UNIVERSIDAD
POLITÉCNICA
DE VALENCIA

Application of Sound Source Separation Methods to Advanced Spatial Audio Systems

DOCTORAL THESIS

by
Máximo Cobos Serrano

Supervisor:
Dr. José Javier López Monfort

Valencia, Spain
June 2009

To Sandra

Abstract

This thesis is related to the field of *Sound Source Separation* (SSS). It addresses the development and evaluation of these techniques for their application in the resynthesis of high-realism sound scenes by means of *Wave Field Synthesis* (WFS). Because the vast majority of audio recordings are preserved in two-channel stereo format, special up-converters are required to use advanced spatial audio reproduction formats, such as WFS. This is due to the fact that WFS needs the original source signals to be available, in order to accurately synthesize the acoustic field inside an extended listening area. Thus, an object-based mixing is required.

Source separation problems in digital signal processing are those in which several signals have been mixed together and the objective is to find out what the original signals were. Therefore, SSS algorithms can be applied to existing two-channel mixtures to extract the different objects that compose the stereo scene. Unfortunately, most stereo mixtures are *underdetermined*, i.e., there are more sound sources than audio channels. This condition makes the SSS problem especially difficult and stronger assumptions have to be taken, often related to the *sparsity* of the sources under some signal transformation.

This thesis is focused on the application of SSS techniques to the spatial sound reproduction field. As a result, its contributions can be categorized within these two areas. First, two underdetermined SSS methods are proposed to deal efficiently with the separation of stereo sound mixtures. These techniques are based on a *multi-level thresholding* segmentation approach, which enables to perform a fast and unsupervised separation of sound sources in the time-frequency domain. Although both techniques rely on the same clustering type, the features considered by each of them are related to different localization cues that enable to perform separation of either instantaneous or real mixtures. Additionally, two post-processing techniques aimed at improving the isolation of the separated sources are proposed. The performance achieved by several SSS methods in the resynthesis of WFS sound scenes is afterwards evaluated by means of listening tests, paying special attention to the change observed in the perceived *spatial attributes.* Although the estimated sources are distorted versions of the original ones, the masking effects involved in their spatial remixing make artifacts less perceptible, which improves the overall assessed quality. Finally, some novel developments related to the application of time-frequency processing to source localization and enhanced sound reproduction are presented.

Keywords: Wave Field Synthesis, Sound Source Separation, Time Frequency Processing, Direction of Arrival, Spatial Audio Quality.

Resumen

Esta tesis se enmarca dentro del campo de la *Separación de Fuentes Sonoras* (SSS), donde se ha trabajado en el desarrollo y evaluación de estas técnicas para aplicarlas a la resíntesis de escenas sonoras de alto realismo utilizando *Síntesis de Campo de Ondas* (WFS). Dado que la gran mayoría de grabaciones sonoras se almacena en un formato estéreo de dos canales, es necesario disponer de sistemas especiales de conversión con el fin de utilizar sistemas avanzados de reproducción de sonido espacial, como por ejemplo WFS. Esto se debe al hecho de que WFS necesita las señales originales de las fuentes para sintetizar de forma precisa el campo acústico dentro de una amplia zona de escucha, requiriendo un proceso de mezcla basado en objetos.

Los problemas de separación de fuentes en el tratamiento digital de la señal son aquellos en los que, a partir de una mezcla de varias señales, se trata de encontrar las señales originales que dieron lugar a la mezcla. Por tanto, los algoritmos de SSS pueden aplicarse a mezclas estéreo ya existentes para extraer los distintos objetos que componen la escena sonora. Desafortunadamente, la mayoría de las mezclas estéreo son *subdeterminadas*, es decir, están compuestas por un número de fuentes mayor al número de canales. Esta condición hace que el problema de SSS sea especialmente difícil y lleva a asumir ciertas propiedades de las señales, normalmente relacionadas con la *escasez* (*sparsity*) de éstas bajo alguna transformación.

Esta tesis se centra en la aplicación de las técnicas SSS al campo de sonido espacial. Es por esto que sus contribuciones pueden ser clasificadas en estas dos áreas. En primer lugar, se proponen dos métodos de SSS subdeterminados que tratan de forma eficiente y no supervisada la separación de mezclas estéreo. Estas técnicas están basadas en la segmentación por *umbralización multinivel*, la cual permite separar fuentes sonoras de forma rápida en el dominio tiempo-frecuencia. Aunque ambas técnicas se basan en el mismo tipo de agrupación, las características consideradas por cada una de ellas están relacionadas con diferentes aspectos de localización que permiten separar las fuentes de mezclas instantáneas y reales. Adicionalmente, se proponen dos técnicas de post-procesado enfocadas a mejorar el aislamiento de las fuentes separadas. Las prestaciones obtenidas por varios métodos de SSS en la resíntesis de escenas sonoras con WFS son posteriormente evaluadas por medio de tests subjetivos, prestando especial atención al cambio observado en los atributos de percepción espacial. Aunque las fuentes estimadas son versiones distorsionadas de las originales, los efectos de enmascaramiento que se producen en la remezcla espacial provocan que los artefactos sean más dificilmente percibidos, mejorando la calidad subjetiva global. La Tesis finaliza con una serie de nuevos desarrollos relacionados con la aplicación del procesamiento tiempo-frecuencia a la localización de fuentes y a la mejora espacial de la reproducción de sonido.

Palabras Clave: Síntesis de Campo de Ondas, Separación de Fuentes Sonoras, Procesado Tiempo Frecuencia, Dirección de Llegada, Calidad de Sonido Espacial.

Resum

Aquesta tesi s'emmarca dins del camp de la *Separació de Fonts Sonores* (SSS), on s'ha treballat en el desenvolupament i l'avaluació d'aquestes tècniques per aplicar-les a la resíntesi d'escenes sonores d'alt realisme, mitjançant *Síntesi de Camp d'Ones* (WFS). Tot i que la gran majoria dels enregistraments sonors s'emmagatzemen en un format estèreo de dues canals, és necessari disposar de sistemes especials de conversió, amb la finalitat d'utilitzar sistemes avançats de reproducció de so espacial, com per exemple WFS. Això es deu al fet que WFS necessita els senyals originals de les fonts per sintetitzar, de forma precisa, el camp acústic dins d'una àmplia zona d'escolta, requerint un procés de barreja basat en objectes.

Els problemes de separació de fonts en el tractament digital del senyal són aquells en els que d'una barreja de diverses senyals, es tracta de trobar els senyals originals que van donar lloc a la mescla. Per tant, els algorismes de SSS poden aplicar-se'n a mescles estèreo ja existents per extraure els objectes sonors que composen l'escena sonora. Malauradament, la majoria de les mescles estèreo són *subdeterminades*, és a dir, estan compostes per un nombre de fonts major al nombre de canals. Aquesta condició fa que el problema de SSS siga especialment difícil i porta a assumir fortes propietats dels senyals, normalment relacionades amb l'*escassetat* (*sparsity*) d'aquestes sota alguna transformació.

Aquesta tesi se centra en l'aplicació de les tècniques de SSS al camp de so espacial. És per això que les seues contribucions poden ser classificades en aquestes dues àrees. En primer lloc, es proposen dos mètodes de SSS subdeterminats que tracten de manera eficient la separació de mescles estèreo. Aquestes tècniques estan basades en la segmentació i aplicació de *llindars multinivell*, la qual permet separar fonts sonores de forma ràpida en el domini temps-freqència. Tot i que les dues tècniques es basen en el mateix tipus d'agrupació, les característiques considerades per cadascuna d'elles estan relacionades amb diferents aspectes de localització que permeten separar les fonts en mescles instantànies i reals. Addicionalment, es proposen dues tècniques de post-processament enfocades a millorar l'aïllament de les fonts separades. Les prestacions obtingudes per diversos mètodes de SSS a la resíntesi d'escenes sonores amb WFS és posteriorment avaluada mitjançant tests subjectius, posant especial atenció al canvi observat en els atributs de percepció espacial. Encara que les fonts estimades són versions distorsionades de les originals, els efectes d'emmascarament que es produeixen en la remescla espacial provoquen que els artefactes siguen més difícilment percebuts, millorant la qualitat subjectiva global. La tesi finalitza amb una sèrie de nous desenvolupaments relacionats amb l'aplicació del processament temps-freqüència a la localització de fonts i a la millora espacial de la reproducció de so.

Paraules Clau: Síntesi de Camp d'Ones, Separació de Fonts Sonores, Processament Temps-Freqüéncia, Direcció d'Arribada, Qualitat de So Espacial.

Acknowledgements

This thesis would have not been possible without significant advice, stimulation and encouragement from many people. First and foremost I offer my sincerest gratitude to my supervisor, Dr. José Javier López, who has supported me throughout this thesis with his knowledge and advice whilst allowing me the room to work in my own way.

I am very grateful to Prof. Dr. Rudolf Rabenstein from the University of Erlangen-Nuremberg and to Dr. Nicolás Ruiz from the Universidad de Jaén, for serving as reviewers of this thesis and for providing me with very useful comments that helped to improve the final manuscript.

I would like to thank Dr. Sascha Spors, who apart from being a member of the committee, hosted me at the Deutsche Telekom Laboratories in Berlin. I really appreciate the opportunity he gave me and his kind support during the months I spent working with his team.

It is a pleasure for me to thank those people who were involved in the AnClaS^3 project, especially to Dr. Manuel Rosa. A special mention goes to the people from the Universidad de Jaén: Dr. Pedro Vera, Francisco Quesada, Julio Carabias and to my good friend Dr. José Escolano.

A person who has taught me many important things during my whole academic career is Dr. Basilio Pueo from the Universidad de Alicante. I owe my deepest gratitude to him for his valuable advice and his continuous technical support, which have contributed in a very important way to the development of this thesis.

I would like to show my gratitude to all the people at the Universidad Politécnica de Valencia that shared my daily work at the Institute of Telecommunications and Multimedia Applications. In particular, I would like to thank Prof. Alberto González, Dr. Gema Piñero, Dr. María de Diego, Dr. Miguel Ferrer and Dr. Germán Ramos. Of course, I do not forget my current and former colleagues at the GTAC: Laura Fuster, Emanuel Aguilera, Carmen Botella, Pedro Zuccarello, Sergio Tejero, Eliseo Villanueva, Fernando Domene, Luis Maciá, Maria Cabanes and Beatriz Padilla. I really enjoyed the friendly atmosphere in the lab, especially the sound of Laura's triangle, the time for our daily crossword and our tasty “meriendas”.

I am very thankful to the people I met in Berlin, mainly to my colleagues at the Deutsche Telekom Labs, Jens Ahrens, Mattias Geier, Tania Villa, Raúl Salamanca and Pablo Vidales, and to my friends Juan, José Antonio, Borja, David and Sonsoles. Thank you for making me feel like being at home. Moreover, I also want to thank my colleagues at the Technical University of Vienna, Dr. Andreas Rauber and his team, for our interesting discussions during my stay at their department.

A special word of appreciation goes to Dr. Juan José Burred from IRCAM and Dr. Diemer de Vries from the Technical University of Delft. Also to Jesús “Suso” Ramallo, who provided me with some of the music tracks used in the experiments described in this thesis.

I wish also to thank the Spanish Ministry of Science and Innovation for the received financial support under the FPU program.

Thanks to my friends Ariana Salieto, Jose Vicente Galán and Fulvio Grassi for their friendship and favors during my university years in Valencia. Also thanks to my friends at the

Universidad de Alicante: Roger, Carlitos, Majo, Cris, Fer, Roda, Sergio, Parra and Sandra. I will always remember the wonderful moments we have spent together.

Special thanks to my good friend Jorge Francés for all the good times we had during our engineering studies.

I would also like to thank the members of *The Crashed Bones*, not only because they are probably the best rock band in the world, but also because they always allow me to forget all my hardest moments.

It is a pleasure for me to acknowledge the help and encouragement given by my loving family.

I am deeply indebted to my parents, Máximo and Carmen, and to my sisters, Silvia and Mari Carmen, for their endless confidence and fondness.

A warm hug to my grandparents and kisses to my little nephew Jorge.

Finally, I would like to express my deepest gratitude and affection to Sandra, whose continuous patience and support has made it possible for me to complete this thesis. Finding you was the best thing that could have ever happened to me when I decided to start this challenge. I love you.

Máximo Cobos
June 2009

Contents

List of Figures

List of symbols

$\mathbf{X}$	Matrix
$\mathbf{x}$	Vector
x	Scalar
$(\cdot)^{\mathrm{T}}$	Transpose
$(\cdot)^{*}$	Complex conjugation
$(\cdot)^{\mathrm{H}}$	Conjugate transpose
$(\cdot)^{+}$	Moore-Penrose pseudoinverse
$\lvert \cdot \rvert$	Absolute value
$\lVert \cdot \rVert_p$	ℓ_p norm
n	Source index
m	Sensor index
t	Discrete time
N	Number of sources
M	Number of mixtures/sensors
T	Number of signal samples
$x(t)$	Time-domain signal
K	Number of expansion functions
k	Frequency index
r	Time-frame index
R	Number of time-frames
$\mathbf{X}(k,r)$	Time-frequency signal matrix
$X(k,r)$	Time-frequency signal samples (elements of $\mathbf{X}(k,r)$)
$\lvert X(k,r)\rvert$	Magnitude spectra
$\angle X(k,r)$	Phase spectra
$\hat{(\cdot)}$	Estimation
$\langle\cdot\rangle$	Dot (scalar) product
$*$	Element-wise convolution
$\circ$	Hadamard (element-wise) product
$./$	Element-wise division
$\lVert \cdot \rVert_F$	Frobenius norm
$P(A)$	Marginal probability of A
$P(A\vert B)$	Conditional probability of A, given B
$(\cdot)^{+} \leftarrow$	Value update

Abbreviations and Acronyms

ADRess	Azimuth Discrimination and Resynthesis
ASW	Auditory Source Width
BAQ	Basic Audio Quality
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival
DR	Detection Rate
DTS	Digital Theater Systems
DWT	Discrete Wavelet Transform
DUET	Degenerate Unmixing Estimation Technique
EC	Equalization Cancellation
EM	Expectation Maximization
ENSIR	Energy Normalized Source to Interference Ratio
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HOA	High Order Ambisonics
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
IBM	Ideal Binary Mask
IC	Interaural Coherence
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
ILD	Interaural Level Difference
ISR	source to Image Spatial distortion Ratio
ISW	Individual Source Width
ITD	Interaural Time Difference
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
LFE	Low Frequency Effects
LMM	Laplacian Mixture Model
MAP	Maximum A Posteriori
MDCT	Modified Discrete Cosine Transform
MOS	Mean Opinion Score
MuLeTS	Multi Level Thresholding Separation
MUSHRA	MUltiple Stimuli with Hidden Reference and Anchor
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PIW	Panning Index Window
SBSS	Semi-Blind Source Separation
SAOC	Spatial Audio Object Coding
SAR	Source to Artifacts Ratio
SASSEC	Stereo Audio Source Separation Evaluation Campaign
SIR	Source to Interference Ratio

SiSEC	Signal Separation Evaluation Campaign
SDR	Signal to Distortion Ratio
SRP	Steered Response Power
SSS	Sound Source Separation
STCF	Short Time Coherence Function
STFT	Short Time Fourier Transform
TIR	Target to Interference Ratio
TDOA	Time Delay Of Arrival
VBAP	Vector Base Amplitude Panning
WDO	W-Disjoint Orthogonality
WFS	Wave Field Synthesis

1 Introduction and Scope

Introduction and Scope 1

1.1 Background

Over the last few decades, surround sound, or multichannel sound reproduction systems, have played an increasingly important role in the entertainment industry, as well as in the multimedia field. While high-fidelity audio devices currently demonstrate flat frequency responses and very low levels of noise and distortion, reproduction issues involving spatial quality or spatial fidelity are still an active research field. In fact, multichannel audio systems try to reproduce sound in a way which is more "natural" with the aim of enhancing the listening experience [1]. This is possible due to the improved spatial attributes of these systems, such as the ability to facilitate the perception of sound sources coming from different directions.

Although five channel systems are a consolidated standard in multichannel audio today, there is increasing interest in emerging reproduction systems based on sound field rendering. The most popular of these systems is *Wave Field Synthesis* (WFS), a spatial reproduction system capable of synthesizing an acoustic field in an extended area by means of loudspeaker arrays. This makes the reproduced sound scene independent from the listening position, and therefore the relative acoustic perspective perceived by a listener changes as he or she moves. The idea of an *acoustic curtain* aimed at transporting the acoustics of the recording venue to a reproduction room using microphone and loudspeaker arrays was described by Snow in 1953 [2]. However, although this may be considered as a WFS "*avant la lettre*", it misses the physical insight and background that forms the basis of the real invention. It was not until the late 80s when the theory of WFS was introduced in the works published by the Delft University of Technology [3][4][5] and which led to the first WFS prototypes.

Despite all of the advances made in spatial sound reproduction over the last few years, the vast majority of musical recordings are stored and supplied in a two-channel (stereo) format, making it necessary to listen to them on a two-loudspeaker reproduction system. In this context, audio signal processing systems for converting stereo recordings into four or five channels are gaining attention. These *up-mixers* are used for reproducing conventional stereo recordings with more advanced spatial reproduction systems, taking advantage of the spatial properties of multichannel audio reproduction. Most stereo-to-5.1 up-mixers are usually based on a matrix scheme, which generates the additional channels by simple adding and subtracting the input channels with altered gain and phase.

As WFS systems are not yet widely deployed, up-mixing processors fully designed to convert stereo recordings into synthesized scenes have been rarely discussed in the literature. The main objective of stereo-to-WFS up-mixers would be the same as in the case of five-channel up-mixing: to enhance the spatial quality of conventional stereo recordings. However, the spatial properties of WFS, which are ideally suited to be combined with virtual and augmented reality systems and other applications, open a new door to go further than the conventional home-theater oriented up-mixing. Moreover, WFS needs the signals of each source to be available before rendering the sound field, thus an object-based processing becomes necessary. From this point of view, more sophisticated up-mixing schemes must be considered, being *source separation* algorithms a potential solution to this problem.

Source separation problems in digital signal processing are those in which several signals have been mixed together and the objective is to find out what the original signals were. Algorithms for source separation have been shown to be very useful in many areas, ranging from image and video processing to biomedical applications. In the audio field, algorithms aimed at extracting different sound sources from a set of audio mixtures are usually denoted as *Sound Source Separation* (SSS) algorithms. Since the introduction of *Independent Component Analysis* (ICA) in the early 1990's [6], the source separation field has become one of the most active research areas in signal processing. Together with the statistical/mathematical framework set by ICA methods, the development of biologically-inspired computational models for source separation have also led to another popular discipline in SSS, known as *Computational Auditory Scene Analysis* (CASA) [7]. In fact, these systems try to mimic one of the most surprising properties of human hearing: the ability to distinguish individual sound sources from complex mixtures of sound. This human ability is usually related to the well known *"cocktail party effect"* discussed by Cherry in 1953 [8], which describes the ability to focus one's listening attention on a single talker among a mixture of conversations and background noises, ignoring other conversations and enabling humans to talk in a noisy place.

The difficulty underlying source separation problems is mainly related to the amount of information that is known about the sources and the nature of the mixtures. When very little information about the mixing process or the sources is known in advance, the term *"blind"* is normally used. That is the reason why the label *Blind Source Separation* (BSS) has been widely accepted to denote statistical algorithms, such as ICA-based approaches. However, strictly speaking, there are no fully blind algorithms, since general assumptions about their statistical behavior are always taken, most of them related to the *independence*, *non-Gaussianity* or *sparsity* of the sources. Another factor that determines the difficulty of the problem is the proportion between the number of mixture channels and the number of sources. When the number of sources is less (*overdetermined case*) or equal (*determined case*) than the number of observation channels, separation is more easily achieved. When the number of sources exceeds the number of available mixtures (*underdetermined case*), the problem becomes more difficult and stronger assumptions have to be made, often related to the sparsity of the sources. Moreover, the mixing process that generated the mixture channels has also great influence on the separation difficulty, leading to *instantaneous*, *anechoic* and *convolutive* approaches.

One of the most demanding applications of SSS algorithms is audio up-mixing. This application is intended to generate a high-quality multitrack recording from a final mix as contained in a conventional stereo CD. Current algorithms are far from providing separated sources with a quality similar to that obtained by recording sources separately, making most of these applications unfeasible. However, if the separated tracks are mixed again with different gains or spatial distributions, artifacts on the separated tracks are greatly masked by the rest of sources in the final reproduced scene and the overall quality is substantially improved.

As pointed out, WFS needs to have separate audio signals for the different virtual sources that compose a sound scene, thus, it is strictly needed an object-oriented up-mixing based on source separation. There are some benefits for SSS algorithms to be used in the WFS framework. Obviously, the first one is the possibility to synthesize spatially enhanced scenes from audio material stored in conventional formats, such as stereo or 5.1. The second one is that the listening experience would not only be enhanced in terms of spatial fidelity, but in terms of interactivity, enabling the user to modify the level and spatial position of the sources. Finally, this possibility would help WFS systems to be more easily adopted by the audio industry, since WFS systems have experienced a considerable slow development in the market during the last decade.

1.2 Motivation

This dissertation focuses on separation of audio mixtures in the context of stereo to WFS up-mixing. The main motivation is the use of SSS algorithms as a powerful tool for the resynthesis of sound scenes by means of advanced spatial audio reproduction systems. This was one of the main goals set in the *AnClaS*^{3} (Analysis and Classification for Sound Source Separation) project, which was supported by the *Spanish Ministry of Science and Innovation* and was carried out with the collaboration of five Spanish universities.

The considered application scenario of SSS algorithms is indeed one of the most complex situations among the different source separation tasks, since most of stereo audio mixtures are underdetermined, i.e. there are more sources than observation channels. As already introduced, algorithms for underdetetermined SSS are based on strong assumptions of the sources and the mixing process. Fortunately, time-frequency transformations provide a sparse representation of audio sources, making valid most of these assumptions. Moreover, it is desirable to perform separation in a fast and unsupervised manner, with a potential for real-time implementation.

On the other hand, although WFS has its own artifacts and practical imperfections that make it impossible to render a desired sound field perfectly, the degradation caused by the separated sources can be far greater. These degradations include timbre modification, burbling artifacts, musical noise and inter-source residuals. As a result, it becomes necessary to evaluate how different separation methods influence the spatial perception of sound scenes, not only in terms of overall quality, but also in the subjective assessment of other attributes related to spatial sound reproduction.

Separation can be optionally improved by means of other post-processing techniques aimed at restoring the estimated signals or eliminating residuals from other sources. These techniques can be thought as *separation-after-separation* algorithms and can be usually applied to the output of several unmixing methods.

1.3 Scope of the Thesis

Taking into account the above context, the main scope of this thesis is as follows:

To contribute new methods for the separation of underdetermined stereo mixtures (instantaneous and anechoic/convolutive), that can be applied to synthetic and real recordings of music and speech. To apply several separation algorithms in the context of stereo-to-WFS up-mixing and to evaluate the spatial quality of the resynthesized sound scenes. To contribute with new post-processing methods aimed at improving the isolation of the separated sources and to evaluate the performance of the proposed approaches. To develop new applications related to the separation framework considered in this thesis, especially those related to the spatial localization and the spatial resynthesis of sound sources recorded with small microphone arrays.

Some particular aims emerge from this main scope, which are presented as follows:

- To study the suitability, advantages and disadvantages of source separation techniques to be used in the context of WFS up-mixing.
- To explore the main problems that arise in this application context and consequently propose new solutions.
- To examine the relationship between objective performance measures used by the source separation community and spatial sound attributes.
- To propose new applications related to the processing used in underdetermined SSS and to integrate them into the spatial sound field.

1.4 Organization of the Thesis

The remainder of this thesis describes the research that has been undertaken to develop the aims stated above. Since the contributions of this thesis fall within two different areas, which are source separation and spatial sound, it seems reasonable to structure the contents of this dissertation into two parts. Note that this two-part division has not been applied to the introductory and the concluding chapters.

The chapters are then organized and presented as follows:

Part I: Sound Source Separation

- Chapter 2: This chapter is intended to give a comprehensive overview of SSS principles and algorithms. It starts by presenting the motivations underlying source separation techniques and the main types of algorithms that have arisen in the audio signal processing field. The different signal models used in source separation are subsequently introduced to the reader, concentrating on the case of underdetermined mixtures and sparsity-based approaches. Some popular SSS methods that are later evaluated in Chapter 5 are also presented. The chapter ends with an overview of the objective performance measures that will be used to evaluate the quality of the separated sources throughout this thesis.

- Chapter 3: This chapter presents a novel approach for the separation of underdetermined stereo mixtures in the time-frequency domain. Inspired by image segmentation techniques, separation is achieved by using a maximum between-class variance criterion between the estimated mixing factors corresponding to the sources present in the mixture. The proposed method computes a set of thresholds that define the different time-frequency masks that separate the sources with little computational cost. The first part of the chapter describes how the proposed approach is applied to the case of underdetermined instantaneous audio mixtures. The second part provides a detailed description of the method for the separation of real mixtures by using a small two-microphone array.

- Chapter 4: This chapter describes two approaches for improving the isolation of separated sources under the framework considered in this thesis. These methods are aimed at removing residuals from other sources in the final extracted signals and they are based on a further analysis of the separated sources in the time-frequency domain. Specifically, an energy-based ratio and a source reassignment technique are introduced, discussed and evaluated.

Part II: Spatial Sound

- Chapter 5: In this chapter, an overview of spatial sound reproduction systems is provided. Stereo, multichannel surround systems and other advanced techniques based on sound field rendering are presented, with special emphasis in the fundamentals of WFS technology. Finally, some concepts involving audio up-mixing and object-oriented coding are also introduced.

- Chapter 6: This chapter deals with the evaluation of the spatial perception of resynthesized sound scenes in WFS using source separation. Experiments carried out to evaluate the performance of several algorithms in the context of WFS up-mixing are explained in detail. The way that source signals distorted by the separation process influence different spatial attributes is analyzed.

- Chapter 7: This chapter presents a set of developments related to the processing used in underdetermined source separation. Without the need for separating sound sources, these algorithms are aimed at localizing multiple sound sources of sound and provide enhanced sound reproduction systems (stereo enhancement and binaural synthesis) based on array processing.

- Chapter 8: Finally, the conclusions obtained throughout this thesis are presented, including some guidelines for future research lines. A list of published work related to this thesis is also given.

Part I

Source Separation

2

Sound Source Separation

2 Sound Source Separation

Source Separation algorithms currently constitute one of the most active research fields in signal processing. Algorithms for source separation have been applied to many areas, ranging from image and video processing to biomedical applications. In the audio context, *Sound Source Separation* (SSS) aims at recovering each source signal from a set of audio mixtures of the original sources, such as those obtained by a microphone array, a binaural recording or an audio CD. Therefore, several applications can emerge from the development of advanced SSS techniques, including music remixing, automatic karaoke, speech enhancement, automatic music transcription or music information retrieval systems. In this chapter, a global framework for source separation is presented, with an emphasis on the underdetermined case, i.e. when there are more sources than mixture channels. A set of popular approaches for the separation of audio mixtures is also described.

2.1 Introduction

Humans are surrounded by sound. If we try to concentrate and listen carefully to the things that happen in our environment, probably we will be able to identify more than one source of sound. This fact reveals an important property of human hearing: the ability to distinguish individual sound sources from complex mixtures of sound. When talking about the perception of speech in complex acoustic environments, this human ability is usually related to the well known "*cocktail party problem*", first described by Cherry in 1953 [8]. The cocktail party effect describes the ability to focus one's listening attention on a single talker among a mixture of conversations and background noises, ignoring other conversations and enabling us to talk in a noisy place [9]. In addition to speech understanding, this special ability plays also a major role in the way humans perceive and feel music. Most of musical compositions are mixtures of different instruments playing simultaneously and we are usually able to concentrate on hearing one of these instruments. A trained subject can even describe the melodies or musical lines played by each instrument and translate them into musical language and notation. Therefore, we can generally talk about an special ability of understanding *sound scenes*. Dan Ellis defined a sound scene as a complex entity of acoustic stimuli that is processed by our brain, resulting in a symbolic description corresponding to a listener's perception of the different auditory events or sources that are present [10].

The complex understanding of sound scenes is so familiar to humans that we usually take it for granted and we perform this task without being aware of it. However, computational methods aimed at imitating our hearing sense and brain for understanding complex sound material are far from performing with such an accuracy and are still an active research line that involves mathematical, physical and psychological issues. In this context, a common approach to achieve understanding of complex sound scenes is usually *Sound Source Separation* (SSS). Generally, the objective of SSS methods is the extraction of each individual audio signal that constitutes a sound mixture, as depicted in Figure 2.1. However, it is important to notice that human understanding of sound scenes is related to the perception of some properties that suggest the existence of certain auditory events from which they can obtain a description of what they are listening. Therefore, to date, there is no necessary evidence for the human auditory system to extract the individual audio signals corresponding to each auditory event [11]. Moreover, it is not clear that SSS is necessary for understanding a sound scene and it could be possible to follow an *understanding without separation* approach as proposed by Scheirer [12]. Nevertheless, having estimations of the sources that conform a sound mixture is for sure a help for analyzing the features corresponding to different auditory events, and therefore, for providing a description of a sound scene. SSS techniques can be roughly divided into three categories: *Blind Source Separation* (BSS), *Semi-Blind Source Separation* (SBSS) algorithms and *Computational Auditory Scene Analysis* (CASA) techniques.

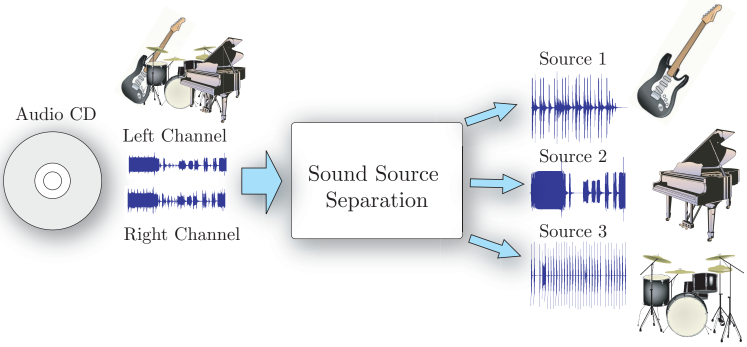


Figure 2.1. Sound source separation scheme.

BSS approaches are very popular in the statistical signal processing and machine learning areas. The term *blind* is used to emphasize that very little information about the sources or the mixing process is known. Generally, several observations of the mixture are available in these kind of methods, but their performance is closely related to the mixing environment. Since Hérault and Jutten published their work in the mid 1980's [13], BSS has always been closely related to *Independent Component Analysis* (ICA). Popular approaches based on ICA assume that the source signals are statistically independent and non-Gaussian [14]. These assumptions are usually sufficient to carry out separation in the linear complete case, when there are as many mixture channels as sources available. When the problem is *underdetermined*, i.e. there are more sources than mixture channels, the difficulty is even higher and stronger assumptions are taken, generally related to the *sparsity* of the sources under some signal transformation [15].

When more than the above statistical assumptions are considered, algorithms are usually classified into SBSS methods. Model-based and hybrid approaches that assume some information about the structure of the sources are easily found in the literature. For example, methods that take into account the harmonic spectral structure of pitched musical instruments or other source-specific priors [16][17] [18]. Moreover, many algorithms make use of other common techniques for learning the underlying spectral structure of musical sources, such us the widely known *Non-Negative Matrix Factorization* (NMF) algorithm [19].

In contrast to pure mathematical techniques, CASA methods are aimed at designing machines capable of hearing the way humans do [7]. The processes underlying the perceptual interpretation of sound sources have been studied for decades, being the work by Bregman [11] one of the most well-known references in this field. These studies on human sound perception have revealed useful information on how sound is processed in the inner ear, leading to engineering prototypes and algorithms that can, with relatively good accuracy, perform tasks like segregation of speech and musical instruments [20][21], automatic music transcription or estimation of multiple pitches from noisy mixtures [22]. The biologically-inspired models used by CASA methods have the capability to perform separation of sound sources in the monaural case, i.e. with only one mixture channel. However, the features used by these models are so specific that separation is only successfully achieved under very specific situations. Therefore, their applicability is not as wide as the one of BSS approaches.

Throughout this chapter, it will be seen that existent SSS approaches are very varied. The chapter begins introducing the different types of audio sources and audio mixtures in Section 2.2, making special emphasis on stereo audio mixtures. The mixing process, which has a great influence on the design of separation algorithms, is mathematically expressed by means of a *mixing matrix*, which is usually assumed to be unknown. Section 2.3 describes the mixing models that usually appear in BSS problems: *instantaneous* (or *linear*), *anechoic* (or *delayed*) and *convolutive* (or *echoic*). In Section 2.4, the source separation problem is presented under a probabilistic framework, leading to the two general approaches to BSS: the *joint approach* and the *staged approach*. In the joint approach, both the mixing matrix and the sources are estimated at the same time. In the staged approach, the mixing matrix is estimated in the first place and the demixing is tackled in a second stage. Section 2.5 provides the basics of underdetermined SSS. Signal decompositions and transformations are reviewed, paying close attention to time-frequency representations, specially to the *Short-Time Fourier Transform* (STFT). Time-frequency models will appear many times throughout this thesis, as they provide the basic processing frontend for almost all the contributions contained in this work. Section 2.6 and Section 2.7 describe some well-known SSS algorithms for stereo and monaural source separation, respectively. The basic ideas underlying some popular separation approaches are here described, including the *Degenerate Unmixing Estimation Technique* (DUET), *Azimuth Discrimination and Resynthesis* (ADRess) or NMF. Finally, Section 2.8 introduces the performance evaluation measures used throughout this thesis to evaluate the quality of the separated source tracks.

The general framework for source separation presented in this chapter has been mostly inspired by other excellent overviews on the topic, especially the ones published by Burred [23], O'Grady *et al.* [15] and Vincent *et al.* [24].

2.2 Sources and Mixtures

2.2.1 Audio Sources

The temporal and spectral features of audio sources are essential for many separation algorithms. Next, the basic structures of the most important audio sources (music and speech) are briefly described.

Speech Sources

Speech is the most important sound produced by humans using the vocal folds [25]. These, in combination with the articulators, are capable of producing highly intricate arrays of sound and can suggest emotions such as anger, surprise, or happiness. Speech sources can be thought as a sequence of discrete units called *phonemes*. Speech signals are non-stationary and have a characteristic structure that includes a periodic part containing harmonic sinusoidal *partials* and a transient or noisy part. Sinusoidal partials are multiples of a single frequency called the fundamental frequency, or *pitch*. The pitch varies over time, but stays within a range of about 40 Hz centered around an average of 140 Hz for male and 200 Hz for female speakers.

Music Sources

Musical instruments and singers produce sequences of *notes*. The signals produced by notes also follow a basic structure, where a transient part is usually followed by a near-periodic part of harmonic sinusoidal partials. In the musical context, the term pitch denotes the listener's subjective judgement as to where a note is located on the musical scale. This perception not only depends on the fundamental frequency but on other factors related to the amplitude and loudness of the sound [26]. Nevertheless, and for the sake of simplicity, the term pitch is usually assumed to be equivalent to the fundamental frequency. Fundamental frequencies in music vary slower in time than speech does, and usually on discrete steps of the semitone scale, which spans logarithmically the range from 30 Hz to 4 kHz.

On the other hand, the term *timbre* denotes the quality or color of the sound of a musical note and it basically depends on the harmonic content of the note, i.e. which harmonics are present and what their relative strengths are. Other factors, such as the duration of onset transients, also have an influence on the subjective perception of timbre. As western harmony rules are based on specific frequency ratios (2/1, 3/2 or 5/4), there is usually some overlapping of frequencies when musical instruments play together.

