UNIVERSITY OF MIAMI

DESIGN OF AN UNASSISTED BLIND AUDIO SOURCE SEPARATION ALGORITHM FOR STEREO MUSIC

By

Karthik Palanichamy

A THESIS

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science in Music Engineering Technology

Coral Gables, Florida

May 2007

UNIVERSITY OF MIAMI

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in Music Engineering Technology

DESIGN OF AN UNASSISTED BLIND AUDIO SOURCE SEPARATION ALGORITHM FOR STEREO MUSIC

Karthik Palanichamy

Approved:

______________________ Ken C. Pohlmann Professor of Music Engineering	______________________ Dr. Edward P. Asmus Associate Dean of Graduate Studies
______________________ Dr. Colby N. Leider Assistant Professor of Music Engineering	______________________ Dr. James D. Shelley Assistant Vice President of IT

PALANICHAMY, KARTHIK (M.S., Music Engineering Technology)
Design of an Unassisted Blind Audio Source Separation Algorithm for Stereo Music (May 2007)

Abstract of a Master’s Thesis at the University of Miami.

Thesis supervised by Professor Ken C. Pohlmann.
No. of pages in text. (109)

Audio signals occur as mixtures comprising several sources in the form of speech, music and noise. Blind Audio Source Separation (BASS) algorithms attempt to recover one or more sources from a given observation mixture without prior knowledge or learning of the constituent individual sources. There are many ways to classify the mixture based on the sources that comprise it: broadly as music and speech or specifically as live recordings and synthetic mixtures, and furthermore as mono signals and stereo mixtures. Also, research conducted in the field of BASS has demonstrated that understanding the composition of the mixtures to be separated plays a pivotal role in not only designing the algorithm itself but also in achieving the desired efficiency for a proposed application. In this paper, an original algorithm is described that performs unassisted source separation on stereo music by selection of appropriate panning coefficients. Unlike some previous stereo separation algorithms where separation is based on human-assisted approaches, the main goal of this algorithm is to detect and separate sources from a stereo mixture automatically. Based on the assumptions of an “ADRess” algorithm, a unified frequency-azimuth plane is created for both channels to display all the sources. Using peak and envelope detectors that discriminate panning values corresponding to the azimuth location of different sources in the stereo field, the automatic detection and separation is achieved.

DEDICATION

To my family

for their ever lasting support

&

to Krithika

for being there with me always

no matter the distance.

>|<

ACKNOWLEDGEMENTS

Writing this page is the best feeling as it makes one think and reflect on the entire journey that one has experienced and all the people that came along the way to make it possible. First and foremost, thank you Mom and Dad. If there's any reason I'm here writing this, it is undoubtedly because of your selfless dedication in making me the person I am today. To my sister, for sharing her support and her mind with me, whenever I needed it. To my adorable niece for always making those weekly long distance phone calls rejuvenating. Instead of teaching her I've learnt so much from her about life. To all my engineering buddies in Chennai, who were excited and supportive of my MUE venture. Your friendship was invaluable. Professor Colby Leider, Joe Abbati and Dr. Asmus for making it an ultimate MUE experience. Special thanks to Ken Pohlmann and Dr. James Shelley for expressing their enthusiasm for my work and encouraging me to reach new heights. To my senior Jon Boley for planting the seeds of source separation in me. To my virtual Gurus, Cédric Févotte and Emmanuel Vincent for being there unconditionally whenever those bulky text books didn't make much sense. To my best friends, Oveal, Nicolas and Arun. You guys were my energy and I do hope we get back together sometime soon. Thank you ArC family! My roomies Ajay and Mark for being my home away from home. My sincere regards to Olivier Gillet for being my mentor, philosopher and a great friend. You showed me that in a crazy world of competition, humanity still lives. I would not have been able to complete this thesis without you. And finally, to the person who was close to my heart and spirit regardless of the thousands of miles of land and water that separated us. Krithika, you will always be my greatest teacher of friendship and true love. You continue to amaze me. *Keep the faith.*

TABLE OF CONTENTS

1 Introduction..1

1.1 Audio Source Separation – The Concept……………………………………………1

1.2 Understanding the Nomenclature…………………….…………………………...3

1.2.1 Sources as Auditory Streams……………………………………………...3

1.2.2 Sensors and the Mixing System…………………………………………...5

1.3 Blind Audio Source Separation (BASS)……………………………………………6

1.3.1 Classification of Mixtures……………………………………………………8

1.4 The BASS Problem……………………………………………………………10

1.5 Building the Typology.………………………………………………………...11

1.5.1 Audio Quality Oriented (AQO) Separation Tasks………………………………12

1.5.2 Significance Oriented (SO) Separation Tasks……………………………….…..13

1.5.3 Multi-Channel and Single Channel Analyses……………………………………14

2 Development of the Research Idea………………....……………………….…………...16

2.1 Azimuth Discrimination and Resynthesis (ADRess)…………………..……………...17

2.2 Demixing Commercial Music Productions via Human-Assisted Time Frequency Masking…………………….……………………………………………..…...19

2.3 Human-Assisted versus Unassisted Separation Approach…….……………….….…...20

3 The Proposed Approach……………………………………………………….………21

3.1 Assumptions……………………………………………………………………...21

3.2 Creating the Stereo Track – Mathematical Formulation………………………........22

3.3 Pan Laws – Theoretical Design of the Unassisted Approach…………………………23

3.4 Estimation of Pan Information……………………………………………………...25

4 Implementation……………………………………………………………….……..30

4.1 Processing Tasks………………………………………………………………......30

4.2 Audio Frame-by-frame Analysis.………………………………………………….....31

4.2.1 Fragmentation……………………………………………………………......32

4.2.2 Windowing and Fourier Transform………………………………………….....34

4.2.3 Pan Estimation Per Frequency…………………………………………….......36

4.2.4 Source Estimation.………………………………………………………........37

4.2.5 The Frequency-Azimuth Plane…………………………………….………….38

4.2.6 Pan Values Window Selection……………………………………………….....40

4.3 Pan Tracking…………………………………………………………………….....46

4.3.1 Pan Window Store………………………………………………………...…..46

4.3.2 Pan Window Tracking………………………………………………………....47

4.4 Source Separation and Reconstruction…………………………………………….49

5 Performance Analysis….……………………………………………………….........51

5.1 Experiment 1……………………………………………………………………….51

5.2 Experiment 2……………………………………………………………………….54

5.3 Experiment 3……………………………………………………………………….58

5.4 Experiment 4……………………………………………………………………….62

5.5 Experiment 5……………………………………………………………………….65

5.6 Experiment 6……………………………………………………………………….67

5.7 Performance Analysis……………………………………………………………...70

5.8 Experiment with Computer Music……………………………………………………73
5.9 Experiment with Sine Waves…….…………………………………………………...76

6 Evaluation.…………………………………….…………………...……………..…..78
6.1 MUSHRA Listening Test………………………………………………………….80
6.1.1 Test Description….………………………………………………………………80
6.1.2 Results…………………………………………………………………………......81
6.2 Separation Quality Assessment Listening Test..……………………………………..82
6.3 Analysis of Results.………………………………………………………………...84

7 Conclusions………………………………………………………………………......86
7.1 Final Evaluation – Importance of Listening Tests ………………………………..87
7.2 Future Work.…………………………………………………………………………89
7.2.1 Expansion to Multi-channel Recordings………………………………………..95

References……………………………….………………………………………………...96

INDEX OF FIGURES

Figure 1: Most basic form of Blind Audio Source Separation problem………………….9
Figure 2: Block Diagram for Audio Frame-by-Frame Analysis Stage………………….32
Figure 3: The left and right channels of a 4096 sample audio frame of a stereo signal...33
Figure 4: Attenuated first and last samples of an audio frame after windowing………..35
Figure 5: Frequency spectrum of left and right channels of a windowed audio frame....36
Figure 6: Estimated Pan versus Frequency in an audio frame…………………………..37
Figure 7: Modulus of estimated source at each frequency……………………………...38
Figure 8: Frequency-Azimuth Plane…………………………………………………39
Figure 9: Frequency-Azimuth plane (2-D); x axis - pan position, y axis – source magnitude…………………………………………………………………40
Figure 10: Magnification of the plane for pan values between 0.4 and 0.7…………….41
Figure 11: Output of the Causal (up) and Anti-causal (down) Envelope Detectors……42
Figure 12: Final combined output from Causal and Anti-causal Detectors…………….43
Figure 13: Magnification of the smoothened envelope for pan values between 0.4 and 0.7…………………………………………………………………………...44
Figure 14: Block Diagram for the Pan Tracking and Source Reconstruction stage…….46
Figure 15: Left and right channels of the original track's waveform – Experiment 1.....52
Figure 16: Separated Mono Track – Saxophone………………………………………..53
Figure 17: Separated Mono Track – Piano……………………………………………...53
Figure 18: Separated Mono Track – Drums…………………………………………….54
Figure 19: Left and right channels of the original track's waveform – Experiment 2.....55
Figure 20: 1st un-clustered separated mono track - Lead guitar………………………. 56
Figure 21: 2nd un-clustered separated mono track - Lead guitar……………………….56
Figure 22: 3rd un-clustered separated mono track - Lead guitar………………………..57
Figure 23: 4th un-clustered separated mono track - Lead guitar………………………..58
Figure 24: Left and right channels of the original track's waveform – Experiment 3.....59
Figure 25: 1st un-clustered separated mono track – Drums…………………………….60
Figure 26: 2nd un-clustered separated mono track – Drums…………………………...61
Figure 27: 1st un-clustered separated mono track – Lead guitar……………………….61
Figure 28: 2nd un-clustered separated mono track – Lead guitar………………………62
Figure 29: Left and right channel waveforms of the original track in experiment 4…...63
Figure 30: The six separated un-clustered mono tracks of the vocals………………....64
Figure 31: Clustered vocal mono track obtained from the six un-clustered versions…..65
Figure 32: Left and right channels of the original mix in experiment 5………………..66
Figure 33: Separated mono track consisting of acoustic guitar and drums…………….67
Figure 34: Left and right channel waveforms of the original mix in experiment 6…….68
Figure 35: The three un-clustered separated mono tracks of the vocals………………..69
Figure 36: Separated vocals obtained after clustering the three mono tracks…………..70
Figure 37: Left and right channels of the original MIDI track………………………….74
Figure 38: The separated mono tracks - drums (above) and piano (below)…………….75

INDEX OF TABLES

Table 1: Classification of Mixtures…………………………………………………...7
Table 2: Typical effects applied during stereo mixing……………………………..….8
Table 3: Comparison of Source Separation Techniques………………………………..15
Table 4: Detected Pan Windows…………………………………………………......45
Table 5: MUSHRA Listening Results…………………………………………………. 82
Table 6: Separation Quality Assessment (SQA) Listening Test Results……………….84

1

Introduction

1.1 Audio Source Separation – The Concept

In our daily lives, we are exposed to a variety of unique and diverse sounds that are constantly processed by our complex auditory system. The process by which we are able to select certain sounds from a mixture, treating them as relevant information while disregarding the rest as noise, forms the foundation of the field of study called Auditory Scene Analysis. This study is comprised of a combination of psychoacoustic concepts such as Localization, Binaural Masking and Spatial Hearing, which attempt to explain the various aspects of the innate human ability to process auditory information based on certain criteria. With the growth of digital audio technology, the need for designing systems that could simulate the human auditory system to perform Auditory Scene Analysis for various applications also arose. This led to a new area of research known as Audio Source Separation. This area primarily tackles the problem of separating sound sources that occur simultaneously, the sources being speech, musical instruments or noise. Though it is possible with the help of psychoacoustic concepts and theories such as the "cocktail party effect" to understand how we are able to separate sounds from a mixture and filter out those that are relevant to us, replicating the entire process using computer systems has proved to be an extremely challenging task. The difficulties lie not only in the replication process itself but also in studying the properties of the sources that need to be separated. Newer complications emerge when exploring the methods and acoustical conditions under which the sounds are recorded. Studying the nature of the sources and the recording conditions provide vital information to the design of a separation algorithm. Some of the notable factors that determine the recording conditions are number of microphones used, distance between microphones, number of sound sources, room reverberation and size of room. Though absolute knowledge of these factors cannot guarantee an ideal solution, they certainly help in defining the separation problem at hand and in building the solution framework for that particular application. There are a large number of applications for which source separation algorithms can be developed. Though they all come widely under the same category, the approach and methodology employed in each case is usually different.

In the previous decades, more emphasis has been laid on developing algorithms for purely research interests. The results from these experiments have contributed to understanding the source separation problem and all the various factors that accompany it. And with time, the field of audio source separation has evolved into a special domain with various sub-categories under it, each of them defined by the outcomes of previous studies. As a result, today a general audio source separation typology exists that is able to explain the many different types of separation tasks in the field. Yet, it is still difficult to perfectly classify all the possible models and algorithms that are available due to a large number of complexities and overlaps. To eliminate the hindrance of these complexities when performing source separation research, one needs to make certain assumptions while defining the problem in order to achieve the target results.

1.2 Understanding the Nomenclature

The first distinguishing factor between Audio Source Separation algorithms is whether they employ a “blind” or “unblind” approach. As the name suggests, an “unblind” approach is when complete information about the source signals, the sensors and mixing system is available. But practically, in a majority of the scenarios, this is highly unlikely to be the case. Almost all previous approaches have been “blind” in nature. This is because it is impossible to estimate or be aware of the exact relationship between the sources and the sensors that form the mixing matrix from the observation mixture. Before going into detail of the classification of Blind Audio Source Separation methods, there are certain terms that need to be understood as they form the nomenclature.

1.2.1 Sources as Auditory Streams

The term “source” is used to refer to the audio signals that need to be separated by the algorithm. Though “source” may not be the right term conveying the correct implication, it has been used over the years consistently in Audio Source Separation. But to make it clearer, by “source” one actually means Auditory Streams which is to be understood the same way as Bregman used it decades ago [1]. An auditory stream is produced by a continuous activity of a physical source in the form of waves by interaction with the environment. For example, in the case of a piano played in a closed reverberant room, the sound waves that are produced are not due to the instrument alone, but due to each of the keys that are played along with the reverberation that is produced due to reflection of the waves from the walls and so on. In this case, though the sounds are produced by one musical instrument, logically we would have ‘*n*’ sources, if each key played is considered a source, in addition to the reverberant room itself. In such a case, the use of the term “source” to include all the factors producing the waves is inappropriate. The term “source” is often also misunderstood to be a single physical audio source, which is clearly not the case. It is more correct to use the term Auditory Stream to denote the continuous activity produced by the piano in conjunction with its immediate surrounding. Separation of various auditory streams produced by different musical instruments including the human voice is what Audio Source Separation algorithms deal with. As these auditory streams are perceived to be single entities, the term “source” is used to denote them.

The concept of Auditory Streams was first created by Albert S. Bregman in his book, “Auditory Scene Analysis” where he explained it to be the perceptual grouping of similar units that formed a single event [1]. Though the word “sound” was simpler in usage, he dismissed it for two reasons:

> First of all, a physical happening can incorporate more than one sound. A series of footsteps, for instance, can form a single experienced event, despite the fact that each footstep is a separate sound.

He explained that the word “stream,” suited the purpose better in addressing our mental representations of acoustical events.

> A second reason for preferring the word stream is that the word sound refers indifferently to the physical sound in the world and to our mental experience of it. It is useful to reserve the word stream for a perceptual representation, and the phrase acoustic event or the word sound for the physical cause.

1.2.2 Sensors and the Mixing System

The next term that is commonly used in audio source separation is “sensors.” This is a relatively simpler concept to understand than sources. Sensor is used to denote the physical entities that are used to detect the audio signals or sources. In real world terminology, sensors could be microphones used to record the audio signals or the channels of an audio mixture. For a stereo mixture, there would be two sensors, since there are two channels, left and right. It can also be understood that the sensors form what is known as the mixing system or the mixing matrix in a source separation problem. The relationship between this mixing system and the sources forms the observation mixture or the output mixture. In essence, audio data in the form of a song or a music piece comprising ‘n’ instruments or sources would be referred to as the observation mixture. Mathematically, this can be denoted by the following equation,

$$X = A*S$$

where “S” represents the sources, “A” represents the mixing system comprising of mixing filters, and “X” is the final observation mixture.

Though there are various approaches and algorithms available, the above mentioned concepts remain fundamental to Blind Audio Source Separation (BASS). It is then essential to see the need and use of these algorithms and study how the classification of BASS tasks is done.

1.3 Blind Audio Source Separation (BASS)

Speech enhancement in mobile phones, noise reduction in hearing aids, removal of vocals from audio for karaoke, restoration of corrupted audio data and re-mastering of a stereo CD on multi-channel devices are some of the derived applications of BASS algorithms performed for remixing or modifying the observation mixture. Real-time speaker separation for simultaneous translation and aids for electronic music composition by sampling of instrumental sounds are some of the prominent direct applications. The separated sources can also be used for other purposes such as single source indexing, transcription and coding techniques needed for various other applications such as improved automatic indexing of audio documents, multi-speaker speech recognition in a cock-tail party scenario or object based coding.

Different BASS algorithms have to address different problems based on their application and goal. Certain tasks require finding the number of sources in a mixture given the observation mixtures, certain other require acquiring one or most of the source signals from the mixture and the mixing system. An important point to be remembered is that a particular algorithm cannot be expected to do well on all tasks. Each task has a unique problem and different criteria are considered to evaluate its performance. Before creating an algorithm it is important to understand the problems that have to be specifically solved to ensure success of the algorithm.

1.3.1 Classification of Sound Mixtures

Before learning about the separation problem itself, it is important to classify the sources that the algorithms have to deal with. One such classification is whether the sources to be separated are live sources or synthetic mixtures. Live sources can be recorded separately or together based on the music genre or occasion. The sound setup and microphones determine the amount of interferences and reverberation on each of the channels. On the contrary, synthetic mixtures can include live recordings that are mixed down to two channels (stereo) with effects such as panning, reverb, equalization, and so on. In signal processing literature, sound mixtures can be classified as:

- under-determined vs. over-determined
- instantaneous vs. convolutive
- time varying vs. time invariant

Table 1 explains the meaning of the vocabulary used to classify sound mixtures in source separation terminology.

Table 1: Classification of Mixtures [4]

Term	*Equivalent Meaning*
Over-determined (or under-complete) Determined Under-determined (or over-complete)	More mixture channels than sources As many mixture channels as sources Less mixture channels than sources
Instantaneous Anechoic Convolutive (ehoic) Reverberant	Trivial mixing filters (gains, no delays) Trivial mixing filters (gains and delay pairs) Non-trivial mixing filters Mixing filters exhibiting realistic reverberation
Time-invariant Time-varying	Mixing filters constant over time Mixing filters slowly varying with time

Live recordings are over-determined time-varying convolutive mixtures as they involve a number of microphones, reverberation and moving sources whereas synthetic mixtures are under-determined time-invariant convolutive mixtures which are mixed down to stereo using synthetic reverberation.

Table 2 lists the common synthesis techniques or effects that are applied during the creation of the mixture. The effects in the top half of the table modifies spatial properties of sources whereas the bottom half modifies spectro-temporal properties.

Table 2: Typical effects applied during stereo mixing [4]

Effect	*Use*	*Processing*
Pan	Creates a point stereo image	Scales a mono signal by two constant or slow time-varying positive gains
Auto-pan	Creates an extended stereo image	Scales a mono source signal by two fast time-varying positive gains
Echo/Reverb	Mimics natural reverberation or echo	Filters mono source signal by two different synthetic filters
Polarity	Makes a stereo image sound un-natural	Inverts the sign of one channel
Compressor	Reduces dynamic range	Scales an image by a slowly time-varying gain
Equalizer	Modifies timbre	Scales each sub-band of an image by a constant or slowly time-varying gain
Tremolo/Vibrato	Increases expressiveness	Applies amplitude/frequency modulation
Chorus	Multiplies the number of perceived sources	Adds to an image a few time-delayed versions of itself with slowly time-varying delays

1.4 The BASS Problem

The BASS problem can be formulated in its simplest form as follows. For this we bring back the equation,

$$X = A*S \tag{1.1}$$

When written in the time domain, it can be re-written as,

$$X(t) = A{*}S(t) \tag{1.2}$$

A set of unknown source signals that are mutually independent of each other can be considered and denoted by $s_1(t)$, $s_2(t)$,…, $s_M(t)$, which form the source vector S(t). With respect to audio, these signals would be the various auditory streams from the musical instruments in a music piece. These signals are recorded using sensors and are then linearly mixed using an unknown matrix of mixing filters A, in an unknown environment to give the observation mixture X(t). The source and observation vectors can be written as,

$$S(t) = [s_1(t), s_2(t),\ldots s_M(t)]^T \tag{1.3}$$

$$X(t) = [x_1(t), x_2(t),\ldots x_M(t)]^T \tag{1.4}$$

Figure 1 illustrates a basic form of the Blind Audio Source Separation problem.

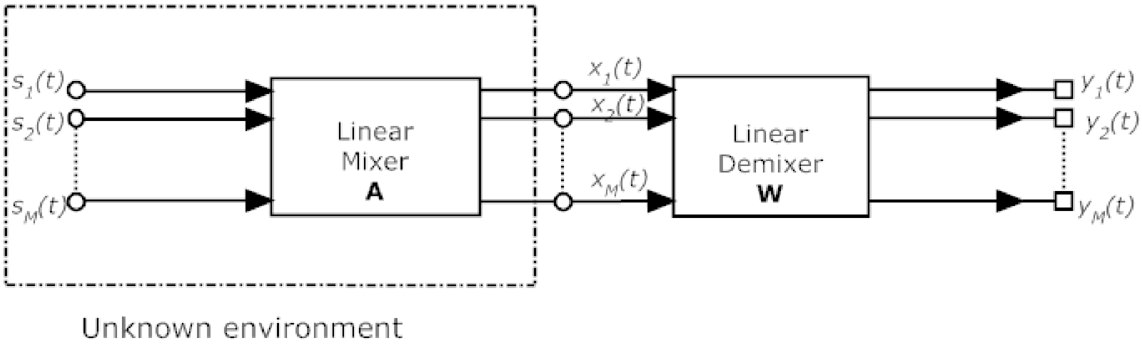

Figure 1: Basic form of Blind Audio Source Separation problem

Conceptually, W is the inverse system of A, and Y would consist of the separated source image signals. The estimation of the source signals is performed on the basis of the output signals $Y(t) = [y_1(t), y_2(t),\ldots y_m(t)]^T$ and the sensor signals along with some *a* *priori* knowledge of the mixing system. In certain cases when the inverse system does not exist or the observation signals are less than the source signals, it is easier to estimate an unknown mixing system and then estimate the source signals implicitly by using some *a priori* information about the system and applying a suitable optimization procedure. One of the most basic assumptions that BASS algorithms are based on is that there are as many sensors as the number of sources, although this many not be true in many cases. All these assumptions are made to be able to design the framework for a BASS algorithm.

1.5 Building the Typology

An important criterion before developing a BASS algorithm is to evaluate whether the output of the algorithm returns a set of extracted sources that need to be listened to or not. Based on this condition, we have two types of applications:

-Audio quality oriented (AQO) and

-Significance oriented (SO)

AQO applications extract only those sources that are required to be listened to after separation or after some audio post-processing treatment. Some of the techniques that are used to perform this are Independent Component Analysis (ICA) and Spectral Decomposition (SD). In SO applications, the extracted sources and/or the mixing parameters are processed to obtain information at more abstract levels to find a representation of the observations related to human perception. For example, if we need to analyze the number and kind of instruments in a mixture, that would be the scope of an SO application. The difference between both the applications is that separation quality is of lesser importance in SO than in AQO because SO aims at keeping only certain specific features of the source. This means feature extraction algorithms constitute a large part of SO applications.

1.5.1 Audio Quality Oriented (AQO) Separation Tasks

There are two main types of applications in AQO separation. In the first kind, we are interested in studying each individual extracted source, while in the second one the objective is to listen to a new mixture of the source by creating a different mixing matrix system.

1. One Versus All

This method consists of extracting one sort of sound from a cluster considering the remaining sounds as noise. The extracted source is the target source s_m. Some applications of this method are restoration of old monophonic recordings, speech de-noising, de-reverberation for auditory prostheses and mobile phones and extraction of certain music samples for electronic music composition. The method of extraction here is to estimate the source with the highest signal to noise ratio (SNR). But again, this is dependent on the task that has to be performed. In case of convolutive mixtures where it is highly unlikely to extract an exact estimation of the source due to indeterminacies, it may be sufficient to obtain filtered versions of the target source and not the target itself. In some other applications, it may be of interest to study the contribution of ‘s_m’ to each sensor. In such cases, quality criteria would include the differences in spatial direction of ‘s_m’ when listening to the source image signal ‘s_{im}’ and the observation signal ‘X.’ A logical extension of the *one versus all* problem would be to extract each and every source from a cluster using different SNR values for each source if all of them need to be listened to.

