolution or vice versa. The time and frequency resolutions are dependent on the analysis window and in the case of the STFT, the time and frequency resolutions are fixed because it has a fixed analysis window. Multi-resolution transforms on the other hand, have time and frequency resolutions that change throughout the representation.

The continuous wavelet transform (WT) is a multi-resolution transform given by

$$WT(\alpha, \beta) = \int x(t)\psi_{\alpha,\beta}(t)dt \tag{2.8}$$

where $x(t)$ is the signal and $\psi_{\alpha,\beta}(t)$ is the basis function given by

$$\psi_{\alpha,\beta}(t) = \frac{1}{\sqrt{\alpha}}\psi\left(\frac{t-\beta}{\alpha}\right) \tag{2.9}$$

While the basis functions of the STFT are derived by varying the frequency of a sinusoid, the basis functions of the WT are time translated and scaled versions of a primary wavelet, which are translated and scaled by α and β respectively. There is also an additional normalisation factor $\frac{1}{\sqrt{\alpha}}$ to ensure that the wavelet has unit energy.

The wavelet packet transform (WPT) is another multi-resolution transform which like the WT, uses basis functions that are time translated and scaled versions of a primary wavelet. However the WPT offers greater flexibility in the way the time-frequency tiling occurs.

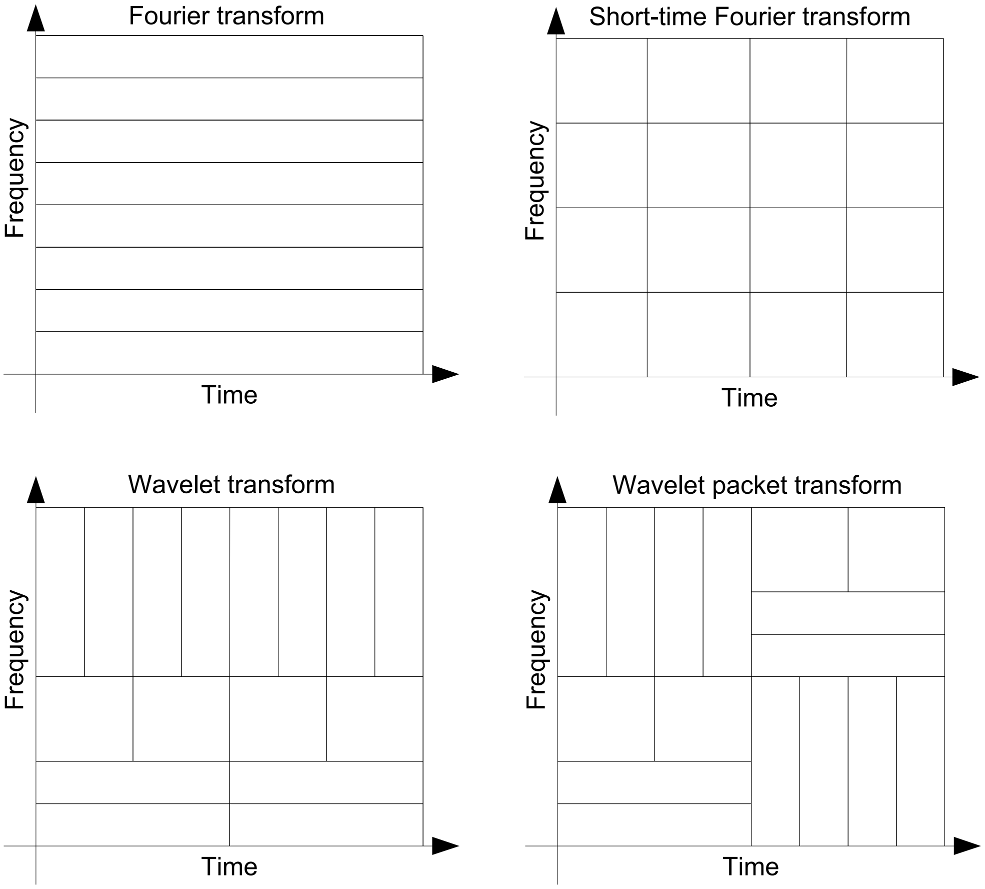

Figure 2.5: Time-frequency tiling of the Fourier transform, the short-time Fourier transform, the wavelet transform and the wavelet packet transform

Figure 2.5 compares the time-frequency tilings of the Fourier transform (FT), the STFT with the WT and WPT which are multi-resolution transforms that have been adopted for applications in audio and image processing. While the STFT improves on the FT by offering time-localised frequency information, the WT and WPT offer flexibility over the STFT by varying the time-frequency tiling. The WT apportions the time and frequency axes to have resolutions that vary logarithmically, so that in the lower frequencies, there is higher frequency resolution but less time resolution, while in the higher frequencies there is higher time resolution but less frequency resolution. The WPT offers even greater flexibility, allowing the time-frequency tiling to be designed according to the signal statistics.

The flexibility of the WT and WPT requires that an investigation be made into the appropriate time-frequency tiling for music source separation. In a typical music mixture, the harmonics of pitched musical instruments are approximately linearly spread across the frequency axis determined by the the fundamental frequencies. While the spectral energy of pitched musical instruments are known to decrease at higher frequencies, the presence of these high frequency harmonics are still evident (see Figure 2.6) and perceptually relevant [50]. Thus, the representation of a music source separation system must not only have good frequency resolution at low frequencies, but also at high frequencies in order to separate harmonics across the entire spectrum. This obviously requires a trade-off in temporal resolution, but there are other methods of analysing the aspects of musical sounds which require high temporal resolution such as note attack transients. This will be discussed Section 2.3.

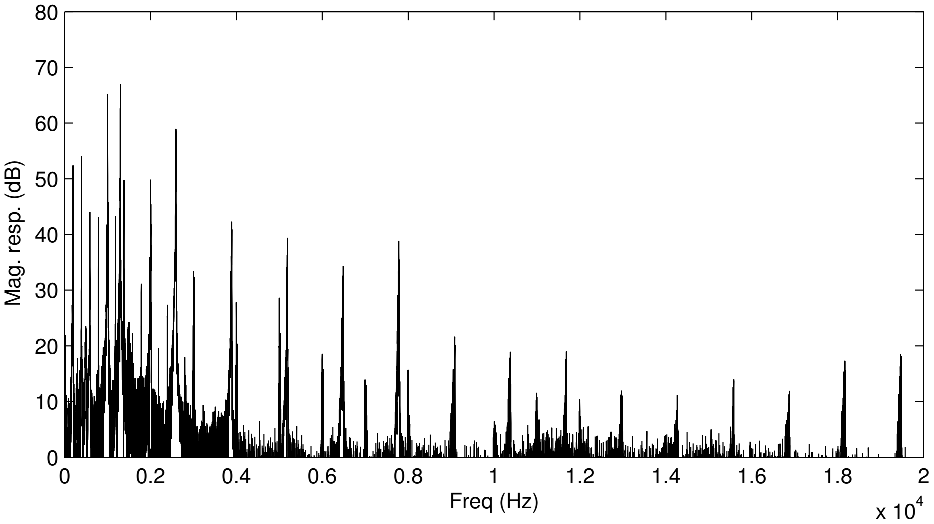

Figure 2.6: Magnitude response of a music mixture containing a flute, clarinet and piano playing the notes B5, E6 and G3 respectively. The mixture illustrates the linear distribution of harmonics and prominent high frequency components well past 10 kHz.

2.2.3 Physiologically Motivated Representations

The human auditory system is one of the most extensive and complex systems of audition and there has been extensive research invested into understanding it and modelling it. Given the capabilities of the human auditory system, it seems logical to explore the representation of sound that it uses, as a potential representation for a music source separation system.

The human auditory system can be divided into the peripheral (comprising of the outer, middle and inner ear) and the central auditory nervous system. The outer ear channels the acoustic pressure waves into the middle ear, which performs an impedance matching to the inner ear. The vibrations from the middle ear are then sent through the oval window where the basilar membrane performs a spectral analysis of the sound. Each point along the basilar membrane resonates at a different characteristic frequency, with the amplitude of the resonance corresponding to the intensity of the frequency. This spectral information is then set to the central auditory nervous system for higher level processing.

Central to the human auditory system's representation of sound is the cochlea. While the precise mechanics of the cochlea are still being investigated, there have been a few models which have been proposed. One such model that has been used extensively in audio analysis, views the primary function of the cochlea as an array of band-pass filters. The development of these filters was initiated by the experimental work of Fletcher [36] who, through a noise-bandwidth experiment, proposed the concept of the critical band - the auditory filter bandwidths. The Bark scale was later introduced [97, 132, 133], named after Heinrich Barkhausen, corresponding to the first 24 critical bands of hearing, where the critical band rate (CBR) is given by the analytical expression

$$CBR = 13arctan\left(0.76\frac{f}{1000}\right) + 3.5arctan\left(\frac{f}{7500}\right)^2 \tag{2.10}$$

and the critical bandwidth Δf is related to the band centre frequency f_c by

$$\Delta f = 25 + 75\left[1 + 1.4\left(\frac{f_c}{1000}\right)^2\right]^{0.69} \tag{2.11}$$

Moore and Glasberg [85] later revised Zwicker's model and proposed the equivalent rectangular bandwidth (ERB) which, while related to the critical bandwidth, better accounted for loudness. At moderate sound levels, the ERB in Hz is defined by [85]

$$ERB = 24.7\,(4.37 f_c + 1) \tag{2.12}$$

A very approximate but efficient implementation of the auditory filterbank as a 4th order gammatone filterbank, was presented by [88, 107] and the magnitude responses of these filters are illustrated in Figure 2.7.

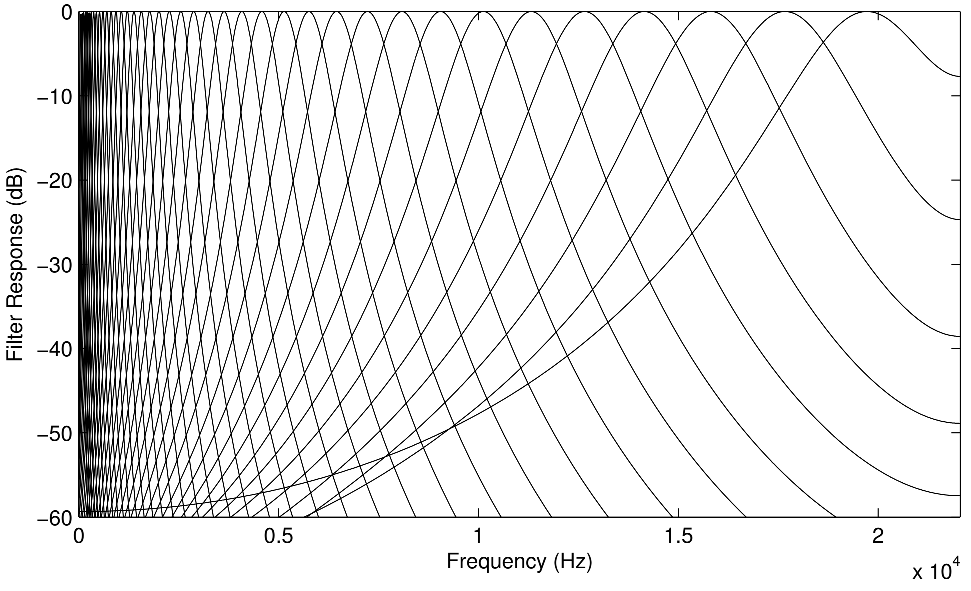

Figure 2.7: Gammatone filterbank

Other models such as Lyon's model [105] offer more accurate transmission-line modelling with automatic-gain control, to account for the large range of intensities that can be handled by the auditory system. However, while accurate modelling of the human auditory system offers benefits in many areas of audio processing such as compression, using such a representation for music separation does not seem to be particularly advantageous. Firstly, the approximately logarithmically spaced centre frequencies of the filters are similar to the WT, and for the same reasons as mentioned in Section 2.2.2, such filters do not offer the frequency resolution required at higher frequencies to ensure component exclusivity. In addition to this, filter-bank models of the cochlear have heavily overlapping band-pass filters, degrading component exclusivity even further. Finally, while the human auditory system is able to efficiently perform auditory scene analysis, there is no indication that it is in fact performing a separation of sound mixtures. Recent studies also seem to infer that due to the limitation of neural resources, mechanisms exist to bias attention towards salient events rather than actually performing source separation [61]. This implies that perceptual segregation of an auditory scene is a weighted representation of our environment, which is also consistent with visual saliency models [28, 68]. Such a model also seems more consistent with the type of representation delivered by the cochlear - one from which objects and events can be identified without truly separating sources and wasting unnecessary neural resources.

2.3 Music Signal Modelling

As there is a high level of statistical dependency between source sounds in music mixtures (refer to Section 2.1.2), signal models of the sources can be used to facilitate separation. If sources can be adequately modelled, the models can aid in the identification of sources in mixtures. Models can also be used to infer information when attributes of a source are obstructed by another source.

The difficulty with modelling musical instrument sounds for source separation is that there are a vast number of instruments, which each produce sounds in different ways. For example, the flutist produces a musical note by blowing a rapid jet of air across the embouchure hole. This in cooperation with the resonances of the air in the flute, produces oscillations, which is radiated as sound. A guitar on the other hand, produces a musical note by an entirely different physical mechanism. The guitar string is displaced by the player's finger and when it is released, the string is set into motion, resulting in a vibration that is radiated as sound. There are numerous instruments whose sounds are generated by different means. In fact most instruments can be excited in a number of different ways which increases the complexity of parametrising the generation of sounds from musical instruments.

The modelling of musical instruments has been extensively researched. Helmholtz's [51] early investigations in music acoustics have laid the foundations for much of the research we know today. His discovery of the harmonic structure of pitched sounds and his work on sound perception were pivotal, and have formed the basis of many of the modern, complex music analysis and synthesis systems. In this section, we discuss two different modelling approaches: the instrument specific approach of physical modelling synthesis (Section 2.3.1), and the more general modelling approach of sinusoidal modelling (Section 2.3.2).

2.3.1 Physical Modelling Synthesis

Physical modelling synthesis [112] is an instrument modelling approach that models individual instruments using digital signal processing formulations of physical models from musical acoustics. The models are instrument specific and the intricacies of each instrument are modelled using digital waveguides, digital filters and memoryless non-linearities. Models have been calculated for a vast number of instruments including slap bass [93], brass instruments [21], string and wind instruments [57] and even traditional Finnish instruments such as the kantele [56]. Models of the speech production mechanism also have been used extensively in speech processing applications, particularly in speech coding where they have played a pivotal role in achieving very low bit rates.

A separation system that could leverage the efficient parametrisations of physical modelling synthesis would provide numerous benefits, such as excellent coding gain and good individual sound reproduction. However, such a system would require the non-trivial intermediate step of determining what instruments are in the mixture. While there are a plethora of techniques available for monophonic musical instrument classification[3] with very high rates of correct identification, polyphonic music instrument classification systems are still in their infancy [79] and have yet to achieve identification scores comparable to polyphonic fundamental frequency estimation systems [66]. With this in mind, we explore a more generic model of parametrising the sounds of musical instruments.

2.3.2 Sinusoidal Modelling

An efficient parametrisation for pitched musical instruments is the sinusoidal model. First applied to speech signals by McAulay and Quatieri for speech coding [81],

---

[3] A comprehensive review of classification techniques can be found in [52]

the sinusoidal model is based on the observation that signals (particularly pitched sounds) can be decomposed into a sum of sinusoids. The efficient parametrisation only requires the sinusoid frequencies, amplitudes and the phases to be estimated. Sinusoidal modelling was later applied to music signal processing by Serra and Smith [101, 111], who extended the model to include a stochastic component to capture the non-periodic components of musical sounds such as the noise transients at the beginning of a sound. Since then, the sinusoidal model has been a powerful tool for estimating musical instrument parameters in order to ascertain higher-level information such as the notes played by each source [76]. Aspects of sinusoidal modelling such as the estimation of harmonic amplitudes, have also been used to estimate parameters for the synthesis of instrument models [119] and also for parametric coding [72, 120].

The sinusoidal model for a pitched note $x[n]$, is mathematically expressed as

$$x[n] = s[n] + r[n] \tag{2.13}$$

where

$$s[n] = \sum_{h=1}^{H} a_h[n] cos\left(2\pi f_h[n]n/f_s + \theta_h[n]\right) \tag{2.14}$$

where $s[n]$ is the deterministic sinusoidal component at time n, $r[n]$ is the stochastic residual component, $\{f_h, a_h, \theta_h\}$ are the time-varying parameters of the sinusoid's frequency, amplitude and phase respectively, f_s is the sampling frequency, h is the harmonic index and H is the number of harmonics. The majority of the analysis is typically devoted to precise estimation of the sinusoidal parameters and once these are found, the residual signal is found by subtracting out the sinusoidal signals.

Figure 2.8 illustrates the sinusoidal modelling paradigm. The time-domain signal $x[n]$ is first transformed into the time-frequency domain. The STFT is the representation of choice for most sinusoidal modelling systems such as the McAulay-Quatieri (MQ) system [81] and PARSHL [111], as it is a linear, invertible representation that provides the time-localised frequency domain behaviour of the signal. The parameters of the STFT that are selected are dependent on the source properties. For musical instrument sources, the parameters chosen should be able to account for the frequency and amplitude modulation effects from vibrato and glissando, and there should be sufficient time-resolution to model the amplitude changes such as the attack and decay of the notes.

Once the signal is transformed into the time-frequency domain, sinusoidal components are then identified. As sinusoids appear as peaks in the magnitude of a DFT, the sinusoidal components are typically estimated by picking the peaks of the magnitude response at each time frame [81, 111, 34]. Other methods of estimating sinusoidal components have also included the F-test [114, 125] and the cross-correlation method [45].

Sinusoid trajectories are then formed over time from the estimated peaks. At each time frame, decisions are made as to when to 'birth' a trajectory, add peaks to a trajectory and when to mark the 'death' of a trajectory [81]. A robust method of defining these trajectories normally includes a set of heuristics [81, 111, 34], however hidden Markov models (HMM) [26] and recursive least-squares [69] have also been used for the formation of trajectories. The frequencies of the trajectories are then estimated using either phase derivative methods [46], interpolation methods [111, 4], maximum likelihood estimation [94, 75] or least-squares optimisation [27]. This will be explored in more detail in Section 3.2.

The phase of each sinusoid is then estimated and in the MQ algorithm [81], this is accomplished by using a cubic interpolation function

$$\phi_h[n] = \zeta + \gamma n + \alpha n^2 + \beta n^3 \tag{2.15}$$

where the coefficients are found by matching sinusoidal frequencies at consecutive time frames with the additional constraint that the phase interpolation function be 'maximally smooth'. Phase estimation of the sinusoids is paramount for time-domain re-synthesis, for while the auditory system is known to be insensitive to the phase of non-periodic signals, it is sensitive to the phase of periodic signals.

Given the frequency, amplitude and phase estimates of the sinusoids, it is then possible to construct the complete deterministic component of the sinusoidal model.

This deterministic component is then subtracted from the original signal to obtain the stochastic residual component of the signal. This subtraction is performed either in the time-domain or the frequency-domain and the residual component is typically modelled as noise that is shaped by a magnitude spectral envelope [103, 38].

2.4 Conclusions

This chapter has explored the various representations and models of sound that are appropriate for music sound source separation. Music mixtures are unique to many other types of sound mixtures as they are arranged to be harmonically and temporally concordant. This produces a significant challenge for music separation systems.

As a representation of music source separation, the STFT is not a sophisticated representation, but it does fulfil many of the requirements outlined in Section 2.2. It is a linear, invertible, computationally efficient representation that offers a comparably significant amount of component separability, particularly for pitched sounds. Other multi-resolution transforms and physiologically motivated representations may better approximate the representation of the auditory system, but ultimately they fall short of the STFTs level of component separability, particularly in the higher frequencies.

The STFT also provides a good foundation on which to perform sinusoidal modelling. The sinusoidal model adequately parametrises the salient aspects of periodic signals, making it a good candidate for modelling the pitched notes found in music mixtures. While physical modelling synthesis provides efficient and accurate modelling of musical instruments, the sinusoidal model provides an instrument generality that is necessary for the separation of music mixtures that encompass a broad range of musical instruments.

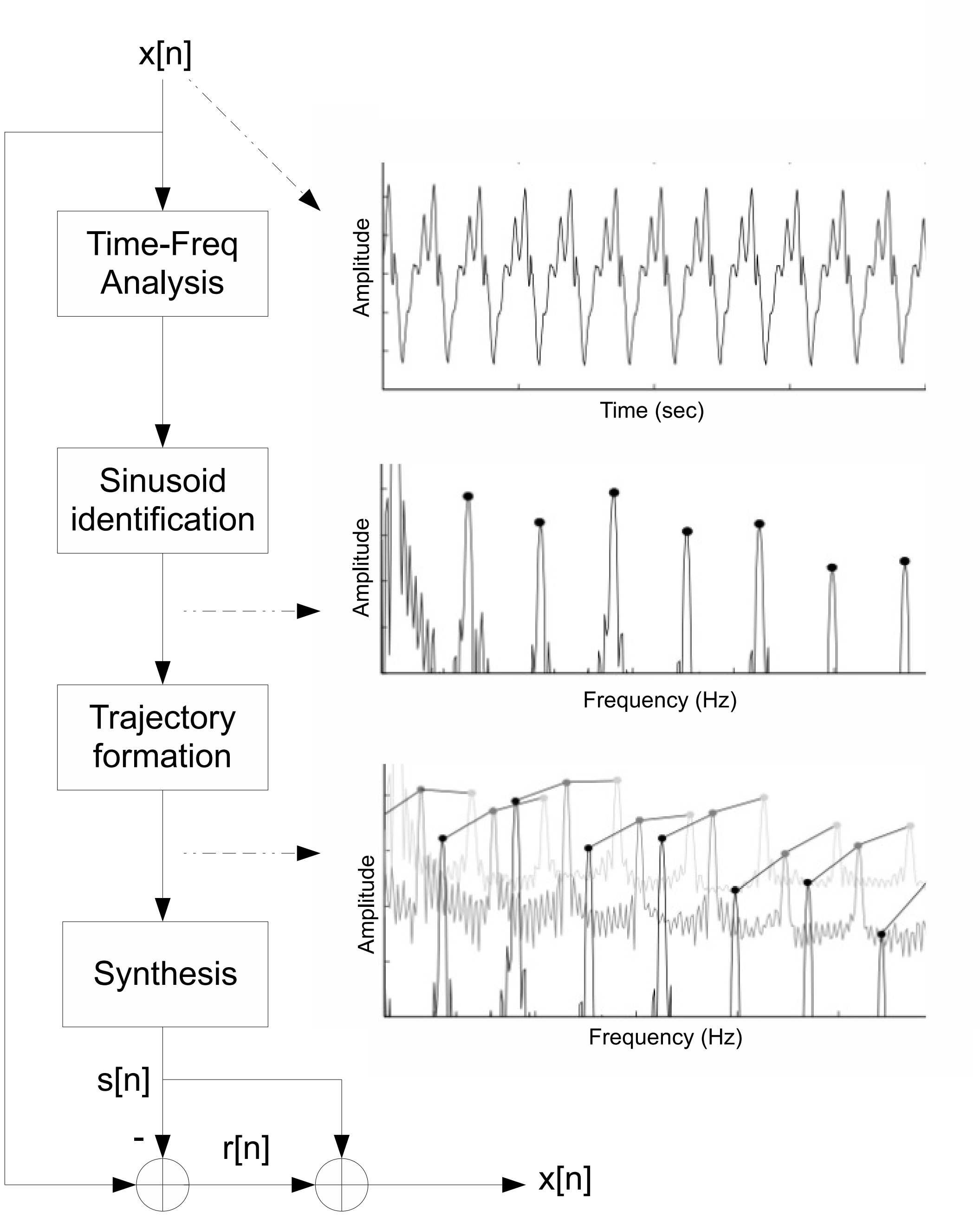

Figure 2.8: Sinusoidal modelling paradigm

Chapter 3

Harmonic Signal Modelling for Musical Mixtures and Separation

A music mixture typically contains a combination of pitched notes and non-pitched sounds. The pitched notes play an important role in music, as they combine in series to form the melodies of songs, and their simultaneous combination forms harmonies. While pitched notes can be decomposed into a deterministic component and a stochastic component, it is the deterministic component that greatly influences the perception of pitch [113, 82]. The deterministic component, being quasi-periodic, manifests itself in the spectral domain as a series of harmonically spaced sinusoids that are known as *harmonics*, which are adequately modelled using a sinusoidal model (Section 2.3.2).

In this chapter, various aspects of signal modelling for mixtures of pitched notes are explored, with the aim of producing parametric representations to facilitate source separation. Section 3.1 begins with an overview of methods for detecting harmonics in a STFT representation, and this is followed by the introduction of a novel, computationally efficient algorithm to refine the frequency estimates of the detected harmonics in Section 3.2. In a model-based source separation architecture, the estimated harmonic frequencies and magnitudes are then grouped to form sinusoidal trajectories, or *harmonic tracks*, which are then grouped to form sources. An assumption often exploited in this process, whether for track formation or track grouping, is the postulation that the harmonic tracks of a given source are highly correlated. While this assumption is frequently employed in separation systems, there are a lack of formal investigations into the nature of these correlations over a variety of instruments. In Section 3.3, a large set of musical instrument samples are analysed and a novel generalised model to predict harmonic magnitude tracks from neighbouring harmonic tracks is presented. The model is then evaluated in the context of harmonic mixtures and is shown to provide improvements compared to existing methods.

3.1 Spectral Peak Picking

The initial objective for modelling harmonic sources as a sum of sinusoids is the detection of the harmonics. Once the harmonics have been identified, their sinusoidal parameters (frequency, amplitude and phase) can then be refined and the harmonics can be grouped into their relevant sources (Section 2.3.2).

There have been various approaches to identifying harmonics within a mixture, including simple peak picking methods [81, 111], sinusoidal likeness measures using the F-test [114, 125] and the cross-correlation method [45], and peak picking methods with spectrum dependant thresholds [34]. In the peak picking methods of [81, 111], the detection of harmonics is based on the observation that sinusoids appear as peaks in a magnitude spectrum. However, there are other peaks which arise from non-sinusoidal, noise-like components particularly at lower levels, and a simple peak picking algorithm usually results in the identification of many false positives.

Sinusoidal likeness measures [114, 125, 45] are an alternative to peak picking, and aim to identify sinusoids by searching the magnitude spectrum for shapes that resemble the magnitude response of sinusoids. In [45], the cross-correlation between the short-time spectrum of the signal and the spectrum resulting from an ideal sinusoid is obtained, and the result is scaled by the overall spectral shape. This produces a sinusoidal likeness measure where regions of the spectrum resembling a sinusoid result in a higher values, and regions of the spectrum that do not resemble a sinusoid result in a lower values. Since cross-correlation is the same as convolution with one signal having an inverted time-scale, this method can be efficiently implemented in the time domain by observing that convolution in the frequency domain is equivalent to multiplication in the time domain. The F-test [114, 125], uses discrete prolate spheroidal sequences to provide a sinusoidal likeness measure for each frequency, and it performs well under ideal conditions, discriminating between sinusoids and noise peaks. The computational cost however is high, due to the calculation of several FFTs.

All of the sinusoidal likeness measures perform well when sinusoids appear in isolation in the spectrum, but in the case of music mixtures where harmonics are frequently overlapping, these methods fail to consistently identify the necessary harmonics. When harmonics overlap, the resulting shape in the spectrum can be far from ideal, producing a considerable number of false negatives.

The method described in [34] addresses these issues by obtaining the spectral peaks of each frame which lie above a spectrum-dependant threshold $E(k)$. This peak picking method provides robust identification of harmonic sinusoid-like components in music mixtures while eliminating undesirable peaks due to windowing effects of the STFT and other noise.

Consider a discrete short-time Fourier transform (STFT) computed on a mixture of harmonic musical instrument notes $y[n]$, where the FFT length is N samples, sampling rate f_s , using a Hanning analysis window and hop size H samples. $Y_r(k)$ denotes the complex STFT coefficient at frequency k, at the rth frame and at frequency bin $k \in [0, N-1]$, which corresponds to the frequency $f_k(r) = \frac{kf_s}{N}$ Hz.

The threshold $E(k)$ is calculated by convolving the magnitude spectrum $|Y_r(k)|$ with a normalised Hamming window $G(k)$ of length $1 + N/64$ samples, and raising the result to the power of a compression constant c as given by Equation 3.1 [34].

$$E(k) = \left(\sum_{l} |Y_r(l)| \, G(k-l) \right)^{c} \tag{3.1}$$

where l are the samples over which the convultion is performed. Thus $E(k)$ is a smoothed amplitude envelope threshold where the flatness of the curve is determined by the compression constant c. Suitable values of c lie in the range $[0.5, 1]$. Figure

3.1 illustrates the effect of varying the compression constant c. Smaller values of c result in a flatter threshold $E(k)$ thereby increasing the number of peaks detected particularly at lower frequencies. Conversely, larger values of c result in a smaller number of peaks detected, reducing the probability of detecting a noise peak, but also increasing the probability of missing low frequency harmonics. Once $E(k)$ has been calculated, the peaks which lie above the threshold $E(k)$ are classified as the harmonic candidates of the short-time spectrum.

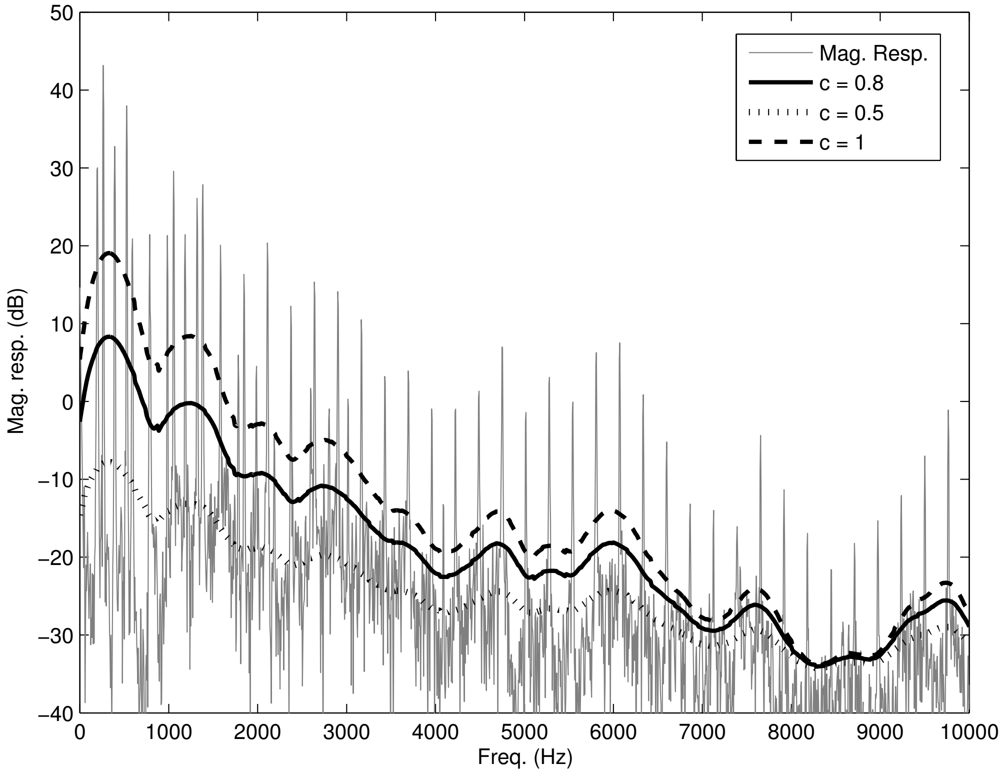

Figure 3.1: Threshold $E(k)$ for various compression constants c, calculated for a mixture of an alto saxophone and a piano playing the notes C4 and G3 respectively.

3.2 Frequency Estimation

The estimation of sinusoidal parameters is a widely studied area and has been extensively employed in many audio applications. Sinusoidal modelling in particular [81, 111, 38, 103] has been widely used to represent the dominant harmonic components found in musical signals and a major component of such modelling requires the accurate estimation of sinusoidal parameters.

Most sinusoidal parameter estimation algorithms obtain their estimates of sinusoid parameters by analysing the magnitude spectrum obtained typically from a Fast Fourier Transform (FFT). Frequency bins corresponding to a potential frequency component are then interpolated according to a predefined model or curve [111], and further refinement can take place in the form of optimisation by Newton's method [2], or by other maximum likelihood estimation methods [94, 75, 92]. Other estimation methods such as least-squares optimisation [27] and the Hilbert Transform [5] have also been used.

Most of the latter methods are quite computationally expensive and it is the simplicity of methods such as the Quadratically Interpolated FFT (QIFFT) [111] that make them popular in audio processing [64].

In this section, we propose the Phase Derivative FFT (PDFFT) [46] - a computationally efficient method for estimating the frequency of a sinusoid using the time derivative of the phase response obtained from the Short Time Fourier Transform (STFT). We show that the algorithm's accuracy outperforms the QIFFT zero-padded 2.5 times and computationally only requires 4 multiplies per peak. Estimation of multiple closely spaced frequencies is then attempted and shown to be successful even from a single peak in the magnitude spectrum.

3.2.1 Phase-Derivative FFT

Consider the sinusoid $x(t) = \cos(2\pi f_0 t + \phi_0)$, where f_0 is the frequency in hertz and ϕ_0 is the phase in radians. Let $H = \alpha T_0$, where $T_0 = \frac{1}{f_0}$ and α is a constant, and define

$$p = H - mT_0 \tag{3.2}$$

where $m = \left\lfloor \frac{H}{T_0} \right\rfloor$ is an integer denoting the number of whole periods of T_0 there are in H as illustrated in figure 3.2.

The instantaneous phase $\phi(r)$ of the sinusoid at a given point $t = rH$ is given by,

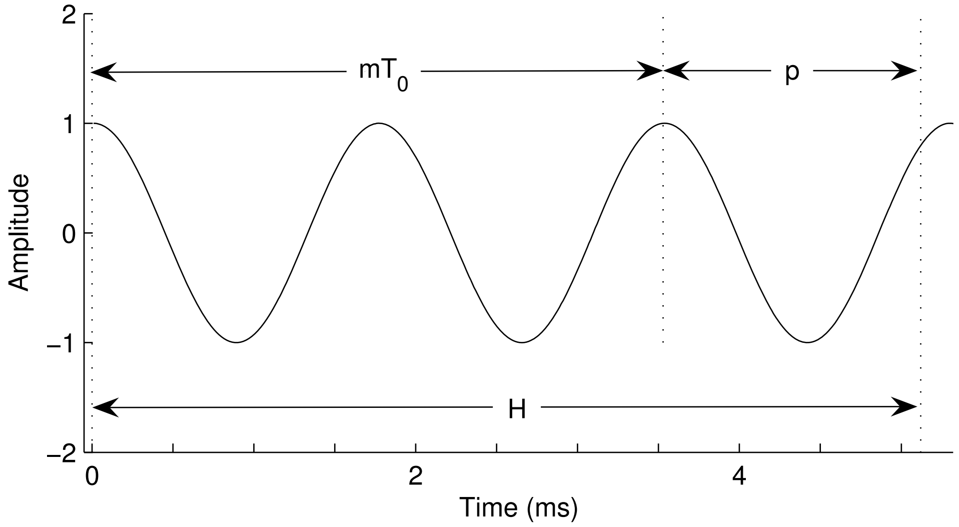

Figure 3.2: Segmentation of a sinusoid where $p = H - mT_0$

$$\phi(r) = 2\pi \frac{p}{T_0} r + K \tag{3.3}$$

where r is an integer and K is the arbitrary phase of the sinusoid at $r = 0$. Taking the derivative with respect to r, we obtain,

$$\frac{d\phi}{dr} = 2\pi \frac{p}{T_0} \tag{3.4}$$

where $\frac{d\phi}{dr} \in [0, 2\pi)$. Substituting equation (3.2) and $m = \left\lfloor \frac{H}{T_0} \right\rfloor$ into (3.4) we obtain

$$\frac{d\phi}{dr} = 2\pi \left(\frac{H}{T_0} - \left\lfloor \frac{H}{T_0} \right\rfloor \right) \tag{3.5}$$

Now if the signal $x(t)$ is then sampled at f_s (Hz), then equation (3.5) becomes,

$$\frac{d\phi}{dr} = 2\pi \left(\frac{Hf_0}{f_s} - \left\lfloor \frac{Hf_0}{f_s} \right\rfloor \right) \tag{3.6}$$

Observe that equation (3.6) can be related to the STFT by defining H as the hop size between frames and r as the rth frame. If a N point DFT is used, a peak in the magnitude response will be observed at bin b and the true frequency will be found at $b_0 = b + \Delta_b$, where Δ_b is the deviation from bin b. Writing equation (3.6) in terms of b_0 such that $f_0 = \frac{b_0 fs}{N}$, gives

$$\frac{d\phi_b}{dr} = 2\pi \left(\frac{Hb_0}{N} - \left\lfloor \frac{Hb_0}{N} \right\rfloor \right) \tag{3.7}$$

and rearranging for b_0 produces

$$b_0 = b + \Delta_b = \frac{N}{H} \left(\frac{d\phi_b}{dr} \frac{1}{2\pi} + \left\lfloor \frac{Hb_0}{N} \right\rfloor \right)$$

Finally, due to the non-linearities of the floor function we can assume that

$$\left\lfloor \frac{Hb_0}{N} \right\rfloor \approx \left\lfloor \frac{Hb}{N} \right\rfloor \tag{3.8}$$

and hence obtain an expression to approximate the true frequency of a sinusoid in terms of the DFT bins:

$$b_0 = \frac{N}{H} \left(\frac{d\phi_b}{dr} \frac{1}{2\pi} + \left\lfloor \frac{Hb}{N} \right\rfloor \right) \tag{3.9}$$

Thus equation (3.9) shows that simply using parameters obtained from the STFT and 4 multiplies, we can obtain an accurate estimate of the true frequency f_0.

3.2.1.1 Frequency Estimation Errors

Analytically, the frequency estimation error, can be found by analysing:

$$b_0 + \delta_b = \frac{N}{H} \left\{ \left(\frac{d\phi_b}{dr} + \delta_{\phi_b} \right) \frac{1}{2\pi} + \left\lfloor \frac{Hb}{N} \right\rfloor + \delta_{floor} \right\} \tag{3.10}$$

Thus the error component can be denoted by:

$$\delta_b = \frac{N}{H} \left(\frac{\delta_{\phi_b}}{2\pi} + \delta_{floor} \right) \tag{3.11}$$

where δ_b is the frequency estimation error in bins, δ_{ϕ_b} is the estimation error of $\frac{d\phi_b}{dr}$ and δ_{floor} is the error incurred by the assumptions made in the floor function. Equation (3.11) indicates that there are two possible sources of error within the algorithm. The first is due to errors from the floor function $\lfloor \frac{Hb}{N} \rfloor$ while the second arises from erroneous estimation of $\frac{d\phi_b}{dr}$.

To observe the errors in the floor function, we recall the assumption made in equation (3.8) and see that

$$\left\lfloor \frac{Hb_0}{N} \right\rfloor = \left\lfloor \frac{Hb}{N} + \frac{H\Delta_b}{N} \right\rfloor = I_b + \left\lfloor F_b + \frac{H\Delta_b}{N} \right\rfloor \tag{3.12}$$

where $I_b = \left\lfloor \frac{Hb}{N} \right\rfloor$ and $F_b = \frac{Hb}{N} - \left\lfloor \frac{Hb}{N} \right\rfloor$.