2.2.2 Audio Mixtures

Mixtures of audio sources can be acquired in many ways, having the recording set-up a big influence on the different separation approaches found in the literature. In this context, *synthetic audio mixtures*, which are artificially created, are very different in nature from *real live recordings*, where the mixtures are obtained by capturing several sources that have been physically mixed. Microphone arrangements may involve near-field and far-field microphones with certain directional patterns. Moreover, *binaural recordings* obtained by means of dummy heads are also of interest for the development of new hearing-aid applications. Thus, the spatial properties of the observed signals highly depend on their mixing environment.

In the case of real live recordings, the room where the sound is recorded plays also an important role, as the recorded signals are the total contribution of direct path signals from the sources and the reflections that occur inside the room [27]. Therefore, the observed mixtures are filtered versions of the sources due to successive reflections on the room surfaces. These reflections can be divided into *early reflections* and *late reflections.* Early reflections may be calculated easily from the room geometry, but late reflections make off the room *reverberation*, which has an stochastic structure. Reverberation is usually characterized by the room *reverberation time*, RT_{60}, which is the time required for reflections of a direct sound to decay by 60 dB below the level of the direct sound. Reverberation time is defined for wide band signals. When talking about the decay of an individual frequency, the term *decay time* is used. Normal RT_{60} values range from 100 ms to 500 ms in office rooms and more than 1 s in large venues. Reverberation is also highly related to the *wall reflection factor* ρ. If a plane wave strikes a plane and uniform wall of infinite extent, in general a part of the sound energy will be reflected from it in the form of a reflected wave originating from the wall, the amplitude and the phase of which differ from those of the incident wave. The intensity of the reflected wave is smaller by a factor $|\rho|^2$ than that of the incident wave.

Music productions from studios (such as pop music or movie soundtracks) are often made by recording separately the sources in a near-anechoic room with a single microphone and applying posterior effects to each source. Common effects include panoramic and reverberation. The *panoramic pot* modifies the spatial location of the sources in the mix by applying a different factor scale in each channel. Reverberation simulates the effect introduced by a room by applying filters to the sources. In the *down-mix* step, all the sources are added synthetically by means of an audio workstation or a mixer. Synthetic mixing can also be used to add the channels of a live recording. Figure 2.2 shows several mixing set-ups classified according to their type of mixing, which will be described in Section 2.3.

Stereo Mixtures

Stereophony is still the most common format for sound recording and reproduction. Although multichannel recordings for 5.1 reproduction systems have been widely available since the arrival of DVDs, they have not displaced the classic stereo format yet. The vast majority of CDs, MP3s, FM radios and TV broadcasts are in stereo.

The motivation of stereo recording and reproduction relies on the fact that the physical superposition of two loudspeakers enables the building of a *phantom source*, which is understood as a substitute sound source. In Section 5.3.1 the insights of stereophonic reproduction will be further described.

The position of the phantom source can be modified according to the gains applied to a source in the left and right channels during the mix-down process, a technique known as *amplitude panning.* This effect on the perceived position is managed by means of the panoramic parameter $\phi_n \in [0, 1]$, which defines the scaling factors that multiply the source signal in each of the mixture channels (α^L and α^R). Two panning laws are popular in this context:

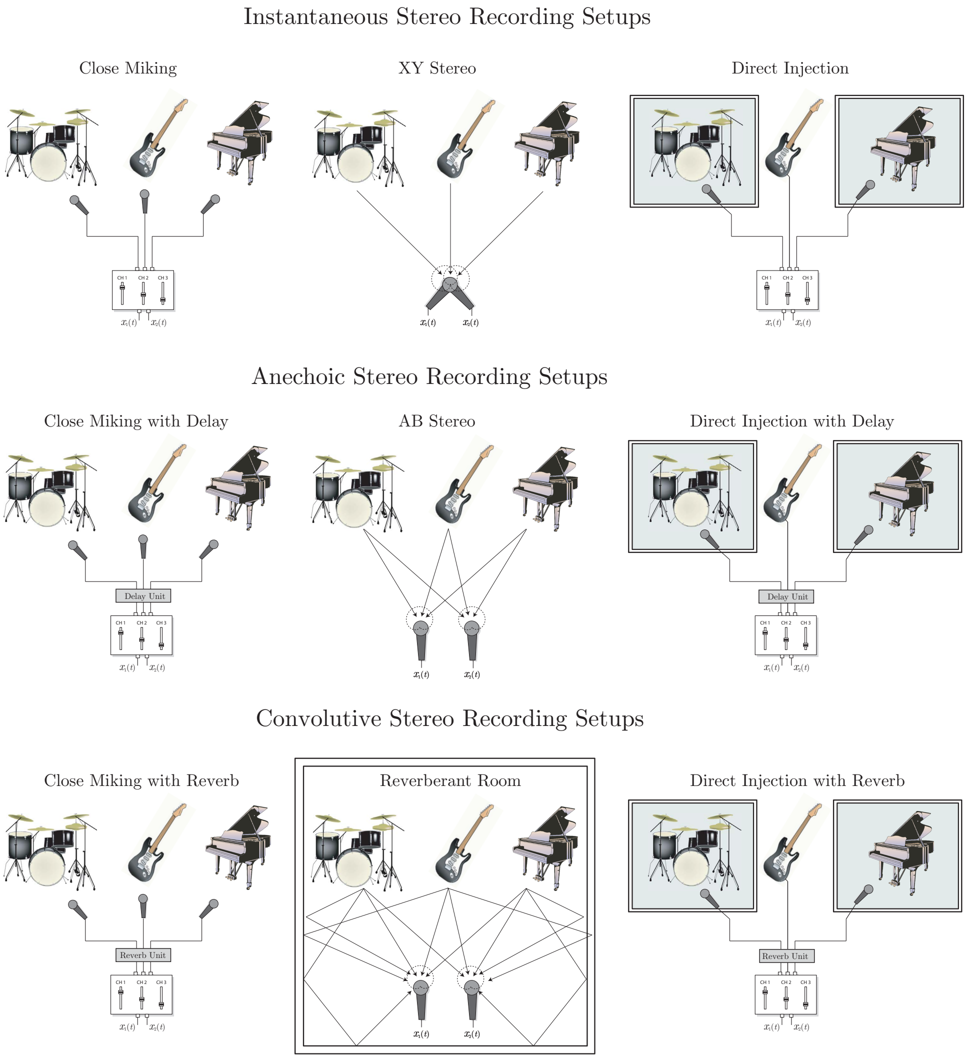

Instantaneous Stereo Recording Setups
Close Miking
XY Stereo
Direct Injection
Anechoic Stereo Recording Setups
Close Miking with Delay
AB Stereo
Direct Injection with Delay
Delay Unit
Convolutive Stereo Recording Setups
Close Miking with Reverb
Reverberant Room
Direct Injection with Reverb
Reverb Unit
CH 1
CH 2
CH 3
$x_1(t)$
$x_2(t)$

Figure 2.2. Different recording setups classified according to their mixing model.

- **Amplitude constant law**: $\alpha_n^L + \alpha_n^R = 1$

$$\begin{aligned} \alpha_n^L &= (1 - \phi_n), \\ \alpha_n^R &= \phi_n, \\ \phi_n &= \frac{\alpha_n^R/\alpha_n^L}{1 + \alpha_n^R/\alpha_n^L}. \end{aligned} \tag{2.1}$$

- **Power constant law**: $(\alpha_n^L)^2 + (\alpha_n^R)^2 = 1$

$$\begin{aligned} \alpha_n^L &= \cos\left(\frac{\phi_n \pi}{2}\right), \\ \alpha_n^R &= \sin\left(\frac{\phi_n \pi}{2}\right), \\ \phi_n &= \arctan\left(\frac{\alpha_n^R}{\alpha_n^L}\right) \cdot \frac{2}{\pi}. \end{aligned} \tag{2.2}$$

The constant power law is the most used configuration in digital audio workstations and mixing desks [28] [29]. In addition to amplitude panning, there are a set of popular stereo recording techniques based on specific microphone setups. These techniques achieve the stereo effect due to the directional characteristics of the microphones and their relative placement, which provide an *Interaural Level Difference* (ILD) and/or and *Interaural Time Difference* (ITD) that enhance the spatial properties of the recorded scene. This is due to the fact that ILD and ITD constitute the basic azimuth localization cues used by the human auditory system, as will be explained in Section 5.2.1. The interested reader can find more information on these kind of arrangements for stereo recordings in [30].

In the following section, the mathematical models used to characterize different types of mixtures are presented. Most BSS methods are based on the these models, being one of the criteria used for the classification of separation algorithms.

2.3 Mixing models

As previously commented in the last section, mixing scenarios are very varied and they determine the nature of the resulting observed mixtures. Generally, the different mixing situations can be mathematically expressed by means of a model that describes how the observations are generated. This is the reason why these models are called *generative models*. Before introducing these models, it is important to clarify the notation used in this thesis for sampled signals. Although an academic distinction between continuous (t) and discrete $[n]$ time variables is normally used in signal processing works ($s[n] = s(nT_s)$, being T_s the sampling interval), all the signals considered hereafter are assumed to be discrete. Therefore, this distinction is not necessary and the notation (t) has been chosen as the widely adopted in the source separation field, where $t = 1, \ldots, T$ denotes discrete time observations and source signals are indexed by $n = 1, \ldots, N$.

We consider a general setup where M sensors are exposed to N sound sources. As established by the *principle of superposition*, the electrical signal at the m-th channel resulting from this setup can be mathematically expressed as the scalar addition of the instantaneous amplitudes corresponding to the different *source images*:

$$x_m(t) = \sum_{n=1}^{N} s_{mn}(t), \quad m = 1, \ldots, M, \tag{2.3}$$

where $s_{mn}(t)$ is the image of the n-th source in the m-th microphone at time sample t. These images of the sources represent how the original source signals $s_n(t)$ are recorded at each sensor after being modified by the mixing process (which in the general case can be modeled by a filter $h_{mn}(t)$). Figure 2.3 shows the relations existent between all these signals with an example with two microphones ($M = 2$) and two speakers ($N = 2$).

In the following subsections the different models in source separation are described. The images of the sources vary depending on the type of mixing considered, which is mathematically represented by a *mixing matrix*. According to the mixing conditions and the nature of this matrix, three mathematical formulations of the mixing process can be defined: the *instantaneous* (or *linear*), the *anechoic* (or *delayed*) and the *convolutive* (or *echoic*) *mixing models*.

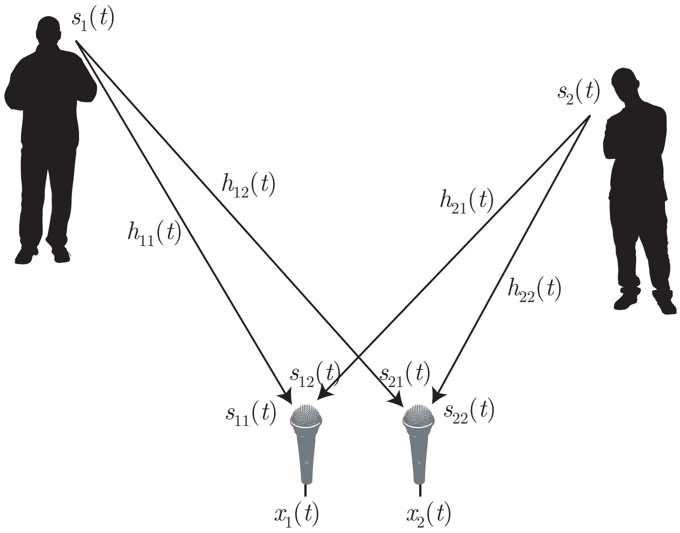


Figure 2.3. Two microphones picking up the signals from two speakers and the signals involved in the mixing process.

2.3.1 Instantaneous Model

The simplest mixing model is the *instantaneous* or *linear* model. In this model, the mixtures are formed by linear combinations of the sources. Therefore, the mixtures are obtained by summing scaled versions of the sources:

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t), \quad m = 1, \ldots, M, \tag{2.4}$$

where a_{mn} are scalar factors. Thus, the images of the sources are then given by

$$s_{mn}(t) = a_{mn} s_n(t). \tag{2.5}$$

Alternatively, the instantaneous model can be expressed as a system of linear equations in the form

$$\begin{array}{rcl} x_1(t) & = & a_{11}s_1(t) + a_{12}s_2(t) + \cdots + a_{1N}s_N(t) \\ x_2(t) & = & a_{21}s_1(t) + a_{22}s_2(t) + \cdots + a_{2N}s_N(t) \\ \vdots & & \\ x_M(t) & = & a_{M1}s_1(t) + a_{M2}s_2(t) + \cdots + a_{MN}s_N(t). \end{array} \tag{2.6}$$

Taking into account the above system, it is usual to find the mixing models in a compact matrix formulation:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{bmatrix}, \tag{2.7}$$

or equivalently

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.8}$$

where $\mathbf{x} = [x_1(t), \ldots, x_M(t)]^{\mathrm{T}}$ is a $M \times 1$ vector of mixtures, $\mathbf{A}$ is the $M \times N$ mixing matrix and $\mathbf{s} = [s_1(t), \ldots, s_N(t)]^{\mathrm{T}}$ is a $N \times 1$ vector of sources. If a collection of individual time samples of the mixture and source signals are considered, the model can be represented as:

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{2.9}$$

where $\mathbf{X}$ is the $M \times T$ matrix corresponding to the sensor data at times $t = 1, \ldots, T$:

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ x_M(1) & x_M(2) & \cdots & x_M(T) \end{bmatrix}, \tag{2.10}$$

and $\mathbf{S}$ is the $N \times T$ matrix of source signals:

$$\mathbf{S} = \begin{bmatrix} s_1(1) & s_1(2) & \cdots & s_1(T) \\ s_2(1) & s_2(2) & \cdots & s_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ s_N(1) & s_N(2) & \cdots & s_N(T) \end{bmatrix}. \tag{2.11}$$

Note that under this notation, each row of $\mathbf{X}$ and $\mathbf{S}$ corresponds to one of the mixture and source signals, respectively. The notation used in Eq.(2.8) is often referred as the model in *instantaneous notation*, as it represents the generation of the mixtures in a single time sample. On the other hand, the notation of Eq.(2.9) is referred as the model in *explicit notation* and it describes the generation of the mixtures in the whole observation time.

2.3.2 Anechoic Model

The *anechoic* or *delayed model* can be thought as an extension of the instantaneous model where, in addition to different gain factors, different transmission delays between the sources and the sensors are considered. This is equivalent to an anechoic mixing scenario, where only the direct path between each source and sensor has influence on the mixture. The generative model is

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn}), \quad m = 1, \ldots, M, \tag{2.12}$$

where δ_{mn} is the arrival delay between source n and sensor m, and a_{mn} stands for the amplitude factor corresponding to the path between source n and sensor m.

The images of the sources are scaled and delayed versions of the sources:

$$s_{mn}(t) = a_{mn} s_n(t - \delta_{mn}). \tag{2.13}$$

The mixing matrix has the form

$$\mathbf{A} = \begin{bmatrix} a_{11}\delta(t-\delta_{11}) & a_{12}\delta(t-\delta_{12}) & \cdots & a_{1N}\delta(t-\delta_{1N}) \\ a_{21}\delta(t-\delta_{21}) & a_{22}\delta(t-\delta_{21}) & \cdots & a_{2N}\delta(t-\delta_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}\delta(t-\delta_{M1}) & a_{M2}\delta(t-\delta_{M2}) & \cdots & a_{MN}\delta(t-\delta_{MN}) \end{bmatrix}, \tag{2.14}$$

where $\delta(t)$ are Kronecker[1] deltas. Note that the operator $\delta(t-\delta_{mn})$ is used to denote a delay between source n and sensor m. With this notation, the model can be compactly expressed by

$$\mathbf{x} = \mathbf{A} * \mathbf{s}, \tag{2.15}$$

where $*$ denotes the element-wise convolution operation.

2.3.3 Convolutive Model

In the *convolutive* or *echoic model*, reflections occurring in the mixing environment are considered too. Mathematically, the process can be written as

$$x_m(t) = \sum_{n=1}^{N} \sum_{\tau=1}^{L_{imp}} a_{mn\tau} s_n(t - \delta_{mn\tau}), \quad m = 1, \ldots, M, . \tag{2.16}$$

where L_{imp} is the number of paths the source signal can take to the sensors. Therefore, the images of the sources are filtered versions of the original sources:

$$s_{mn}(t) = \sum_{\tau=1}^{L_{imp}} a_{mn\tau} s_n(t - \delta_{mn\tau}). \tag{2.17}$$

The mixing matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} \sum_{\tau=1}^{L_{imp}} a_{11\tau}\delta(t-\delta_{11\tau}) & \cdots & \sum_{\tau=1}^{L_{imp}} a_{1N\tau}\delta(t-\delta_{1N\tau}) \\ \vdots & \ddots & \vdots \\ \sum_{\tau=1}^{L_{imp}} a_{M1\tau}\delta(t-\delta_{M1\tau}) & \cdots & \sum_{\tau=1}^{L_{imp}} a_{MN\tau}\delta(t-\delta_{MN\tau}) \end{bmatrix}, \tag{2.18}$$

thus, a convolutive formulation of the form $\mathbf{x} = \mathbf{A} * \mathbf{s}$ is also used here. Note that the anechoic and instantaneous models can be thought of as particular cases of the convolutive model.

2.3.4 Noisy models

In real life, there is always some kind of noise present in the observations. Noise can come from measuring devices or from any inaccuracies in the model used. Therefore, a noise term is sometimes included in the above models:

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} + \mathbf{n}, \tag{2.19}$$

where $\mathbf{n} = [n_1(t), n_2(t), \ldots, n_M(t)]^{\mathrm{T}}$ is the $M \times 1$ noise vector and $\star$ denotes the model dependent operator (matrix product in instantaneous mixtures and element-wise convolution in the anechoic and convolutive models). Noise is often assumed to be white, Gaussian and uncorrelated, i.e. having diagonal covariance matrix in the form $\sigma^2\mathbf{I}$, where σ^2 is the variance of one of its M components.

The separation methods presented in this thesis are based on noise-free models. However, the probabilistic approach to BSS described in Section 2.4.2 assumes that additive noise is added to the observed mixtures, resulting in the above noisy model.

[1] The Kronecker delta is defined in signal processing as $\delta(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}$. The alternate notation for Kronecker deltas found in other works, δ_{ij}, must not be here confused with the source-sensor delay δ_{mn}.

2.4 Source Separation Tasks and Approaches

The source separation problem consists in estimating the source spatial images of the sources $s_{mn}(t)$, from the mixture signals $x_m(t)$. Note that the estimation of the single-channel source signals $s_n(t)$ involves undoing the filtering effect of the mixing process (dereverberating), which is an additional problem that will not be considered in this thesis.

The instantaneous mixing model $\mathbf{X} = \mathbf{AS}$, has the form of a conventional system of linear equations. Although it seems that the problem of extracting the sources $\mathbf{S}$ from the mixtures $\mathbf{X}$ can be completely solved by traditional algebraic techniques, this is only possible if the mixing matrix $\mathbf{A}$ is known. However, source separation tries to give solution to this problem in the case were both $\mathbf{S}$ and $\mathbf{A}$ are unknown. Moreover, even if the mixing matrix $\mathbf{A}$ can be accurately estimated, the system is only invertible if $\mathbf{A}$ is square and has full rank, thus, the number of equations must be equal to the number of unknowns ($M = N$). When the problem is overdetermined ($M > N$), dimensionality reduction techniques such as *Principal Component Analysis* (PCA) [31] are usually employed. If the problem is underdetermined ($M < N$), there is an infinite number of solutions and demixing the sources from the mixtures becomes a very challenging task.

In the next subsections, source separation approaches are presented in detail. First, several criteria that are commonly used to classify separation problems are introduced, followed by a description of the joint and staged BSS approaches.

2.4.1 Problem Classification

The separation difficulty is mainly related to three different aspects: the relative number of mixture channels and sources, the length of the mixing filters and the time variation of the mixing filters [15]. These three criteria are used to characterize the mixtures in the following way:

- Relative number of mixture channels and sources:
    1. $M > N$ *Overdetermined mixture.*
    2. $M = N$ *Determined mixture.*
    3. $M < N$ *Underdetermined mixture.*
- Mixing filters:
    1. Scalars (zero delay): *Instantaneous mixture.*
    2. Scalars and/or Delays (possibly fractional): *Anechoic mixture.*
    3. Otherwise: *Convolutive mixture.*
- Time variation of the mixing filters:
    1. Static sources or fixed filters: *Time-invariant mixture.*
    2. Moving sources or time-varying filters: *Time-variant mixture.*

Note that the formulation of the mixing processes described in Section 2.3 only considers static sources, i.e. fixed filters. Musical mixtures are usually underdetermined, since the most popular format for music recordings is stereo and there are usually more than two instruments playing. Overdetermined and determined situations usually appear in microphone array processing techniques, which are usually used for source localization and tracking [32].

2.4.2 Joint and Staged Approaches

This section formulates the source separation problem from a probabilistic point of view. This formulation has similarly appeared in several BSS works, including the ones by Olshausen and Field [33], Zibulevsky [34], Virtanen [35], Abdallah [36] or Burred [23].

Let us define the separation problem as an optimization problem by setting an appropriate *cost function*. Such a cost function can be constructed using a measure of distance between $\mathbf{X}$ and the product $\mathbf{AS}$. One typical measure is the square of the Euclidean distance between $\mathbf{X}$ and $\mathbf{AS}$. Using the explicit notation of the instantaneous mixing model, the separation problem can be thought as a minimization of the Frobenius[2] norm of the error matrix, yielding

$$\min_{\mathbf{A},\mathbf{S}} \|\mathbf{X} - \mathbf{AS}\|_F^2. \tag{2.20}$$

Approaches formulated in this way are termed *joint source separation* methods, since they estimate both unknown quantities, $\mathbf{S}$ and $\mathbf{A}$, at the same time. This is insufficient to fully constrain the solution, since an optimum can be transformed into another equivalent optimum by $\mathbf{S} \rightarrow \mathbf{US}$, $\mathbf{A} \rightarrow \mathbf{AU}^{-1}$, being $\mathbf{U}$ any invertible matrix. One popular approach to further constrain the problem is to formulate it under a Bayesian perspective. Thus, a *Maximum A Posteriori* (MAP) formulation can be applied with the aim of maximizing the posterior probability:

$$\max_{\mathbf{A},\mathbf{S}} P(\mathbf{A}, \mathbf{S}|\mathbf{X}). \tag{2.21}$$

According to Bayes' theorem, and assuming that $\mathbf{A}$ and $\mathbf{S}$ are statistically independent, $(P(\mathbf{A}, \mathbf{S}) = P(\mathbf{A})P(\mathbf{S}))$ this posterior is given by

$$P(\mathbf{A}, \mathbf{S}|\mathbf{X}) = \frac{P(\mathbf{X}, \mathbf{A}, \mathbf{S})P(\mathbf{A})P(\mathbf{S})}{P(\mathbf{X})} \propto P(\mathbf{X}|\mathbf{A}, \mathbf{S})P(\mathbf{A})P(\mathbf{S}). \tag{2.22}$$

If $\mathbf{A}$ is assumed to be uniformly distributed (i.e., all mixing weights are equally probable), then $P(\mathbf{A})$ will not have an influence on the optimization, and thus the problem reduces to

$$\max_{\mathbf{A},\mathbf{S}} P(\mathbf{A}, \mathbf{S}|\mathbf{X}) \propto \max_{\mathbf{A},\mathbf{S}} P(\mathbf{X}|\mathbf{A}, \mathbf{S})P(\mathbf{S}). \tag{2.23}$$

The sources are assumed to have a probability density function $p_n(s_n(t))$ of the exponential form

$$p_n(s_n(t)) = \frac{1}{Z} e^{-f(s_n(t))}, \tag{2.24}$$

where Z is a normalization factor that forces the density function sum to unity. The function f is used to control the shape of the distribution.

If it is assumed now that the sources are statistically independent and their samples also are, the joint prior $P(\mathbf{S})$ is factorial:

$$P(\mathbf{S}) = \prod_{n,t} p_n(s_n(t)), \tag{2.25}$$

[2] The Frobenius norm of a matrix $\mathbf{X}$ is given by $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} |x_{ij}|^2}$.

To compute the $P(\mathbf{X}|\mathbf{A},\mathbf{S})$ it is useful to consider the noisy model described in Section 2.3.4. The probability of observing the mixture matrix $\mathbf{X}$ given $\mathbf{A}$ and $\mathbf{S}$ would only rely on the noise distribution, since noise is the only element that adds uncertainty. Thus, assuming Gaussian white noise of covariance $\sigma^2\mathbf{I}$, this likelihood is given by the Gaussian distribution of the noise matrix $\mathbf{N} = \mathbf{X} - \mathbf{AS}$. If sources and samples are again considered to be statistically independent, the likelihood can be written as

$$P(\mathbf{X}|\mathbf{A},\mathbf{S}) \propto \prod_{m,t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_m(t) - (\mathbf{AS})_{mt})^2}{2\sigma^2}\right), \tag{2.26}$$

where $(\mathbf{AS})_{mt} = \sum_{n=1}^{N} a_{mn}s_n(t)$. Substituting Eqs.(2.26) and (2.25) into (2.23) and by taking the logarithm, the products become summations, and the $\exp(\cdot)$ operators and scaling terms can be discarded. This can be done since the logarithm is order-preserving and therefore does not affect the maximization. The sign is changed to obtain a minimization problem, obtaining the following MAP cost function:

$$\min_{\mathbf{A},\mathbf{S}} \left\{ \frac{1}{2\sigma^2} \|\mathbf{X} - \mathbf{AS}\|_F^2 + \sum_{n,t} f(s_n(t)) \right\}. \tag{2.27}$$

The MAP formulation of the BSS problem has a similar form when signal decompositions are considered. As will be seen in Section 2.5.4, the function f has considerable importance, since assuming different types of functions leads to several demixing approaches.

This joint MAP formulation to BSS is very general, thus it can be used for many types of separation problems [35][36]. However, it is computationally demanding and convergence is not always guaranteed [34]. That is the reason why, rather than using a joint optimization approach where the sources and the mixing matrix are estimated at the same time, most separation methods first estimate the mixing matrix and afterwards estimate the sources. These *staged approaches* support more freedom in their design, since different methods for mixing matrix estimation and recovery of the sources can be combined, leading to more efficient separation methods.

Estimation of the Mixing Matrix

If the M rows in the mixture matrix $\mathbf{X}$ are considered the components of an M-dimensional random vector, their empirical joint distribution can be represented by means of a *scatter plot*. Each point in the scatter plot lies in a position related to the value of that particular signal sample between the mixture channels. Denoting the columns of the mixing matrix by $\mathbf{a}_n = [a_{1n}, a_{2n}, \ldots, a_{Mn}]^{\mathrm{T}}$, the instantaneous model $\mathbf{x} = \mathbf{As}$ can be rewritten as

$$\mathbf{x} = \sum_{n=1}^{N} \mathbf{a}_n s_n(t). \tag{2.28}$$

If a mixture sample is the result of the contribution of only one source, i.e. $s_n(t) \neq 0$ and $s_{n'}(t) = 0$ for all $n' \neq n$, the point $\mathbf{x}$ will follow the direction defined by the column of the mixing matrix $\mathbf{a}_n$ corresponding to the only active source that has generated the observation. Consider, for example, the following instantaneous mixture with only two sources:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}. \tag{2.29}$$

If only one of the sources is active, for example s_1, the resultant mixture will be

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = s_1(t) \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}, \tag{2.30}$$

and so, the points of the scatter plot of $x_2(t)$ versus $x_1(t)$ would lie on a line crossing the origin with the direction given by the vector $[a_{11}\ a_{21}]^{\mathrm{T}}$, showing a very clear structure imposed by the linear mixing. Therefore, if the sources are active in few points, i.e. they are sparse, there will be low probability for more than one source being active at the same time and the scatter plot will constitute a mixture of lines following the directions given by $\mathbf{a}_n$. For convenience, unit-length mixing directions are always assumed when estimating the mixing matrix: $\|\mathbf{a}_n\| = 1$. From a geometrical point of view, the goal is therefore to estimate these line orientations (also known as *basis vectors*) from the observed data. Thus, the sparser the signals, the more their coefficients will be concentrated around the mixing directions and the easier will be the detection of line orientations.

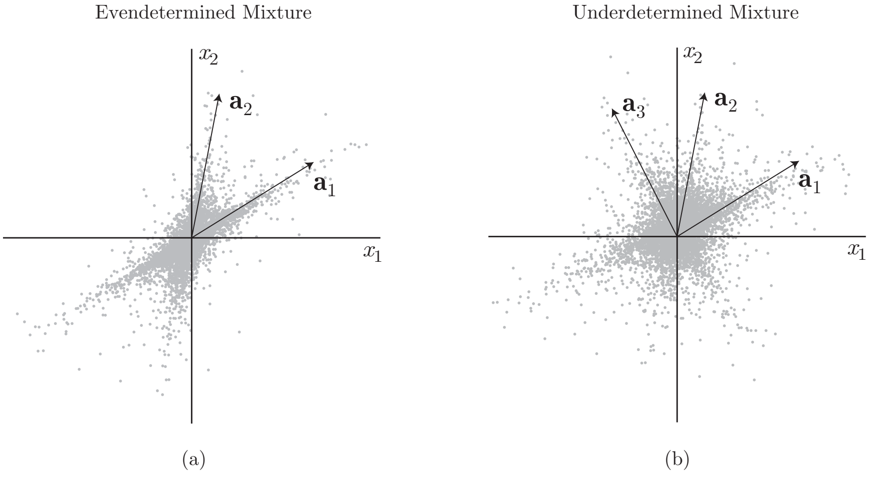


Figure 2.4. Scatter plot for two different mixtures. (a) Determined mixture with $N = 2$ sources. (b) Underdetermined mixture with $N = 3$ sources.

As an example, Figure 2.4 shows two scatter plots of a determined and an underdetermined mixture of independent signals following a sparse distribution. The corresponding mixing directions have also been included, showing how the mixture points tend to cluster along these vectors. However, it can be observed that with the same sparsity, an increment of the number of sources decreases the clustering effect, making more difficult the estimation of these directions. There are many algorithms to estimate the mixing matrix that rely on source sparsity. For example, Bofill and Zibulevsky [37] proposed a method for the estimation of the mixing matrix in stereo instantaneous mixtures using a potential function over the angles formed by the mixture data points.

In this section, the estimation of the mixing matrix has been discussed for the instantaneous BSS problem. However, when the mixtures are anechoic or convolutive, the problem also involves the estimation of delay terms and/or mixing filters. For example, the DUET algorithm proposed by Yilmaz and Rickard [38], which is described in Section 2.6.1, considers the estimation of the mixing matrix in the anechoic case for underdetermined mixtures.

Estimation of the Sources

As previously explained, when the mixing matrix $\mathbf{A}$ is square (determined problem) and has full rank, the sources can be directly obtained in the instantaneous case by

$$\hat{\mathbf{S}} = \hat{\mathbf{A}}^{-1}\mathbf{X} = \hat{\mathbf{W}}\mathbf{X}, \tag{2.31}$$

and the source separation problem is reduced to the estimation of the mixing matrix and its inversion ($\hat{\mathbf{A}}^{-1} = \hat{\mathbf{W}}$). ICA algorithms address the problem by assuming that the sources are statistically independent and have a non-Gaussian distribution [14]. Using ICA-based methods, the sources can be estimated up to a permutation and scaling ambiguity (the energies and order of the sources cannot be determined). When the problem is overdetermined ($M > N$) it is possible to apply dimensionality reduction techniques such as PCA for reducing the problem to a determined one [39].

In the underdetermined case, $\mathbf{A}$ is rectangular and thus non-invertible, which means that the sources can not be extracted even if the mixing matrix has been estimated without error. Under this situation, stronger assumptions than independence are required to find a solution to the problem, which are based on the sparse structure of audio sources under some signal transformation. Section 2.5.5, an overview of common approaches to source estimation for the solution of underdetermined BSS problems will be presented in detail, showing the relation existent with the already described MAP formulation.

2.5 Underdetermined Source Separation

The underdetermined (or *degenerate*) case in SSS is the most challenging one. The challenge resides in the fact that the mixing matrix is not invertible and the traditional method of demixing by estimating the inverse mixing matrix can not be applied in this case. Unfortunately, most commercial music productions and audio material can be categorized as underdetermined mixtures and perfect separation of instruments or singers from a stereo track is not a solved problem so far.

Sparsity refers to the property by which most of the sample values of a signal are zero or close to zero. This property is the fundamental piece supporting most underdetermined separation algorithms, including all the techniques presented throughout this thesis. In Section 2.4.2 it was shown how if the sources are sparse, the mixing directions of a linear instantaneous mixture can be easily observed in the scatter plot. Moreover, in the underdetermined case, higher sparsity is a requirement for good separability of the sources, even in the case when the mixing matrix is known. Thus, an increasingly popular and powerful assumption that has led to many practical algorithms is to assume that the sources have a sparse representation under a given basis. These approaches are motivated by the fact that very often the desired data in the time domain do not represent the required sparsity. Therefore, they have come to be known as *sparse methods.*

The advantage of a sparse signal representation is that the probability of two or more sources being simultaneously active is low. Thus, sparse representations are potentially good for achieving high-quality separation due to the fact that most of the energy in a basis coefficient belongs to a single source. The sparse representation of an audio signal has an interpretation in information theoretic terms: a signal represented by a small number of coefficients corresponds to transmission of information using a code with a small number of bits [40]. Sparse representation of information is a phenomenon that also occurs in the natural world. In the brain, neurons are said to encode data in a sparse way, if their firing pattern is characterized by long periods of inactivity [41].

Throughout this section the general framework for underdetermined source separation will be presented. Most approaches rely on signal transformations that enhance the sparse structure of the sources. Therefore, special attention is paid to the basics of signal decomposition, mainly to time-frequency representations. Sparse distributions and sparsity measures are also introduced, followed by a description of the most common approaches to source estimation.

2.5.1 Additive Expansions

Generally, a time-domain signal $s(t)$ can be expressed as an *additive expansion* or *decomposition* by means of a weighted sum of expansion functions:

$$s(t) = \sum_{k=1}^{K} c_k b_k(t), \tag{2.32}$$

where K denotes the number of expansion functions, c_k are the expansion coefficients and $b_k(t)$ are the time-domain expansion functions. In general terms, the application context usually determines the choice of the decomposition functions. However, this choice is specially relevant in underdetermined SSS, since most approaches are based on a sparse decomposition of the audio signals. Therefore, the aim is to find a decomposition that allows to represent the signals with most of the expansion coefficients equal or close to zero. Spectral transforms such as the *Discrete Fourier Transform* (DFT) or the *Discrete Cosine Transform* (DCT) are additive expansions with a finite set of frequency-localized fixed expansion functions. On the other hand, time-frequency representations are decompositions which are localized both in time and frequency. The most used time-frequency representations are the *Short Time Fourier Transform* (STFT) and the *Discrete Wavelet Transform* (DWT). Time-frequency representations have been shown to be a powerful approach for achieving sparsity and have been widely used in underdetermined BSS. The sparsity achieved by several types of frequency-warped representations was deeply covered in Juan José Burred's thesis [23].

Adaptive or *data-driven expansions* are those in which the set of expansion functions is not fixed. For example, the basis expansion functions can be selected out of a signal *dictionary* so that only the ones that best match the observed signal take part in the decomposition. Overcomplete and sparse decomposition methods such as *Basis Pursuit* [42] and *Matching Pursuit* [43] work under this principle. Adaptive signal decompositions using PCA and ICA have also been proposed in the literature in order to extract the expansion functions directly from the input signals [44].