The number of sources that need to be extracted play an important role when tackling the one versus all problem. Algorithms have to be robust enough to take into account the noise levels, dependency between sources and the kind of mixing used. ICA and SD are used in blind case separation and in other semi-blind and unblind cases where *a priori* information is used in a model of the target source which can be achieved through Hidden Markov Models (HMM) [2].

2. Audio Scene Modification

Audio Scene Modification consists of obtaining a new mixture,

$X_{remix} = B^* [f_1(s_1),\ldots f_M(s_M)]^T$. This task is performed by extracting all the sources 's_m' where m=1 to M, from the original observation mixture 'X,' applying an adapted audio processing 'f_m' to each extracted source and remixing the tracks using a possibly different mixing matrix 'B', so as to listen to the new result '$X_{remix.}$' In this method, prior extraction of each source is not a requirement as more emphasis is laid on the audio quality of the mixture after audio processing. Some applications include re-mastering of a stereo CD, blind multi-channel diffusion of stereo recordings, spatial interpolation, and cancellation of the voice in a song for "automatic karaoke." One way to evaluate the remixed output is to calculate the SNR of the estimated remixed scene with respect to the expected result i.e. the scene constructed by remixing true sources. When compared to the *one versus all* problem, separation quality may not be an issue. For example, only when a single instrument needs to be made "brighter" in a CD, the distortion and crosstalk formed in the extracted instrument will affect the entire result. Moreover due to auditory masking effects, it is most likely that the noise is masked by the presence of the other sources.

The difficulties of audio scene modification are similar to that of *one versus all*. The number of sources in the mixture significantly affects the mixing process. Based on the type of mixture that needs to be remixed, an appropriate filtering method has to be selected which may not always be effective for the audio source separation task. The type of audio processing 'f_m' that is imposed on every extracted source and the structure of the new mixing matrix also play an important role in the final result. Mono recordings are much more difficult to deal with than stereo. Because in mono, it is unlikely to be able to cancel only one source or increase the output of another source alone. Blind or unblind remixing can be done to perform audio scene modification. In blind mixing, one has to rely on directly computable features such as instantaneous power from each source or directionality or panning coefficients of the sources. In case of unblind remixing, modeling of each source can be done using HMMs or physical modeling of each source instrument.

1.5.2 Significance Oriented (SO) Separation Tasks

The objective of SO applications is to retrieve specific source features and/or mixing parameters with the aim of studying and describing complex audio signals at various cognitive levels and focusing on different aspects of sound. In such applications, ASA plays a major role for feature extraction and the parameters for extraction. Main applications of SO separation include:

-indexing of audiovisual (AV) databases and

-construction of intelligent hearing systems.

Depending on the area of application, descriptors of low or high levels are required. Some examples of descriptors are musical score of each instrument in a musical excerpt, the text pronounced by a speaker in a noisy surrounding, or the spatial position of the sources with respect to the sensors in a live recording session. Other descriptors can also be naming of instruments and genre classification, speaker identification and linking of audio to corresponding visual objects on a video screen.

1.5.3 Multi-Channel and Single Channel Analyses

BASS algorithms can also be classified based on the number of channels that are to be separated. In this context, channels also imply the number of observation mixtures available for separation.

1. *Multi-Channel Identification based on Sparsity*

Sparsity is the core of source separation methods for simple multi-channel sound mixtures. Time-frequency sparsity is an important method that allows separate sources in each sub-band based on spatial diversity and localization. These approaches aim at exploiting the azimuth and panning information of various sources in the mixture. Complementary assumptions are then needed to link sub-band signals that belong to the same source. Common methods are:

- Independent Component Analysis (ICA)
- Degenerate Unmixing Estimation Technique (DUET-like) methods.

2. Single Channel Identification based on Advanced Models

The previously mentioned time-frequency sparsity methods are insufficient for the separation of single-channel mixtures. There are more factors that need to be taken into consideration before separating single channel sources. For this purpose, higher level models are used which represent the short term magnitude spectrum of the mixture and make note of the discrete structure of the sources along with periodicity, spectral envelope and temporal continuity. Some of the approaches are:

- Hidden Markov Models (HMM)
- Spectral Decomposition (SD)
- Computational Auditory Scene Analysis (CASA)

Every source separation approach has its own set of limitations and strengths depending on the type of intended application and mixture to be separated. A summary of the various approaches along with their performances is given in Table 3. X denotes major limitation, O denotes minor limitation and √ denotes no limitations.

Table 3: Comparison of Source Separation Techniques [4]

Separation Technique	*Fewer channels than sources*	*Long reverberation*	*Sources with similar spectral properties*	*Limited prior information*
ICA	X	X	√	√
DUET-like methods	√	O	√	√
HMM	√	√	X	O
SD	√	√	X	O
CASA	√	√	X	O

2

Development of the Research Idea

After reviewing audio source separation research, it can be seen that emphasis has been placed in achieving goals defined by a rigid framework made possible by a set of initial assumptions pertinent to a particular problem. Most algorithms have been less robust in nature and very inflexible. Though the outcomes of this research have furthered knowledge about source separation, they lack practical usefulness. For example, various source separation algorithms need information such as size of the room in which recording was done, room reverberation, distance between the microphones and so on. The algorithm's performance is then measured based on audio signals that are created specifically for that research under pre-defined conditions. When the same algorithm is applied to audio signals created from outside these conditions, the algorithm performs poorly. This pattern can be observed commonly with BASS algorithms too. In other words, knowledge about the sources and the mixing conditions has proven to be two major criteria in designing a BASS algorithm. As seen in the previous chapter, the other framework of designing BASS algorithms is for particular practical applications which can be broadly categorized as direct applications or derived applications.

Focusing on music-based applications in this thesis research, CD quality audio is considered to be a superior standard when it comes to commercial music productions. But when it comes to designing source separation algorithms, little research has been done.

Taking the above into consideration, in this paper the objective is to design a robust BASS algorithm for source separation on stereo audio.

After defining the algorithm's objective, the next step is to review and study previous approaches that focus on stereo audio to have a clear idea of the progress in this area of source separation. Several papers [6], [7], [11] and [12] have been found to be of significant interest for the following reasons:

1) The algorithms aim at separating sources from stereo music tracks without prior knowledge of the sources, the number of sources, or mixing conditions making them highly robust practical application based algorithms.
2) Their separation quality and performance is superior not only when compared to their predecessors but also when studied objectively.
3) The algorithms are built on the same fundamental concepts and assumptions that are conducive to analyzing and separating sources in stereo music.

For these reasons, the proposed algorithm is based on these previous approaches with an aim of incorporating novel ideas into the current framework to achieve substantial results. This would also provide simpler methods for performance measurement.

2.1 Azimuth Discrimination and Resynthesis (ADRess)

ADRess is an efficient source separation algorithm based on azimuth discrimination of sources within the stereo field. It was developed by Dan Barry and Eugene Coyle of Dublin Institute of Technology and Bob Lawlor of National University of Ireland [6]. It achieves image localization in stereo recordings by exploiting the pan positions of the various musical instruments that make up the mix. Most stereo recordings have an inter-aural intensity difference (IID) between the left and right channels as the different instruments are panned to various degrees in the azimuth plane. The algorithm uses gain scaling and phase cancellation to obtain left and right channel mixtures where the effect of the left channel on the right is nullified and vice versa. This is followed by constructing a frequency-azimuth plane that displays the frequency dependent nulls of the various sources. Using this information, sources are separated and resynthesis is carried out.

The algorithm is designed for stereo recordings made in the fashion where ‘N’ sources are first recorded individually as mono tracks, and then summed and spread across the two channels, left and right, using a mixing console. In the mixing process, a panoramic potentiometer is used to achieve localization of the various sources by dividing them into the two channels with different intensity ratios that are continuously variable. A source can be “positioned” at any location between two speakers by creating an inter-aural intensity difference between the two channels. This is done by attenuating a source signal in one of the channels which causes it to be localized in the other channel thereby causing the source to come from a particular location in the azimuth plane. In commercial stereophonic recordings only the intensity of the sources between the two channels differs but the phase information is exactly the same. The ADRess algorithm is created for recordings made using this methodology. The algorithm also exhibits limited success with stereo-pair, mid-side and binaural type recordings.

This algorithm is found to be effective and dynamic for practical applications because the majority of recordings are made using mixing consoles. Also, unlike other algorithms, prior knowledge of the sources, sensors or the recording conditions are not required to perform the separation task. The ADRess algorithm succeeds fairly well in separating sources from commercial recordings. The degree of separation largely depends on the number of sources present, the proximity of the sources in the azimuth plane and the intensity level of the sources. Experiments reveal that a low number of sources results in a low signal to noise ratio whereas a high number of sources results in missing overlapping partials.

2.2 Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking

This research aims at source separation using a real-time user-defined interface. It was developed by MarC Vinyes, Jordi Bonada and Alex Loscos of Pompeu Fabra, University in Barcelona, Spain [7]. The algorithm exploits the panning of various musical sources in the stereo field that are defined by a set of pan laws used in most mixing consoles today. Performing time-frequency masking and extracting the DFT coefficients of the sources using human-assisted selection based on certain criteria, a real-time graphical interface is used to perform the demixing process. The methodology shows considerable success in demixing sources from commercial recordings.

The authors propose a new version of audio source separation where they believe there may be infinite solutions to separating an audio signal from a mixture in which a human finds only a subset of these solutions perceptually meaningful. Based on this principle they aim to create a real-time interface for source separation where the selection criteria is human-assisted, or in other words, establishing a real-time feedback loop for the user performing source separation.

They make the same assumptions as in the ADRess paper where recordings are made in which sources are first individually recorded and then electrically summed and panned using a mixing console in addition to applying specific reverberation, equalization and other effects. They also propose the solution of their audio source separation framework to be ideal if the separated tracks are perceived most similar to the tracks that are used to make the original mix. This conversely means that the original source tracks would be required as a reference to compare to the separated ones in order to assess and evaluate the separation quality.

2.3 Human-Assisted versus Unassisted Separation Approach

The two methods discussed above are successful separation techniques. Both techniques require a user interface that provides aural and visual feedback where the various separation parameters are manually adjusted in real-time until a perceived separation quality is reached. The quality of separation is achieved on a subjective basis and the need for human assistance in the selection process is inevitable. Though these research techniques serve as the foundation for this thesis, they differ significantly on the premise that this thesis aims at developing a fully automatic separation algorithm which would require no type of live human interaction or assistance in the separation process. In this thesis research, a completely novel and unique approach is formulated and applied to perform blind audio source separation on commercial stereo music. This involves estimation of the panning coefficients applied to the various sources at each frequency and automatically extracting the sources carrying the same pan information thereby separating them from the mixture based on their spatial location in the stereo field.

3
The Proposed Approach

Once the algorithm's framework has been established, the next step in blind audio source separation is to make assumptions based on which the algorithm is designed to perform its targeted task. The first six assumptions stated below are based mainly on the above algorithms.

3.1 Assumptions

1) The algorithm requires no prior learning or knowledge of the type of sources or number of sources in order to separate an observation mixture.
2) The algorithm assumes that the stereo mixtures that it separates are comprised of sources that are recorded as individual mono tracks and then electrically summed and panned (using a pan-pot) across two channels resulting in a stereo track– the technique followed in most modern day commercial recordings.
3) The various sources in the stereo track are panned differently from each other in the stereo field.
4) The mixing conditions such as room reverberation and room dimensions are unknown and are assumed to have no effect on the algorithm's separation capability.
5) The individual mono tracks do not have any processed artificial stereo reverberation.

6) The mono tracks do not undergo any phase processing after being mixed to create the stereo track.
7) The algorithm can identify the number of ‘meaningful’ sources in a recording and separate them on its own without any assistance.
8) The algorithm can separate sources that use automated pan when they are first recorded as mono tracks.

It is important to note that not all assumptions are incorporated in the actual design of the algorithm. In BASS research the main purpose of assumptions is to provide a framework within which an algorithm is built thereby also providing a means for understanding and evaluating its behavior and performance. The need for such assumptions arises as a result of unavailability of adequate knowledge about the mixture to be separated thereby preserving “blindness” in the approach. Assumptions serve as a starting point in the algorithm’s design procedure but are not always requirements for its functioning. With respect to this thesis it is the second assumption that is strictly incorporated in the theoretical and practical design process of the unassisted algorithm.

3.2 Creating the Stereo Track – Mathematical Formulation

Most music recordings are created by electrically summing and distributing ‘N’ sources that are first recorded as mono tracks onto two channels using a mixing console. Image localization of the source in the stereo field is achieved by means of a panoramic potentiometer which divides a single source into two channels using continuously variable intensity ratios [6]. Using the concept of inter-channel intensity difference (IID), a source is placed at any particular point within the stereo field. In cases of automated pan, a source can be perceived to be sweeping across or moving across a stereo field for a short duration in the recording. This is done by feeding the source onto one channel thereby attenuating it in another channel causing it to be more localized in a channel. As a result, only the intensity of the source differs between the two channels and the phase remains unaltered.

Assuming we have 'N' mono tracks which can be called the sources, $s_1(t)$, $s_2(t)$,…$s_N(t)$ when mixed by a mixing console result in a stereo track with left channel L(t) and right channel R(t). The stereo track can be seen as a linear combination of the various source signals with different panning coefficients:

$$\mathrm{L}(t) = \lambda_1 s_1(t) + \lambda_2 s_2(t) + \ldots + \lambda_N s_N(t) \tag{3.1}$$

$$\mathrm{R}(t) = \rho_1 s_1(t) + \rho_2 s_2(t) + \ldots + \rho_N s_N(t) \tag{3.2}$$

The above equations are a simplification assuming that there is no phase processing such as spatialization or reverberation applied while making the stereo track.

3.3 Pan Laws – Theoretical Design of the Unassisted Approach

The unassisted approach involves estimation of pan information applied to various sources during the creation of the stereo track. In this thesis, a new approach of the pan estimation process is presented where based on the mathematical pan laws that are used in mixing consoles, the maxima and minima are calculated for every frequency giving two solutions that help perform source separation. If the third assumption stated in the previous section holds completely true, then each source will carry a unique pan coefficient. This implies that estimation of a particular pan coefficient throughout all frequencies would lead to separation of that source from the stereo mixture.

In typical sound-mixing boards the panning coefficients applied to a source on the left and right channel are not independent of each other and are mathematically related using the pan-pot [7]. The pan-pot follows a sine-cosine curve where:

$$\lambda = \cos \frac{\pi x}{2} \tag{3.3}$$

$$\rho = \sin \frac{\pi x}{2} \tag{3.4}$$

'*x*' denotes the pan position, where $x = 0$ corresponds to a full-left pan, $x = 0.5$ corresponds to perfectly centered source and $x = 1$ corresponds to a full-right pan. The main advantage of the pan-pot following a sine-cosine curve is that the output power stays constant when the source is panned from left to right or vice versa. This is because the power is proportional to the square of the sum of the magnitudes which is given by the following basic trigonometric law:

$$\sin^2\alpha + \cos^2\alpha = 1 \tag{3.5}$$

Assuming that we have a stereo track that is generated due to mirroring one source mono track using sine-cosine pan-pot, disregarding the dependence on time for ease of mathematical computation, we have:

$$\mathrm{L} = \cos\frac{\pi x}{2} s \tag{3.6}$$

$$\mathrm{R} = \sin \frac{\pi x}{2} s \tag{3.7}$$

where '*s*' is the single source. From the above equations, calculating the squared modulus of the linear combination of left and right channels as a function of '*y*', the coefficients:

$$\alpha_{\mathrm{L}}(y) = \cos \frac{\pi y}{2} \tag{3.8}$$

$$\alpha_{\mathrm{R}}(y) = \sin\,\frac{\pi y}{2} \tag{3.9}$$

The squared modulus function is denoted by *f*(*y*):

$$f(y) = |\,\alpha_{\mathrm{R}}(y).\mathrm{R} + \alpha_{\mathrm{L}}(y).\mathrm{L}\,|^{2} \tag{3.10}$$

Rewriting the above using equations 3.6, 3.7, 3.8 and 3.9, we get:

$$f(y) = \left|\cos\frac{\pi y}{2}\cos\frac{\pi x}{2}s + \sin\frac{\pi y}{2}\sin\frac{\pi x}{2}s\right|^{2}$$

$$= \left|\cos\frac{\pi y}{2}\cos\frac{\pi x}{2} + \sin\frac{\pi y}{2}\sin\frac{\pi x}{2}\right|^{2}|s|^{2} \tag{3.11}$$

Rewriting equation 3.11 using trigonometric identities, we get:

$$f(y) = \left|\frac{1}{2}(\cos(\frac{\pi y}{2} - \frac{\pi x}{2}) + \cos(\frac{\pi y}{2} + \frac{\pi x}{2})) + \frac{1}{2}(\cos(\frac{\pi y}{2} - \frac{\pi x}{2}) - \cos(\frac{\pi y}{2} + \frac{\pi x}{2}))\right|^{2}|s|^{2}$$

$$= \frac{1}{4}\left|\cos\frac{\pi(y-x)}{2} + \cos\frac{\pi(y+x)}{2} + \cos\frac{\pi(y-x)}{2} - \cos\frac{\pi(y+x)}{2}\right|^{2}|s|^{2}$$

$$= \frac{1}{4}\left|2\cos\frac{\pi(y-x)}{2}\right|^{2}|s|^{2}$$

$$= \cos^{2}\frac{\pi(y-x)}{2}|s|^{2}$$

$$\therefore\ f(y) = \frac{1}{2}(1+\cos(\pi(y-x)))|s|^{2} \tag{3.12}$$

3.4 Estimation of Pan Information

After obtaining the simplified version of the square modulus function *f*(*y*), the minima and maxima are computed by taking the first derivative with respect to '*y*' and equating it to zero. Performing this on equation 3.12 we get:

$$\frac{df}{dy}(y) = -\frac{\pi}{2}\sin(\pi(y-x))|s|^2 = 0 \tag{3.13}$$

This implies that, $\pi(y-x) = k\pi$, where $k \in Z$. Further simplifying we obtain:

$$y_{k\,=}\, x + k \tag{3.14}$$

To determine whether equation 3.14 represents a solution that is a maxima or minima, we proceed by computing the second derivative:

$$\frac{d^2 f}{dy^2}(y) = -\frac{\pi^2}{2}\cos(\pi(y-x))|s|^2 \tag{3.15}$$

Applying the solution 3.14 in equation 3.15, we get:

$$\frac{d^2 f}{dy^2}(y_k) = -\frac{\pi^2}{2}\cos(\pi(x+k-x))|s|^2$$

$$= -\frac{\pi^2}{2}\cos(\pi k)|s|^2 \begin{cases} > 0, k = odd \\ < 0, k = even \end{cases} \tag{3.16}$$

From equation 3.16 it can be observed that the function has a maximum when '*k*' is even, and a minimum when '*k*' is odd. The values of these two points can be determined by substituting for $y = x + k$ to get 'f_{max}' and '$f_{min.}$' Since the function is periodic in nature, any odd or even value of '*k*' in equation 3.12 will give the same result. Substituting $k = 0$ (even) and $k = 1$ (odd) we get:

$$f_{max} = f(x) = \frac{1}{2}(1+\cos(\pi(x-x)))|s|^2$$

$$= \frac{1}{2}(1+\cos(0))|s|^2$$

$$\therefore\ f_{\max} = |s|^2 \tag{3.17}$$

$$f_{min} = f(x{+}1) = \frac{1}{2}(1+\cos(\pi(x+1-x)))|s|^2$$

$$= \frac{1}{2}(1+\cos\pi)|s|^2$$

$$\therefore f_{\min} = 0 \tag{3.18}$$

From equations 3.17 and 3.18 it can be inferred that for the equation $y = x + k$ if $y = x$ (as $k = 0$) is selected (where 'x' is the pan value applied to the source during mixing) then the magnitude of the source signal is reconstructed which is the maximum. And if $y = x + 1$ (as $k = 1$) is selected then zero is obtained which is the minimum. This result forms the main principle in estimating the pan information needed to separate the sources.

In the above procedure of determining the minima and maxima, the dependence on time is neglected for ease in computation. The above derivations hold true only if 's', 'L' and 'R' are taken to be the Fourier Transform of the source at a given frequency. Considering we have the complex Fourier Transform at a given frequency, we determine the pan value at that particular frequency. At a given frequency 'Ω', the left and right channel Fourier Transforms can be written as:

$$\mathrm{R} = r_r + ir_i \tag{3.19}$$

$$\mathrm{L} = l_r + il_i \tag{3.20}$$

Rewriting the squared modulus function in equation 3.10 in terms of 'R' and 'L' from the above equations, we have:

$$f(y) = |\ \alpha_{\mathrm{R}}(y).\mathrm{R} + \alpha_{\mathrm{L}}(y).\mathrm{L}\ |^2$$

$$= \left|\cos\frac{\pi y}{2}(r_r + ir_i) + \sin\frac{\pi y}{2}(l_r + il_i)\right|^2$$

$$= \left|\left(r_r\cos\frac{\pi y}{2} + l_r\sin\frac{\pi y}{2}\right) + i\left(r_i\cos\frac{\pi y}{2} + l_i\sin\frac{\pi y}{2}\right)\right|^2$$

$$= \left(r_r \cos\frac{\pi y}{2} + l_r \sin\frac{\pi y}{2} \right)^2 + \left(r_i \cos\frac{\pi y}{2} + l_i \sin\frac{\pi y}{2} \right)^2 \tag{3.21}$$

Differentiating with respect to '*y*' in order to calculate the maxima and minima:

$$\frac{df}{dy}(y) = 2\frac{\pi}{2}\left(r_r \cos\frac{\pi y}{2} + l_r \sin\frac{\pi y}{2} \right)\left(- r_r \sin\frac{\pi y}{2} + l_r \cos\frac{\pi y}{2} \right)$$

$$+ 2\frac{\pi}{2}\left(r_i \cos\frac{\pi y}{2} + l_i \sin\frac{\pi y}{2} \right)\left(- r_i \sin\frac{\pi y}{2} + l_i \cos\frac{\pi y}{2} \right)$$

$$= \pi\left(- r_r^2 \cos\frac{\pi y}{2}\sin\frac{\pi y}{2} + r_r l_r \cos^2\frac{\pi y}{2} - r_r l_r \sin^2\frac{\pi y}{2} + l_r^2 \cos\frac{\pi y}{2}\sin\frac{\pi y}{2} \right)$$

$$+ \pi\left(- r_i^2 \cos\frac{\pi y}{2}\sin\frac{\pi y}{2} + r_i l_i \cos^2\frac{\pi y}{2} - r_i l_i \sin^2\frac{\pi y}{2} + l_i^2 \cos\frac{\pi y}{2}\sin\frac{\pi y}{2} \right)$$

$$= \pi\left(- r_r^2 + l_r^2 - r_i^2 + l_i^2 \right)\cos\frac{\pi y}{2}\sin\frac{\pi y}{2} + \pi(r_r l_r + r_i l_i)\left(\cos^2\frac{\pi y}{2} - \sin^2\frac{\pi y}{2} \right) \tag{3.22}$$

Simplifying by using trigonometric identities we get:

$$\frac{df}{dy}(y) = \frac{\pi}{2}(-r_r^2 + l_r^2 - r_i^2 + l_i^2)\sin \pi y + \pi(r_r l_r + r_i l_i)\cos \pi y \tag{3.23}$$

Equating 3.23 to zero to find the solution that would give the minima and maxima:

$$\frac{df}{dy}(y) = \frac{\pi}{2}(-r_r^2 + l_r^2 - r_i^2 + l_i^2)\sin \pi y + \pi(r_r l_r + r_i l_i)\cos \pi y = 0$$

$$\frac{\pi}{2}(-r_r^2 + l_r^2 - r_i^2 + l_i^2)\sin \pi y = -\pi(r_r l_r + r_i l_i)\cos \pi y$$

$$\frac{\sin \pi y}{\cos \pi y} = 2\frac{r_r l_r + r_i l_i}{r_r^2 + r_i^2 - l_r^2 - l_i^2}$$

$$\tan \pi y = 2\frac{r_r l_r + r_i l_i}{r_r^2 + r_i^2 - l_r^2 - l_i^2} \tag{3.24}$$

tan πy is a periodic function with period = 1. In order to determine the maximum and minima values, we calculate the second derivative of equation 3.23:

$$\frac{d^2 f}{dy^2}(y) = \frac{\pi^2}{2}(-r_r^2 + l_r^2 - r_i^2 + l_i^2)\cos \pi y - \pi^2 (r_r l_r + r_i l_i)\sin \pi y \qquad (3.25)$$

The minimum value is obtained when the solution to equation 3.24 gives a positive result in the second derivative, and the maximum is obtained when the same solution gives a negative result in the second derivative. Thus as in the previous case (in the time neglected calculation of maxima and minima), the minimum would be the estimated pan position for a particular frequency and the maximum would be the magnitude of the originally mixed source signal.