It is easy to see from equation (3.12) that the floor function error δ_b is an integer that is purely dependent on $F_b + \frac{H\Delta_b}{N}$. Eradicating δ_{floor} therefore simply involves testing values of δ_{floor} in equation (3.10) choosing Δ_b such that

$$\Delta_b = \min\{|\Delta_{b-1}|, |\Delta_b|, |\Delta_{b+1}|\} \tag{3.13}$$

where $\Delta_b = \frac{N}{H}\left(\frac{d\phi_b}{dr}\frac{1}{2\pi} + \left\lfloor \frac{bH}{N} \right\rfloor\right) - b$; $\Delta_{b-1} = \Delta_b - \frac{N}{H}$; $\Delta_{b+1} = \Delta_b + \frac{N}{H}$.

The entire frequency estimation error is thus purely attributed to δ_{ϕ_b} and its severity is dependent on the inverse proportion of the hop size used as seen in equation (3.14).

$$\delta_b = \frac{N\delta_{\phi_b}}{H2\pi} \tag{3.14}$$

Estimation of $\frac{d\phi_b}{dr}$ can be made by differencing between frames $\frac{d\phi_b}{dr} = \phi_b(r) - \phi_b(r-1)$, and this has been used for the duration of this thesis.

3.2.1.2 The PDFFT Method

The estimation of a sinusoid's frequency can be therefore summarised as follows:

1. Calculate the STFT of a signal with windowed overlapping frames with hop size H, using an N point FFT. Note that windowing has negligible effects on the phase response if we use a symmetric window.

2. For a given STFT frame r, locate the bin number b corresponding to a peak in the magnitude spectrum.

3. Compute the time derivative of the phase $\frac{d\phi_b}{dr}$.

4. Solve equation (3.9) choosing appropriate Δ_b.

5. Iterate steps 2 to 4 for all peaks in the magnitude spectrum.

3.2.2 PDFFT Performance

3.2.2.1 Sinusoidal Frequency Estimation

The PDFFT estimation is based on an initial coarse estimate from the peaks of the magnitude response of an N point FFT (where N is a power of 2). The initial FFT estimation error of bin b is simply the deviation Δ_b where the true frequency is $f_0 = \frac{(b+\Delta_b)fs}{N}$. The deviation Δ_b thus reduces for increasing N, having a maximum deviation of $\Delta_b = \frac{fs}{2N}$. Figure 3.3 illustrates the performance of the PDFFT for various sized N point FFT where $H = \frac{N}{4}$ and $fs = 44100$ Hz, with an $N = 2048$ FFT provided for reference. Additive White Gaussian Noise (AWGN) is added to a pure sinusoids to obtain the various SNR levels. At $SNR = 50$ dB, a $N = 2048$ point PDFFT provides almost 1000 orders of magnitude greater accuracy than a standard FFT. This accuracy comes only at the computational expense of 4 multiplications per peak as opposed to the $\widehat{N}log_2\widehat{N} - Nlog_2N$ extra multiplications required to achieve the same frequency resolution from zero-padding to a $\widehat{N}$ point FFT.

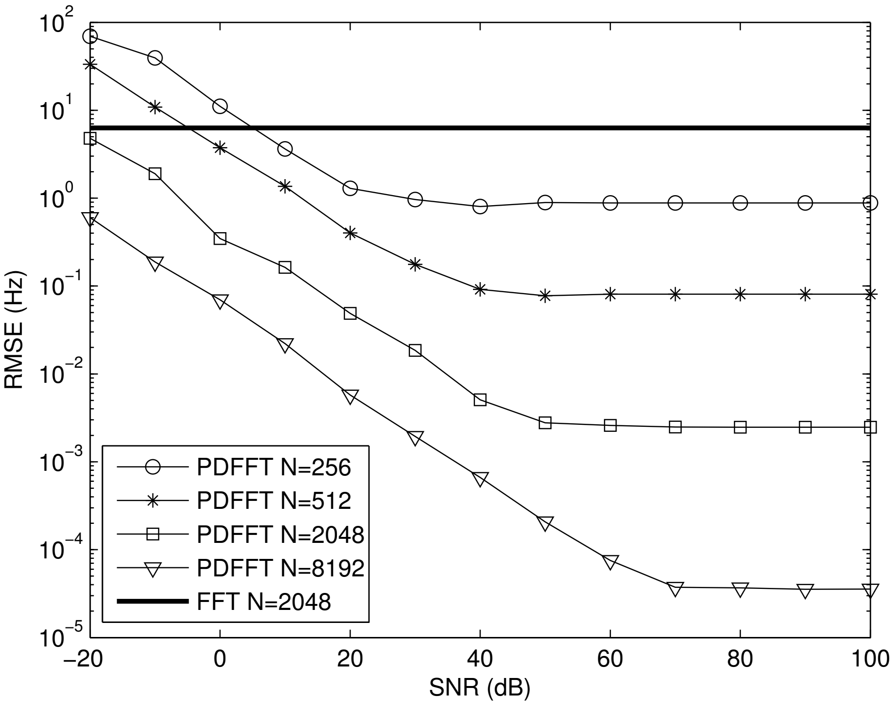

Figure 3.3: Comparison of PDFFT algorithm performance using various N point FFTs for coarse estimates. $N = 2048$ FFT provided for reference.

Figure 3.3 also illustrates the asymptotic nature of the RMS errors with respect to a given SNR for different values of $N = 256, 512, 2048, 8192$. As Additive White Gaussian Noise (AWGN) is added to a pure sinusoid, the algorithm reaches a certain threshold of minimum error for a given value of N. This minimum error decreases as the values of N increase and the asymptotic characteristic commences at higher SNR for higher values of N. Thus an application specific compromise must be made by balancing computational efficiency (choice of N) with estimation accuracy for a given SNR.

As seen in Equation 3.14 the PDFFT is highly dependent on the estimation of $\frac{d\phi_b}{dr}$. Figure 3.4 shows the reliability of the unwrapped phase response ϕ_b subject to different SNR. For $SNR > -10dB$, the phase response is quite reliable; while for lower SNR, the phase response is corrupted by noise, thereby producing erroneous estimates of $\frac{d\phi_b}{dr}$.

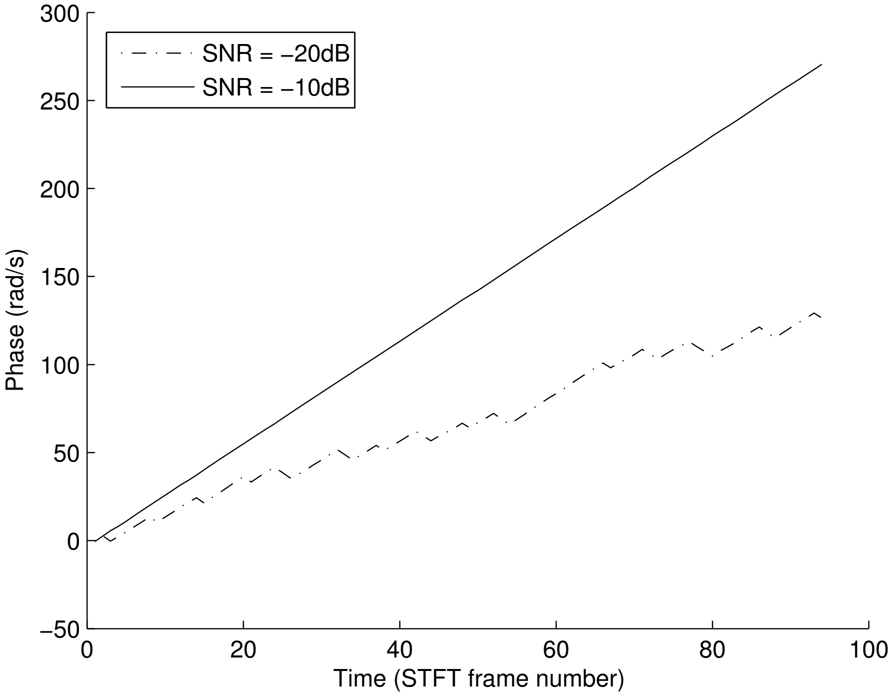

Figure 3.4: Unwrapped phase response of an FFT bin b subject to noise.

3.2.2.2 Comparison of the PDFFT to the QIFFT

The QIFFT method is widely used in audio applications for its simplicity and accuracy [64]. The bias errors incurred however require that for a hamming window, zero-padding factors of 2.4 or greater [4] need to be used, thereby adding to the computational costs of the method. The PDFFT on the other hand does not require any zero-padding and as figure 3.5 illustrates, the performance of a $N = 2048$ point FFT provides more than 10 orders of magnitude better frequency estimation than a $N = 2048 * 2.5 = 5120$ point QIFFT at higher SNR. At lower SNR, the PDFFT only provides marginally better frequency estimation than the QIFFT but at a far less computational cost since the QIFFT is zero-padded by a factor of 2.5.

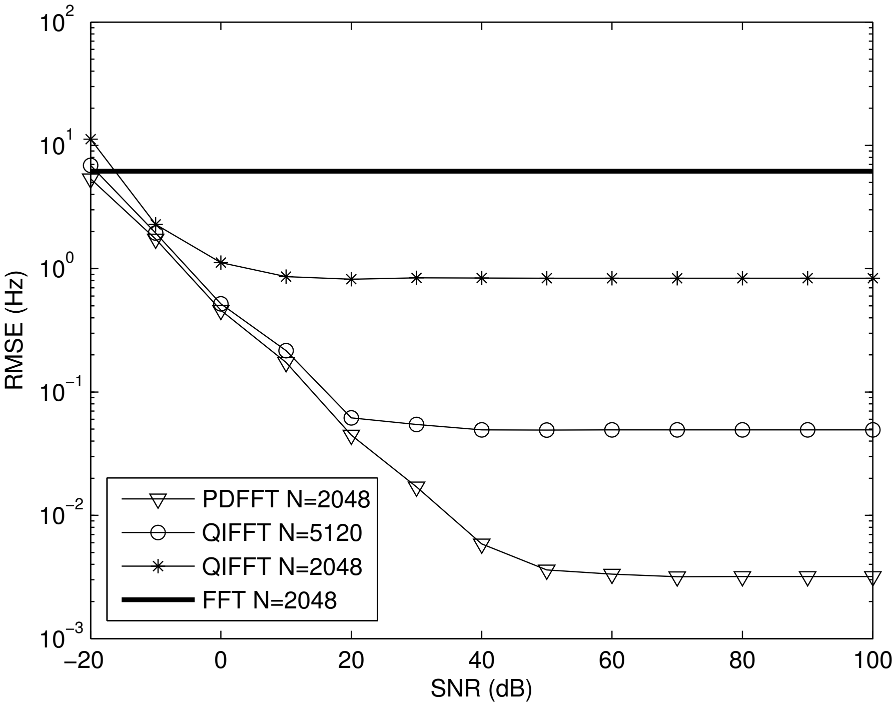

Figure 3.5: Comparison of PDFFT and QIFFT with FFT as reference.

3.2.3 Multiple Sinusoid Frequency Estimation

The limitations of interpolation methods such as the QIFFT are obviously their inability to resolve multiple sinusoidal frequencies that are in close frequency proximity. Typically, frequencies separated by less than 3 FFT bins result in a single peak in the magnitude spectrum and thus interpolation methods estimate erroneous peaks which are actually the superposition of two sinusoids.

Simple modification of the PDFFT however, allows it to resolve two frequency components even if they are presented as a single peak in the magnitude spectrum. Two sinusoids of differing frequencies will each have different phase time derivatives and thus by carefully choosing the bin from which to obtain the phase time derivative, it is possible to obtain reasonable estimates of both frequencies.

3.2.3.1 Two Tone Detection

Given a peak in the magnitude spectrum, a distinction between a single prominent sinusoid and multiple sinusoids in close proximity must be made. This detection task can be performed by analysing the variance of $\frac{d\phi}{dr}$ over adjacent bins, i.e.

$$v_\Phi = var\left(\Phi\right) \tag{3.15}$$

where Φ is the set

$$\Phi \in \left\{\frac{d\phi_{b-1}}{dr}, \frac{d\phi_b}{dr}, \frac{d\phi_{b+1}}{dr}\right\} \tag{3.16}$$

For a sinusoid at frequency f and a second sinusoid of equal amplitude at $f+\Delta_f$, a single peak in the magnitude spectrum will be observed if $|\Delta_f| \lessapprox 3$ bins. Within this range however, the phase information is sufficiently dissimilar to be able to determine if there is a single frequency or multiple frequencies. If there is only a single frequency present in the peak, the adjacent bins will have similar phase time derivatives (i.e. v_Φ is small); on the other hand, if there are multiple frequencies which have superimposed to form a single peak in the magnitude spectrum, then the phase time derivatives of adjacent bins will be very different (i.e. v_Φ is large - see also Figure 3.6). Figure 3.6 illustrates this by plotting v_Φ over various Δ_f. Results were obtained by averaging the results of 500 randomly selected frequencies where $fs = 44100$ Hz, $N = 2048$, $H = \frac{N}{4}$.

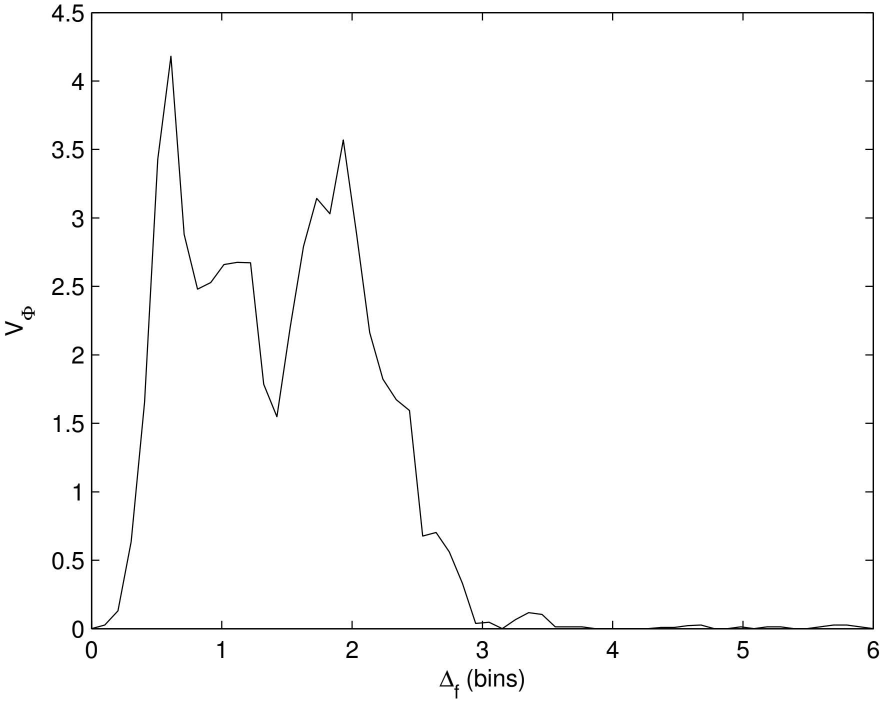

Figure 3.6: v_Φ vs Δ_f

3.2.3.2 Two Tone Frequency Estimation

Having verified that two sinusoids are present forming a single peak at bin b, good phase time derivative estimates for two frequencies of equal amplitude can be obtained at bins $b-1$ and $b+1$. At these bins, informal experiments have shown that the phase responses over time are usually sufficiently unique, providing the required phase information to resolve the two frequencies.

Figure 3.7 illustrates the unwrapped phase of bins $b-1$, b and $b+1$; where a peak exists in the magnitude spectrum at bin b. The phase responses over time are clearly linear for bins $b-1$ and $b+1$ resulting in constants for $\frac{d\phi_{b-1}}{dr}$ and $\frac{d\phi_{b+1}}{dr}$. In this particular case, this yields frequency estimates of $\widehat{f} = 1999.0176$ Hz and $\widehat{f} + \widehat{\Delta}_f = 2026.9696$ Hz which are comparable to the true frequencies $f = 2000$ Hz and $f + \Delta_f = 2026.9709$ Hz.

Figure 3.8 illustrates the PDFFT sinusoidal estimation capabilities when presented with two sinusoids of equal amplitude separated by Δ_f, where Δ_f is swept from 0 to 6 bin spacings, with $fs = 44100Hz$, $N = 2048$, $H = \frac{N}{4}$. Without in-

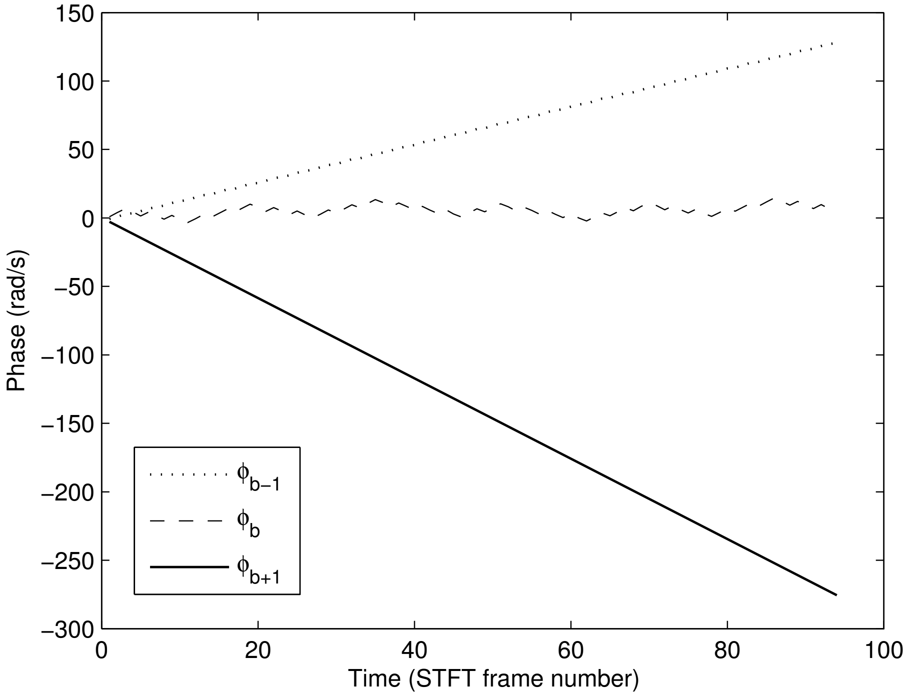

Figure 3.7: Unwrapped phase responses at bins $b-1$, b and $b+1$. $f_0 = 2000$ Hz and $\Delta_f = 26.97$ Hz.

creasing the number of computations per peak, the PDFFT is able to provide good frequency estimation even when the frequency separation is small.

3.2.4 Conclusions

The PDFFT - a novel sinusoidal frequency estimation algorithm is presented. The algorithm builds upon the coarse frequency estimate provided by the FFT and utilising only parameters found from the STFT, computes a highly accurate estimation of the frequency of a sinusoid using the time derivative of the phase response. For single sinusoid frequency estimation, the accuracy of the PDFFT outperforms the frequently employed QIFFT even with zero-padding, at the expense of only 4 multiplies per peak. Unlike other interpolation methods, the PDFFT is shown to also perform well at resolving multiple frequencies from a single peak in the magnitude spectrum.

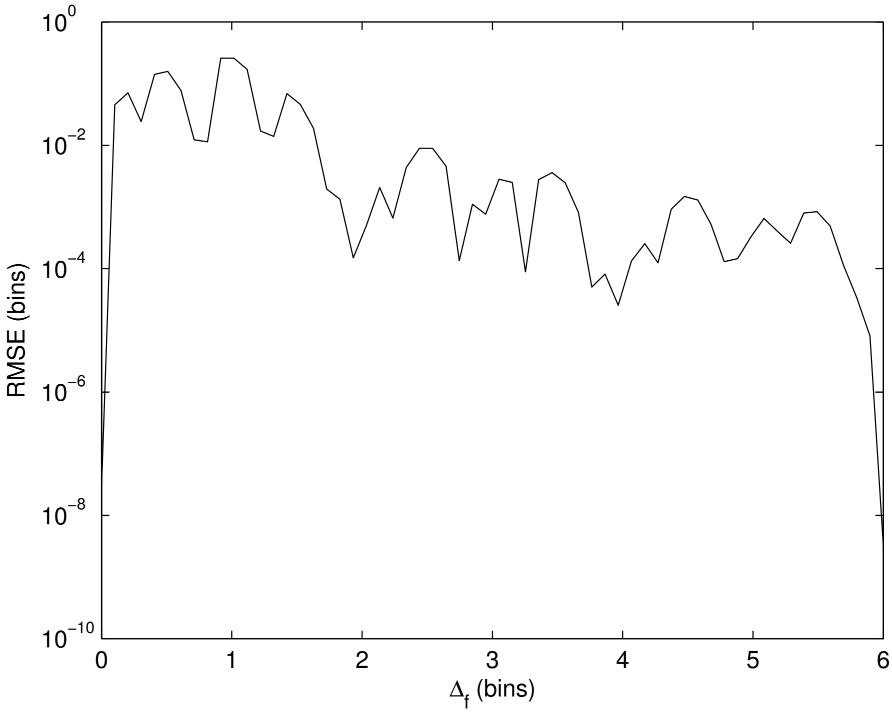

Figure 3.8: PDFFT estimation errors when a tone ($f = 777Hz$) is subject to a second tone of equal amplitude Δ_f apart.

3.3 Harmonic Track Analysis

3.3.1 Harmonic Tracks

In sinusoidal modelling, once the magnitudes and the frequencies of the harmonics within a mixture have been determined, the harmonic peaks are formed into sinusoidal trajectories or *tracks*. These tracks form the fundamental parametrisation of the sinusoidal model, and summarise a vast proportion of the deterministic component of harmonic sources. Figure 3.9 illustrates the tracking of the harmonics of an alto saxophone over 3 adjacent frames.

The formation of harmonic tracks in a music source separation system, is often not a straightforward task of tracking the trajectory of an isolated harmonic over time. Due to the vast number of overlapping components, as well as source-specific nuances such as inharmonicities, estimating harmonic trajectories for musical mixtures is not trivial. This has resulted in a number of approaches to harmonic track formation, many of which have been heuristic in their approach [81, 125, 34]. McAulay and

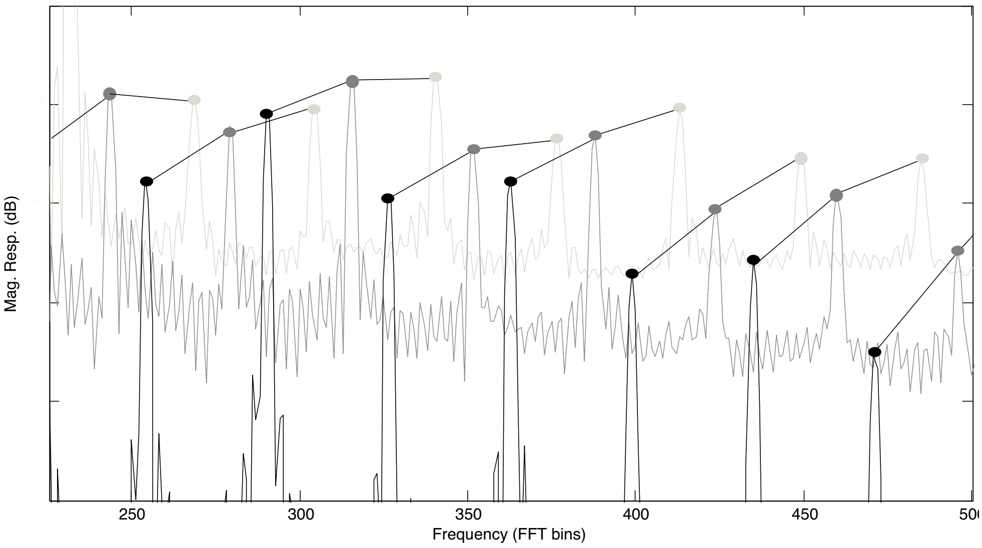

Figure 3.9: Harmonic track formation for an alto saxophone over 3 successive frames. The dots indicate the harmonic peaks whose frequencies correspond to F_j^h, and whose amplitudes correspond to M_j^h. The straight lines joining the dots represent the formation of tracks from the harmonic peaks.

Quatieri [81] first presented a heuristic method for defining sinusoidal trajectories, prescribing a systematic method for determining the birth and death of sinusoidal tracks. Similar methods were later employed in separation systems, such as those by Virtanen [125] and Every [34]. Other methods of harmonic tracking include the linear prediction approach presented by Lagrange [70], which is based on the assumption that harmonic tracks should be slow time-varying and predictable. Nunes [69] opted for a recursive least-squares approach to harmonic tracking, while Depalle employed the use of hidden Markov models [26].

As different as each of these methods are, fundamentally, they all estimate harmonic tracks independent of the surrounding tracks. In case of real instruments, the sound production mechanics are often non-linear, giving rise to the natural assumption that there are substantial correlations between the harmonic tracks originating from the same source. These correlations could be used to develop better harmonic track estimates, however there have been no formal studies investigating and modelling the track dependencies for musical instruments. In the following sections, we explore the intra-instrument correlations which exist between harmonic musical sources, and we develop a generalised musical instrument model to characterise these correlations. The model is then evaluated in the context of harmonic track identification in harmonic musical mixtures.

Developing a deeper understanding of the nature of these correlations provides useful insights into the temporal similarities which exist between harmonic tracks. This knowledge is potentially useful for the development of robust tracking algorithms, and as Chapter 6 will show, modelling these intra-instrument correlations can be used to facilitate the separation of harmonic sources.

3.3.2 Intra-Instrument Correlation of Harmonic Magnitude Tracks

The temporal evolution of the harmonic track magnitudes of a source are often assumed to be correlated and in this section, we quantify this assumption. 3000 musical instrument samples from the University of Iowa instrument database [1] were used to investigate the similarity of the temporal envelopes of each sample's harmonic magnitude tracks. For each sample, the STFT was calculated and the harmonic peaks of the magnitude response for each frame were identified (see Section 3.1). The temporal envelopes for each harmonic magnitude track were then defined as:

$$M_s^i(r) = \sum_{k \in \kappa} 20log_{10} \left| X_r(k) \right| \tag{3.17}$$

where $M_s^i(r)$ is the temporal envelope for the ith harmonic magnitude track of source s, at frame r, $X_r(k)$ is the short-time Fourier transform of the sample and $\kappa \in [\rho_{ir} - b, \rho_{ir} + b]$ where ρ_{ir} is the ith harmonic's FFT bin corresponding to the peak at frame r and $b = 1$. The parameters of the STFT are: $N = 4096$, hop size $H = 1024$, Hanning window and sampling frequency $f_s = 44.1$ kHz.

The temporal envelope contains information about the note onset, offset and amplitude modulation. These attributes are illustrated in Figure 3.10, showing the first 6 harmonic magnitude tracks of an alto saxophone playing an E3 with vibrato. The envelopes are clearly correlated, having distinctly similar onset times, offset times and amplitude modulation. The similarity between the temporal envelope harmonics can be described as the normalised dot product [127]:

$$S\left(M_s^i, M_s^j\right) = \frac{\sum_r M_s^i(r) M_s^j(r)}{\sqrt{\sum_r M_s^i(r)^2}\sqrt{\sum_r M_s^j(r)^2}} \tag{3.18}$$

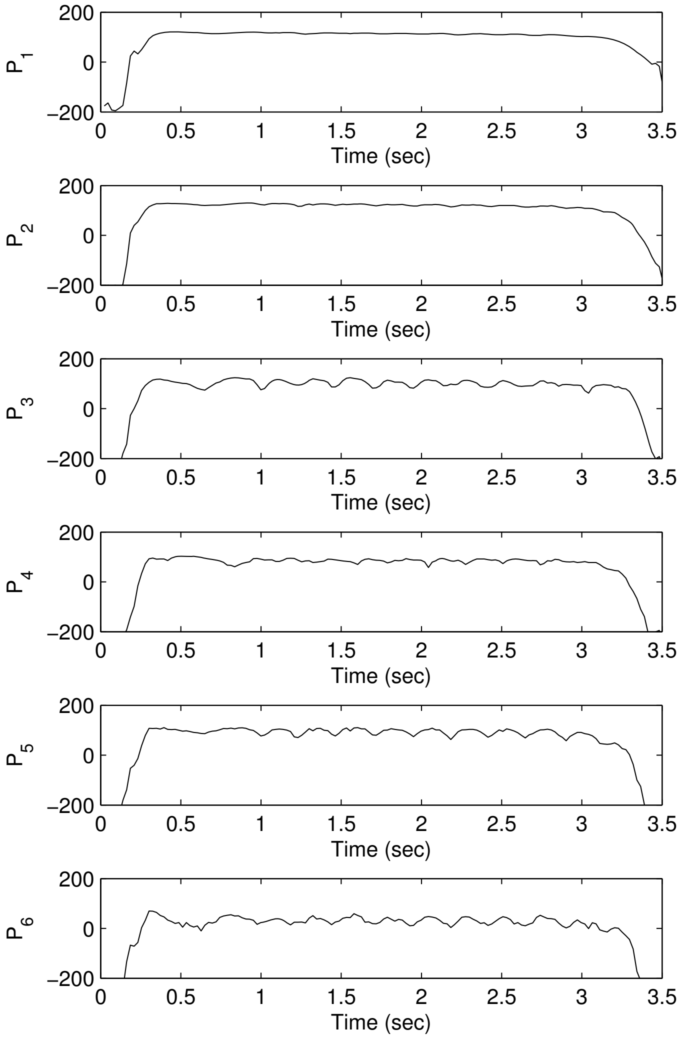

Figure 3.10: First 6 harmonic magnitude tracks of an alto saxophone sample playing an 'E3' with vibrato.

Similarity matrices were calculated for each of the samples limiting the number of partials to the first 50. The samples included notes recorded at a variety of pitches, intensities and styles (e.g. vibrato, pizzicato, ...), from a large range of instruments including the alto flute, alto saxophone, bass clarinet, bass flute, bassoon, bass trombone, Bb clarinet, cello, Eb clarinet, flute, horn, oboe, piano, soprano saxophone, tenor trombone, trumpet, tuba, viola and violin. To investigate the general similarities over the large number of instruments, the 3000 matrices were then averaged to obtain the results shown in Figure 3.11. The temporal envelopes of the harmonic magnitude tracks are clearly similar, particularly between adjacent harmonics. The first few harmonics also have very strong similarity with each other.

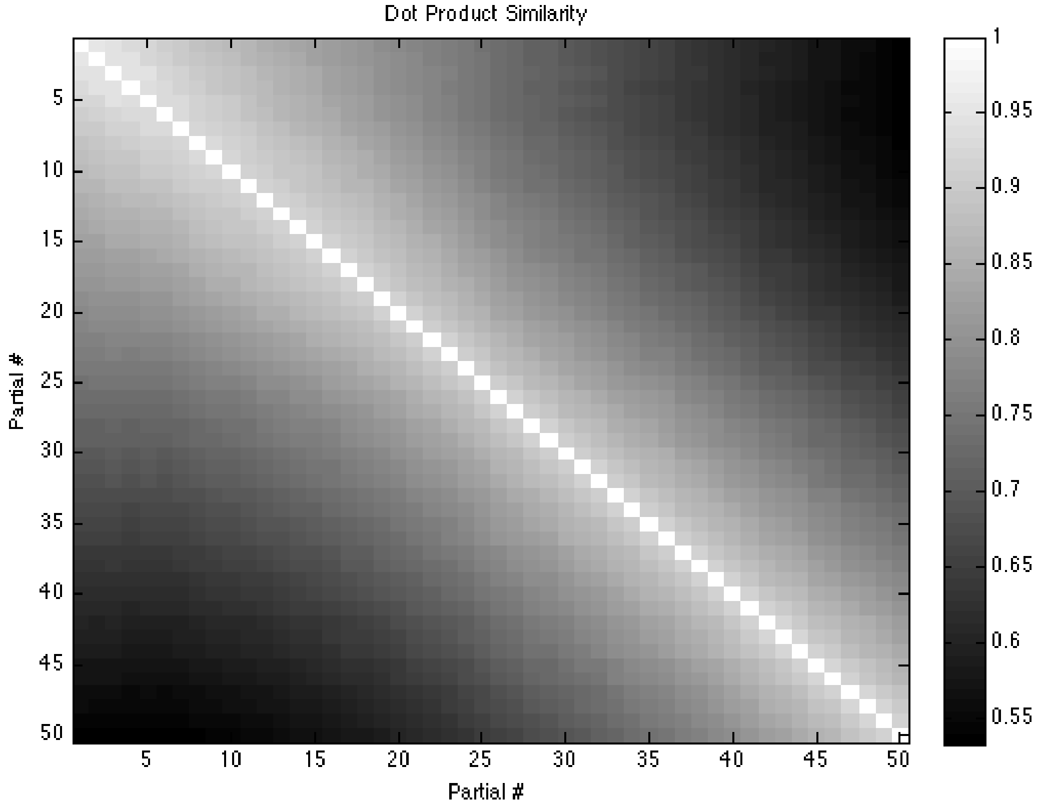

Figure 3.11: Similarity matrix of temporal envelopes over 3000 samples. Lighter regions indicate high similarity.

3.3.3 Modelling the Correlation of Magnitude Tracks

Knowing that neighbouring harmonics are highly correlated with one another is useful for identifying ambiguous harmonics in polyphonic mixtures and estimating corrupted, overlapped harmonics. If harmonics can be predicted by a linear combination of neighbouring harmonics with high accuracy, then harmonics can be identified by simply measuring the distance between ambiguous harmonics and the predicted harmonics. Furthermore, the estimates can be used to predict corrupted harmonics. We call these predicted harmonic temporal envelopes *model magnitude tracks* and we aim to find a general set of weighting functions w that will optimally predict a harmonic magnitude track $M_s^q(r)$ for the qth harmonic:

$$\hat{M}_s^q(r) = \frac{\sum_i w_{q,i} M_s^i(r)}{\sum_i w_{q,i}}, i \neq q \tag{3.19}$$

An optimal solution for these weights in a least squares sense can be found by regularised least squares, which has the solution:

$$w_q = \left(\Upsilon^T \Upsilon + \delta I\right)^{-1} \Upsilon^T M_s^q \tag{3.20}$$

where Υ is a matrix with columns containing the harmonic temporal envelopes but excluding the harmonic envelope to be predicted $\hat{M}_s^q(k)$, δ is the regularisation parameter, I is the identity matrix and ^{T} denotes the conjugate transpose. The weights were calculated for each harmonic of the 3000 instrument samples and then general weighting functions were found by averaging them. The shapes of these weighting functions are shown in Figure 3.12 and emphasise the similarity findings found earlier. Model magnitude tracks are best predicted as a linear combination of the neighbouring harmonic magnitude tracks and the first few harmonics exhibit very high similarity between each other. The shapes of these weighting functions are quite distinctive and were then found to be adequately modelled by:

$$w_{q,i} = \begin{cases} \frac{-c}{i-q}, & i < q \\ \\ \frac{c}{i-q}, & i > q \end{cases} \tag{3.21}$$

for the weight of the qth harmonic to model the ith harmonic, where $c = \frac{1}{q+5} + 0.1$. The error resulting from modelling the average weights, was a root mean squared error of less than -36 dB.

Figure 3.13 shows the prediction performance of each method. When the weights directly from the regularised least squares calculation for each instrument are used as the model magnitude tracks, they obviously yield the lowest error. The Noise

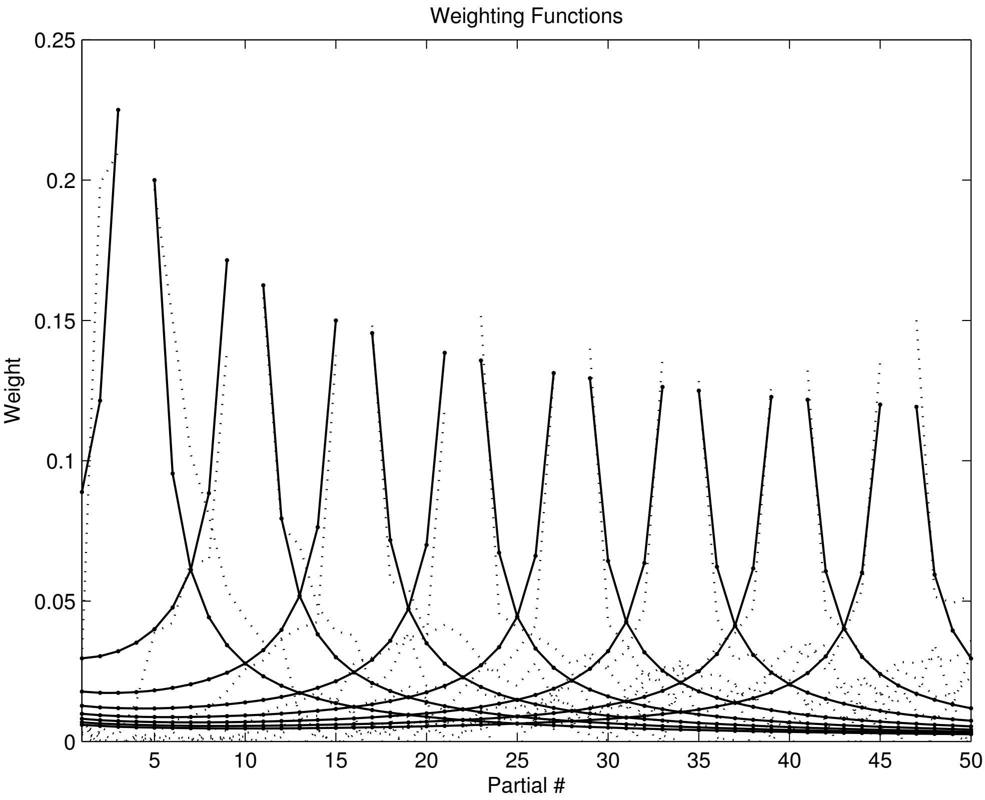

Figure 3.12: Weighting functions (dotted lines) and the modelled weighting functions (solid lines with dots). For clarity, only every 6th weighting function is illustrated.

to Signal Ratios lie around −35 dB, indicating that using linear combinations of neighbouring harmonics is a good predictor of harmonic magnitude tracks. When the weighting functions are generalised, the errors increase the Noise to Signal Ratio to approximately −17 dB. This is still a reasonable figure, considering that the weights are now generalised over all the musical instruments. Note that these Noise to Signal Ratios are indicative only of the prediction errors and should not be related to perceptual Noise to Signal Ratios. The performance of the modelled weighting functions given by Equation 3.21 are very similar in performance with only a very slight increase in errors in the first few harmonics. This slight increase is probably due to the fact that the similarity between the first few harmonics is more variable for different instruments. This leads to the weighting functions for the first few harmonics to be not as smooth as the latter harmonics, which therefore contribute to the modelling error increase. Finally the errors of a reference model using only a single adjacent harmonic as the model magnitude track are shown. This approach as employed in [125] and [127] does not perform as accurately as the other methods.