In the next subsections, fixed (frequency and time-frequency localized) decompositions are described in detail, especially the STFT, since this transformation constitutes the basic frontend used throughout this thesis.

2.5.2 Basis decompositions

Considering a finite-length interval $t = 0, \ldots, T-1$, and using the vector notation $\mathbf{s} = [s(0), \ldots, s(T-1)]^{\mathrm{T}}$ for the signal and $\mathbf{c} = [c_1, \ldots, c_K]^{\mathrm{T}}$ for the coefficients corresponding to the K expansion functions $b_k(t) = \mathbf{b}_k = [b_k(0), \ldots, b_k(T-1)]^{\mathrm{T}}$, it is possible to express Eq.(2.32) in matrix notation:

$$\mathbf{s} = \mathbf{B}\mathbf{c}, \tag{2.33}$$

or equivalently

$$\mathbf{s} = \sum_{k=1}^{K} c_k \mathbf{b}_k, \tag{2.34}$$

where $\mathbf{B}$ is a $T \times K$ matrix whose columns are the functions $\mathbf{b}_k$.

Equation (2.33) reveals that the decomposition is a linear transformation from the coefficient space to the signal space, being $\mathbf{B}$ the transformation matrix and $\mathbf{b}_k$ the transformation bases. Note that such a linear decomposition model is of the same form as the linear mixing model of Eq.(2.8). In fact, there is a strong analogy between source separation and signal decomposition [23].

When the number of expansion functions equals the number of signal samples ($T = K$), and the columns of $\mathbf{B}$ are linearly independent, then the set of expansion functions constitutes a *basis* of the signal space, meaning that each signal vector $\mathbf{s}$ can be represented as a unique linear combination of the functions $\mathbf{b}_k$, which are then called *basis functions*. In this case, the basis decomposition is said to be *complete*. However, when $T < K$, the matrix $\mathbf{B}$ contains linearly dependent vectors and the representation is said to be *overcomplete*.

In the complete case, the transformation matrix is invertible, and the expansion coefficients can be readily obtained by

$$\mathbf{c} = \mathbf{B}^{-1}\mathbf{s}. \tag{2.35}$$

In the context of signal transformation, Eq.(2.35) is called the *analysis equation* and Eq.(2.33) is called the *synthesis equation*. By convention, the analysis equation is considered the direct transformation and the synthesis equation is considered the inverse transformation.

When considering the transformation of multiple signals using explicit notation, the $N \times T$ matrix $\mathbf{S}$ contains N signals of length T in its rows. Thus, the formulation of basis decomposition becomes

$$\mathbf{S} = \mathbf{C}\mathbf{B}^{\mathrm{T}}, \tag{2.36}$$

with the the coefficient vectors arranged as the rows of the matrix $\mathbf{C}$ (size $N \times K$), respectively.

Orthogonal Basis

A further simplification is possible if the basis are orthogonal[3]. In this case, the coefficients are given by

$$\mathbf{c} = \mathbf{B}^{\mathrm{H}}\mathbf{s}, \tag{2.37}$$

where $^{\mathrm{H}}$ denotes the Hermitian (complex conjugate) transpose. Strictly speaking, such bases are referred to as *orthonormal bases*; however, since most applications involve unit-norm basis functions, there has been a growing tendency in the literature to use the terms orthogonal and orthonormal interchangeably [44].

Each coefficient is directly given by projecting the signal upon each one of the basis functions:

$$c_k = \langle \mathbf{b}_k, \mathbf{s} \rangle = \mathbf{b}_k^H \mathbf{s} = \sum_{t=0}^{T-1} b_k^*(t) s(t), \tag{2.38}$$

[3] Orthogonality implies $\langle \mathbf{b}_i, \mathbf{b}_j \rangle = \delta(i-j)$ or, in matrix form, $\mathbf{B}^{-1} = \mathbf{B}^{\mathrm{H}}$

where $\langle\cdot\rangle$ denotes the scalar (or dot) product and * denotes conjugation. Substituting the above equation in (2.34), this results in an expansion of the form

$$\mathbf{s} = \sum_{k=1}^{K} \langle \mathbf{b}_k, \mathbf{s} \rangle \mathbf{b}_k, \tag{2.39}$$

which is called the *orthogonal projection* of $\mathbf{s}$ onto the set of bases $\mathbf{b}_k$.

The Discrete Fourier Transform (DFT)

The previously described framework is useful for defining common signal transformations that are widely used in signal processing. The most popular orthogonal transformation with invariant bases is the DFT. The basis functions used by this transformation are complex exponentials of the form $b_k(t) = e^{j\frac{2\pi}{T}kt}$. The analysis equation (2.38) then yields the following complex coefficients (the usual notation for the DFT is $c_k = S(k)$):

$$S(k) = \sum_{k=0}^{T-1} s(t) e^{-j\frac{2\pi}{T}kt}, \quad k = 0, \dots, T-1. \tag{2.40}$$

The quantities $|S(k)|$ and $\angle S(k)$ constitute the magnitude and phase spectrum of the signal, respectively. Note that the number of basis functions is equal to the number of samples of the signal, thus it is a complete basis decomposition. The *Fast Fourier Transform* (FFT) algorithm can efficiently calculate the DFT [45].

The *Inverse Discrete Fourier Transform* (IDFT) is given by

$$s(t) = \frac{1}{T} \sum_{k=0}^{T-1} S(k) e^{j\frac{2\pi}{T}kt}, \quad t = 0, \dots, T-1. \tag{2.41}$$

Note that the normalization factor multiplying the DFT and IDFT (here 1 and $1/T$) and the signs of the exponents are mereley conventions, and differ in some treatments. The only requirements of these conventions are that the DFT and IDFT have opposite-sign exponents and that the product of their normalization factors be $1/T$. A normalization of $1/\sqrt{T}$ for both the DFT and IDFT makes the transforms unitary, which has some theoretical advantages, but it is often more practical in numerical computation to perform the scaling all at once as above.

2.5.3 The Short-Time Fourier Transform (STFT)

Spectral transforms such as the DFT are *frequency localized*, which means that the basis functions have a definite position in frequency. For example, the *frequency support* of the DFT, i.e., the set of positions in frequency of the basis functions, is given by $f_k = \frac{k}{T} f_s$, where $k = 1, \dots, K$ and f_s is the sampling frequency. However, frequency localized decompositions do not provide accurate temporal information and thus they are not very useful to manage real-world non-stationary signals. This is due to the fact that the *time support* for all the basis functions is equal to the time support of the signal to be analyzed, i.e. $t = 0, \dots, T-1$. If the analyzed signal is highly non-stationary (which is the case of speech and music signals), a certain time granularity is required to obtain a useful representation. This is especially important to fulfill the sparsity requirements of underdetermined source separation: higher time localization leads to higher time resolution and, thus, to higher temporal sparsity, since each meaningful temporal component of the signal will be represented by only one or few coefficients. The same reasoning applies for the frequency localization.

A general time-frequency decomposition is a generalization of the basic additive model of Eq.(2.33) with a set of expansion functions both localized in time (index r) and frequency (index k):

$$s(t) = \sum_{r=1}^{R}\sum_{k=1}^{K} c_{kr} b_{kr}(t). \tag{2.42}$$

The *Short-Time Fourier Transform* (STFT) is the most well known time-frequency decomposition. It has proven useful for many speech processing and speech communication applications, including time scaling, pitch shifting, noise reduction, and echo cancellation [46]. In practical applications, the STFT is typically implemented in a sliding-window fashion, as will be next described.

Before starting to describe the insights of the STFT, it is worth to clarify the different notations that appear in the following sections. As in the case of the DFT, there is a conventional notation for STFT transformed signals: $S(k,r) = c_{kr}$. Therefore, in discussions where keeping the two-dimensional time-frequency meaning is important, such as in time-frequency masking (Section 2.5.6) a signal will be denoted with element-wise notation with explicit indexing $S(k,r)$. The whole time-frequency matrix will be denoted as $\mathbf{S}(k,r)$, keeping the indices (k,r) in order to avoid confusion with the multi-source matrices of the mixing model $\mathbf{X} = \mathbf{AS}$ (with time-domain signals as the rows). When the indexing of the coefficients is not necessary, they will be assumed to be *lexicographically ordered* and arranged in a $C \times 1$ vector $\mathbf{c}$, being $C = KR$ the total number of coefficients. In a multi-signal context, the coefficients of each signal $\mathbf{c}_n$ will be arranged as the rows of the coefficient matrix $\mathbf{C}$ (with elements denoted as c_{nk}).

STFT analysis

Given an input signal $s(t)$ of arbitrary duration, data segments are extracted at regular intervals using a time-limited window $w(l)$; these signal segments or frames are expressed as

$$s_r(l) = w(l)s(l + rH), \quad 0 \leq l \leq L - 1, \tag{2.43}$$

where L is the window length, r is a frame index, and H is the hop size, i.e., the spacing in samples between consecutive applications of the sliding window; the new index l is a local time index, which is relative to the start of the sliding window. For each frame, a K point DFT is carried out, which yields for each frame r

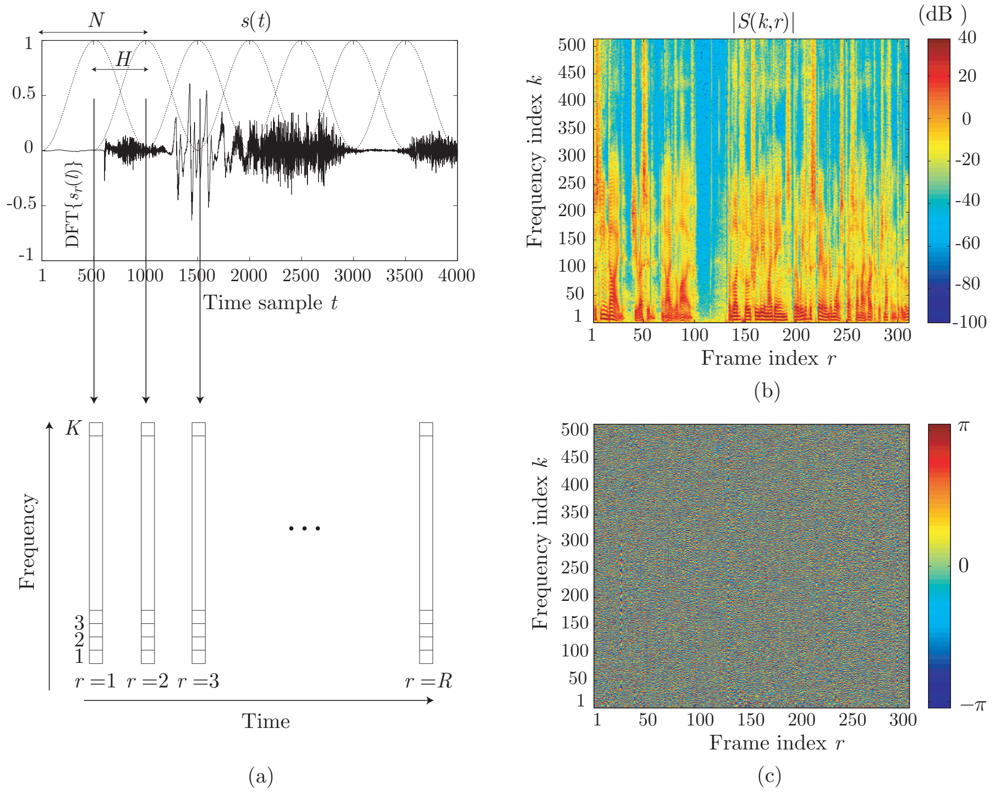
$$S(k,r) = \sum_{l=0}^{L-1} s_r(l) e^{-j\frac{2\pi k}{K}l} = \sum_{l=0}^{L-1} w(l)s(l + rH)e^{-j\omega_k l} \tag{2.44}$$

where k is the frequency index. Therefore, the frequency support of the STFT is $f_k = \frac{k}{K} f_s$ and the time support is $t_r = \frac{r}{f_s} H$.

Figure 2.5(a) shows an scheme for STFT synthesis. In Figure 2.5(b) and 2.5(c) are represented the STFT magnitude and phase of a male speech signal, respectively.

STFT synthesis

The STFT is usually modified so as to create a desired effect in a new time-domain signal. To generate the modified time-domain signal is necessary a *synthesis* operation, which ideally



(b)

(a)

(c)

Figure 2.5. STFT analysis. (a) A time signal is windowed in segments of length N with a hope size of H samples, obtaining a K point DFT for each segment. (b) Magnitude STFT. (c) Phase STFT.

reconstructs the original signal perfectly in the case that the original STFT has not been altered. This property is known as *perfect reconstruction*. The reconstruction framework is basically the opposite of the analysis: first, an inverse DFT (IDFT) of each local spectrum is carried out; then, the signals are added to synthesize the signal. If the DFT is large enough ($K \geq L$) the IDFT simply returns the windowed signal segment:

$$\hat{s}_r(l) = w(l)s(l + rH), \quad 0 \leq l \leq L - 1. \tag{2.45}$$

If the IDFT of each segment is expressed with respect to the global time $t = l + rH$, then Equation (2.45) becomes

$$\hat{s}_r(t - rH) = w(t - rH)s(t), \quad rH \leq t \leq rH + L - 1. \tag{2.46}$$

Note that if $K < L$, time-domain aliasing is introduced, to the condition $K \geq L$ is imposed. Therefore, reconstruction can be simply carried out by an *overlap-add* process, possibly with a synthesis window $v(l)$:

$$\hat{s}(t) = \sum_r v(t - rH)\hat{s}_l(t - rH) = \sum_r v(t - rH)w(t - rH)s(t). \tag{2.47}$$

Since $s(t)$ is not a function of r, Equation (2.47) can be rewritten as

$$\hat{s}(t) = s(t)\left(\sum_r w(t - rH)v(t - rH)\right), \tag{2.48}$$

so perfect reconstruction is achieved if the the analysis and synthesis windows satisfy

$$\sum_r w(t - rH)v(t - rH) = 1. \tag{2.49}$$

When a synthesis window $v(l)$ is not explicitly specified, the equivalent synthesis window is a rectangular window leading to the following perfect reconstruction constraint:

$$\sum_r w(t - rH) = 1. \tag{2.50}$$

Perfect reconstruction windows have been proposed in the literature, for example rectangular and triangular windows, or Hamming and Hanning windows [47][48]. In practice, the STFT analysis window is generally chosen based on its frequency resolution and sidelobe behavior.

One of the downfalls of the STFT is that it has a fixed resolution. The width of the windowing function determines how the signal is represented. Intuitively, long time windows are needed in order to resolve low frequency components, with longer periods. Inversely, short time windows offering better time resolution can only resolve frequency components whose periods are shorter than the time interval they span. Therefore, a wider window gives better frequency resolution but poor time resolution. A narrower window (said to be compactly supported) gives good time resolution but poor frequency resolution.

2.5.4 Sparsity

The reason why high sparsity of source representations is desired in source separation problems is straightforward: the less coefficients are needed to adequately describe a particular source signal, the less degree of overlapping will occur when mixed with other signals. Sparsity is crucial in most underdetermined situations, specially when very little a priori information is available and the ratio between number of sources and number of mixtures is high.

In Section 2.4.2 it was already shown how the probability distribution of the sources had an influence on the MAP formulation to BSS. The sources were assumed to have an exponential probability density function (see Eq. 2.24) whose shape is dependent on the function f. In fact, if the sources are decomposed by $\mathbf{S} = \mathbf{C}\mathbf{B}^{\mathrm{T}}$ (as in Equation 2.36), the MAP formulation becomes

$$\min_{\mathbf{A},\mathbf{C}} \left\{ \frac{1}{2\sigma^2} \|\mathbf{X} - \mathbf{A}\mathbf{C}\mathbf{B}^{\mathrm{T}}\|_F^2 + \sum_{n,k} f(c_{nk}) \right\}, \tag{2.51}$$

and the problem is now directly related to the shape of the probability density function of the coefficients of the sources in the transformed domain. These are also assumed to have an exponential distribution $p(c) = \frac{1}{Z} e^{-f(c)}$, which can represent different degrees of sparsity depending on the chosen function $f(c)$ (considering c the random variable). For example, a function of the form $f(c) = |c|^\nu$ can result in a Gaussian distribution for $\nu = 2$ (in the usual form, with mean and scaling terms):

$$p(c) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|c-\mu|^2}{2\sigma^2}}. \tag{2.52}$$

For $\nu = 1$ a Laplacian distribution is obtained:

$$p(c) = \frac{1}{2b} e^{-\frac{|c|}{b}}, \tag{2.53}$$

where the variance is given by $2b^2$.

Note that the term $f(c_{nk})$ can be interpreted as a cost function that penalizes "active" (non-zero) coefficients. For example, if a Laplacian distribution is considered, values of $\mathbf{C}$ near zero will be given a smaller cost and a higher probability. This is illustrated in Figure 2.6, showing the cost function and the corresponding Laplacian prior distribution.

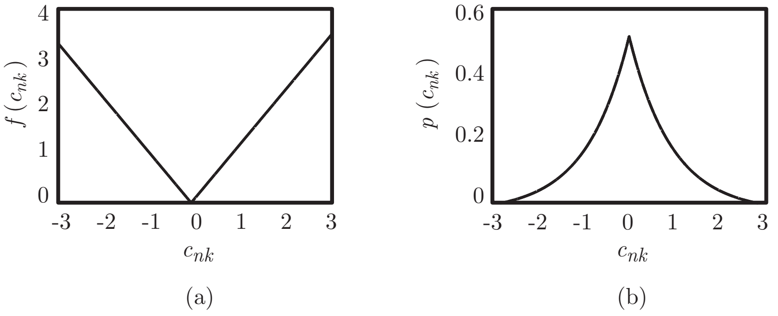


Figure 2.6. (a) The cost function $f(c_{nk})$. (b) Corresponding laplacian distribution.

From Equation (2.27) and the above definitions of $f(c)$, it can be seen that a sparse representation can be obtained by minimizing a cost function which is the weighted sum of the reconstruction error term $\|\mathbf{X} - \mathbf{ACB}^{\mathrm{T}}\|_F^2$ and the term which incurs a penalty on non-zero elements of $\mathbf{C}$. The variance σ^2 is used to balance between these two.

Measures of sparsity

The sparsity ξ of a signal is usually measured by means of the ℓ_p norm of its coefficient vector $\mathbf{c}$ with the constraint $0 \leq p \leq 1$:

$$\xi = \|\mathbf{c}\|_p = \left(\sum_{i=1}^{C} |c_i|^p\right)^{1/p}, \quad 0 \leq p \leq 1. \tag{2.54}$$

Depending on the value of p, several well-known sparsity measures appear:

- **The ℓ_0 norm**. This measure gives the number of non-zero coefficients in $\mathbf{c}$:

  $$\|\mathbf{c}\| = \#\{i, c_i \neq 0\}, \tag{2.55}$$

  where $\#\{\cdot\}$ denotes the counter operator. This norm is rarely used since it is highly sensible to noise: a slight addition of noise will make a representation completely nonsparse.

- **The ℓ_ϵ norm**. A thresholded version of the ℓ_0 norm in order to be more robust against noise:

  $$\|\mathbf{c}\|_\epsilon = \#\{i, |c_i| \geq \epsilon\}. \tag{2.56}$$

  However, determining a reasonable noise threshold ϵ for unknown signals is a difficult task [49].

- **The ℓ_1 norm**. This measure gives the summation of the modulus of the coefficients:

  $$\|\mathbf{c}\|_1 = \sum_{i=1}^{C} |c_i|. \tag{2.57}$$

> The ℓ_1 norm is a popular choice since some algorithms can be implemented with linear programming techniques. In Section 2.5.5 it will be seen how under a Laplacian prior and assuming that $\mathbf{A}$ has been previously estimated, the estimation of the sources becomes a minimization of the ℓ_1 norm. The ℓ_2 norm $\|\cdot\|$, for which the order index is usually omitted, corresponds to the traditional Euclidean norm, and to the square root of the energy.

These and others measures of sparsity such as the *normalized kurtosis* were analyzed by Karvanen and Cichoki [49] in the context of the BSS MAP approach, showing that very different results can be obtained by using different sparsity measures if the distribution does not have a unique mode at zero.

Overcomplete decompositions

An overcomplete signal decomposition has a redundant dictionary of expansion functions, i.e. matrix $\mathbf{B}$ in the expansion model $\mathbf{s} = \mathbf{Bc}$ is unsquare with $T < K$. This makes the representation problem no longer invertible and the general analysis equation (2.35) can not be applied. However, overcomplete dictionaries have been shown to be very useful in source separation due to the fact that, if the dictionary is composed of well designed time-frequency localized functions, a high sparse representation of audio signals can be derived. The problem with overcompleteness is that the nice orthogonality principle that grants the uniqueness of the decomposition is lost: indeed, for a given signal there is an infinity of possible decompositions. Thus, the problem is to find, amongst these decompositions, the one that is the most sparse, or that admits a sparse approximation, according to one of the definitions of sparsity. Figure 2.7 shows three time-frequency atoms for a dictionary of *Modified Discrete Cosine Transforms* (MDCT) functions.

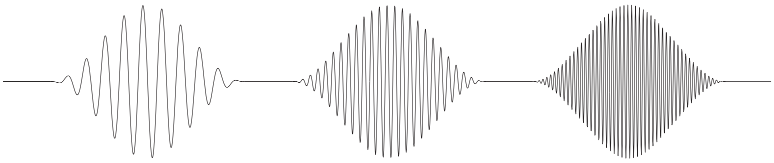


Figure 2.7. Three time-frequency atoms for a dictionary of MDCT functions.

Several algorithms with different complexities have been proposed in the literature to find such decompositions [50][51]. In the context of audio time-frequency processing the most popular is the *Matching Pursuit* method, proposed by Mallat and Zhang [43]. This is a fast suboptimal iterative algorithm. At each iteration, Matching Pursuit chooses the atom in the dictionary most correlated with the signal, subtracts it, and iterates until some stopping condition is met. This algorithm has not only been used in the source separation framework, but also in other areas such as efficient audio coding [52].

2.5.5 Estimation of the Sources

As opposed to the determined instantaneous problem ($M = N$), in the underdetermined problem ($M < N$), the estimation of the sources is not trivial, due to the fact that the mixing matrix is not invertible and thus, the equation system is ill-posed. Therefore, more sophisticated approaches are needed to find a correct solution.

1-D and M-D projection

One of the simplest ways to invert the underdetermined problem consists in the 1-D projection approach proposed by Vielva [53]. This 1-D approach is carried out in a two-step manner:

1. Given the data point $\mathbf{x}$, select the mixing direction $\hat{\mathbf{a}}_n$ closest to this point.
2. Project the data point over the selected direction to estimate the contribution to the mixture:
$$s_n = \hat{\mathbf{a}}_n^{\mathrm{T}}\mathbf{x}. \tag{2.58}$$

Note that this hard-assignment assumes that each mixture coefficient has been generated by only one of the sources, an assumption that is closely related to the ideas underlying time-frequency masking and W-Disjoint Orthogonality, which are later discussed in the text.

The M-D ($M > 1$) approach, selects M mixing directions for each data point $\mathbf{x}$ according to a given criterion, and then inverts the problem by means of a $M \times M$ square reduced mixing matrix $\hat{\mathbf{A}}_\rho$ with the selected vectors as its columns:

$$\hat{\mathbf{s}}_\rho = \hat{\mathbf{A}}_\rho^{-1}\mathbf{x}, \tag{2.59}$$

where $\hat{\mathbf{s}}_\rho$ are the estimated components of the source vector $\mathbf{s}$ corresponding to the selected mixing directions (the others are assumed to be zero). If the criterion is to select the columns that minimize the ℓ_1 or ℓ_2 norms of the projection, the M-D approach is equivalent to the methods presented in the next section.

ℓ_1 and ℓ_2 minimization

In Section 2.4.2 it was described how the general joint MAP formulation of Eq.(2.27) was quite difficult to solve without any a priori knowledge. In an staged approach, $\mathbf{A}$ has been estimated beforehand ($\hat{\mathbf{A}}$) and it is easier to deal with the problem. The sources, can then be estimated as

$$\hat{\mathbf{S}} = \arg \min_{\mathbf{X}=\hat{\mathbf{A}}\mathbf{S}} \left\{ \frac{1}{2\sigma^2}\|\mathbf{X} - \hat{\mathbf{A}}\mathbf{S}\|_F^2 + \sum_{n,t} f(s_n(t)) \right\}. \tag{2.60}$$

In the noise-free case, it is possible to omit the first term, in contrast to the joint optimization problem, in which the term was needed to include the mixing information in the minimization process. Thus, the problem further reduces to

$$\hat{\mathbf{S}} = \arg \min_{\mathbf{X}=\hat{\mathbf{A}}\mathbf{S}} \left\{ \sum_{n,t} f(s_n(t)) \right\}. \tag{2.61}$$

Note that considering a signal decomposition, the problem is the same but working in the coefficients space:

$$\hat{\mathbf{C}} = \arg \min_{\mathbf{X}=\hat{\mathbf{A}}\mathbf{C}} \left\{ \sum_{n,k} f(c_{nk}) \right\}. \tag{2.62}$$

Therefore, the goal is to minimize the contribution of the function f, which penalizes the "active" (non-zero) values of the sources. If the sources are assumed to be Gaussian, $f(x) = x^2$, the problem becomes

$$\hat{\mathbf{S}} = \arg \min_{\mathbf{X}=\hat{\mathbf{A}}\mathbf{S}} \left\{ \sum_{n,t} s_n(t)^2 \right\}, \tag{2.63}$$

which, assuming that the source samples (or coefficients) are real, equals to minimizing the Euclidean (ℓ_2) norm of the signals.

It can be shown that there is a closed solution to this problem [53], given by

$$\hat{\mathbf{S}} = \hat{\mathbf{A}}^{+}\mathbf{X}, \tag{2.64}$$

where $\hat{\mathbf{A}}^{+}$ is the Moore-Penrose pseudoinverse, computed as

$$\hat{\mathbf{A}}^{+} = \hat{\mathbf{A}}^{\mathrm{T}} \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{T}} \right)^{-1}. \tag{2.65}$$

However, if the sources are sparse, which is a more realistic assumption when working with audio sources in the time-frequency domain, the pseudoinverse solution is not optimal. If the sources are assumed to be Laplacian, i.e. $f(x) = |x|$, the formulation becomes

$$\hat{\mathbf{S}} = \arg \min_{\mathbf{X}=\hat{\mathbf{A}}\mathbf{S}} \left\{ \sum_{n,t} |s_n(t)| \right\}, \tag{2.66}$$

and the solution is reached by the minimization of the ℓ_1 norm, which can be solved by linear programming techniques. The *shortest path* algorithm prosposed by Bofill and Zibulevsky [37] provides an intuitive way to understand this minimization problem from a geometrical point of view.

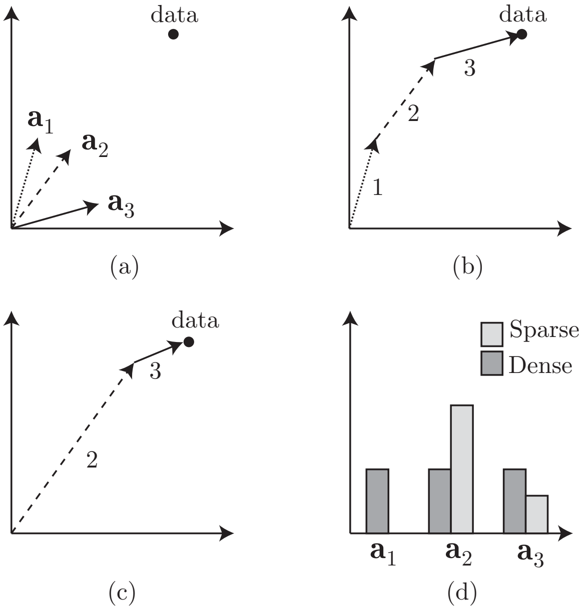


Figure 2.8. Representation of a sparse recovery problem. (a) Data point and basis vectors. (b) Dense solution obtained by ℓ_2 minimization. (c) Sparse solution obtained by ℓ_1 minimization. (d) Decomposition obtained for both solutions.

An intuitive idea of sparse recovery is shown in Figure 2.8 for a two-dimensional case, reproduced from [54]. An observed data point can be obtained by multiple combinations of the three features shown in Figure 2.8(a) , so there are many possible ways to represent a data point with no error. In Figure 2.8(b) is depicted the conventional solution given by the pseudoinverse, which yields a dense representation because it minimizes the squared sum of the coefficient activity $\sum_{i=1}^{C} c_i^2$. This representation invokes all features about evenly. It can be observed in Figure 2.8(c) that the sparse solution invokes at most $N = 2$ features because it minimizes $\sum_{i=1}^{C} |c_i|$. The weights obtained for each feature under the two solutions are shown in 2.8(d).

The last source estimation technique is time-frequency masking, which is more extensively covered in the next section. This technique is the one used in most of the separation methods described in this thesis, including those presented as novel contributions in Chapter 3.

2.5.6 Time-Frequency Masking

Although the sparsity achieved by signal transformations provide a way to deal with underdetermined BSS, the factor that ultimately determine the separation performance is the degree of overlapping that occurs during the mixing process. Sparsity is not the only thing to consider since sparsity alone is useless if there is high overlap among the sources in the mixture. Two sources that are closely positioned in the stereo panoramic will be very hard to separate even if they are sufficiently sparse. Moreover, the correlation properties of the sources also play a role in the degree of overlap of the mixture. The *disjointness* of a mixture can be defined as the degree of non-overlapping of the mixed signals.

Time-frequency masking is another powerful approach for the separation of underdetermined mixtures, especially for the separation of single-channel mixtures. Techniques based on time-frequency masking use a time-frequency representation of the signal, taking profit from the disjointness provided by sparse transformations, such as the STFT. Their aim is to identify the dominating source in each time-frequency unit, obtaining a mask that indicates which are the active points of each source in the time-frequency domain. Most common techniques are usually based either on a CASA approach (using perceptual principles of auditory scene analysis) or on BSS methods.

Formally, the time-frequency source image $\hat{\mathbf{S}}_{mn}(k, r)$ is produced from the m-th mixture $\mathbf{X}_m(k, r)$ by

$$\hat{\mathbf{S}}_{mn}(k, r) = \mathbf{M}_n(k, r) \circ \mathbf{X}_m(k, r), \tag{2.67}$$

where $0 \leq M_n(k, r) \leq 1$, $\forall (k, r)$ and the $\circ$ operator denotes the Hadamard (element-wise) product. Note that this corresponds to filtering the mixture with a set of time-varying frequency responses. The solution to the separation problem consists in deriving the masks from the mixture.

The Ideal Binary Mask

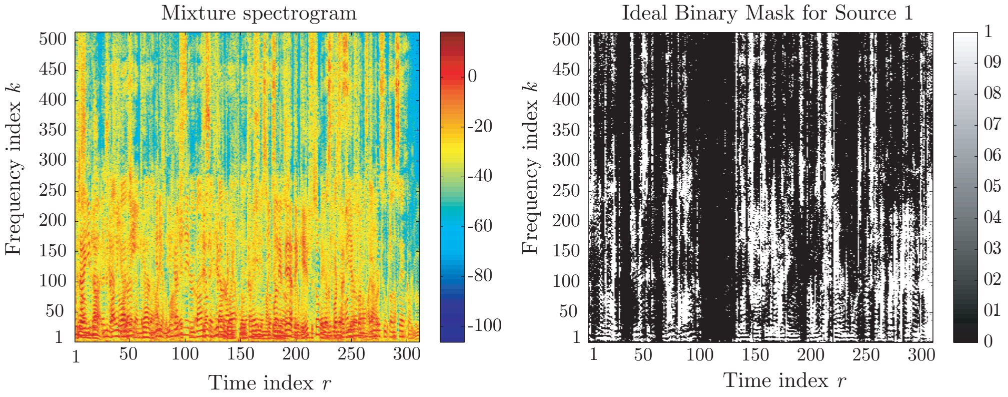
Consider the sum of all signals that interfere with source n in the STFT domain:

$$\mathbf{U}_n(k, r) = \sum_{n'=1, n' \neq n}^{N} \mathbf{S}_{n'}(k, r). \tag{2.68}$$

The *ideal binary mask* (IBM) for a source $\mathbf{S}_n(k, r)$ is defined as the binary time-frequency mask that is 1 for time-frequency bins where its energy is higher than all the interfering sources:

$$\mathrm{IBM}_n(k,r) = \begin{cases} 1 & \text{if} \quad 20\log\left(\frac{|S_n(k,r)|}{|U_n(k,r)|}\right) \geq 0 \\ 0 & \text{elsewhere} \end{cases}, \quad \forall (k,r). \tag{2.69}$$

This mask has been shown to be optimal when applied to the mixture and this is why the IBM has been suggested as a major computation goal of sound source separation algorithms, specially in the CASA community, since it has proven to be highly effective for robust *automatic speech recognition* and human speech intelligibility in noise [55]. An example of ideal binary mask for a 4 source mixture is shown in Figure 2.9.



(a) (b)

Figure 2.9. Example of ideal binary mask for an instantaneous mixture of 4 sources. (a) Magnitude STFT of one of the mixture channels. (b) Ideal binary mask for one of the sources.

The properties of the IBM are further explored in Chapter 4, where a source-reassignment technique aimed at improving the isolation of the separated sources is proposed.

W-Disjoint Orthogonality

Binary time-frequency masking is the special case in which $M_n(k,r)$ can only take the values 0 or 1. It is based on the assumption that every time-frequency point in the mixture with significant energy is dominated by the contribution of one source. This assumption is widely known as the *W-Disjoint Orthogonality* (WDO) assumption. Mathematically, the sources are said to be WDO if

$$S_n(k,r)S_{n'}(k,r) = 0, \quad \forall n \neq n', \forall (k,r), \tag{2.70}$$

where $S_n(k,r)$ and $S_{n'}(k,r)$ are the STFT of any two sources in the mixture. In matrix notation:

$$\mathbf{S}_n(k,r) \circ \mathbf{S}_{n'}(k,r) = \mathbf{0} \quad \forall n \neq n', \tag{2.71}$$

where $\mathbf{0}$ is the zero matrix.

Based on the IBM, a disjointness measure is given by the *average approximate* WDO. Burred provided an excellent analysis of the disjointness properties of speech and music mixtures considering both STFT and frequency-warped representations, showing the advantages of using a non-uniform time-frequency resolution [23][56].

Yilmaz and Rickard [38] applied binary time-frequency masks to mixtures of several speech sources in the STFT domain considering a two-sensor arrangement. They observed that speech sources are sufficiently disjoint under time-frequency representations and showed that they are approximately WDO in mixtures of up to 10 signals [57]. Note that this aspect is very important, since source sparsity alone is useless if the sources overlap to a high degree.

Time-Frequency Masking Limitations

In practice, perfect separation is very difficult to be achieved and the estimated source spatial images may contain different distortions: musical noise, interference from other sources, timbre distortion and spatial distortion. Musical noise or burbling artifacts appear with time-frequency masking algorithms, being one of the most common distortions in source separation. Musical noise can be reduced by using small STFT hop sizes [58] and non-binary time-frequency masks. To this end, Araki et al. proposed a set of smoothed masks in [59], showing the tradeoff between source distortion and interference. Nevertheless, the performance achieved by time-frequency masking is sufficient for most practical applications of SSS [60].

2.6 Algorithms for Stereo Mixtures

In this section we review the algorithms that are evaluated in Chapter 6 in the context of WFS scene resynthesis. These algorithms are designed to work with underdetermined stereo mixtures.