The above derivation is based on the hypothesis that there exists only one source signal in the observation mixture. For such a case, the above equations hold true. But in most practical scenarios where there are a number of instruments playing together at any given instant, the estimation of pan values at each frequency and the reconstruction of the source signal will not be completely accurate due to overlapping of frequencies. A minor improvement in source estimation can be achieved by subtracting from the maximum signal modulus (magnitude of the estimated source) the minimum signal modulus ($y = x$ or the estimated pan) which would result in the estimation of the extraneous sounds contributed by the other sources in that particular frequency.

4
Implementation

In blind audio source separation research, there have been few algorithms aimed at separating sources in commercial music productions in the stereo format. Moreover, the unassisted approach to source separation based on pan estimation has not been implemented before. To ensure the success of this new approach, it is important to choose a platform which aids in the implementation process along with reduced computational load. Taking this into consideration, the practical implementation of the algorithm is done using Matlab.

4.1 Processing Tasks

Based on the theoretical formulations presented in the previous chapter, the implementation of the algorithm can be broadly classified into three stages:

1. Audio Frame-by-Frame Analysis

In this stage, the input stereo mixture signal is fragmented into short time frames and then each frame is processed to obtain the Fourier Transform. From each frame, pan information per frequency is extracted. The pan positions that contain the constituent sources are then detected.

2. Pan Tracking

The detected pan positions are tracked from frame to frame and the Fourier Transforms of the input signal are separated into Fourier Transforms corresponding to the detected pan positions.

3. Source Separation and Reconstruction

The similar pan position information are then extracted which would thereby correspond to the position of a particular panned source in the stereo field. The source is then reconstructed by performing an Inverse Fourier Transform followed by the overlap-and-add technique from frame to frame.

4.2 Audio Frame-by-Frame Analysis

The frame-by-frame analysis process is the first step in the algorithm. The main objective is to fragment the incoming signal into short time frames and to extract:

1. Estimated Pan Value Per Frequency

This is the pan-pot value applied to a source that is estimated at a particular frequency where the source is present.

2. Estimated Source Per Frequency

This is the estimated magnitude and phase of a source at a particular frequency.

3. Estimated Pan Windows

A pan window corresponds to the range of pan values within which a target source is present. In theory, a source that is positioned at a particular pan value will carry that value in the frequencies in which it is present. In ideal scenarios, it is possible to extract the source based on that single pan value. But in real world scenarios due to the effect of other sources present in the stereo signal, as well as noise and interference, it is nearly impossible to find a single source containing the same pan value at different frequencies. Due to this, the estimated pan value for a source may differ from frequency to frequency. This spread or close range of pan values for a particular source is termed as the pan window. Figure 2 shows the various blocks comprising the audio frame-by-frame analysis stage.

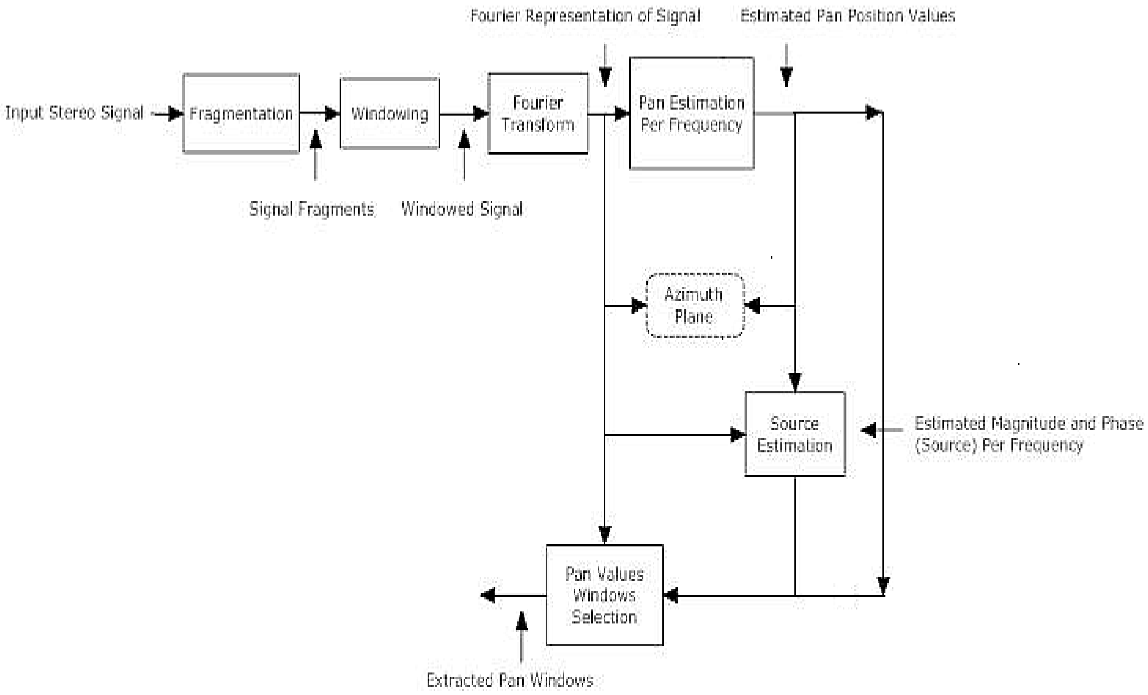
Fourier Representation of Signal
Estimated Pan Position Values
Input Stereo Signal
Fragmentation
Windowing
Fourier Transform
Pan Estimation Per Frequency
Signal Fragments
Windowed Signal
Azimuth Plane
Source Estimation
Estimated Magnitude and Phase (Source) Per Frequency
Pan Values Windows Selection
Extracted Pan Windows

Figure 2: Block Diagram for Audio Frame-by-Frame Analysis Stage

4.2.1 Fragmentation

The incoming stereo signal is fragmented into frames of audio consisting of 4096 samples each. This process allows the signal to be effectively analyzed by Fourier transform. The size of each frame is selected considering the trade-off that exists between time-domain resolution and frequency-domain resolution. A shorter audio frame results in better time resolution at the expense of worse frequency resolution, whereas a longer audio frame gives better frequency resolution but worse time resolution. Another factor when choosing the frame length is the computational efficiency when computing the Fast Fourier Transform (FFT) for each frame. Efficient frame lengths are powers of 2 where the FFT is computed over N.log N (where ‘N’ = frame length). Taking this into consideration, a frame length of 4096 samples gives reasonable resolution in both time and frequency domains for a sampling rate of 44100 samples per second (CD quality audio) with time and frequency resolutions of 92 ms and 10.8 Hz respectively.

To reduce artifacts during source reconstruction and to improve time resolution without affecting the frequency resolution, overlapping of frames by a specific sample length is an effective method. Choosing this overlap length and the number of frames to be processed involves another trade-off between the required buffer storage and the processing time. An overlap of five consecutive frames is chosen where each frame is displaced from the previous one by 1024 samples, implying that successive frames share an overlap of 3072 samples, giving a time resolution of 23.2 ms. Figure 3 shows the left and right channels of an audio frame consisting of 4096 samples of a stereo signal.

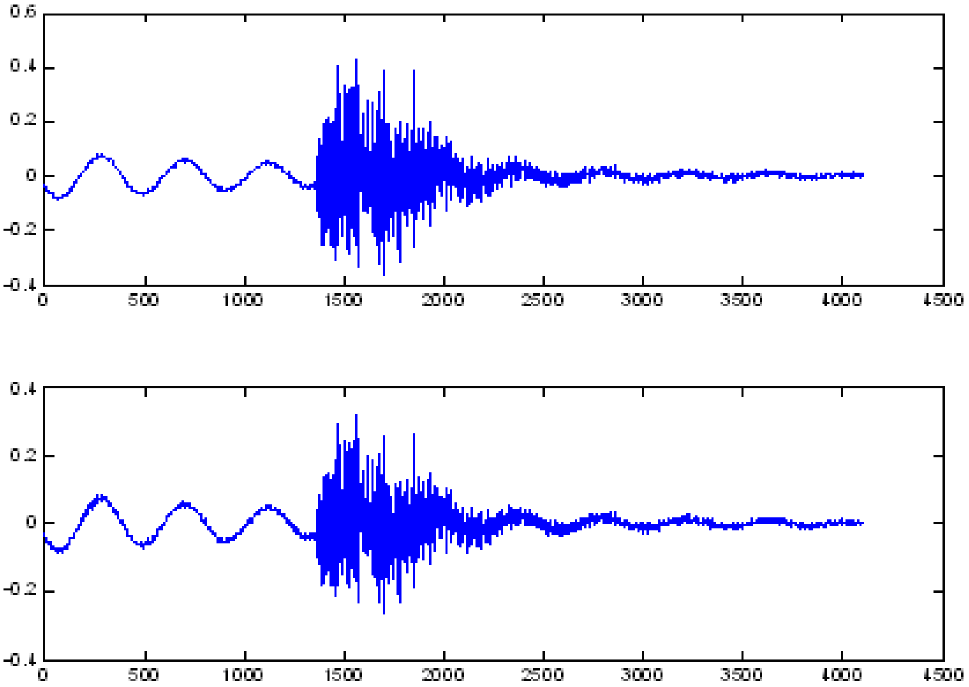

Figure 3: The left and right channels of a 4096 sample audio frame of a stereo signal

4.2.2 Windowing and Fourier Transform

The Discrete Fourier Transform is the process of applying the Fourier Transform of a periodic signal coincident with the signal being transformed in an integer period. In cases of non-integer periods, the end points of the spectrum are discontinuous resulting in high side lobes or what is known as spectral leakage. Windowing attenuates these artifacts by connecting the endpoints of the spectrum in a more smooth fashion before computing the FFT resulting in better spectral resolution. This is achieved by attenuation of the first and last samples of a signal resulting in improved estimation of the frequency spectrum while compromising on the information provided by these samples. This effect is further reduced in case of frame overlapping, as the attenuated end samples occur in the middle of successive frames thereby reducing the loss of information. In this research, the Hann window is employed as it provides a good trade-off between the main lobe width and side lobe energy. It should be noted that other windows can also be used to perform windowing but the Hann window is preferred due to its standard use in source separation research. Figure 4 shows the left and right channels of an audio frame after windowing. As a result, the first and last samples of the signals are attenuated.

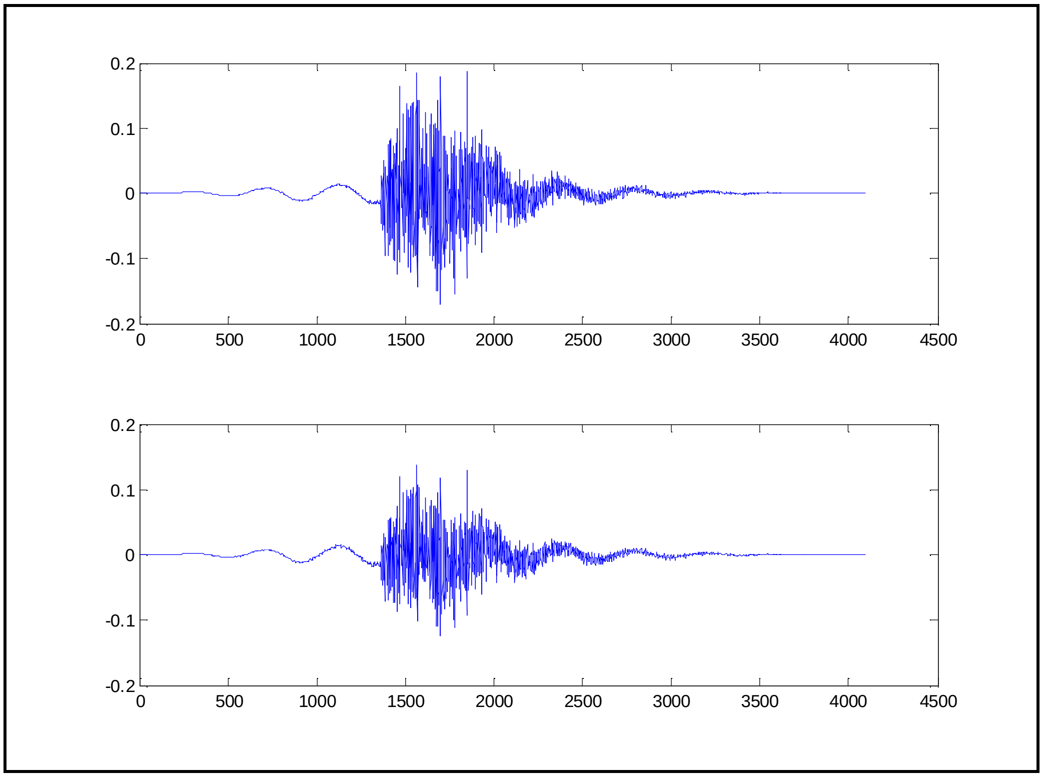

Figure 4: Attenuated first and last samples of an audio frame after windowing

The Fourier Transform is then computed for both the channels of the windowed signal using the FFT algorithm resulting in a complex function of frequency. The following figure shows the magnitude of the Fourier Transform of the left and right channels of the audio frame. Figure 5 shows the frequency spectrum of the audio frame after windowing.

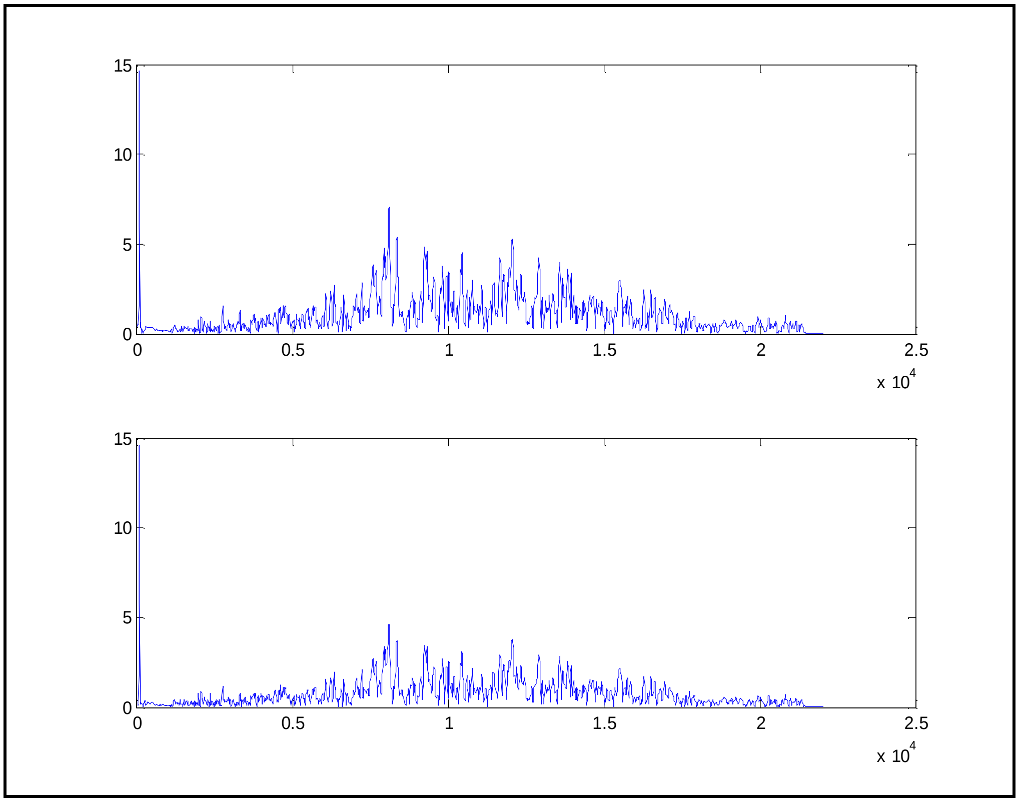

Figure 5: Frequency spectrum of the left and right channels of a windowed audio frame

4.2.3 Pan Estimation Per Frequency

In this step, a pan position is detected and estimated for each frequency under the assumption that the stereo signal consists of mono source tracks mixed to stereo where each source is panned differently. The pan estimation process is done using equations 3.23, 3.24 and 3.25. But in real world scenarios where there are a number of sources with overlapping frequencies, the pan estimation for each frequency is an approximate reconstruction of the actual value applied during mixing. From Figure 6 it can be inferred that interpretation of a particular pan for each frequency is complicated due to the overlapping of various pan values in adjacent frequencies. This leads to the approximation of pan values. This overlap is largely influenced by the number of sources present in the stereo signal, their proximity in the stereo field and their performance overlap in the time-frequency domain.

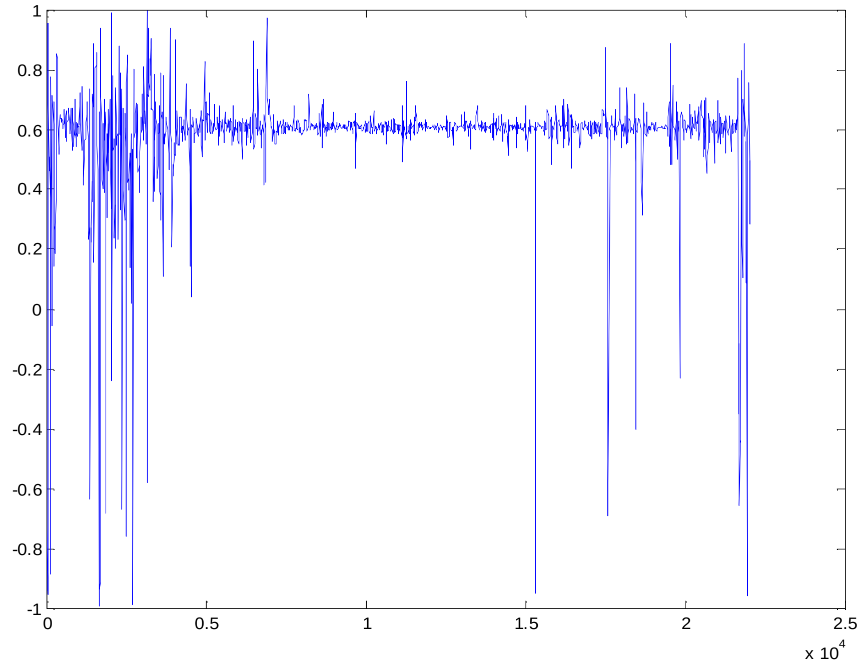

Figure 6: Estimated Pan versus Frequency in an audio frame

4.2.4 Source Estimation

Using the solutions obtained in Equations 3.17 and 3.18, the magnitude of the source is also estimated for each frequency. Due to the presence of many sources in most commercial music recordings, the minimum estimated at each frequency is not zero as shown theoretically in equation 3.18, but is a remainder that consists of noise and other artifacts caused due to other sources operating at the same frequency. Figure 7 shows the modulus of the estimated source plotted for each frequency.

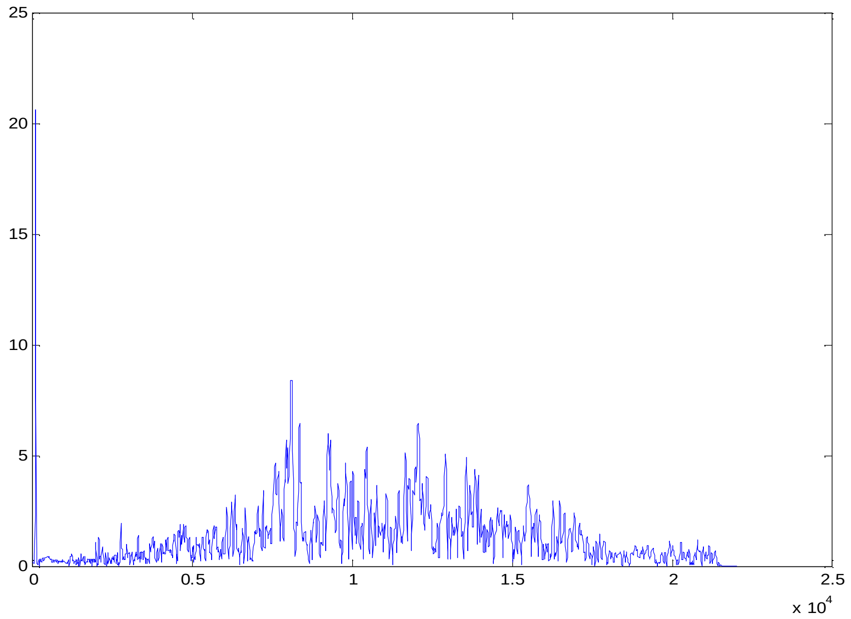

Figure 7: Modulus of estimated source at each frequency

4.2.5 The Frequency-Azimuth Plane

Based on the information obtained in the previous steps, a frequency-azimuth plane similar to the plane employed in the ADRess algorithm is constructed for better visualization of the location of the different sources panned in the stereo field. But unlike the ADRess frequency-azimuth plane, the plane in this thesis displays the unified left and right channels of the stereo signal along with the pan positions of all the sources present in it in their respective frequency ranges. Figure 8 shows the frequency-azimuth plane for a stereo mixture. The plane displays both the left and right channels together. The pan axis ranges from 0 to 1, the frequency axis ranges from 0 to 12 kHz and the magnitude axis ranges from 0 to 3. The plane displays the magnitude peaks of all the sources in the stereo mixture corresponding to their location in the stereo field.

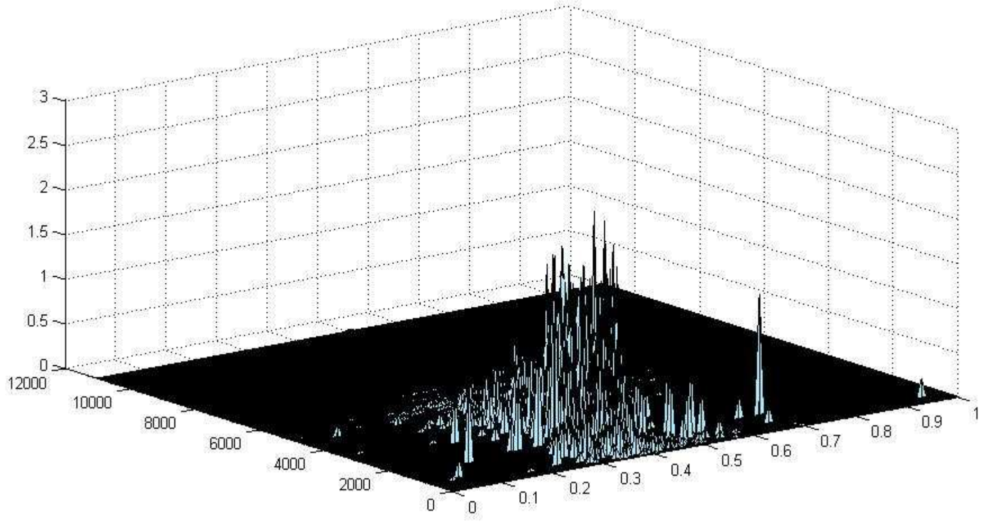

Figure 8: Unified frequency-azimuth plane displaying both left and right channels

It can be seen from Figure 9 that there are about two or three sources at pan values around 0.45, 0.5 and 0.6 (where 0.5 corresponds to centered signal). Sources with negative pan values are also present and this phenomenon is due to artifacts resulting from two or more sources combining at the same frequency point.

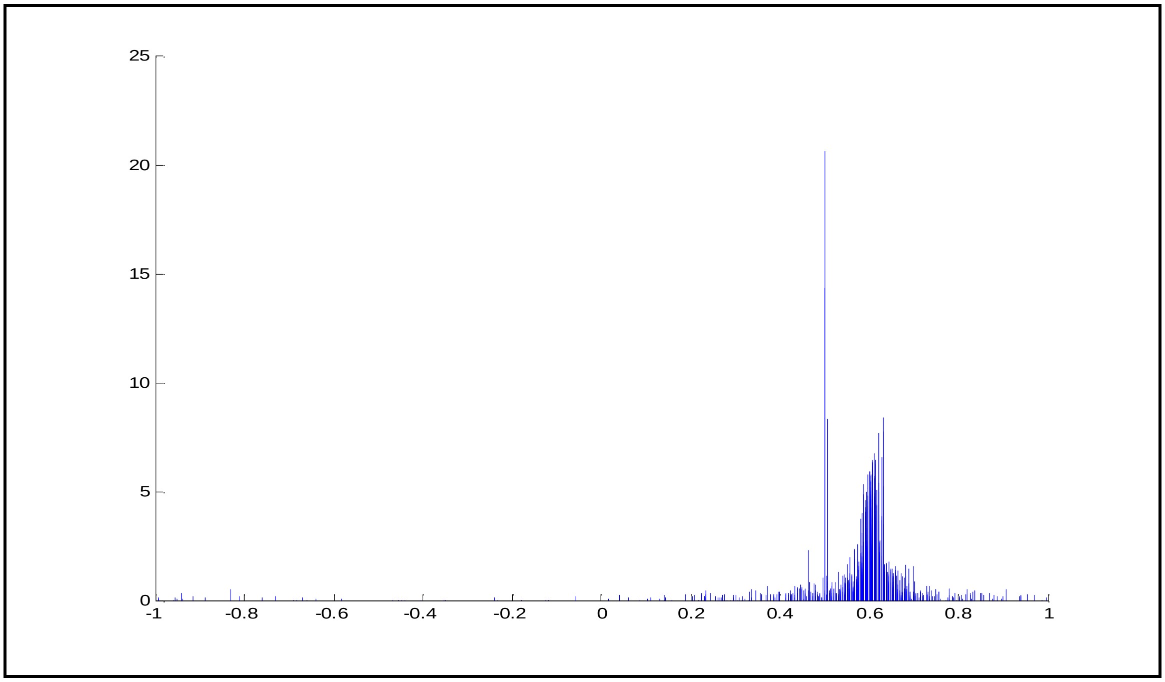

Figure 9: Frequency-Azimuth plane (2-D); x axis - pan position, y axis – source magnitude

4.2.6 Pan Values Windows Selection

Once the pan values at each frequency have been determined along with the source magnitudes, pan windows are calculated for each frequency that corresponds to a range of pan values that belong to a particular source. The objective of this step is to determine “meaningful” pan information that would help in separating the sources more effectively. It is easy for a human being to observe the frequency-azimuth plane and determine such pan information, but the task is complicated if the process has to be an automatic detection. To achieve this, an envelope detector is employed which smoothens the multiple peaks of different sources that appear in the frequency-azimuth plane. Figure 10 shows a magnified version of the plane for pan values between 0.4 and 0.7 where high and low amplitude levels of sources at a very close range of pan values can be observed.