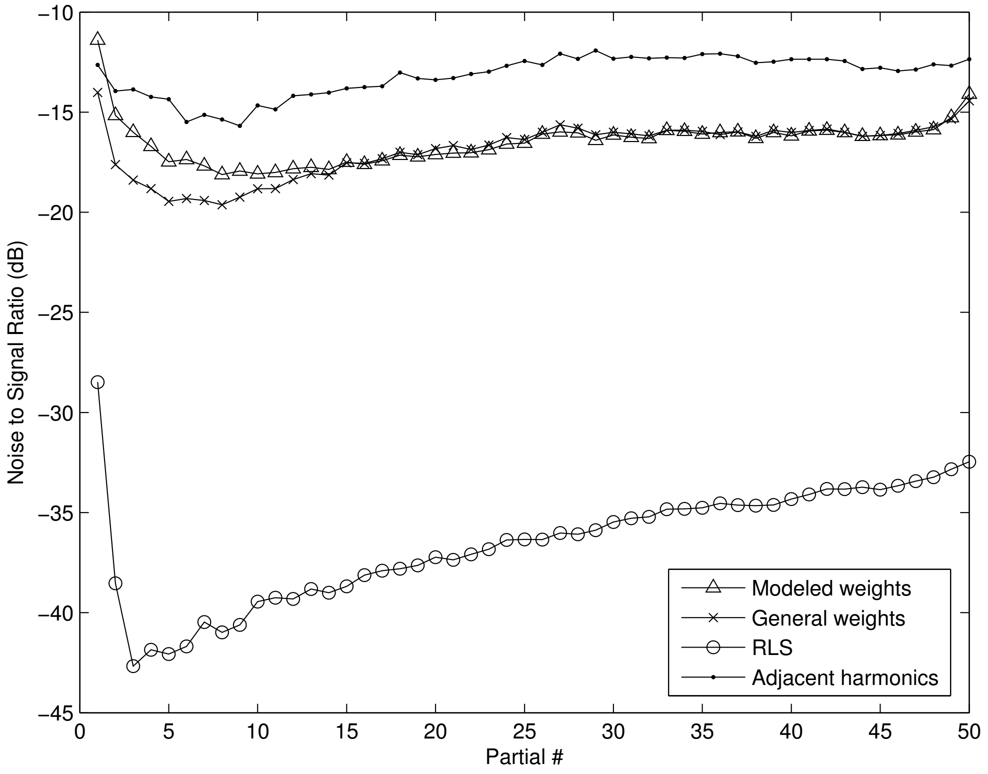

Figure 3.13: Noise to Signal Ratio comparing the prediction error of regularised least squares, general weighting functions, modelled weighting functions and adjacent harmonics.

3.3.4 Harmonic Track Identification

The effectiveness of the model magnitude tracks for harmonic track identification were then tested in simulated polyphonic mixtures. 2000 sound mixtures for polyphonies of 2, 3, 4, 5 and 6, were created by randomly summing together instrument samples from the Iowa instrument database [1]. Prior to the summation, the samples were individually analysed to determine the true harmonic magnitude tracks as described in Section 3.3.2. After the sources were summated, the short-time Fourier transform of the mixture was then analysed to determine the harmonic magnitude tracks found in the mixture. Harmonics that were in the frequency range $[\delta_{c1}, \delta_{c2}]$ Hz of each other were considered ambiguous and were selected for the experiment.

Harmonics less than δ_{c1} from another harmonic were considered overlapping, and harmonics further than δ_{c2} were assumed to be far enough to be adequately resolved by harmonicity constraints in a source separation system. Based on the STFT parameters, $\delta_{c1} = 30$ Hz, and $\delta_{c2} = 100$ Hz were used for the evaluation. Model magnitude tracks were then calculated for each ambiguous harmonic in the mixture using only unambiguous harmonics weighted by the appropriate weighting function as described in Equation 3.21. Each ambiguous harmonic was then tested against model harmonic tracks using the similarity measure of Equation 3.18. The ambiguous harmonic was then classified to belong to the source whose model magnitude tracks achieved the highest similarity score. This result was then compared to the ground truth found in the pre-summation analysis and thereby classified as correctly identified or incorrectly identified.

2000 mixtures were calculated for each polyphony and the results are shown in Figure 3.14. This is compared to a reference using the closest unambiguous adjacent harmonic as the model magnitude track [127]. The harmonic identification scores using the weighting functions were all higher than the reference adjacent harmonic scores. Using the model magnitude tracks built from the weighting functions results in consistently better performance even as the polyphony increases because the estimate is based on a number of neighbouring harmonics. As the polyphony increases, the identification scores decrease, and this is due to the fact that there are a fewer number of unambiguous harmonics from which to build reliable estimates. The more instruments there are, the more harmonics there are in the mixture, so the more likely a harmonic will have another close by. The results found in Section 3.3.2 highlight that harmonic magnitude track similarity is highest around the directly neighbouring harmonics, with similarity decreasing quite rapidly thereafter. Thus if there are more ambiguous harmonics, forming reliable harmonic estimates becomes more difficult. The results of Figure 3.14 clearly show that forming estimates from a number of neighbouring harmonics using the modelled weights provides consistently better performance than using a single unambiguous neighbouring harmonic, regardless of the polyphony.

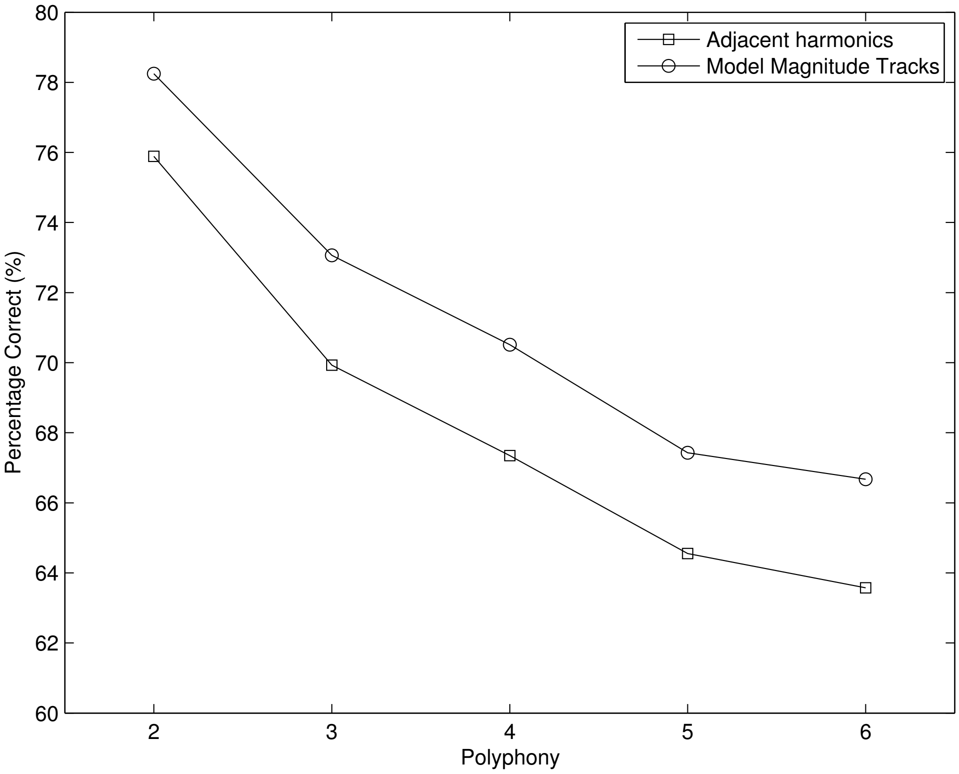

Figure 3.14: Partial identification scores averaged over 2000 samples using model magnitude tracks and the reference adjacent harmonics.

3.3.5 Conclusion

The harmonic magnitude tracks of sources are highly correlated, and in our investigation of 3000 musical instrument samples, it was found that the highest similarity occurs consistently between adjacent harmonics. Using this information, the relationships between harmonics tracks were used to predict model magnitude tracks using a linear combination of neighbouring harmonics. Least squares optimised weighting functions were found for each of the 3000 samples and were averaged to give a general set of weighting functions. These functions were then modelled and shown to perform comparably. The modelled weighting functions were then used to construct model magnitude tracks that improved the identification of ambiguous partials in polyphonic mixtures compared to using adjacent partials.

Chapter 4

Synthesis of Separated Sources

In a sinusoidal model based source separation architecture, various data parameters relating to each source are estimated. Magnitude tracks describe the amplitudes of the temporal envelopes of the source harmonics, while frequency tracks model the frequency location of each of the harmonics over time. However while these parametric representations are a summary of the separated sources, a complete source separation requires each source to be synthesised from the sinusoidal parametrisation.

Synthesis from the sinusoidal parametrisation is typically accomplished using either sinusoidal synthesis (Section 4.1.1), or binary masking of the short-time Fourier transform (Section 4.1.2). However, both of these methods employ an open-loop architecture where sources are typically synthesised independently. In Section 4.2, a multiple source synthesis method is presented, engaging the strengths of both sinusoidal synthesis and binary masking for the estimation of the magnitude spectra, while the phase spectra of the sources is iteratively estimated using a novel algorithm which aims to minimise the difference between the superposition of the source estimates and the mixture.

4.1 Source Synthesis

4.1.1 Sinusoidal Synthesis

One of the most intuitive ways of synthesising sources from a sinusoidal parametrisation, is to reconstruct each of the sinusoid tracks from the frequency $\hat{F}_s^h$, magnitude

$\hat{M}_s^h$ and phase $\hat{P}_s^h$ information (where h is the harmonic index and s is the source index). In most sinusoidal synthesis algorithms, each of the parameters $\left(\hat{F}_s^h, \hat{M}_s^h, \hat{P}_s^h\right)$ is interpolated over time to ensure that there are no discontinuities at the frame boundaries. The instantaneous amplitudes of the sinusoids at each frame are typically linearly interpolated, so that the amplitude of a sinusoid at time sample n into the rth frame is given by

$$\hat{a}_s^h(n) = \hat{M}_s^h(r) + \frac{\left(\hat{M}_s^h(r) - \hat{M}_s^h(r-1)\right)}{H} n \tag{4.1}$$

where H is the hop size of the STFT, $n = 0, 1, 2, ..., H-1$, for the hth harmonic of source s.

The phases are also interpolated, but as frequency and phase are related (frequency is the phase derivative, see Section 3.2.1), there are four variables involved in the interpolation: $\hat{F}_s^h(r), \hat{M}_s^h(r), \hat{F}_s^h(r+1), \hat{M}_s^h(r+1)$. This requires at least three degrees of freedom, which is obtained using a cubic polynomial as the interpolation function:

$$\hat{\phi}_s^h[n] = \zeta + \gamma n + \alpha n^2 + \beta n^3 \tag{4.2}$$

whose solution is described extensively in [81, 100].

Once the parameters have been interpolated, a bank of oscillators are then used to re-synthesise the sound:

$$\hat{x}_s[n] = \sum_{h=1}^{H} \hat{a}_s^h[n] cos\left(\hat{\phi}_s^h[n]\right) \tag{4.3}$$

where H is the total number of harmonics for source s.

A computationally efficient alternative to the bank of oscillators is presented in [102]. The synthesis method begins by using the estimated parameters $\left(\hat{F}_s^h, \hat{M}_s^h, \hat{P}_s^h\right)$ to represent each frame r in the complex frequency domain. Given the estimated frequency $\hat{F}_s^h(r)$ and magnitude $\hat{M}_s^h(r)$ track information, the estimated magnitude spectrum $\left|\hat{X}_{s,r}(k)\right|$ is obtained by translating the magnitude spectrum of the analysis window to be centred at frequency $\hat{F}_s^h(r)$ and scaled to have a magnitude of $\hat{M}_s^h(r)$.

The phase spectrum $\angle\hat{X}_{s,r}(k)$ is obtained using the phase estimates $\hat{P}_s^h$ and the complete spectrum estimate $\hat{X}_{s,r}$ for the rth frame is given by

$$\hat{X}_{s,r} = \left|\hat{X}_{s,r}\right| exp\left(j\angle\hat{X}_{s,r}\right) \tag{4.4}$$

The inverse FFT is then calculated for each frame of the spectrum and the frames are then combined using the overlap-add method. Any number of sinusoids can therefore be synthesised in a computationally efficient manner by simply placing the magnitude spectrum of the analysis window at the appropriate locations and computing an inverse FFT.

4.1.2 Binary Masking

Another re-synthesis method used particularly in source separation, is *binary masking* [129, 34, 127]. Binary masking uses the estimated parameters $\left(\hat{F}_s^h, \hat{M}_s^h, \hat{P}_s^h\right)$ to derive masks of the STFT for each source, based on the assumption that for pitched sounds, the energy is concentrated around the harmonic locations. The binary masks (with values of either 0 or 1), define the regions in the STFT that belong to particular source and the re-synthesis of separated sources is obtained by multiplying the binary mask for a source with the STFT coefficients, and taking the inverse STFT. Figure 4.1 is an example of a binary mask for a saxophone playing the note G#3.

Binary masking is essentially spectral filtering. The frequency location of the harmonics of a source are defined by $\hat{F}_s^h$, and the binary mask of a source is a conglomerate of time-varying narrow band filters that are positioned to filter each of the harmonics (see Figure 4.2). The width of the narrow band filters are dependant on the STFT parameters and particularly the analysis window, as this defines the sinusoid's main lobe width. The width of each filter must be larger than the main lobe in order to capture all of the harmonic information, but it must also be narrow enough to exclude other surrounding harmonics.

In [33], the separation quality of the sinusoidal synthesis method and the binary masking method was compared. The study found that the sinusoidal model performed better at separating synthesised sinusoids in additive white noise, partic-

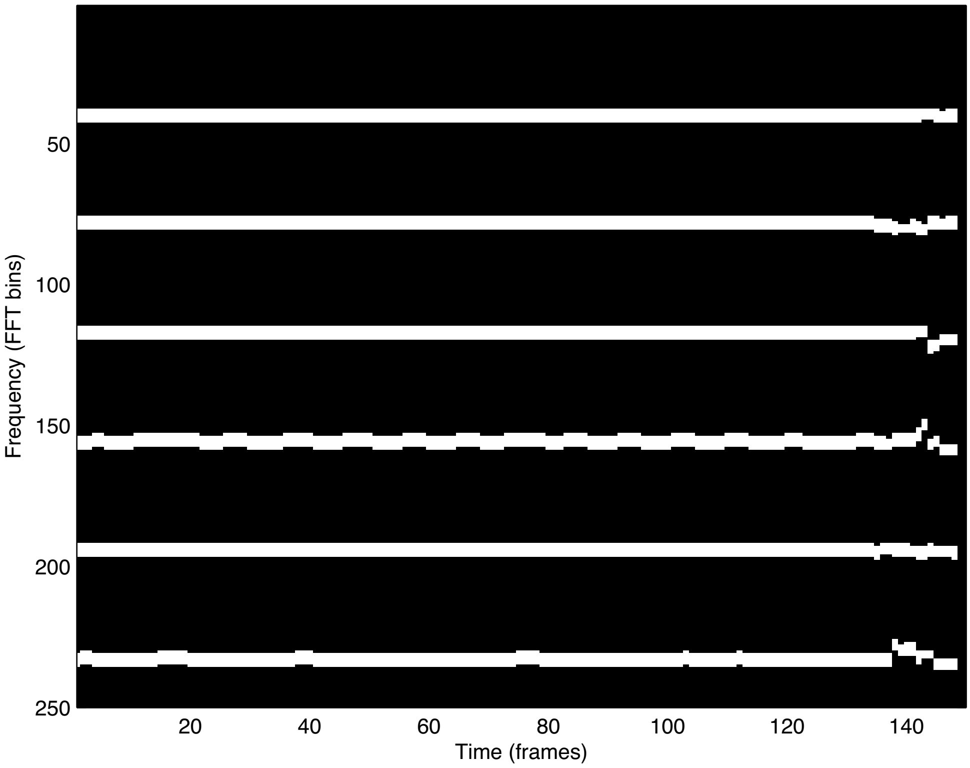

Figure 4.1: Binary mask for a saxophone playing a G#3. Regions belonging to the source are shaded white.

ularly at lower signal-to-noise ratios (SNR). However when using real monophonic musical instrument samples, the binary masking method performed consistently better, especially when using longer DFT lengths. Sinusoidal synthesis is a more flexible synthesis method, which is useful for altering the pitch of a sound or time-stretching sounds because of its compact parametrisation. Binary masking on the other hand, captures more of the surrounding nuances of the harmonics, resulting in better separation when there is less interference.

4.2 Multiple Source Synthesis

When sources are mixed together, the overlapping regions contain magnitude and phase information from multiple sources. Merely setting a binary mask to include or exclude a particular time-frequency element does not separate the mix of information in the overlapping regions. This is particularly problematic in musical mixtures and

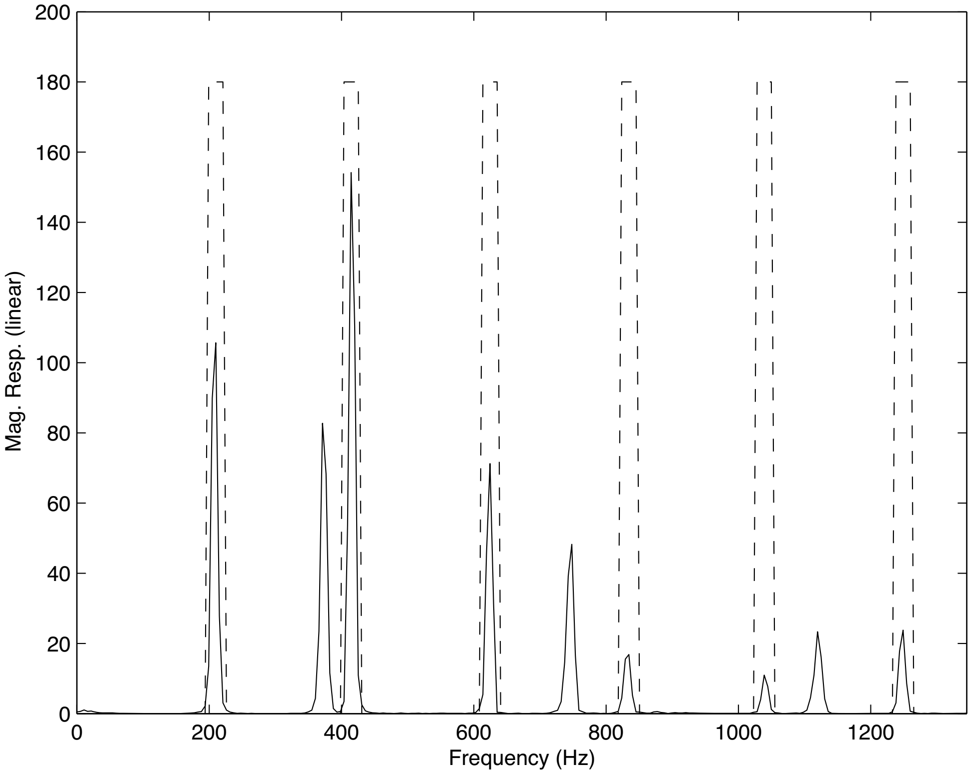

Figure 4.2: Binary mask as a spectral filter. The solid line is the magnitude response of a saxophone playing a G#3, and the dotted line denotes the binary mask.

the mixed information results in audible distortions due to source interference, in the separated sources.

In this section, we propose a method of synthesising separated sources from a mixture given their parametric representations. The synthesis problem is approached with the dual intent of both capturing source information, while minimising extraneous source interference in regions of overlap. First, the magnitude spectrum of each source is estimated by combining the flexibility of the sinusoidal synthesis method with the benefits of binary masking as detailed in Section 4.1. The time-domain estimates of each source are then calculated in a closed-loop architecture, using Multiple Input Spectrogram Inversion (MISI) - a novel algorithm that iteratively estimates the phase of each source given the individual magnitude spectra (Section 4.2.4). The performance of the MISI algorithm is evaluated, and is shown to provide significant improvements in the estimation of time-domain sources, given sufficiently accurate source magnitude spectra.

4.2.1 Multiple Source Magnitude Estimation

In the comparison between sinusoidal synthesis and binary masking described in Section 4.1.2, Every [33] showed that when using real monophonic musical instrument samples, the binary masking method performed consistently better. This is a reasonable result for monophonic samples, because while harmonic partials of real instruments are predominantly sinusoidal in nature, they often contain additional salient spectral information around the sinusoids. However in polyphonic mixtures, creating source binary masks becomes problematic when harmonics overlap with each other. Magnitude and phase information is corrupted by overlapping harmonics and the information becomes unreliable for source estimation. To evade these issues, a combination of both binary masking [129] and sinusoidal synthesis (using spectral synthesis [102]) was used to estimate the source magnitude spectra.

Given the parametrised harmonic tracks of the sources, it is necessary to define the regions where tracks overlap with other tracks. An awareness of this prior to estimating the magnitude spectra of the sources, facilitates better separation by allowing different strategies to be employed for both overlapping and non-overlapping regions. Overlapping harmonics can be classified by observing their proximity to other source harmonics, and an overlapping vector O_j^h for each frame r, was defined such that,

$$O_j^h[r] = \begin{cases} 1, & if \left| F_j^h[r] - F_s^g[r] \right| < \delta_{overlap} \\ 0, & otherwise \end{cases} \tag{4.5}$$

where F_j^h is the frequency track corresponding to the hth harmonic of source j and $s \neq j$, and $\delta_{overlap}$ is the overlap distance in Hz dependent on the width of the primary spectral lobe related to the analysis window. Thus overlapping regions within a frequency proximity $\delta_{overlap}$ were classified as overlapping, while all other regions were considered as non-overlapping.

For non-overlapping tracks, the magnitude spectrum $\left|\hat{X}_s\right|$ of source s was estimated using binary masking [129]. Fixed-width, time-varying, narrow band masks centred around the harmonic frequencies were used to create the source binary masks. The width of the narrow band masks was set to be large enough to capture the harmonic's main lobe, but narrow enough to exclude neighbouring harmonics as shown in Figure 4.3. The binary masks for each source B_s were then used to obtain magnitude estimates for the non-overlapping tracks using element-wise multiplication of the magnitude spectrum of the mixture $|Y[k]|$, where

$$\left|\hat{X}_{s,r}[k]\right| = B_{s,r}[k]\,|Y[k]|$$

for each frame r, and every DFT frequency bin k.

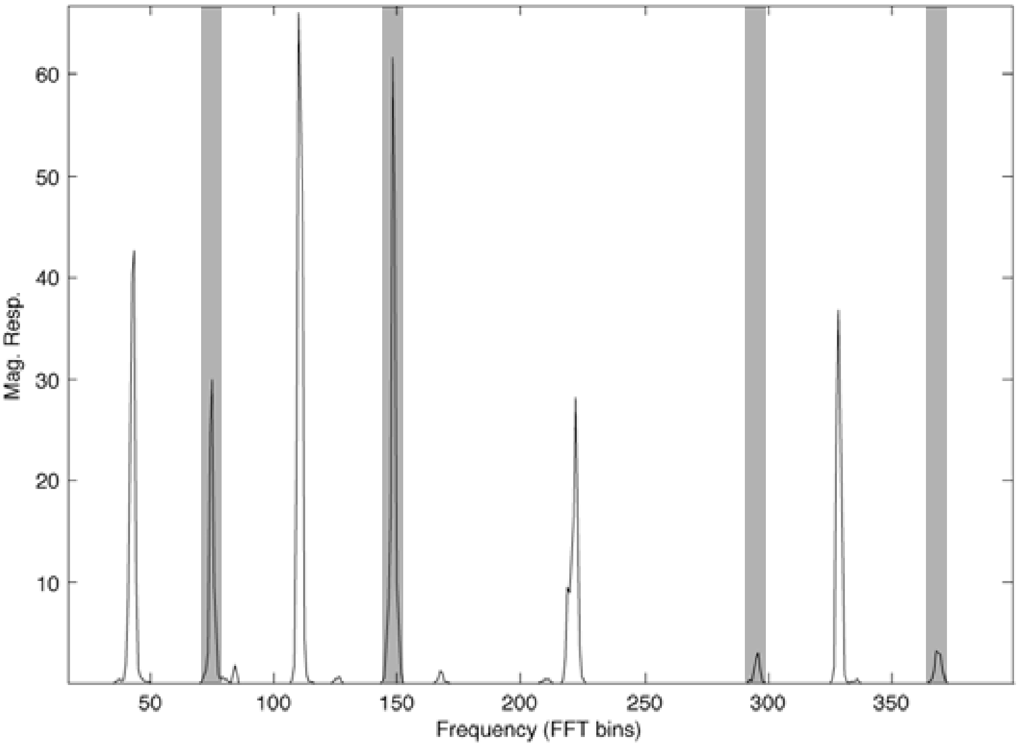

Figure 4.3: Binary mask for the non-overlapping harmonics of a bassoon playing a G4, in a mixture of 3 sources. The solid line is the magnitude response and the grey shaded regions indicate the binary masked regions.

For overlapping tracks, the spectral synthesis [102] method was used to define the magnitude spectra. The magnitude spectrum of overlapping harmonics was estimated at each frame by translating the magnitude spectrum of the analysis window (in this case, the Hanning window) to be centred at the harmonic frequency and scaling it to its estimated amplitude. Figure 4.4 illustrates the estimation of an overlapping harmonic using this method.

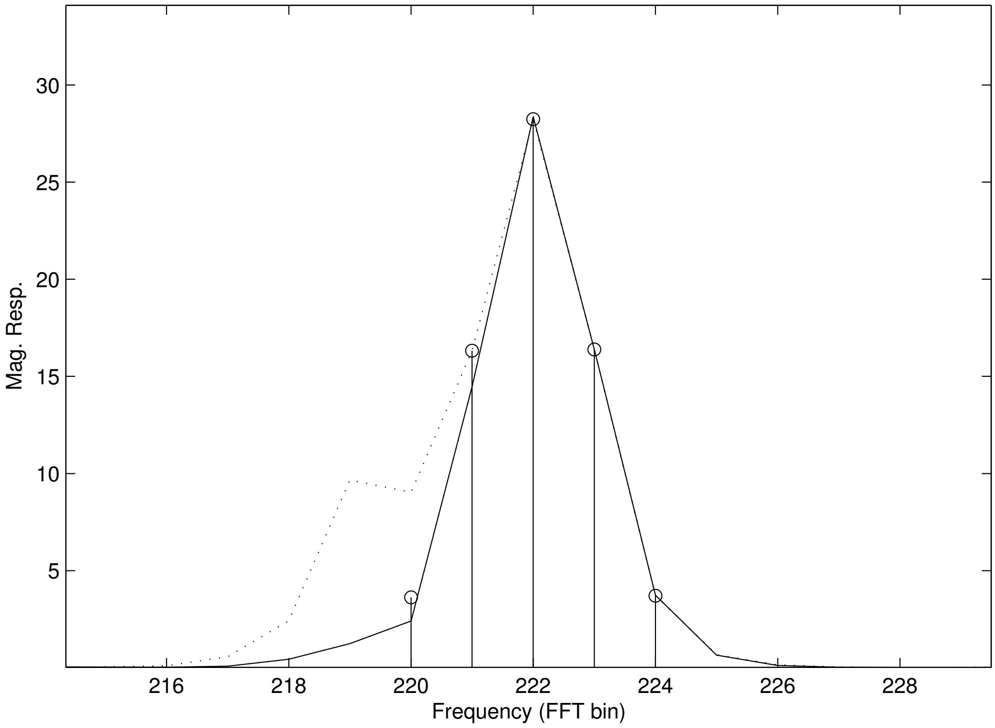

Figure 4.4: Magnitude estimation of an overlapping harmonic (in close proximity to another harmonic) by translating and scaling the magnitude spectrum of a Hanning analysis window. Solid line is magnitude response of mixture signal, dotted line is magnitude response of source and estimated magnitude response is given by the stem plot.

The combination of both binary masking and sinusoidal synthesis facilitated the estimation of the magnitude spectra of the sources. In regions where the STFT described a unique source, binary masking was used to capture the harmonic, and in regions of overlap where binary masking would be unreliable, sinusoidal synthesis was used to estimate the magnitude response of the harmonic. The source magnitude estimates were therefore formed by engaging the most effective method for each harmonic, resulting in accurate magnitude estimation of the sources, which will be shown in Chapter 6.

4.2.2 Multiple Input Spectrogram Inversion

Source synthesis from parametric information is typically achieved using either sinusoidal synthesis [81, 102], binary masking [129], spectral filtering [34], or a hybrid approach [130]. With the exception of [81], the other methods begin by estimating the magnitude spectra of the sources, followed by the binary masking of the phase spectra of the mixture in regions where the magnitude spectra is salient. The reconstruction of the sources is then formed by inverting the spectral magnitude and phase information into the time-domain. Most of the methods primarily focus on the rigorous estimation of the source magnitude spectra, while merely masking the mixed phase spectrum to obtain the phase spectra of the sources. While using the mixed phase spectrum may suffice in applications where the sources in the mixture are uncorrelated, the concordant nature of musical mixtures limits the effectiveness of using the mixed phase spectrum.

Each of the methods also employ an open-loop estimation method for generating time domain source estimates. In the typical scenario where all the sources are to be estimated in a mixture, we propose that a closed-loop estimation system can improve the quality of synthesis by successively refining the source estimates. In the following sections, we describe spectrogram inversion (Section 4.2.3) as a precursor to a novel, closed-loop approach to source synthesis (Section 4.2.4). The synthesis algorithm separates sources by iteratively minimising the aggregate error of the sources, constraining the minimisation to a set of estimated parameters. The performance of the algorithm is then evaluated with respect to harmonic music mixtures and is shown to provide significant improvements over binary masking of the mixed phase spectrum.

4.2.3 Spectrogram Inversion

Spectrogram inversion [44, 131] algorithms are a subset of reconstruction algorithms which aim to estimate signals by recovering the missing phase information through an iterative process that converges towards a signal with a magnitude-constrained spectrum. A comprehensive overview of iterative reconstruction algorithms can be found in [7].

Consider the discrete-time signal $x[n]$ whose short-time Fourier transform (STFT) is given by

$$X_r(\theta) = \sum_{n=-\infty}^{\infty} x[n]w[n-rH]e^{-j\theta n} \tag{4.6}$$

where w is the analysis window, H is a positive integer denoting the hop size, θ is the digital frequency, and r is the frame index of the STFT. From the STFT, the short-time Fourier transform magnitude (STFTM) is defined as $|X_r(\theta)|$.

The estimation of a signal by spectrogram inversion is obtained by minimising the mean square error (MSE) function given by

$$D_M[x[n], x'[n]] = \sum_{r=-\infty}^{\infty} \frac{1}{2\pi} \int_{\theta=-\pi}^{\pi} \left[|X_r(\theta)| - \left|X_r^{'}(\theta)\right|\right]^2 d\theta \tag{4.7}$$

with respect to $x[n]$, where $|X_r(\theta)|$ is the STFTM of the original signal and $\left|X_r^{'}(\theta)\right|$ is the STFTM of the estimated signal.

The spectrogram inversion algorithm described by Griffin and Lim [44] minimises Equation 4.7 by iterating between the frequency and time domains. At each iteration i, the following function is used to update the estimate

$$x^{i+1}[n] = \frac{\sum_{r=-\infty}^{\infty} w[n-rS]\frac{1}{2\pi}\int_{\theta=-\pi}^{\pi} \bar{X}_r^i(\theta)e^{j\theta n}d\theta}{\sum_{r=-\infty}^{\infty} w^2(n-rS)} \tag{4.8}$$

where $\bar{X}_r^i e^{j\theta n}$ is the STFT of $x^i[n]$ with the magnitude constraint

$$\bar{X}_r^i(\theta) = |X_r(\theta)| \frac{X_r^i(\theta)}{|X_r^i(\theta)|} \tag{4.9}$$

The algorithm obtains the $i+1$th estimate $x^{i+1}[n]$ by replacing the magnitude of $X_r^i(\theta)$ with the given magnitude $|X_r(\theta)|$, taking the inverse Fourier transform and then overlap-adding the frames to generate the time-domain signal. By enforcing the constraint on the magnitude over time, the time evolution of the spectral phase is also constrained, resulting in signal estimates that not only decrease the MSE with each iteration, but which also minimise distracting audible phase artifacts.

Figure 4.5 illustrates the iterative phase estimation process in segments α, β,

γ. At the ith iteration, the complex frequency estimate at the rth frame is given by $X_r^i(\theta)$. In this example, there is a unit magnitude constraint (i.e. $|X_r(\theta)| = 1$), and this is applied using Equation 4.9, which in turn yields $\bar{X}_r^i(\theta)$. In the next iteration $i+1$, the time domain estimate $x^{i+1}[n]$ is calculated using Equation 4.8 and since this equation minimises Equation 4.7, the Fourier transform of $x^{i+1}[n]$ (denoted by $X_r^{i+1}(\theta)$), improves on the previous phase estimate. The magnitude is again constrained and the process iterated until a satisfactory error level is obtained.

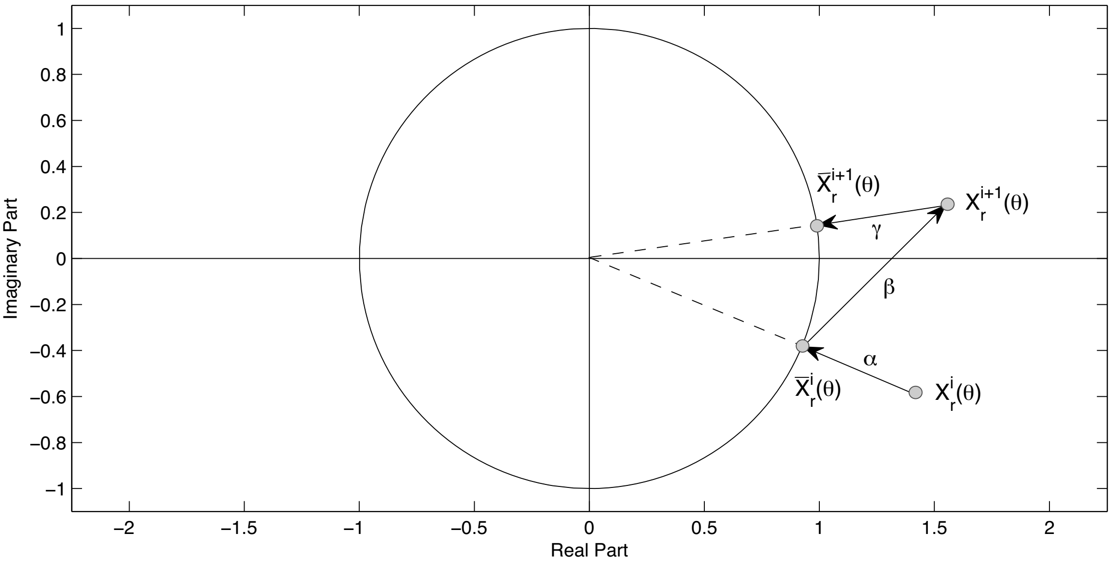

Figure 4.5: A phase estimation iteration. Segment α: Magnitude constraint at ith iteration. Segment β: Phase refinement. Segment γ: Magnitude constraint at the $i+1$th iteration.

4.2.4 Multiple Input Spectrogram Inversion Algorithm

In a typical source separation scenario, the observed acoustic waveform $y[n]$ is a superposition of source signals $y[n] = \sum_{j=1}^{J} x_j[n]$, where $x_j[n]$ is the jth source signal and J is the number of sources. Given the magnitude spectra estimations of the sources from Section 4.2.1, we present the multiple input spectrogram inversion (MISI) algorithm, which iteratively estimates the time-domain source signals $x_j[n]$ in a mixture $y[n]$ given the corresponding magnitude spectra of the source signals (Figure 4.6).

If the STFTM of the source is known, then using Equations 4.8 and 4.9 it is

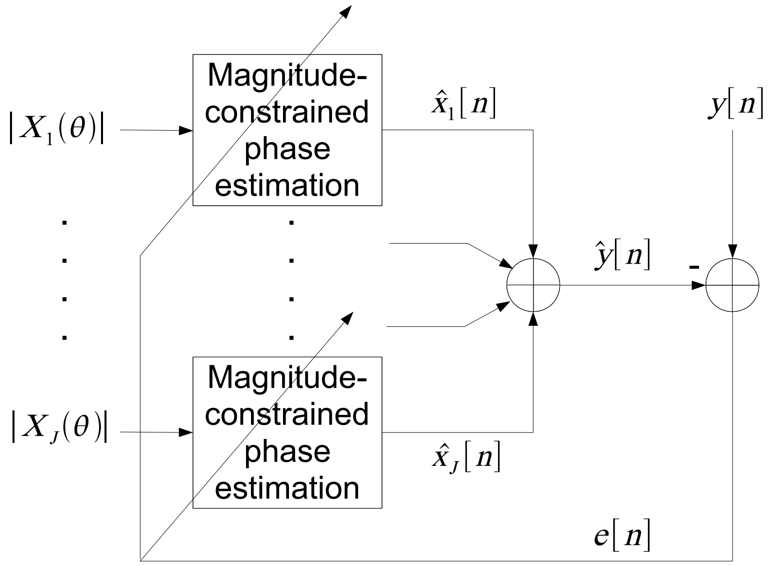

Figure 4.6: Overview of the MISI algorithm. Given the magnitude estimates of each source, the phase responses are estimated and the signal transformed into the time domain. The source estimates are then subtracted from the original mixture and the error is used to refine the phase estimates.

possible to calculate an estimate for the $i+1$th iteration of the jth source signal,

$$\hat{x}_j^{i+1}[n] = \frac{\sum_{r=-\infty}^{\infty} w[n-rH]\frac{1}{2\pi}\int_{\theta=-\pi}^{\pi} \bar{X}_{j,r}^{i}(\theta)e^{j\theta n}d\theta}{\sum_{r=-\infty}^{\infty} w^2[n-rH]} \tag{4.10}$$

$$\bar{X}_{j,r}^{i}(\theta) = |X_{j,r}(\theta)| \frac{\bar{\bar{X}}_{j,r}^{i}(\theta)}{\left|\bar{\bar{X}}_{j,r}^{i}(\theta)\right|} \tag{4.11}$$

where $\bar{\bar{X}}_j^i$ is obtained by taking the STFT of

$$\bar{\bar{x}}_j^i[n] = \hat{x}_j^{i-1}[n] + \frac{e^i[n]}{J} \tag{4.12}$$

$$e^i[n] = y[n] - \sum_{j=1}^{J} \hat{x}_j^i[n] \tag{4.13}$$

The algorithm is initialised with each of the initial source estimates set to $\hat{x}_j^0[n] = y[n]$ and the error set to $e^0[n] = 0$. The algorithm then constrains each estimate with the known STFTM $|X_{j,r}(\theta)|$ and then calculates an inverse Fourier transform, appropriately overlap-adding each frame. The MISI algorithm then accounts for the total error $e^i[n]$ between $y[n]$ and the superposition of the estimate source signals $\hat{y}[n]$, adding a scaled version of the error to each of the source estimates before the next iteration. Scaling the error and adding it back to the source estimates, aids in the minimisation of the total error.

As long as the sum of the scaled errors equals the total error, the energy of the summed sources will be conserved at each iteration. In the MISI algorithm, this is achieved by simply dividing the error equally among the sources, so that the scaled error assigned to each source is given by $\frac{e^i[n]}{J}$, where J is the total number of sources. At each iteration, the scaled errors that are fed back, combined with the magnitude constraints, ensure that the phase estimates of each source approach the true phases. The source magnitude constraints shape the errors so that the phase is re-estimated in the frequency regions where the error is large, and the phase is retained in the regions where the error is low.