2.6.1 DUET

One of the most popular algorithms for stereo underdetermined mixtures is the *Degenerate Underdetermined Estimation Technique*, widely known as the DUET algorithm [38][61], which appeared as the first practical approach for the separation of anechoic mixtures. Each observation channel is transformed into the STFT domain, and the relative attenuation and delay values between two observations are calculated from the ratio of the corresponding time-frequency points R_{21}, written as

$$R_{21}(k,r) = \frac{X_2(k,r)}{X_1(k,r)}, \quad \forall (k,r). \tag{2.72}$$

The symmetric attenuation estimate is defined as

$$\hat{\alpha}(k,r) = \hat{a}(k,r) - \frac{1}{\hat{a}(k,r)}, \quad \forall (k,r), \tag{2.73}$$

where $\hat{a}(k,r) = |R_{21}(k,r)|$ and the instantaneous delay estimate is defined as

$$\hat{\delta}(k,r) = -\frac{1}{\omega_k} \angle R_{21}(k,r), \quad \forall (k,r), \tag{2.74}$$

where ω_k is the angular frequency associated to frequency index k. The relative mixing parameters, $\hat{a}_n$ and $\hat{\delta}_n$, for each source in the mixture are estimated from peaks in a 2D power weighted histogram constructed from $\hat{\alpha}(k,r)$ and $\hat{\delta}(k,r)$. Figure 2.10 shows the histogram obtained for the example mixture of Figure 2.9, where the 4 sources can be clearly observed as prominent peaks that give an estimation of the mixing parameters. A set of separation masks $\mathbf{M}_n(k,r)$ is obtained assigning each time-frequency point to the closest mixing parameter estimate by the likelihood function:

$$\mathcal{L}_n(k,r) = \frac{\left| \hat{a}_n e^{-j\hat{\delta}_n 2\pi f_k} X_1(k,r) - X_2(k,r) \right|^2}{1 + \hat{a}_n^2}, \tag{2.75}$$

and

$$M_n(k,r) = \begin{cases} 1 & n = \underset{n'}{\operatorname{argmin}} \ \mathcal{L}_{n'}(k,r) \\ 0 & \text{otherwise} \end{cases} \qquad \forall(k,r) \tag{2.76}$$

The separation of the sources is carried out using time-frequency masking and a *Maximum Likelihood* criterion:

$$\hat{S}_n(k,r) = M_n(k,r)\left(\frac{X_1(k,r) + \hat{a}_{n'}e^{j\hat{\delta}_{n'}\omega}X_2(k,r)}{1+\hat{a}_{n'}^2}\right) \quad \forall(k,r). \tag{2.77}$$

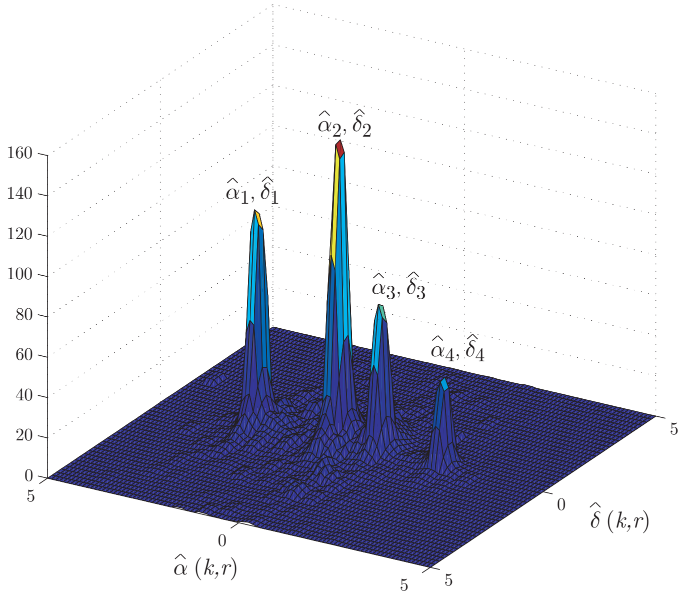


Figure 2.10. 2D Power weighted histogram for estimating the separation masks. As the mixture is instantaneous, all the peaks are aligned in the value $\hat{\delta} = 0$.

It is important to remark that a phase ambiguity problem may appear in the upper frequency range: the inter-microphone spacing d must satisfy the condition $d < \frac{c}{2f_{\max}}$ to avoid the spatial aliasing problem, being $f_{\max}$ the maximum frequency of the source signals. Therefore, it is guaranteed that there is less than one wavelength difference between two sources with a *direction-of-arrival* of $+\pi/2$ and $-\pi/2$ from the equal-delay direction [38].

2.6.2 Panning Index Window (PIW)

Avendano proposed a method [62] designed with the aim of identifying, selecting and processing a source or sources panned to a particular direction of the stereo mix-down (Eq.(2.2)). He defined the *panning index* as

$$\Psi(k,r) = \left[1 - \frac{|X_1(k,r)X_2^*(k,r)|}{|X_1(k,r)|^2 + |X_2(k,r)|^2}\right]\hat{\Delta}(k,r), \quad \forall(k,r), \tag{2.78}$$

where

$$\hat{\Delta}(k,r) = \begin{cases} 1 & \text{if} \ \left(\frac{|X_1(k,r)X_2^*(k,r)|}{|X_1(k,r)|^2} - \frac{|X_2(k,r)X_1^*(k,r)|}{|X_2(k,r)|^2}\right) > 0 \\ 0 & \text{if} \ \left(\frac{|X_1(k,r)X_2^*(k,r)|}{|X_1(k,r)|^2} - \frac{|X_2(k,r)X_1^*(k,r)|}{|X_2(k,r)|^2}\right) = 0 \\ -1 & \text{if} \ \left(\frac{|X_1(k,r)X_2^*(k,r)|}{|X_1(k,r)|^2} - \frac{|X_2(k,r)X_1^*(k,r)|}{|X_2(k,r)|^2}\right) < 0. \end{cases} \tag{2.79}$$

Negative and positive values in $\Psi(k,r)$ will correspond to sources panned to the left and sources panned to the right, respectively. For example, time-frequency bins with values close to $\Psi = -1$ will correspond to a top-left panned source. The relation between the mixing parameters of a source and its panning index is given by:

$$\Psi_n = \left[1 - \frac{a_{1n}a_{2n}}{a_{1n}^2 + a_{2n}^2}\right] \operatorname{sign}(a_{2n} - a_{1n}) \tag{2.80}$$

The estimation of the different $\hat{\Psi}_n$ associated to each source was not considered in the original work, but a histogram of $\Psi(k,r)$ reveals the mixing structure of the mixture. A soft time-frequency masking demixing is then applied to estimate the sources:

$$M_n(k,r) = \upsilon_0 + (1-\upsilon_0)\exp\left(\frac{-(\Psi(k,r) - \hat{\Psi}_n)^2}{2\zeta_0}\right) \quad \forall (k,r), \tag{2.81}$$

where ζ_0 controls the width of the window, and υ_0 is a floor value necessary to avoid setting STFT values to zero, which might result in musical-noise artifacts. The STFT image of the source s_n panned at Ψ_n is obtained by applying the Gaussian smoothing window to the mixture:

$$\hat{S}_{mn}(k,r) = M_n(k,r)X_m(k,r) \quad \forall (k,r). \tag{2.82}$$

2.6.3 ADRess

As in the case of PIW separation, the ADRess (*Azimuth Discrimination and Resynthesis*) algorithm [63] assumes an instantaneous stereo mixing process. ADRess uses gain scaling and phase cancellation techniques to expose frequency dependent nulls across the azimuth domain, from which source separation and resynthesis is carried out. The appropriate gain required for the gain-scaling operation is determined by creating a frequency-azimuth plane for each time frame. The discrete Fourier transform (DFT) of the two channels is obtained for a given time frame, and the right mixture is scaled by a range of values corresponding to 0° (far left) to 180° (far right) azimuth. The resultant frames are subtracted from the left mixture frame and the results are used to construct one frequency-azimuth plane:

$$AZ(k,u,r)_2 = |X_1(k,r) - g(u)X_2(k,r)| \quad \forall (k,u,r), \tag{2.83}$$

where $g(u)$ is a gain factor defined as

$$g(u) = \frac{l}{\beta}, \quad u = 0, 1, ..., \beta \tag{2.84}$$

and β refers to how many equally spaced gain scaling values of g are used to construct the frequency-azimuth plane. A similar procedure is also carried out for the left channel, being this one scaled and subtracted from the right channel:

$$AZ(k,u,r)_1 = |X_2(k,r) - g(u)X_1(k,r)| \quad \forall (k,u,r). \tag{2.85}$$

Note that in this algorithm, the term "azimuth" is loosely used, as it is not dealing with real angles of incidence. The azimuth is considered here as a function of the intensity ratio created by the stereophonic panning law. Due to the overlap between sources, the apparent frequency dependent null drifts away from a source position and may be at a minimum in a position where there is no source at all. To overcome this problem, an “azimuth subspace width” is defined. This allows to recover nulls within a given neighborhood around the apparent position of the source. A wide azimuth subspace will result in worse rejection of nearby sources, whereas a narrow azimuth subspace will lead to poor resynthesis and missing frequency information. The final resynthesis is carried out using the short-time spectrum estimated from the nulls in the frequency-azimuth plane, following an overlap-add scheme.

2.7 Monaural Separation

Blind Audio Source Separation methods are generally based on statistical assumptions. The use of several observation channels make possible to estimate several sources with very little information about them. However, the use of advanced source models can improve the separation quality or even perform separation from monaural mixtures. In this context, some a priori knowledge is necessary to extract the sources from a single mixture. Algorithms for one-channel separation are often classified between *supervised* and *unsupervised* approaches [64]:

- *Supervised*: Include methods based on training the model with a sound example database.
- *Unsupervised*: No source-specific prior knowledge.

Unsupervised methods are also divided into three categories:

- *Model-based Inference*: This methods use a parametric model of the sources and the parameters are estimated from the observed mixture signal. Implicit information can be used to design different types of estimators. In music applications, the most commonly used parametric model is the sinusoidal model, suitable for the separation of pitched musical instruments [65].
- *Unsupervised Learning*: These methods usually apply a simple non-parametric model and use less prior information of the sources. They learn the source characteristics from the data. Some algorithms are based on the independence and sparsity assumptions already discussed. One of the most popular approaches here is *Non-Negative Matrix Factorization* (NMF).
- *Psychoacoustically motivated methods*: This category includes almost all CASA methods based on auditory perception, which are briefly described in Section 2.7.2.

2.7.1 Non-Negative Matrix Factorization

Consider the spectrogram of a single signal mixture of several sources represented by the $K \times R$ matrix $|\mathbf{X}(k,r)|$, with columns $|\mathbf{x}^r|$ containing the magnitude DFT of a time-frame of the input mixture in the time domain. Now, let us express the observed magnitude DFT at time-frame r as an additive expansion of basis functions $\mathbf{b}_j$, $j = 1, \ldots, J$:

$$|\mathbf{x}^r| = \sum_{j=1}^{J} g_{j,r}\mathbf{b}_j \quad r = 1, \ldots, R, \tag{2.86}$$

where J is the number of basis functions, and $g_{j,r}$ is the gain of the j-th basis function in the r-th frame. Note that the notation c_{kr} has not been used here in order to avoid confusion with the actual time-frequency coefficients and to be consistent with the classical NMF nomenclature. Moreover, the basis functions $\mathbf{b}_j$ should not neither be confused with the DFT basis functions $\mathbf{b}_k$ used for computing the spectrogram of the mixture signal.

Following this model, each sound source can be modeled as a sum of several components (where the term component denotes here a basis function together with its time varying gain):

$$\mathbf{y}_n^r = \sum_{j \in \mathcal{S}_n} g_{j,r} \mathbf{b}_j, \tag{2.87}$$

where $\mathcal{S}_n$ is the set of components within source n. In a matrix form, the model becomes:

$$|\mathbf{X}(k,r)| = \mathbf{BG}, \tag{2.88}$$

where $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_J]$ is the $K \times J$ *basis matrix*, and $\mathbf{G}$ is the $J \times R$ *gain matrix*. Although several data representations can be used for the observed signals, absolute values of the DFT are often used (as considered here). As magnitude spectra are non-negative by definition, it is advantageous to restrict the basis functions to be entry-wise non-negative. In addition, it is useful to allow only positive gains (components purely additive). Thus, in the signal model $|\mathbf{X}(k,r)| = \mathbf{BG}$, the element-wise non-negativity of $\mathbf{B}$ and $\mathbf{G}$ alone is sufficient condition for the separation of sources in many cases (without an explicit assumption of their independence).

The *non-negative matrix factorization* (NMF) algorithm is based on the following optimization problem:

$$\min \| |\mathbf{X}(k,r)| - \mathbf{BG} \|^2 \quad \text{subject to} \quad \mathbf{B}, \mathbf{G} \geq 0. \tag{2.89}$$

The algorithm of Lee and Seung [19] minimizes the Euclidean distance by initializing the entries of $\mathbf{B}$ and $\mathbf{G}$ with random positive values, and then by updating them iteratively using multiplicative rules:

$$\mathbf{B} \leftarrow \mathbf{B} \circ (|\mathbf{X}(k,r)|\mathbf{G}^{\mathrm{T}})./(\mathbf{BGG}^{\mathrm{T}}) \tag{2.90}$$

and

$$\mathbf{G} \leftarrow \mathbf{G} \circ (\mathbf{B}^{\mathrm{T}}|\mathbf{X}(k,r)|)./(\mathbf{B}^{\mathrm{T}}\mathbf{BG}), \tag{2.91}$$

where $\circ$ and ./ denote the element-wise multiplication and division, respectively.

To summarize, the NMF algorithm is composed of the next steps:

1. Initialize each entry of $\mathbf{B}$ and $\mathbf{G}$ with the absolute values of Gaussian noise.
2. Update $\mathbf{G}$ using Eq.(2.91).
3. Update $\mathbf{B}$ using Eq.(2.90).
4. Repeat Steps 2 and 3 until the values converge.

Note that the NMF can be used only for a non-negative observation matrix, thus, it is not suitable for separating time-domain signals. If a mixture spectrogram is a sum of sources with a static spectrum with a time-varying gain, and each of them is active in at least one frame and frequency in which the other components are not active, NMF provides perfect separation.

However, real-world music signals rarely fulfill these conditions, and two or more sources that are simultaneous at all times would be represented by a single component. NMF with a sparsness constraint is referred as *non-negative sparse coding* (NSC). This technique uses also minimization of Euclidean distances but adding an additional term that enforces sparseness on the activity of the basis functions. For example, Virtanen [64] proposed NSC for monaural sound separation based on minimizing a cost function which is a weighted sum of three terms: a reconstruction error term (Euclidean distance), a temporal continuity term, and a sparseness term.

Figure 2.11 shows the result of applying the algorithm to a signal made up of two notes played by two different instruments (trumpet and oboe). Two components were calculated corresponding to the harmonic spectrum of the two notes played. Their temporal gain is showed in the two upper plots and they give information about when the notes are being played and their level over time.

2.7.2 Computational Auditory Scene Analysis

Another popular approach to monaural separation is *Computational Auditory Scene Analysis* (CASA). CASA algorithms belong to the category of psychoacoustically motivated methods in monaural separation and are aimed at imitating the human mechanisms of hearing for identifying and understanding sound objects from a complex sound scene. The processes underlying the perceptual interpretation of sound sources have been studied for decades, being the work by Bregman [11] one of the most well-known references in this field. Based on Bregman's findings, CASA systems look for specific features related to the stages of psychoacoustical perception, from acoustical processing in the outer and inner ear, to neural and cognitive processes in the brain.

Computational systems for auditory scene analysis usually follow a similar structure, as shown in Figure 2.12 (reproduced from [7]). Firstly, the recorded mixture enters the peripheral analysis block, which results in a time-frequency representation of auditory activity. The most common front-end used in CASA systems is the *cochleagram* [7]. Using this representation, an analysis step is carried out with the aim of extracting relevant features, such as periodicity, onsets, offsets, frequency and amplitude modulation, etc. With the aid of these features, mid-level representations are formed, usually called *segments*. The grouping cues of Auditory Scene Analysis combined with specific source models are then used to perform separation of the mixture into different streams. In the last block, the final audio waveform is resynthesized from the separated streams.

The main advantage of CASA against BSS is its easier applicability to more realistic mixtures, even when the statistical constraints of BSS are not fully fulfilled. However, CASA methods usually need large training data and the processing is quite complicated, which makes difficult to apply them to large problems.

2.8 Performance measurement

The evaluation method proposed by Vincent *et al.* [66] allows to obtain a meaningful insight into the type of errors that affect the separated sources. This evaluation method has been widely adopted by the separation community, being the usual performance measures used in

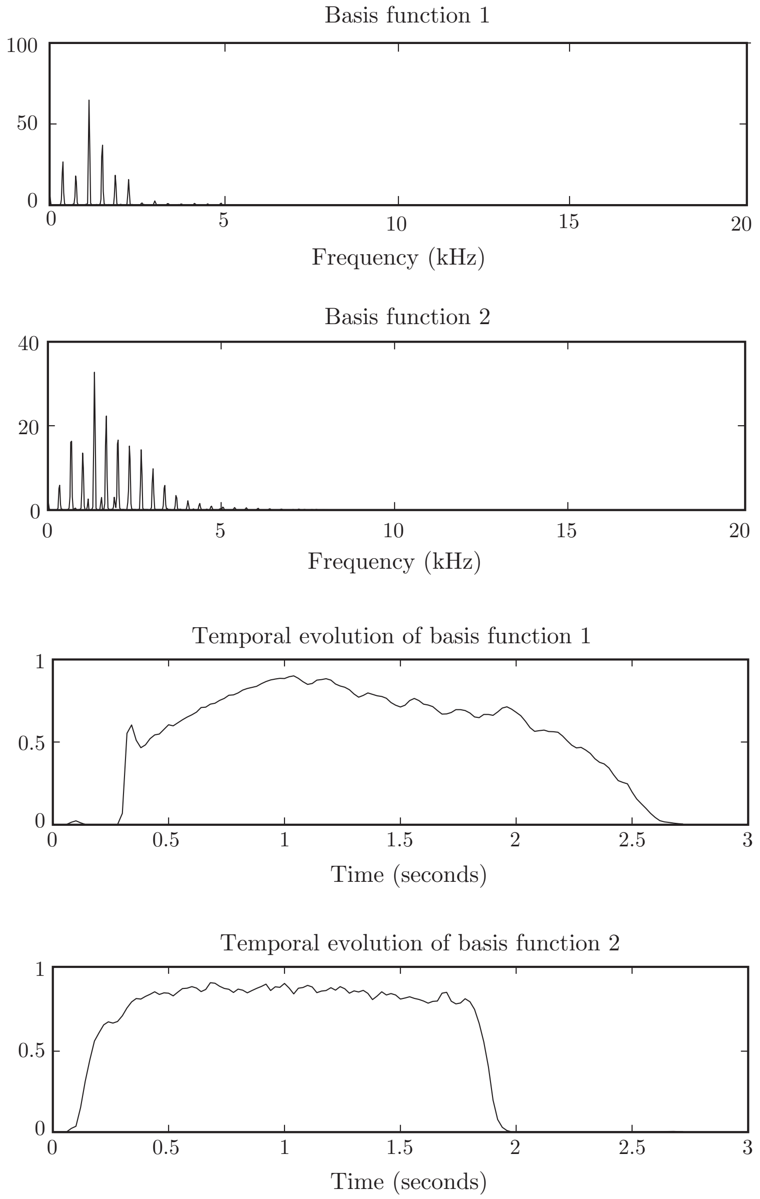


Figure 2.11. NMF components estimated from a mixture signal of two notes (trumpet and oboe). Basis functinos are plotted on top and their gains at the bottom.

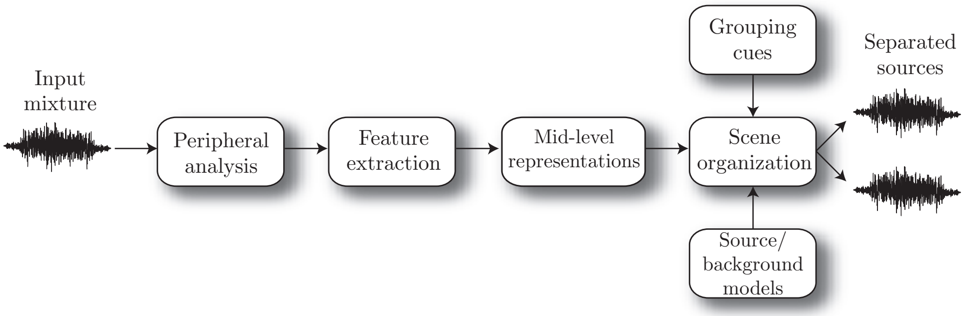


Figure 2.12. System architecture of a typical CASA system.

underdetermined source separation problems [67][68]. This is the reason why these measures have also been selected to evaluate the performance of the different approaches discussed throughout this thesis.

The method was first formulated using single-channel signals as follows. In the noiseless case, it consists of decomposing each estimated source $\hat{s}_n$ as the sum

$$\hat{s}_n = s_{\text{target}} + e_{\text{interf}} + e_{\text{artif}}, \tag{2.92}$$

where s_{target} is an allowed deformation of the target source $s_n(t)$, e_{interf} is an allowed deformation of the sources which accounts for the interferences of the unwanted sources, and e_{artif} is an artifact term that may correspond to artifacts of the separation algorithm such as musical noise, etc. or simply to deformations induced by the separation algorithm that are not allowed. Depending on the allowed transformations, there are several ways of computing such a decomposition. The allowed distortion can be a constant gain, a constant filter, a time-varying gain or a time-varying filter.

With the purpose of carrying out a global evaluation campaign involving researchers from all over the world, Vincent extended the above ideas to evaluate multichannel signals corresponding to estimates of the source images $\hat{s}_{mn}$. The decomposition in this case is:

$$\hat{s}_{mn} = s_{\text{true},mn} + e_{\text{spat},mn} + e_{\text{interf},mn} + e_{\text{artif},mn}, \tag{2.93}$$

where $s_{\text{true},mn}$ is the true source image, and $e_{\text{spat},mn}$, $e_{\text{interf},mn}$ and $e_{\text{artif},mn}$ are distinct error components representing spatial (or filtering) distortion, interference and artifacts. The spatial distortion and interference components are modelled as filtered versions of the true source images, computed by least-squares projection of the estimated source image onto the corresponding signal subspaces [66]:

$$e_{\text{spat},mn} = P_n^{L_{\text{fil}}}(\hat{s}_{mn}) - s_{\text{true},mn} \tag{2.94}$$

$$e_{\text{interf},mn} = P_{\text{all}}^{L_{\text{fil}}}(\hat{s}_{mn}) - P_n^{L_{\text{fil}}}(\hat{s}_{mn}) \tag{2.95}$$

$$e_{\text{artif},mn} = s_{\text{true},mn} - P_{\text{all}}^{L_{\text{fil}}}(\hat{s}_{mn}) \tag{2.96}$$

where $P_n^{L_{\text{fil}}}$ is the least-squares projector onto the subspace spanned by $s^{\tau}_{\text{true},mn}$, $1 \leq m \leq 2$, $0 \leq \tau \leq L_{\text{fil}}$, and $P_{\text{all}}^{L_{\text{fil}}}$ is the least-squares projector onto the subspace spanned by $s^{\tau}_{\text{true},mn'}$, $1 \leq m \leq 2$,$1 \leq n' \leq N$, $0 \leq \tau \leq L_{\text{fil}} - 1$. The superindex $^{\tau}$ denotes a signal delay $s^{\tau}_{\text{true},mn} = s_{\text{true},mn}(t-\tau)$, and the filter length L_{fil} is set to 512 coefficients (32 ms at $f_s = 16$ kHz).

Once such decomposition has been performed, the following objective measures can be defined using a set of energy ratio criteria expressed in decibels:

- *Source to Image Spatial distortion Ratio* (ISR):

$$\text{ISR} = 10\log_{10}\frac{\sum_{m=1}^{2}\sum_t s^2_{\text{true},mn}}{\sum_{m=1}^{2}\sum_t e^2_{\text{spat},mn}}. \tag{2.97}$$

- *Source to Interference Ratio* (SIR):

$$\text{SIR} = 10\log_{10}\frac{\sum_{m=1}^{2}\sum_t (s_{\text{true},mn} + e_{\text{spat},mn})^2}{\sum_{m=1}^{2}\sum_t e^2_{\text{interf},mn}}. \tag{2.98}$$

- *Source to Artifacts Ratio* (SAR):

$$\mathrm{SAR} = 10\log_{10}\frac{\sum_{m=1}^{2}\sum_t (s_{\mathrm{true},mn} + e_{\mathrm{spat},mn} + e_{\mathrm{interf},mn})^2}{\sum_{m=1}^{2}\sum_t e_{\mathrm{artif},mn}^2}. \tag{2.99}$$

A total distortion (reconstruction error) is measured by the *Signal to Distortion Ratio* (SDR):

$$\mathrm{SDR} = 10\log_{10}\frac{\sum_{m=1}^{2}\sum_t s_{\mathrm{true},mn}^2}{\sum_{m=1}^{2}\sum_t (e_{\mathrm{spat},mn} + e_{\mathrm{interf},mn} + e_{\mathrm{artif},mn})^2}. \tag{2.100}$$

Notice that the SDR weights the three error components equally. It should be noted that, in practice, each component should be given a different weight depending on the application. For instance, spatial distortion is of little importance for most applications, except for karaoke where it can result in imperfect source cancellation, even in the absence of interference or artifacts. As will be seen in Chapter 6, interference errors play an important role in audio up-mixing for spatial sound reproduction, since localization accuracy seems to be influenced by interference terms. However, artifacts are greatly masked in the reproduced scene. This is not the case of hearing aid applications, where "gurgling" noise should be avoided at the cost of increased interference.

2.9 Conclusion

In this chapter, an overview of sound source separation techniques has been carried out. The properties of audio mixtures highly depend on the recording set-up and different mixing models are commonly used to mathematically describe their generation. In this context, synthetic mixtures characterized by a linear instantaneous model are very different in nature from real live recordings picked up by a set of microphones, which are more accurately expressed by means of a convolutive mixing model. The relation between the number of sources N and the number of mixtures M, has also a big influence in the separation difficulty and the problem is classified as evendetermined ($M = N$), overdetermined ($M > N$) or underdetermined ($M < N$). The source separation problem has been presented under a general probabilistic framework, with special attention to the underdetermined case. The challenge of separating underdetermined mixtures resides in the fact that the sources can not be extracted by system inversion. Thus, separation algorithms usually follow a staged approach: first estimate the mixing matrix and afterwards estimate the sources according to a selected criterion. Both stages are usually performed assuming that the signals have a sparse structure when applying a proper signal decomposition. Therefore, the sparsity assumption is the fundamental piece that supports most of the algorithms designed to deal with underdetermined mixtures, including all the methods presented in the following chapters of this thesis.

3
Multi-Level Thresholding Separation

Multi-Level Thresholding Separation 3

Thresholding is an important technique for image segmentation, which is used for the identification and extraction of targets from their background on the basis of the distribution of pixel intensities in image objects. Image segmentation has many similarities with underdetermined BSS. Binary time-frequency masking aims at decomposing the observed mixture in the time-frequency domain into several objects defined by a set of binary masks. Therefore, the spectrogram of a sound mixture can be thought of as an image composed of different non-overlapping sound objects. In this chapter, we present two time-frequency masking techniques based on image segmentation. Both techniques perform clustering based on multi-level thresholding but from different localization cues, which enables to perform separation of both instantaneous and anechoic/convolutive underdetermined mixtures. The segmentation approach presented in this chapter is a powerful tool for the separation of stereo mixtures in real-time.

3.1 Introduction

As seen in the last chapter, sparse methods for audio separation are able to provide a solution to the underdetermined problem taking advantage of the sparsity achieved by signal decompositions, especially by time-frequency transformations. Most sparse methods follow an staged approach where first the mixing matrix is estimated and afterwards the sources are obtained by using one of the techniques seen in Section 2.5.5. In this context, time-frequency masking was introduced as one of the most common approaches to source estimation, where the estimated images of the sources are directly obtained by applying the computed masks over the mixture spectrogram. The source separation problem is therefore reduced to the estimation of these masks by analyzing the available observations. In this chapter, time-frequency masking is addressed from an image processing perspective, using some well-known *image segmentation* algorithms.

Image segmentation is a powerful tool used in computer vision. Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). This way, segmentation algorithms are aimed at simplifying or changing the representation of an image into something that is more meaningful and easier to analyze [69]. More precisely, image segmentation is the process of assigning a label to every pixel in an image so that pixels with the same label share certain visual characteristics. In this context, the result of applying a segmentation process to an image is a set of segments that collectively cover the entire image. All the pixels that belong to the same segment are similar with respect to some characteristic or computed property. For example, segmentation can be performed on features such as color, intensity, or texture. One efficient segmentation technique used for the extraction of multiple objects from an image is *multi-level thresholding*. This technique computes a set of thresholds that splits the image into different objects according to the statistics of a given feature, generally the gray-level. Figure 3.1 shows the results of applying a segmentation process to a medical image by considering the extraction of two classes (*bi-level thresholding*) and three classes (*tri-level thresholding*).

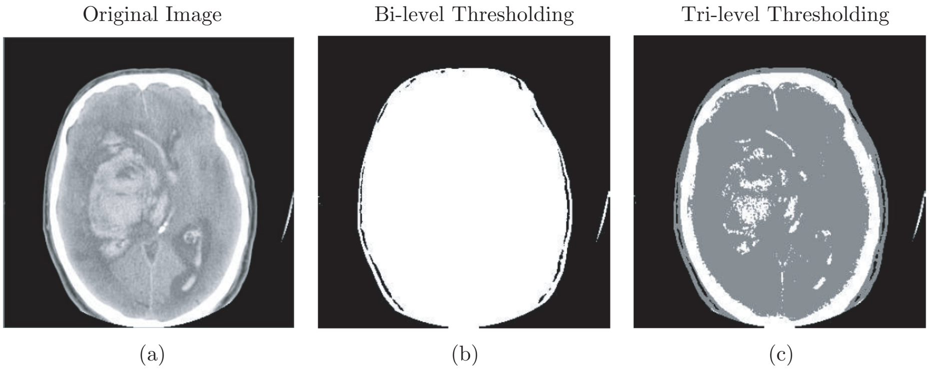


Figure 3.1. Image thresholding example. a) Original image. b) bi-level thresholding with two classes represented in black and white. c) tri-level thresholding with three classes represented in black, white and gray.

Throughout this chapter, two stereo source separation techniques based on multi-level thresholding and time-frequency masking are presented. The computation of time-frequency masks is performed by analyzing the mixture representation in the STFT domain, but segmentation is carried out according to different computed features. These are directly related to the *Interaural Level Difference* (ILD) and *Interaural Time Difference* (ITD) binaural localization cues (see Section 5.2). If level differences between the observation channels are considered in the segmentation, the approach is shown to be useful for the separation of underdetermined instantaneous mixtures. Segmentation based on the computation of time differences (which are directly related to the *direction-of-arrival*, or DOA, of the sources) is carried out to separate real audio mixtures recorded by a pair of closely spaced microphones. The separation time-frequency masks are directly estimated from the observed features, without the need for performing a previous estimation of the mixing matrix. Moreover, it is not necessary to search for peaks (as in the DUET algorithm, explained in Section 2.6.1) nor to specify initial centroid values as in other clustering approaches [70][59], which results in a more robust and unsupervised algorithm. In addition, the proposed approach admits an efficient implementation, which allows to perform separation with minimal computation time.

The chapter is structured in two sections. Section 3.2 describes how the multi-level thresholding approach is applied to separate instantaneous mixtures, with a detailed description of all the steps involved in the computation of the separation masks. Similarly, Section 3.3 explains how the method is extended with the aim of separating real stereo mixtures using two microphones. The common steps involved in each approach are the following: the computation of a time-frequency representation of the separation feature, the formation of a *weighted histogram* and the application of the *Otsu* multi-level thresholding algorithm. The details concerning these two approaches are pointed out in each section, including the description of a set of experiments aimed at evaluating the performance of the algorithm under different conditions and mixing configurations.

3.2 Separation of instantaneous mixtures

Studio recordings are usually made via the instantaneous mixing of recorded mono or stereo tracks which usually correspond to different sound sources. These tracks are mixed using amplitude panning to create the stereophonic effect. Therefore, many commercial stereo recordings can be categorized as instantaneous mixtures where the mixing ratios have been set using any of the existing panning laws (see Section 2.2.2).

This section introduces the proposed separation approach based on multi-level thresholding. An overview of the algorithm is depicted in Figure 3.2, which can be summarized in the following steps:

1. **Pan map calculation** (Section 3.2.1). An analysis of amplitude differences between the left and right channels is performed in the STFT domain. This analysis results in a *pan map*, a matrix containing level differences estimates for the whole time-frequency plane.

2. **Weighted histogram** (Section 3.2.2). The histogram that shows the distribution of values that conform the pan map can be thought as an spatial representation of the mixed sources and shows a clear structure imposed by the mixing matrix due to the sparsity provided by the STFT. Weighted histograms help to enhance this mixing structure, which has an influence on the subsequent thresholding process.

3. **Multi-level Thresholding** (Section 3.2.3). The distribution given by the histogram is processed for calculating a set of thresholds that maximize the *between-class variance*, as explained in Section 3.2.3. These thresholds are used for segmenting the pan map into the final separation masks.

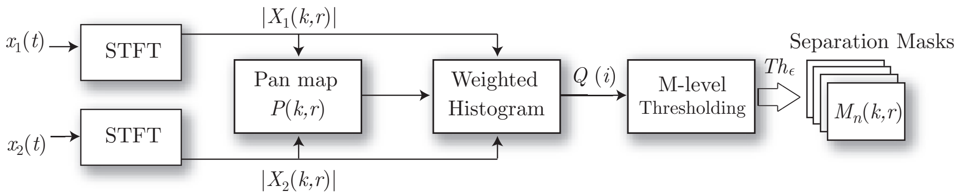


Figure 3.2. Processing scheme for the separation of instantaneous mixtures.

3.2.1 Pan Map

Given the linearity of the STFT, the instantaneous mixing expressed by Equation (2.4) can be directly expressed in the time-frequency domain as:

$$X_m(k,r) = \sum_{n=1}^{N} a_{mn} S_n(k,r), \quad m = 1, \ldots, M, \tag{3.1}$$

where $S_n(k,r)$ are the STFT of the sources. Note that, if a stereo panning law is used in the mix-down, the mixing matrix is directly obtained by the panning gains defined by the panoramic parameter ϕ_n (Section 2.2.2):

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ a_{21} & \cdots & a_{2N} \end{bmatrix} = \begin{bmatrix} \alpha(\phi_1)^L & \cdots & \alpha(\phi_N)^L \\ \alpha(\phi_1)^R & \cdots & \alpha(\phi_N)^R \end{bmatrix}. \tag{3.2}$$

A source with $a_{1n} > a_{2n}$, i.e. $\phi_n < 0.5$, is said to be panned to the left. Similarly, a source with $a_{1n} < a_{2n}$, i.e. $\phi_n > 0.5$ is said to be panned to the right. When $a_{1n} = a_{2n}$, or $\phi_n = 0.5$, the source is said to be center-panned.