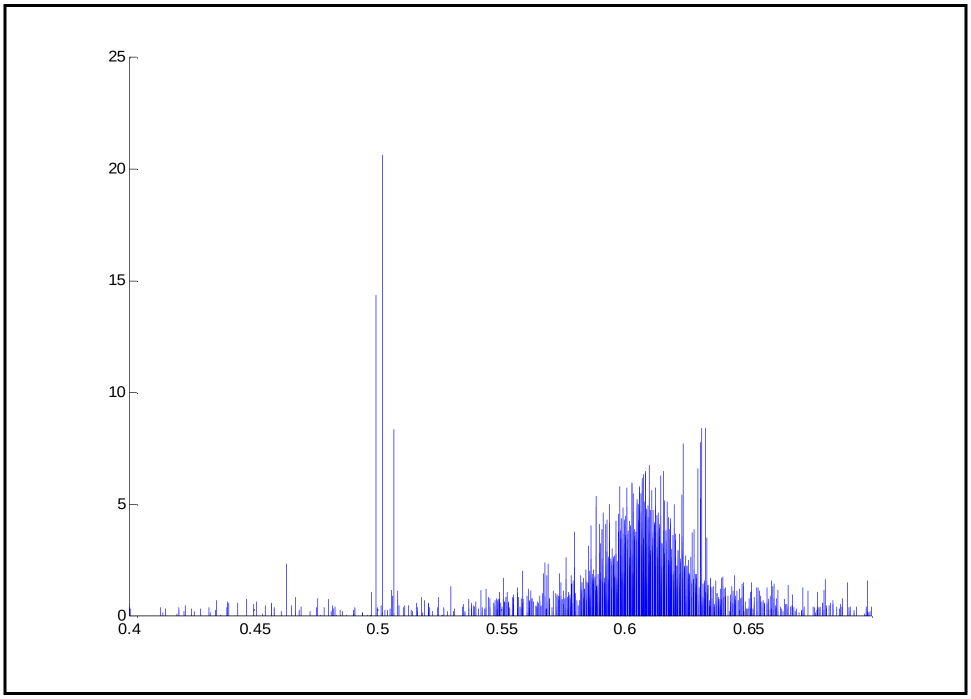

Figure 10: Magnification of the plane for pan values between 0.4 and 0.7

The plot can also be visualized as an envelope formed by the various peaks that belong to all the sources. This envelope information can be extracted from the plane by using a peak detector that operates on the same principle of a traditional AM-demodulator.

The implementation of the peak detector consists of two steps:

1) Discard all pan values that are negative or greater than one as these cannot correspond to the pan values applied by the pan-pot.

2) Arrange the left over values in ascending order of estimated pan position keeping as reference the frequency at which the pan was estimated and the reconstructed source magnitude at the same frequency.

A function (source magnitude versus pan position) is created. Two envelope detectors are applied at first, one causal and the other anti-causal. The terms causal and anti-causal are strictly with reference to pan position as the causal detector moves from low to high pan and the anti-causal from high to low pan. The envelope detectors output at each pan value the maximum value between the source magnitude and an exponentially decaying function which takes the output at the previous pan position and decreases it exponentially. The resulting and final envelope detector output is the maximum between the causal and anti-causal envelope detectors. This is done to eliminate any irrelevant local minima that may be present during processing. Figure 11 shows the output envelopes generated by the causal and anti-causal detectors.

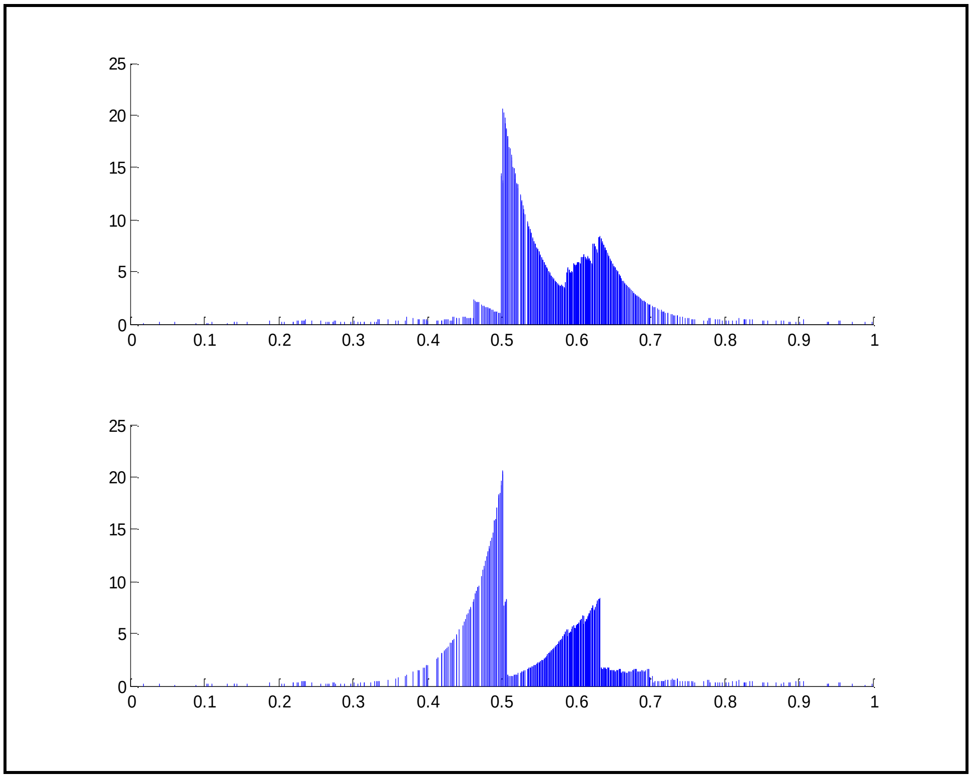

Figure 11: Output of the causal (up) and anti-causal (down) envelope detectors

Figure 12 displays the final combined output of the two detector outputs from Figure 11.

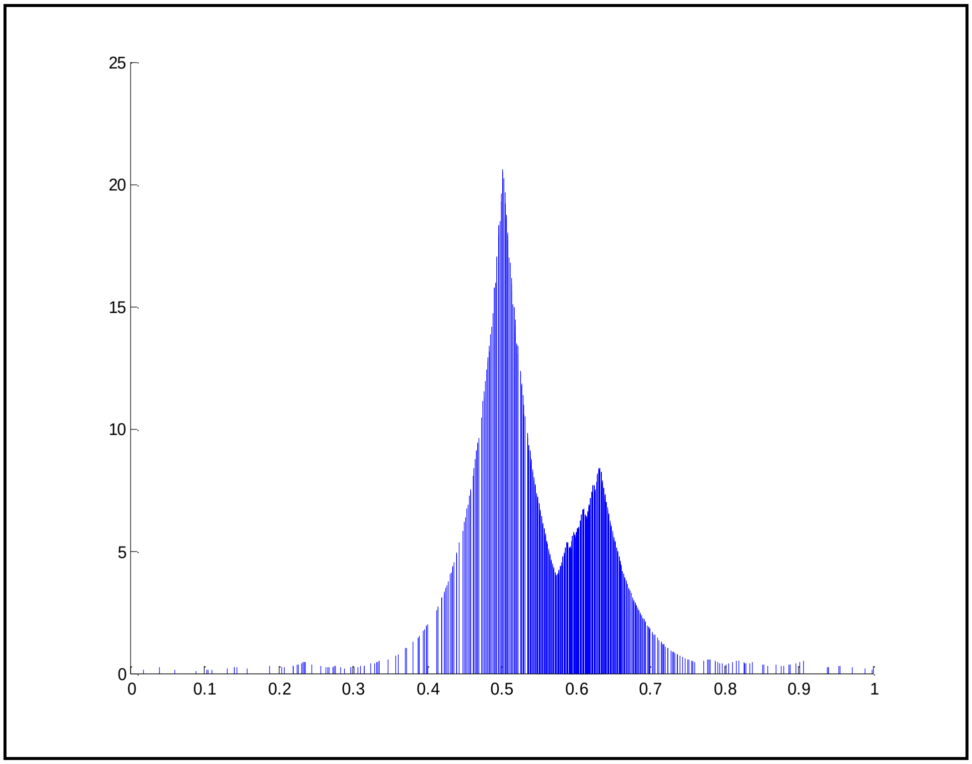

Figure 12: Final combined output from causal and anti-causal detectors

Figure 13 is a magnification of the region in Figure 12 for pan values between 0.4 and 0.7. The curves are smoother due to the effect of the envelope detectors:

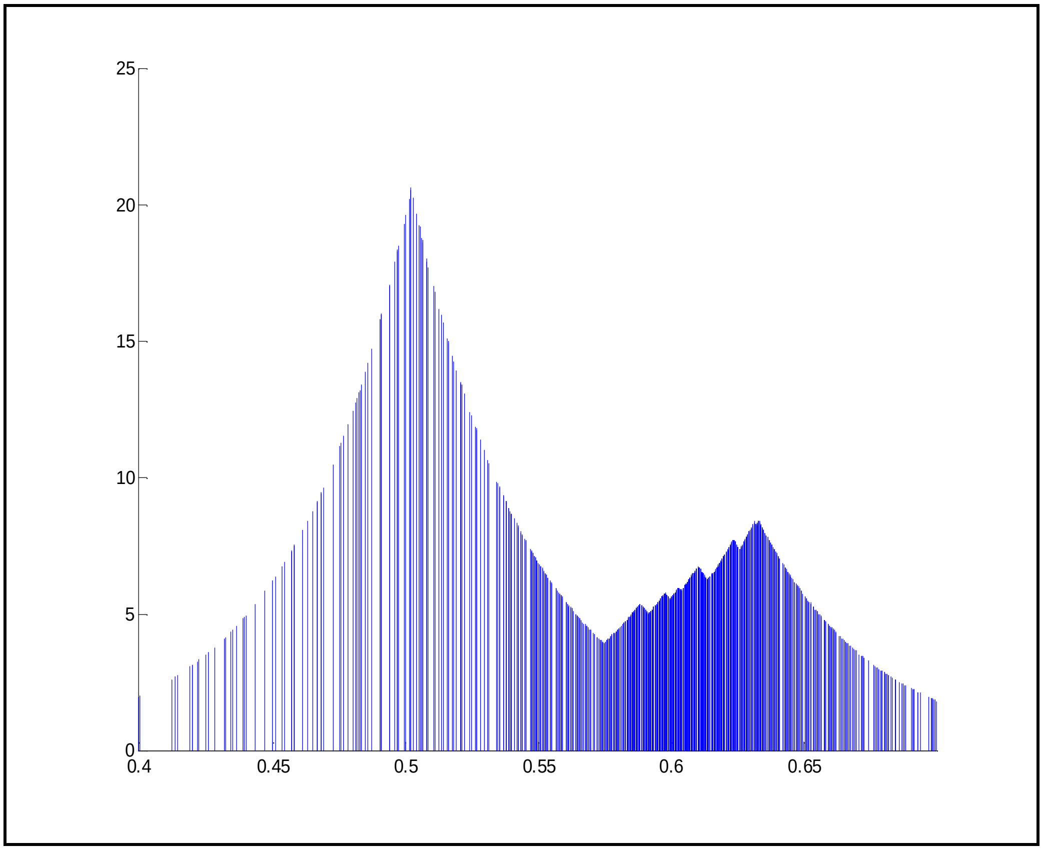

Figure 13: Magnification of the smoothened envelope for pan values between 0.4 and 0.7

The relevant pan windows can be detected by searching for local maxima (window center) and local minima (window width) of the envelope. Irrelevant peaks do occur in most cases due to the effect of two or more sources at a certain frequency or due to artifacts and noise produced from stereo processing. These effects can be minimized to a certain degree by considering maxima only above a certain threshold value. This threshold value is calculated as the mean plus the standard deviation of the envelope. The audio frame in Figure 13 results in the following detected pan values for three relevant peaks as given in Table 4:

Table 4: Detected Pan Windows

Peak #	**Window Center**	**Window Width**	**Left Limit**	**Right Limit**
1	0.4404	0.1341	0.3063	0.5745
2	0.6288	0.0032	0.6256	0.6320
3	0.6964	0.0644	0.6320	0.7608

The pan window width cannot be determined accurately in case of most real world signals because the estimated pan values have deviations around the real pan value due to interference with other sources at the same frequency, noise and artifacts due to stereo processing as mentioned before. The complexity in selecting the window width lies in capturing most frequencies that belong to a single source minimizing the extraneous effects. Since the approach is unassisted, it is very difficult for an algorithm to “know” what width is to be considered for the pan window to achieve efficient selection. The pan values that are finally accepted are assumed to be the values applied to one source in that particular window. This is the most crucial step in the source separation algorithm as this is decisive in what the algorithm considers as “meaningful” sources to be separated.

On manually verifying the values obtained in Table 4 with Figure 13 it is observed that the detection of pan values is good but not perfect. The small irregularities on the left slope of the second peak affected the detected pan values generating spurious values in the process. Summarizing the outputs of the audio frame-by-frame analysis stage, the outputs are:

1) Estimated pan value for each frequency,

2) Estimated source magnitude and phase for each frequency and

3) Estimation of pan windows where relevant sources are found.

4.3 Pan Tracking

The results of the previous stage serve as the input for tracking the pan windows from frame to frame which is instrumental to the source separation and reconstruction process. Figure 14 shows the various blocks that comprise the pan tracking and source reconstruction stage of the separation algorithm.

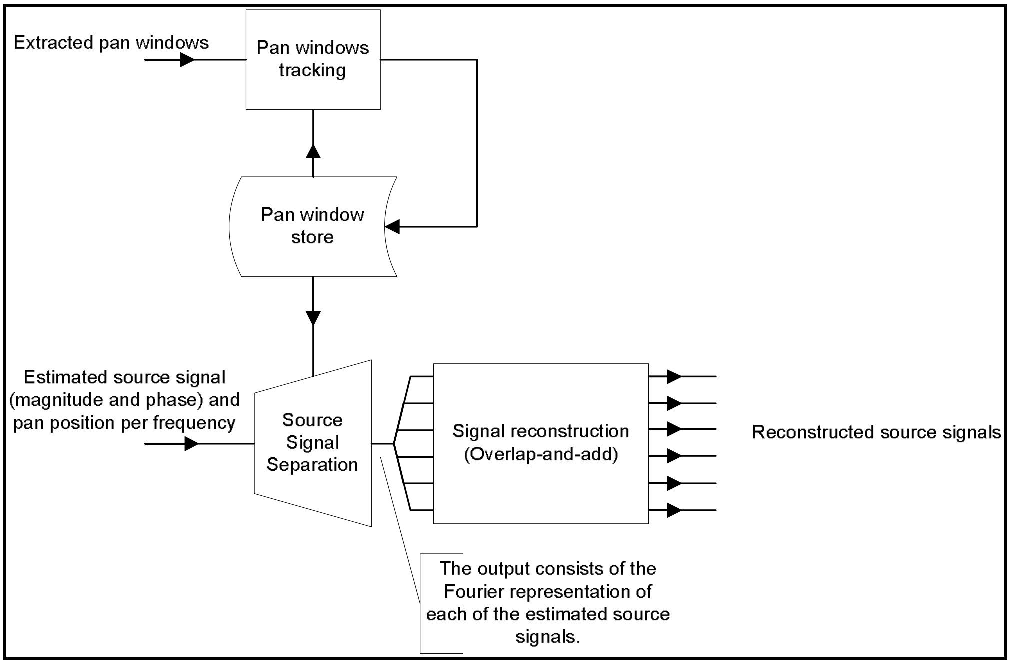
Extracted pan windows
Pan windows tracking
Pan window store
Estimated source signal (magnitude and phase) and pan position per frequency
Source Signal Separation
Signal reconstruction (Overlap-and-add)
Reconstructed source signals
The output consists of the Fourier representation of each of the estimated source signals.

Figure 14: Block Diagram for the Pan Tracking and Source Reconstruction stage

4.3.1 Pan Window Store

This block forms the main component of the pan tracking and source reconstruction stage. The pan window store block stores all the information about the estimated pan windows that have been detected for each of the sources being processed. For each detected pan window, the following information is used to perform separation; they are referred to by their assigned variable names:

1. *Estimated pan window center – "azimuth"*
2. *Estimated pan window width – "az_width"*
3. *Partially reconstructed source – "signal"*
4. *Number of frames where the pan window is detected – "nSegments"*
5. *Previous frame where the same pan window was last detected – "lastFound"*

For every frame that is processed, the pan windows store is updated with the estimated pan windows information from the previous stage thereby providing details needed to perform source reconstruction and separation.

4.3.2 Pan Windows Tracking

For each frame that is processed, this block is responsible for updating the pan window store information. This is done using the current values available in the pan windows store along with the estimated pan windows information from the audio frame-by-frame analysis stage thus acting as a feedback loop. The *lastFound* variable acts as an indicator whenever a new frame is processed in search of a new pan window.

Each estimated pan window from the previous stage is analyzed to find an overlap with the pan window information available in the pan windows store block; this is done by comparing the corresponding *azimuth* and *az_width* variables. If an overlap is found, the pan windows store variables are updated as follows:

- A new pan window center is computed by applying an auto-regressive filter to the stored information and the incoming information:

$$\text{azimuth}_{\text{new}} = (1 - \alpha).\text{azimuth}_{\text{old}} + \alpha.\text{azimuth}_{\text{detected}} \qquad (4.1)$$

This allows the tracking of the pan window if the pan position of a source changes over time.

- A new pan window width or *az_width* is computed using an auto-regressive filter using the stored and incoming information as in the above case.

- The *lastFound* variable is set to zero to indicate that the particular pan window under inspection was found in the frame being processed.

- The *nSegments* variable is incremented as this counts the number of frames in which the particular pan window was found.

If an overlap does not occur, this would imply that the incoming information belongs to a newly detected source, and a new entry is made in the pan windows store:

- *azimuth* and *az_width* are set equal to the incoming values.

- *lastFound* is set to zero.

- *nSegments* is set to one.

This process acts as a continuous update routine as incoming frames are being analyzed. It helps to store the latest information on all pan windows detected up to the current frame. The autoregressive filter applied to the pan window center and pan width helps to combine past information obtained from each pan window to the present incoming information from the audio frame being processed. The parameter ‘α’ controls the relative weight of the past versus the current information. This implies that if α is close to zero the output is more influenced by past values, and if it is close to one then the output is more influenced by current values.

4.4 Source Separation and Reconstruction

The data for every frequency is sorted according to the pan position, as this corresponds to the various sources present in the stereo mixture. For every frequency where the estimated pan position falls within a particular window in the pan window store it is assigned to an estimated source. If some estimated pan positions do not fall within any of the windows, then those frequencies do not get assigned to any source and are neglected. This procedure is carried out until all frequencies and pan windows are exhausted. This results in a number of sources separated from each other based on their pan information. In essence, if the different sources in a stereo mixture are panned differently from each other, the sources are separated after this step.

The separated sources are then reconstructed to obtain their respective time domain versions. This is done using an inverse FFT for each of the detected pan windows and then adding them together to reconstruct the source using the overlap-and-add technique which applies to four successive frames at a time. The resultant fragments of partial sources are added together to obtain the complete source. The source reconstruction procedure is performed for each of the estimated sources, and as each source carries only certain frequencies the remaining frequencies are set to zero.

The lastFound and nSegments are important variables decisive in the reconstruction of meaningful sources. A source is reconstructed only if it is detected recently. This is done by comparing the lastFound value with a threshold value (presently set at 3) and updating only those sources where this value is less than the threshold. This ensures that spurious signals that may occur sporadically are not treated as sources. The sources that are not detected in enough frames are discarded by comparing the nSegments value with a threshold (presently set at 5). This implies that for a signal to be considered as a meaningful source it must be present in the recording for a reasonable duration thereby minimizing the possibility of treating noise and artifacts as sources.

5

Performance Analysis

The performance of the algorithm was tested on stereo music files that were of CD quality audio. Selection of the test data was an important factor as this would shape the assessment of the algorithm and promote better understanding of its working. Though the objective of this research is not to propose an improved successor to previous attempts at separating music, an algorithm's effectiveness can be evaluated by using test data that have been employed in other research. Since there has been little research dedicated towards stereo music, the set of available test data was small. For this reason, music files that were used to study the algorithm's performance in [6] and [7] were selected due to similarity in the research framework and nature of the assumptions. Five test files were selected for the algorithm to perform source separation on, in six experiments.

5.1 Experiment 1

Name of Music Track/Artist: In a sentimental mood/John Coltrane

Duration of test file: 12 seconds

Genre: Jazz

Type of sources: Saxophone, Piano, Drums, Bass

The algorithm was able to detect and separate three of the four available sources. Along with the main sources, the algorithm generated three spurious signals that consisted primarily of artifacts and noise. Figure 15 shows the left and right channels of the original stereo mix consisting of the four sources.

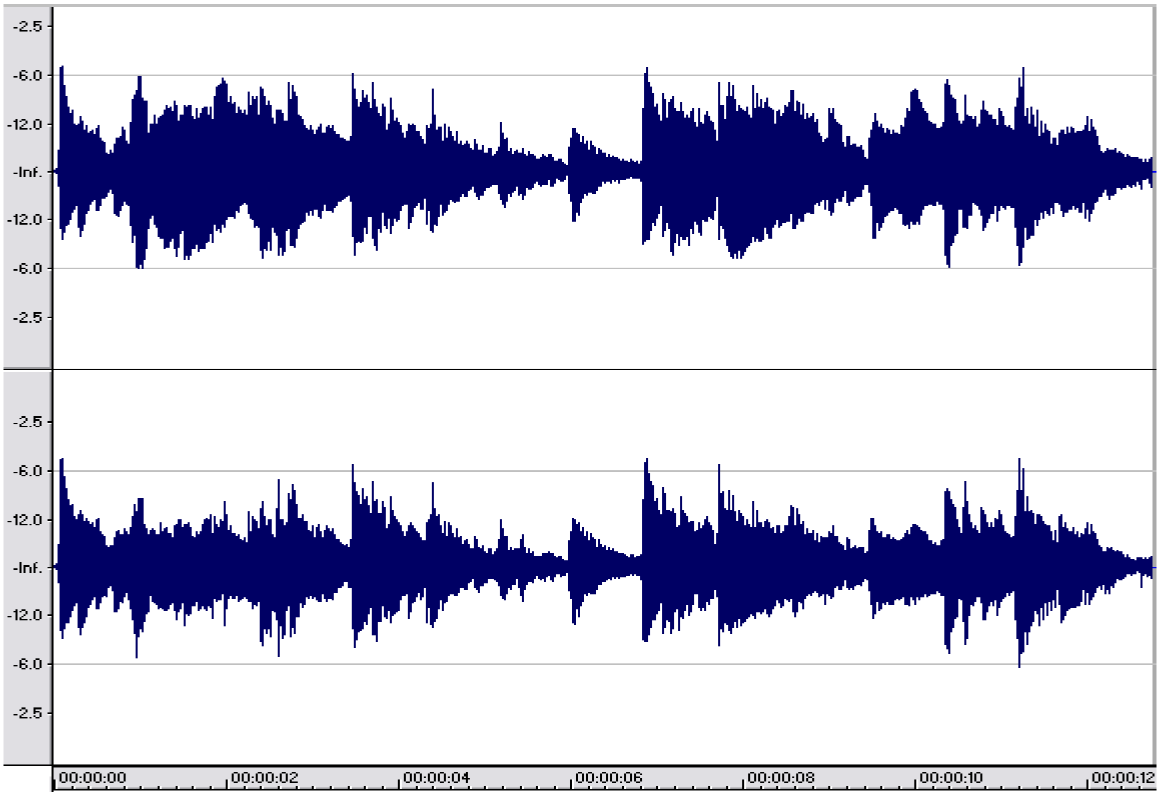

Figure 15: Left and right channels of the original track's waveform in experiment 1

Figure 16 shows the separated saxophone mono track. Since the saxophone was played intermittently in the original, there are portions where there are no peaks in the separated track which signify moments of silence. This also implies that the saxophone was separated without picking up noise from other instruments or artifacts.