Thus the MISI algorithm essentially computes a spectrogram inversion for each source with the additional constraint of minimising the error between the mixture and the superposition of the estimated sources at each iteration.

4.2.5 MISI Performance

In this section, we evaluate the separation performance of the MISI algorithm in a variety of polyphony and additive noise scenarios.

We evaluated the source separation capabilities of the MISI algorithm using mixtures of musical instrument samples from the University of Iowa musical instrument samples database [1]. Single note mixtures consisting of 2, 3, 4, 5, and 6 instruments, were randomly created by summing time-domain waveforms of different instruments The individual STFTMs of each source were then calculated using a Hanning analysis window, with frame size $N = 4096$, step size $H = 1024$, and sampling frequency $f_s = 44.1$ kHz. 50 mixtures were created for each polyphony p. The root mean square error (RMSE) of the total errors at each iteration were then calculated and averaged over the 50 mixtures to obtain the results illustrated in Figure 4.7.

It can be observed from Figure 4.7 that the RMSE decreases with every iteration and this was observed to be true for every mixture tested. Furthermore, the indi-

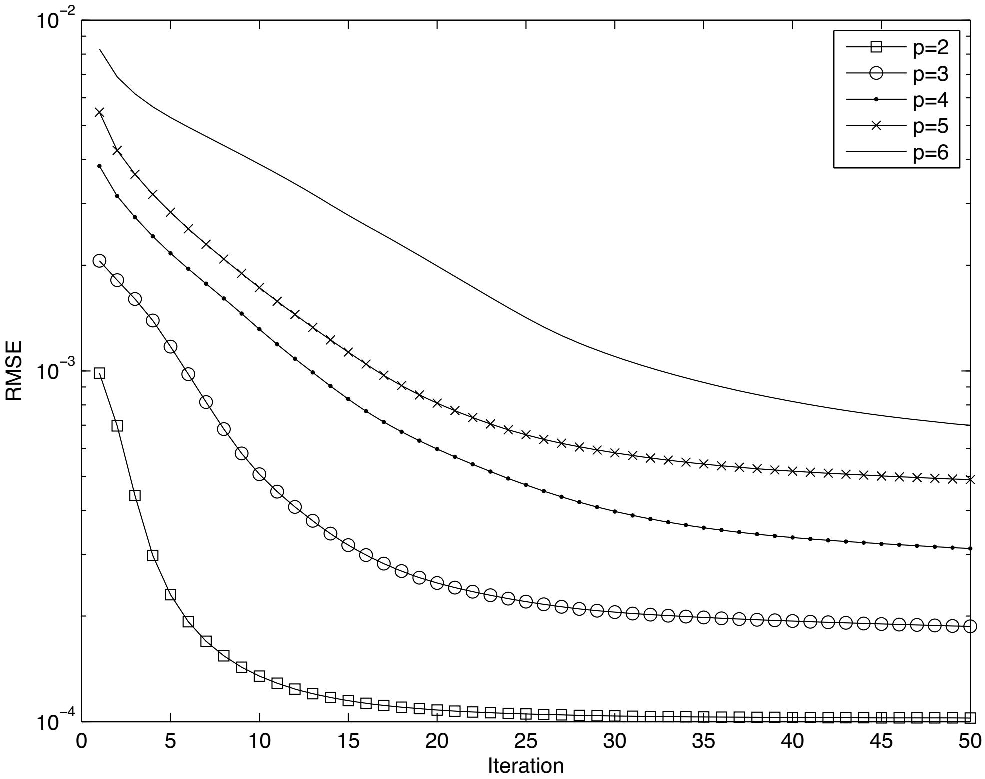

Figure 4.7: Average RMSE over 50 iterations for instrument mixtures with polyphony $p = 2, 3, 4, 5, 6$. 50 mixtures were used for each polyphony p.

vidual source estimates in each mixture were also found to monotonically decrease. As the polyphony increases, the RMSE also increases and takes longer to converge. However, informal observations (under conditions as those described in Section 5.1) revealed that even after 10 iterations and $p = 6$, the distortions were found to be minimal and often imperceptible, with no audibly distracting artifacts. Thus given the true source STFTMs, the MISI algorithm provides good separation results, with each of the tested source RMS errors decreasing with every iteration.

In a real separation system where the true STFTMs are unknown, estimates must be made based on the knowledge of the source signals. To simulate the effect of estimation errors, various levels of additive white Gaussian noise were added to each of the source signals such that $\tilde{x}_j[n] = x_j[n] + \gamma\eta[n]$ where $\eta[n]$ is white Gaussian noise and γ is a scaling factor to raise the noise to the desired signal-to-noise ratio (SNR) level. The STFTMs of $\tilde{x}_j[n]$ were then calculated and with $p = 4$, the performance of the MISI algorithm was evaluated for SNR values ranging from -30 dB to 30 dB over 50 mixtures. Figure 4.8 illustrates the effect of using erroneous STFTMs in the MISI algorithm. The introduction of additive noise still results in the asymptotic behaviour of the RMSE for all mixtures, albeit being quicker and at a much higher level as the SNR decreases. The quality of the source STFTM estimates therefore limit the quality of time-domain estimation that can be achieved with the MISI algorithm.

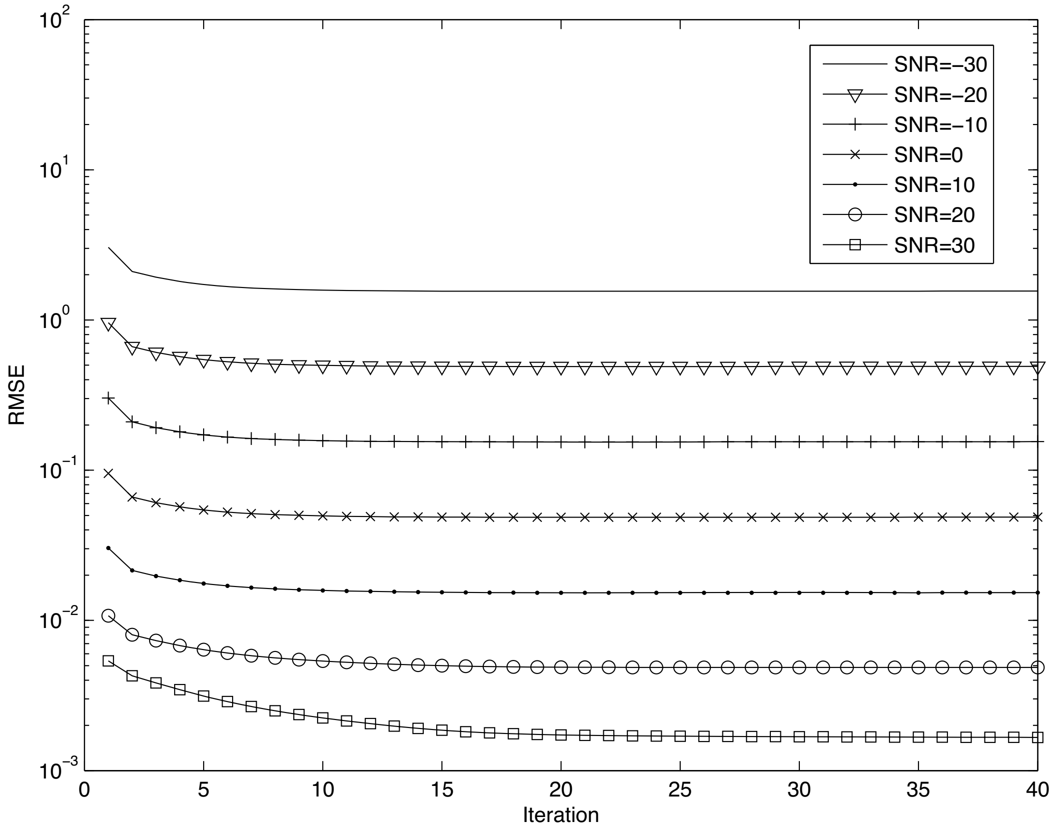

Figure 4.8: Average RMSE over 50 iterations with additive white Gaussian noise for SNRs of -30, -20, -10, 0, 10, 20 and 30 dB.

While the total errors are asymptotic in nature even in the presence of noise, this does not imply that the individual estimated sources $\hat{x}_j[n]$ converge to the original individual sources $x_j[n]$. At a certain SNR, the individual source estimations would be sufficiently erroneous that each iteration of MISI would in fact increase the RMSE with respect to the original individual sources. Figure 4.9 illustrates the individual source SNRs obtained using phase binary masking[1] (PBM) and various iterations

---

[1] Note that phase binary masking (PBM) is equivalent to one iteration of the MISI algorithm.

of MISI for the same set of mixtures. Thus for mixtures of 4 sources, $\hat{x}_j[n]$ only converges to $x_j[n]$ when all of the source estimate STFTM SNRs are larger than 15 dB. When the STFTM estimates are below this, PBM provides better estimates.

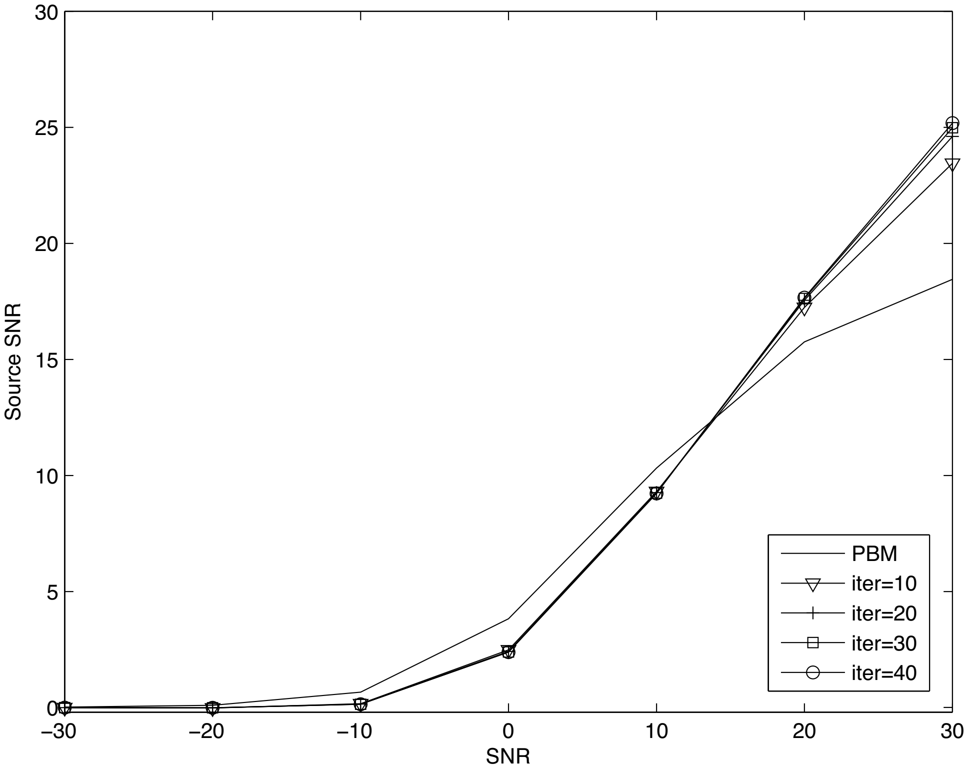

Figure 4.9: Comparison of MISI and PBM over various source SNRs.

Figure 4.10 illustrates the performance of the MISI algorithm as a function of note F0s grouped into musical octaves. The separated source SNRs were averaged over 100 random mixtures containing 4 sources, using the true STFTMs of the sources over 20 iterations. Notes were partitioned into musical octaves where A2, A3, A4, A5, A6, A7 are the notes which correspond to the F0's: 110 Hz, 220 Hz, 440 Hz, 880 Hz, 1760 Hz, 3520 Hz. The results clearly illustrate a consistent improvement in the SNR of the MISI algorithm over PBM, of approximately 13 dB over all musical octaves. There is also a clear upward trend in the separation quality of the sources as the F0s of the notes increase in frequency. This can be attributed to the increased number of harmonic components of sources that have lower F0s, which in turn increases the probability of overlap with other sources.

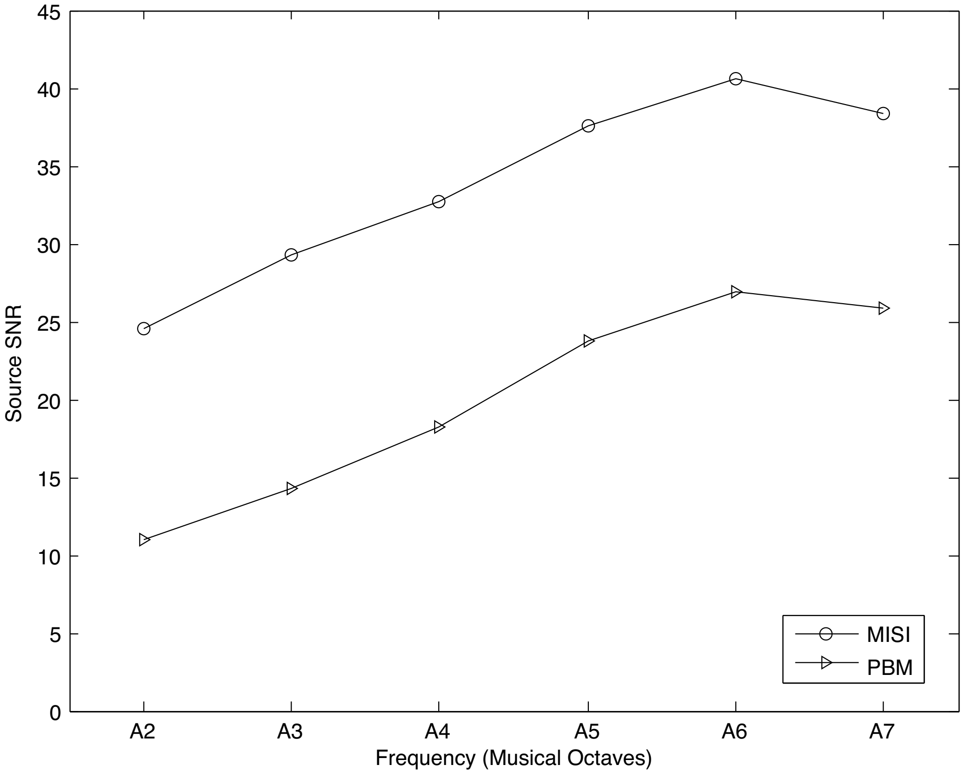

Figure 4.10: Comparison of MISI and PBM as a function of musical octaves.

4.2.6 Conclusions

The Multiple Input Spectrogram Inversion algorithm highlights the merits of a closed-loop synthesis algorithm, which contrasts with the predominant use of open-loop algorithms for synthesising sources in model-based separation systems. The iterative phase estimation of the MISI algorithm produces significant gains, minimising the synthesis errors with respect to the time-domain mixture, given sufficiently accurate source magnitude spectra.

Chapter 5

Perceptual sensitivity of Timbre: Towards an objective distortion metric

In the quest to separate sources from a mixture of sounds, it is necessary to evaluate systems using a distortion metric that is perceptually relevant. Ultimately the quality of any sound is subject to human perception, making it appropriate to define the quality of a separation system with reference to human perception.

When considering the quality of musical sources, it is necessary to consider the perceptual sensitivity to the attribute known as *timbre*. Timbre is the perceptual attribute that enables the ability to distinguish between instruments and has been defined more formally as "that attribute of auditory sensation in terms of which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different" [8]. This negative definition of timbre characterises the development of timbre research over the years as researchers have investigated the development of an adequate model of timbre.

The classical view of musical timbre proposed by Helmholtz [51], assumed that timbre was exclusively associated with the spectral energy distribution of a tone. Through ingenious experimentation, Helmholtz deduced that the tone quality of musical sounds were largely governed by the relative amplitudes of the harmonics, with phase having a minimal effect. His investigations of the timbre space led to verbal descriptors of variations of the spectrum. Complex tones consisting only of the odd harmonics were *hollow*, while the predominance of the fundamental produced a *full* tone. Tones with moderately loud lower harmonics up to the 6th were classed as musical and *rich*, while tones with loud harmonics beyond the 6th were sharp and *rough*. However Helmholtz's experimentation only considered the steady-state portion of periodic waveforms, neglecting the temporal variations which exist in real instrument sounds. In the research that followed on from Helmholtz, there was an increasing trend towards the view that timbre was multifaceted [13, 99] and dependent on temporal aspects such as the attack transient [74].

Unlike pitch, which is primarily dependent on a tone's fundamental frequency, and loudness, which depends on tone intensity, the consensus is now that timbre is a multidimensional property of sound [90]. In addition to the verbal scales, like that used by Helmholtz, the dimensions of timbre have been explored using multidimensional scaling (MDS) techniques [90, 42, 43, 17]. The majority of these studies have found that the perception of timbre is dominated by spectral energy and temporal variation.

The experiments presented in the subsequent sections explore various salient attributes of timbre with the objective of understanding the timbre space and developing a robust objective timbre distortion measure. In the first experiment (Section 5.1), we investigate the sensitivity to changes made to musical instrument spectral envelopes [50]. In Section 5.2, the spectral envelope space is explored using novel linear-logarithmic morphing techniques [47]. In the final experiment (Section 5.3), stimuli are decomposed into harmonic and noise components, to investigate the discrimination thresholds for changes to the noise component [49].

5.1 Experiment 1: Spectral Envelope Sensitivity

Timbre research has found the spectral envelope to be a salient attribute [80][89][43][17]. In musical acoustics, the spectral envelope can be described in the frequency domain as an interpolation between the amplitudes of the sinusoidal components of a signal

[80, 89]. Sufficient modification of the spectral envelope of an instrument produces a change in perception of that instrument's timbre, and in some cases significant modification can lead to the instrument sounding similar to a different instrument. Grey's [43] work in developing perceptual spaces of timbre using multidimensional scaling led to the identification of the spectral energy distribution being one of the important dimensions of timbre. More recently, McAdam's et al. [80] have identified the spectral envelope shape as being the most salient parameter in timbre discrimination when performing various simplifications to instrument spectro-temporal parameters. Caclin et al. [17] also verified the spectrum's importance in their confirmatory study using synthetic tones.

A thorough understanding of timbre therefore requires knowledge of how much spectral deviation is required before there is a perceptible change in timbre. The primary objectives of this section are to analyse the discrimination thresholds of spectral change for various instruments and observe the sensitivity to change as a function of centre frequency and bandwidth. We have chosen to study three instruments (trumpet, clarinet and viola) which represent the brass, woodwind and string families. While previous studies have analysed sensitivity to musical instrument spectral envelopes [90, 53, 86], none of them have investigated the sensitivity as a function of centre frequency and bandwidth. Other studies have studied sensitivity as a function of frequency but not in the context of musical instruments. Due to the complex nature of musical instrument signals, the results of such studies are very difficult to translate into a musical instrument context.

Early studies by Plomp [90] investigated perceptual sensitivity to spectral change for static musical instrument and vowel spectra and found that spectral differences were good predictors of differences in timbre. Horner et al. [53] extended this work by observing instrument discrimination for random alterations to time-varying instrument spectra. The spectra of instruments were modified by various error levels (8%, 16%, 24%, 32% and 48%) by randomly altering the amplitudes of individual sinusoids. They observed that discrimination was very good for 32% and 48% error levels, moderate for the 16% and 24% error levels and poor for the 8% error levels. However the spectral modifications were performed randomly over time and frequency and did not account for the varying sensitivities that may be apparent as a function of frequency.

Similar work has been done in the field of speech processing particularly for the purposes of speech coding. Paliwal [86] divided speech signals into frames of approximately 20 ms and observed that the average spectral distortion difference limen for perceptual indistinguishabilty is 1 dB, ensuring that no frames have average spectral distortions greater than 4 dB and less than 2% of the frames have average spectral distortions between 2-4 dB. These results have been used extensively in the design of vector quantisers for speech coders. However, once again, these observations are based on the entire spectrum and do not reveal sensitivity as a function of frequency.

Auditory profile analysis is a field concerned with in observations on the discrimination thresholds of spectrally modified sounds. Green [41] performed an analysis of discrimination thresholds for 21 component complexes; however, like most auditory profile analysis experiments, the stimuli considered were sums of sinusoids that were spectrally flat and with log-spaced frequencies. Thus, the stimuli were very different from realistic musical instrument spectra which are harmonically spaced and non-uniform. The results are therefore difficult to extrapolate to a musical instrument context.

In the present study, we aim to investigate the discrimination thresholds for changes to musical instrument spectral envelopes. Previous studies have often assumed that spectral envelope sensitivity is unchanged as a function of frequency [43, 53], however we hypothesise that there will be variations in the discrimination thresholds for modifications made as a function of centre frequency and bandwidth. The experimental results are compared to a number of spectral distortion measures and then are discussed with reference to other experimental findings as well as predictions from a psychoacoustic model.

5.1.1 Experimental Method

In order to investigate the sensitivity to the spectral envelope, we endeavoured to keep all other physical parameters constant. These included fundamental frequency, level and duration - the details of which are described in the following section. With the intent of understanding how sensitivity varies as a function of centre frequency and bandwidth, each stimulus was modified by attenuating a band of frequencies by various amounts. Subjective tests were conducted to determine discrimination thresholds for different instruments.

5.1.1.1 Stimuli

Three musical instrument sounds were selected for analysis. Samples of trumpet, clarinet, and viola taken from a University of Iowa website [1] were used. The samples were chosen for their representation of three different instrument families - brass, woodwind and string. The sounds were recorded using 16 bits, and a 44100 Hz sampling rate, and each sound was played at a pitch of E^{b}4, corresponding to a fundamental frequency of approximately 311.1 Hz - a frequency within the normal playing range of these instruments and commonly used in timbre experiments for this reason [80, 53]. Average spectra of the three sounds are illustrated in Figure 5.1. The duration of each sound was standardised to 1.5 seconds using a 100 msec half-Hanning window to taper the offsets. The onsets of each sample were left unmodified. The level of each sound was adjusted by a gain factor such that five independent subjects perceived them to be of equal loudness.

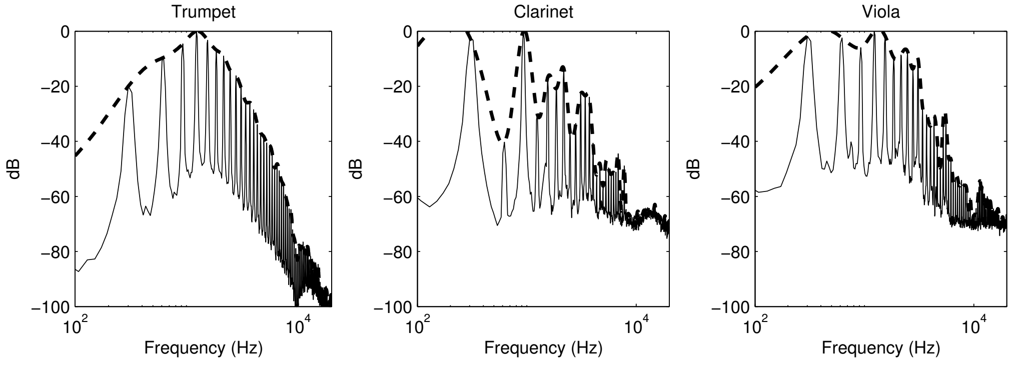

Figure 5.1: Average musical instrument spectra (solid lines). Dashed lines illustrate the spectral envelope calculated using the SEEVOC method. Averages were taken over 32 frames, each of 2048 samples.

The three sounds were then each modified such that various bands across the frequency spectrum were attenuated by various amounts. The stimuli presentation was controlled using MATLAB on an Intel PC with an RME Multiface sound card. Each of the stimuli was presented monaurally at an average level of approximately 65 dB SPL through Beyerdynamic DT770pro headphones in a sound-insulated (Acoustic Systems) anechoic chamber.

5.1.1.2 Stimuli modification

The system illustrated in Figure 5.2 was employed to make the relevant modifications. As the stimuli are time-varying in nature, time-invariant filters were employed to preserve the time resolution. Each stimulus was passed through a zero-phase band-pass filter and the output of the filter was then attenuated and subtracted from the original stimulus. Using 14 zero-phase filters of differing centre frequencies and bandwidths, we compiled a set of stimuli where the output was the original signal with a certain frequency band attenuated. Note that the modified stimuli were not equalised for loudness as this would produce more audible changes than not equalising the loudness.

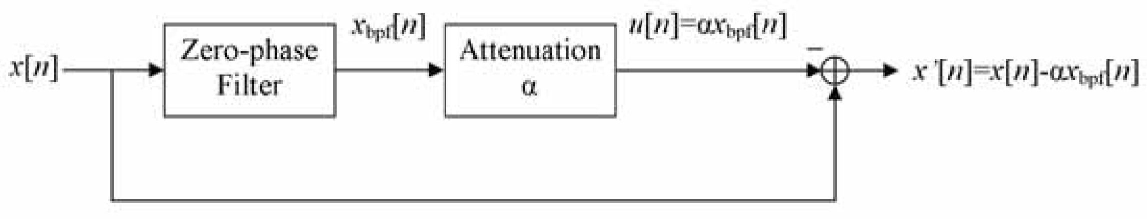

Figure 5.2: System used for stimuli modification.

The zero-phase filters were designed by taking 256-tap linear-phase band-pass filters (designed by the window method based on a Hamming window) and advancing the output signal by the group delays of the filters. Since the human auditory system has an non-linear frequency resolution [84] which can be approximated by a logarithmic-like function, 14 logarithmically-spaced filters were used as illustrated in Figure 5.3. More low frequency filters with narrower bandwidths were selected to analyse the lower frequencies with higher resolution in similar fashion to the auditory system. The filters are labelled 1 to 14.

As an example, the magnitude response of filter 4, which has a bandwidth of 5512.5 Hz and a centre frequency of 8268.75 Hz, is plotted in Figure 5.4. An atten-

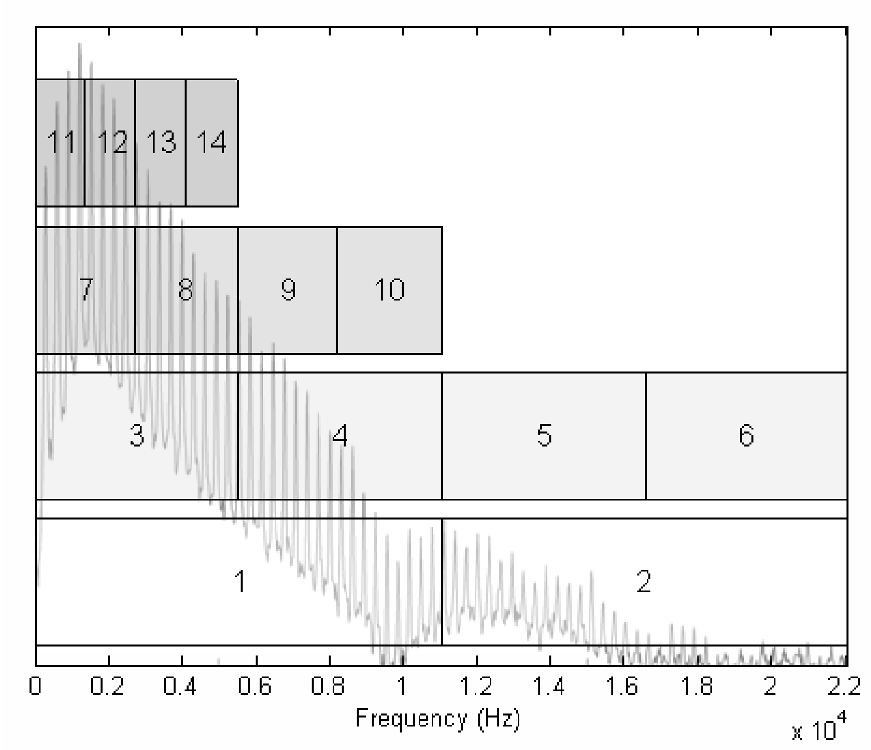

Figure 5.3: Bandwidths of the 14 zero-phase filters with trumpet spectrum overlaid. (The rectangular boxes only indicate bandwidth and should not be associated with the y axis which indicates the spectral magnitude of the trumpet).

uated viola spectrum using this filter is illustrated in Figure 5.5. Preliminary tests using ERB gammatone filters similar to those in [107] resulted in measurements being dependent on harmonic content rather than the spectral envelope. The ERB gammatone filters had bandwidths that were too fine for spectral envelope analysis and thus wider logarithmically spaced bandwidth filters were used to observe the effects of spectral envelope modification.

5.1.1.3 Participants

Five listeners aged between 20 to 26 years participated in the experiment. Four participants were male and one was female and all were tested and found to have normal hearing. Three of the participants had musical training with experience ranging between 5-10 years.

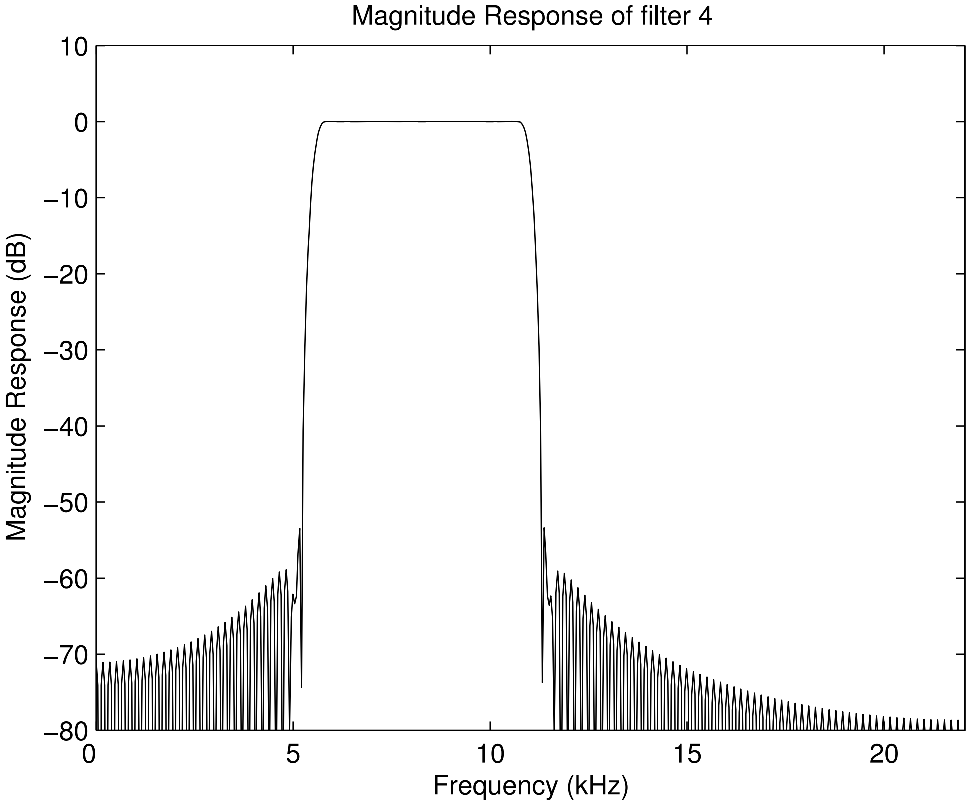

Figure 5.4: Magnitude response of filter 4.

5.1.1.4 Procedure

A two-alternative forced-choice (2AFC) Reference AB, 1-up 2-down paradigm [73] was used for all our experimentation. For each trial, the participant heard three sounds: the reference sound (original, unfiltered) followed by two other sounds - one of which was filtered and the other which was the same as the reference. The order of presentation of the two latter sounds were independently randomised for each trial and 300 msec silence periods separated the presentation of each sound. For each trial, the participant was prompted with “Which sound has a different timbre to the reference?” and had to respond by clicking buttons marked A and B on the screen. Once a response was submitted, feedback was provided to the participant in the form of “Correct” or “Incorrect”.

The first trial presented for each centre frequency was always with the most attenuation and the attenuation was incrementally decreased to include more of the contents of the band. The attenuation step sizes changed from 4 dB to 2 dB and

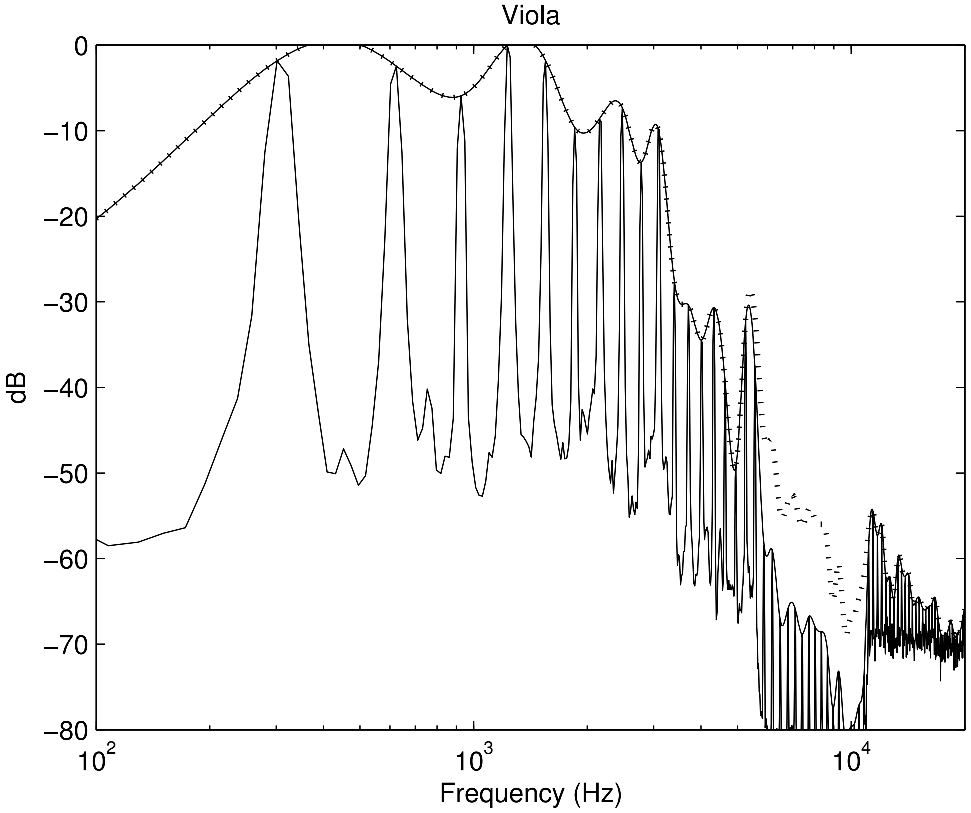

Figure 5.5: Attenuation of a band of frequencies by filter 4 on the viola. The spectrum and spectral envelope of the attenuated signal are illustrated by the solid line while the dashed line illustrates the original unattenuated viola spectrum.

finally to 0.5 dB. The last 3 reversals were averaged to estimate the discrimination threshold. Listeners were trained for 15 mins to familiarise themselves with the task prior to the experiment. Thresholds for the 14 filtered bands were recorded in a single 50 minute block per instrument.

5.1.2 Results

The results from the experiment were analysed in four different ways. The first was a measurement of sensitivity which analysed the individual Band Attenuations (BA). Following that, we computed two different distortion measures as employed in [53] and [86] to compare the data to previous studies. Finally in Section 5.1.3, the results are discussed with what is predicted by a psychoacoustic difference limen model.

5.1.2.1 Band Attenuation

If a listener is more sensitive to a change in a signal parameter, then a smaller change of that parameter is needed to hear the effect of the change. We define $x[n]$ to be the original stimulus, $x'[n]$ to be the modified stimulus (as illustrated in Figure 5.2), $x'^*[n]$ to be the just-noticeable modified stimulus and α^* to be the just-noticeable attenuation that produces $x'^*[n]$. If only a small change in the energy of a band is required before it is detected, sensitivity is considered to be high for that band and the ratio of the band signal energy to the distortion will be large. If we define the Band Attenuation (BA) to be the original energy of the band, divided by the minimum difference required to observe a change in that band, then sensitivity is simply proportional to the BA and can be written simply as a function of the attenuation α (Equation 5.1). Thus the BA can be used as a measure of sensitivity.

$$\begin{aligned} BA &= 10\log_{10}\left(\frac{\sum x_{bpf}[n]^2}{\sum u[n]^2}\right) &(5.1)\\ &= 20\log_{10}\left(\frac{1}{a^*}\right)(dB) &(5.2)\end{aligned}$$

where α^* is the just-noticeable attenuation of a particular band (in linear units) and $u[n] = x[n] - x'^*[n] = \alpha^* x_{bpf}[n]$.

The BA results are shown in Figure 5.6, clearly indicating that there are obvious differences in the sensitivities for different bandwidths and centre frequencies. Qualitatively, it can be observed that smaller changes at lower frequencies consistently trigger a perceptual change in timbre compared to changes at higher frequencies. The lower frequencies are therefore more sensitive than higher frequencies.

Another important observation is that the bands that include the first few harmonics also tend to set the upper bound for sensitivity (filters 1, 3, 7, 11), for all other bands have lower sensitivity than these. This implies that the maximum sensitivity can be estimated from the sensitivities of the lower frequencies and no other region of the spectral envelope will have higher sensitivity.

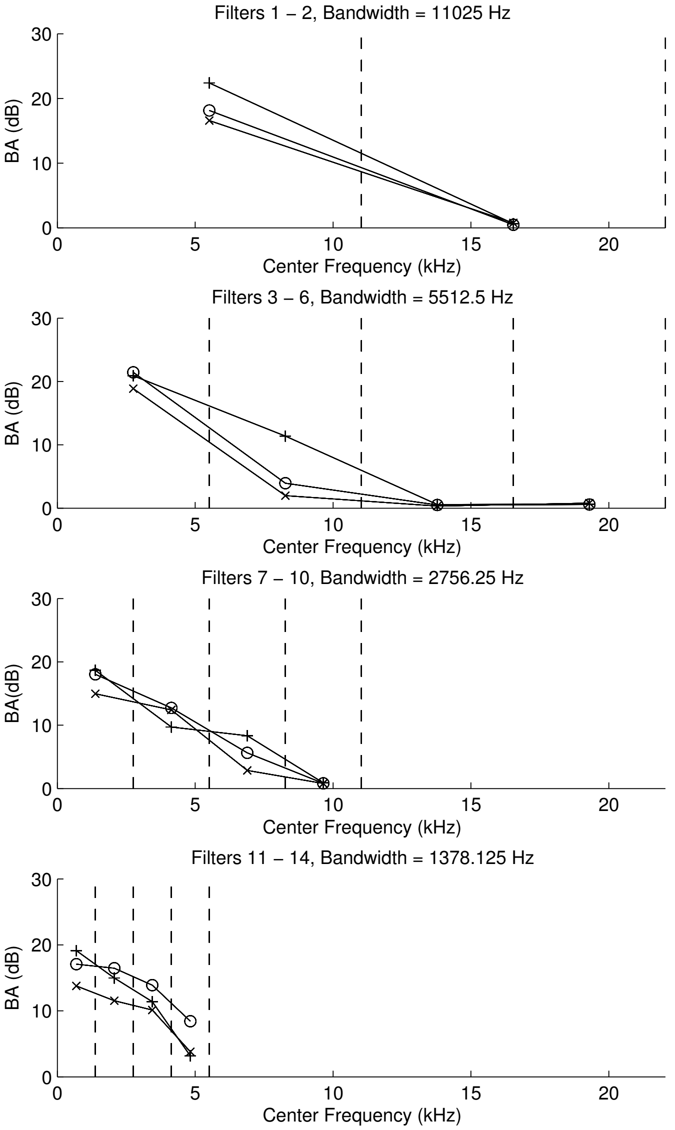

Figure 5.6: BA plots for the Trumpet (o), Clarinet (+), Viola (x) positioned at the centre frequencies of filters 1 to 14. Dashed lines indicate the filter bandwidths.