Following the usual approach of many sparse methods for stereo separation, and similarly to the symmetric attenuation estimate used in DUET, we define the pan map matrix as

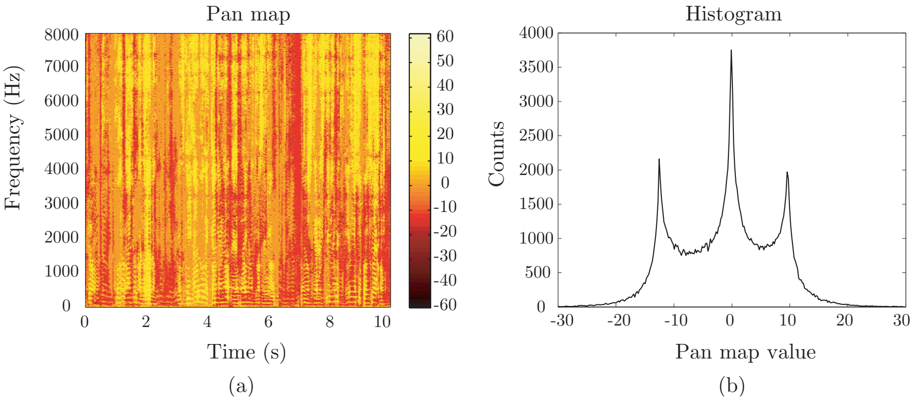
$$P(k,r) = 20 \log_{10}\left(\frac{|X_2(k,r)|}{|X_1(k,r)|}\right), \quad \forall (k,r), \tag{3.3}$$

which is symmetric with respect to the center of the panoramic parameter range ($\phi_n = 0.5$). The pan map represents the log-ratio of the magnitudes of the left and right mixtures in the STFT domain and is expressed in matrix notation as $\mathbf{P}(k,r)$. As the STFT provides a sparse representation of the sources, the calculated pan map values tend to be clustered in regions around $20\log_{10}(a_{2n}/a_{1n})$. If the energy preserving panning law is used in the mixing, the peaks can be directly related to the panoramic parameter by $20\log_{10}\left(\tan\left(\frac{\phi_n \pi}{2}\right)\right)$. Figure 3.3 shows the pan map and its histogram obtained for a mixture of three speech sources, that were mixed with pan parameters $\phi_1 = 0.15$, $\phi_2 = 0.5$ and $\phi_3 = 0.8$. It can be clearly observed how three peaks are located at pan map values close to -12.4, 0 and 9.8. Notice how the histogram provides also a graphical description of the spatial properties of the mixture, where time-frequency points dominated by a source panned to the left tend to have negative $P(k,r)$ values, while those dominated by a source panned to the right are more likely to have $P(k,r)$ values greater than zero.

Pan map deviations

Perfect *W-Disjoint Orthogonal* (WDO) sources, i.e. sources that do not interfere in the time-frequency domain, would always show their correct mixing directions when computing the pan map. However, real-world sources always have some spectral overlap, resulting in a deviation of their expected pan map values. In order to provide an idea of how these deviations occur, consider a time-frequency point where a target source $S_\mathrm{t}(k,r) = |S_\mathrm{t}(k,r)|e^{j\Phi_\mathrm{t}}$ is mixed with an interfering source $S_\mathrm{i}(k,r) = |S_\mathrm{i}(k,r)|e^{j\Phi_\mathrm{i}}$, so that the observed mixtures can be written as

$$\begin{aligned} X_1(k,r) &= a_{1\mathrm{t}} S_\mathrm{t}(k,r) + a_{1\mathrm{i}} S_\mathrm{i}(k,r) = S_{1\mathrm{t}}(k,r) + S_{1\mathrm{i}}(k,r) \\ X_2(k,r) &= a_{2\mathrm{t}} S_\mathrm{t}(k,r) + a_{2\mathrm{i}} S_\mathrm{i}(k,r) = S_{2\mathrm{t}}(k,r) + S_{2\mathrm{i}}(k,r). \end{aligned} \tag{3.4}$$



(a) (b)

Figure 3.3. Pan map example obtained from a stereo mixture of 3 sources mixed with pan parameters $\phi_1 = 0.15$, $\phi_2 = 0.5$ and $\phi_3 = 0.8$. a) Pan map. b) Histogram.

The ratio of their magnitudes can be then expressed as

$$\begin{aligned}\frac{|X_2(k,r)|}{|X_1(k,r)|} &= \frac{|S_{2\mathrm{t}}|}{|S_{1\mathrm{t}}|} + \left(\frac{|S_{2\mathrm{t}}+S_{2\mathrm{i}}||S_{1\mathrm{t}}| - |S_{1\mathrm{t}}+S_{1\mathrm{i}}||S_{2\mathrm{t}}|}{|S_{1\mathrm{t}}+S_{1\mathrm{i}}||S_{1\mathrm{t}}|}\right) \\ &= \rho_W + \xi, \end{aligned} \tag{3.5}$$

where ρ_W is the correct mixing ratio without interference, i.e. $a_{2\mathrm{t}}/a_{1\mathrm{t}}$, and ξ is the error term due to the interference:

$$\xi = \frac{\left|\frac{|S_\mathrm{t}|}{|S_\mathrm{i}|}\frac{a_{2\mathrm{t}}a_{1\mathrm{t}}}{a_{1\mathrm{i}}}e^{j(\Phi_\mathrm{t}-\Phi_\mathrm{i})} + \frac{a_{2\mathrm{i}}a_{1\mathrm{t}}}{a_{1\mathrm{i}}}\right| - \left|\frac{|S_\mathrm{t}|}{|S_\mathrm{i}|}\frac{a_{2\mathrm{t}}a_{1\mathrm{t}}}{a_{1\mathrm{i}}}e^{j(\Phi_\mathrm{t}-\Phi_\mathrm{i})} + a_{2\mathrm{t}}\right|}{\left|\frac{|S_\mathrm{t}|}{|S_\mathrm{i}|}\frac{a_{1\mathrm{t}}^2}{a_{1\mathrm{i}}}e^{j(\Phi_\mathrm{t}-\Phi_\mathrm{i})} + a_{1\mathrm{t}}\right|} \tag{3.6}$$

This error term depends on several factors: the panning position of the target and the interference source, the phase difference between the sources and the *Target-to-Interference Ratio* ($\mathrm{TIR} = 20\log 10(|S_\mathrm{t}|/|S_\mathrm{i}|)$). In the same way, the calculated pan map suffers a deviation Δ_P due to this error term:

$$P(k,r) = 20\log_{10}(\rho_W) + 20\log_{10}\left(1 + \frac{\xi}{\rho_W}\right) = 20\log_{10}(\rho_W) + \Delta_P. \tag{3.7}$$

In Figure 3.4 are depicted different graphs showing the deviation Δ_P produced in the pan map with respect to the WDO case, as a function of the panoramic parameter of the target source ϕ_t and the panoramic parameter of the interfering source ϕ_i. At the top panel the deviations are presented for different values of TIR, considering that the sources interfere constructively. At the bottom panel the sources are assumed to interfere destructively.

In general, it is observable that source deviations are produced towards the interfering source when the sources add constructively. Therefore, an interfering source that is located in the panoramic at the left of the target source will produce negative deviations in the estimated pan map, while interfering sources located at the right of the target will result in positive deviations. The magnitude of these deviations is greater when the interfering source is stronger and its spatial distance (in terms of panoramic) is higher. When the sources add destructively the behavior is more complex, since the deviations tend to be greater, with a complicated pattern that is highly dependent on the relative energy of the sources.

Figure 3.5, shows the same 3 sources of Figure 3.3 mixed with higher spatial distance: $\phi_1 = 0.05$, $\phi_2 = 0.5$ and $\phi_3 = 0.95$. Notice how the peaky central source remains but the other two peaks slightly disappear because a large amount of points have been severely affected by the interference and have been spread towards the other sources due to the increased spatial distance. This effect can become a serious problem when estimating the mixing directions directly from a normal histogram. In the following section it is described how to enhance the observed mixture directions by means of a weighted histogram.

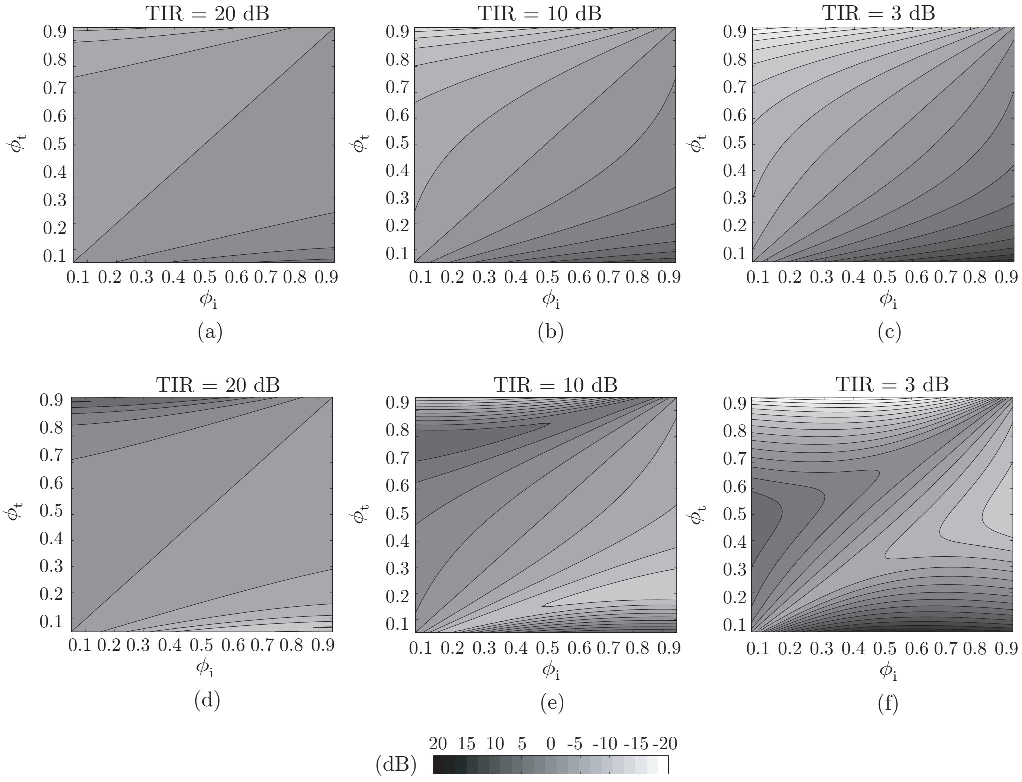


Figure 3.4. Pan map deviation Δ_P for different cases. Top: deviations assuming constructive interference with a) TIR= 20 dB, b) TIR= 10 dB and c) TIR= 3 dB. Bottom: deviations assuming destructive interference with a) TIR= 20 dB, b) TIR= 10 dB and c) TIR= 3 dB.

3.2.2 Weighted histograms

The visual inspection of a scatter plot from mixtures made of sparse sources usually reveals the mixing directions associated to the columns $\mathbf{a}_n$ of the mixing matrix (see, for examples, Figure 2.4). Due to sparsity, many of the points that can be seen in a scatter plot are close to the origin and their noisy distribution does not bring any useful information for estimating the mixing directions. Some approaches for estimating the mixing matrix using clustering techniques can be found in the literature [37][71]. Nevertheless, the separation approach here described does not need to accurately estimate the mixing directions. In fact, the clustering performed in the following step will directly depend on the shape of the observed histogram, so it is important to enhance the presence of the sources before carrying out the separation. Weighted histograms have been previously proposed in the separation field as a means for enhancing the sparse structure of audio mixtures [38][72][73][74]. Next, we describe the formation of a weighted

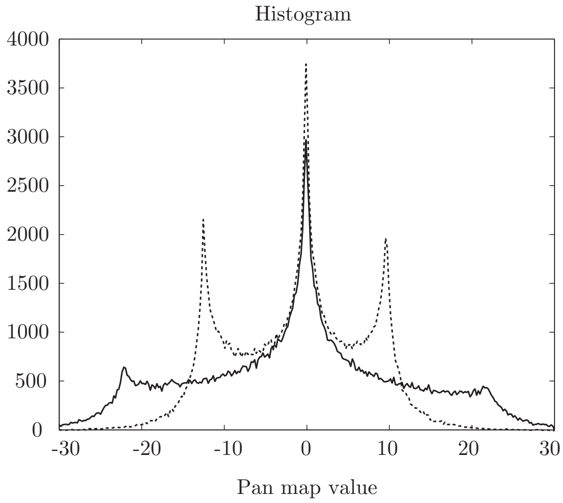


Figure 3.5. Histogram for a stereo mixture of 3 speech sources with pan parameters $\phi_1 = 0.05$, $\phi_2 = 0.5$ and $\phi_3 = 0.95$. The histogram of Fig. 3.3(b) has been plotted in dotted line for comparison purposes.

histogram and the use of different weighting functions.

First, we define L_c uniform containers in the boundary $[-P_{\max}, P_{\max}]$, where $P_{\max}$ is a maximum value that defines the range of the pan map values considered in the formation of the histogram. The centers z_i of these containers are:

$$z_i = P_{\max}\left(\frac{2i+1}{L_c} - 1\right), \quad i = 0, \ldots, L_c - 1. \tag{3.8}$$

The weighted histogram Q is calculated by summing all the weighting factors of the time-frequency points with pan map values lying on each of the previously defined L_c containers:

$$Q(i) = \sum_{(k,r)\in\mathcal{S}_i} g_w(k,r), \quad i = 0, \ldots, L_c - 1, \tag{3.9}$$

where $g_w(k,r)$ is the weighting factor for the time-frequency point (k,r) and $\mathcal{S}_i = \{(k,r)|\ \frac{L_c}{P_{\max}} \cdot |P(k,r) - z_i| < 1\}$, i.e. the set of time-frequency points with $P(k,r)$ in the value range defined by container i.

The function $g_w(k,r)$ can be used to select different types of weighting. For example, Rickard and Yilmaz proposed to use a cross power weighted histogram in the DUET algorithm with the aim of making easier the search for peaks used in the separation [38]. The peak picking stage was necessary to estimate the mixing parameters of the sources. Some possibilities explored by the author for computing weighted histograms include:

- **Non-weighted histogram**: No weighting is applied, so it corresponds to a normal histogram that only counts occurrences:

$$g_w(k,r) = 1. \tag{3.10}$$

- **Magnitude weighting**: The weighting factor is given by the mean amplitude in the left and right channels:

$$g_w(k,r) = \frac{1}{2}\left(|X_1(k,r)| + |X_2(k,r)|\right). \tag{3.11}$$

- **DUET weighting**: The weight of each point is computed as

  $$g_w(k,r) = |X_1(k,r)X_2(k,r)|^p \omega_k^q, \tag{3.12}$$

  where the p and q parameters can be varied according to several criteria and $\omega_k = 2\pi f_k$ is the angular frequency corresponding to frequency index k. The authors suggest a default value of $p = 1$ and $q = 0$ [61].

- **1/log weighting**: Human hearing has a more like logarithmic behavior, as much on the frequency axis as on the magnitude axis. This function gives a greater weighting factor to points in the lower part of the frequency range of interest, which can result in a more perceptually enhanced distribution:

  $$g_w(k,r) = \frac{\log_{10}(10)}{\log_{10}(10 + p \cdot \omega_k)}, \quad \forall r, \tag{3.13}$$

  where the p parameter is here used to adjust the weighting function across the frequency axis. A default value $p = 0.01$ is experimentally suggested.

- **Exponential weighting**: Similarly, a negative exponential weighting function can be also used to enhance low frequencies, where the sources are supposed to concentrate their energy:

  $$g_w(k,r) = e^{-(p \cdot \omega_k)}, \quad \forall r. \tag{3.14}$$

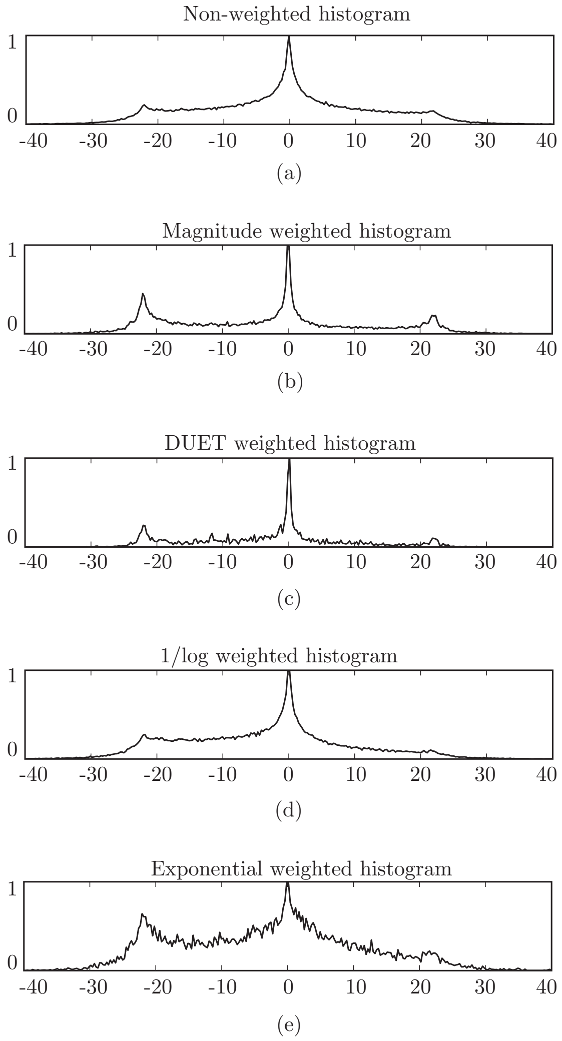
  The p parameter can be used again to modify the shape of the function and obtain different decay envelopes, for example, $p = 0.0001$.

In Figure 3.6, the normalized weighted histograms for the 3 source example mixture are represented. Notice that the shapes obtained by using different weighting functions are very varied. The normal (non-weighted) histogram does not show the prominent peaks achieved by magnitude or DUET-like weighting and, therefore, estimating the mixing matrix **A** by peak picking is more difficult. However, these weighting functions are completely dependent on the relative energy of the sources and, therefore, the distributions correspondent to low-energy sources can be hidden by those pertaining to stronger sources. The $1/\log$ and exponential functions do not depend on the specific source energies, since the weighting function affects only the frequency axis. This can help to enhance the source distributions by reducing the effect of low-energy points, which are usually located at high frequencies. The performance achieved by different types of weighting functions in the separation task will be discussed in Section 3.2.7.

3.2.3 Thresholding

As explained throughout Chapter 2, the usual underdetermined BSS is carried out by means of an staged approach, where the mixing matrix is estimated in the first place and the sources are later obtained by using the knowledge on the mixing process. The proposed *Multi-Level Thresholding Separation* (MuLeTS) approach estimates the sources by means of binary time-frequency masking and the masks are directly computed from the distribution given by the weighted histogram. Therefore, there is not an explicit estimation of the mixing parameters and the problem is solved at once.

Thresholding is a very powerful technique used in image segmentation with the aim of extracting different objects according to any meaningful feature, the most common being the



(a)

(b)

(c)

(d)

(e)

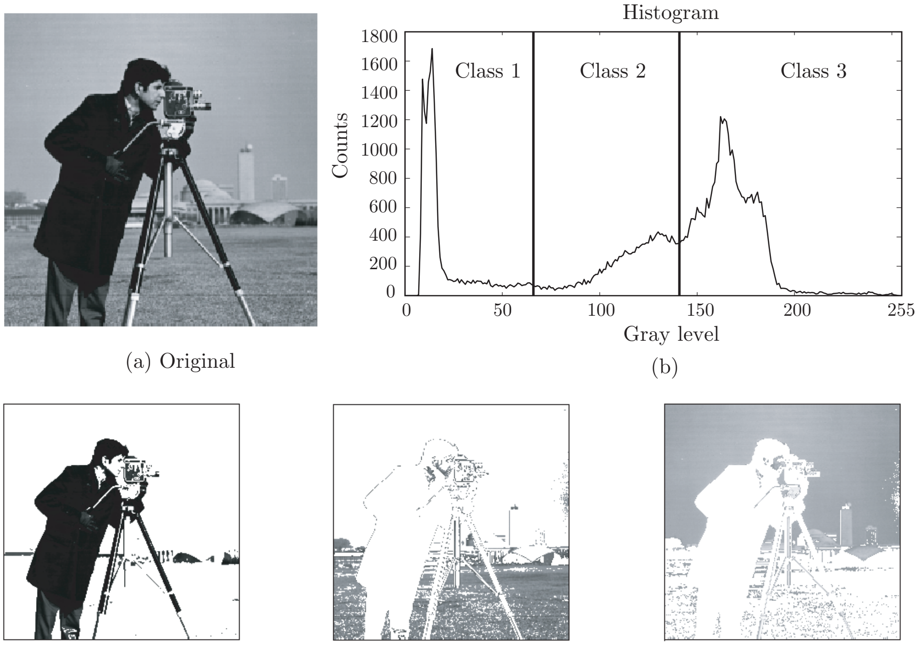
Figure 3.6. Normalized weighted histograms using different weighting functions $g_w(k, r)$. a) Non-weighted. b) Magnitude weighted. c) DUET weighted with $p = 1$, $q = 0$. d) $1/\log$ weighted with $p = 0.01$. e) Exponential weighting with $p = 0.0001$.

gray-level. In many applications of image processing, the gray levels of pixels belonging to the object are substantially different from the gray levels of the pixels belonging to the background. The basic idea of automatic thresholding is to automatically select an optimal or several optimal gray-level threshold values for separating objects of interest in an image from the background based on their gray-level distribution. Examples of thresholding applications include:

- Document image analysis: where the goal is to extract printed characters logos, graphical content, or musical scores [75].
- Map processing: where lines, legends, and characters are to be found [76].
- Scene processing: where a target has to be detected [77].
- Quality inspection of materials, where defective parts must be delineated [78].
- Medical image segmentation: extraction of objects from ultrasonic images, thermal images, x-ray computed tomography, endoscopic images, laser scanning, etc [79].

- Spatio-temporal segmentation of video images [80].

The input of the thresholding algorithm is the histogram of the observed distribution. In the case of *bi-level thresholding* the output is a binary image where entries with value 1 indicate the foreground objects, that is, printed text, a legend, a target, defective part of a material, etc., while zero entries correspond to the background. The gray level which splits the image into two classes (foreground and background) is given by the threshold. When multiple thresholds are computed for splitting the image into more than 2 classes, the process is called *Multi-level thresholding*.



(a) Original

(b)

(c) Class 1

(d) Class 2

(e) Class 3

Figure 3.7. Multi-level thresholding example. a) Original image. b) Histogram. c-d) Separated objects considering 3 classes.

Figure 3.7 shows a gray-level image and the histogram corresponding to the intensity distribution. The vertical dotted lines in the histogram are the segmentation thresholds that divide the original image into the three classes shown at the bottom of the figure. Notice that these classes correspond to three "objects" in the photograph: (c) the man, (d) the ground and (e) the sky.

One of the most referenced thresholding methods in image processing is the Otsu method [81]. Several works on evaluation of thresholding algorithms have shown that, despite it was published a long time ago (1979), the Otsu method is one of the better threshold selection algorithms for general real world images [82][83]. Moreover, this clustering technique has been selected in the proposed audio separation framework for some remarkable advantages:

- It is very general: no specific histogram shape is assumed. It has already been seen that the shape of the histogram may change a lot depending on the mixing configuration, the spectral overlap and the weighting function applied.

- It has been shown to work well and reliably under many different situations: different types of images, additive noise, etc.

- Its extension to multi-level thresholding is straightforward, which enables to carry out underdetermined BSS.
- Admits an efficient implementation, thus opening the possibility of real-time processing.
- There is no need for detecting peaks nor for specifying initial centroid values, which can lead to a more unsupervised approach.

In the following subsections the bi-level and multi-level Otsu algorithms are introduced.

Otsu bi-level thresholding

Let us briefly describe the Otsu algorithm. Consider the probability of a value in the middle of container i of the histogram Q, given by:

$$p_i = \frac{Q(i)}{\sum\limits_{i=1}^{L_c} Q(i)}. \tag{3.15}$$

Note that the above equation satisfies the conditions of a pdf: $\sum_{i=1}^{L_c} p_i = 1$ and $p_i \geq 0$. The mean of the total distribution is therefore given by

$$\mu_T = \sum_{i=1}^{L_c} i p_i. \tag{3.16}$$

In the case of bi-level thresholding, the time-frequency points are divided into two classes by means of a threshold value l. The first class, C_1, has values in the range given by the histogram bins $i \in [1, \ldots, l]$ and the second class C_2 has values within the bins $i \in [l+1, \ldots, L_c]$. Therefore, the probability of occurrence for each class depends on the threshold l and is given by:

$$\omega_1(l) = \sum_{i=1}^{l} p_i, \tag{3.17}$$

$$\omega_2(l) = \sum_{i=l+1}^{L_c} p_i = 1 - \omega_1(l). \tag{3.18}$$

The probability distributions of C_1 and C_2 can be expressed as:

$$p_i^{C_1} = \frac{p_i}{\omega_1(l)}, \tag{3.19}$$

$$p_i^{C_2} = \frac{p_i}{\omega_2(l)}, \tag{3.20}$$

and their means are therefore given by

$$\mu_1(l) = \sum_{i=1}^{l} i p_i^{C^1} = \sum_{i=1}^{l} i \frac{p_i}{\omega_1(l)} = \frac{\mu(l)}{\omega_1(l)}, \tag{3.21}$$

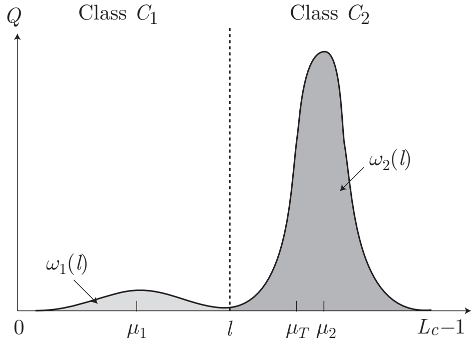


Figure 3.8. Histogram divided into two classes.

$$\mu_2(l) = \sum_{i=l+1}^{L_c} ip_i^{C^2} = \sum_{i=l+1}^{L_c} i\frac{p_i}{\omega_2(l)} = \frac{\mu_T - \mu(l)}{1-\omega_1(l)}, \tag{3.22}$$

where $\mu(l) = \sum_{i=1}^{l} ip_i$ is the first-order cumulative moment of the histogram up to the level l.

It is easy to verify that the following relations stand for any choice of the threshold l:

$$\omega_1(l)\mu_1(l) + \omega_2(l)\mu_2(l) = \mu_T, \quad \forall l \tag{3.23}$$

$$\omega_1(l) + \omega_2(l) = 1, \quad \forall l. \tag{3.24}$$

The class variances are given by

$$\sigma_1^2(l) = \sum_{i=1}^{l}(i-\mu_1(l))^2 p_i^{C_1} = \sum_{i=1}^{l}(i-\mu_1(l))^2 p_i/\omega_1(l), \tag{3.25}$$

$$\sigma_2^2(l) = \sum_{i=l+1}^{L_c}(i-\mu_2(l))^2 p_i^{C_2} = \sum_{i=l+1}^{L_c}(i-\mu_2(l))^2 p_i/\omega_2(l). \tag{3.26}$$

In order to evaluate the "goodness" of a threshold l, some discriminant criterion measures (or measures of class separability) used in discriminant analysis are introduced [84]:

$$\lambda = \sigma_B^2/\sigma_W^2, \quad \kappa = \sigma_T^2/\sigma_W^2, \quad \eta = \sigma_B^2/\sigma_T^2, \tag{3.27}$$

where

$$\begin{aligned} \sigma_W^2(l) &= \omega_1(l)\sigma_1^2(l) + \omega_2(l)\sigma_2^2(l), && (3.28)\\ \sigma_B^2(l) &= \omega_1(l)(\mu_1(l)-\mu_T)^2 + \omega_2(l)(\mu_2(l)-\mu_T)^2 \\ &= \omega_1(l)\omega_2(l)(\mu_2(l)-\mu_1(l))^2, && (3.29) \end{aligned}$$

(due to (3.23)) and

$$\sigma_T^2 = \sum_{i=1}^{L_c}(i-\mu_T)^2 p_i \tag{3.30}$$

are the within-class variance, the between-class variance, and the total variance, respectively. The class separability problem becomes an optimization problem to search for a threshold l

that maximizes one of the criterion measures in Equation (3.27). However, the problems of maximizing λ, κ, and η for l are equivalent, since the following basic relation always holds:

$$\sigma_W^2(l) + \sigma_B^2(l) = \sigma_T^2. \tag{3.31}$$

For example, both κ and η can be expressed in terms of λ: $\kappa = \lambda + 1$ and $\eta = \lambda/(\lambda + 1)$. It is noted that σ_W^2 is based on second-order statistics (class variances), while σ_B^2 is based on first-order statistics (class means). Therefore, η is the simplest measure with respect to l and it is adopted as the criterion measure to evaluate the "goodness" of the threshold. The optimal threshold l^* that maximizes η, or equivalently σ_B^2, is selected in the following sequential search by using simple cumulative quantities ($\omega_1(l)$ and $\mu(l)$) from Equations (3.21)-(3.24):

$$\sigma_B^2(l) = \frac{\left(\mu_T\omega_1(l) - \mu(l)\right)^2}{\omega_1(l)\left(1 - \omega_1(l)\right)} \tag{3.32}$$

The optimal threshold l^* is chosen so that the between-class variance σ_B^2 is maximized, that is:

$$l^* = \arg\max_l \left\{\sigma_B^2(l),\right\} \quad 1 \leq l \leq L_c. \tag{3.33}$$

Otsu multi-Level thresholding

The previous bi-level approach can be extended to multi-level thresholding as follows. Assuming that there are $N-1$ thresholds, $\{l_1, l_2, \ldots, l_{N-1}\}$, which divide the histogram into N classes: C_1 for $[1, \ldots, l_1]$, C_2 for $[l_1 + 1, \ldots, l_2]$, ..., C_ϵ for $[l_{\epsilon-1} + 1, \ldots, l_\epsilon]$ and C_N for $[l_{N-1} + 1, \ldots, L_c]$. The optimal thresholds $l_1^*, l_2^*, \ldots, l_{N-1}^*$ are chosen by maximizing σ_B^2 as follows:

$$\{l_1^*, l_2^*, \ldots, l_{N-1}^*\} = \arg\max_{l_1, l_2, \ldots, l_{N-1}} \left\{\sigma_B^2(l_1, l_2, \ldots, l_{N-1})\right\}, \tag{3.34}$$

for $1 \leq l_1 < l_2 \cdots l_{N-1} < L_c$, and

$$\sigma_B^2(l_1, l_2, \ldots, l_{N-1}) = \sum_{\epsilon=1}^{N} \omega_\epsilon(\mu_\epsilon - \mu_T)^2, \tag{3.35}$$

with

$$\omega_\epsilon = \sum_{i \in C_\epsilon} p_i, \tag{3.36}$$

and

$$\mu_\epsilon = \sum_{i \in C_\epsilon} i\frac{p_i}{\omega_\epsilon}. \tag{3.37}$$

The ω_ϵ in Eq.(3.36) is regarded as the zeroth-order cumulative moment of the ϵ-th class C_ϵ, and the numerator in the definition of μ_ϵ is regarded as the first-order cumulative moment of C_ϵ, that is

$$\mu(\epsilon) = \sum_{i \in C_\epsilon} ip_i. \tag{3.38}$$

Regardless of the number of classes being considered during the thresholding process, the sum of the cumulative probability functions of N classes equals one ($\sum_{\epsilon=1}^{N} \omega_\epsilon = 1$) and the total mean is equal to the sum of the means of N classes weighted by their cumulative probabilities:

$$\mu_T = \sum_{\epsilon=1}^{N} \omega_\epsilon\mu_\epsilon. \tag{3.39}$$

Thus, the interclass variance in Eq.(3.35) can be rewritten as

$$\sigma_B^2(l_1, l_2, \ldots, l_{N-1}) = \sum_{\epsilon=1}^{N} \omega_\epsilon \mu_\epsilon^2 - \mu_T^2. \tag{3.40}$$

Because the second term in Eq.(3.40) is independent from the choice of the thresholds, the optimal thresholds can be chosen by maximizing a modified between class variance $\sigma_B'^2$:

$$\{l_1^*, l_2^*, \ldots, l_{N-1}^*\} = \arg \max_{l_1, l_2, \ldots, l_{N-1}} \{\sigma_B'^2\}, \tag{3.41}$$

where σ_B' is defined as the summation term on the right-hand side of Eq.(3.40):

$$\sigma_B'^2 = \sum_{\epsilon=1}^{N} \omega_\epsilon \mu_\epsilon^2. \tag{3.42}$$

3.2.4 Efficient implementation

Liao *et al.* [85] proposed a faster algorithm based on the recursive calculation of $\sigma_B'^2$. Let us define the look-up tables for the $u - v$ interval:

$$\mathcal{P}(u, v) = \sum_{i=u}^{v} p_i, \tag{3.43}$$

$$\mathcal{S}(u, v) = \sum_{i=u}^{v} i p_i. \tag{3.44}$$

For index $u = 1$, equations (3.43) and (3.44) can be rewritten as

$$\mathcal{P}(1, v+1) = \mathcal{P}(1, v) + p_{v+1}, \tag{3.45}$$

and $\mathcal{P}(1, 0) = 0$.

$$\mathcal{S}(1, v+1) = \mathcal{S}(1, v) + (v+1) p_{v+1}, \tag{3.46}$$

and $\mathcal{S}(1, 0) = 0$.

From equations (3.45) and (3.46), it follows that

$$\mathcal{P}(u, v) = \mathcal{P}(1, v) - \mathcal{P}(1, u-1), \tag{3.47}$$

and

$$\mathcal{S}(u, v) = \mathcal{S}(1, v) - \mathcal{S}(1, u-1). \tag{3.48}$$

Now, the modified between-class variance $\sigma_B'^2$ can be rewritten as

$$\begin{aligned} \sigma_B'^2 &= \mathcal{G}(1, l_1) + \mathcal{G}(l_1 + 1, l_2) + \\ \ldots &+ \mathcal{G}(l_{N-1} + 1, L_c), \end{aligned} \tag{3.49}$$

where the modified between-class variance of class C_ϵ is defined as

$$\mathcal{G}(l_{\epsilon-1} + 1, l_\epsilon) = \frac{\mathcal{S}(l_{\epsilon-1} + 1, l_\epsilon)^2}{\mathcal{P}(l_{\epsilon-1} + 1, l_\epsilon)}. \tag{3.50}$$

The search range for the maximal $\sigma_B'^2$ is $1 \leq l_1 \leq L_c - N + 1, l_1 + 1 \leq l_2 \leq L_c - N + 2, \ldots, l_{N-1} + 1 \leq l_{N-1} \leq L_c - 1$.

Once the thresholds have been calculated, the separation masks can be computed as explained in the next subsection.

3.2.5 Estimation of the sources

The optimum thresholds resulting from the Otsu algorithm are the ones that maximize the between-class variance, which was shown to be a measure of class separability. Time-frequency masking is therefore used for estimating the sources directly from the calculated thresholds, without the need for considering any further information regarding the mixing matrix or any other prior information. However, note that the number of classes in the histogram has been denoted N, the same as the number of sources assumed in the mixing model. In fact, the classes that we want to separate from the weighted histogram are the spatial distributions generated by sparse sources in the time-frequency domain. Unfortunately, the thresholding algorithm needs to know a priori the number of classes to be separated. Nevertheless, alternative methods for source counting in BSS [74][86] and threshold number selection [87] can be found in the literature, although they have not been considered in the development of this thesis.