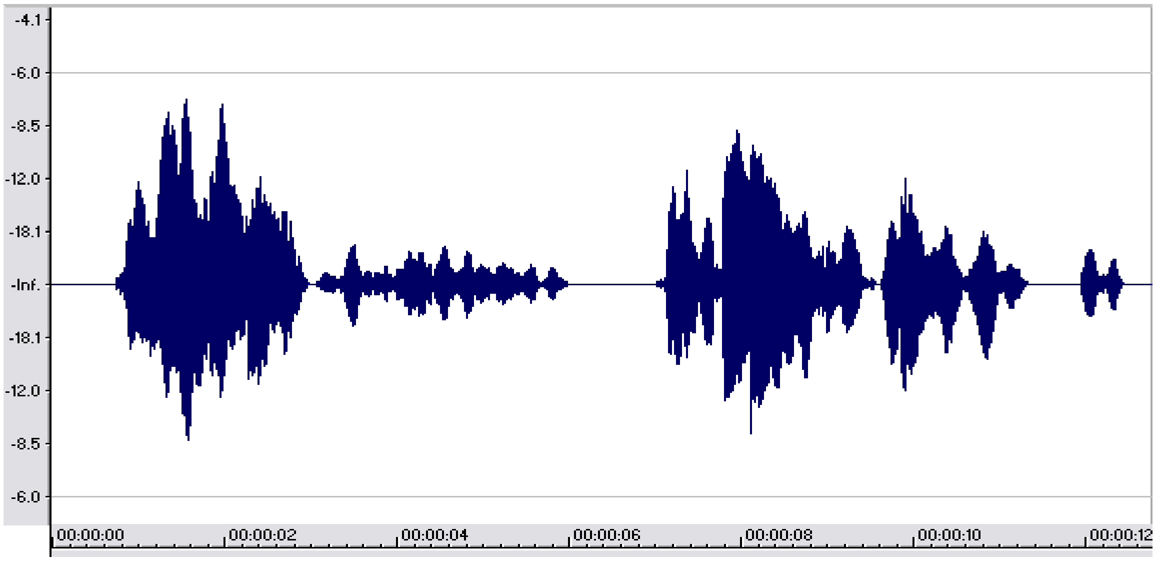

Figure 16: Separated Mono Track – Saxophone

Figure 17 shows the separated piano mono track. The separation quality of the piano was less when compared to the saxophone and the presence of the bass and drums was high. But it can be observed that the prominent peaks of the piano were well separated.

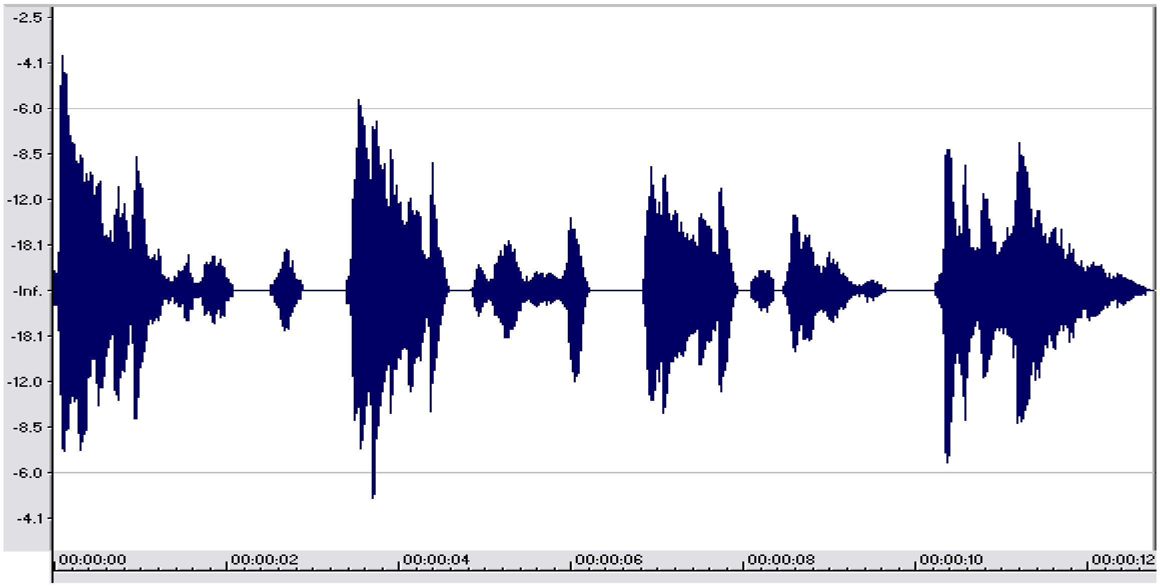

Figure 17: Separated Mono Track – Piano

Figure 18 shows the separated drums mono track. It can be observed from the waveform that the drums were not separated completely. The presence of the other instruments was highest in this case along with noise and artifacts.

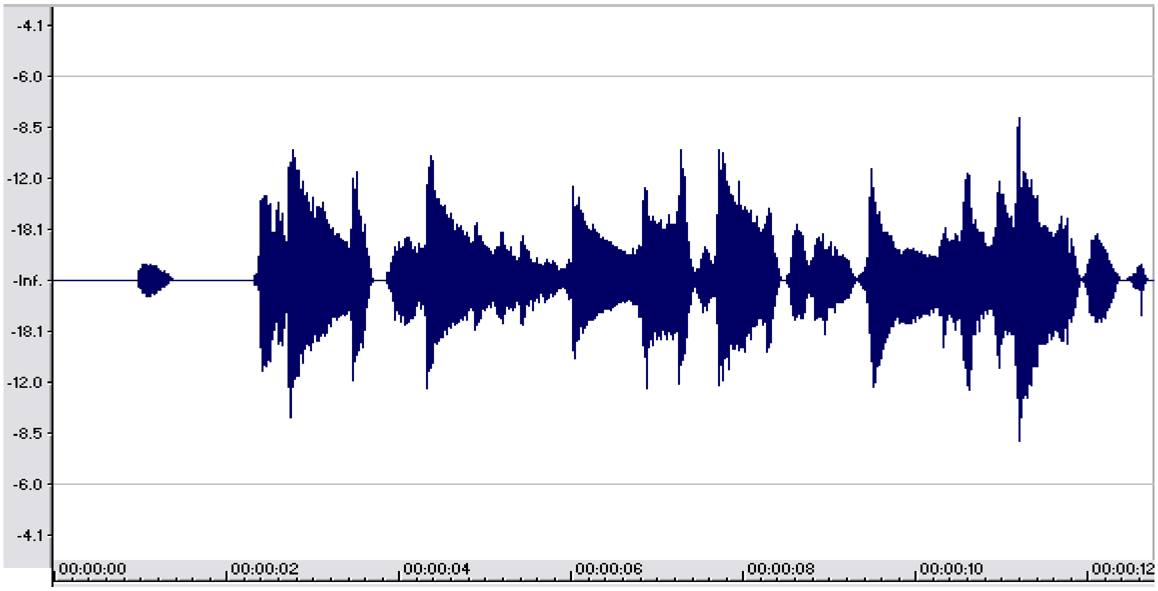

Figure 18: Separated Mono Track – Drums

5.2 Experiment 2

Name of Music Track/Artist: Wind Cries Mary/Jimi Hendrix

Duration of test file: 17 seconds

Genre: Rock

Type of sources: Lead Guitar, Bass Guitar, Drums

The algorithm detected and separated the lead guitar. But in this case, a total of eleven sources were separated; four of which were of the lead guitar and the remaining were spurious signals. From hereon, all the separated sources other than the spurious signals shall be referred to as “un-clustered” sources. Figure 19 shows the left and right channels of the original stereo mix consisting of three sources. When compared to Figure 15, it can be observed that the overlap between the sources is higher in this case making separation more difficult. The drums was played using a 4/4 rhythm pattern which can be observed from the periodic distribution of certain peaks in the waveform.

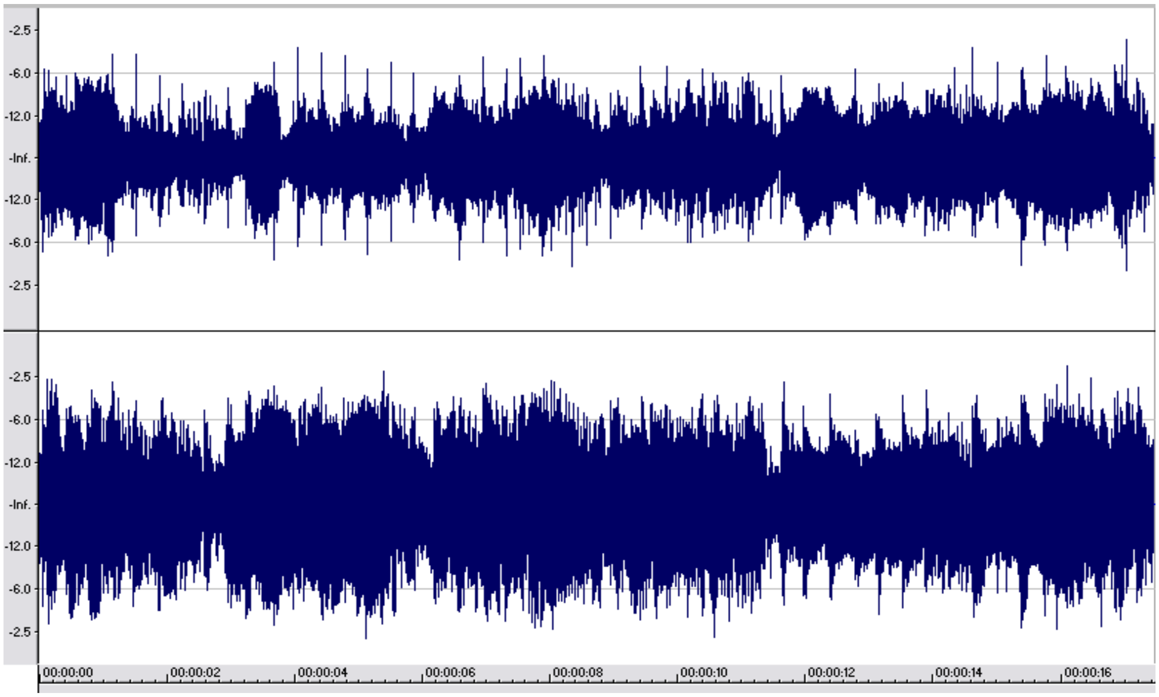

Figure 19: Left and right channels of the original track's waveform in experiment 2

Figure 20 shows the first separated un-clustered guitar track. It can be observed that the peaks were of lesser amplitude when compared to the original mix. On listening to the track, it was seen that the guitar track was not separated completely and was missing certain notes when compared to the original.

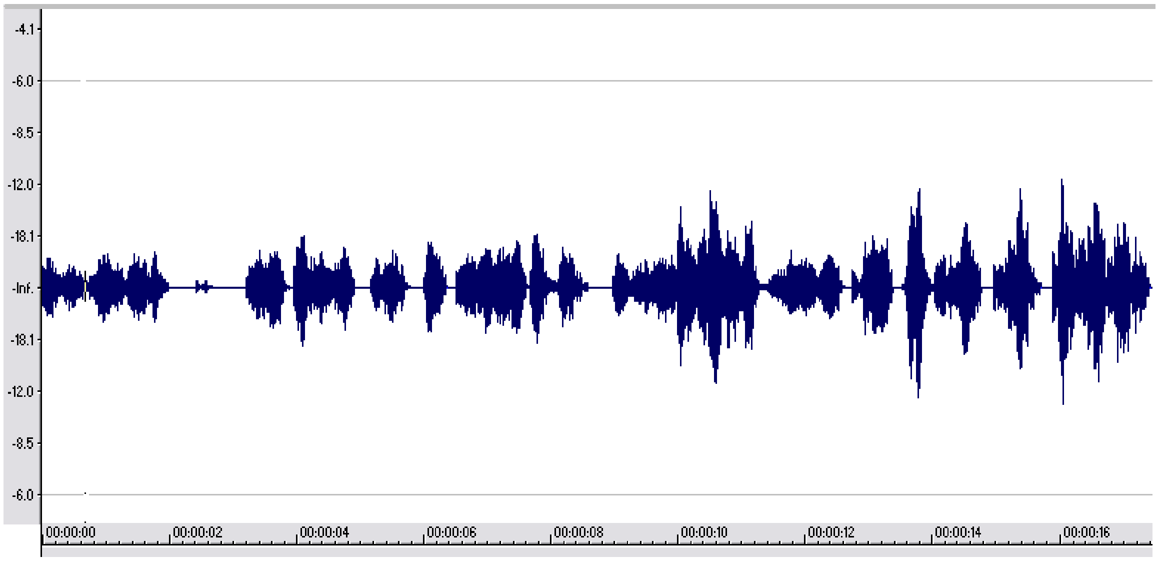

Figure 20: 1st un-clustered separated mono track - Lead guitar

Figure 21 shows the second un-clustered guitar track separated by the algorithm. In this case the amplitude levels were relatively higher than the peaks in Figure 20. This separated track consisted of certain peaks that were not present in Figure 20.

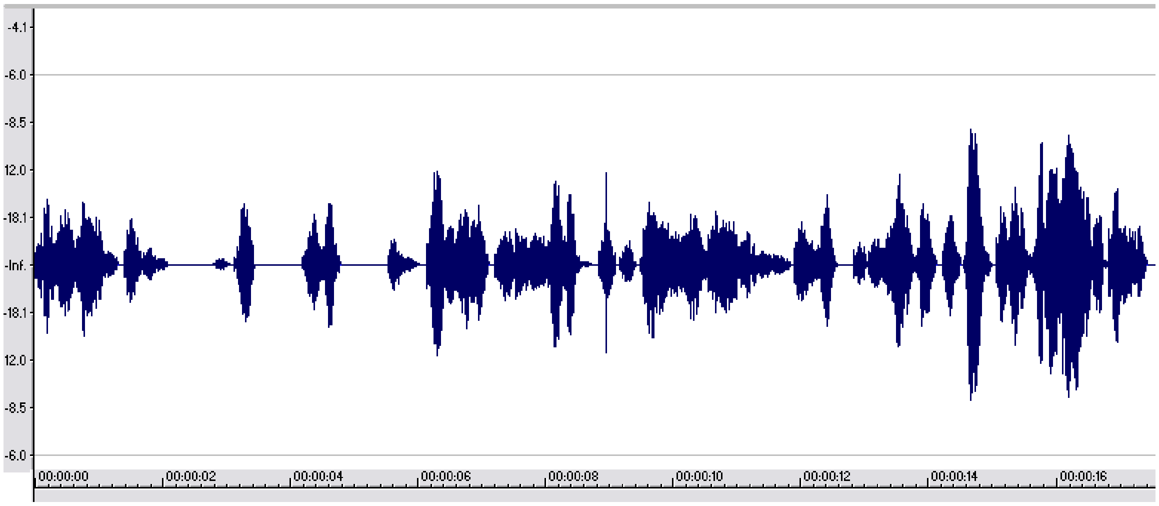

Figure 21: 2nd un-clustered separated mono track - Lead guitar

Figure 22 shows the third un-clustered separated guitar track. As in Figure 21, the amplitude levels were higher in this case, and the waveform shows enhanced peaks that were missing in Figure 20 and 21.

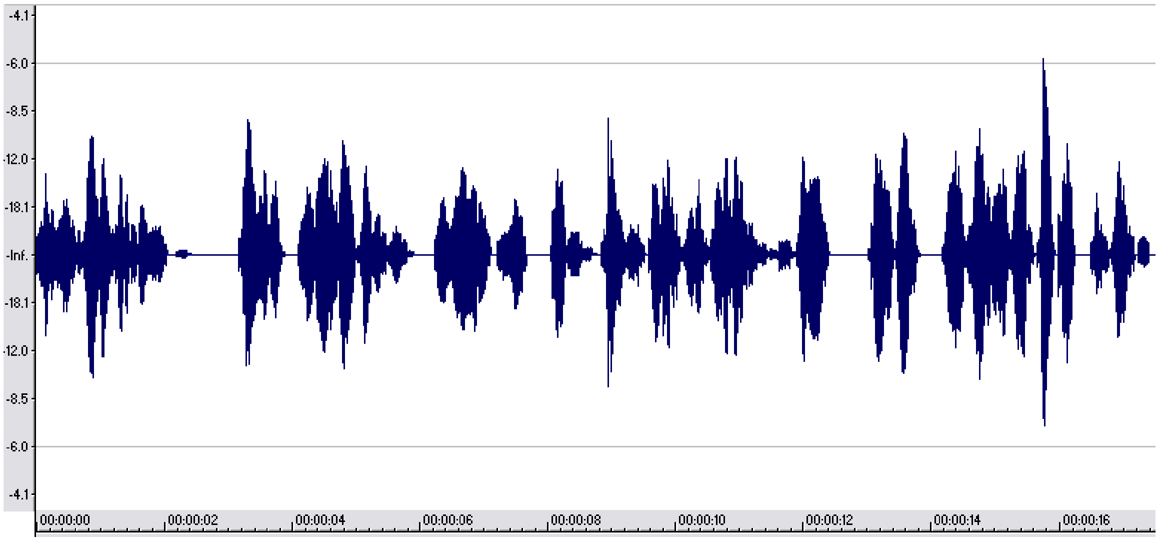

Figure 22: 3rd un-clustered separated mono track - Lead guitar

Figure 23 shows the forth un-clustered version of the separated guitar track. It can be observed that the peaks were most prominent in this case. This track had the best separation quality when compared to the other three un-clustered tracks. Overall, the presence of the other instruments was low with minimum noise, but the effect of artifacts in all the clustered tracks was noticeable.

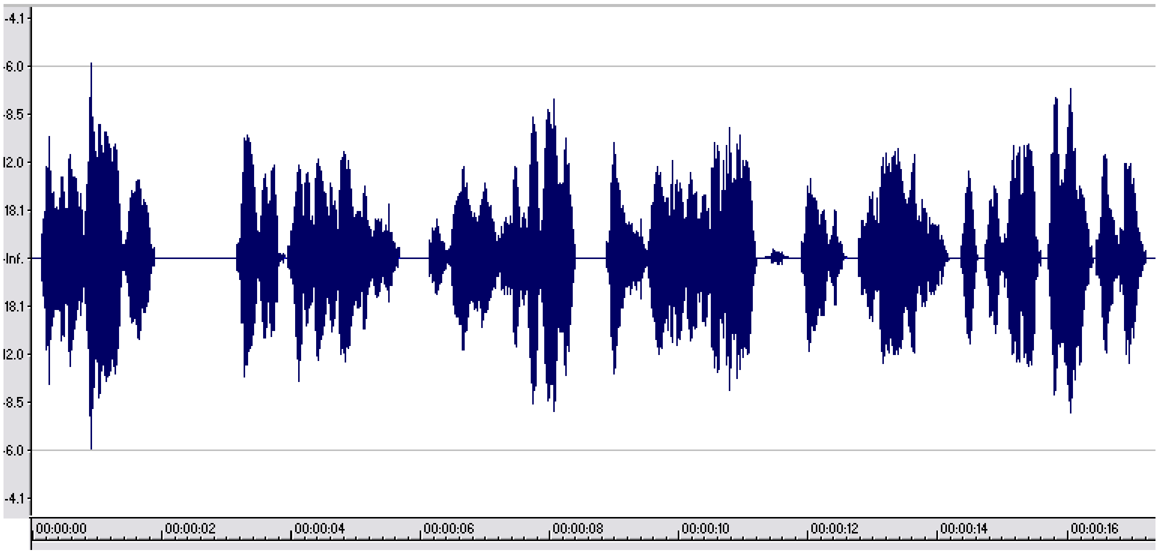

Figure 23: 4th un-clustered separated mono track - Lead guitar

5.3 Experiment 3

Name of Music Track/Artist: Unknown

Duration of test file: 20 seconds

Genre: Rock

Type of sources: Lead Guitar, Piano, Drums, Bass Guitar

The lead guitar and the drums were detected and separated. A total of eleven sources were separated; two un-clustered versions of the lead guitar, two un-clustered versions of the drums, and the remaining seven sources were spurious signals. Figure 24 shows the left and right channels of the original stereo mix. The overlap of instruments was lower than the original mix in Experiment 2, but the drums followed a 4/4 rhythm pattern in this case too which is evident in the waveform.

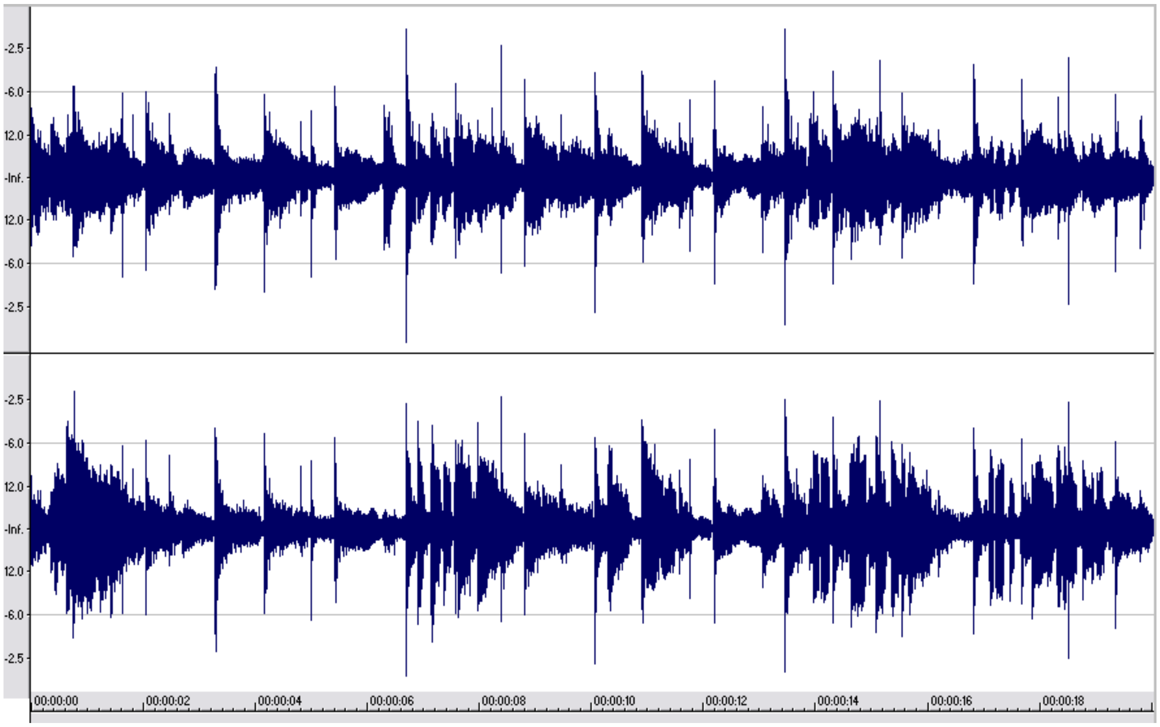

Figure 24: Left and right channels of the original track’s waveform in experiment 3

Figure 25 shows the first un-clustered separated drums track. In this case, the snare and the bass drum were separated efficiently where as the peaks corresponding to the hi-hats were not completely separated.

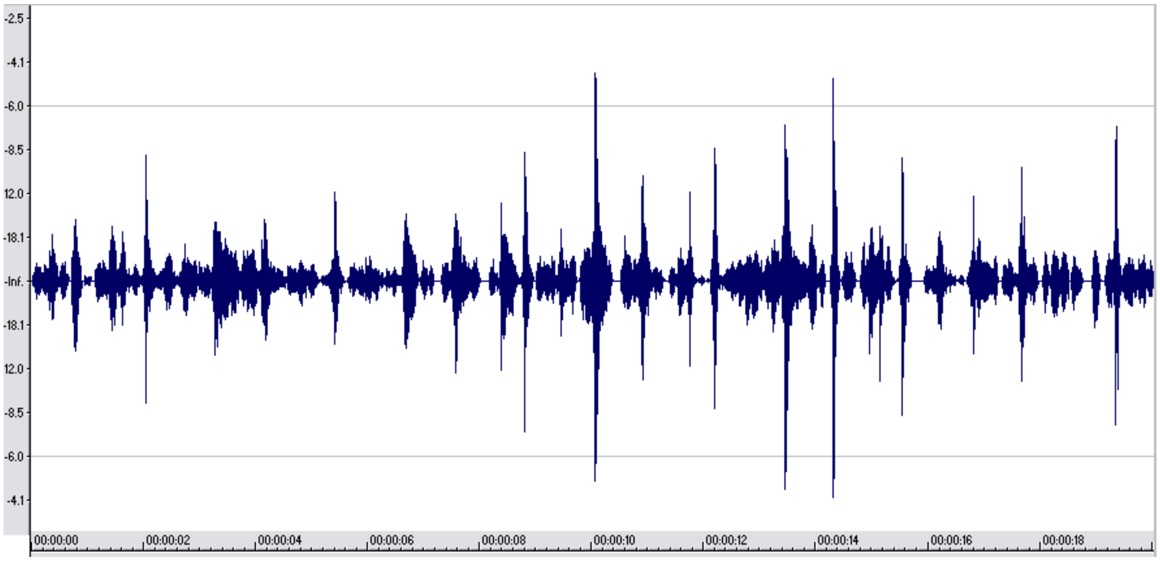

Figure 25: 1st un-clustered separated mono track – Drums

Figure 26 shows that the amplitude levels of the peaks corresponding to the bass and snare drum are comparatively lower than in Figure 25 but the hi-hat has been separated more effectively. In both un-clustered tracks, the overall quality of separation was low due to the presence of other sources and artifacts.