5.1.2.2 Distortion measures

The results can also be expressed in terms of the amount of modification required to perceive a change. Here we compare our results to two other studies from [53] and [86].

Error Level

In a study by Horner et. al. [53], the spectra were altered randomly and the spectral deviation was measured by observing average error levels as a percentage of the deviation from the original. Alteration of the harmonic spectra was performed by multiplying each amplitude of the kth harmonic at time t, $A_k(t)$, with a randomly selected scalar r_k:

$$A'_k(t) = r_k A_k(t) \tag{5.3}$$

The scalars $\{r_k\}$ were selected uniformly in the range $[1-2\epsilon, 1+2\epsilon]$, where ϵ denotes the error level.

The calculation of the relative error level whether in the frequency domain or time domain is analogous, so for simplicity, the error levels (EL) were calculated in the time domain using:

$$EL = \sqrt{\frac{\sum_{n=1}^{N} u[n]^2}{\sum_{n=1}^{N} x[n]^2}} \times 100\% \tag{5.4}$$

$$= \alpha^* \sqrt{\frac{\sum_{n=1}^{N} x_{bpf}[n]^2}{\sum_{n=1}^{N} x[n]^2}} \times 100\% \tag{5.5}$$

where $x[n]$ is the original stimulus, $x'^*[n]$ is the just noticeable modified stimulus, α^* is the just-noticeable attenuation of a particular band, $u[n] = x[n] - x'^*[n] = \alpha^* x_{bpf}[n]$, n is the sample number and N is the total number of samples. Thus for fixed bandwidth and centre frequency, EL varies linearly with α.

The percentage errors in Figure 5.7, correspond to 70.7% discrimination on the psychometric curve [73]. These results indicate that the discrimination for the bands with low centre frequencies (containing most of the signal) is around 13%. This agrees with the results in [53] where it was found that discrimination was approximately 16% at the 75% discrimination level. While the analyses for the bands with low centre frequencies concur with [53], the additional analysis for various bandwidths in this study reveals that error levels vary for different bandwidths and centre frequencies. Bands with higher centre frequencies and with wider bandwidths can only undergo smaller changes relative to the entire signal before discrimination.

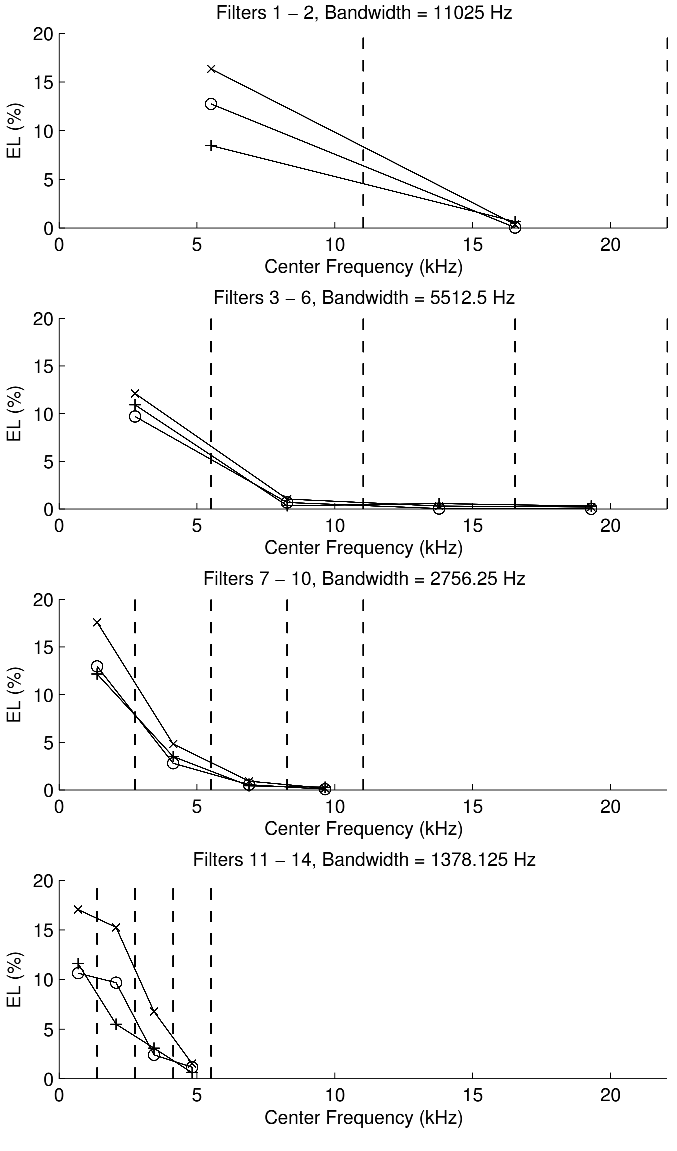

Figure 5.7: Error Level plots for the Trumpet (o), Clarinet (+), Viola (x) positioned at the centre frequencies of filters 1 to 14. Dashed lines indicate the filter bandwidths.

Spectral Distortion

The spectral envelope analysis by Paliwal [86] employed a spectral distortion error metric to define the maximum error before the altered spectrum could be distinguished from the original. The spectral envelopes for each frame were calculated by the SEEVOC method [89] which proceeds as follows: A periodic signal is divided into frames and the DFT of each frame is calculated. Using the $F0$ of the signal, the harmonic peaks are located and a smooth curve is fitted through them. In this analysis, 1024 frequency points for the DFT were chosen and cubic interpolation was chosen to join the harmonic peaks.

The spectral distortion (defined for a given frame as the root-mean-square difference between the original log-power spectral envelope and the modified log-power spectral envelope), is averaged over a large number of frames to give the average spectral distortion:

$$SD = \frac{1}{N_k}\sum_{k=1}^{N_k}\sqrt{\frac{1}{M}\sum_{\omega=0}^{M-1}\left(s(\omega) - s'^{*}(\omega)\right)^2} \tag{5.6}$$

where N_k is the number of frames, k is the frame number, M is the number of frequency points, $s(\omega)$ is the original log-power spectral envelope, $s'^{*}(\omega)$ is the just noticeable modified log-power spectral envelope and ω is the DFT frequency number.

Figure 5.8 illustrates the results with respect to spectral distortion. The spectral distortion for these envelopes were then calculated by Equation 5.6 yielding a spectral distortion measure that was averaged over a number of frames. Interestingly, the results for the bands with lower centre frequencies concur with the 1 dB value of distortion found for the spectral transparency of speech [86]. The present analysis sheds further insight into the spectral distortions allowable for various bandwidth modifications. A significantly larger amount of spectral distortion of up to 17 dB is allowable before discrimination occurs for bands with higher centre frequency.

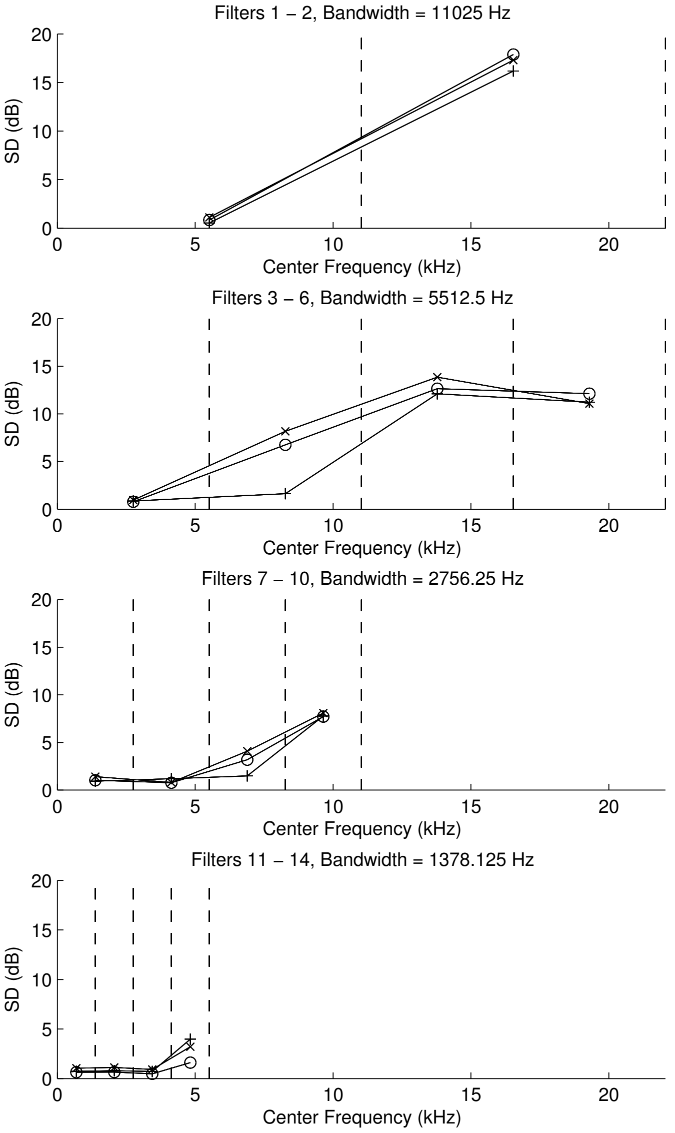

Figure 5.8: Spectral Distortion (SD) plots for the Trumpet (o), Clarinet (+), Viola (x) positioned at the centre frequencies of filters 1 to 14. Dashed lines indicate the filter bandwidths.

5.1.2.3 Difference limen model comparison

It would be of interest to compare the subjective experimental data to that predicted by a psychoacoustic difference limen model. In particular, we would like to see threshold levels as a function of both stimuli level, centre frequency and bandwidth. Thus, a simple frequency or level difference limen model would not suffice. A simultaneous auditory masking model, on the other hand, does provide approximate threshold levels for the complex stimuli such as those used in this study. Masking thresholds were calculated using [55] for each of the three original stimuli using overlapping frames of 512 samples. For each stimulus, the masking thresholds were then averaged over all the frames, and then for each band (see Figure 5.3) the average Signal-to-Masking Ratio (SMR) was calculated to represent the band's SMR. The SMR describes the relationship between the stimuli and the minimum distortion that is perceivable. A high SMR indicates that only a small deviation from the original signal can be tolerated, while a low SMR suggests the opposite. SMR can thus be viewed as an indication of sensitivity.

Figure 5.9 illustrates the average SMR for each of the instruments. The results clearly show that the bands with lower centre frequency are more sensitive to change than the bands with higher centre frequency and therefore agree with the BA results found in Figure 5.6. The results also show that the lower bands indeed dominate the sensitivity and the higher bands become increasingly more sensitive as the bandwidth narrows. However, the SMR model is not an extremely accurate predictor of sensitivity in the lower bands, for while the experimental findings suggest a more consistent sensitivity as the bandwidth narrows, the SMR model clearly suggests an increase in sensitivity as the bandwidth narrows.

5.1.3 Discussion and Conclusion

The results from the experiment highlight a number of important attributes about perceptual sensitivity to the spectral envelope. The BA plot (Figure 5.6) clearly shows that any assumption of sensitivity being equal over centre frequency and bandwidth is inaccurate. The spectral envelope's sensitivity to change varies considerably over centre frequency and bandwidth, and further studies that manipulate the spectral envelope of an instrument ought to consider such effects.

The experiment highlights that there are clear discrepancies between the amount of distortion tolerable over frequency, but also accentuates the importance of clarifying the reference for the measure of distortion. This can be seen by comparing the results from Figure 5.7 and Figure 5.8. What initially seems contradictory in

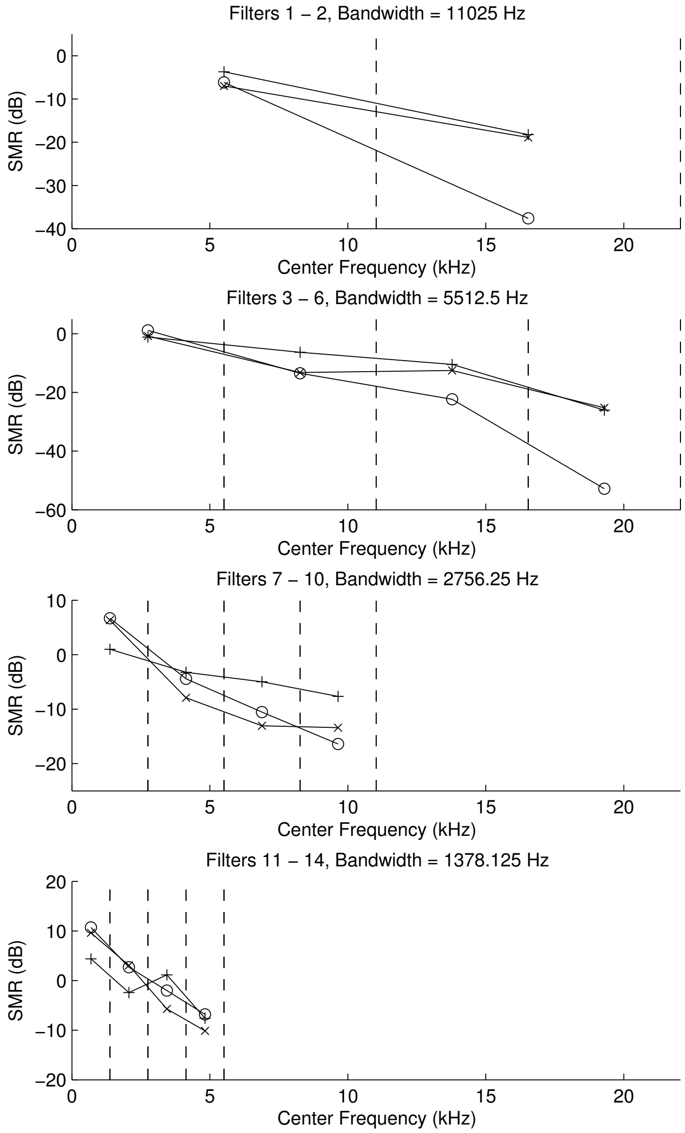

Figure 5.9: Signal-to-Masking Ratio (SMR) plots for the Trumpet (o), Clarinet (+), Viola (x) positioned at the centre frequencies of filters 1 to 14. Dashed lines indicate the filter bandwidths.

fact proves to be complementary. Figure 5.7 shows that the error required to discriminate changes for higher band decompositions is much smaller than lower band decompositions. However, this is relative to the entire signal energy. For musical instruments the higher frequencies generally have much lower amplitudes than the lower frequencies and thus their relative errors are smaller. Figure 5.8 on the other hand gives the spectral distortion in dB. Because the higher frequencies have around 40 dB less power than the lower frequencies, a greater level of distortion is required for discrimination of the higher bands.

The study in [53] sought to quantify how much spectral envelope modification could be made before a change in timbre was observed. The error level for 75% discrimination in [53] was approximately 16% and this result is similar to the 13% error level at 70.7% discrimination for bands with low centre frequencies found in this experiment. The spectral distortion threshold result of 1 dB found in [86] is a criterion that is frequently employed in the design of speech vector quantisers. Interestingly in the context of musical instruments, this 1 dB result also aligns well with the 1 dB spectral distortion threshold for bands containing the lower harmonics as calculated in this experiment. Thus, the results in this experiment agree with previous results for bands with low centre frequency, but shed further light into the nature of discriminability when considering change to only a certain bandwidth.

The comparison with a masking analysis model illustrates that our sensitivity measurements generally agreed with psychoacoustic masking theory. Despite some differences particularly in the bands with lower centre frequency, Figures 5.6 and 5.9 seem to have the same fundamental appearance and would therefore suggest that sensitivity to the spectral envelope can be crudely approximated using the average SMR value for the band in question. However, the experimental results suggest a more consistent sensitivity of the lower bands than the masking model infers.

In summary, distortion of different portions of musical instrument spectral envelopes using band attenuation with different bandwidths and centre frequencies results in different discrimination levels. This implies that sensitivity varies as a function of frequency and bandwidth. Sensitivity is maximum for the lower frequencies and decreases as the centre frequency moves higher. For bands with lower centre frequency, the sensitivity remains approximately the same while the bands with higher centre frequency consistently decrease in sensitivity. Thus, from a perceptual standpoint, sensitivity has an upper bound governed by the first few harmonics and our sensitivity does not improve when extending the bandwidth any higher. However, if changes are made only to the higher harmonics, then our sensitivity is decreased and reduces further as the bandwidth distorted is widened.

5.2 Experiment 2: Spectral Envelope Morphing

While the majority of musical timbre research has focused extensively on real musical instrument sounds, the timbre space can be explored in greater depth by considering sounds that lie in between real instrument sounds. Grey's [42] investigation of the timbre space was facilitated by the use of *morphed* or *interpolated* sounds and in this section, we propose a novel extension of morphing strategies that will aid in the development of timbre space models.

This study explores the timbre space by using a number of different linear-logarithmic morphing permutations of the spectral envelopes of a trumpet sound and a clarinet sound. The results of a two-alternative forced-choice experiment are compared with existing spectral envelope classification parametrisations and a psychoacoustic masking model, providing insight into considerations for developing models of the timbre space.

5.2.1 Experimental Method

5.2.1.1 Stimuli

Two musical instrument sounds, of a trumpet and a clarinet from [1], were selected for the experiment. The samples were recorded using 16 bits, and a 44100 Hz sampling rate, and each sound was played at a pitch of Eb4, corresponding to a fundamental frequency of approximately 311.1 Hz. The duration of each sound was standardised to 1.5 s using a 100 msec half-Hanning window to taper the offsets, while the onsets were left unmodified. The level of each sound was adjusted by a gain factor and presented monaurally at an average level of approximately 65 dB SPL through Beyerdynamic DT770pro headphones in a sound-insulated (Acoustic systems) anechoic chamber.

The spectral envelopes of each sample using the spectral envelope estimation vocode (SEEVOC) method [89], which proceeds as follows: A periodic signal is divided into frames and the DFT of each frame is calculated. Using the F0 of the signal, the harmonic peaks are located and a smooth curve is fitted through them.

In this analysis, 4096 frequency points for the DFT were chosen, with a step size of 1024 samples and cubic interpolation was chosen to join the harmonic peaks.

The trumpet and clarinet samples were selected for their contrasting spectral envelopes and their similar *residual*[1] spectra. The contrasting envelopes of the two instruments, illustrated in Figure 5.10, allow for the investigation of the prominent parameters of the spectral envelope which dominate timbre classification.

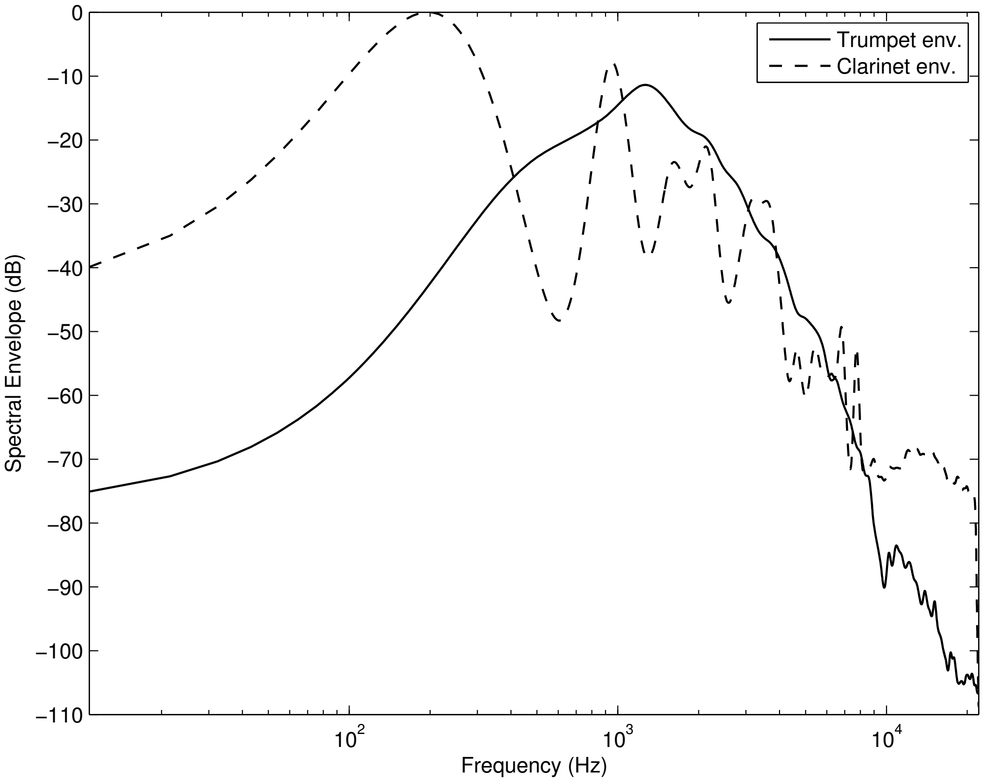

Figure 5.10: Average spectral envelopes of the trumpet and clarinet.

5.2.1.2 Morphing

Each morphed stimulus was created by dividing the magnitude spectrum by the spectral envelope and multiplying the result with the desired spectral envelope. Six sets of morphs from the trumpet to the clarinet were created, consisting of 30 morphs per set. Each set was assigned a cut-off frequency f_c, which denoted the boundary

---

[1] The *residual* is defined as the stochastic component of the signal obtained by subtracting the deterministic sinusoidal component (see Section 2.3.2).

at which different morphing strategies were employed. The envelope frequencies lower than f_c were morphed linearly (see the left-hand-side of Figure 5.11), while the envelope frequencies higher than f_c were morphed logarithmically (see the right-hand-side of Figure 5.11). The values of f_c for the 6 sets are given in Table 5.1.

Set #	1	2	3	4	5	6
f_c (Hz)	22050	11025	5512.5	2756.25	1378.125	689.0625

Table 5.1: f_c values for the 6 sets.

The variable cut-off frequencies and linear-logarithmic morphing strategy facilitates the exploration of the timbre space. The logarithmic morphing of the higher frequencies in particular, assists in revealing the salient aspects of the spectral envelope that relate to instrument classification.

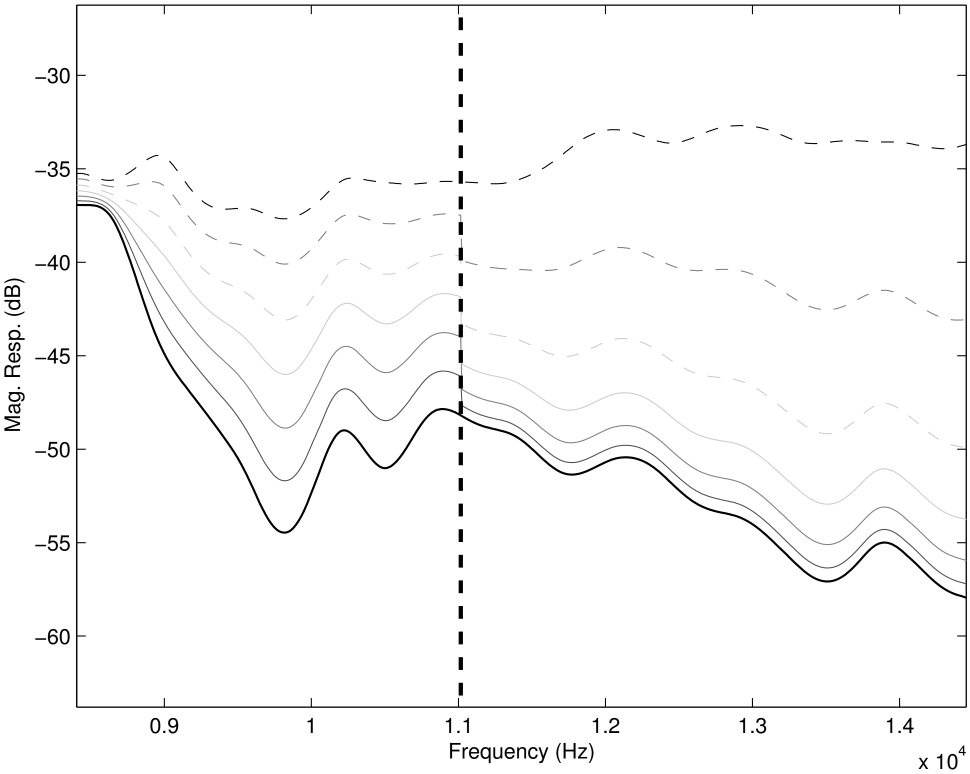

Figure 5.11: Linear and logarithmic morphing with $f_c = 11025$ Hz. 30 morphs were used per set, however only 5 morphs are illustrated for clarity.

5.2.1.3 Procedure

Six listeners participated in the experiment, which was conducted using a two-alternative forced-choice (2AFC) paradigm. Three participants were musicians with experience ranging between 5-10 years and the remaining participants were non-musicians. All participants were tested and were found to have normal hearing. Sufficient training was then provided until they could consistently distinguish between the trumpet and clarinet samples.

The stimuli from each set were presented monaurally and randomised. For each stimulus, participants were prompted to select “What instrument name best describes the sound?”, selecting from buttons labelled “Trumpet” and “Clarinet”.

5.2.2 Results

5.2.2.1 Psychometric functions

The results from the participants were averaged and psychometric functions were calculated using a least squares fit to a cumulative Gaussian curve. Figure 5.12 illustrates the results, where the dotted horizontal line indicates the threshold for classification as a trumpet at 75%. As the cut-off frequencies f_c decreased, the morph numbers at which the thresholds for classification as a trumpet were generally found to increase.

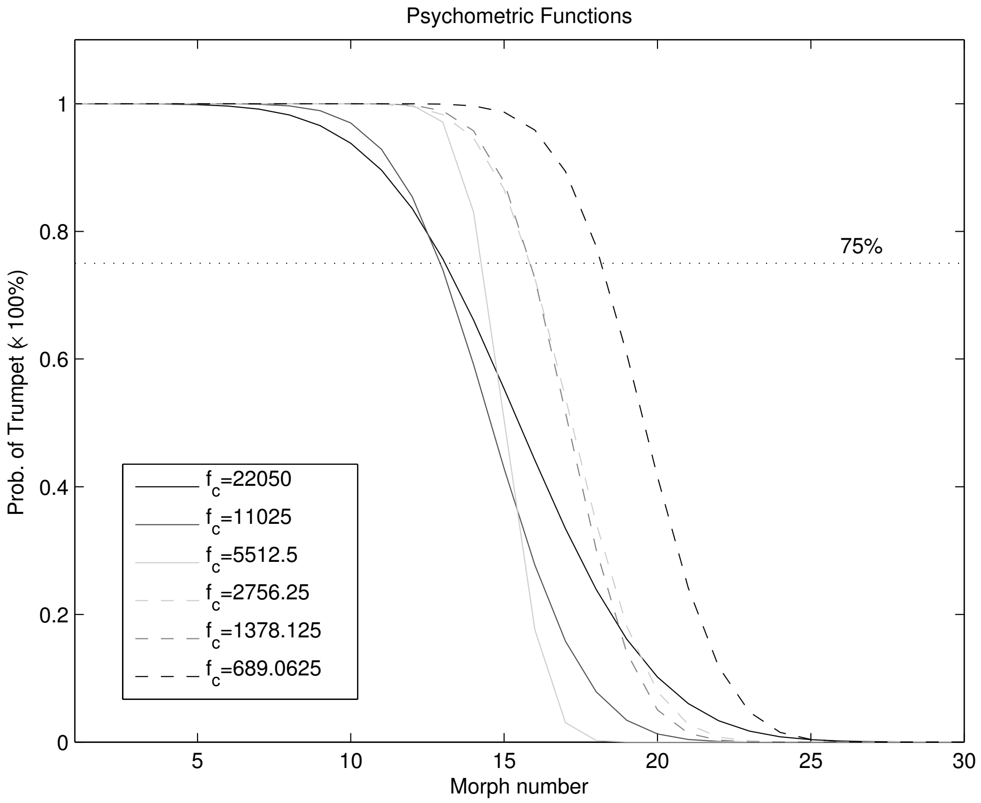

Figure 5.12: Psychometric functions of the 6 sets.

Figure 5.13 compares the psychometric functions of each set with normalised distances using Mel-frequency cepstral coefficients (MFCC) [24] and the spectral centroid [43, 80, 17]. Both the MFCCs and the spectral centroid are parametrisations of the spectrum and have been used extensively in instrument timbre classification.

MFCCs are computed by taking the cosine transform of the log-amplitude spectrum calculated along the Mel-frequency scale. The calculation of the normalised MFCC distance for a morphed sound was obtained by computing a 13 point MFCC vector [104], finding the euclidean distance relative to the MFCCs of the trumpet sample, and normalising the result by the distance between the trumpet sample and the clarinet sample.

The spectral centroid is the amplitude-weighted mean frequency of the energy spectrum given by

$$SC_m = \frac{\sum_k f_k \left|X_m[k]\right|}{\sum_k \left|X_m[k]\right|} \tag{5.7}$$

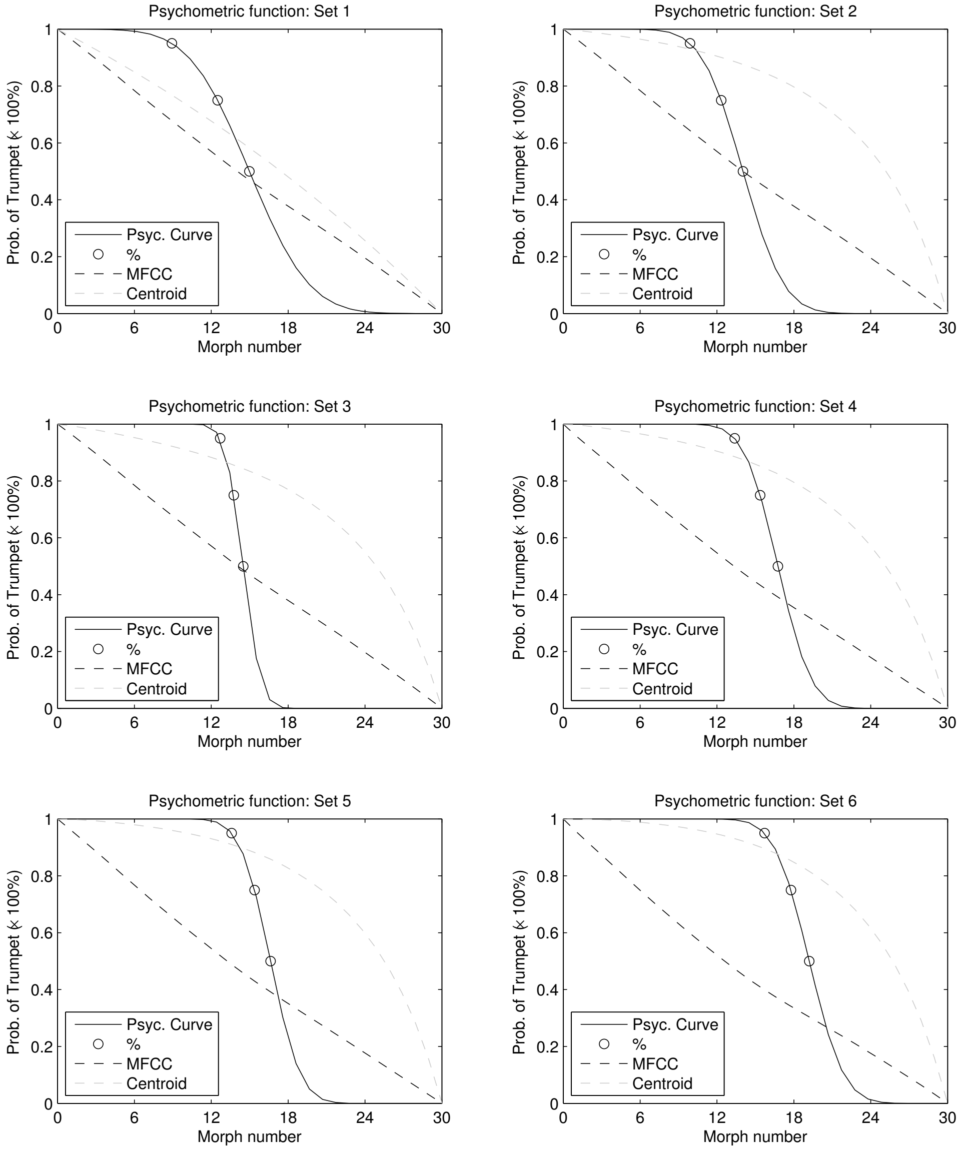

Figure 5.13: Comparison of the psychometric functions and normalised MFCC and spectral centroid distances. The circles indicate the 95%, 75% and 50% probabilities.

where $|X_m[k]|$ is the magnitude of the kth DFT bin at the mth frame, where f_k is the corresponding frequency of the kth DFT bin. The calculation of the normalised spectral centroid distance for a morphed sound was obtained by computing the relative distance between the centroid of the morphed sound and the trumpet sample, and normalising the result by the distance between the trumpet sample and the clarinet sample.

For Set 1, the normalised distances for both the MFCCs and the spectral centroid, align with the 50% probability point of the psychometric function. However as f_c decreases, the spectral centroid quickly deviates from this point while the MFCCs correspond well until Set 4 and 5. This can be attributed to the linear frequency axis employed in the calculation of the spectral centroid. The MFCCs on the other hand, are based on the human-auditory inspired Mel-frequency axis, which has higher frequency resolution at lower frequencies and lower frequency resolution at higher frequencies. Thus the initial high frequency modifications of Sets 2 and 3, which play a perceptually minimal role in influencing the 50% probability point, are also less influential on the MFCCs distance measure, compared to the spectral centroid.

The averaged morphed spectral envelopes corresponding to the 50%, 75% and 95% trumpet identification probabilities are illustrated in Figure 5.14. The figure highlights that the spectral envelopes of what is perceived to be a trumpet even at 95% probability for the various morphs, contain a large proportion of the clarinet resonances. This implies that timbre perception is dependent on more than just the resonances and the anti-resonances of the spectral envelope.

5.2.2.2 Masking Analysis

A psychoacoustic masking model [55] was applied to the morphed sounds and Signal-to-Mask Ratios (SMR) were calculated and time-averaged, resulting in the plots given in Figure 5.15. In each of the 6 sets, trumpet classification (75%) occurs consistently when the SMRs of certain frequency components increase above zero (e.g. 500-700 Hz, 1100-1300 Hz).

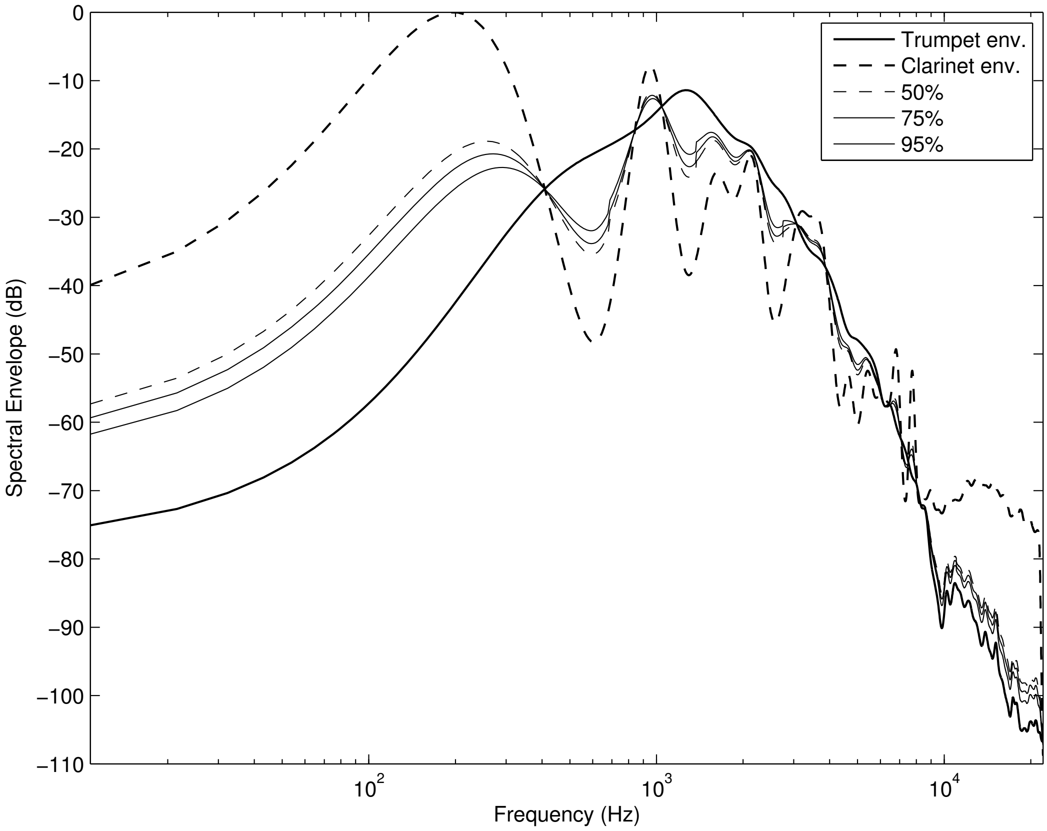

Figure 5.14: Probability for trumpet identification. Averaged results of 6 sets.

5.2.3 Discussion and Conclusion

The results of this experiment provide a few insights into the nature of the timbre space with respect to the spectral envelope. The psychometric functions for each set shown in Figure 5.12, indicate that in the case of morphing from the trumpet to the clarinet, the chance-level discrimination point (50%) corresponds to the morphed sounds that have envelopes that lie approximately half-way between the trumpet and clarinet envelopes for sets 1-3. These sets only have the higher envelope frequencies (>5 kHz) logarithmically morphed, implying that for these samples, these frequencies do not impact timbre discrimination as much as the lower frequencies. Once frequencies below 5 kHz are logarithmically morphed (sets 4-6), the envelopes require more morphing to achieve the chance-level discrimination point, reaching a maximum morph at the lowest cut-off frequency.