Given the calculated thresholds l_ϵ^*, the final pan map values that are used to divide the mixture spectrogram are

$$Th_\epsilon = z_i|_{i=l_\epsilon^*} \quad \epsilon = 1, \ldots, N-1. \tag{3.51}$$

The binary masks $\mathbf{M}_n$ defined by the thresholds are:

$$M_n(k,r) = \begin{cases} 1 & \text{if} \quad Th_{n-1} < P(k,r) \le Th_n \\ 0 & \quad \text{elsewhere} \end{cases} \qquad \forall (k,r). \tag{3.52}$$

with $n = 1, \ldots, N$, being N the number of sources, $Th_0 = -P_{\max}$ and $Th_N = P_{\max}$. The estimates of the images of the sources in each of the observation channels can be directly obtained by applying the separation masks to the STFT of the mixtures:

$$\mathbf{Y}_{mn}(k,r) = \hat{\mathbf{S}}_{mn}(k,r) = \mathbf{M}_n(k,r) \circ \mathbf{X}_m(k,r), \quad m = 1, 2. \tag{3.53}$$

3.2.6 Computational Complexity

$\sigma_B'^2$	$\mathcal{P}$ Table	$\mathcal{S}$ Table	$\mathcal{G}$ Table	$\sum_{\epsilon=1}^{N} \mathcal{G}(l_{\epsilon-1}+1, l_\epsilon)$
addition	L_c	L_c		$N-1$
substraction	$L_c(L_c-1)/2$	$L_c(L_c-1)/2$		
multiplication		L_c		
division			$L_c(Lc+1)$	
direct index				N
combination		1		$(L_c-N+1)^{N-1}$

	Total Computation
addition	$(N-1)(L_c-N+1)^{N-1}2L_c$
substraction	$L_c(L_c-1)$
multiplication	L_c
division	$L_c(L_c+1)$
direct index	$N(L_c-N+1)^{N-1}$

Table 3.1. Number of operations needed in the efficient implementation of the multi-level thresholding algorithm.

Table 3.1 shows the number of operations that must be computed in the efficient implementation of the thresholding algorithm proposed by Liao *et al.* It can be observed that the number of operations depends on the number of histogram bins L_c and the number of classes considered N. In [88], two different histograms where independently considered in order to further reduce the computational complexity of the method, one for left panned sources and another one for right panned sources. However, our experience in the use of the algorithm has shown that more accurate results are obtained with a single pan map histogram that takes into account the whole pan map distribution.

Computation Time

The times required by the thresholding algorithm for deciding the separation areas as a function of the number of sources N and the number of histogram bins L_c are shown in Table 3.2. The processor used for measuring these times was an *Intel Core2* 2.0 GHz running Matlab 7. Note that the number of sources considered has the greatest influence on the computation time. The number of histogram bins L_c can be modified depending on the computational power of the processor, which is another advantage for real-time implementations.

Computing Times (ms)						
L_c	10	20	100	200	300	500
$N = 2$	1.2	1.2	1.8	3.3	7.7	19.0
$N = 3$	1.5	1.5	2.0	4.4	10.1	24.4
$N = 4$	1.6	1.6	9.2	52.2	164.2	743.4

Table 3.2. Computing times for different number of histogram bins L_c and number of sources N.

3.2.7 Performance Evaluation

In this section, the performance achieved by the proposed approach is studied using the objective measures described in Section 2.8. Several aspects are considered: the effect of the weighting functions on the final separation masks, the capability of the method to separate sources that are very close together and the performance achieved over the test data set used in the *Signal Separation Evaluation Campaign 2008* (*SiSEC 2008*)[68].

Effect of Histogram Weighting

Several possibilities to obtain a weighted histogram were examined in Section 3.2.2. In order to show the effect of the weighting functions on the calculated thresholds that lead to the final separation masks, we consider an example mixture consisting of 4 male speech sources of length 10 s ($f_s = 16$ kHz). The STFT of the mixtures was computed using Hann windows, with 1024 samples of length and 50% overlap. The different weighting functions are denoted as: Non-Weighted (NW), Amplitude-weigthing (AW), DUET-weighting (DW), 1/ log-weighting (LW), Exp-weighting (EW).

Figure 3.9 shows the histograms obtained with different weighting functions (with default parameters) and the thresholds calculated by the segmentation algorithm. Table 3.3 shows the different objective performance measures (SDR, ISR, SIR and SAR) obtained for each weighting type and for each estimated source. It can be observed how, although the shape of the histogram is substantially changed by applying the different weighting functions, the thresholds do not vary significantly. These small changes are also reflected on Table 3.3, where only slight changes between the different cases can be appreciated in all the performance measures. This fact shows how the proposed approach is quite robust to changes in the histogram shape, being able to decide correct source regions for a wide range of histogram shapes, given that the sources are sufficiently disjoint in the time-frequency domain. The results provided by the *ideal binary mask* (IBM) are here provided for comparison purposes and should be understood as an upper limit in the separation quality using time-frequency masking.

A more complete evaluation was conducted by using a larger set of stereo mixtures with $N = 3$ and $N = 4$ speech sources extracted from the development data set provided in the in SiSEC 2008[1]. The sources are audio files of 10 s length, sampled at 16 kHz and quantized with 16 bits. They were randomly selected from the available male and female speech signals, and 15 source combinations were randomly mixed down using 10 different mixing configurations. Therefore, 300 different stereo mixtures were generated (150 mixtures with 3 sources and 150 mixtures with 4 sources). As each mixture was separated 5 times using different weighting functions, this makes a total of 1500 separation experiments. As in the previous example, the STFT processing was carried out using half overlapping Hann windows with 1024 samples of length. Table 3.4 shows the average results over all the sources. Besides the objective performance measures, the percentage of experiments in which correct source regions were detected (detection rate, DR) is given as an additional measure of robustness. We consider that a source region has been correctly detected if the real mixing parameters of the sources lie within separate thresholded areas.

In order to compare the results of speech mixtures with those obtained with correlated music mixtures, a similar evaluation with examples containing 3 instruments was also conducted using the music source signals of the SASSEC and SiSEC data sets. For each example, 10 different mixing configurations were randomly chosen and separation was carried out with the same 5 previously considered weighting functions. The results of this evaluation are in Table 3.5.

The results of these experiments can be summarized into the following conclusions:

- The number of sources has a big influence on the separation performance. Between $N = 3$ and $N = 4$, there is a decrease in SDR and SAR of approximately 3.5 dB and 3 dB, respectively. The decrease in SIR is around 4 dB, while the decrease in ISR is approximately 5 dB.

- There are only slight changes in the separation performance for different weighting functions. This is in concordance with the results of the example in Table 3.3. The LW weighting seems to provide the better performance, but the small changes observed may have little influence on the perceived subjective quality of the sources.

- The DR for all the cases is very high. This result confirms that the thresholding algorithm performs perfectly well in almost every situation, although some errors are found when the weighting is related to the energy of the mixture channels, since low energy sources can be hidden by others with greater energy. Although equal mixing configurations were avoided in the random mix-down, a deeper analysis of the results showed that detection errors were only produced under extremely close sources.

[1] This development data set was also the test data set of the first Stereo Audio Source Separation Evaluation Campaign (SASSEC)[67]

- The results considering speech mixtures are better than those with correlated music mixtures. This is probably due to the fact that mixtures of correlated sources share more spectral content than speech, resulting in a less degree of disjointness, and therefore, they affect the demixing performance of time-frequency masking.

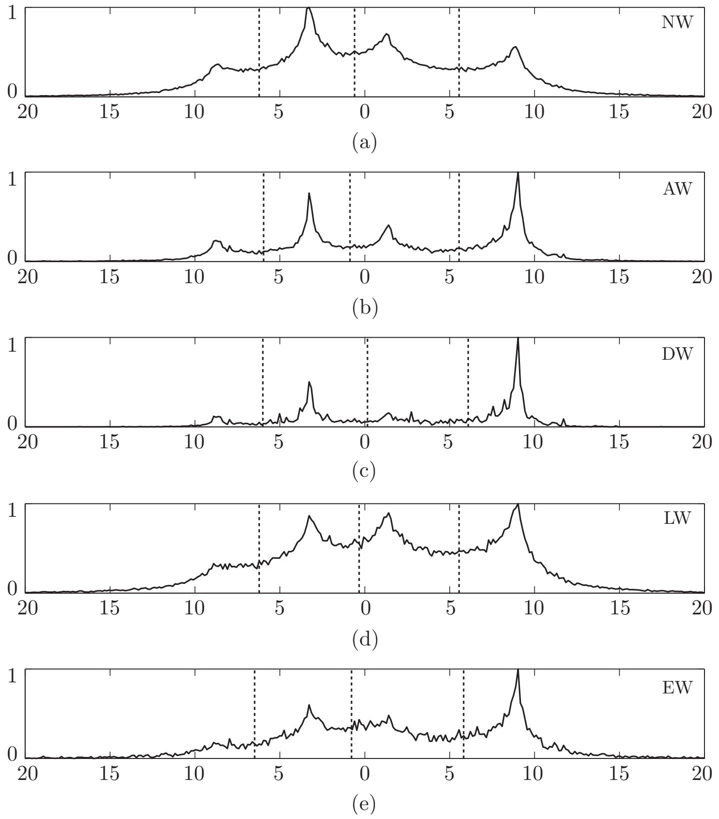


Figure 3.9. Weighted histograms and thresholds obtained for a mixture of 4 speech sources. (a) Non-weighted histogram. (b) Amplitude-weighted histogram. (c) DUET-weighted histogram. (d) 1/ log-weighted histogram. (e) Exp-weighted histogram.

Effect of Source Proximity

As seen in Section 3.2.1, the mixing parameters of the sources have a big influence on the histogram shape due to the increased deviation in the estimated values when there is overlap between the sources in the time-frequency domain. The influence on the segmentation process and the separation performance is studied considering the example shown in Figure 3.10. Three speech sources are repeatedly mixed with a decreasing panoramic distance $\Delta\phi = \phi_n - \phi_{n-1}$, each time resulting in a histogram (NW) with closer peaks. In the same figure, the thresholds provided by the MuLeTS algorithm are depicted as dotted vertical lines. Notice how the peaks remain within the different regions defined by the thresholds, showing that a correct source detection is possible even in the case when the sources have been closely mixed. The results for the separation performance measures are given in Table 3.6. Although there are no big differences, the best case is the one in which the sources have a panoramic separation of $\Delta\phi = 0.2$. However, it is worth to observe that the worst case is that in which the panoramic distance of the sources is very high. This result indicates that negative effects appear in extreme cases when the sources are very close together or very distant regarding their panoramic mixing configuration.

Non-weighted (NW)				
	s_1	s_2	s_3	s_4
SDR	2.9	8.5	8.0	7.9
ISR	12.7	20.3	12.8	13.1
SIR	8.3	14.6	20.4	21.4
SAR	3.4	9.4	8.4	7.9

1/ log-weighted (LW)				
	s_1	s_2	s_3	s_4
SDR	2.9	8.6	8.0	7.9
ISR	12.7	20.2	12.8	13.1
SIR	8.3	14.9	20.4	21.4
SAR	3.4	9.5	8.4	7.9

Magnitude-weighted (MW)				
	s_1	s_2	s_3	s_4
SDR	3.0	8.7	8.1	7.9
ISR	12.6	19.9	12.6	13.1
SIR	8.0	14.3	20.2	21.4
SAR	3.6	9.7	8.5	8.0

Exp-weighted (EW)				
	s_1	s_2	s_3	s_4
SDR	2.8	8.5	8.2	7.5
ISR	12.5	20.4	12.6	12.9
SIR	8.6	15.0	20.6	21.7
SAR	3.2	9.4	8.7	7.7

DUET-weighted (DW)				
	s_1	s_2	s_3	s_4
SDR	3.0	8.6	8.3	7.5
ISR	12.6	20.2	12.5	13.0
SIR	8.0	14.8	19.9	21.0
SAR	3.6	9.3	8.7	7.8

Ideal Binary Mask				
	s_1	s_2	s_3	s_4
SDR	7.7	11.6	10.4	10.0
ISR	14.1	24.4	20.0	19.3
SIR	21.4	20.2	22.8	22.4
SAR	7.7	12.2	10.7	10.2

Table 3.3. Performance evaluation measures for different histogram weightings in an example mixture of 4 speech sources.

	$N = 3$					$N = 4$				
	DR (%)	SDR	ISR	SIR	SAR	DR (%)	SDR	ISR	SIR	SAR
NW	99.8	9.8	19.3	19.3	10.1	98.9	6.4	14.2	15.1	6.9
AW	99.5	10.1	18.9	19.3	10.3	95.1	6.6	13.9	14.9	7.1
DW	98.1	9.9	18.9	19.1	10.0	94.9	6.4	14.0	14.9	7.2
LW	99.9	10.2	19.1	19.5	10.3	99.1	6.6	14.1	15.5	7.4
EW	99.3	9.8	19.4	19.9	10.0	99.0	6.3	14.3	15.3	7.0

Table 3.4. Average performance evaluation measures for different weighting functions with speech mixtures

	$N = 3$				
	DR (%)	SDR	ISR	SIR	SAR
NW	98.9	6.5	15.3	15	7.5
AW	97.5	6.6	16	15.2	7.4
DW	96.4	6.6	15.8	15.3	7.1
LW	99.1	6.8	16.1	16	7.5
EW	98.8	6.5	16	15.4	7.4

Table 3.5. Average performance evaluation measures for different weighting functions with music mixtures

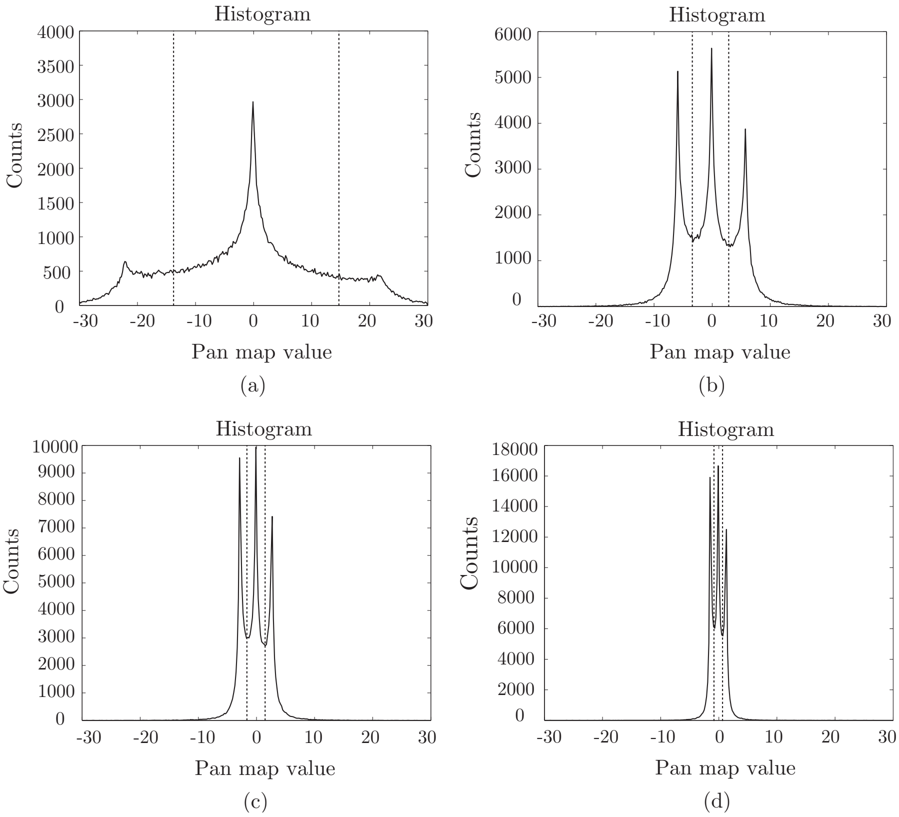


Figure 3.10. Mixture of 3 speech sources and separation thresholds. (a) $\Delta\phi = 0.45$ ($\phi_1 = 0.05$, $\phi_2 = 0$, $\phi_3 = 0.95$). (b) $\Delta\phi = 0.2$ ($\phi_1 = 0.3$, $\phi_2 = 0.5$, $\phi_3 = 0.7$). (c) $\Delta\phi = 0.1$ ($\phi_1 = 0.4$, $\phi_2 = 0.5$, $\phi_3 = 0.6$). (d) $\Delta\phi = 0.05$ ($\phi_1 = 0.45$, $\phi_2 = 0.5$, $\phi_3 = 0.55$)

	$\Delta\phi = 0.45$	$\Delta\phi = 0.2$	$\Delta\phi = 0.1$	$\Delta\phi = 0.05$
SDR	9.4	10.0	9.7	9.7
ISR	18.6	19.5	19.3	19.2
SIR	18.0	18.1	18.2	18.1
SAR	10.2	10.8	10.6	10.6

Table 3.6. Average performance evaluation measures for different mixing configurations with decreasing panoramic distance.

Results from SiSEC 2008

The MuLeTS algorithm was also a participant in SiSEC 2008. In the campaign, different separation algorithms were tested over the same data set. The average results for speech and music corresponding to this data set are shown in Table 3.7. The results for speech are quite similar to those of Table 3.5, but the results using music mixtures are considerably better. This difference is probably due to the fact that the campaign only considered two example mixtures with two different mixing configurations. The results provided by the IBM are given as well for comparison purposes. As expected, the IBM performance is also affected by the number of sources. The average performance for all the measures is around 2 dB lower than the IBM in the speech case with $N = 3$. When the number of sources is increased, the results are around

3.5 dB worse than the IBM for the SDR and SAR, but the SIR is more severely affected, with a difference of approximately 6 dB with respect to the performance achieved by the IBM. The average results for music are quite similar to those of speech with the same number of sources in SDR and SAR, but the ISR and SIR are considerably lower in the music case.

	Speech $N = 3$		Speech $N = 4$		Music $N = 3$	
	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM
SDR	9.6	11.4	5.9	9.3	9.15	11.5
ISR	19.2	21.3	13.8	18.4	16.3	20.3
SIR	20.6	22.7	14.9	20.8	17.2	21.3
SAR	10.1	11.8	6.3	9.5	10.0	13.3

Table 3.7. Average results obtained from the SiSEC 2008 Campaign.

3.3 Separation in Real Environments

In this section, we show that multi-level thresholding can be also exploited to achieve fast separation in real scenarios by identifying different angular areas wherein the speakers are located with a strong likelihood. For this purpose, a pair of closely spaced microphones are considered in the recording setup. Although the mixing process is assumed to be anechoic or delayed, a coherence-based selection of time-frequency points enables the method to perform separation with acceptable quality in rooms with moderate reverberation.

In contrast to the instantaneous case described in Section 3.2, amplitude differences are only used to improve the robustness against reverberation, and phase information is the key feature to separate sources that are angularly spaced in the real azimuth plane (as opposed to the panoramic space considered in the previous section). Although the basic multi-level thresholding clustering remains, the key feature used in this section is the estimation of the *Direction-Of-Arrival* (DOA) by means of the observed time differences between the microphones. This working principle will be used and recalled again in Chapter 7.

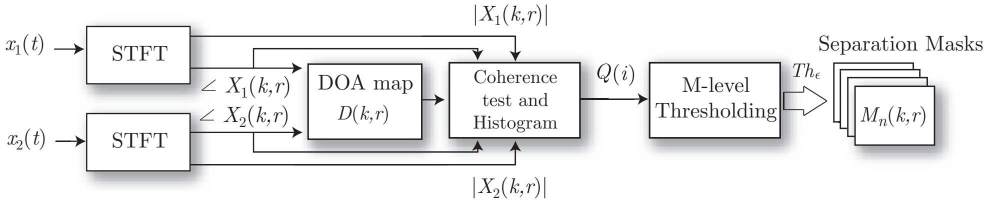


Figure 3.11. Processing scheme for the separation of real mixtures.

The proposed separation approach, depicted in Figure 3.11, can be summarized in the next steps:

1. **DOA map** (Section 3.3.1). The input channels are transformed into the STFT domain. The phase information in each time-frequency point is used for estimating the DOA of a given frequency in the current analysis window. Similarly to the pan map used in the instantaneous approach, a DOA map is formed with the value of the estimated DOA for

each time-frequency point.

2. **Coherence Test** (Section 3.3.2). A coherence test is performed with the aim of identifying reliable DOA estimates and increase the robustness against reverberation.

3. **Weighted histogram** (Section 3.3.3). The STFT magnitudes of both input channels are used for constructing an amplitude-weighted histogram from the selected values of the DOA map. This histogram is the input of the multi-level thresholding algorithm.

4. **Multi-level Thresholding** (Section 3.3.4). The histogram obtained in the previous step is processed for calculating a set of thresholds that maximize the between-class variance of the time-frequency points, according to their estimated DOAs. These thresholds are used for segmenting the DOA map into the final separation masks.

In the next subsections we will describe the above steps in detail, paying special attention to the main differences with respect to the instantaneous mixing approach. Finally, a complete set of examples is presented using real speech mixtures, considering different degrees of reverberation, different number of sources and a range of angular separations of speakers.

3.3.1 DOA map

Consider the anechoic model of Equation (2.12). The STFT representation of this model is given by

$$X_m(k,r) = \sum_{n=1}^{N} a_{mn} S_n(k,r) e^{-j\omega_k \delta_{mn}}, \quad m = 1,2, \tag{3.54}$$

where δ_{mn} is the time delay of the path from source n to sensor m. Throughout this section, the above simplified model will be considered.

The DOA map $\mathbf{D}(k,r)$ is calculated from the phase difference of the two observation channels. Figure 3.12 shows a pair of microphones capturing the signal arriving from a source s_n with angle θ. The inter-microphone distance is d. Assuming that the source is sufficiently distant to consider plane wavefront incidence, the path difference between the two microphones is $d\cos\theta$.

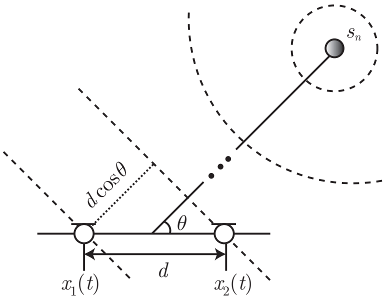


Figure 3.12. Two microphones capturing a signal from a source s_j with direction-of-arrival θ. The path difference between both sensors is $d\cos\theta$

Given the model of Equation 3.54, the relation between the phase difference between the two sensors and the incidence angle of the source for frequency ω_k and time frame r is given by the next equation:

$$\angle X_2(k,r) - \angle X_1(k,r) = \omega_k(\delta_{2n} - \delta_{1n}) = \omega_k \left(\frac{d\cos\theta}{c} \right), \tag{3.55}$$

where $\angle()$ denotes the phase of a complex number and c is the speed of sound ($\approx$ 340 m/s). The DOA θ is then directly given by the time difference $(\delta_{2n} - \delta_{1n})$, which is observed as a frequency-dependent phase difference between the two channels.

The DOA map represents the cosine of the direction-of-arrival from each frequency at each time window, and it is calculated from the STFT of the mixture channels as

$$D(k,r) = \frac{c}{\omega_k d} \angle \left(\frac{X_2(k,m)}{X_1(k,m)} \right) \quad \forall (k,r). \tag{3.56}$$

Note that frequencies above $\omega_k > \frac{\pi c}{d}$ will suffer from spatial aliasing [38]. Nevertheless, speech sources concentrate most of their energy below 5 kHz and small distances ($d \sim 3$ cm) easily satisfy this condition for the effective bandwidth of speech.

In Figure 3.13(a) and 3.13(b) the left channel and right channel spectrograms of a mixture of three speakers (with directions 135°, 90° and 20°) in an anechoic environment are represented. The amplitude relationship (in dB) between the two channels is represented in Figure 3.13(c), showing that very little information is provided when the microphones are very close to each other. In fact, when the inter-microphone distance verifies $d << R_s$, being R_s the distance between the source and the array, $a_{1j} \approx a_{2j}$ and it holds that the modulus $|X_1(k,m)|/|X_2(k,m)| \approx 1$. However, the DOA map extracted from the phase information of the mixtures (Figure 3.13.d) shows clear zones related to the activity of the four speakers in the STFT domain. It is also important to notice that the use of the phase difference has another advantage over the amplitude ratio: it avoids the division by zero when $|X_2(k,m)| = 0$.

The representation shown in Figure 3.13.(d) brings to the eye the DOA estimates for each frequency at different time frames. Under anechoic conditions, the different zones corresponding to each speaker are highly visible as different colored zones which support the presence of predominant speech signals. The histogram of the DOA map is depicted in Figure 3.14(a), which shows strong peaks at the values coinciding with the cosine of their DOA. The effect that reverberation has in this histogram is next described.

3.3.2 Coherence-Based Selection

If a single source in free-field is considered, sound from only one direction arrives at the microphones with a deterministic time difference and thus, the DOA of the source can be easily determined by the observed phase difference, as explained in the previous section. However, in complex listening situations, i.e., in the presence of other sound sources or room reflections, it often occurs that sound from different directions reaches the sensor array concurrently, resulting in erroneous DOA estimations. In fact, when the mixing is convolutive, multiple reflections of the source signal arrive at both microphones. These reflections, and the fact that the sources are not completely WDO, cause deviations in the DOA estimation, and histogram peaks corresponding to each speaker become blurred due to the stochastic nature of reverberation. Figure 3.14 shows the distributions obtained for the example mixture of three speakers under different room conditions. The room has been simulated as a "shoe box" shape, with all their surfaces having the same reflection coefficient ρ [89].

The robustness of the method against these complex environments can be improved by discarding time-frequency bins where other sources and reverberation are dominant. This is done by means of a *coherence-test*. In fact, from an auditory perspective, *Interaural Coherence*

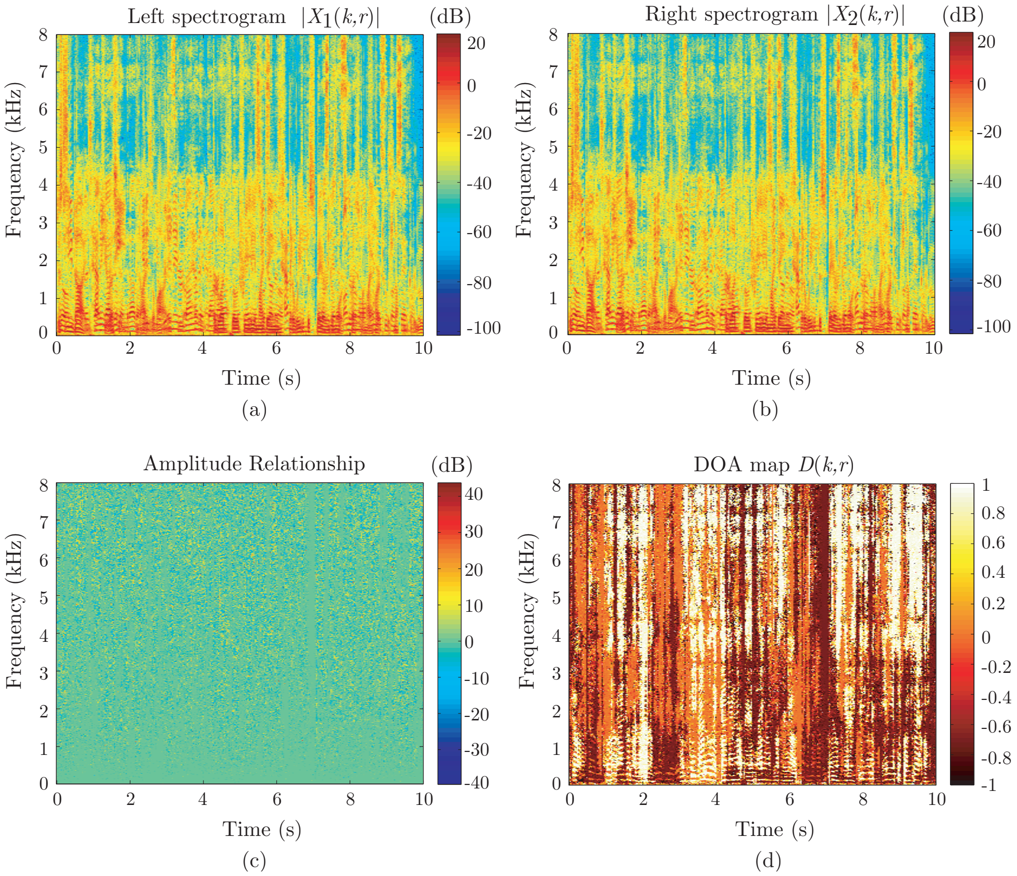
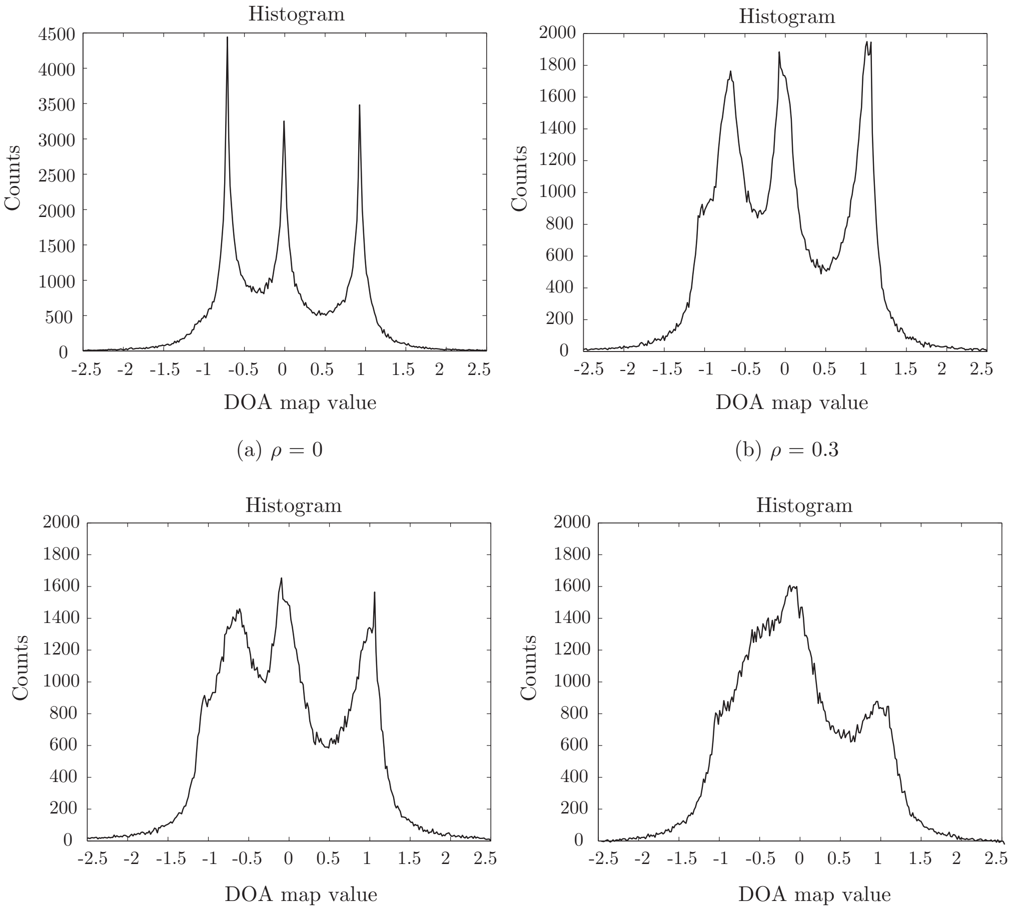


Figure 3.13. (a) Left observation channel spectrogram. (b) Right observation channel spectrogram. (c) Amplitude relationship between the two channels, calculated as $20\log_{10}(|X_1(k,m)|/|X_2(k,m)|)$. (d) DOA map of the mixture.

(IC) plays an important role on the localization of sound events in complex listening situations. Thus, the degree of "similarity" between the left and right ear entrance signals has been shown to provide a very important cue for dealing with the localization of sources in difficult environments. A measure for IC usually ranges between 0 and 1, where 0 means that two signals are coherent (signals are equal with possibly a different scaling and delay) and 1 means that the signals are independent [90]. For example, Faller proposed a successful localization system based on the time difference and IC estimated from the normalized cross-correlation function [91]. From an algebraic perspective, Mohan *et al.* proposed recently a *coherence test* in the time-frequency domain based on the identification of low-rank bins where the time-frequency covariance matrix of the input mixtures has effective low-rank (ideally 1 when a time-frequency point is dominated by only one source) [92]. Avendano, also propose the use of a *short-time coherence function* in the context of automatic up-mixing [93]. Next, we describe the IC measures used by Mohan and Avendano.

Mohan Coherence-Test

Recently, Mohan *et al.*[92] showed the improvement achieved in source localization by discarding low-SNR or higher-rank bins containing corrupt spatial-spectral estimates by means of a coherence test. In fact, the $M \times M$ time-varying time-frequency covariance matrix $\mathbf{R}(k,r)$ will have effective full rank if the number of active sources is greater than or equal to the number of



(a) $\rho = 0$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Figure 3.14. Histograms for a mixture of 3 speech sources with DOA 135°,90° and 20° in different room conditions considering several reflection coefficients. (a) $\rho = 0$ (anechoic). (b) $\rho = 0.3$. (c) $\rho = 0.5$. (d) $\rho = 0.7$.

sensors. In contrast, if only few sources are active, it will have low effective rank or be poorly conditioned. Speech non-stationarity leads to the use of the next estimation of the covariance matrix at time-frequency bin (k, r):

$$\hat{\mathbf{R}}(k,r) = \frac{1}{C_r} \sum_{v=l-C_r+1}^{l} \mathbf{X}(k,v)\mathbf{X}(k,v)^{\mathrm{H}}, \quad \forall (k,r) \tag{3.57}$$

where $\mathbf{X}(k,r) = [X_1(k,r), X_2(k,r)]^{\mathrm{T}}$, C_r is the number of averaged time-frames and $^{\mathrm{H}}$ denotes complex conjugation and transposition. The coherence function that allows to identify rank-1 time-frequency bins is given by:

$$\Phi(k,r) = \frac{|\mathrm{R}(k,r)_{12}|^2}{\mathrm{R}(k,r)_{11}\mathrm{R}(k,r)_{22}} \tag{3.58}$$

where $\mathrm{R}(k,r)_{mn}$ is the (m,n) entry of the estimated covariance matrix $\hat{\mathbf{R}}(k,r)$. Matrices close to be Rank-1 will show a value of $\Phi(k,r)$ close to unity. Therefore, this pre-selection involves estimating the covariance matrix at each time-frequency bin, computing $\Phi(k,r)$ and checking which time-frequency bins are above a defined threshold $\Phi_T \approx 1$, for example $\Phi_T = 0.9$.

Avendano Short-Time Coherence Function

Alternatively, Avendano proposed the use of the following short-time coherence function (STCF) [93]:

$$\Phi(k,r) = \frac{|\Phi_{12}(k,r)|}{\sqrt{\Phi_{11}(k,r)\Phi_{22}(k,r)}}, \quad \forall(k,r), \tag{3.59}$$

where the statistics $\Phi_{mm'}(k,r)$ are a practical way of computing the inter-channel correlation $\mathrm{E}\{X_m(k,r)X^*_{m'}(k,r)\}$, given by:

$$\Phi_{mm'}(k,r) = (1-\lambda)\Phi_{mm'}(k,r-1) + \lambda X_m(k,r)X^*_{m'}(k,r), \tag{3.60}$$

where * denotes complex conjugation. Due to the non-stationarity of speech, the forgetting factor $\lambda \in [0,1]$ is introduced to compute the cross-correlation between the observation channels over a block of time frames, which can be adjusted to balance current and previous estimations. The coherence function $\Phi(k,r)$ has values close to one in time-frequency regions where a source is present, and it is usually smaller when sounds from different directions overlap. Therefore, high-coherence time-frequency bins above a certain threshold Φ_T are selected as having reliable DOA estimates (again, for example $\Phi_T > 0.9$). This pre-selection is used to identify time-frequency points that are more likely to give a better estimation of the DOA, which results in a sharper histogram as it will be seen in the next subsection.

3.3.3 Histogram formation

In Section 3.2.2, the utility of weighted histograms was already discussed in the context of instantaneous mixing. In the case of real audio mixtures picked up by a microphone array, amplitude-weighted histograms not only provide a way for enhancing the sparse structure of audio sources in the time-frequency domain, but they provide a way of weighting the quality of DOA estimates according to the magnitude of time-frequency bins.