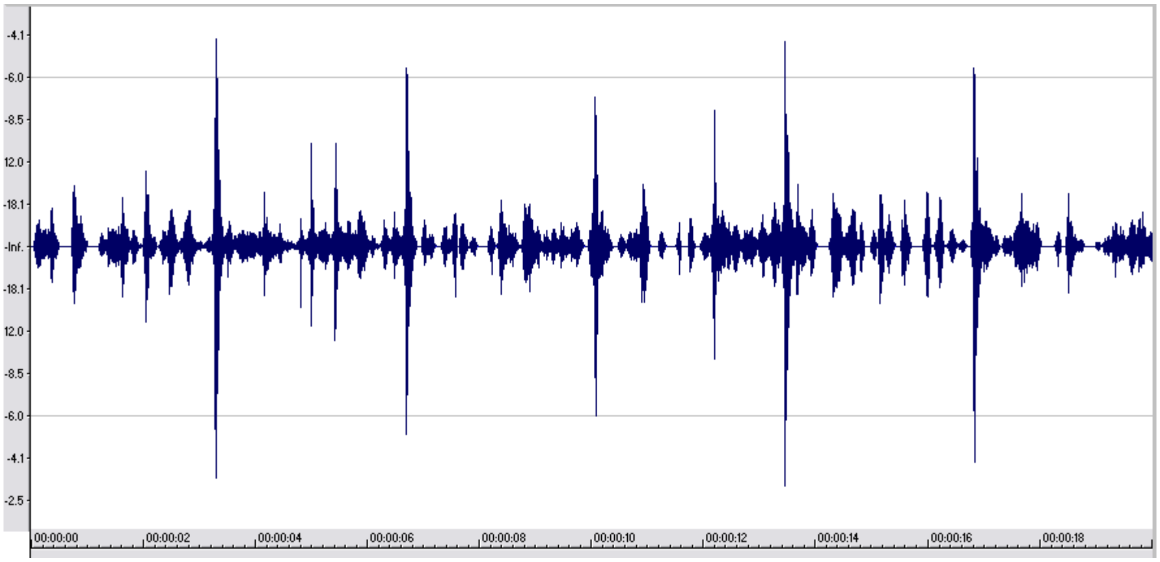

Figure 26: 2nd un-clustered separated mono track – Drums

Figure 27 shows the first separated un-clustered guitar track which consists of the prominent peaks with less noise from the other sources. The amplitude levels of the peaks were comparable to the original mix but were not fully separated.

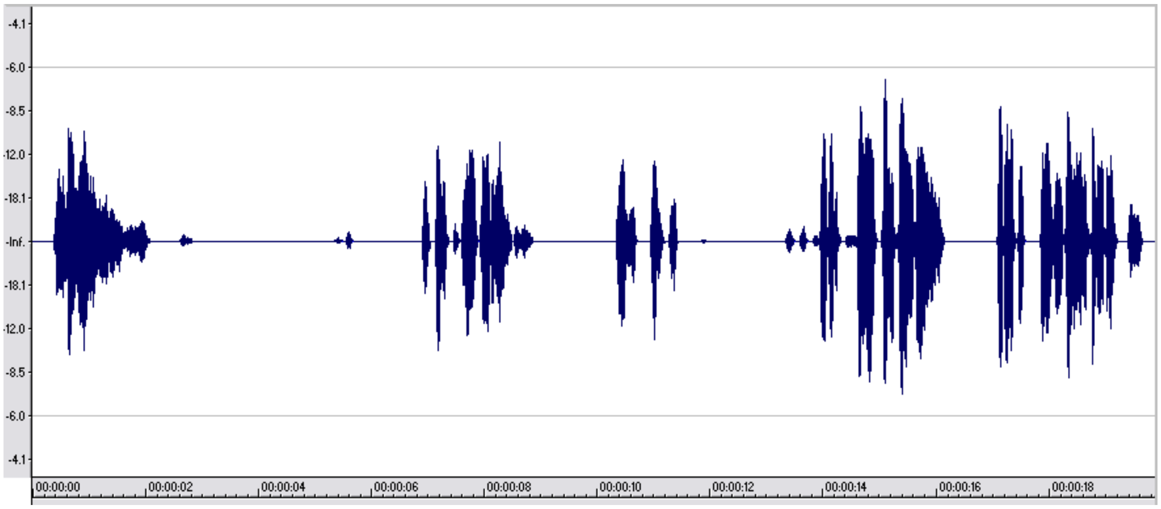

Figure 27: 1st un-clustered separated mono track – Lead guitar

Figure 28 shows the second un-clustered separated guitar track. The separation quality was lower than in Figure 27 due to low amplitude levels and missing peaks. But the presence of other sources and noise was low in this case too.

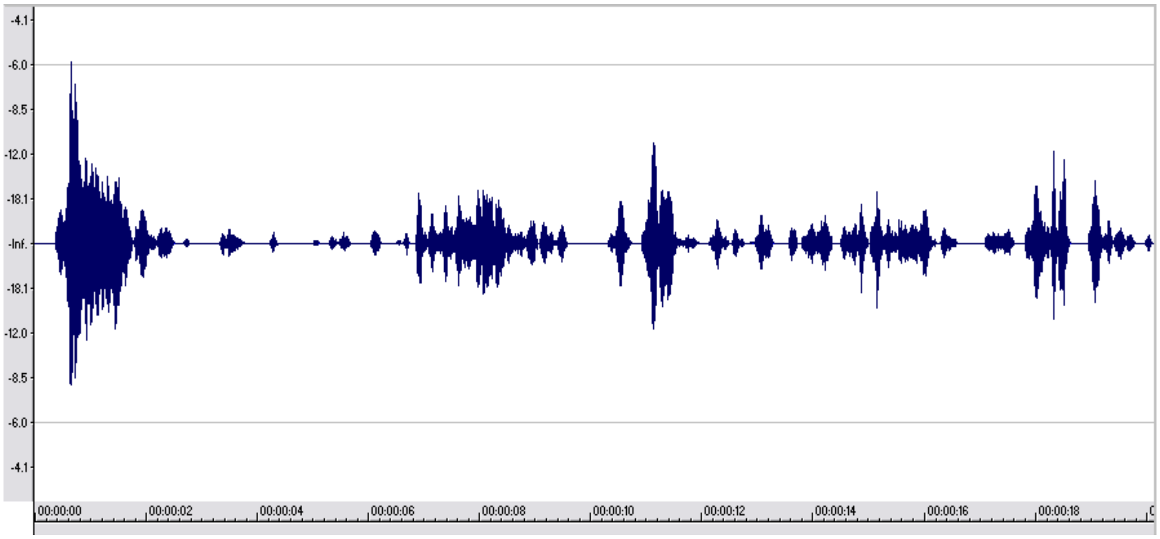

Figure 28: 2nd un-clustered separated mono track – Lead guitar

5.4 Experiment 4

Name of Music Track/Artist: Help/Beatles

Duration of test file: 10 seconds

Genre: Rock

Type of sources: Lead Guitar, Bass Guitar, Vocals, Drums

This track consisted of a lead vocalist with back up singers and also included continuous playing of lead and bass guitars with drums. All the sources could be heard at every second of the recording implying high temporal overlap. The algorithm separated the vocals successfully while generating ten un-clustered versions of which four were spurious in nature. The other instruments could not be recovered due to high level of artifacts and noise. Figure 29 shows the left and right channels of the original mix.

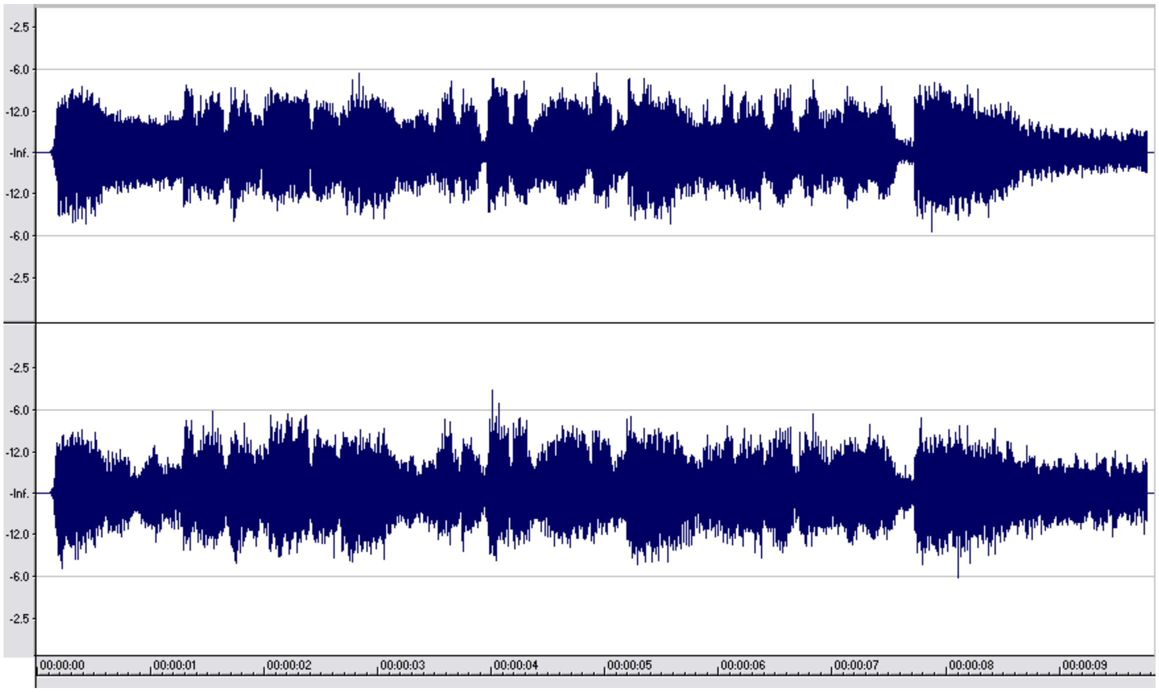

Figure 29: Left and right channel waveforms of the original track in experiment 4

The six un-clustered versions of the vocals contained minimum artifacts and were informally selected as the significant sources. The amplitude levels of the vocals were also higher than in the other un-clustered sources. Figure 30 shows the six separated un-clustered mono tracks of the vocals. In all cases, the lead and back up vocals were present in varying degrees.

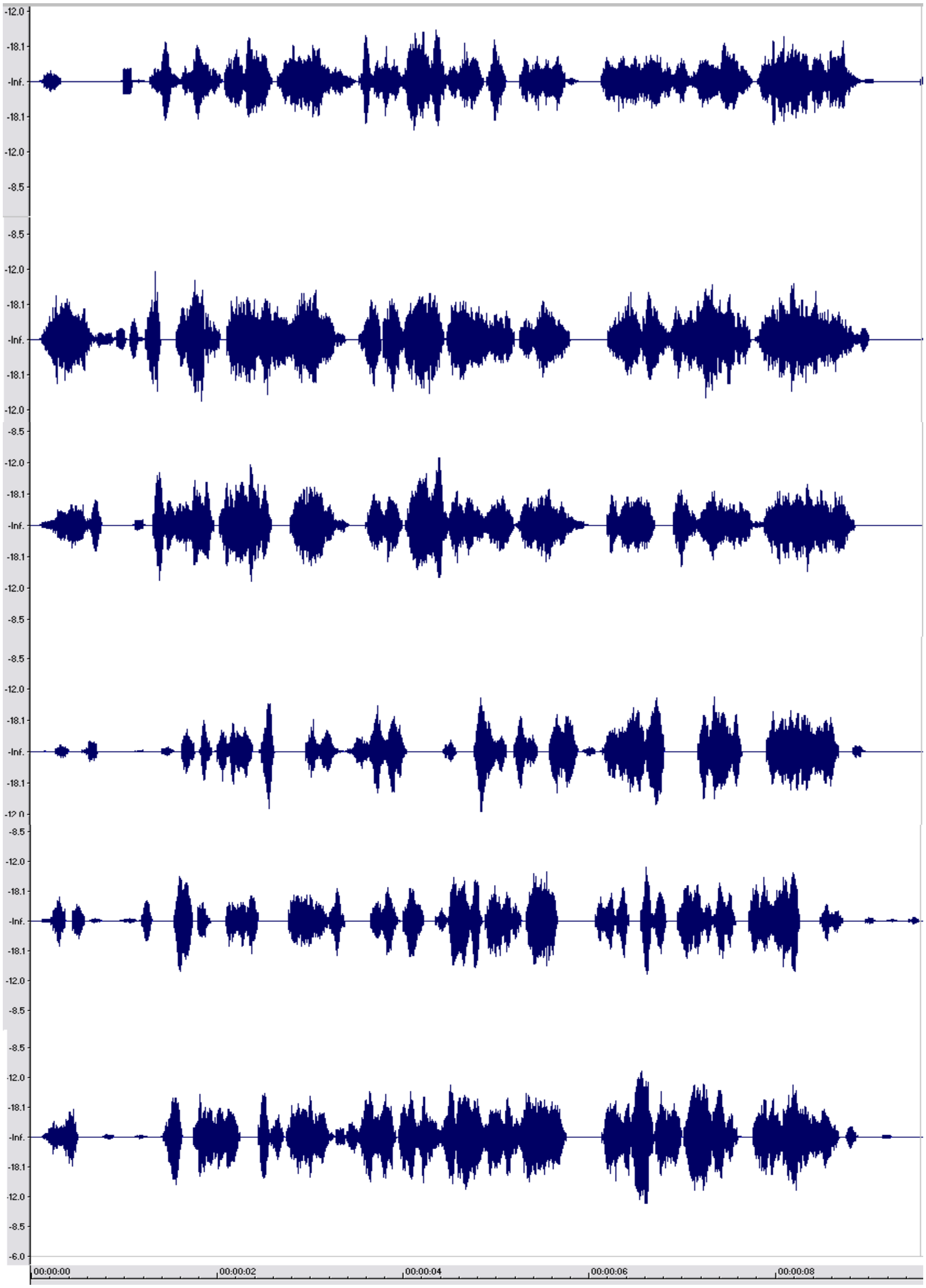

Figure 30: The six separated un-clustered mono tracks of the vocals

Figure 31 shows the separated vocals after clustering the six versions together. The clustering was performed by simple overlapping of the sources in the time domain.

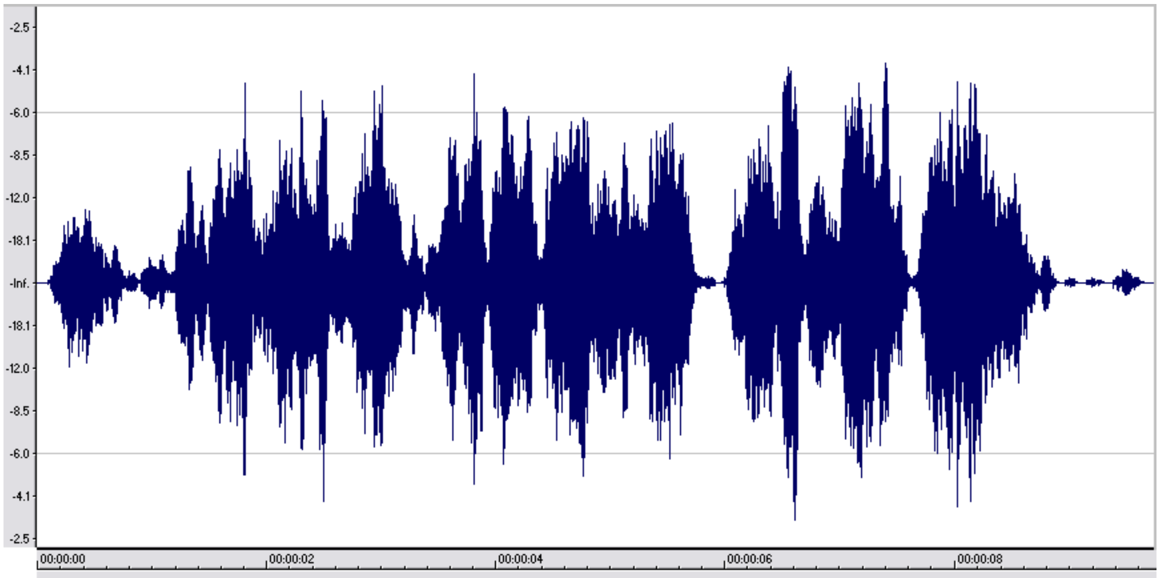

Figure 31: Clustered vocal mono track obtained from the six un-clustered versions

5.5 Experiment 5

Name of Music Track/Artist: SO3/MarC Vinyes

Duration of test file: 15 seconds

Genre: Rock

Type of sources: Lead Guitar, Bass Guitar, Drums

This track was a composition by MarC Vinyes also one of the authors of the “Demixing Commercial Music Productions” algorithm [7]. The track was more complex in nature in comparison to the tracks in the previous experiments. There were two acoustic guitars and one electric guitar along with drums. All the instruments were played continuously in the recording and due to the nature of the sources, the frequency overlap was also high. The algorithm separated one of the acoustic guitars along with the drums as a single source. This result was also obtained in [7]. Figure 32 shows the left and right channels of the original mix.

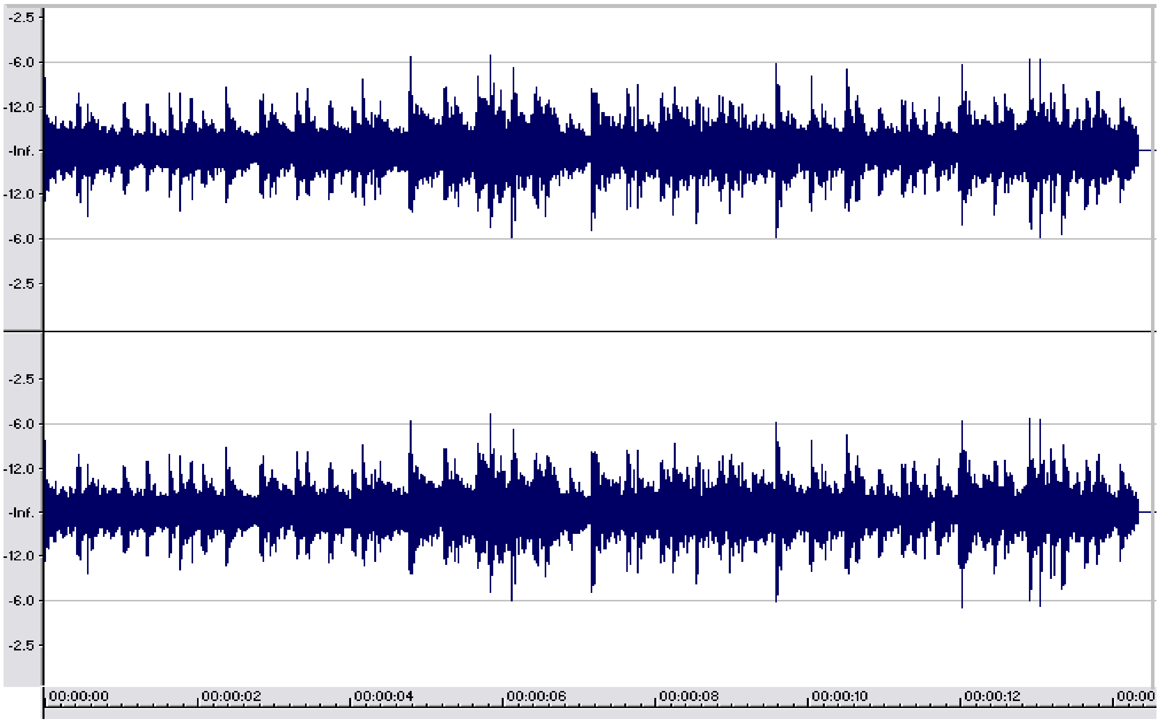

Figure 32: Left and right channels of the original mix in experiment 5

Figure 33 shows the separated mono track comprising the drums and the acoustic guitar together. The algorithm detected the significant peaks successfully. This can be observed from the amplitude levels of the waveform.

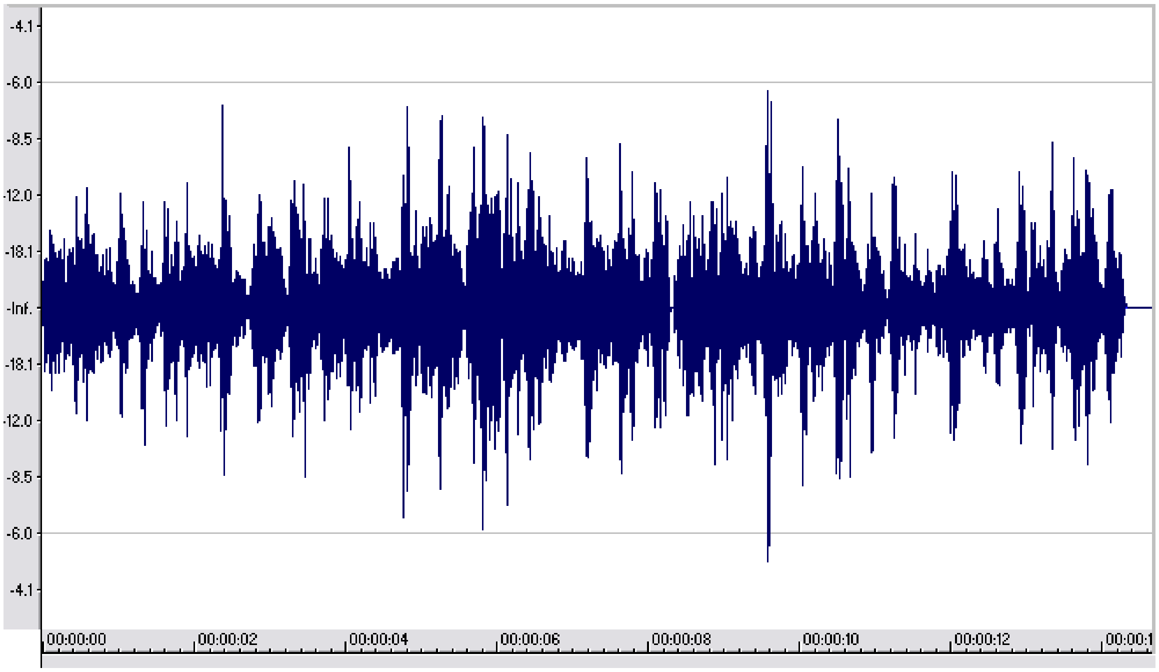

Figure 33: Separated mono track consisting of acoustic guitar and drums

5.6 Experiment 6

Name of Music Track/Artist: Wind Cries Mary/Jimi Hendrix

Duration of test file: 11 seconds

Genre: Rock

Type of sources: Vocals, Lead Guitar, Bass Guitar, Drums

The recording was a sample of the same track used in Experiment 2. It comprised of the same instruments with the addition of vocals. In this case, the vocals were more prominent than the other sources as the amplitude levels were higher. The overlap between all the sources was also high as all the sources were continuously performed throughout the recording. The algorithm generated nine un-clustered sources of which three were the vocals. Figure 34 shows the left and right channels of the original mix.