The psychometric functions were then compared with two spectrum parametri-

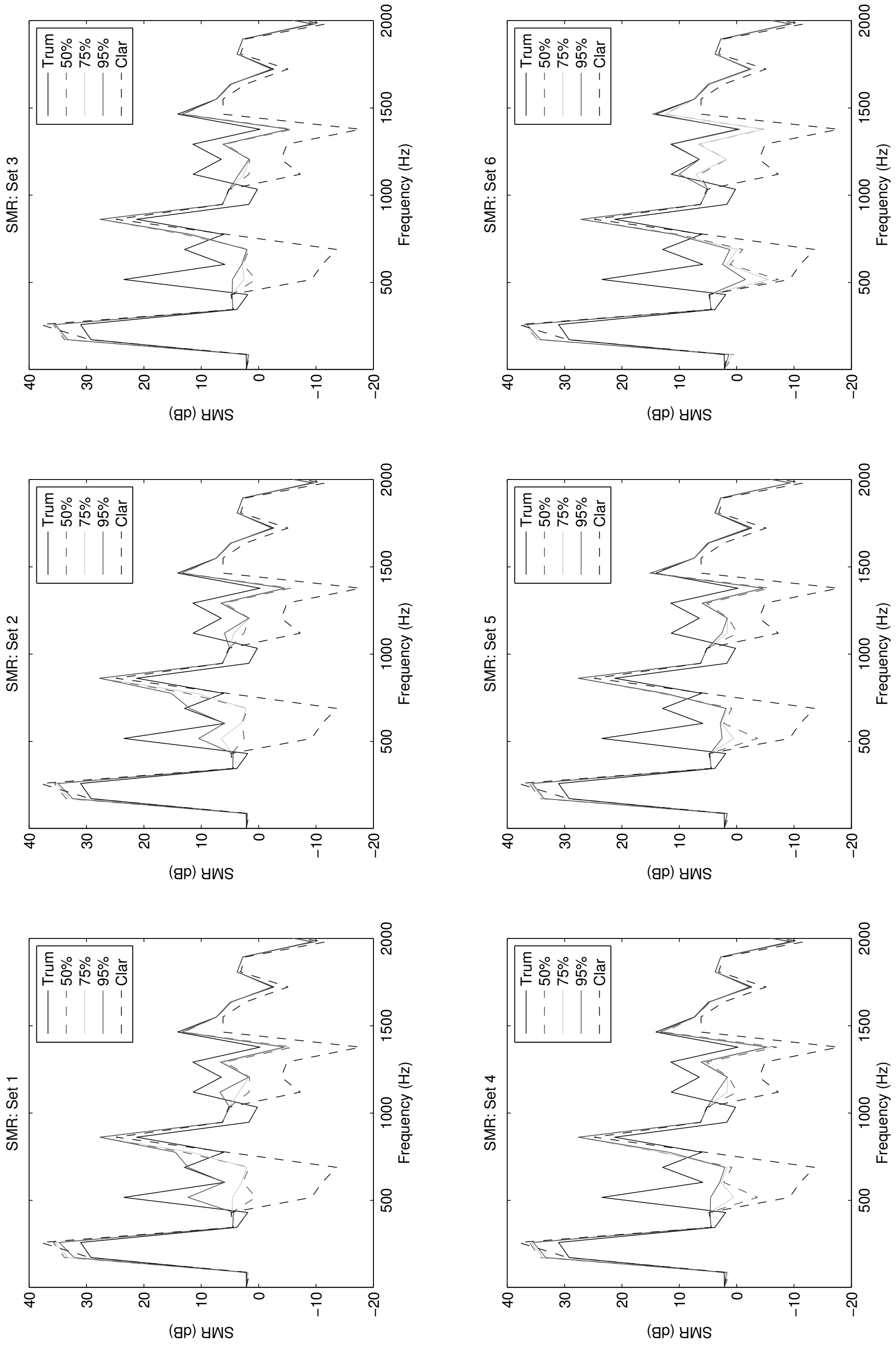

Figure 5.15: Signal-to-Mask Ratio of the 6 sets for 50%, 75% and 95% probabilities.

sations that are used extensively in classification - MFCC and spectral centroid. The MFCCs provided better correspondence to the chance-level discrimination points on the psychometric functions, particularly between sets 1-3. This can be attributed to the approximately logarithmic frequency axis which better approximates the manner in which listeners perceive sound. The spectral centroid has a linear frequency axis, equally weighting all frequencies, which results in a higher sensitivity to the initial higher frequency changes.

The SMR results of the 6 sets reveal potential reasons for the discrepancies found between the normalised MFCC distances and the psychometric functions for sets 4-6. The SMR results highlight that the trumpet classification point (75%) occurs consistently when the SMR at certain frequencies, increases above zero (e.g. 500-700 Hz, 1100-1300 Hz). These frequencies correspond to masked frequencies, which implies that the perceptual timbre space is dependent on masking. Perceptual instrument classification therefore requires these frequencies which provide important timbre cues, to be unmasked before they can be classified as certain instruments. Thus models of the timbre space would benefit by accounting for the masking that occurs in the auditory system.

This experiment has presented a novel strategy for exploring the timbre space using different permutations of linear-logarithmic morphing between instrument spectral envelopes. Upon generating psychometric functions and comparing them with normalised MFCC and spectral centroid distance measures, logarithmic frequency axes, like that of the Mel-frequency scale, were shown to provide a more accurate representation of the timbre space than a linear axis. Furthermore, the SMR analysis revealed that incorporating psychoacoustic masking models into models of timbre may produce models that are more coherent with perception.

5.3 Experiment 3: Noise Sensitivity

Musical timbre is a multidimensional property of sound and while the spectral envelope is a salient attribute of timbre, temporal changes (such as the attack transient) and non-harmonic components (such as the breathiness of wind instruments) have been found to be perceptually relevant [42]. Instrument synthesis experiments in particular have revealed that *real* instrument sounds are usually perceived as having more *roughness* than sounds that are synthesised by merely modelling the harmonic component [29].

To investigate the perceptual salience of the non-harmonic or *noise* components of pitched instruments, musical sounds were decomposed into a harmonic component and a noise component. The perceptual sensitivity to the noise component was then investigated by attenuating various bands of the noise component in a two-alternative forced-choice experiment.

5.3.1 Experimental Method

5.3.1.1 Stimuli

Four musical instrument sounds were selected for analysis. Samples of an alto saxophone, clarinet, viola and flute taken from a University of Iowa website [1] were used. The sounds were recorded using 16 bits, and a 44100 Hz sampling rate, and each sound was played at a pitch of Eb4, corresponding to a fundamental frequency of approximately 311.1 Hz - a frequency within the normal playing range of these instruments and commonly used in timbre experiments for this reason [80, 53]. The duration of each sound was standardised to 1.5 seconds using a 100 msec half-Hanning window to taper the offsets. The onsets of each sample were left unmodified.

5.3.1.2 Stimuli Modification

The *harmonic* component of each stimulus was extracted using the spectral filtering method described in Section 4.1.2. The *noise* component was then obtained by subtracting the harmonic component from the original stimulus in the time domain. The harmonic and noise decomposition of the 4 stimuli are shown in Figure 5.16. The stimuli for the experiment were created by passing the noise component of each stimulus through a zero-phase band-pass filter and the output of the filter was then attenuated and added to the harmonic component of the stimulus. Using 7 zero-phase filters of differing centre frequencies and bandwidths, we compiled a set of stimuli where the output was the original signal with the noise component of a certain frequency band attenuated.

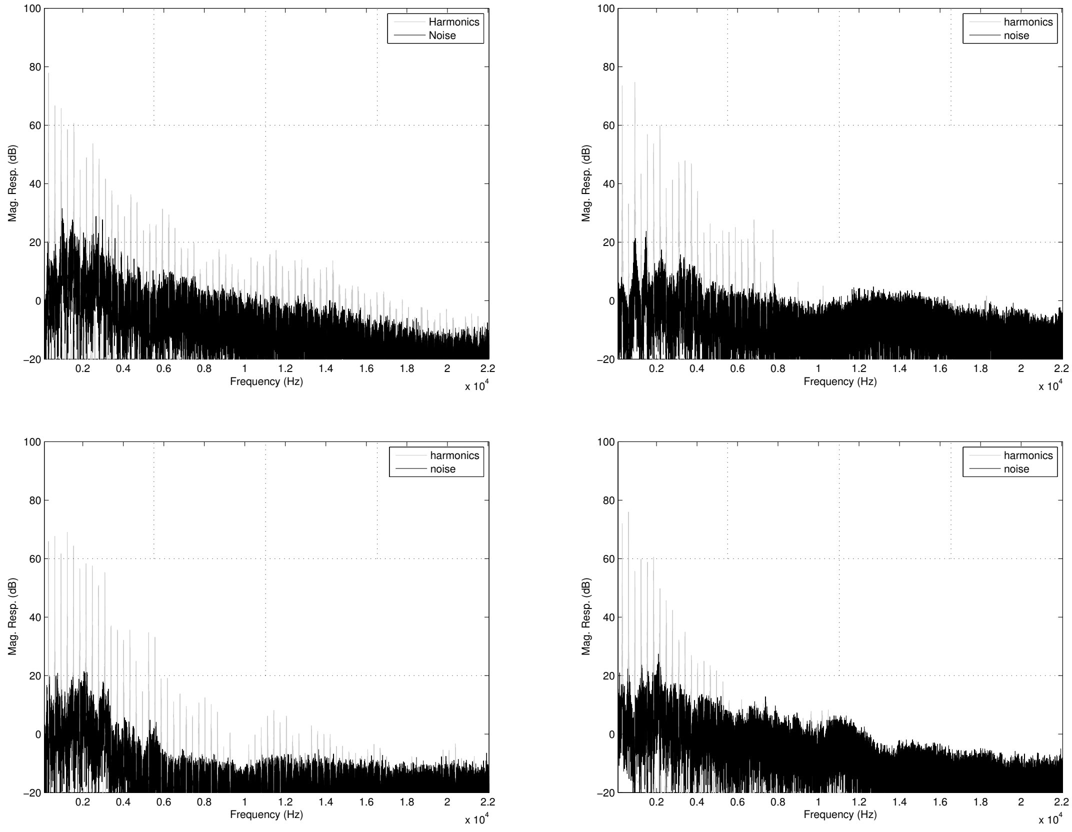

Figure 5.16: The harmonic and noise spectrum of an alto saxophone (top left), clarinet (top right), viola (bottom left) and flute (bottom right). Dotted lines indicate frequency bands attenuated.

The zero-phase filters were designed by taking 256-tap linear-phase band-pass filters (designed by the window method based on a Hamming window) and advancing the output signal by the group delays of the filters. Seven filters were created, consisting and an all-pass filter and the first 6 filters of the experiment in Section 5.1 (see Figure 5.3).

The stimuli presentation was controlled using MATLAB on an Intel PC with an RME Multiface sound card. Each of the stimuli was presented monaurally at an average level of approximately 65 dB SPL through Beyerdynamic DT770pro headphones in a sound-insulated (Acoustic Systems) anechoic chamber.

5.3.1.3 Procedure

Four listeners aged between 20 to 26 years participated in the experiment and all were tested and found to have normal hearing. Two of the participants had musical training with experience ranging between 5-10 years.

A two-alternative forced-choice (2AFC) Reference AB, 1-up 2-down paradigm [73] was used for all our experimentation. For each trial, the participant heard three sounds: the reference sound (original, unfiltered) followed by two other sounds - one of which was filtered and the other which was the same as the reference. The order of presentation of the two latter sounds were independently randomised for each trial and 300 msec silence periods separated the presentation of each sound. For each trial, the participant was prompted with “Which sound has a different timbre to the reference?” and had to respond by clicking buttons marked A and B on the screen. Once a response was submitted, feedback was provided to the participant in the form of “Correct” or “Incorrect”.

The first trial presented for each centre frequency was always with the most attenuation and the attenuation was incrementally decreased to include more of the contents of the band. The attenuation step sizes changed from 4 dB to 2 dB and finally to 0.5 dB. The last 3 reversals were averaged to estimate the discrimination threshold. Listeners were trained for 15 mins to familiarise themselves with the task prior to the experiment.

5.3.2 Results

The sensitivity of each frequency band to the attenuation of band-pass filtered noise was calculated using the band attenuation (BA) measure described in Section 5.1.2.1, Equation 5.1. The BA is defined to be the original energy of the band, divided by the minimum difference required to observe a change in that band. The BA results from all the participants were averaged and are illustrated in Figure 5.17.

The sensitivity to noise attenuation varies for each instrument and also varies as a function of frequency. Sensitivity to noise attenuation is generally higher around 6-11 kHz and as the bandwidth narrows, sensitivity decreases. The flute and clarinet, which have relatively large noise components, are the most sensitive to noise attenuation for both broadband and narrow-band attenuation. The harmonic component of the saxophone and the viola is more prominent, with harmonics extending up to the higher frequencies, resulting in lower sensitivity to noise attenuation particularly around the higher frequencies.

5.3.3 Discussion and Conclusion

The BA results indicate that the noise component of a harmonics plus noise decomposition is a salient attribute of timbre. Sensitivity to noise attenuation varies as a function of frequency, and sensitivity is at a maximum around 6-11 kHz. This implies that low frequency harmonics are effective maskers of noise, resulting in lower sensitivity to noise below 6 kHz. Sensitivity was also found to decrease as the bandwidth of the noise being attenuated narrowed, suggesting that sensitivity to noise is governed by broadband sensitivity.

Compared to the sensitivity results for the spectral envelope (Section 5.1), sensitivity to instrument noise is lower and exhibits higher variability as a function of frequency for different instruments. Future investigations should explore the temporal nature of the noise in conjunction with spectral variations.

The results of this experiment highlight the need to consider the noise component in the development of timbre models. Timbre models have traditionally focused on the spectral and temporal variations of the harmonics, however the results presented in this study indicate that the non-harmonic components, for certain instruments such as the flute and the clarinet, are salient, and sensitivities to these components vary as a function of frequency and bandwidth.

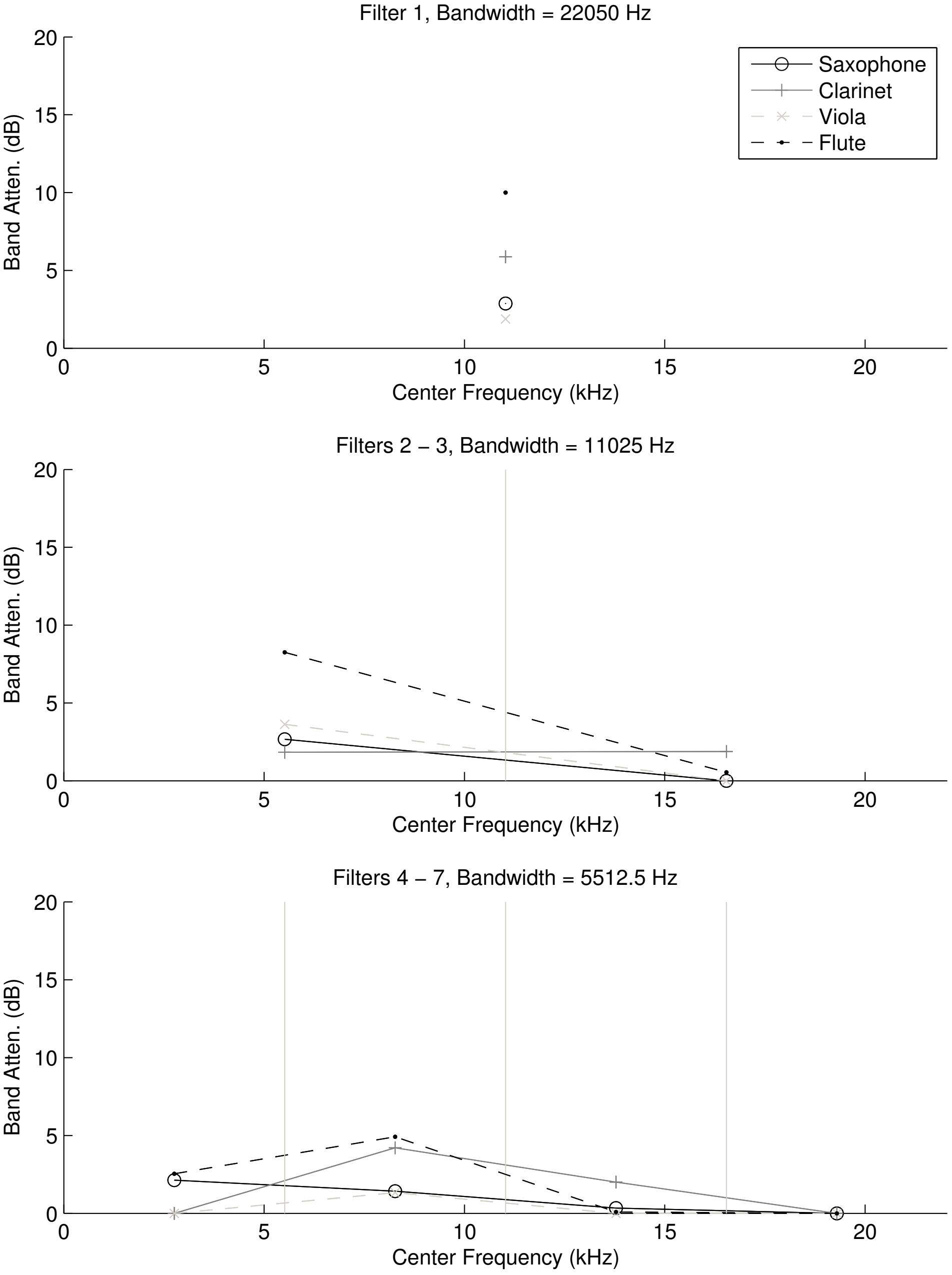

Figure 5.17: Error level results.

Chapter 6

Separation of Harmonic Musical Instrument Notes using Spectro-Temporal Modelling of Harmonic Magnitudes and Multiple Input Spectrogram Inversion

Resolving overlapping harmonics and the re-synthesis of accurate source signals remain persistent and unsolved issues when separating individual sources from a single channel mixture of harmonic musical instruments. In this chapter we bring together concepts described in previous chapters and present novel methods that address both of these issues.

In Section 3.3, it was shown that the spectro-temporal correlations which exist between instrument harmonics can be modelled using linear combinations of neighbouring harmonics. In this chapter, we elaborate on this model, deriving a novel method which facilitates source separation by resolving overlapping harmonics. The performance of the method is evaluated against other approaches, and is shown to provide better and more robust estimates of the harmonic magnitudes as a function of time. The multiple input spectrogram inversion algorithm (Section 4.2.4) is then employed for source synthesis and additional results are presented demonstrating its potential in source separation. The methods are then combined in a source separation framework, and the overall performance is evaluated using a variety of objective distortion measures, as well as a subjective evaluation, motivated by the perceptual investigations presented in Chapter 5.

6.1 Introduction

The separation of individual sources from a single channel mixture of multiple sources involves amongst other considerations, an insight into the nature of correlations and inter-dependencies between the sources. Musical mixtures in particular, are noteworthy, because they are a practical example of spectro-temporally concordant mixtures. The extensive dependencies within such a mixture inherently coerce the need to engage source-specific information in order to produce exceptional source separation.

In the context of music source separation, a system capable of separating musical mixtures from a single channel also has a vast number of applications. Some of these include source modelling based audio compression; greater flexibility in the recording, mixing and mastering of audio; robust automatic transcription; and the facilitation of music education. Source separation would also open many avenues for the processing of polyphonic music signals, by the numerous monophonic audio processing algorithms which have been developed.

The separation of musical mixtures is a non-trivial task due to the vast proportion of overlapping components. Western music in particular, is often arranged so that sounds are not only played simultaneously, but are also harmonically related. This results in numerous overlapping harmonic trajectories, which make source separation problematic. The work described in this thesis presents a novel method of estimating regions of overlap by exploiting spectro-temporal intra-instrument dependencies, integrating the spectral and temporal approaches which are currently employed in a mutually exclusive manner in existing systems. Subsequent to the harmonic magnitude extraction using this method, we present a unique, closed-loop approach to source synthesis, separating sources by iteratively minimising the aggregate error of the sources, constraining the minimisation to a set of estimated parameters.

Existing single channel source separation systems can be broadly classified into either model-based systems or unsupervised learning based systems. The latter are usually built on a simple linear model and aim to find decompositions where the sources are statistically independent or non-redundant. Proposed systems have been based on Independent Subspace Analysis (ISA) [18], Non-negative Matrix Factorisation [110] and sparse coding [123, 3]. Based on the principles of Independent Component Analysis (ICA), ISA provides a framework in the power spectrogram domain for the estimation of sources from a single mixture that are as statistically independent as possible. By contrast, Non-negative Matrix Factorisation (NMF) algorithms employ non-negativity constraints to the basis functions, which has been shown to be sufficient for the separation of sources [110]. In sparse coding algorithms, redundancy reduction is the central motivation. Sources are formed using a small number of elements from a large set. The selection of these elements is typically based on a cost function which aims to minimise the estimation error and maximise the efficiency of the representations.

While unsupervised learning algorithms have produced good results of late, model-based single-channel separation systems [34, 124] have been able to achieve greater separation in mixtures that are predominantly harmonic with a limited number of simultaneous sources. Harmonic sources are well modelled using a sinusoidal model [81, 111, 101], where sinusoids are approximately harmonically spaced and each sinusoid trajectory is parametrised by frequency, amplitude and phase. Grouping principles are then applied to form sources by superimposing sinusoidal trajectories, and the parametric sinusoidal information is then used to aid in the re-synthesis of the sources.

Model-based, single-channel source separation is a complex problem which is comprised of a number of non-trivial problems, ranging from parameter selection and modelling, to mixture analysis and source synthesis. Of the myriad of problems to be solved in source separation, the resolution of overlapping harmonics and source synthesis are particularly challenging issues. Overlapping harmonics are both prolific and problematic in the separation of musical mixtures, and their resolution is non-trivial due to the high spectro-temporal correlations which exist between sources. Once the harmonics have been estimated, the sources need to be synthesised and obtaining high-quality separated sources from parametric information remains another challenging task. In the following, some of the existing techniques for resolving overlapping harmonics and synthesising sources are reviewed.

In musical mixtures, harmonic and rhythmic concordance is problematic, resulting in many overlapping sinusoidal trajectories. Attempts to resolve these overlapping harmonics can be broadly classified as either a spectral approach or a temporal approach. Spectral approaches [87, 67, 34], resolve harmonics exclusively based on spectral information at a given time, disregarding information of the spectrum at other points in time. Parsons [87] and Klapuri [67] both exploited the expectation that spectral envelopes of real sound sources tend to be continuous, and overlapping harmonics were resolved by interpolating between the amplitudes of adjacent harmonics. In [34], Every and Szymanski designed spectral filters to filter harmonics from their estimated frequency locations. The filters worked to effectively partition the energy of shared spectral peaks to the relevant sources based on frequency location. In all of these methods, overlapping harmonics were estimated by utilising the information of spectrally neighbouring harmonics in a given frame. By contrast, the methods which employ a temporal approach [127, 130], exploit the temporal correlations found between harmonics but do not utilise the spectral information. In [127], Viste and Evangelista combined knowledge of the temporal nature of harmonic tracks with multichannel separation techniques to resolve overlapping harmonics. In regions of harmonic overlap, the closest non-overlapping harmonic track was used to aid in the spatial demixing of the sources. Woodruff et al. [130], proposed a method for resolving overlapping harmonics using the strongest harmonic to model amplitude changes and pitch information to predict phase changes. These temporally based methods exclusively employ a selected trajectory to aid in separation, thereby ignoring other spectral information. Given these two approaches to harmonic magnitude estimation, the work described within is based on the hypothesis that a unified spectro-temporal model will provide better performance by exploiting both the spectral and temporal dependencies which are known to exist between harmonic magnitude tracks.

Individual sources are synthesised from their parametric representations (including harmonic magnitudes). There have been a variety of approaches to estimating the source magnitudes and phases from parametric models and these have included using banks of oscillators [81], spectral synthesis [102], binary masking [129], and a hybrid approach [130]. The approaches employed by McAulay and Quatieri [81], as well as Serra et al. [102], are based on sinusoidal synthesis. Directly synthesising sinusoids in the context of source separation, produces estimates that are devoid of interfering artifacts, however this can also result in a loss of perceptual 'naturalness' as the sidelobes of the harmonics contain perceptually salient information. An alternative approach presented by Wang [129], relies on the construction of binary masks to denote regions of the time-frequency representation which correspond to a particular source. This method captures these spectral nuances of the harmonics, but only when the harmonics do not overlap with any other harmonics. In the regions where there is an overlap, binary masking produces unreliable source estimates corrupted by the presence of other sources. In [130], Woodruff et al. presented a hybrid approach to source synthesis. After using least-squares estimation to approximate sinusoidal parameters, sources were estimated by using binary masking for non-overlapping harmonics and spectral synthesis for overlapping harmonics. Each of the methods reviewed above employ an open-loop estimation method for generating time domain source estimates. In the typical scenario where all the sources are to be estimated in a mixture, we propose that a closed-loop estimation system can improve the quality of synthesis by successively refining the source estimates.

In following sections, we present novel methods for spectro-temporally resolving overlapping harmonics and synthesising sources using a closed-loop estimator. In Section 6.2, a method for resolving overlapping harmonics using spectro-temporal harmonic magnitude track prediction is presented. A diverse selection of harmonic musical instruments are analysed and a generalised-instrument magnitude track prediction model is derived that utilises both the spectral and temporal information to generate magnitude track estimates. In Section 6.3, a novel closed-loop algorithm is presented to refine the synthesis of the separated sources. Given the source magnitudes, the algorithm iteratively refines the phase estimates of the sources using a spectrogram inversion algorithm for multiple inputs. In Section 6.4, the performance of the harmonic magnitude track prediction model and the phase estimation algorithm are evaluated independently. They are then utilised in a separation system to assess the objective and subjective performance gains achieved when separating a mixture of harmonic sounds. The results are then discussed in Section 6.5, followed by conclusions in Section 6.6.

6.2 Harmonic Magnitude Track Prediction

6.2.1 Sinusoidal Modelling of Harmonic Tracks

The modelling in this work begins with a transformation of the time domain signal into a time-frequency representation, from which sinusoid frequencies and amplitudes are estimated. Sinusoid trajectories or *tracks* are then defined and the tracks are grouped into sources. This kind of sinusoidal modelling [81, 111, 101] has been widely adopted, particularly in music signal processing due to its efficient and convenient parametrisation of the harmonic components of pitched sounds.

Consider a mixture $y[n] = \sum_{j=1}^{J} x_j[n]$, comprising of a superposition of J sources. Each source $x_j[n]$ is assumed to be a harmonic source and is therefore adequately modelled as a sum of sinusoids

$$x_j[n] = \sum_{h=1}^{H_j} a_{h,j}[n] cos \left(2\pi f_{h,j}[n] n / f_s + \theta_{h,j}[n]\right) \tag{6.1}$$

where $(a_{h,j}[n], f_{h,j}[n], \theta_{h,j}[n])$ are the time-varying amplitudes, frequencies and phases for the h^{th} harmonic of the j^{th} source respectively, f_s is the sampling frequency and H_j is the number of harmonics for the j^{th} source. The objective of any separation system is to obtain estimates of the sources $\hat{x}_j[n]$ that best approximate the true sources $x_j[n]$, from a single channel observation of the mixture, $y[n]$.

The process begins with the computation of the Short-time-Fourier-transform (STFT) for the mixture $y[n]$, which gives the time-frequency representation

$$Y_r[k] = \sum_{n=-\infty}^{\infty} y[n]w[n - rH]e^{-i\frac{2\pi k}{N}n} \tag{6.2}$$

where r is the time frame index, k is the DFT frequency bin index, $w[n]$ is an analysis window, H is the hop size in samples and N is the size of the DFT. In this work a Hanning window is used for the analysis window.

The magnitude spectrum $|Y_r|$ of the mixture is then calculated and the harmonics estimated using spectral peak picking [34]. Refined estimates of the frequencies corresponding to each harmonic peak are then calculated using the Phase Derivative Fast Fourier Transform (PDFFT) (Section 3.2, [46]), and the peaks are used to form harmonic tracks. Each harmonic track is parametrised as a vector of magnitudes M_j^h, which correspond to a vector of frequencies F_j^h, where h is the harmonic number of the jth source. The parameters are estimated by tracking the harmonic peaks based on frequency proximity, harmonicity and magnitude continuity [81, 34]. For the purposes of evaluating the proposed methods, the fundamental frequency, F_0 , of each note are assumed to be known and the validity of each track in the mixture is verified using the original sources.

Having parametrised the harmonic tracks of the sources in the mixture, overlapping harmonics were classified by observing their proximity to other source harmonics. An overlapping vector O_j^h for each frame r, was then defined such that,

$$O_j^h[r] = \begin{cases} 1, & if \left|F_j^h[r] - F_s^g[r]\right| < \delta_{overlap} \\ 0, & otherwise \end{cases} \tag{6.3}$$

where g is the harmonic number of source s, $s \neq j$, and $\delta_{overlap}$ is dependent on the width of the primary spectral lobe related to w.

6.2.2 Harmonic Track Prediction

Previous approaches to resolving overlapping harmonics have relied on either interpolating spectral information, or alternatively utilising a single temporal magnitude trajectory to estimate the magnitudes of overlapping harmonics. In this section, we describe a spectro-temporal approach to resolving overlapping harmonics that utilises both the spectral and the temporal information to predict harmonic track estimates. The method is based on a generalised-instrument track weighting model presented in Section 3.3.3, where harmonic tracks are modelled as a weighted linear combination of neighbouring tracks.

The magnitude tracks of musical instrument harmonics are known to be highly correlated to each other and knowledge of this has been used in a number of separation systems [125, 127, 130]. Gunawan and Sen [48] quantitatively investigated this correlation on a large database of musical instrument notes and found that the correlation between harmonic magnitude tracks was highest between directly adjacent partials, with the correlation decreasing exponentially with increasing distance from a particular harmonic. The significant correlation between harmonic tracks, suggests that harmonic tracks can be predicted as a weighted linear combination of normalised neighbouring harmonic tracks. The prediction of the h^{th} harmonic magnitude track $\hat{M}^h$ can be expressed as,

$$\hat{M}^h = \frac{\sum_{q=1}^{H} v_{h,q} \bar{M}^q}{\sum_{q=1}^{H} v_{h,q}} \tag{6.4}$$

where $q \neq h$, H is number of harmonics of the source, and $v_{q,h}$ is the weight contribution of the qth normalised harmonic magnitude track $\bar{M}^q$. An optimal solution for these weights in a least squares sense can be found using regularised least squares, which has the solution:

$$v_h = (\Gamma^T \Gamma + \delta_{rls} I)^{-1} \Gamma^T \bar{M}^h \tag{6.5}$$

where $\bar{M}^h$ is the h^{th} normalised harmonic magnitude track which is to be predicted, Γ is the matrix with columns containing the normalised harmonic magnitude tracks but excluding $\bar{M}^h$, δ_{rls} is the regularisation parameter, I is the identity matrix and ^{T} denotes the conjugate transpose. To construct a generalised-instrument track weighting model, Equation 6.5 was used to compute the weights for 3000 musical instrument samples from the University of Iowa instrument database [1]. The weights for each harmonic were then averaged over all the samples, and a curve was fit to the averaged weights. The resulting parametrisation is given by,

$$v_{h,q} = \begin{cases} \frac{-\left((h+\beta_1)^{-1}+\gamma_1\right)}{q-h}, & q < h \\ \frac{(h+\beta_2)^{-1}+\gamma_2}{q-h}, & q > h \end{cases} \tag{6.6}$$

for the weight of the qth harmonic to model the hth harmonic, where the optimum values for $\beta_1, \gamma_1, \beta_2, \gamma_2$ were empirically found to be $\beta_1 = 0.994366, \gamma_1 = 0.092848, \beta_2 = 1.880769, \gamma_2 = 0.060059$, by minimising the root mean squared error as illustrated in Figure 6.2. Figure 6.1 illustrates the weights determined using regularised least squares, and the corresponding model of the weights.

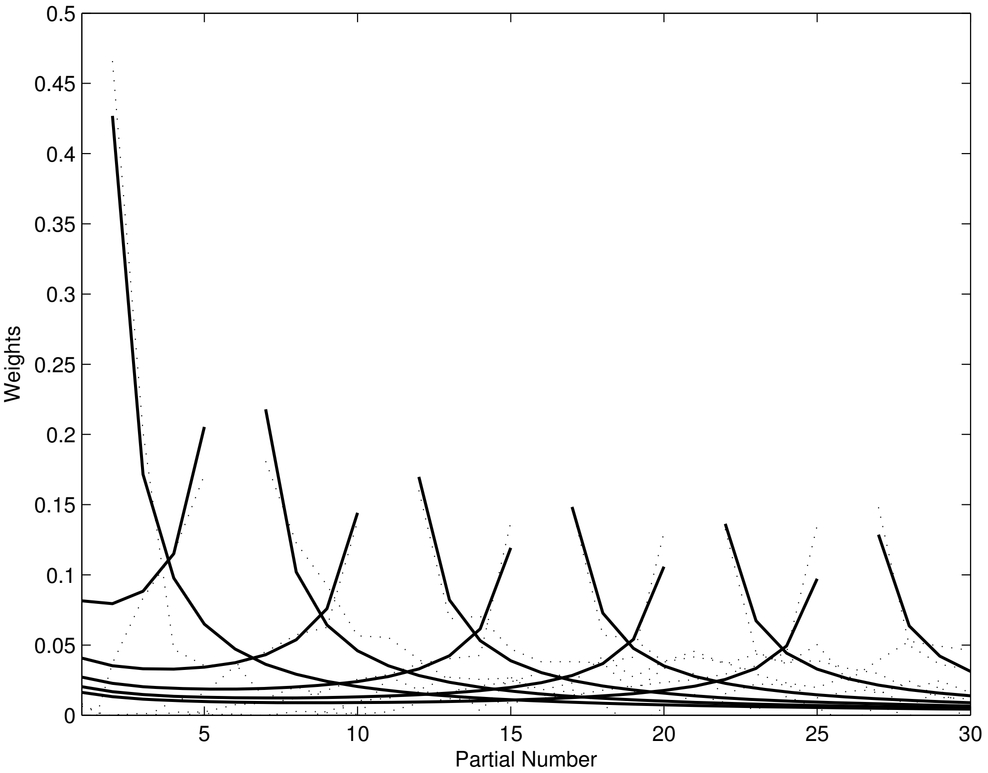

Figure 6.1: Weighting functions (dotted lines) averaged over 3000 musical instrument samples, and the modelled weighting functions (solid lines). For clarity, only every 5th set of weights is illustrated.

The harmonic track estimate $\hat{M}^h$, is therefore a prediction of the h^{th} harmonic track, which is derived by exploiting the spectral and temporal correlations which

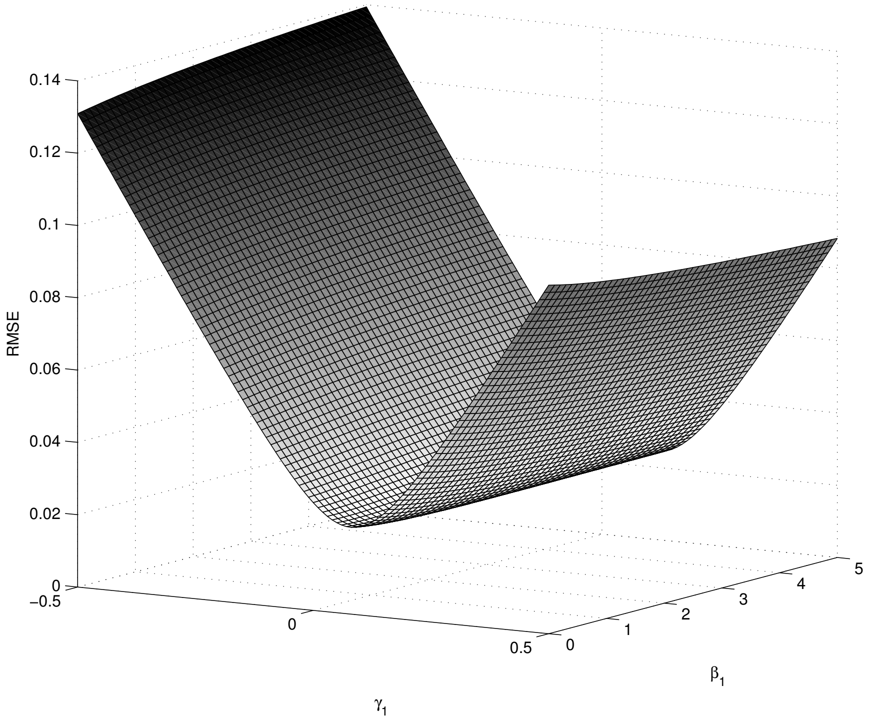

Figure 6.2: Empirical determination of model parameters β_1 and γ_1. The parameters for the model given in Equation 3.21, were found by minimising the root mean squared error between the model and the averaged weights determined by regularised least squares.

exist between the neighbouring harmonic tracks. In the source separation context, this provides a robust method of estimating the magnitude trajectories of harmonic tracks which may have been corrupted due to the interference of other sources.

6.3 Iterative Source Synthesis

In separation systems based on sinusoidal modelling, a widely-adopted approach for source synthesis, involves the estimation of the magnitude spectra of the sources, followed by the masking of the phase spectra of the mixture in regions where the magnitude spectra is salient [129, 34]. The sources are then synthesised in an open-loop manner, inverting the spectral magnitude and phase information into the time-domain. While the binary masking of the mixed phase may suffice in applications where the sources in the mixture are uncorrelated, using the mixed phase for source synthesis in music mixtures produces significant distortions due to sources being spectrally and temporally concordant.

In this section, we propose a novel closed-loop algorithm to synthesise separated sources. The method is an extension of the spectrogram inversion algorithms [44, 131], iteratively producing time-domain source estimates from magnitude spectra estimates and the time-domain mixture signal.

In a typical source separation scenario, the observed acoustic waveform $y[n]$ can be written as a superposition of source signals $y[n] = \sum_{j=1}^{J} x_j[n]$, where $x_j[n]$ is the jth source signal and J is the number of sources. Given the short-time Fourier transform magnitude (STFTM) estimates of the sources, we describe the Multiple Input Spectrogram Inversion (MISI) algorithm, which iteratively estimates the time-domain source signals $x_j[n]$ in a mixture $y[n]$ given the corresponding STFTM of the source signals. An overview of the algorithm is illustrated in Figure 6.3.

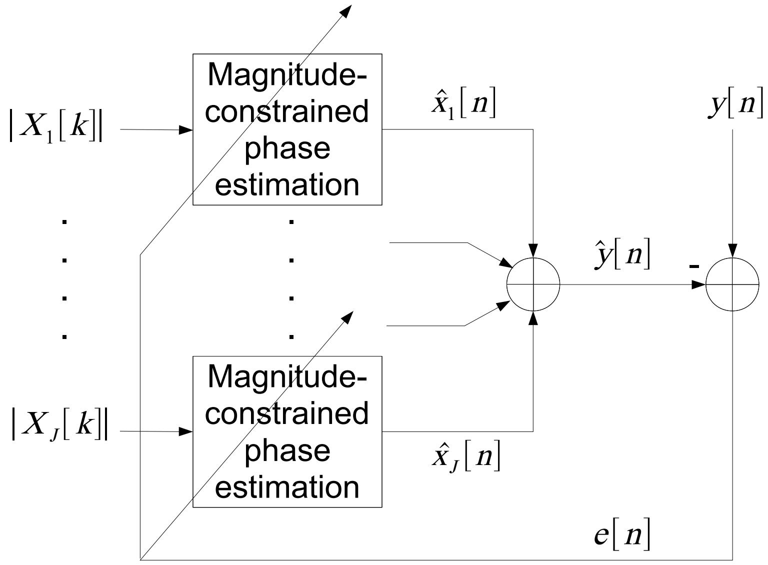
$|X_1[k]|$
Magnitude-constrained phase estimation
$\hat{x}_1[n]$
$y[n]$
$\hat{y}[n]$
-
$|X_J[k]|$
Magnitude-constrained phase estimation
$\hat{x}_J[n]$
$e[n]$

Figure 6.3: Overview of the MISI algorithm. Given the magnitude estimates of each source, the phase responses are estimated and the signal is transformed into the time domain. The source estimates are then subtracted from the original mixture and the error is used to refine the phase estimates.