Let us explain why it is important to consider different weights for each DOA estimate[2] in the formation of the histogram. As commented in the previous subsections, the effect of a non-ideal scenario affects negatively the estimation of the DOA at each time-frequency point. The reason is the phase distortion caused in the observation channels due to other sources contributions and the reverberation. If the phase is not very much altered in a given time-frequency bin, the estimated DOA calculated from the phase difference at that bin will be similar to the real one. In contrast, if the phase has been distorted by the other sources or the room reflections, the quality of the estimation will be poorer. This effect is graphically depicted in Figure 3.15. In this figure, it can be seen the addition of two phasors. One phasor represents the source of interest s_t, and the other, s_i, represents the total contribution of the interference (other sources and reflections). It can be observed that the phase deviation $\Delta\varphi$ caused by the interference is highly related to the magnitude of the observation channel. In Figure 3.15(a), the deviation produced by the total interference is significantly smaller than in Figure 3.15(b).

Considering the above relationship between magnitude and phase stability in a given time-frequency bin, a more accurate segmentation of the DOA map will be achieved if points with high magnitude are given a greater weight in the formation of the histogram. To this end, an amplitude-weighted histogram formed by using Equation (3.11), is an excellent option for improving the accuracy of the last segmentation step.

[2] Strictly speaking, the real DOA is the θ, but the estimates are of its cosine $\cos\theta$

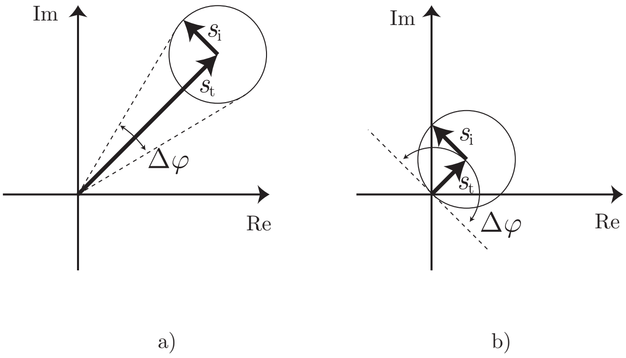


Figure 3.15. a) Possible phase deviation $\Delta\varphi$ when the source of interest $s_{\rm t}$ has big amplitude in comparison to the total interference $s_{\rm i}$. b) Possible phase deviation $\Delta\varphi$ when the source of interest $s_{\rm t}$ has low amplitude in comparison to the total interference $s_{\rm i}$

Figure 3.16 shows different normalized histograms for the example mixture of 3 sources and different amounts of reverberation simulated with different wall reflection factors ρ. It can be clearly observed that the peaks corresponding to the presence of the three speakers are clearly distinguishable in the anechoic case, but they broaden and overlap among themselves as soon as reverberation increases (dotted lines). However, note that the peaky structure is highly enhanced when the histogram is computed from high-coherence time-frequency bins selected by means of a coherence test (solid line).

3.3.4 Multi-Level Thresholding and Source Estimation

The thresholding procedure is carried out without alterations with respect to the separation of instantaneous mixtures. This is a great advantage for a complete source separation system dealing with both instantaneous and delayed/convolutive mixtures, as the same clustering block can be used for performing the two tasks. Moreover, all the advantages discussed in Section 3.2.3 also apply for this situation, which justify the use of this segmentation method as a powerful tool for audio separation.

The binary masks $\mathbf{M}_n$ are then given by:

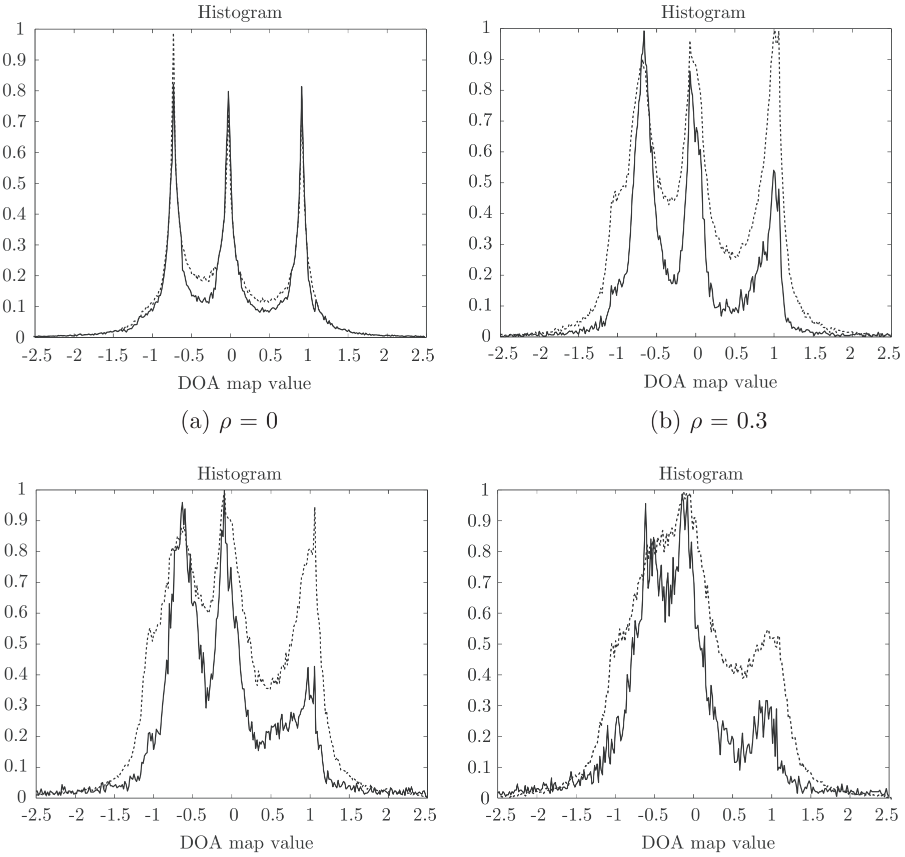
$$M_n(k,r) = \begin{cases} 1 & \text{if} \quad Th_{n-1} < D(k,r) \leq Th_n \\ 0 & \quad \text{elsewhere} \end{cases} \qquad \forall (k,r). \tag{3.61}$$

with $n = 1, \ldots, N$, being N the number of sources, $Th_0 = -D_{\max}$ and $Th_N = D_{\max}$, where $D_{\max}$ is a boundary value for the estimates considered in the formation of the histogram ($D_{\max} \sim 1$). Finally, the images of the sources are estimated in each of the observation channels by applying the separation masks to the STFT of the mixtures:

$$\mathbf{Y}_{mn}(k,r) = \mathbf{M}_n(k,r) \circ \mathbf{X}_m(k,r), \quad m = 1, 2. \tag{3.62}$$

Dynamic Separation

Note that a dynamic approach of the described processing is straightforward. In the previous subsections, a DOA map was introduced and several thresholds were calculated considering a histogram that covered estimates over a set of time frames. However, the time length of the DOA map can be modified in real-time applications, allowing to re-calculate the separating



(a) $\rho = 0$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Figure 3.16. Weighted histograms (normalized) for a mixture of 3 sources without preselection (dotted line) and with pre-selection (solid line) for different wall reflection factor ρ.

thresholds in a dynamic way. This approach opens the possibility of separating moving sources. In the following subsection, several experiments will be carried out to evaluate the performance of the algorithm under different mixing situations and recording setups, including the case of moving speakers.

3.3.5 Performance Evaluation

In this section, we evaluate the performance of the MuLeTS method under different mixing situations. The validity of the proposed approach is shown by means of a set of experiments designed with the aim of:

- discussing the performance of the algorithm when the speakers are very close to each other or vary their angular separation,
- analyze the performance of the algorithm in complex environments with different degrees of reverberation,
- showing the accuracy of the separation thresholds with different number of speakers,
- evaluating how a dynamic approach derived from the presented separation framework is useful for the separation of moving sources,

- testing the algorithm under real-world situations, with mixtures of speech recorded in a real room.

The processing parameters and the input signals are as follows:

- Audio sources (male and female speech) of 10 s duration, sampled at 16 kHz, 16 bits. As in Section 3.2.7, they correspond to the development data set of SiSEC 2008.
- STFT: Hann windows of 1024 samples of length, with hop size 512 samples.
- Algorithm parameters: $L_c = 200$, STCF with $\Phi_T = 0.95$, $\lambda = 0.6$, $d = 0.02$.

Angular Separation of Speakers

The capability of the algorithm to decide the angular areas wherein the speakers are located is graphically shown in Figure 3.17, where an anechoic scenario and 3 speech sources have been considered in order to better visualize how the method works (later experiments will consider room reflections and more speakers). The separation thresholds (marked as dotted vertical lines) are accurately positioned between the source peaks, segmenting the azimuth plane into angular areas that denote the presence of each speaker. Notice how the segmentation is successfully performed even in the case when the angular separation between the sources is very small (10 degrees).

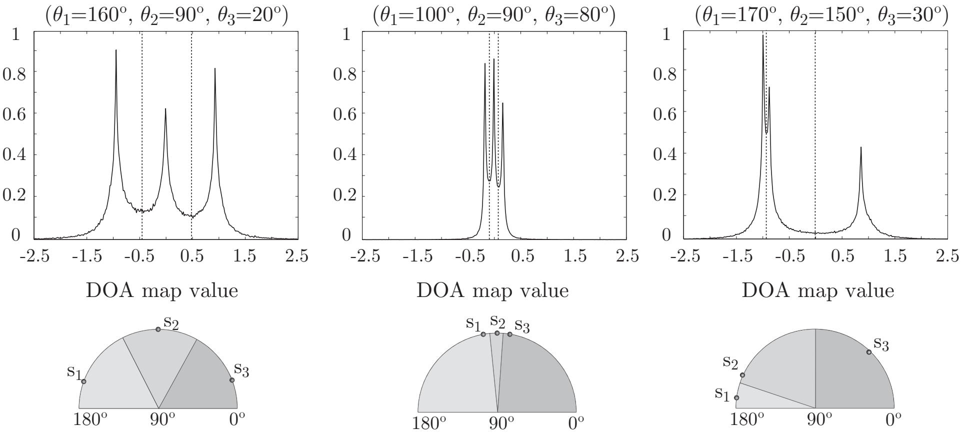


Figure 3.17. Histograms and separation thresholds for different source arrangements with $\rho = 0$ (anechoic). The angular segmentation of the azimuth plane for each case is depicted at the bottom of each graph, showing how the sources are effectively clustered in different angular sections even when they are close to each other.

Robustness Against Reverberation

The degradation introduced by reverberation with respect to the anechoic case is shown in Figure 3.18. The peaks that appear in the histogram are more distorted when the wall reflection factor ρ is increased, making more difficult the detection of the speakers in the presence of room reflections. Note that, despite the observed degradation, only slight differences between the angular source areas are produced. It is important to remark that the histograms were constructed considering only high-coherence bins, so the effect that pre-selection has on improving the robustness against reverberation is here clearly appreciated.

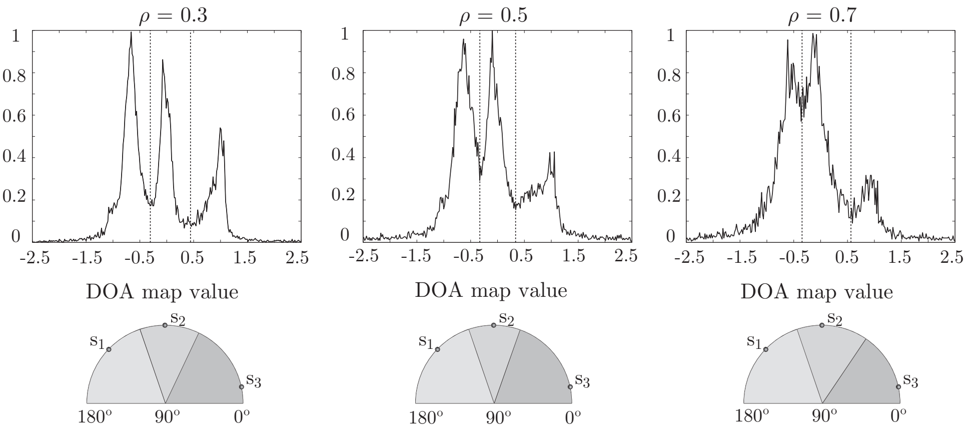


Figure 3.18. Histograms and separation thresholds for different degrees of reverberation as a function of the wall reflection factor ρ. The speakers are at locations $((\theta_1 = 135°, \theta_2 = 90°, \theta_3 = 20°))$. Note that despite the degradation of the histograms with reverberation, only slight differences between the angular source areas are produced.

Number of Speakers

The number of speakers in the mixture affects the performance of the algorithm. Nevertheless, Figure 3.19 shows how the thresholds correctly separate the source distributions even when 4 sources are present and room reflections are considered. When $N = 5$, two sources are clustered into the same region, but the other sources are correctly detected. Therefore, even in the case when many sources are present, the algorithm is still able to provide correct results for most of the sources.

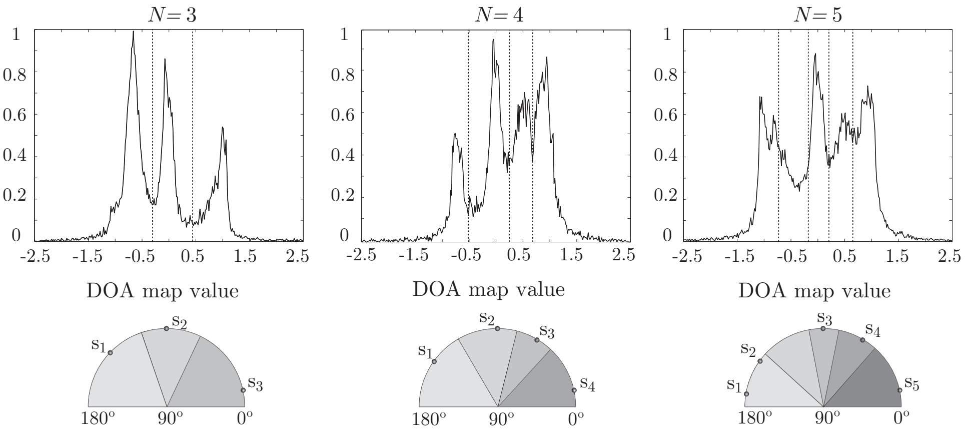


Figure 3.19. Histograms and separation thresholds for different number of speakers in a room with moderate reverberation $\rho = 0.3$. a) $(\theta_1 = 135°, \theta_2 = 90°, \theta_3 = 20°)$. b)$(\theta_1 = 135°, \theta_2 = 90°, \theta_3 = 60°, \theta_4 = 20°)$. c) $(\theta_1 = 170°, \theta_2 = 135°, \theta_3 = 90°, \theta_4 = 60°, \theta_5 = 20°)$.

Moving Sources

The dynamic approach of Section 3.3.4 has been considered in an example where there are two static sources and one moving source going from $\theta = 0°$ to $\theta = 150°$ in a time period of 4 s. The thresholds are calculated every 0.8 s and the separation masks are consequently updated taking into account the new histograms. Figure 3.20 shows the histograms obtained in each update step, together with the thresholds obtained as outputs of the segmentation process. Note that the thresholds are correctly updated in each step. At $t = 2.4$ s the moving source is at the same location as one of the static sources. Although the angular distributions are correctly identified, the overlapping sources can not be separated if they are at the same angular location. Obviously, when there are crossings between angular source positions, the separation masks are interchanged, thus, a tracking stage must be included in order to preserve the identity between each source and its corresponding mask. The solutions to these problems are out of the scope of this thesis.

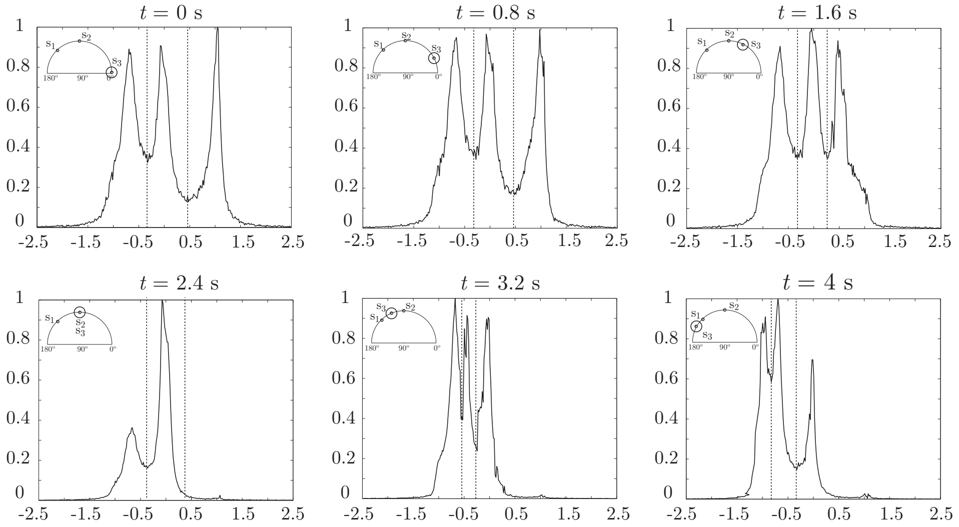


Figure 3.20. Histograms and separation thresholds for the moving source example at different time instants

Separation Performance

The objective performance measures for different separation examples with $N = 3$ and $N = 4$ sources are shown in Table 3.8 and Table 3.9, respectively. The values given correspond to the average results over all the sources for each separation experiment. The angular positions of the sources were ($\theta_1 = 135°, \theta_2 = 90°, \theta_3 = 20°$) in the 3 source example, and ($\theta_1 = 135°, \theta_2 = 90°, \theta_3 = 60°, \theta_3 = 20°$) in the 4 sources case. For comparison purposes, the performance measures achieved by the IBM in each experiment are also included. As expected, the performance of the algorithm is decreased when either the number of sources in the mixture or the degree of reverberation gets higher. This is not only true for the MuLeTS algorithm, but also for the case of the IBM, which also shows a degradation in the demixing performance due these two factors. Of all the performance measures, SAR and SDR are mostly affected by an increment in the number of sources, while reverberation has its greatest influence on the SIR and ISR measures.

Results from SiSEC 2008

Experiments using mixture signals recorded in a real office room have also been conducted in the context of SiSEC 2008. The room had a reverberation time of $\mathrm{RT}_{60} = 130$ ms and the experiments were carried out by considering mixtures of male speech sources and mixtures of female speech sources. Table 3.10 shows the average results for these experiments together with the

$N=3$								
	$\rho=0$		$\rho=0.3$		$\rho=0.5$		$\rho=0.7$	
	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM
SDR	6.4	9.5	2.6	9.2	1.9	8.9	0.9	8.2
ISR	15.1	20.0	10.1	18.9	7.8	18.2	5.6	17.4
SIR	15.3	22.0	10.7	20.6	8.3	19.7	2.8	18.3
SAR	6.9	9.7	5.6	9.6	4.6	9.3	4.4	8.5

Table 3.8. Average Objective Performance Measures for $N = 3$ sources.

$N=4$								
	$\rho=0$		$\rho=0.3$		$\rho=0.5$		$\rho=0.7$	
	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM
SDR	3.7	8.1	1.5	7.9	0.1	7.3	-0.1	6.9
ISR	11.1	17.8	7.3	17.2	5.3	16.5	3.4	15.4
SIR	10.6	20.3	5.8	19.5	3.5	18.3	1.7	17.1
SAR	3.7	8.1	3.4	8.1	2.3	7.8	2.1	6.5

Table 3.9. Average Objective Performance Measures for $N = 4$ sources.

results of the IBM. Note that the results with real mixtures are even better than those obtained in previous simulations with high reverberation. A complete description of the experiments, audio files and results of other algorithms can be found at http://www.irisa.fr/metiss/SiSEC08/SiSEC_underdetermined/test_eval.html. The measures shown in the table confirm that the method is able to separate real mixtures of speech with little computation time.

	$N=3$				$N=4$			
	Female		Male		Female		Male	
	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM	MuLeTS	IBM
SDR	2.4	12.1	2.2	9.6	2.2	8.4	2.1	7.6
ISR	6.5	21.8	6.8	18.3	6.6	16.0	6.1	15.5
SIR	6.4	22.6	5.9	20.2	6.5	17.4	6.6	17.5
SAR	7.8	12.8	5.2	9.9	5.5	9.0	4.5	7.8

Table 3.10. Average Objective Performance Measures for the results of SiSEC 2008.

3.4 Conclusion

This chapter has addressed the underdetermined source separation problem by applying some well-known techniques from image processing. From this perspective, the time-frequency representation of a computed feature is thought of as an image where different non-overlapping objects are present. Multi-level thresholding has been proposed as a useful segmentation technique that, maximizing the between class variance of the computed feature, provides efficiently a set of time-frequency masks that separate these objects into the different source estimates that conform the mixture. The chosen feature is a key aspect in the separation, since amplitude differences have been shown to be useful for the separation of instantaneous mixtures, while phase differences turn into DOA estimates that are used for the separation of real stereo mixtures with two-microphones.

In the instantaneous case, amplitude differences are analyzed in each time-frequency point to form a pan map. A weighted histogram of the pan map values is then used as the input of the thresholding algorithm, which gives as a result a set of thresholds that divide the spectrogram into the separation binary masks. A set of experiments were carried out for evaluating the performance of the algorithm under several mixing situations and source configurations, showing the validity of the approach to separate sources without the need for estimating the mixing matrix, searching for peaks or specifying initial clustering values.

The extension of the method for achieving separation in real environments has also been studied. Although the method is based on the anechoic mixing model, two features have been proposed to improve the robustness against reverberation: an interaural coherence test and a weighted histogram. These features allow to perform separation in realistic scenarios. The results obtained from the experiments confirm that separation of several sources with moderate reverberation is possible, although the demixing performance is severely affected by the number of sources and the degree of reverberation.

Post-Processing Separated Sources 4

4 Post-Processing Separated Sources

The sources obtained by separation methods usually have residuals from other sources that severely degrade their quality, especially in the underdetermined case. Most sound source separation methods do not specify any post-processing for eliminating these residuals and the isolation of the sources is restricted to the performance achieved by an specific algorithm. In this chapter, two techniques for improving source isolation are proposed. The aim of these techniques is to identify residuals from interfering sources and to eliminate them by using a time-frequency masking approach.

4.1 Introduction

Source separation methods are aimed at recovering the sources observed from a set of mixture signals. Separation is almost perfect in the determined instantaneous case, where the problem can be perfectly inverted if the mixing matrix is known. However, due to the difficulty inherent in the separation of underdetermined mixtures, the quality of the extracted sources using sparse methods is considerably worse than in the linear complete case. For example, the performance achieved by staged approaches (Section 2.4.2) can be degraded by both errors in the estimation of the mixing matrix and by errors in the source estimation stage. Therefore, audible artifacts and source residuals can be usually found in the estimated sources.

The sparsity achieved by time-frequency transformations has considerable importance in the quality of the separated sources. Speech and music signals concentrate most of their energy in the middle-low part of the spectrum, thus, spectral overlap is more likely to occur in this frequency range. As shown by Burred [23], the use of non-uniform time-frequency representations that consider the logarithmic behavior of the human auditory system provides an improved separation performance in stereo separation. This is especially true in the reduction of artifact errors measured in terms of Source to Artifacts Ratio (SAR). However, the reduction of interference terms (SIR) tends to be less related to the sparsity/disjointness of the time-frequency frontend.

In the case of binary time-frequency masking algorithms, there are always some time-frequency points that are misassigned to the incorrect source, being the problem greater when the number of sources or the amount of reverberation is increased. As a result, the estimated masks contain isolated time-frequency points that belong to other sources, affecting negatively the quality of the recovered signals. For instance, in a speech communication system, these points may have a considerable influence on the perceived speech intelligibility.

This chapter addresses the problem of interference removal and musical noise reduction in the time-frequency domain. The main motivation is to reduce the interference effects produced by other sources, so that the use of the proposed post-processing techniques can help to improve the isolation of the estimated sources. To this end, two techniques are presented. In Section 4.2, a residual extraction method based on energy masking is introduced. Section 4.3 describes another post-processing method based on the reassignment of isolated points that commonly appear when using time-frequency masking algorithms. Experiments that evaluate the performance of these approaches are reported at the end of each section.

Part of the contents of this chapter are published in [94] and [95].

4.2 Residual Extraction by Energy Masking

When a target source is extracted from a mixture using any separation algorithm, although it may have residuals from other sources, it is usually more clearly perceived than the rest of sources in the track. It seems reasonable to think that, if the interference signal and the estimated source signal have the same energy and are sufficiently disjoint, time-frequency points dominated by the interference will tend to have greater energy than those dominated by the target source. This idea is illustrated in Figure 4.1. The original interference spectra and the original source spectra are depicted in Figures 4.1(a) and (b), respectively. Figures 4.1 (c) and (d) show the estimated spectra of both signals, containing some residuals one from the other. The spectra of the energy-normalized signals are shown in (e) and (f). A comparison between the energy-normalized signals makes able to identify the points (circled) where the interference was present in the estimated source signal, since these points have greater energy in (e) than in (f). Unfortunately, the interference signal is not usually available and detecting the interference points is not an easy task.

This section presents a residual removal technique based on energy masking. First, an estimation of the interference signal is carried out by minimizing the squared error between the original mixture and a scaled version of the estimated source. Afterwards, time-frequency masks used to separate residuals from the estimated source are computed by means of an *Energy-Normalized Source to Interference Ratio* (ENSIR). The following subsections describe in detail the steps involved in this refinement technique. Finally, some experiments that evaluate the performance of the method with both instantaneous and convolutive mixtures are presented.

4.2.1 Energy-Normalized Source to Interference Ratio

The aim of the approach presented in this section is to obtain a set of binary time-frequency masks that, given the estimated source signals and the original mixtures, enable to improve the

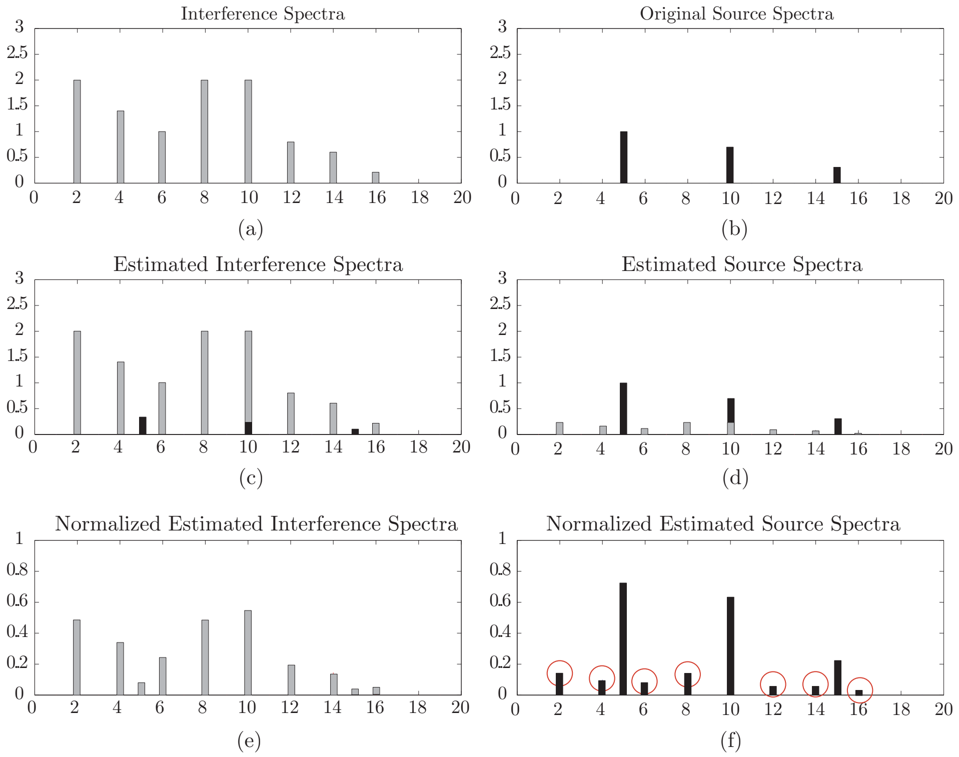


Figure 4.1. Residual Extraction by Energy Masking. (a) Original interference spectra. (b) Original source spectra. (c) Spectra of the estimated interference containing residuals from the target source. (d) Spectra of the estimated source, containing residuals from the interference. (e) Energy normalized spectra of the estimated interference. (f) Energy normalized spectra of the estimated source. Circles denote the points were (e) is greater than (f).

isolation of the sources in terms of SIR. These time-frequency masks are formed by comparing the separated sources with a difference signal between the original mixture and the estimated source in each time-frequency point. The comparison is made via the ENSIR, which is calculated as follows.

Firstly, a separated source image is subtracted from the corresponding observation signal:

$$r_{mn}(t) = x_m(t) - \mu y_{mn}(t), \tag{4.1}$$

where $y_{mn}(t) = \hat{s}_{mn}(t)$ is the recovered image of source n in sensor m, and μ is a scaling factor that minimizes the mean square error

$$\mu = \arg\min_{\mu} \left\{ \sum_t \left(x_m(t) - \mu y_{mn}(t)\right)^2 \right\} = \frac{\sum_t x_m(t) y_{mn}(t)}{\sum_t (y_{mn}(t))^2}. \tag{4.2}$$

Notice that $r_{mn}(t)$ is an estimation of the total interference signal in sensor m. The separated source image and the interference are energy-normalized:

$$\bar{y}_{mn}(t) = \frac{y_{mn}(t)}{\sqrt{\sum_t (y_{mn}(t))^2}}, \tag{4.3}$$

$$\bar{r}_{mn}(t) = \frac{r_{mn}(t)}{\sqrt{\sum_t (r_{mn}(t))^2}}. \tag{4.4}$$

Using the above signals, the Energy-Normalized Source to Interference Ratio is defined as

$$\mathrm{ENSIR}_{mn}(k,r) = 10\log_{10}\left(\frac{|\bar{Y}_{mn}(k,r)|}{|\bar{R}_{mn}(k,r)|}\right), \quad \forall(k,r), \tag{4.5}$$

where $|\bar{Y}_{mn}(k,m)|$ and $|\bar{R}_{mn}(k,m)|$ are the amplitude values of the STFT of $\bar{y}_{mn}(t)$ and $\bar{r}_{mn}(t)$, respectively.

4.2.2 Binary Masks for Residuals Removal

The binary masks that give place to a better isolated version of the sources are obtained from the ENSIR values as follows:

$$M^{\mathrm{src}}_{mn}(k,r) = \begin{cases} 1 & \text{if} \quad \mathrm{ENSIR}_{mn}(k,r) \geq \rho_{th} \\ 0 & \text{if} \quad \mathrm{ENSIR}_{mn}(k,r) < \rho_{th} \end{cases}, \quad \forall(k,r), \tag{4.6}$$

and the masks corresponding to the residuals are then obtained as

$$M^{\mathrm{res}}_{mn}(k,r) = \begin{cases} 0 & \text{if} \quad \mathrm{ENSIR}_{mn}(k,r) \geq \rho_{th} \\ 1 & \text{if} \quad \mathrm{ENSIR}_{mn}(k,r) < \rho_{th} \end{cases}, \quad \forall(k,r), \tag{4.7}$$

where ρ_{th} is a threshold that defines when a given time-frequency point is considered to be the interference or the target source signal. The new source estimates are calculated applying the corresponding masks:

$$Y'_{mn}(k,r) = M^{\mathrm{src}}_{mn}(k,r)Y_{mn}(k,r), \quad \forall(k,r). \tag{4.8}$$

Obviously, the similarity between $Y'_{mn}(k,r)$ and the real image of the source n in sensor m, $S_{mn}(k,r)$, will depend on the accuracy obtained by the algorithm used in the estimation of $Y_{mn}(k,r)$.

The choice of the threshold ρ_{th} is a tradeoff between source isolation and artifacts. As the value of ρ_{th} increases, the number of points eliminated by the mask gets higher. The cancellation of these points can result in a better intelligibility of the target source, due to the elimination of residuals from other sources. However, when ρ_{th} has a high value, the cancellation of points can severely degrade the target source and the overall quality is substantially affected. The examples described in the next section suggest that a good value for the threshold is $\rho_{th} = 0$, which means that only points with normalized source energy lower than the normalized interference energy ($|\bar{Y}_{mn}(k,r)| < |\bar{R}_{mn}(k,r)|$) are suppressed.

The residuals can be added to another source estimate in order to further improve the separation. However, if more than two sources are present, it can be difficult to know which source the residuals should be added to. In that case, a correlation-based technique can be used for adding the residuals to the most convenient source.

4.2.3 Experiments

In this section we evaluate how the described approach improves the isolation of separated sources obtained by means of different SSS algorithms. With this purpose, the separated sources made available by the participants of the *Stereo Audio Separation Evaluation Campaign* (SASSEC) and the *Signal Separation Evaluation Campaign 2008* (SiSEC 2008) were used. Two cases are discussed using the SASSEC test data set [67] (which was also given as a development data set in SiSEC 2008 [68]). Firstly, the performance achieved in stereo instantaneous mixtures is analyzed. With this purpose, a separation example considering several separation algorithms is presented. In addition, the average improvement over all the separated sources available from the campaigns are computed for speech and music mixtures. In a second case, the same procedure is followed with the separated sources corresponding to the convolutive mixtures of the test data sets.

Instantaneous mixtures

The results of applying the proposed technique over the outputs of several source separation algorithms are shown in Table 4.1. These results correspond to an instantaneous mixture of 4 speech sources and are organized into two columns: the original measures with no processing (NP) and the new performance measures after applying the refinement technique (WP). The elimination of residuals was performed using Hann windows of 1024 samples of length and 50% overlap. The threshold was set to $\rho_{th} = 0$.

The separated sources were uploaded by the participants of the SASSEC: Bofill [96], Sawada [70], Vincent [97], and Barry [63]. The results obtained by the MuLeTS algorithm (Cobos) are included as well (it was also evaluated over the same data set in SiSEC 2008). All of these algorithms were developed for dealing with underdetermined mixtures. Bofill carried out the separation by minimization of the l_1 norm of the real and imaginary parts of the source STFT. Vincent used also a sparsity based method, minimizing in this case the l_0 norm of the source STFT. Barry estimated the source magnitude spectra from the minima observed in the ADRess frequency-azimuth plane (see Section 2.6.3), while Sawada and Cobos applied binary STFT masking for recovering the sources.

The results of this example show that energy masking always improves the SIR, with an average gain (over all the sources and separation methods) of 2.25 dB. This improvement in isolation makes artifacts slightly more apparent due to the cancellation of more time-frequency points. However, the SAR only decreases 0.86 dB in average.

Figure 4.2(a) shows the average gain[1] in SDR, ISR, SIR, and SAR for different values of the threshold ρ_{th} after processing all the sources from all the participants of the two campaigns corresponding to the instantaneous speech mixtures included in the data set. A total of 160 separated sources were processed for the experiment, corresponding to 20 different separation algorithms (see [67] and [68] for details on the participant algorithms). It can be observed that a significant improvement in SIR is obtained when increasing the value of ρ_{th} while the SAR and SDR gains remain almost constant. When ρ_{th} has bigger values (around $\rho_{th} = 0$), SAR and SDR start to fall down making the artifacts more apparent. The maximum average SIR gain is achieved when $\rho_{th} = 1$ (6.5 dB), with -1 dB gain in SDR and ISR, and -2 dB gain in SAR.

Similarly, Figure 4.2(b) shows the average gains in the performance measures for the instantaneous music mixtures included in the data set (with $N = 3$ sources). In this case, 57 separated sources corresponding to all the results available in the campaign were processed and evaluated.