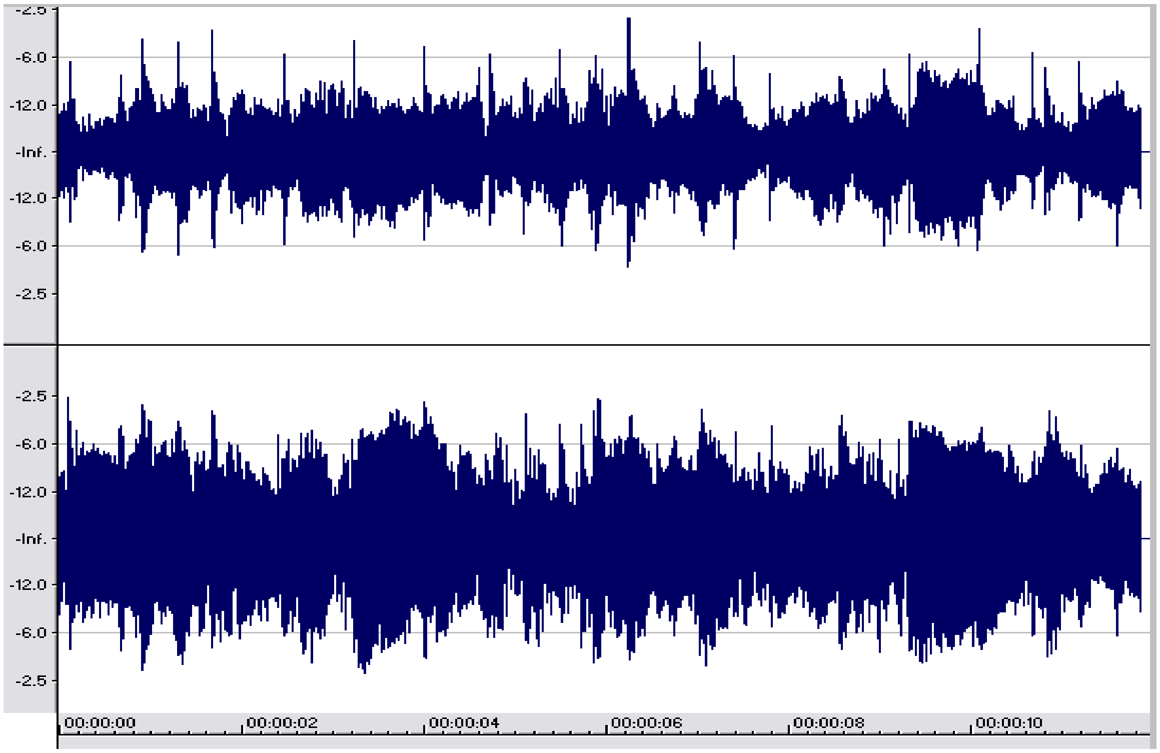

Figure 34: Left and right channel waveforms of the original mix in experiment 6

Figure 35 shows the three un-clustered separated mono tracks of the vocals. One of the un-clustered tracks had the lowest amplitude level with least temporal information. All three tracks had low noise and artifacts due to other interfering sources. When compared to the amplitude levels in the original mix, the algorithm recovered the same level of information in the second and third un-clustered sources.

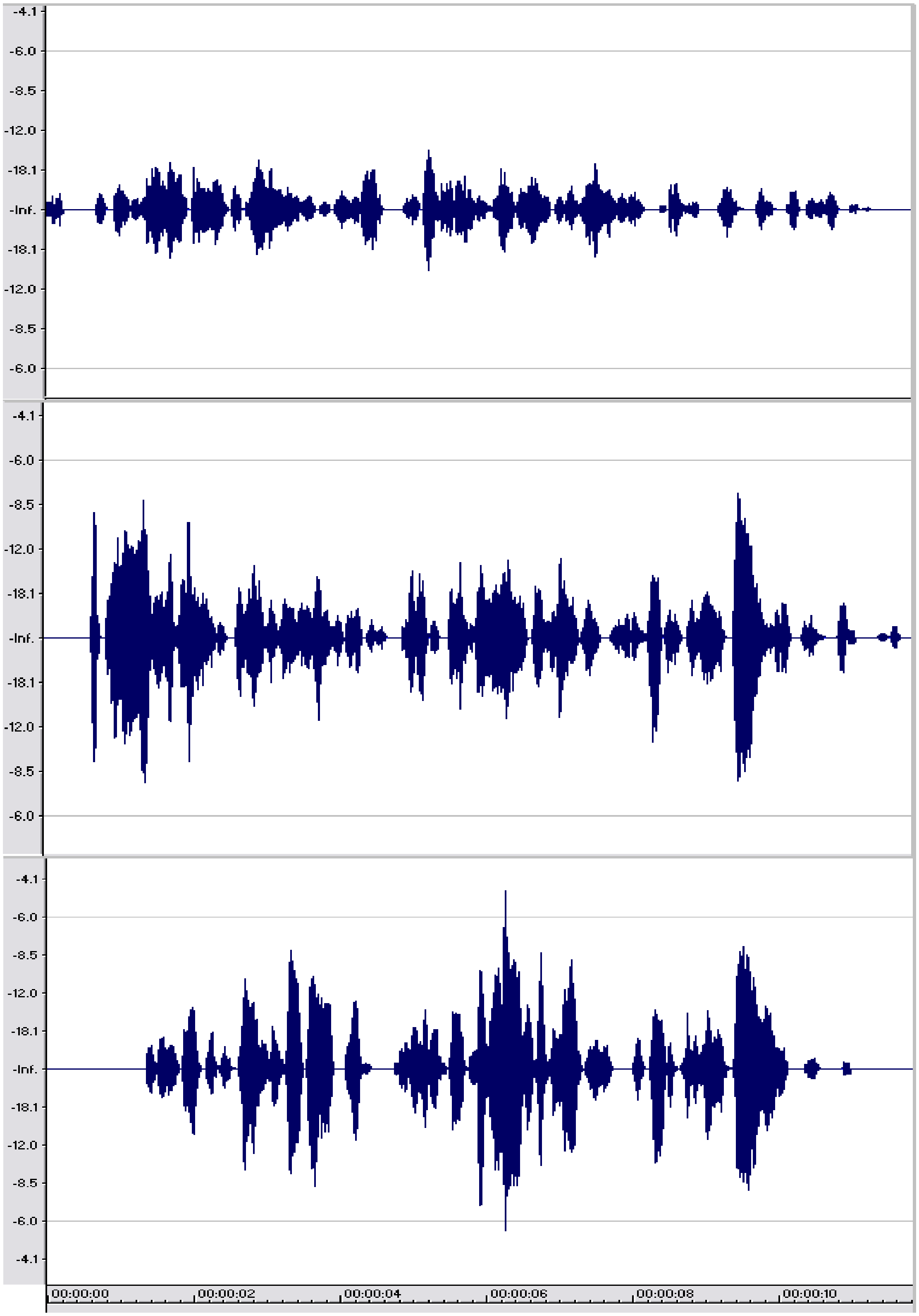

Figure 35: The three un-clustered separated mono tracks of the vocals

Figure 36 shows the separated mono track of the vocals obtained after clustering the three sources by overlapping in the time domain. After clustering it was observed that the effect of noise and musical artifacts was slightly higher than before clustering but had a minimal effect on the separated source.

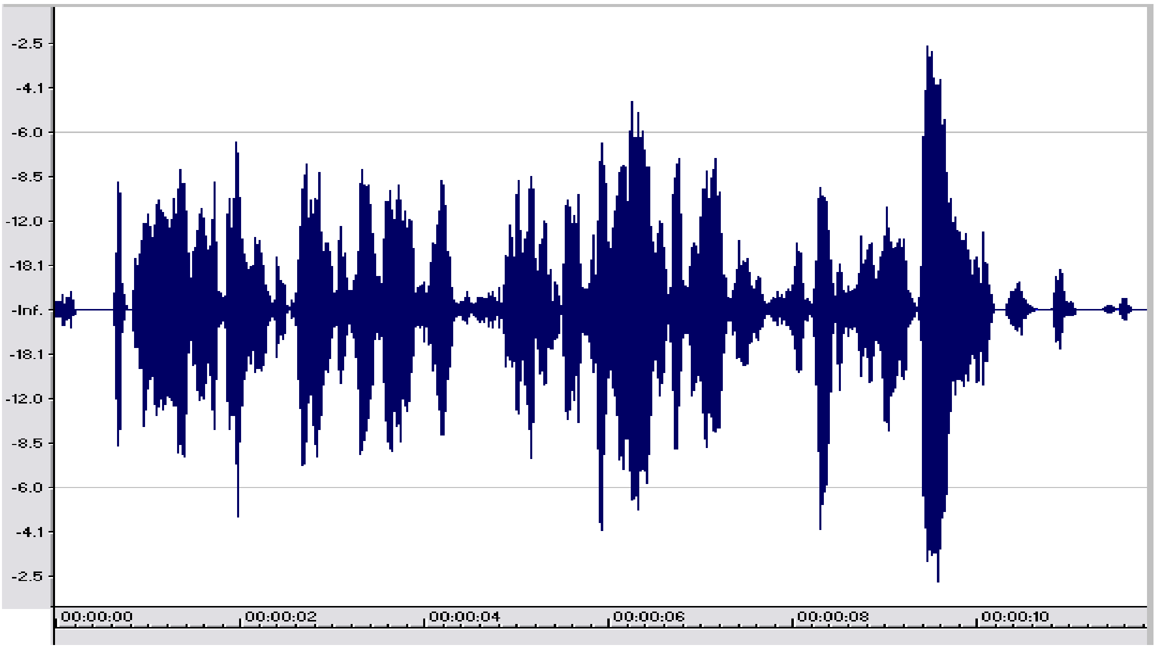

Figure 36: Separated vocals obtained after clustering the three mono tracks

5.7 Performance Analysis

The six test files involved similar musical instruments with the number of instruments always less than six. The algorithm performed variably in separating sources in each experiment. The degree of overlap between the sources in the time domain was the least in the first experiment. All the instruments played intermittently and were panned differently in the stereo track thereby allowing for better pan tracking and detection of the sources. Also, the number of spurious signals being generated as sources by the algorithm was the least in this experiment.

As the overlap of sources increased in the second and third experiment, the number of spurious signals being generated became higher and the number of sources being separated became less. The more interesting phenomena though were the un-clustered versions of sources that were being generated. In these experiments, there were always un-clustered versions of whichever sources that were separated and no source was separated as a whole unlike the first experiment. This phenomenon can be explained as "over-estimation" which is common in source separation. On listening to the un-clustered versions of the same source, one of them sounded close to the original source whereas the other would sound "wet" or "cloudy." This could be explained in the following way. The effects of stereo processing such as reverberation that may have been applied during the mixing made the algorithm overestimate the number of sources. In the case of the two un-clustered versions of drums or guitar separated, the algorithm treated the original sound and the reverberation (wet or cloudy) as two different sources. This could also be the reason for the generation of more spurious signals in the process.

In the fourth and fifth experiments, the degree of temporal overlap was the highest. Also, the number of instruments in these experiments was greater than in previous cases. In experiment 4, there were twenty three separated sources, of which thirteen consisted of signal fragments along with noise and artifacts. The remaining ten separated sources comprised predominantly of vocals of which six contained meaningful information. It was also observed that due to high degree of overlap, the number of separated un-clustered sources was the highest in this case though no other instrument could be recovered due to the increased number of spurious signals.

The interesting result was that the algorithm separated both the lead vocals and the back up singers as one single source in all the un-clustered versions in experiment 4. Another important aspect to note was that the recording was of a different era – 1960s. The recording technique used in this track is unknown and could have been responsible for the algorithm's success in vocal extraction from a complex track. The same applies to the separation of the vocals alone in experiment 6 even though the vocals and lead guitar were performed by the same person. In both these experiments, vocals were the only sources successfully separated. Due to lower degree of overlap, the number of extracted un-clustered sources was nine in experiment 6, three of which were predominantly the vocals and the rest being spurious.

Over-estimation or estimating more number of sources can happen as a source is split across several other estimated sources. There are two typical cases. In spectral split, a range of frequencies appear in one of the separated sources, and another set of frequencies appear in the other separated sources. In temporal split, a source moves in the stereo field, resulting in the instrument sounds being split among two or more separated sources. With respect to this thesis, the term separated sources in the above two cases directly corresponds to the un-clustered sources that were obtained in the experiments. Studying the un-clustered sources waveforms, it can be seen that the sources were temporally split. Peaks that were missing in one un-clustered waveform appeared to be enhanced in the other and vice versa. The spurious signals generated could be due to spectral splitting of the sources, as some of the artifacts resembled the sound from the sources that were separated.

Another case of over-estimation exists where the snare, hi-hat and kick drum could be detected as three different sources or the same with each attack of a piano note being separated as one source and the sustained part of the note as another source. This is more common in one sensor separation problems, and the algorithm was able to separate the drums in three of the experiments and the piano in the first experiment as single sources.

The algorithm’s inability to extract all sources successfully could be due to the limitations of the main assumption itself- that all the instruments were recorded as mono tracks and then mixed to stereo using linear combinations. Many contemporary recordings also carry out binaural recording techniques along with natural and artificial reverb to add to the “liveliness” of the playing instrument. This factor proves to be crucial in the success of the algorithm.

5.8 Experiment with Computer Music

The algorithm was tested on a recording created by MIDI software to study its operation on different types of signals. The recording was made in the following manner. Two sources were selected comprising of piano and drums panned left and right respectively. The recording started with the drums playing in solo for 4.5 seconds followed by the piano in solo for 5 seconds after which the drums was re-introduced to play with the piano till the end of the 13 second recording. Figure 37 shows the left and right channels of the MIDI track.

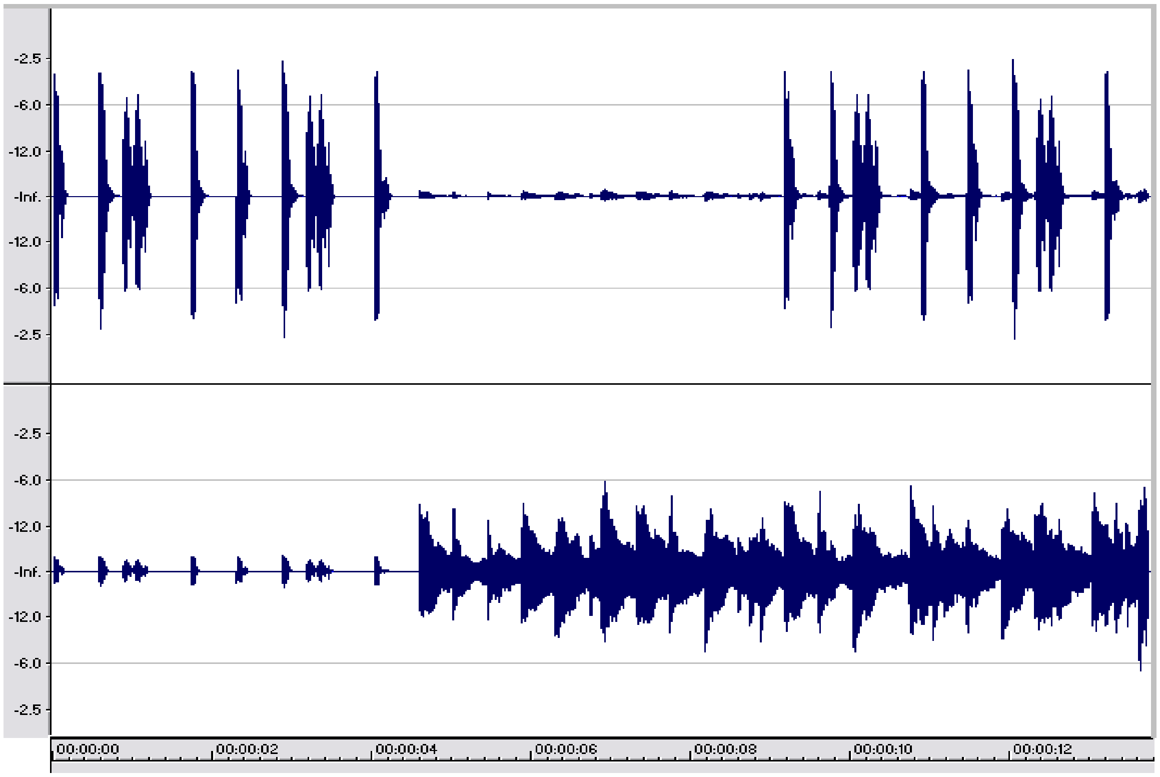

Figure 37: Left and right channels of the original MIDI track

From figure 37 it can be observed that the left channel consisted predominantly of the drums and right channel consisted of the piano. The algorithm separated both instruments with no noise or artifacts and there were no un-clustered sources generated. Figure 38 shows the separated drums and piano mono tracks respectively.

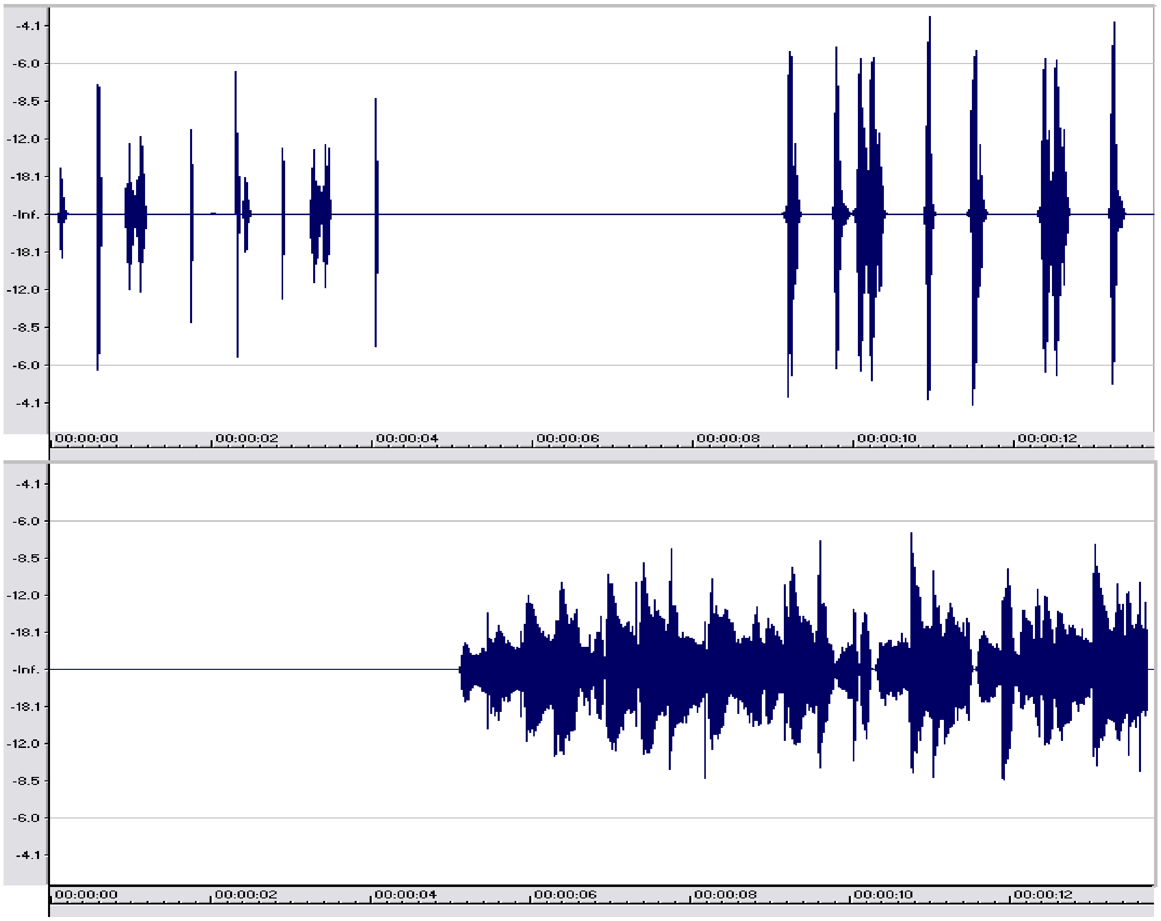

Figure 38: The separated mono tracks - drums (above) and piano (below)

From figure 38 it can be observed that there was minor loss of information in the first 4.5 seconds of the separated drums and the first 0.3 seconds of the piano was not detected. The algorithm also did not detect a snare hit that occurred at the twelfth second of the recording. In both the separated tracks the amplitude levels were well reproduced when compared to the original mix.

When a synthesizer (third instrument) was center panned in the same recording in the first 4.5 seconds, the algorithm generated seven sources of which three were spurious. One of the sources was the separated drums that was comparable to the drums obtained in the previous case of two instruments. The remaining three sources consisted of un-clustered versions of the piano and synthesizer. On clustering the three, the piano was well separated whereas there was loss of information to reproduce the synthesizer.

The algorithm performed more efficiently in case of two instruments when there was clear distinction in the pan values applied to the sources. In case of the third instrument when a center pan was applied, the source was split between both channels making it difficult for the algorithm to separate that particular source. In the absence and presence of the third instrument, the algorithm was successful in separating the sources that were panned left and right though more spurious signals were generated in the latter case.

5.9 Experiment with Sine Waves

The algorithm was tested on sine waves where three sine waves of different frequencies were mixed to stereo as follows: 300 Hz tone panned left, 440 Hz tone panned center and 600 Hz tone panned right. The algorithm extracted all three tones without generating any spurious signals or un-clustered sources. The separated tones consisted of certain constant clipping that was observed uniformly but there were no occurrences of any kind of musical noise or artifacts.

This experiment with sine tones as opposed to musical recordings helped to further understand the functioning of the algorithm. It worked efficiently for sine tones due to clear distinction between three constant frequencies panned differently from each other. This implied that the pan estimation and source separation process were performed perfectly in such an ideal case scenario. However the clipping phenomenon could have been due to improper phase reconstruction which was observed even in the previous experiments with musical recordings. Since the overlap of frequencies was far more complex in music, musical noise and artifacts were also generated in the separated sources but not in case of sine tones.

6

Evaluation

After performance analysis, a measure of the algorithm’s performance is necessary to characterize its success. For blind audio source separation, there is no definite method to do so. The evaluation can be based on the application of the algorithm but in this thesis more emphasis is laid on the possibility of separating sources in an unassisted fashion. Going by the BASS typology, this algorithm can be categorized as being an Audio Quality Oriented (AQO) task with either application- One Versus All problem or Audio Scene Modification.

As mentioned in [29], even though listening tests can provide statistically significant results with less than ten non-expert subjects, they have not been conducted in most previous research due to certain misconceptions and time constraints.

To achieve a substantial evaluation and analysis of the unassisted algorithm, two different types of listening tests were conducted. The tests were aimed at bringing out different characteristics of the separation algorithm’s performance:

1) MUSHRA listening test

The MUSHRA format of having multiple signals with a hidden reference and an anchor was followed. The MUSHRA listening test only applies to cases where the hidden reference is of superior quality. This listening test had to be structured to select a suitable reference sound. Even though the algorithm was successful in detecting and separating sources, they were of inferior quality when compared to the source in the original stereo mix. This was evident due to the presence of noticeable artifacts and clips in the sound in the separated sources. In cases where un-clustered sources were separated (as in Experiment 2 and 3), the artifacts and clips were more prominent due to the occurrence of temporal split. For this reason, the separated sources from the first three experiments could not be used as the hidden reference sound as per the MUSHRA requirements but the need for evaluating the algorithm had to be satisfied. Taking all these factors into consideration, the test was structured on a purely subjective basis with reasonable test duration [29], where the separated sources from the ADRess research were selected as the hidden reference sound for the corresponding separated sounds obtained from the first three experiments. Since the experiments were conducted taking the test files from [6] and [7] and due to the availability of the separated sources for the same experiments conducted in this thesis research, it provided a means of comparison of the performance of the algorithm. The differences in average scores between the hidden reference sounds and the separated tracks would provide more information than the individual scores themselves. This was the first objective of the listening test.

In case of the second and third experiments where un-clustered sources were obtained, the listening test compared the un-clustered sources with the clustered ones. The un-clustered results were clustered together by simple overlapping of the sources in the time domain. Through informal listening tests it was observed that the separation and audio quality of the clustered sources were improved than their respective un-clustered versions. Temporal information was added thereby giving a much enhanced version of the source. As a result, the comparison of the clustered and un-clustered sources became the second objective of the listening test.

2) Separation Quality Assessment (SQA) Listening Test

In this test, the “separation and audio quality” of the separated sources was assessed on an absolute scale. The objective of this test was to analyze the performance of the algorithm without any comparison with other algorithms that performed source separation. The assessment terms and conditions had to be defined to achieve the objective of the test. The main difficulty was structuring the sequence of the test and the questions that needed to be asked to the subjects since it was to be based on an absolute scale for assessing source separation for which there is currently no clear methodology.

6.1 MUSHRA Listening Test

The test comprised 26 subjects between the age group of 18 – 38 years of which 13 subjects were musically trained and current performers. Of the remaining 13 subjects, 8 had not received any previous musical training and the 5 that did receive musical training quit performing at least 3 years ago. The test was conducted by the recommendation ITU-R BS.1534-1 and the guidelines of the MUSHRA testing format. The following has to be noted. The hidden reference sound was selected to be the separated source files available from the ADRess research as this was assumed to be of best quality. The anchor sound was the hidden reference low pass filtered at 3.5 kHz.

6.1.1 Test Description

The complete test consisted of four sections. Each section consisted of a hidden reference, anchor, separated sources from this research and a degraded quality (noise like) file of the reference. The perceived difference between the hidden reference and the anchor was negligible. The four sections were: the separated saxophone track (from experiment 1), the piano track (from experiment 1), the lead guitar track (from experiment 2) and lead guitar track (from experiment 3). In the last two sections, the un-clustered separated guitar tracks were also included. The structure of the test and the objectives were clearly explained to subjects along with the details about the research. This was done to minimize the number of any possible discrepancies in the grading procedure.