Consider the short-time Fourier transform of the jth source $x_j[n]$

$$X_{j,r}[k] = \sum_{n=-\infty}^{\infty} x_j[n]w[n - rH]e^{-i\frac{2\pi k}{N}n} \tag{6.7}$$

where w is the analysis window, H is a positive integer denoting the hop size, N is the size of the DFT, and r is the frame index of the STFT. If the STFTM of the source $|X_{j,r}[k]|$, is known or can be well estimated, then it is possible to iteratively estimate each source using spectrogram inversion. Spectrogram inversion [44, 131] algorithms are a subset of reconstruction algorithms [7] which aim to estimate signals by recovering the missing phase information through an iterative process that converges towards a signal with a magnitude-constrained spectrum.

Using the update equation described by Griffin and Lim [44], it is possible to calculate an estimate for the $i_t + 1$th iteration of the jth source signal,

$$\hat{x}_j^{i_t+1}[n] = \frac{\sum_{r=-\infty}^{\infty} w[n-rH]\frac{1}{N}\sum_{k=0}^{N-1} \bar{X}_{j,r}^{i_t}[k]e^{i\frac{2\pi k}{N}n}}{\sum_{r=-\infty}^{\infty} w^2[n-rH]} \tag{6.8}$$

where

$$\bar{X}_{j,r}^{i_t}[k] = |X_{j,r}[k]| \frac{\bar{\bar{X}}_{j,r}^{i_t}[k]}{\left|\bar{\bar{X}}_{j,r}^{i_t}[k]\right|} \tag{6.9}$$

Since multiple inputs are involved, it is necessary to account for the error term obtained by subtracting the superposition of the estimated sources $\hat{y}[n] = \sum_{j=1}^{J} \hat{x}_j^{i_t}[n]$, from the mixture $y[n]$. $\bar{\bar{X}}_j^{i_t}$ is thus obtained by taking the STFT of

$$\bar{\bar{x}}_j^{i_t}[n] = \hat{x}_j^{i_t}[n] + \frac{e^{i_t}[n]}{J} \tag{6.10}$$

$$e^{i_t}[n] = y[n] - \sum_{j=1}^{J} \hat{x}_j^{i_t}[n] \tag{6.11}$$

The algorithm is initialised with each of the initial source estimates set to $\hat{x}_j^0[n] = y[n]$ and the error set to $e^0[n] = 0$. The algorithm then constrains each estimate with the known STFTM $|X_{j,r}[k]|$ and then calculates an inverse STFT. The MISI algorithm then accounts for the total error $e^{i_t}[n]$, between $y[n]$ and $\hat{y}[n]$, adding a scaled version of the error to each of the source estimates before the next iteration. As long as the sum of the scaled errors equals the total error, the energy of the summed sources will be conserved at each iteration. In the MISI algorithm, this is achieved by simply dividing the error equally among the sources, so that the scaled error assigned to each source is given by $\frac{e^{i}_{t}[n]}{J}$, where J is the total number of sources. At each iteration, the scaled errors that are fed back, combined with the magnitude constraints, ensure that the phase estimates of each source approach the true phases. The source magnitude constraints shape the errors so that the phase is re-estimated in the frequency regions where the error is large, and the phase is retained in the regions where the error is low. Thus the MISI algorithm essentially computes a spectrogram inversion for each source with the additional constraint of minimising the error between the mixture and the superposition of the estimated sources at each iteration.

6.4 Results

To demonstrate the performance of the presented methods, each method was evaluated individually and also in the context of a source separation system. The performance of the harmonic magnitude track prediction method is evaluated in Section 6.4.1, while the performance of the Multiple Input Spectrogram Inversion algorithm is evaluated in Section 6.4.2. Both methods are then placed within the context of a source separation system, and the separation of a harmonic mixture is assessed in Section 6.4.3.

All of the results presented use the samples from the University of Iowa musical instrument samples database [1], with each sample recorded at 44.1 kHz sampling rate and 16 bits. The samples range in duration from 1.5-5 seconds and have fundamental frequencies between 65-2100 Hz. Sample mixtures were created by randomly summing p samples in the time domain, where p is the polyphony of the mixture. The power of each sample was randomly scaled between 0-10 dB to simulate the varying levels of sources found in musical mixtures. The parameters for the STFT were $N = 8192$, $H = 1024$, Hanning analysis window, $f_s = 44.1$ kHz, and the parameter defining overlapping tracks was $\delta_{overlap} = 2\frac{fs}{N}$.

6.4.1 Harmonic Magnitude Track Prediction Results

The harmonic magnitude track prediction method was first assessed by analysing the prediction performance for individual instrument samples. Sinusoidal models were generated for 360 monophonic samples, comprising of 20 samples from 18 different instruments. Magnitude track prediction estimates were calculated for the first 20 harmonics of each sample, using the method described in Section 6.2. The correlation coefficient was then calculated to compare the estimates to the true magnitude tracks. To comparatively evaluate the performance of the prediction method, the results were assessed relative to existing methods of magnitude track estimation, which include the spectral approach of linear interpolation which is employed in [87, 34], as well as the temporal approaches of the strongest harmonic [130], and the adjacent harmonic[127].

Box plots of the averaged correlation coefficients for each method are illustrated in Figure 6.4. The upper and lower edges of each box correspond to the upper and lower quartiles, the middle line denotes the median value, and the whiskers extend to the largest observations excluding the outliers. For clarity, the outliers have been excluded from the plot. On average, the prediction method produced better estimates of the magnitude track compared to the other methods, and did so with a lesser degree of variability.

Figure 6.5 is a plot of the correlation coefficients averaged over 18 different instruments. The figure illustrates the consistent performance of the magnitude track prediction method across a range of different musical instruments. While the other methods vary in their performance as a function of instrument, the prediction method leverages both the spectral and temporal information to provide consistently accurate estimates. By contrast, the estimation consistency of the methods varies more so as a function of instrument. The spectral interpolation method performs well for instruments that have a smooth spectral envelope, such as the trumpet, however for instruments where this is not the case, such as the clarinet, the interpolation method does not perform as well as the other methods. The methods employing the strongest harmonic track and the adjacent harmonic track also vary in performance,

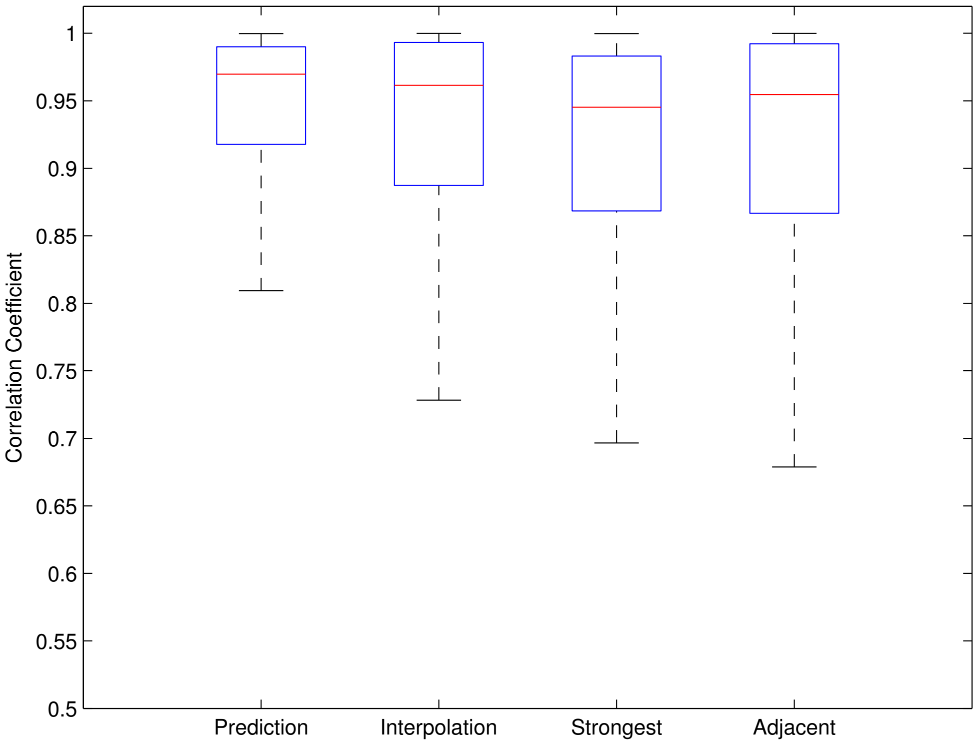

Figure 6.4: Box plots of the correlation coefficient comparing the proposed harmonic magnitude track prediction, spectral interpolation, strongest track and adjacent track methods. The results are an average over the first 20 harmonics of 360 monophonic instrument samples. The proposed prediction method produces the optimum estimates on average, with the minimum variance.

each being highly dependent on the particular instrument being estimated.

The harmonic magnitude track prediction method was also assessed in the context of random mixtures of varying polyphonies. Polyphonies ranged from 2 to 6 samples per mixture, and for each polyphony, 50 random mixtures were assessed. Harmonic track prediction estimates were calculated for each overlapping harmonic, and were scaled by α_h, which was estimated in one of two ways depending on the what proportion of the track was overlapping. If $\frac{\sum_{\forall r \in R^h_{no}} O^h[r]}{\sum_{\forall r} O^h[r]} \leq \delta_{toverlap}$, where R^h_{no} is the set of frame indices corresponding to the non-overlapping frames for harmonic track h, then the non-overlapping temporal regions of the track were used to estimate the scale factor, $\alpha_h = \frac{\sum_{\forall r \in R^h_{no}} M^h[r]}{\sum_{\forall r \in R^h_{no}} \hat{M}^h[r]}$. If $\frac{\sum_{\forall r \in R^h_{no}} O^h[r]}{\sum_{\forall r} O^h[r]} > \delta_{toverlap}$, then due to the insufficient non-overlapped portions of the track, the scale factor was calculated using

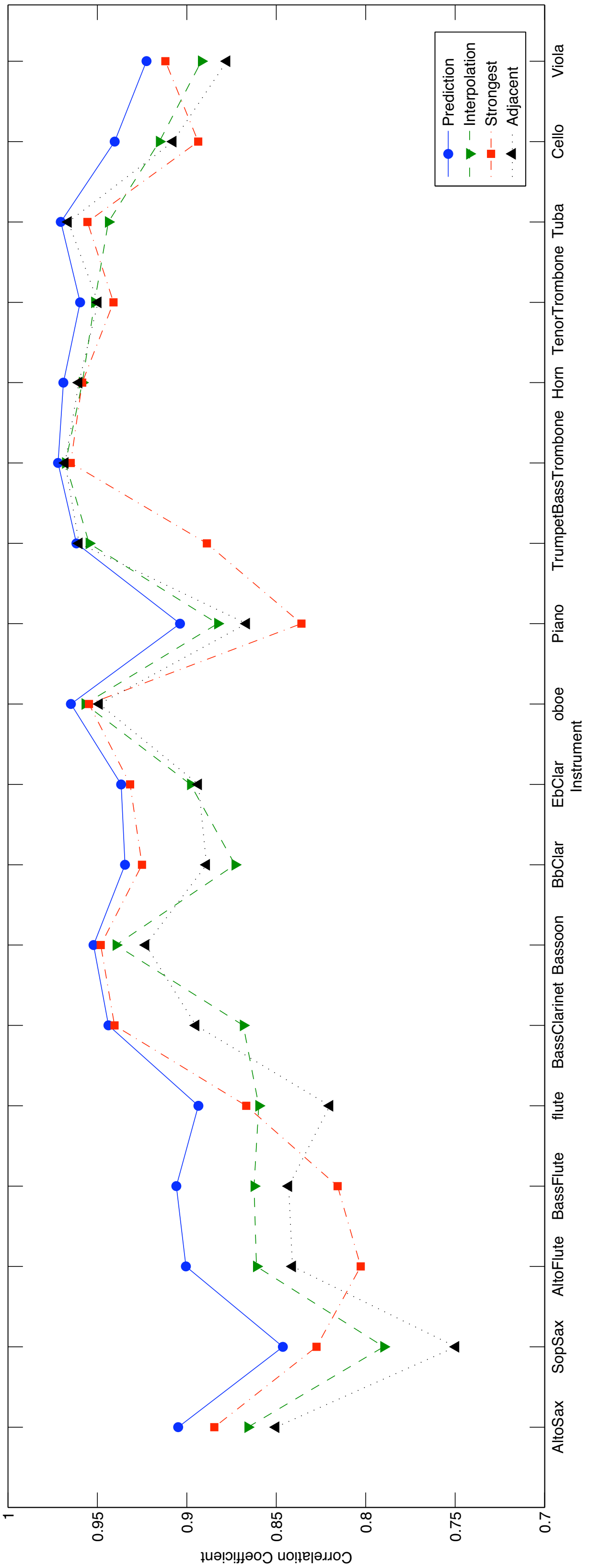

Figure 6.5: Correlation comparison of harmonic magnitude track estimation methods with the true harmonic tracks as a function of musical instrument. While the performance of the spectral and temporal methods are instrument dependent, the proposed spectro-temporal method consistently provides the best estimate.

a spectrally linear-interpolated track. Scale factors for all other temporally-based approaches, were also calculated in a similar manner, and $\delta_{toverlap} = 0.8$.

The root mean square error (RMSE) was calculated between each harmonic track and the corresponding estimate, and the errors were averaged over the mixtures pertaining to a particular polyphony. The results are illustrated in Figure 6.6, highlighting the performance of the prediction method relative to the previously described spectral interpolation, strongest harmonic and adjacent harmonic methods. The interpolation method produced comparatively poorer results because it does not leverage the non-overlapping temporal information. This resulted in large estimation errors, primarily due to scaling. The methods which leverage the temporal information however, produce more accurate estimates particularly when a sufficient proportion of the track is not overlapped. At lower polyphonies, the performance of the temporal approaches are similar to the prediction methods, however at higher polyphonies, the relative errors between the prediction method and the other methods clearly widen. As the polyphony increases, the number of overlapping harmonic tracks in the mixtures also increase, thereby reducing the number of non-overlapping tracks from which estimates can be made. Since the temporal approaches rely on a single harmonic magnitude track, the reduction of non-overlapping harmonics in each mixture escalates the errors relative to the prediction method, which obtains its estimate from a weighted combination of tracks.

6.4.2 Source Synthesis Results

The performance of the Multiple Input Spectrogram Inversion algorithm for closed-loop source synthesis was evaluated using the Signal-to-Noise Ratio (SNR) distortion metric which is defined as,

$$SNR = 10log_{10}\left(\frac{\sum_{\forall n} x[n]^2}{\sum_{\forall n} \left[x[n] - \hat{x}[n]\right]^2}\right) dB \tag{6.12}$$

where $x[n]$ is the original time-domain source signal, $\hat{x}[n]$ is the estimated time-domain source signal. Randomly selected samples were formed into mixtures containing 2, 3, 4, 5 and 6 instrument samples. The magnitude responses of each source

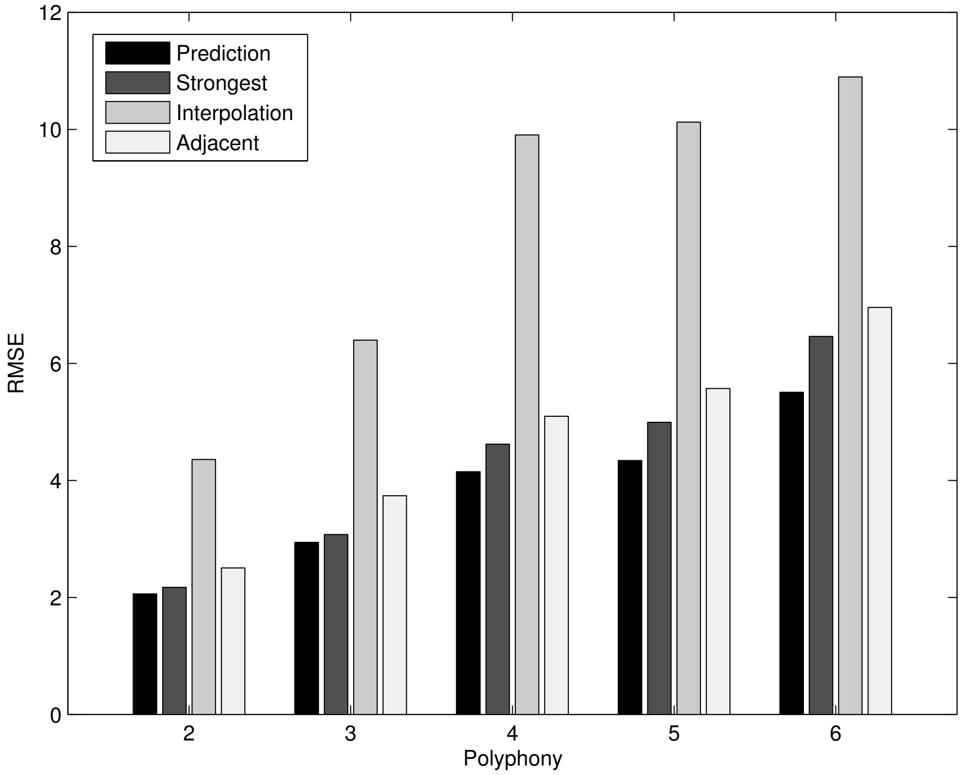

Figure 6.6: Root mean square error (RMSE) performance in polyphonic mixtures, comparing the proposed prediction method, strongest harmonic track, spectral interpolation and adjacent harmonic track. Increasing polyphony is synonymous with increasing number of overlapping tracks.

were then obtained from their respective monophonic samples, and these were used in conjunction with the time-domain mixture signal in the MISI algorithm. The Signal-to-Noise Ratios of the synthesised source estimates obtained from the MISI algorithm, were then calculated and averaged for each polyphony. Figure 6.7 illustrates the SNR performance of the MISI algorithm as a function of the iteration index, for various polyphonies. Due to the manner in which the MISI algorithm is initialised, the first iteration is equivalent to the binary masking of the phase and provides a reference to which the method can be compared. The monotonically increasing SNR values over various polyphonies, indicate that the MISI algorithm provides a closed-loop method of iteratively improving source synthesis estimates given sufficiently accurate magnitude estimates. In mixtures containing 2 sources, the average SNR gains after 50 iterations are approximately 4 dB per source, relative to the binary masking of the phase ($i_t = 1$). As the polyphony increases, the estimation accuracy of the phase binary masked values decrease, since the phase of the mixture contains increasingly more corrupted phase estimates. However after 50 iterations of the MISI algorithm, the relative SNR gains for 6 sources increase by approximately 6 dB per source, to a value which is higher than the average phase binary masking value for 3 sources.

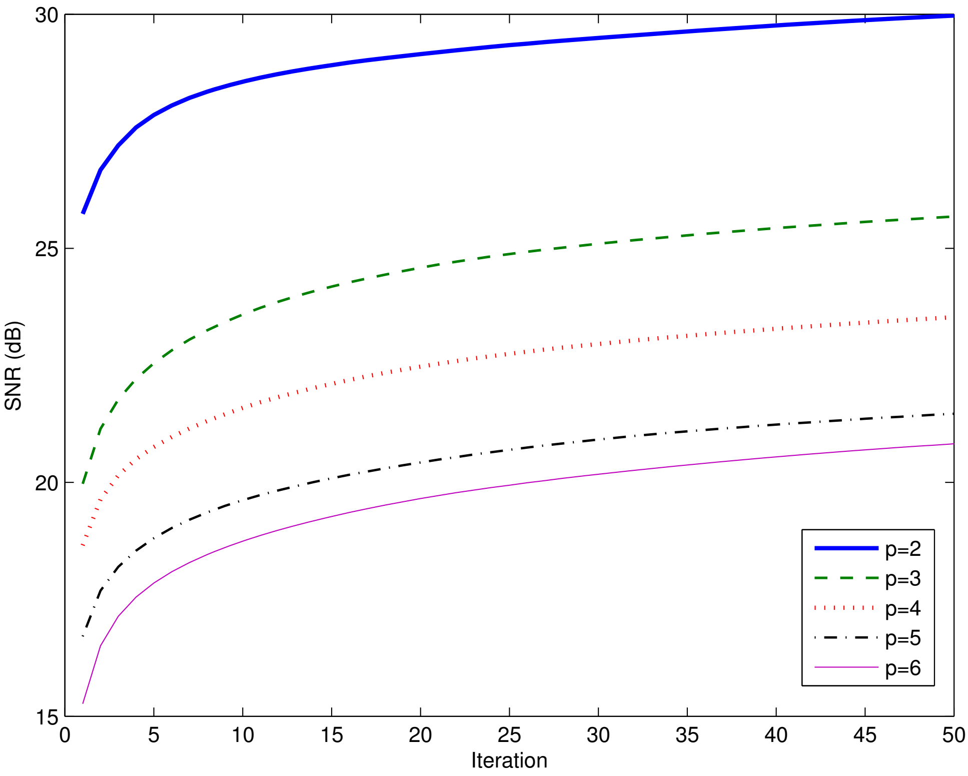

Figure 6.7: Averaged SNR time-domain synthesis performance when using the MISI algorithm given the true magnitude spectra of the sources. The first iteration is equivalent to binary masking of the phase, and the MISI algorithm clearly improves the source estimates with each iteration.

An example of the estimation of a mixture containing 4 sources over 1000 iterations is illustrated in Figure 6.8. The SNR gains after 1000 iterations vary for each source, ranging from 6 to 21 dB. The majority of the gains for sources 2 to 4 occur in the first 50 iterations, whereas the predominant performance gains continue to rise for the first source until the 300th iteration.

Employing the MISI algorithm for source synthesis is clearly advantageous when

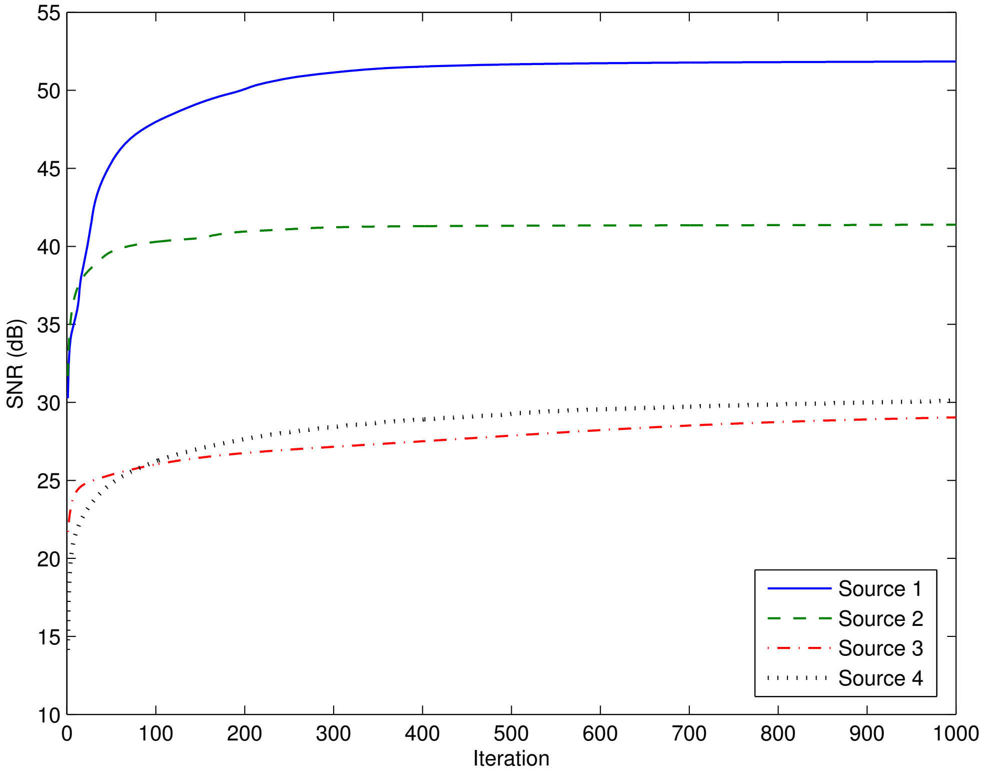

Figure 6.8: SNR improvements of 4 separated sources using the MISI algorithm over 1000 iterations. Asymptotic behaviour plateaus around the 50th iteration for all the sources with the exception of the first source.

the source magnitudes are known, however in the context of a source separation architecture, the magnitudes for each source are often estimated. To simulate the effect of magnitude estimation errors, additive white Gaussian noise (AWGN) was added to each of the source samples and the erroneous magnitudes were fed into the MISI algorithm to observe the performance on source synthesis. Figure 6.9 illustrates the separated source SNRs as a function of the source SNRs due to AWGN, for 20 mixtures containing 4 sources, after 100 iterations of the MISI algorithm. Once again, the first iteration is the phase binary masking baseline, and the results clearly show that given sufficiently accurate estimates, the MISI algorithm provides noticeable improvements. Below a certain source SNR threshold, in this case 13 dB, there is a marginal decrease in performance relative to phase binary masking. Above this value however, the magnitude estimates are sufficiently accurate for the MISI algorithm to converge to the true sources.

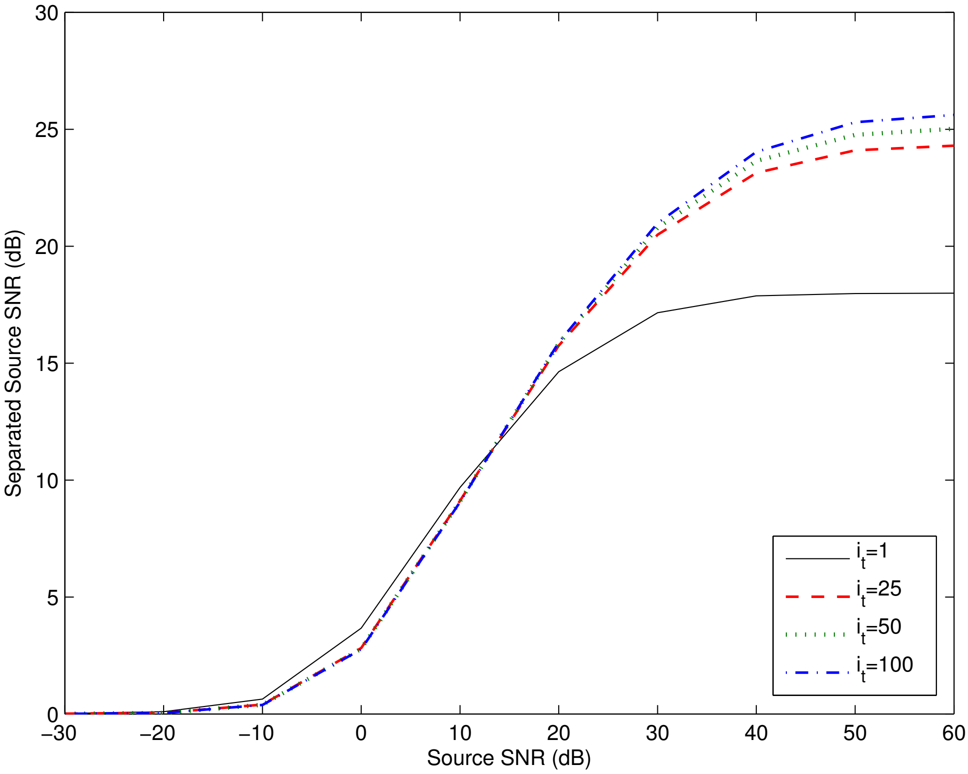

Figure 6.9: Performance of the MISI algorithm with the inclusion of Additive White Gaussian Noise to the source estimates. A threshold determines the point at which the MISI algorithm improves the source estimates.

6.4.3 Separation of 3 sources

To illustrate the potential of the proposed methods, the magnitude track prediction method and the MISI algorithm were placed within the context of a source separation system. A mixture was then separated using the separation system shown in Figure 6.10.

The mixture, comprised of three notes *C4*, *E4* and *G4*, played by the tuba, *Eb* clarinet and alto saxophone, was selected for its perpetual use in Western music and high inter-instrument correlations resulting in a large proportion of overlapping harmonic tracks. To isolate the performance evaluation to that of the proposed methods, ground truth source track estimates were calculated from the original sources. Overlapping tracks were identified using Equation 6.3, and estimated using the magnitude track prediction method. The magnitude spectra of the sources were then estimated

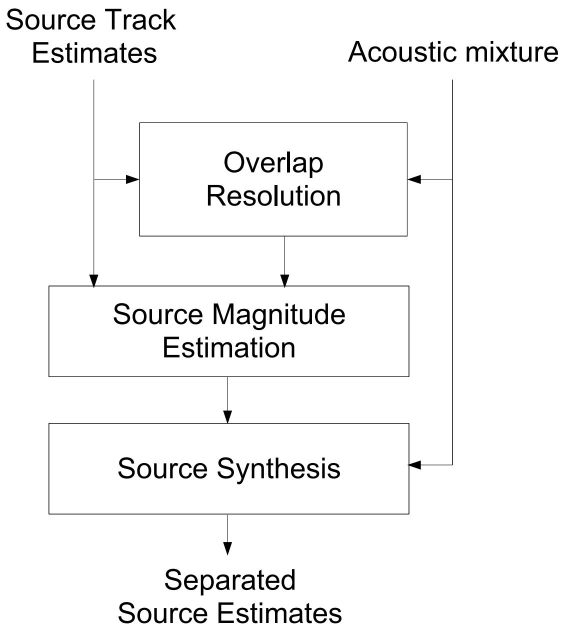

Figure 6.10: System diagram of the separation system for performance evaluation.

using the non-overlapping tracks, the estimated harmonic tracks and the STFT of the mixture. For the non-overlapping tracks, source magnitude estimates were determined using binary masking [129], while the spectral synthesis [102] method was employed for approximating the magnitude spectra of the estimated tracks (see Section 4.2.1). Using the source magnitude estimates and the time-domain mixture signal, the sources were then synthesised using the MISI algorithm. The performance at each iteration of the MISI algorithm was evaluated using the separation measures described in [122], from the *BSS_EVAL* MATLAB toolbox.

The results shown in Figures 6.11, 6.12 and 6.13 correspond to the Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR) and Source-to-Artifacts Ratio (SAR) performance metrics [122]. The results are shown for each source, over 100 iterations of the MISI algorithm. The SDR curves in Figure 6.11, show the distortion reduction of the first and third sources due to the MISI algorithm, increasing the SDR by 1 dB and 5.5 dB respectively, while the SDR of the second source decreased by 1 dB. Since SDR $\approx$ SAR, it can be concluded that the predominant errors are attributed to the signal artifacts which are caused by source magnitude estimation errors. However, since the average SDR and SAR values increase, the source magnitude estimates are adequate, and this is attributed to the effectiveness of the harmonic track prediction method. The errors due to source interference, illustrated in Figure 6.12, clearly confirm the negligible impact of the interference errors. The large SIR values continue to increase with each iteration of MISI algorithm, with the exception of the first source, which decreases after the 18th iteration. This asymptotic decrease is, on average, offset by the significant increase of the third source, which is a compromise the MISI algorithm makes to obtain optimal global error minimisation without implicit knowledge of the sources. At a systemic level, the harmonic track prediction method produced sufficiently accurate estimates from which the MISI algorithm was able to improve the average SIR by 24 dB.

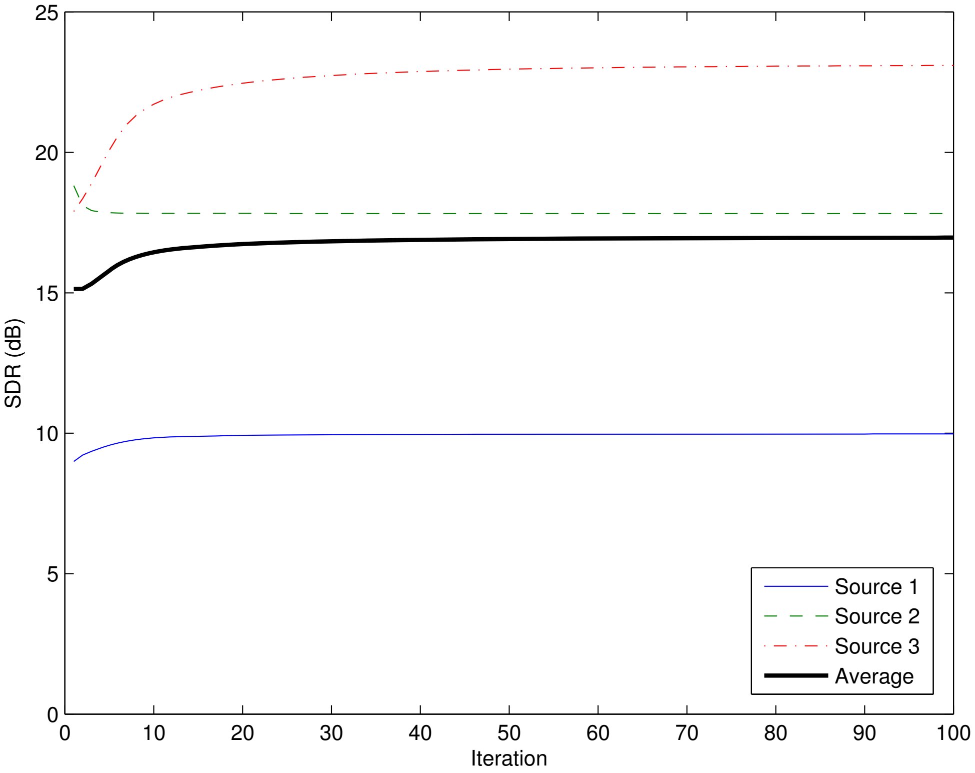

Figure 6.11: Source-to-Distortion Ratio of 3 separated sources as a function of MISI iterations. The magnitude spectra produced by the harmonic magnitude track predictions were sufficiently accurate for the MISI algorithm to improve to average SDR of the sources.

Table 6.1 compares the SDR, SIR and SAR performance, of three approaches with successively increasing levels of ground truth to determine the source magnitude

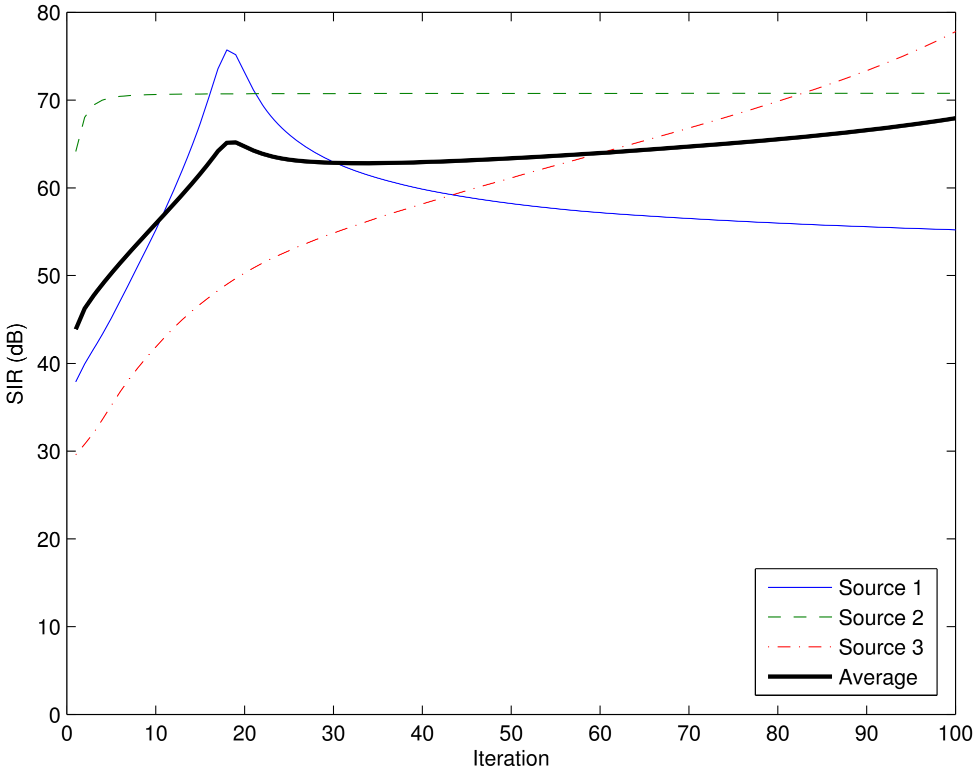

Figure 6.12: Source-to-Interference Ratio of 3 separated sources as a function of MISI iterations. Overall, the MISI algorithm provides drastic reduction of interference errors.

spectra. The first approach employs the prediction method to estimate overlapping tracks, followed by the previously described magnitude estimation. The second approach utilises the ground truth for all the tracks, including the overlapping tracks, followed by magnitude estimation. The final approach uses the ground truth for all the magnitude spectra of the sources. The results show that when using phase binary masking ($i_t = 1$), the harmonic track prediction method produces results that are comparable to that of the ground truth tracks, and the sinusoidal model is a reasonable approximation for estimating the magnitude. After 100 iterations of the MISI algorithm however, the closed-loop estimator is able to capitalise on even small magnitude spectra improvements, producing gains of greater than 10 dB for the SDR of the ground truth magnitude.

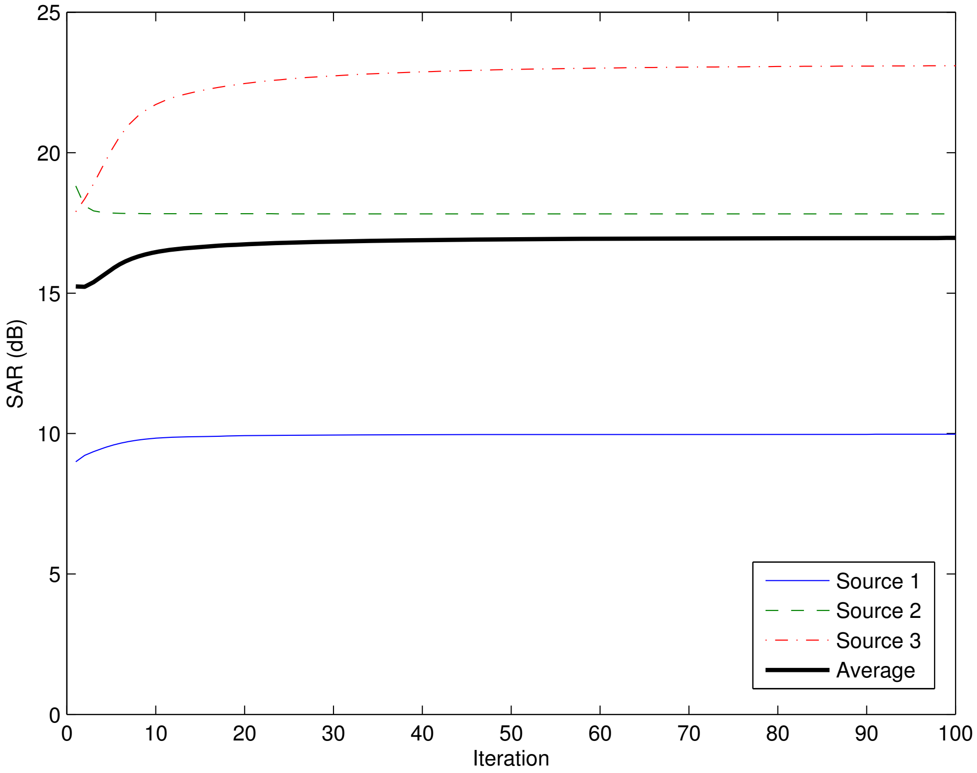

Figure 6.13: Source-to-Artifacts Ratio of 3 separated sources as a function of MISI iterations. The resemblance to the SDR results indicates that the predominant source of error is due to artifacts.

6.4.4 Subjective Evaluation

The results presented in Chapter 5 highlighted the importance of considering the perceptual aspects of source separation. In practise, objective measures provide a means of efficiently evaluating large quantities of separated mixtures, however until a robust objective timbre distortion measure is developed, subjective tests provide the best discriminator of perceptual separation quality for music source separation. In this section we evaluate the subjective quality of the proposed methods in a source separation architecture, using a MUSHRA listening test using MUSHRAM [121].

6.4.4.1 Stimuli

Six musical instrument samples were selected from the University of Iowa musical instrument samples database [1] and are listed in Table 6.2. The duration of the

		SDR (dB)		*SIR* (dB)		*SAR* (dB)	
		$(i_t = 1)$	$(i_t = 100)$	$(i_t = 1)$	$(i_t = 100)$	$(i_t = 1)$	$(i_t = 100)$
1.	*Predicted tracks*	15.14	16.96	43.88	67.93	15.24	16.96
2.	*Ground truth tracks*	15.74	19.49	42.14	70.01	15.9	19.49
3.	*Ground truth magnitude*	16.32	26.50	35.69	60.45	16.56	26.50

Table 6.1: Comparison of the average separation performance of successively increasing levels of ground truth for estimating the magnitude spectra. Performance is evaluated over 100 iterations of the MISI algorithm.

samples were between 1.5-4 s and all of the samples were recorded using 16 bits and a 44100 Hz sampling rate. They were chosen for their representation of different instrument families, thus having different timbres and inharmonicities. The samples are also harmonically related to each other, resulting in a large number of overlapping harmonic tracks.

Instrument #	Instrument	Note	F0 (Hz)
1	Alto Saxophone	C4	261.63
2	Trumpet	E4	329.63
3	Viola	G4	392.00
4	Tenor Trombone	Bb4	466.16
5	Bb Clarinet	D5	587.33
6	Cello	F5	698.46

Table 6.2: Musical instrument samples

Eight mixtures with polyphonies varying between 1-6, listed in Table 6.3, were created to test the subjective quality of the proposed methods in the context of a source separation architecture. The mixtures were separated using an automatic separation framework supplied with the note fundamental frequencies, and note onsets and offsets. A harmonic tracking algorithm based on [34] was used to form harmonic tracks from which overlapping harmonics were resolved using estimates from the harmonic magnitude track prediction method. Source magnitude spectra were then estimated using binary masking for non-overlapping regions and sinusoidal synthesis for overlapping regions, and the synthesised time-domain estimates were obtained after 20 iterations of the MISI algorithm.

For comparative purposes, the overlapping harmonics were also resolved using spectral interpolation and no magnitude track prediction, and the time-domain sep-

Mix #	Polyphony	Instrument # in mix
1	1	1
2	1	2
3	1	3
4	2	1, 2
5	2	3, 4
6	3	1, 2, 3
7	4	1, 2, 3, 4
8	5	1, 2, 3, 4, 5
9	6	1, 2, 3, 4, 5, 6

Table 6.3: Musical instrument mixtures

arated sources were also synthesised using binary masking of the mixed phase spectrum. Three algorithms were evaluated thus used in the evaluation shown in Table 6.4.

Algorithm #	Overlap resolution	Synthesis
1	Predicted Tracks	MISI
2	Spectral Interpolation	Binary Masking
3	None	Binary Masking

Table 6.4: The 3 algorithms evaluated in the subjective tests, with differing methods for overlap resolution and time-domain synthesis.

The resulting separated sounds from instruments 1, 2 and 3 from each polyphony were then evaluated in the subjective tests. Thus the stimuli comprised of 54 sounds from 3 instruments, 6 polyphonies and 3 separation algorithms.

6.4.4.2 Procedure

Five listeners aged between 22 and 27 participated in the experiment, 3 of which had musical training with experience ranging between 5-10 years. The stimuli presentation was controlled using MATLAB on a PC with a MOTU 828mkII audio interface. The levels of the stimuli were adjusted to have equal A-weighted levels and were presented monaurally through Beyerdynamic DT 770 headphones, at a comfortable listening level in a quiet, acoustically treated recording studio.

A MUSHRA paradigm was used for the experiment which comprised of 18 trials. Each trial consisted of a reference sound and 3 test sounds whose quality was to be rated against the reference sound. The reference sounds were the original, unaltered sounds from [1] that were used in creating the mixtures found in Table 6.3. The order of the 4 test sounds were randomised, and comprised of 3 stimuli from the 3 different algorithms at the same polyphony and 1 anchor sound which was the same as the reference sound.

Each participant was briefed on the nature of the experiment, and was told that they had the task of grading the quality of each of the test sounds between 0-100 with respect to the reference sounds. A test sound that was perceptually identical to the reference sound was given a score of 100 and test sounds that deviated significantly from the reference sounds were given lower scores. In the training phase, each participant then listened to all the sounds that they would have to grade, paying attention to the varying levels of distortion between the test sounds and the reference sounds. Once the participants had familiarised themselves with the stimuli, they then proceeded to the evaluation phase. In the evaluation phase, participants were presented with an interface that allowed them to play the reference sound and the test sounds. The interface also had 5 sliders which allowed the participants to grade the tests sounds between 0-100. Once they had graded each of the test sounds in the trial, they then clicked a "Save and proceed" button and continued to grade the tests sounds of the next trial until all the trials were complete.

6.4.4.3 Results

The scores for the stimuli were averaged over all participants and stimuli of the same polyphony and algorithm were then also averaged. These results are illustrated in Figure 6.14.

The results indicate that Algorithm 1, containing both of the proposed methods, produced the least perceptual distortions for all polyphonies. Consistent with the other objective measures, the performance of all 3 algorithms perceptually degraded as the polyphony increased. Algorithm 3 provided the baseline and expectedly produced the most audibly distorted sounds in general, since no harmonic track resolution method was used. However, for a polyphony of 6, the mis-estimations resulting from spectral interpolation in Algorithm 2, proved to be perceptually more detrimental than not estimating the overlapped harmonics at all. The overall quality assessments are generally consistent with the previous results, particularly with the individual instrument correlation results in Figure 6.5. The differences in performance can be explained by the instruments added at each stage, and the predictability of their harmonic magnitude tracks. For example, in the case of the two source mixture, the addition of the trumpet, which is well predicted by both spectral interpolation and the track prediction method, resulted in perceptually better separation compared to Algorithm 3. When the Bb clarinet was added into the 5 source mixture, the poor estimation of the spectral interpolation method resulted in a drop in quality compared to the higher quality prediction offered by the track prediction method.

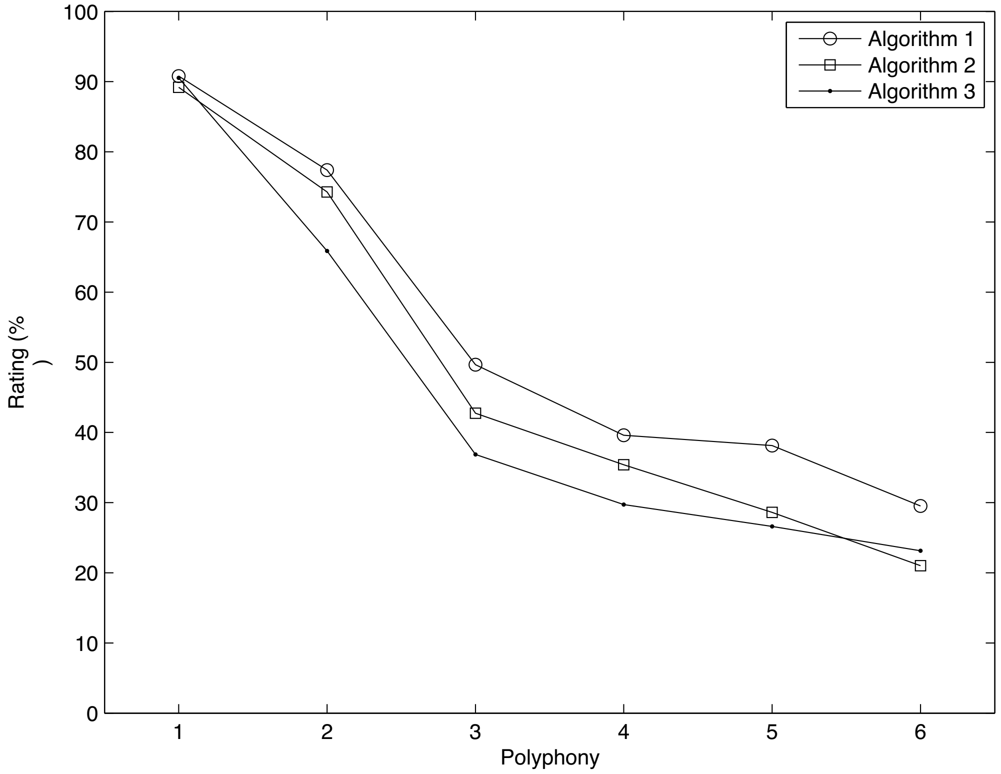

Figure 6.14: Participant scores for the subjective quality assessment. The figure shows the averaged scores from the participants, for sounds 1, 2 and 3 for each polyphony.

6.5 Discussion

The results presented in Section 6.4.1 highlight the gains achieved by exploiting both the spectral and temporal harmonic information in the estimation of overlapping harmonic magnitude tracks. In the evaluations comparing the spectral, temporal and the proposed spectro-temporal approach, the benefits of each method were highlighted. Spectral approaches provide good information about the general shape of the temporal trajectory, and this is due to the inherent averaging of the magnitude tracks. In the case of spectral linear interpolation, the magnitude tracks are predicted to be the average of the adjacent tracks, and this averaging produces results which reduce the variability of the estimated shape, as seen in correlation coefficients of Figure 6.4. In addition to the shape information, the spectral approach also provides scale information, which is obtained in the interpolation process. This scale information is reasonably accurate for certain musical instruments which have a smoother spectral envelope (see Figure 6.5), however for other instruments, this scale information can lead to grossly inaccurate estimation. This is particularly evident in the RMSE results in Figure 6.6, where the spectral interpolation method produced the lowest performance. By contrast, the temporal approaches performed well in the track estimation for various polyphonic mixtures, and this is primarily attributed to the ability to estimate the scale of harmonic tracks from non-overlapping regions. While temporal approaches have this over their spectral counterparts, their performance relies on a single harmonic track, such as the strongest harmonic track [130] or the adjacent harmonic track [127]. As Figure 6.5 illustrates, there are certain instruments for which this produces good results, but there are also instruments for which such an assumption does not hold. The solution is therefore to combine the merits of both the spectral and temporal approaches, creating an estimator which is able to exploit spectral averaging to determine trajectory shape, as well as use non-overlapping temporal regions to determine trajectory scale. The harmonic magnitude track prediction method does both, by linearly combining neighbouring harmonic tracks to produce a trajectory shape estimate, and exploiting non-overlapping regions to estimate the scale of the track. The results verify the combinatorial performance benefits. In general shape estimation (Figure 6.4), the magnitude track predictor achieved the highest correlation coefficient scores with the least amount of variability relative to the other methods, producing the best results consistently over all the instruments tested (Figure 6.5). The RMSE results for track estimation in harmonic mixtures also emphasise the merits of the prediction method, surpassing the interpolation method due to the estimation of scale, and bettering the temporal methods due to the estimation of shape. In the broader context of source separation, the magnitude track prediction method provides a spectro-temporally derived estimate of harmonic magnitude tracks which can be employed to approximate source magnitude spectra estimates using methods such as filtering [34], partial demixing [127] or least-squares [130].

Given these source magnitude spectra, it is then possible to synthesise the sources. Besides sinusoidal synthesis based techniques, the majority of source synthesis is heavily reliant on the phase spectrum of the mixture. This becomes increasingly more problematic as the polyphony of the mixtures increase, since it results in an increasing number of overlapping harmonics and thus an increasing amount of corrupted phase information. The MISI algorithm provides a closed-loop approach to the synthesis of sources, minimising the error between the collective source estimates and the time-domain mixture. Given the true magnitude spectra of the sources, the MISI algorithm is able to substantially improve on phase binary masking, providing increasing performance gains with increasing polyphony (Figure 6.7). The performance benefits vary from source to source, dependent on the proportion of overlap, with source synthesis SNRs ranging from 6 dB to 21 dB in the example presented in Figure 6.8. When presented with erroneous estimates of the source magnitude spectra, the performance of the MISI algorithm is dependent on the estimation error. When the source magnitude estimates are below a certain threshold, the performance of the MISI algorithm suffers a minor degradation relative to phase binary masking. However, when the source magnitude estimates are above a certain error threshold, the MISI algorithm produces performance gains significantly larger than the losses incurred when below the threshold (Figure 6.9). Overall, the MISI algorithm highlights the merits of a closed-loop synthesis algorithm, which contrasts with the predominant use of open-loop algorithms for synthesising sources in sinusoid-based separation systems. The iterative phase estimation of the MISI algorithm produces significant gains, minimising the synthesis errors with respect to the time-domain mixture, given sufficiently accurate source magnitude spectra.

Embedding the proposed methods in a separation architecture in Section 6.4.3, highlighted the importance of precision estimation at every stage of the separation process. The harmonic magnitude track prediction model provides a robust means of estimating sinusoidal magnitude tracks with high precision, performing similarly to the ground truth of the tracks (at $i_t = 1$). After 100 iterations of the MISI algorithm however, the additional improvements to the magnitude estimation become increasingly more significant. Future investigations should focus on utilising the MISI algorithm with magnitude spectra estimated as a combination of sinusoidal and non-sinusoidal components. While the sinusoidal model accounts for most of the salient energy in harmonic instrument signals, the additional gains achieved by estimating non-sinusoidal information appear to be significant with regard to the MISI algorithm.

The subjective tests in Section 6.4.4, offered insight into the overall performance of the proposed methods from a perceptual perspective. The participant assessments were, on the whole, consistent with the objective results, suggesting that the separation quality was dependent on the proportion of overlapped harmonics in the mixtures. Furthermore, the results were also related to the instrument type and the track estimation method employed. The robust estimation of the harmonic magnitude track prediction method across a variety of instruments, resulted in perceptually superior performance across the various mixtures.

6.6 Conclusion

Within the context of source separation systems, two particularly challenging issues are the resolution of overlapping harmonics and source synthesis. To address the issue of overlapping harmonics, a model was presented which exploited the spectro-temporal correlations of harmonic magnitude tracks. The benefits of combining the spectral and temporal information were verified in a performance evaluation, where the prediction method was shown to be superior in all regards.

The Multiple Input Spectrogram Inversion (MISI) algorithm was also presented, which iteratively estimates time-domain sources given source magnitude spectra and the time-domain mixture. Given sufficiently accurate magnitude spectra, the closed-loop architecture of the algorithm effectively minimised the source phase errors, resulting in significant improvements in the Signal-to-Noise Ratio of synthesised sources.

Chapter 7

Conclusions

7.1 Conclusions

This thesis has presented several methods related to the separation of musical sources in polyphonic mixtures. It has specifically focused on the separation of pitched sounds, and methods by which the harmonic component can be successfully segregated from a mixture. Experiments were also conducted exploring perceptual sensitivity to timbre to aid in the formulation of an objective distortion metric for timbre.

7.1.1 Source Separation

The source separation of musical mixtures is a complex problem that involves the collaboration of several mechanisms. In this thesis, a signal model-based separation paradigm was employed. Using a sinusoidal parametrisation to model the harmonics of pitched sounds, the harmonic peaks were identified and the refinement of parameters was accomplished using the PDFFT. This novel, computationally efficient estimation algorithm presented in Section 3.2, computes highly accurate estimates of sinusoid frequencies using coarse frequency estimates provided from the FFT and the time derivative of the phase response. For single sinusoid frequency estimation, the accuracy of the PDFFT outperformed the frequently employed QIFFT, even with zero-padding, at the expense of only 4 multiplies per peak. Unlike other interpolation methods, the PDFFT was shown to also perform well at resolving multiple frequencies from a single peak in the magnitude spectrum.

Due to the inherent properties of musical mixtures (discussed in Section 2.1.2), overlapping harmonics are a prevalent issue in music source separation. To assist in the resolution of these harmonics, a quantitative investigation into the nature of the spectro-temporal correlations between the harmonic tracks of 3000 instrument samples was conducted. It was found that the highest similarity was between directly adjacent harmonics, with similarity decreasing exponentially as a function of distance from the harmonic location. Using regularised least squares, modelled weighting functions were then derived for the prediction of harmonic tracks as a linear combination of neighbouring tracks. The modelled weighting functions were then used to construct model harmonic tracks from which ambiguous harmonics in polyphonic mixtures were identified (Section 3.3.4).

Obtaining parametric estimates of the sources of a mixture is only a portion of the complete source separation objective. The parametric data must then be synthesised in the time-domain (Section 4.1.1). Using estimates of the source magnitudes, a novel multiple input spectrogram inversion (MISI) algorithm was presented to iteratively estimate the phase response of each source. The sum of the source estimates is minimised with respect to the mixture signal, and average RMS errors less than 10^{-3} are achieved for mixtures of 6 sources given accurate source magnitude spectra. The precision of the phase estimation was found to be governed by the fidelity of the magnitude spectra and given accurate magnitudes, the MISI algorithm produced significant synthesis improvements over existing methods.

In Chapter 6, a novel harmonic magnitude track prediction method was introduced, based on the spectro-temporal correlation findings of Section 3.3. This was combined with the closed-loop synthesis approach of the MISI algorithm, to offer solutions to the issues of overlapping harmonics and source re-synthesis for music mixtures. Performance evaluations of the individual methods were conducted exclusively, as well as in the context of a source separation system. The results of the objective and subjective evaluations revealed that the proposed methods provided substantial advancements over existing approaches over a wide variety of instruments and polyphonies.

7.1.2 Exploration of Timbre

The experiments presented in Chapter 5 explored the perceptual sensitivity of timbre. In the first experiment, sensitivity to the modification of the spectral envelopes of musical instruments was investigated. By observing the just-noticeable attenuation levels of several frequency bands, spectral envelope sensitivity was shown to vary as a function of frequency and bandwidth. The first few harmonics were generally found to govern the overall sensitivity, with sensitivity decreasing for higher frequencies and wider bandwidths.

The second experiment presented novel methods for investigating the spectral envelope space. Trumpet and clarinet sounds were morphed using 7 different linear-logarithmic permutations, to explore classification thresholds. Psychometric functions were approximated for the subjective results and classification thresholds were compared to spectrum parametrisations, revealing that the spectral envelope space is best described using logarithmic frequency resolution. Additional masking analysis revealed psychoacoustic masking models should be considered in the development of timbre space models.

The third experiment investigated the perceptual sensitivity to the noise component of pitched sounds. Musical instrument samples were separated into harmonic and noise components, and the sensitivity to the attenuation of several frequency bands of the noise components were measured. Noise sensitivity was generally lower compared to spectral envelope sensitivity, with the sensitivity varying for different instruments. The low frequency harmonics were effective maskers of noise and maximum sensitivity was found to be around 6-11 kHz.

7.2 Future Work

7.2.1 Source separation utilising broader temporal context

The separation methods presented in this thesis primarily address the separation of pitched sounds occurring simultaneously. This is one of the more difficult problems and instrument anonymity adds further complications. While the sounds considered throughout this research were recordings of acoustic instruments, they were considered in isolation. Music signals typically consist of several instruments playing a series of notes. The majority of these notes are played simultaneously with other notes, but there may be occasions when notes are played in isolation or they dominate the mixture for a particular time period. In these time periods, information about the nature of the tones may be used to construct tone models that may aid in the separation of other sounds played simultaneously elsewhere in the mixture. This is an example of integrating broader temporal context into separation systems - an area which has yet to be explored in great detail.

There is also a wealth of contextual information that has yet to be harnessed by incorporating musicological models into separation systems. These models can be used to predict the probabilities of the properties of future notes, such as their onset time and their fundamental frequency. For example, most music pieces have a particular musical “key” they are played in, increasing the probability of notes being played at a certain F0. “Time signatures” also guide the underlying rhythmic structure of the piece, increasing the probability of notes being played at certain times. Utilising this sort of information in separation architectures will be beneficial, particularly in the analysis stages of the system.

Music signals have a vast amount of useful information that can be gleaned by considering broader temporal context and future research should investigate these avenues.

7.2.2 Development of Objective Distortion Metric for Timbre

Several attempts have been made to develop a deeper understanding of the timbre space [42, 43, 17, 50, 47, 49], however the information has yet to be consolidated into an objective distortion metric. The research on timbre is generally converging towards a consensus of the “axes” that define the multidimensional nature of timbre, and these relate to the spectral and temporal variations of musical sounds.

Developing a robust distortion metric for timbre will require further investigation into the perceptual sensitivity of temporal variations, temporal and frequency parametrisations that adequately map to perception, and an exploration into the benefits of incorporating psychoacoustic masking models.

Bibliography

[1] The University of Iowa musical instrument samples. http://theremin.music.uiowa.edu/, Date last viewed 24/9/2008. 50, 55, 70, 79, 93, 103, 115, 120, 132, 135

[2] T. J. Abatzoglou. A fast maximum likelihood algorithm for frequency estimation for a sinusoid based on newton's method. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(1):77–89, 1985. 38

[3] S.A. Abdallah and M.D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. *International Conference on Music Information Retrieval*, 2004. 6, 110

[4] M. Abe and J.O. Smith III. Design criteria for the quadratically interpolated fft method (i): Bias due to interpolation. Technical report, Technical Report STAN-M-114, Dept. of Music, Stanford University, August 2004. 31, 44

[5] S. S. Abeysekera. An efficient hilbert transform interpolation algorithm for peak position estimation. In *Proc. IEEE Signal Processing Workshop on Statistical Signal Processing*, pages 417–420, 2001. 38

[6] J. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238, 1977. 21

[7] L.D. Alsteris and K.K. Paliwal. Iterative reconstruction of speech from short-time fourier transform phase and magnitude spectra. *Computer Speech & Language*, 21(1):174–186, 2007. 66, 119

[8] ANSI. American national standard psychoacoustical terminology. *American National Standards Institute, New York*, 1973. 75

[9] ANSI. American national standard acoustical terminology. *American National Standards Institute, New York*, 1994. 15

[10] H. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253, 2001. 6

[11] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995. 6

[12] A.J. Bell and T.J. Sejnowski. The “independent components” of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997. 5

[13] K.W. Berger. Some factors in the recognition of timbre. *J. Acoust. Soc. Am.*, 36(10):1888–1891, 1964. 76

[14] A.S. Bregman. *Auditory scene analysis.* MIT Press Cambridge, 1990. 3

[15] G. J. Brown. *Computational auditory scene analysis: A representational approach.* PhD thesis, University of Sheffield, 1992. 5

[16] J.C. Brown. Calculation of a constant q spectral transform. *J. Acoust. Soc. Am*, 89(1):425–434, 1991. 20

[17] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *J. Acoust. Soc. Am.*, 118:471–482, 2005. 76, 77, 97, 144

[18] M.A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proc. International Computer Music Conference*, 2000. 110

[19] C. Chafe, K. Kashima, B. Mont-Reynaud, and J. Smith. Source separation and note identification in polyphonic music. Technical report, Stanford University, Report STAN-M-29, 1985. 8

[20] P. Comon et al. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. 6

[21] P. Cook. Tbone: An interactive waveguide brass instrument synthesis workbench for the next machine. *Proceedings of the 1991 International Computer Music Conference, Montreal*, pages 297–299, 1991. 29

[22] M. P. Cooke. *Modelling auditory processing and organisation.* PhD thesis, University of Sheffield, 1991. 5

[23] R. Crochiere. A weighted overlap-add method of short-time fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):99–102, 1980. 22

[24] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(4):357–366, 1980. 97

[25] M. Davy and SJ Godsill. Bayesian harmonic models for musical signal analysis. *Bayesian Statistics*, 7, 2003. 7, 8

[26] P. Depalle, G. Garcia, X. Rodet, and P. IRCAM. Tracking of partials for additive sound synthesis using hidden markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993. 31, 49

[27] P. Depalle, T. Helie, and P. IRCAM. Extraction of spectral peak parameters using a short-time fourier transform modeling and no sidelobe windows. *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997. 31, 38

[28] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995. 28

[29] D. Deutsch, editor. *The Psychology of Music.* Academic Press, 1999. 1, 103

[30] R. Duda, R. Lyon, and M. Slaney. Correlograms and the separation of sounds. In *Proc. IEEE Asilomar conf. on sigs., sys. and computers*, 1990. 5, 20

[31] D. Ellis and D. Rosenthal. Mid-level representations for computational auditory scene analysis. In *Proc. Int. Joint Conference on Artificial Intelligence Workshop on Computational Auditory Scene Analysis*, August 1995. 20

[32] D.P.W. Ellis. *Prediction-driven computational auditory scene analysis.* PhD thesis, Massachusetts Institute of Technology, 1996. 5

[33] M.R. Every. *Separation of musical sources and structure from single-channel polyphonic recordings.* PhD thesis, University of York, 2006. 60, 63

[34] M.R. Every and J.E. Szymanski. Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Trans. on Audio, Speech, Lang. Process.*, 14(5):1845–1856, 2006. 7, 20, 31, 35, 36, 48, 49, 60, 66, 110, 111, 114, 117, 121, 133, 138

[35] D. FitzGerald. *Automatic Drum Transcription and Source Separation.* PhD thesis, Dublin Institute of Technology, 2004. 10

[36] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12(1):47–65, 1940. 26

[37] A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming: musical information retrieval in an audio database. *Proc. ACM International conference on Multimedia*, 1995. 9

[38] M. Goodwin. Residual modeling in music analysis-synthesis. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996. 32, 37

[39] M. Goto. A predominant-f0 estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001. 7

[40] M. Goto and Y. Muraoka. A sound source separation system for percussion instruments. *Trans. Institute of Electronics, Information and Communication Engineers*, 77(5):901–911, 1994. 10

[41] David M. Green and Christine R. Mason. Auditory profile analysis: Frequency, phase, and weber's law. *J. Acoust. Soc. Am.*, 77:1155–1161, 1985. 78

[42] J. M. Grey. *An exploration of musical timbre.* PhD thesis, Stanford University, 1975. 76, 93, 103, 144

[43] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *J. Acoust. Soc. Am.*, 61:1270–1277, 1977. 76, 77, 78, 97, 144

[44] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(2):236–243, 1984. 66, 67, 118, 119

[45] D. Griffin and J. Lim. A new model-based speech analysis/synthesis system. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1985. 31, 35

[46] D. Gunawan and D. Sen. Sinusoidal frequency estimation based on the time derivative of the stft phase response. *Proc. Int. Conf. on Information, Communications and Signal Processing, Bangkok*, pages 1452–1456, 2005. 31, 38, 114

[47] D. Gunawan and D. Sen. An exploration of the spectral envelope space of musical instruments using envelope morphing permutation strategies. *Association of Research in Otolaryngology (ARO) Convention, Denver*, 2007. 76, 144

[48] D. Gunawan and D. Sen. Identification of partials in polyphonic mixtures based on temporal envelope similarity. *123rd Audio Engineering Society Convention, New York*, October 2007. 13, 115

[49] D. Gunawan and D. Sen. Sensitivity to musical instrument noise in harmonics plus noise modelling. *International Conference on Music Communication Science, Sydney*, 2007. 76, 144

[50] D. Gunawan and D. Sen. Spectral envelope sensitivity of musical instrument sounds. *J. Acoust. Soc. Am.*, 123(1):500–506, 2008. 25, 76, 144

[51] H. von Helmholtz. *On the Sensation of Tone.* New York: Dover, 1954. 17, 28, 75

[52] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003. 29

[53] A. Horner, J. Beauchamp, and R. So. Detection of random alterations to time-varying musical instrument spectra. *J. Acoust. Soc. Am.*, 116:1800–1810, 2004. 77, 78, 79, 83, 85, 86, 92, 103

[54] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. 6

[55] J. Johnston, S. Quackenbush, G. Davidson, K. Brandenburg, and J. Herre. MPEG audio coding. In A. Akansu and M. Medley, editors, *Wavelet, Subband, and Block Transforms in Communications and Multimedia.* Kluwer Academic, 1999. 90, 99

[56] M. Karjalainen, J. Backman, and J. Polkki. Analysis, modeling, and real-time sound synthesis of the kantele, a traditional finnish string instrument. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993. 29

[57] M. Karjalainen, UK Laine, TI Laakso, and V. Valimaki. Transmission-line modeling and real-time synthesis of string and wind instruments. *Proceedings of the 1991 International Computer Music Conference, Montreal*, pages 293–296. 29

[58] K. Kashino and S.J. Godsill. Bayesian estimation of simultaneous musical notes based on frequency domain modelling. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004. 20

[59] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism. *Proc. International Joint Conf. on Artificial Intelligence*, 1995. 5

[60] K. Kashino and H. Tanaka. A sound source separation system with the ability of automatic tone modeling. *Proc. International Computer Music Conference*, pages 248–255, 1993. 5

[61] C. Kayser, C.I. Petkov, M. Lippert, and N.K. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005. 28

[62] A. Klapuri. Number theoretical means of resolving a mixture of several harmonic sounds. In *Proc. European Signal Processing Conference*, 1998. 18

[63] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. *Proc. IEEE. Int. Conf. on Acoustics, Speech and Signal Processing*, 1999. 8

[64] A. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–816, 2003. 7, 20, 38, 44

[65] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, 2004. 7, 8

[66] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. *International Conference on Music Information Retrieval, Victoria, Canada*, October 2006. 7, 29

[67] A.P. Klapuri. Multipitch estimation and sound separation by the spectral smoothness principle. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001. 111

[68] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4:219–227, 1985. 28

[69] L. Biscainho L. Nunes, R. Merched. Recursive least-squares estimation of the evolution of partials in sinusoidal analysis. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2007. 31, 49

[70] M. Lagrange, S. Marchand, and J.B. Rault. Using linear prediction to enhance the tracking of partials. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 4, pages 241–244, 2004. 49

[71] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001. 6

[72] S.N. Levine. *Audio representations for data compression and compressed domain processing.* PhD thesis, Stanford University, 1999. 30

[73] H. Levitt. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49:467–477, 1971. 82, 86, 105

[74] D.A. Luce. *Physical correlates of nonpercussive musical instrument tones.* PhD thesis, Massachusetts Institute of Technology, 1963. 76

[75] M. D. Macleod. Fast nearly ml estimation of the parameters of real or complex single tones or resolved multiple tones. In *IEEE Trans. Signal Processing*, volume 46, pages 141–148, 1998. 31, 38

[76] R.C. Maher and J.W. Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *J. Acoust. Soc. Am*, 95(4):2254–2263, 1994. 30

[77] S. Makeig, A.J. Bell, T.P. Jung, T.J. Sejnowski, et al. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8:145–151, 1996. 5

[78] K.D. Martin. Automatic transcription of simple polyphonic music: Robust front end processing. *Massachusetts Institute of Technology Media Laboratory Perceptual Computing Section Technical Report*, (399), 1996. 8

[79] L.G. Martins, J.J. Burred, G. Tzanetakis, and M. Lagrange. Polyphonic instrument recognition using spectral clustering. In *Proc. International Conference on Music Information Retrieval*, 2007. 29

[80] S. McAdams, J. W. Beauchamp, and S. Meneguzzi. Discrimination of musical instrument sounds resynthezied with simplified spectrotemporal parameters. *J. Acoust. Soc. Am.*, 105:882–897, 1999. 76, 77, 79, 97, 103

[81] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Process.*, 34(4):744–754, 1986. 7, 29, 30, 31, 35, 37, 48, 49, 59, 66, 110, 112, 113, 114

[82] R. Meddis and L. O'Mard. A unitary model of pitch perception. *J. Acoust. Soc. Am*, 102(3):1811–1820, 1997. 34

[83] D.K. Mellinger. *Event formation and separation in musical sound.* PhD thesis, Stanford University, 1991. 5

[84] B. C. J. Moore. *Introduction to the Psychology of Hearing.* Macmillan, London, 1977. 80

[85] BCJ Moore and BR Glasberg. A revision of zwicker's loudness model. *Acustica*, 82(2):335–345, 1996. 26, 27

[86] K. K. Paliwal and B. S. Atal. Efficient vector quantization of LPC parameters at 24 Bits/Frame. *IEEE Trans. Speech and Audio Processing*, 1:3–14, 1993. 77, 78, 83, 85, 88, 92

[87] T.W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60:911–918, 1976. 7, 111, 121

[88] RD Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. *Auditory Physiology and Perception*, pages 429–446, 1992. 27

[89] D. B. Paul. The spectral envelope estimation vocoder. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29:786–794, 1981. 76, 77, 88, 93

[90] R. Plomp. Timbre as a multidimensional attribute of complex tones. In R. Plomp and G. F. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing.* Sijthoff, Leiden, 1970. 76, 77

[91] R. Plomp and W.J.M. Levelt. Tonal consonance and critical bandwidth. *J. Acoust. Soc. Am*, 38:548–560, 1965. 18

[92] B. G. Quinn. Estimation of frequency, amplitude, and phase from the dft of a time series. *IEEE Trans. Signal Processing*, 45(3):814–817, 1997. 38

[93] E. Rank and G. Kubin. A waveguide model for slapbass synthesis. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 443–446, 1997. 29

[94] D. C. Rife and R. R. Boorstyn. Single-tone parameter estimation from discrete-time observations. In *IEEETrans. Info. Theory*, volume 20, pages 591–598, 1974. 31, 38

[95] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8(4):14–38, 1991. 23

[96] T.D. Rossing. *The Science of Sound.* Addison-Wesley, 1990. 16

[97] B. Scharf. Critical bands and the loudness of complex sounds near threshold. *J. Acoust. Soc. Am.*, 31:365–370, 1959. 26

[98] E.D. Scheirer. Tempo and beat analysis of acoustic musical signals. *J. Acoust. Soc. Am.*, 103(1):588–601, 1998. 8, 9

[99] JF Schouten. The perception of timbre. *Reports of the 6th International Congress on Acoustics*, 1968. 76

[100] D. Sen and WB Kleijn. Synthesis methods in sinusoidal and waveform-interpolation coders. *Proceedings of the IEEE Workshop on Speech Coding for Telecommunications*, pages 79–80, 1995. 59

[101] X. Serra. *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition.* PhD thesis, Stanford University, 1989. 7, 30, 110, 113

[102] X. Serra, J. Bonada, P. Herrera, and R. Loureiro. Integrating complementary spectral models in the design of a musical synthesizer. In *Proc. International Computer Music Conference*, 1997. 59, 63, 64, 66, 112, 129

[103] X. Serra and J. Smith III. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990. 32, 37

[104] M. Slaney. Auditory toolbox, version 2. *Technical Report No: 1998-010*, 1998. 97

[105] M. Slaney and R.F. Lyon. On the importance of time-a temporal representation of sound. *Visual Representations of Speech Signals*, pages 95–116, 1993. 27

[106] M. Slaney, D. Naar, RE Lyon, A.C. Inc, and CA Cupertino. Auditory model inversion for sound separation. *Proc. IEEE. Int. Conf. on Acoustics, Speech and Signal Processing*, 1994. 20

[107] Malcolm Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. Technical Report 35, Apple Computer, 1993. 27, 81

[108] P. Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach.* PhD thesis, Massachusetts Institute of Technology, 2001. 5, 6

[109] P. Smaragdis. Discovering auditory objects through non-negativity constraints. *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004. 6

[110] P. Smaragdis and JC Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003. 110

[111] J.O. Smith and X. Serra. Parshl: A program for the analysis/synthesis of inharmonic sounds based on a sinusoidal representation. In *Proc. International Computer Music Conference*, 1987. 7, 30, 31, 35, 37, 38, 110, 113

[112] J.O. Smith III. Physical modeling using digital waveguides. *Computer Music Journal*, 16(4):74–91, 1992. 29

[113] E. Terhardt. Calculating virtual pitch. *Hearing Research*, 1:155–182, 1979. 34

[114] DJ Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982. 31, 35, 36

[115] T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Trans. Speech and Audio Processing*, 8(6):708–716, 2000. 7, 8

[116] K. Torkkola. Blind separation for audio signals–are we there yet. *Proc. Int. Workshop on Independent Component Analysis and Blind Separation of Signals (ICA'99)*, pages 239–244, 1999. 6

[117] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech and Audio Processing*, 10(5):293–302, 2002. 9

[118] G. Tzanetakis, G. Essl, and P. Cook. Audio analysis using the discrete wavelet transform. In *Proc. WSES Int. Conf. Acoustics and Music: Theory and Applications*, 2001. 20

[119] V. Valimaki, J. Huopaniemi, M.. Karjalainen, and Z. Janosy. Physical modeling of plucked string instruments with application to real-time sound synthesis. *Journal of the Audio Engineering Society*, 44(5):331–353, 1996. 30

[120] TS Verma and THY Meng. A 6kbps to 85kbps scalable audio coder. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000. 30

[121] E. Vincent. MUSHRAM: A MATLAB interface for MUSHRA listening tests interface for mushra listening tests, 2005. 132

[122] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech and Language Processing*, 14(4):1462–1469, 2006. 129

[123] T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. *Proc. International Computer Music Conference*, pages 231–234, 2003. 6, 110

[124] T. Virtanen. *Sound Source Separation in Monaural Music Signals.* PhD thesis, Tampere University of Technology, 2006. 110

[125] T. Virtanen and A. Klapuri. Separation of harmonic sound sources using sinusoidal modeling. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2:765–768, 2000. 5, 7, 31, 35, 36, 48, 49, 55, 115

[126] T. Virtanen and A. Klapuri. Separation of harmonic sounds using multipitch analysis and iterative parameter estimation. *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 83–86, 2001. 20

[127] H. Viste and G. Evangelista. A method for separation of overlapping partials based on similarity of temporal envelopes in multi-channel mixtures. *IEEE Trans. Speech and Audio Processing*, 14(3):1051–1061, May 2006. 6, 20, 50, 55, 56, 60, 111, 115, 121, 137, 138

[128] P. J. Walmsley. *Signal Separation of Musical Instruments.* PhD thesis, University of Cambridge, 2000. 7

[129] D.L. Wang. On ideal binary mask as the computational goal of auditory scene analysis. *Speech Separation by Humans and Machines*, 2005. 60, 63, 64, 66, 112, 117, 129

[130] J. Woodruff, Y. Li, and D.L. Wang. Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation. *Proc. Int. Conf. on Music Information Retrieval*, 2008. 66, 111, 112, 115, 121, 137, 138

[131] X. Zhu, GT Beauregard, and LL Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Trans. on Audio, Speech, Lang. Process.*, 15(5):1645–1653, 2007. 66, 118, 119

[132] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J. Acoust. Soc. Am.*, 33:248, 1961. 26

[133] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models.* Springer-Verlag, Berlin, 1990. 26