6.1.2 Results

The MUSHRA grading scale is as follows: Excellent (81-100), Good (61-80), Fair (41-60), Poor (21-40) and Bad (0-20).

Section 1: *Saxophone*

In 24 out of 26 cases, the hidden reference or the anchor was given the highest score. The average score of the separated sax track was 57.4%. On the MUSHRA grading scale, this falls in the upper limit of the "Fair" category.

Section 2: *Piano*

In 26 out of 26 cases, the hidden reference or the anchor was given the highest score. The average score of the separated piano track was 45%. On the MUSHRA grading scale, this falls in the lower limit of the "Fair" category.

Section 3: *Lead guitar (from experiment 2)*

In 26 out of 26 cases, the hidden reference or the anchor was given the highest score. The average score of the clustered guitar track was 48.76% (Fair) and the un-clustered tracks were averagely scored as 31% (Poor) and 34.72% (Poor) respectively.

Section 4: *Lead guitar (from experiment 3)*

In 23 out of 26 cases, the hidden reference or the anchor was given the highest score. The average score of the clustered guitar track was 54.2% (Fair) and the un-clustered tracks were averagely scored as 37.4% (Poor) and 31.1% (Poor) respectively.

Table 5 gives an overview of the results of the MUSHRA test in the four sections.

Table 5: MUSHRA Listening Test Results

Section # (No. of cases) / Average score	*Excellent 81-100*	*Good 61-80*	*Fair 41-60*	*Poor 21-40*	*Bad 0-20*
Section 1 (24/26): Saxophone			57.4%		
Section 2 (26/26): Piano			45%		
Section 3 (26/26): Un-clustered Guitar 1 *Un-clustered Guitar 2* *Clustered Guitar*			 48.76%	31% 34.72%	
Section 4 (23/26): Un-clustered Guitar 1 *Un-clustered Guitar 2* *Clustered Guitar*			 54.2%	37.4% 31.1%	

6.2 Separation Quality Assessment Listening Test

In this test, 22 different subjects of the age group 18 – 30 years participated out of which 11 had received no musical training and the remaining were musically trained people currently practicing. The test consisted of two sections:

1) The original track, separated saxophone, separated piano and separated drums from experiment 1.

2) The original track, clustered separated lead guitar and clustered separated drums from experiment 3.

The subjects were not given any background information about the research or the purpose of the test to avoid any bias towards guessing the separated tracks once they understood the pattern of the test.

The first original track was played, and the subjects were asked to identify the different instruments that they perceived. This was done to familiarize the subjects with the sounds that they would have to identify in the next part. After playing the original track, the separated tracks were played one by one and after each track, the subjects were asked to identify the instrument(s) that they perceived and to name the most “prominent” sounding instrument. This was done on the premise that if the subjects perceived one instrument playing then they would name that instrument. For tracks where the separation was not clear and there were leakage of sounds from different instruments, the subjects would not perceive that as a single instrument playing. And if they could not identify the most prominent instrument in such tracks, it would imply that the algorithm was unable to separate that instrument completely.

Once all the tracks were played and identified by the subjects, in the second part of the test the subjects were given the background information about this research. They were then asked to grade the separated tracks based on the following factors:

1) How well they perceived the separated tracks to be a reproduction of what they heard in the original track.

2) The level of sounds they could hear from other instruments in the separated tracks.

3) The noise and artifacts introduced in the separated tracks.

The same grading scale was used as in the MUSHRA format where the subjects were asked to rate the separated tracks absolutely as "Excellent," "Good," "Fair," "Poor" or "Bad" based on the above factors. Where the separated tracks were not perceived to be of a single instrument, they were rated as "Unidentified." Table 6 shows the number of people that rated the separated tracks in each grade category.

Table 6: Separation Quality Assessment (SQA) Listening Test Results

Track/Rating	*Excellent* *81-100*	*Good* *61-80*	*Fair* *41-60*	*Poor* *21-40*	*Bad* *0-20*	*Unidentified*
Saxophone	12	10				
Piano	4	12	6			
Drums 1			11	5	3	3
Lead Guitar	6	10	6			
Drums 2	1	1	12	5	1	2

6.3 Analysis of Results

From the MUSHRA test results it can be inferred that the overall performance of the algorithm was "fair." The saxophone and piano separated track were rated the highest of all the tracks which corroborates the hypothesis that the algorithm performed better when there was clear distinction in the panning and when there was minimum overlap in the simultaneous performance of the instruments.

The next important inference was that the clustered guitar tracks were rated considerably higher than their un-clustered versions. This implies that providing more temporal information to the source does improve its perception and quality.

The SQA tests show that the algorithm was highly successful in its functionality of automatic separation as it was able to perceive the musical sources meaningfully and separate them similar to the human perceptual system. This success can be substantiated with the point that in 105 out of 110 cases, the subjects were able to correctly identify the prominent instrument in the separated track and perceive it as a single instrument separated from the original mix. The other inference from these tests was that the algorithm performed better in separating tonal instruments than non-tonal. The algorithm performed least efficiently when it came to separating drums. In 5 cases, the drums from experiment 1 and 3 were not identified as the most prominent instrument and were not perceived as a single instrument track. This was due to the presence of the piano and a lot of noise and artifacts. Again as observed in the previous test, the piano and the clustered lead guitar track were rated highly with the saxophone being rated the best.

7

Conclusions

There has been significant progress in the area of audio source separation with current research contributing to a better understanding of audio signals and mixtures. Despite this, source separation remains to be a very challenging problem in digital signal processing and there still remains much to be discovered and achieved. Today there exist a number of different approaches that attempt to tackle the problem of separating sources from a mixture with partial success. Though the objectives and methodology vary distinctively from one approach to another, the fundamental complexity of source separation remains the same – programming a machine to perform the functions that are very natural to the human perceptual system. Though these do provide sophisticated means to analyze a variety of unobvious problems that arise in source separation, they also prove to be the main limitation and the decisive factor in the success of any algorithm.

There are various factors that influence the shaping of any proposed method in the digital framework. These could be tuned towards specific applications or to understand the relationship between sources and sensors in different recording environments or a combination of both. Though most of the audio mixtures are created under natural conditions in coherence with the aim of a research, the implementation and final analysis of the algorithm is inevitably done using a digital machine. This has led to the studying of source separation not purely as an independent audio problem with mathematical explanations but also as a real world problem that can be solved using advanced technology.

Over the past decades, there have been advances in source separation thus giving rise to a tentative typology. Certain areas as such source separation for commercial stereo music has been less studied and has been focused on only in the last few years. The attempts so far have had reasonable success but there is no standard way to evaluate this success which remains to be decided in the years to come. The main approach used in most of these researches has been common – separation of sources by pan estimation and discrimination in the stereo field. This further narrows down the "commercial stereo music" category as newer methods are yet to be discovered and experimented.

Due to the underlying challenges in creating source separation algorithms, previous methods have found a human interactive process more effective. Unassisted approaches do face more impediments when compared to human assisted methods, but based on the results from this thesis, an acceptable separation quality can still be achieved by incorporating certain ideas within the current framework.

7.1 Final Evaluation - Importance of Listening Test Results

Due to the novel nature of the unassisted source separation approach, the listening tests that were conducted in this thesis proved to be the best analysis of the algorithm's performance and evaluation. Even though the behavioral patterns of the algorithm were exhibited during the experiments with test samples, they were clearly proved and substantiated only after the listening test results. Listening tests offer great flexibility that can be utilized for in-depth analysis of first time approaches. The methodology of the tests was structured purely to study the unassisted approach which led to strong inferences. The MUSHRA listening test helped to grade the efficiency of the algorithm on a more comparative basis thereby paving way for ideas to improve the algorithm's functioning. On the contrary, the SQA listening test results empirically established the overall success of the unassisted approach.

Certain patterns of the algorithm that were observed during the performance analysis section were:

- As the temporal overlap between sources increased the number and quality of separated sources decreased.
- More temporal overlap of sources led to more un-clustered sources.
- Clustering of sources improved separation quality.
- Non-percussive instruments were well separated than percussive instruments.

The listening test results proved all these observations to be true. The music track in experiment 1 had the least temporal overlap. The saxophone separated from that track was graded the highest among all separated sources in the MUSHRA listening test and also received the best separation quality score in the SQA listening test. In all cases the un-clustered sources not only received the lowest scores when compared to their respective clustered counterparts but also separated sources obtained from other experiments. Another important result was that the clustered guitar separated in experiment 3 (which had relatively lower temporal overlap than experiment 2) was graded higher than the clustered guitar in experiment 2. From both listening test results it was clear that percussive instruments were the most difficult sources to separate as they received the lowest scores.

The SQA listening test results were very significant because the test was conducted on a more perceptual basis without any type of comparison unlike in the MUSHRA format. The condition that whether or not the test subjects perceived the separated sources as one source helped to evaluate the algorithm's success. The performance analysis section alone couldn't have confirmed this aspect. With the 95.5% positive results obtained from the SQA test, the unassisted approach can be considered to highly successful. Thus the listening test results proved to be highly crucial in the overall assessment of the unassisted approach.

7.2 Future work

The first step that can be taken towards improving the separation quality of the separated sources would be to separate them more effectively before applying any post-processing techniques. The present algorithm performs overestimation and through listening tests an improvement was noted after clustering the over estimated sources together. The possible method would be to make use of certain clustering algorithms once the over estimation stage is complete. There are a number of these which can be employed, but this is yet to be experimented in source separation especially for commercial stereo music. A careful analysis is to be done before selecting the appropriate clustering method to suit the requirements of the current framework. Clustering would mainly serve to estimate and separate the sources more completely lessening the effect of temporal and spectral splitting while also reducing the number of spurious signals in the process. The quickest solution of clustering would be to cluster estimated sources together based on overlapping or close ranged pan windows.

Another feasible method is to first group the separated sources together on the basis of certain similarity measures [39]. It was observed that the sources currently suffer from temporal split which implies that they would share similar timbral information. This information could be later used to perform clustering. Clustering has been done for the manipulation of drum loops using a k-means method to cluster extracted drum components together [52]. This could serve to improve the quality of separating drums in the unassisted approach. Though most of the clustering approaches have been in experimented in researches that require prior information of the sources to be separated, the results of such methods on unassisted approaches could provide some undiscovered information and possibilities of newer studies.

To ensure more success of clustering approaches, the algorithm can first be “trained” on a fixed database of music signals of the same genre. This would help in identifying the characteristics and reactions of the algorithm more specifically as the target sources would be of a finite subset. The genres used in this research were mainly from Jazz and Rock. On experimenting with more samples, there could be possibilities of occurrences to substantiate that the algorithm’s success is also influenced by the type of genre of the sources to be separated. Thus starting with a fixed database of signals would pave the way to confirm this hypothesis.

A music signal can be divided into a number of segments based on the instruments playing in that segment. Each segment can be labeled using timbral descriptors for example, “Drums+Guitar+Piano” or “Saxophone+Bass+Drums” and so on. Based on [40] a classifier can be built to perform classification of the instruments. This would basically be building a musical instrument recognition system. Once the algorithm identifies the different sources correctly, this information along with the pan estimation could lead to more effective separation of sources. Improvement of the pan estimation procedure is possible too by applying advanced tracking methods such as Kalman filters and Adaptive Weiner filtering in the pan window tracking step. These methods have found to be very useful in advanced tracking applications such as RADAR and could offer more accuracy in estimating the pan window ranges. Once the characteristics of one music genre have completely been studied, another database of music signals of a different genre can be introduced to the algorithm. Rather than taking a random sample of available music samples to separate, a systematic procedure could provide better results.

In order to minimize the limitations of the assumptions on the algorithm, an important step that can be taken would be to obtain as much information about the manner in which the recording was done. This is useful especially in cases where old recordings are used that employed different recording techniques than modern recordings. In this thesis, this information would help to better explain the success of the algorithm in separation of vocals. In essence, factors such as reverb and room dimensions can be neglected but the needed information would be whether the recordings were made in the fashion where instruments were recorded as mono source tracks and later mixed to stereo. Once this aspect of the music signals database is also defined, it would help to easily identify and focus on any measures that would be needed to further improve the algorithm while satisfying the main assumptions.

Another significant area for improvement in the algorithm’s working is the detailed analysis of the phase reconstruction of the separated sources. Although the algorithm is ideally assumed to recover the original source magnitude and phase information without any type of processing, it is highly unlikely due to the complexity of musical recordings. As several sources occur at any moment in time, there is a high possibility that along with the pan and source magnitude information, the phase at each frequency is also an estimate and not the original phase of the source. Resolving this problem to retain accurate phase information could lead to considerable decrease in the clippings and artifacts that occurred in the separated sources obtained from the experiments conducted in this thesis. As a result this would lead to a much superior separation quality. There are no clear ways to tackle this challenging problem but it certainly opens up new perspectives to be explored in source separation research.

Even after breakthroughs in the field, there is not any solid means of evaluating an algorithm’s ability to separate sources. There have been proposals of objective measurement techniques but they have proved to be insufficient leaving many criteria unaccounted for. For application specific tasks like remixing or removal of vocals for karaoke, listening tests can provide significant results. But when it comes to grading the separation quality alone, there exists no effective method. In addition to new approaches in source separation, attempts should also be made in forming standardized tools that would serve in thorough evaluation of algorithms directed primarily towards separating sources. If listening tests prove to be effective for this purpose too, then efforts should be channeled towards devising robust and standardized testing procedures to obtain meaningful results. This would not only help in individual performance analysis but would also help categorize and compare algorithms more substantially based on their success. An instance of such standardization is needed can be substantiated in cases that occurred within this research thesis. In the tests that were conducted, through informal listening it was clearly observed that some of the factors such as presence of other instruments in a separated source and presence of musical noise and artifacts were higher in the ADRess separated sources than in the sources obtained by the unassisted algorithm. The separation quality can be assumed to better in the ADRess research when focusing on the reproduction of sound of the target instrument alone, but if the other factors are also taken into account in the success of an algorithm, there is no means to decide the "superior" separation quality. For compression algorithms, the questions can be framed much more effectively in order to determine the factors on which subjects need to choose the "better" algorithm. But when it comes to pure separation quality evaluation, there are no certain questions that can be asked to subjects which would guarantee an impeccable evaluation. A possible way to start would be to compare the mono source tracks against the separated source tracks, as in such a case the reference signal could be clearly defined as the ideal case [7]. Thus it would be necessary to use test samples which consisted of the original stereo mix to be separated and the original mono source tracks recorded under the conditions stated in the assumptions for more transparent evaluation purposes.

So far the algorithm was discussed solely from a source separation algorithm perspective. If the unassisted approach were to be used for audio scene modification or more popularly, remixing, then the separation quality of the sources need not ideally be the best. This is because, once separated sources go through synthesis procedures and are mixed back into the original, the effect of musical noise and artifacts is greatly reduced. Currently, the algorithm discards frequencies that do not fall within a detected pan window. This information instead can be used to construct a "remainder" signal in order to perform signal reconstruction thereby including more frequency information. The remainder signal could thus be used to alter the detected sources before reconstruction for remixing applications. For such applications, the synthesis procedures are of more value than the separation quality itself. An example of such an application would be to make a particular instrument “brighter” in a mix or to lessen the presence of another. In that case, a more robust evaluation of the unassisted approach is possible. The algorithm could be assessed on its remixing abilities and listening tests could be conducted to compare the original track versus the remixed track. Assuming that the algorithm does prove to be successful in this area, the next ambitious move would be the creation of source separation based music editing software, which would open up new doors to audio scene modification. This would also provide a novel means to music listening where instruments could be altered individually on a purely subjective basis as opposed to conventional EQ techniques where all instruments in a frequency range are altered to suit a genre which rules out the possibility of varying one particular instrument.

Source separation algorithms still have a long journey ahead to be implemented in real-time applications such as portable devices. Technology being the biggest strength also proves to be the biggest weakness because even with the best algorithms which are based on Bayesian techniques, when implemented using the best processors, take about an hour to process 10 seconds of audio in real-time, making the field of source separation the most challenging problem in audio processing.

7.2.1 Expansion to Multi-Channel Recordings

After improvements in the above mentioned performance characteristics of the unassisted approach, the next logical step would be to expand the algorithm to multi-channel recordings. Since it is strongly based on mathematical pan laws followed in mixing consoles today, it can also be used to perform separation on recordings with more than two channels. The main criteria for this to be possible would be thorough knowledge of the respective pan laws applied for such mixing processes. For example, in case of a 5.1 channel recording, there would be six observation mixtures or channels. Taking the pan laws followed in mixing consoles used to create 5.1 recordings and mathematically formulating the mixing process a similar methodology could be designed to separate sources panned in a surround field.

References

[1] Bregmann, A.S., "*Auditory Scene Analysis: The Perceptual Organization of Sound,*" MIT Press, Cambridge, Massachusetts, 1990.

[2] Vincent, E., Févotte, C., Gribonval, R., "*A Tentative Typology of Audio Source Separation Tasks,*" ICA 2003, Nara, Japan, April 2003.

[3] Vincent, E., "*Blind Audio Source Separation, A review of state-of-the-art techniques,*" Southampton Seminar, September 2005.

[4] Vincent, E., Jafari, M.G., Abdallah, S.A., Plumbley, M.D., Davies, M.E., "*Blind Audio Source Separation,*" BASS Tutorial, Centre for Digital Music, Queen Mary, University of London, 2005.

[5] Haykin, S.,"*Unsupervised Adaptive Filtering, Volume 1, Blind Source Separation,*" Wiley-Interscience, April 2000.

[6] Barry, D., Lawlor, B., Coyle, E., "*Sound Source Separation: Azimuth Discrimination and Resynthesis,*" Conference on Digital Audio Effects (DAFx '04), Naples, Italy, October 2004.

[7] Vinyes, M., Bonada, J., Loscos, A., "*Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking,*" 120th AES Convention, Paris, France, May, 2006.

[8] Barry, D., Lawlor, B., Coyle, E., "*Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm,*" 118th AES Convention, Barcelona, Spain, May 2005.

[9] Apte, S., "*Time-varying Azimuth Discrimination and Resynthesis: A new method for music purposing,*" Department of Computer Science, Brown University, September 2005.

[10] Cooney, R., Cahill, N., Lawlor, R., "*An Enhanced implementation of the ADRess Music Source Separation Algorithm,*" 121st AES Convention, San Francisco, CA, USA, October 2006.

[11] Vinyes, M., "*Auditory stream separation in commercial music productions: implementation of a real-time method based on pan location,*" Music Technology Group, Pompeu Fabra University, Barcelona, Spain, November 2005.

[12] Avendano, C., *"Frequency-Domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression, and Repanning Applications,"* in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 55-58, New Paltz, NY, October 2003.

[13] Cook, P., R., *"Music, Cognition and Computerized Sound, An Introduction to Psychoacoustics"* The MIT Press, March, 2001.

[14] Mehta, M., Johnson, J., Rocafort, J., *"Architectural Acoustics: Principle and Design,"* Prentice Hall, December, 1999

[15] Institut Universitaire de Recherche Clinique, Promenade 'round the Cochlea, [Online], http://www.iurc.montp.inserm.fr/cric/audition/english/ear/fear.htm

[16] Wikipedia Online Encyclopedia, [Online] http://wikipedia.org/

[17] Birchfield, S., T., Gangishetty, R., *"Acoustic Localization by Inter Aural Level Difference,"* IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, Pennsylvania, March 2005.

[18] Rahbar, K., Reilly, J.P., *"A Frequency Domain Method for Blind Source Separation of Convolutive Audio Mixtures,"* in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003.

[19] S. Rickard, R. Balan, J. Rosca, "*Real-Time Time-Frequency Blind Source Separation*," Proc. of ICA 2001, San Diego, CA, USA, December 2001.

[20] Avendano, C., *"Frequency Domain Techniques for Stereo to Multi-Channel Upmix,"* in Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, pp. 121-130, Espoo, Finland, 2002.

[21] Rayleigh, L., *"On Our Perception of Sound Direction,"* Philosophy Magazine, vol. 13, pp. 214-232, 1907.

[22] Casey, M.A., Westner, A., "*Separation of Mixed Audio Sources by Independent Subspace Analysis*," International Computer Music Conference (ICMC), 2000.

[23] Hyvarinen, A., Karhunen, J., Oja, E., "*Independent Component Analysis*," Wiley, New York, 2001.

[24] Parra, L. C., "*An Introduction to Independent Component Analysis and Blind Source Separation*," Princeton, NJ, 1999.

[25] Plumbley, M.D., Abdallah, S.A., Bello, J. P., Davies, M.E., Monti, G., Sandler, M.B., "*Automatic Music Transcription and Audio Source Separation*," Cybernetics and Systems: An International Journal, vol. 33, pp. 603-627, 2002.

[26] Jourjine, A., Rickard, S., Yilmaz, O., "*Blind Separation of disjoint orthogonal signals: Demixing N sources from two mixtures*," in proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 5, Istanbul, Turkey, pp. 2985-2988, June 5-9, 2000.

[27] Stone, J. V., *"Blind Source Separation using Temporal Predictability,"* Psychology Department, Sheffield University, Sheffield, England, 2001.

[28] FitzGerald, D., "*Automatic Drum Transcription and Source Separation*," Ph.D. Thesis, Conservatory of Music and Drama, Dublin Institute of Technology, 2004.

[29] Gillet, O., Gael, R., "*Extraction and Remixing of Drum tracks from Polyphonic Music Signals*," in proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2005.

[30] Gribonval, R., Benaroya, L.,, Vincent, E., Févotte, C., "*Proposals for Performance measurement in Source Separation*," in proc. of the 4th Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003), April 2003.

[31] Radocy, R. E. and Boyle, J. D., "*Psychological foundations of musical behavior," 4th ed.*, Charles C. Thomas, 2003

[32] Roads, C., "*The Computer Music Tutorial,"* MIT Press, Boston, 1996

[33] MATLAB software. The Mathworks, Inc. [Online], http://www.mathworks.com/

[34] Faller, C., Erne, M., "*Modifying Stereo Recordings Using Acoustic Information Obtained with Spot Recordings,"* 118th AES Convention, Barcelona, Spain, May 2005.

[35] Pulkki, V., "*Localization of amplitude panned sources I: Stereophonic panning*," AES Journal, vol. 49, no. 9, pp. 739-752, 2001.

[36] Schroeder, M.R., "*Natural sounding artificial reverberation*," AES Journal, vol. 10, no. 3, pp. 219-223, 1962.

[37] Haykin, S., "*Adaptive Filter Theory*," Prentice Hall Information and System Sciences Series, New Jersey, 2002.

[38] Koptenko, S., *LMAX.m – Find Local Maxima*, (MATLAB). Gugine International Ltd., 1997.

[39] Jafari, M.G., Vincent, E., Abdallah, S.A., Plumbley, M.D., Davies, M., "*Blind Source Separation of Convolutive Audio using an Adaptive Stereo Basis*," Technical Report C4DM-TR-06-04, Centre for Digital Music, Queen Mary University of London, 9 August 2006.

[40] Helen, M., Virtanen, T., "*Separation of Drums from Polyphonic Music using Non-negative Matrix Factorization and Support Vector Machine*," 13th European Signal Processing Conference, 2005.

[41] Essid, S., Richard, G., David, B., "*Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies*," IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, 2005.

[42] Pohlmann, K.C., "*Principles of Digital Audio*," *4th ed.*, McGraw-Hill, New York, 2000.

[43] Cichocki, A., Amari, S., *"Adaptive Blind Signal and Image Processing,"* John Wiley and Sons, West Sussex, 2002.

[44] Apostol, T., M., *"Calculus, Vol. 1: One-Variable Calculus with an Introduction to Linear Algebra," 2nd ed.*, John Wiley and Sons, USA, 1967.

[45] Apostol, T., M., "*"Calculus, Vol. 2: Multi-Variable Calculus with an Introduction to Linear Algebra," 2nd ed.*, John Wiley and Sons, USA, 1969.

[46] Spivak, M., *"Calculus," 3rd ed.*, Cambridge University Press, England, 1994.

[47] Sony Media Software, "Sound Forge," [Online] www.sonymediasoftware.com/products/**soundforge**family.asp

[48] Rowe, D., B., *"Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing,"* Chapman and Hall/CRC, USA, 2005.

[49] Oppenheim, A., V., Schafer, R., W., Buck, J., R., *"Discrete Signal Processing," 2nd ed.,* Prentice Hall, 1999.

[50] Kleczkowski, P., *"Selective Mixing of Sounds,"* in Proc. Signal Processing for Audio, 119th AES Convention, New York, October, 2005.

[51] National Instruments, *"Windowing: Optimizing FFTs Using Window Functions,"* [Online], http://zone.ni.com/devzone/cda/tut/p/id/4844

[52] Bello, J., P., Ravelli, E., Sandler, M., B., *"Drum Sound Analysis for the Manipulation of Rhythm in Drum Loops,"* IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006.