[1] Note that these gains are not absolute values, since they are defined as the difference in dB between the new performance figures after processing and the original ones before processing.

Bofill	SDR		ISR		SIR		SAR	
	NP	WP	NP	WP	NP	WP	NP	WP
s_1	3.6	3.5	3.8	3.6	18.0	22.3	13.5	11.0
s_2	6.2	6.2	10.3	9.6	10.6	13.6	7.6	6.8
s_3	4.0	4.4	7.8	11.2	7.8	10.1	6.0	5.3
s_4	3.4	3.7	12.6	6.1	12.6	18.1	9.7	8.5

Sawada	SDR		ISR		SIR		SAR	
	NP	WP	NP	WP	NP	WP	NP	WP
s_1	9.0	8.0	15.0	14.0	20.8	21.2	9.1	7.9
s_2	5.6	5.5	14.6	13.5	11.2	12.9	6.3	5.6
s_3	4.0	3.9	11.7	10.9	7.7	8.7	4.8	4.1
s_4	6.2	5.6	9.0	8.2	20.3	21.9	7.0	6.2

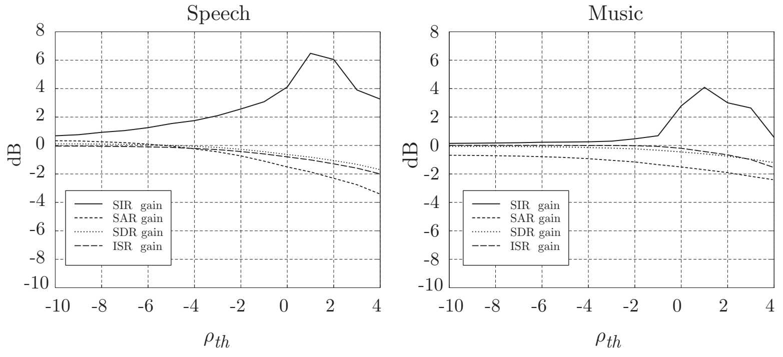
Vincent	SDR		ISR		SIR		SAR	
	NP	WP	NP	WP	NP	WP	NP	WP
s_1	13.3	12.2	26.6	22.5	17.4	21.0	15.4	12.8
s_2	6.8	6.3	11.1	10.3	16.8	18.4	7.1	6.3
s_3	5.8	5.7	11.4	10.7	12.0	14.0	6.1	5.5
s_4	8.4	8.6	18.2	16.0	13.0	17.3	10.0	9.0

Barry	SDR		ISR		SIR		SAR	
	NP	WP	NP	WP	NP	WP	NP	WP
s_1	7.0	6.6	8.5	8.1	18.5	19.9	8.9	7.9
s_2	5.0	4.6	6.0	5.6	16.1	18.0	6.6	5.7
s_3	4.1	4.0	5.8	5.5	11.2	13.1	4.2	3.7
s_4	5.5	5.3	7.2	6.8	15.0	18.2	6.5	6.1

Cobos	SDR		ISR		SIR		SAR	
	NP	WP	NP	WP	NP	WP	NP	WP
s_1	8.9	8.0	18.6	16.9	18.9	21.4	9.0	8.8
s_2	4.3	4.6	19.4	18.2	11.2	12.6	5.4	5.3
s_3	3.3	3.5	15.4	14.4	8.9	9.9	4.2	4.0
s_4	5.8	5.5	9.8	9.3	21.3	21.8	6.2	5.9

Table 4.1. Results obtained for an example mixture of 4 speech sources.

The SIR gain observed is not as high as in the case of speech mixtures. This is probably due to the fact that the separated mixtures are less underdetermined (3 sources versus 4 sources in the speech case). Again, the maximum SIR gain is obtained for $\rho_{th} = 1$, reaching 4 dB in SIR, with an average degradation of 1.5 dB in SAR and 0.5 dB in SDR and ISR.



(a)

(b)

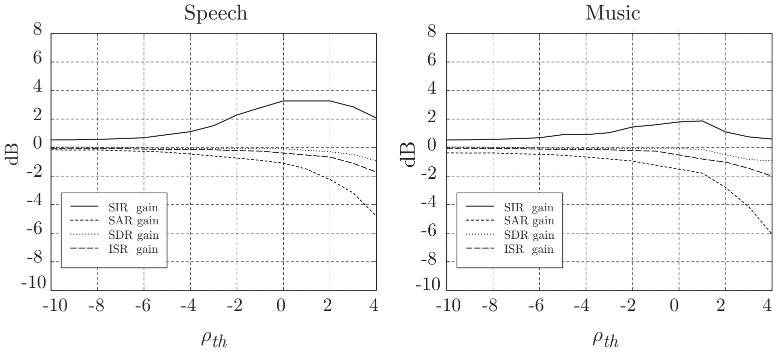
Figure 4.2. Average SDR, ISR, SIR and SAR gains over all the sources and separation algorithms for instantaneous mixtures. (a) speech, $N = 4$. (b) Music, $N = 3$.

Convolutive Mixtures

A similar experiment to the last described for instantaneous mixtures was carried out considering all the available separated sources for convolutive mixtures (live recordings with intermicrophone distance 1 m and 5 cm) corresponding to the SASSEC campaign and those submitted for the same data set in SiSEC 2008.

The average results for speech mixtures (4 sources) are shown in Figure 4.3(a), with a total of 88 separated sources processed corresponding to 11 separation algorithms. In this case, the improvements in SIR are not as high as in the instantaneous case, but still some isolation gain can be obtained. The maximum average SIR gain is for $\rho_{th} = 0$, reaching 3.3 dB. The rest of performance measures follow a behaviour similar to the observed in Figure 4.2, remaining almost constant until $\rho_{th} = 0$, where they start to fall down.

Figure 4.3(b) shows the average results for convolutive music mixtures (with $N = 3$ sources). A set of 54 signals were processed corresponding to 9 algorithms and similarly to the instantaneous case, the gain in SIR achieved is not as high as in the speech experiment (1.9 dB for $\rho_{th} = 1.5$). Therefore, it seems that the energy masking method performs better with instantaneous mixtures than for convolutive mixtures, being more remarkable the benefits when the number of sources is higher.



(a)

(b)

Figure 4.3. Average SDR, ISR, SIR and SAR gains over all the sources and separation algorithms for convolutive mixtures. (a) speech, $N = 4$. (b) Music, $N = 3$.

Subjective quality

As the overall quality of the processed signals is very dependent on the performance of the algorithm used, it is difficult to obtain an overall measure of the perceptual improvement achieved by this method. For that purpose,an informal subjective test was conducted, obtaining the *Mean Opinion Score* (MOS). The listening tests were undertaken by 10 listeners. Each of them awarded a score from 1 (residuals clearly audible and very annoying) to 5 (not audible) for each output signal. The same was done after applying the proposed method to the estimated sources. Only a instantaneous mixture of 4 sources and the outputs of 5 algorithms were considered for the test, so each subject had to evaluate 20 unprocessed sources and 20 processed sources. The average MOS over all the sources and separation algorithms was 1.9 before using the energy masking technique. The average result in the case of applying it, was 2.3. Therefore, an increment of 0.4 was obtained.

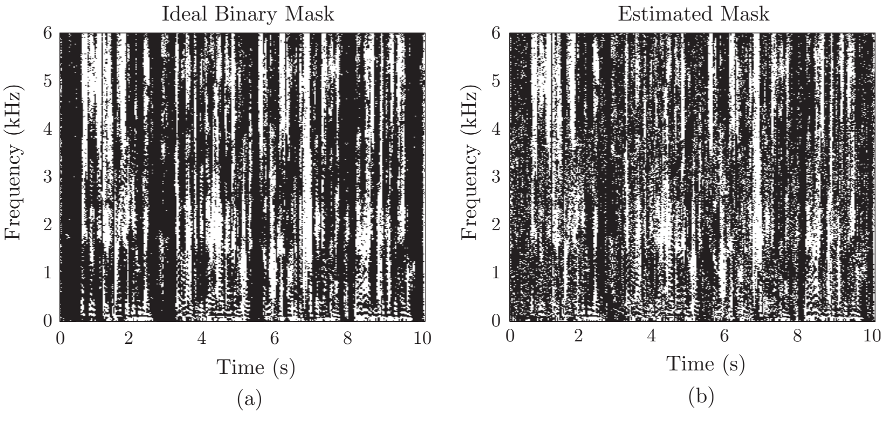
4.3 Neighborhood-Based Reassignment

In the last section, a time-frequency masking method for improving source isolation was presented. The method was based on comparing the normalized energies of the sources and the estimated interference signals in the time-frequency domain. In the present section, another approach for removing source residuals is described.

The estimation of the separation masks is not always perfect in the sense that they differ from the ideal binary masks. In Figure 4.4.(a), the ideal binary mask (black represents zero and white represents one) for the extraction of one source in a two-channel anechoic mixture of three speech sources is represented. Figure 4.4(b) shows the corresponding estimated mask using the MuLeTS algorithm. In the figure it is clear that, whereas most of the non-zero elements in the ideal mask are robustly clustered around harmonic partials and uniform zones, the estimated mask has much more small elements scattered around these areas. These scattered small clusters contribute to musical noise and audible residuals from the other sources when the mask is applied to recover the source.

Musical noise has been widely considered in the field of single channel speech enhancement via spectral subtraction [98], [99]. Few works on source separation are aimed at proposing techniques tackling this problem. For example, Araki *et al.* proposed to use a fine-shift overlap-add analysis of the input mixtures to reduce musical noise [58]. However, the problem is not attacked directly in the estimated masks and the perceptual improvement is only a consequence of smoothing existing errors by averaging more signal frames in the STFT processing.

The approach introduced in this section is intended to be used as a post-processing method for time-frequency masking algorithms. The objective is to reassign isolated and small clusters of non-zero elements in binary masks to the source which has maximum likelihood in the time-frequency neighborhood of the clusters. As will be later explained, the likelihood function represents the closeness of a local estimation of the mixing parameters in each bin to the estimation with highest support given by the rest of time-frequency points, i.e. the actual mixing matrix directions.



(a)

(b)

Figure 4.4. (a) Ideal binary mask. (b) MuLeTS binary mask

This section describes how to identify and reassign these small elements using a neighborhood based criterion. In the following subsections we will describe in detail the steps involved in the proposed reassignment.

4.3.1 Cluster Labelling

In the first step, the estimated masks corresponding to the sources in the mixture are analyzed for finding clusters of non-zero elements in them. The clusters are formed by grouping 4-connected or 8-connected objects in the binary masks, as described in [100]. When a cluster C_{nj} is found in a mask $M_n(k,r)$, the time-frequency bins in the mask that form the cluster are labeled with the cluster size, forming a time-frequency cluster map $\mathbf{C}(k,r)$:

$$C(k,r) = N_{C_{nj}}\big|_{(k,r)\in C_{nj}}, \quad \forall(k,r), \forall n, \tag{4.9}$$

where $N_{C_{nj}}$ denotes the number of elements in the jth cluster C_{nj}. The process is repeated for all the masks until all the time-frequency points have been labeled with their corresponding cluster size.

This way, isolated time-frequency bins in a mask will be labeled as 1, while points in a cluster of 100 connected points will have a label of 100. Figure 4.5(a) shows the cluster map of the example mixture. It can be observed that time-frequency bins with big labels are predominant. Figure 4.5(b) shows the time-frequency bins that form clusters with no more than 3 elements. As it will be explained in the next subsection, these elements will be the candidates for being reassigned in the final masks.

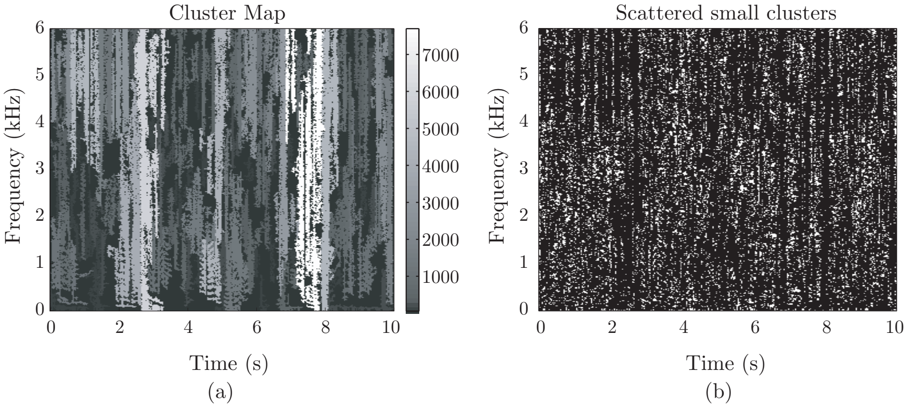


Figure 4.5. (a) Cluster map. (b) Time-frequency bins with label lower than 4.

4.3.2 Source Reassignment

From the observation of Figures 4.4(a) and 4.4(b), it is possible to see that time-frequency bins forming small clusters are more likely to appear in the estimated masks than in an ideal binary mask. This fact is also graphically depicted in the histogram of Figure 4.6. The histogram shows the number of time-frequency bins in $\mathbf{C}(k,m)$ labeled with small numbers for the two masks shown in Figure 4.4. In the case of the MuLeTS mask, a higher number of time-frequency bins form small clusters of nonzero elements.

In order to reduce the number of scattered nonzero elements, time-frequency bins labeled with small numbers are selected as candidates for being reassigned to a different source. A threshold κ can be defined for setting the minimum cluster size. Although this threshold can be modified, experimental results suggest that good values for κ are from $\kappa = 1$ to $\kappa = 10$. This can be also inferred from the histogram of Figure 4.6, which shows that the distribution of elements with label below 10 are more easily found in an estimated mask than in an ideal mask.

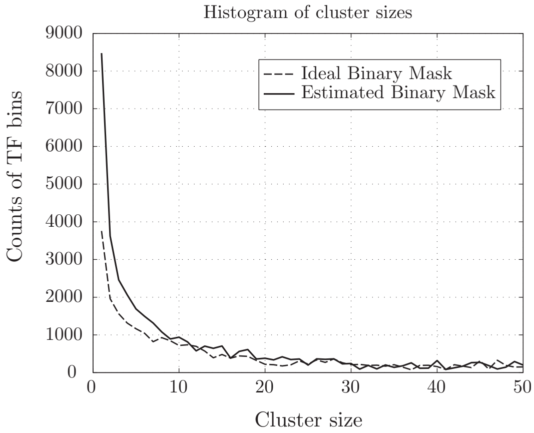


Figure 4.6. Histogram showing the number of time-frequency bins with low cluster label in an estimated mask and the corresponding ideal mask.

The reassignment of the selected time-frequency bins is carried out by exploring the time-frequency neighborhood of the bin, i.e. its neighbors in the time-frequency plane. The span of the neighborhood in rows and columns is defined by the γ parameter. This approach is powerful in the sense that, as small clusters are not easily found in ideal masks, it is likely that the points of small clusters belong to the source with maximum likelihood in their time-frequency neighborhood. The likelihood function should be chosen in agreement with the method used. Usually, this likelihood function represents the closeness of a time-frequency point to each of the estimated mixing parameters. For example, in the DUET algorithm, the distances to the identified peaks in the two-dimensional histogram are used as a likelihood function (see Eq.(2.75)). In algorithms involving other clustering techniques, such as K-means, the likelihood function used can be chosen as the distance of the data to the final cluster centroids. In the MuLeTS algorithm, a likelihood function can be formed using the distance of the local estimations to the maximum values of the thresholded sections in the azimuth (or panoramic) plane.

Denoting $\mathcal{L}_n(k,r)$ as the likelihood matrix related to source s_n, the reassignment algorithm can be described as follows.

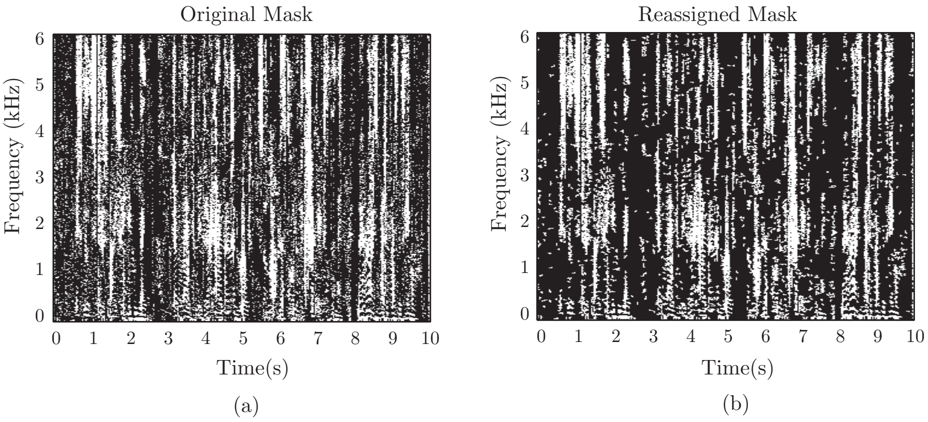
Reassignment algorithm

Inputs: Time-frequency cluster map $\mathbf{C}(k,r)$, estimated separation masks $\mathbf{M}_n(k,r)$, source likelihoods $\mathcal{L}_n(k,r)$, maximum cluster size allowed κ and neighborhood span γ.

1. Initialize the final masks $\mathbf{M}'_n$ with the value of the current masks: $\mathbf{M}'_n = \mathbf{M}_n$. Start to explore each time-frequency point (k,r).
2. If $C(k,r) \leq \kappa$ go to 3, else go to 6.
3. Find the source n to which the point (k,r) belongs, i.e. $M_n(k,r) = 1$.
4. Find the source q with maximum likelihood in the neighborhood of (k,r): $q = \arg\min_n \{\mathcal{L}_n(k_q,r_q)\}$. The span of the neighborhood is defined by γ:
   $$k-\gamma \leq k_q \leq k+\gamma, \quad k_q \neq k$$
   $$r-\gamma \leq r_q \leq r+\gamma, \quad r_q \neq r.$$
5. Set $M'_n(k,r) = 0$ and $M'_q(k,r) = 1$.
6. If all the points were explored: end. Else, update (k,r) to the next point and go to 2.

Outputs: Final masks $\mathbf{M}'_n(k,r)$.

The result of applying the above algorithm to a binary mask can be observed in Figure 4.7. Note that the small scattered elements do not appear in the reassigned mask.



(a)

(b)

Figure 4.7. Binary separation mask. (a) Original. (b) Reassigned with $\kappa = 5$ and $\gamma = 3$

.

4.3.3 Experiments

In this section we evaluate the proposed approach over the masks obtained using the MuLeTS algorithm (with half overlapping Hann windows of 1024 samples length). The set of signals used for the experiments are the instantaneous speech mixtures (4 sources) contained in the development and test data sets of the SASSEC. The MuLeTS algorithm was applied to these mixtures, resulting in 16 masks corresponding to each separated source. The average gains for each performance measure are shown in the graphs of Figure 4.8(a)-(d), where different values for the κ and γ parameters are considered. In (a), the technique is applied by suppressing the selected points, without assigning them to other masks. As it can be seen, a considerable improvement is achieved for $\kappa > 10$, with average SIR gains around 2.5 dB. This improvement is produced at the expense of a 1-2 dB decrease in the average SAR. However, the overall distortion (SDR) is not so severely affected, remaining below 0.5 dB. Surprisingly, the results after applying reassignment (b)-(d) are not as good as with simple suppression, which suggest that it seems better to simply clean the mask by eliminating scattered points than to reassign them to other sources. Another interesting result that was observed when processing the mixtures is that the maximum improvements are always produced for the sources with the lowest quality, with increments both in terms of SIR, SAR and SDR.

4.3.4 Conclusions

Blind Source Separation algorithms can be used in a lot of applications. However, although very acceptable and promising results have been obtained in the last years, the interfering energy present in the extracted sources is sometimes too significant for using them in practical systems. For example, in the case of time-frequency masking, the estimated masks are sometimes strongly corrupted by scattered nonzero points that cause a noticeable degradation of the extracted sources, both in terms of interference and musical noise artifacts.

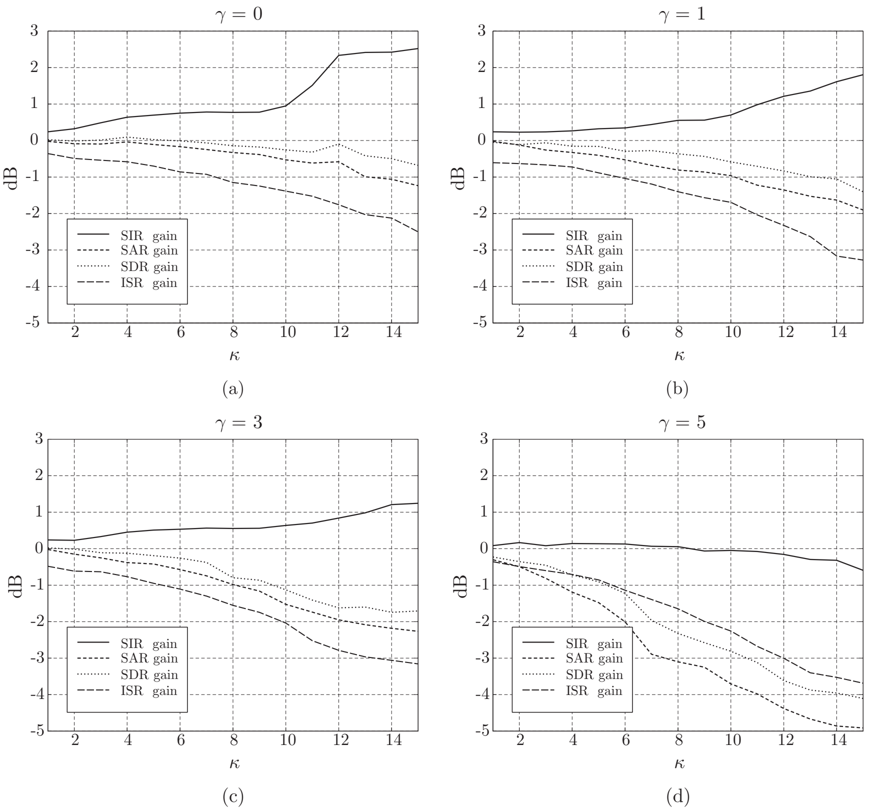


Figure 4.8. Average SDR, ISR, SIR and SAR gains over all the sources and mixtures, $N = 4$. (a) $\gamma = 0$ (only supression). (b) $\gamma = 1$. (c) $\gamma = 3$. (d) $\gamma = 5$.

In this chapter, two post-processing methods intended to be applied over the results obtained by SSS algorithms have been introduced. First, an easy to implement technique has been proposed for removing residuals from other sources given the original mixtures and the estimated sources. This refinement technique is based on analyzing the normalized energy of the sources and the mixtures in the time-frequency domain via the calculation of an Energy-Normalized Signal to Interference Ratio. The second proposed approach is a source reassignment technique that can be applied to post-process the binary masks obtained by time-frequency masking algorithms. This technique allows to reassign isolated and small clusters of non-zero elements in the masks to the source which has maximum likelihood in the time-frequency neighborhood of the elements.

Both methods have been evaluated considering the data sets provided by the two evaluation campaigns SASSEC and SiSEC 2008. The average results show that the proposed post-processing techniques achieve an improvement of the separated tracks in terms of source isolation. Although artifacts appear due to the cancellation of the time-frequency points, the SIR is highly improved. The suitability of these techniques depends on their application context. Removing interference from other sources can be very useful for applications where speech intelligibility is very important, such as automatic speech recognition systems. However, other applications where the timbral quality prevails over the audible interference residual may not find any benefit in their use. Nevertheless, informal listening tests suggest that the subjective quality of the sources estimates improves in most cases.

Part II

Spatial Sound

5 Spatial Sound Systems

Spatial Sound Systems 5

The objective of spatial sound systems is to accurately recreate the acoustic sensations that a listener would perceive inside a particular room or in an environment with certain acoustic properties. This concept, easy to understand, implies a series of physical and technological difficulties that are a current research issue in sound engineering. Stereo sound systems, considered as the simplest approximation to spatial sound, have been utilized throughout the last 50 years as an added value in sound recordings, specially for music material. Together with the entertainment industry, spatial sound evolved to surround sound systems, which provide a better sensation than stereo by using more reproduction channels. Nowadays, the most promising systems for spatial sound reproduction are those based on synthesizing the acoustic field produced by a set of discrete sound sources, especially *Wave Field Synthesis* (WFS). In this chapter, the basics underlying all these audio systems are described, presenting current research lines in spatial sound-reproduction and up-mixing techniques.

5.1 Introduction

Human hearing plays a major role in the way our environment is perceived. One of the most important cues in spatial perception of sound is localization. Generally, sound is perceived in all three dimensions, width, height and depth, which are all necessary to achieve a natural perception of sound [1]. In fact, natural sound scenes are made of different sounds, which may be perceived individually or as different entities where complex grouping mechanisms are involved, some of them directly related to spatial *localization cues*. When a sound scene is recorded by a single microphone, we are still able to recognize the original sound events. However, much of the information corresponding to the spatial properties of these events is lost.

Reproduction using two-channels or *stereo* is the most common way that most people know to convey some spatial content into sound recording and reproduction, and this can be considered as the simplest approximation to spatial sound. On the other hand, *surround sound* systems have evolved and entered homes in order to give a better sensation than stereo by using more reproduction channels and have been widely utilized in theaters since the middle 70s. Surround mixes are mainly intended to enhance the experience in video projections by adding artificial effects in the rear loudspeakers (explosions, reverberation or ambient sound). Both stereo and surround systems have an optimal listening position, known as *sweet spot*. This optimum listening area is almost limited to the central point in the loudspeaker set-up and the spatial sensation degrades considerably outside the central zone [101]. Acoustic simulations of complex loudspeaker setups are able to predict the behavior of a spatial sound system with any room geometry [102].

Another much more realistic strategy is to reproduce directly in the ears of the listener, via headphones, the signal that he would perceive in the acoustic environment that is intended to be simulated. This strategy is widely known as *binaural reproduction*. The signals to be reproduced with headphones can be recorded with an acoustic *dummy head* or they can be artificially synthesized by using a measured *Head-Related Transfer-Function* (HRTF) [103]. There are still some issues to be solved regarding the HRTF variability among different subjects and active research lines are centered on this aspect of binaural reproduction. In addition, the incompatibility in the reproduction of dummy head signals over loudspeakers is another classical problem: loudspeaker reproduction of binaural signals introduces crosstalk, where the left channel signal intended for the left ear will also be heard by the right ear and vice versa. This non desired effect may be eliminated by prefiltering the binaural signal with an inverse system called a crosstalk cancelling filter [104].

Throughout the last decades, a number of different approaches have been proposed to improve spatial sound reproduction over loudspeakers. They can be roughly categorized into: *advanced panning* techniques, *Ambisonics*, and *Wave Field synthesis* (WFS) [105]. Advanced panning techniques are an extension of the stereophony principle to complex loudspeaker setups. An example is the *vector base amplitude panning* technique (VBAP) [106]. On the other hand, Ambisonics systems represent the sound field in an enclosure by an expansion into three-dimensional basis functions [107]. A faithful reproduction of this sound field requires recording techniques for the contributions of all relevant basis functions. The most popular multichannel sound system based on wave-field rendering is Wave Field Synthesis. WFS is a technique for reproducing the acoustics of large recording rooms in smaller sized listening rooms. The most basic difference of this system in comparison to surround systems is that the acoustic field is accurately synthesized using loudspeaker arrays in a broad area, suppressing the sweet spot that characterizes conventional surround systems. Therefore, the most important advantage provided by WFS is that the spatial properties of the acoustical scene can be perceived correctly by an arbitrary large number of listeners which are allowed to move freely inside the listening area without the need for tracking their positions. These features are achieved through a strict foundation on the basic laws of acoustical wave propagation [108].

In this chapter, an overview of spatial sound reproduction systems is carried out. First, the basic localization cues used by humans are described in Section 5.2, providing some concepts on spatial hearing that are useful to understand the fundamentals of these systems. Then, different spatial sound systems are presented in Section 5.3, including stereo reproduction, surround systems and sound field rendering techniques. WFS systems have special relevance in the context of this thesis, so their basics are more deeply explained in Section 5.4. Finally, some basics on

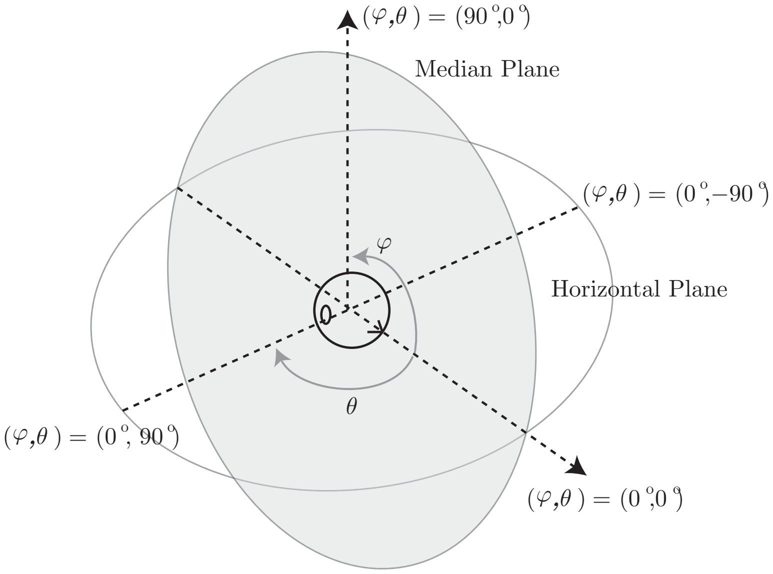

$(\varphi,\theta) = (90°,0°)$
Median Plane
$(\varphi,\theta) = (0°,-90°)$
φ
Horizontal Plane
θ
$(\varphi,\theta) = (0°, 90°)$
$(\varphi,\theta) = (0°,0°)$

Figure 5.1. Interaural coordinate system.

audio up-mixing are provided in Section 5.5, with a description of the forthcoming MPEG standard on *Spatial Audio Object Coding* (SAOC).

5.2 Spatial Hearing

Humans and other animal species have the remarkable ability of identifying the direction of a sound source originating from any point in the three-dimensional space. Throughout evolution, the sense of hearing has helped our survival. As many other mammals, the sense of hearing has played a major role in hunting and avoiding to be hunted, as our hearing sense enables us to identify dangers or targets in the environment. The human auditory system is very sophisticated and, thus, capable to analyze and extract most spatial information pertaining to a sound source using two ears. However, the process of localizing a sound source is dynamic and often aided and complemented by other sensory inputs. The usual coordinate system used in spatial hearing studies is represented in Figure 5.1. Note that the zero azimuth angle is defined in front of the listener, and not at the left side as in Chapter 3. The ear that is closest to a source is called *ipsilateral* and the opposite ear is called *contralateral*.

5.2.1 Interaural Differences

One of the basic binaural processing mechanisms involves the comparison between the time of arrival of the sound to the left and right ears. This difference is commonly known as *Interaural Time Difference* (ITD). If we assume that the average distance between human ears is about 18 cm [109], the ITD has a maximum value of about ±0.75 ms. Notice that the ITD will not uniquely determine the direction of a sound source since there will always exist ambiguity with respect to the front and back hemispheres.

Another consequence of the presence of the head is that higher frequencies are attenuated or shadowed by the head as they reach the contralateral ear. This attenuation produces an

Interaural Level Difference (ILD) which also plays a major role in lateral localization, especially at high frequencies.

The ITD and ILD are considered to be the primary cues for the perceived azimuth angle of a sound source, as proposed by Rayleigh in what is known as the "Duplex theory" [110]. In principle, knowledge of the ITD and ILD would allow one to estimate the azimuth angle, and hence to constrain the location of the source to a particular *cone of confusion.*[1]. Localization in elevation is well developed in humans, but involves other auditory cues as described next.

5.2.2 Spectral Cues

In the median plane (i.e. $\theta = 0°$), the bilateral symmetry of the body implies that both the ITD and the ILD must vanish. However humans are still able to localize sound in the median plane by what is known as *monaural cues*, which are related to the spectral changes introduced by the outer ears (i.e. pinnae) at high frequencies and other body structures like the torso at low frequencies [111]. Some studies have shown that these cues help listeners with complete hearing loss in one ear to localize the azimuth direction of a source with relatively high accuracy. However, this was not the case for fully-binaural subjects with a blocked ear [112].

Spectral cues are also used to discriminate the front from the back when the sound source has sufficient high-frequency energy (above 3 kHz). These cues are believed to be introduced by the front/back asymmetry of the pinna, which results in a pinna shadow for sound sources arriving from the back. In the absence of this cue, head rotation is necessary to resolve front/back ambiguity [113]. In fact, effective localization of unfamiliar sources in the median plane can only be achieved with head motion.

5.2.3 Distance and Dynamic Cues

At large distances, interaural differences and spectral cues are not reliable cues to estimate the distance of a source. One of the most useful cues for range estimation is *loudness*. It is well known that the loudness (and to a lesser degree, the spectra) of a sound source changes with distance [114]. As with median plane localization, the effectiveness of this cue depends on the familiarity of the listener with the source. Other cues for distance perception are those derived from the acoustic environment. Reverberation and/or reflections from nearby surfaces play a major role in distance perception. The ratio of reverberation to direct sound (D/R ratio) is a function of the relative distance between source and listener and the room acoustics. This cue can be more reliably used by listeners even if they have no familiarity with the particular sound source.

Dynamic cues are extremely useful to resolve the ambiguities that static cues cannot handle. Many studies have shown that when subjects are allowed to move the head, localization blur and front/back reversals are significantly reduced [115]. Experiments have shown that listeners evaluate interaural differences at the same time as they move their head in relation to the direction of the source. All cues need to be consistent to produce the correct perception, including other non-auditory cues (e.g. visual cues) that carry information [113].

[1] Notice that in the interaural coordinate system for a constant azimuth angle and range, the trajectory described by the source along the elevation angle corresponds to a slice of the cone of confusion

5.2.4 The HRTF

In an anechoic environment, as sound propagates from the source to the listener, the different structures of the listeners own body will introduce changes to the sound before it reaches the ear drums. The effects of the listeners body are captured by the *Head-Related Transfer Function* (HRTF), which is the transfer function between the sound pressure that is present at the center of the listeners head when the listener is absent and the sound pressure developed at the listeners ear. The HRTF is a function of direction, distance and frequency. The inverse Fourier transform of the HRTF is the *Head-Related Impulse Response* (HRIR), which is a function of direction, distance, and time. In the time domain, the ITD is encoded in the HRIR as differences in the time of arrival of the sound between ipsilateral and contralateral side. Close to the median plane, the time of arrival of the wavefront is similar for both ears. However, as the azimuth angle increases, the time of arrival to the contralateral ear progressively exceeds that of the ipsilateral, thus increasing the ITD. The ILD is encoded as the level differences observed in the HRTF magnitude responses. Notice how the level difference is small near the median plane, and increases with lateral angle. In the median plane, both ITD and ILD are very small, but there are strong spectral variations (i.e. monaural cues) that change with elevation.

While the HRTFs of most humans share many similarities, more detailed examination reveals subtle differences determined mainly by differences in body shape and size among subjects. These subject-dependent differences have been shown to play a major role for precise localization. It is believed that only using ones own HRTF can result in realistic and accurate binaural audio, as evidenced by various experiments [116]. Some research groups have investigated the effects of synthesizing spatial audio using distorted versions of the measured HRTFs. Kulkarni and Colburn [117] conducted experiments with progressively smoothed versions of HRTFs. The results showed that the HRTFs could be smoothed drastically, retaining only the gross features of the original measurements, and still be surprisingly effective at generating spatialized sound. A sample of a smoothed HRTF is depicted in Figure 5.2.

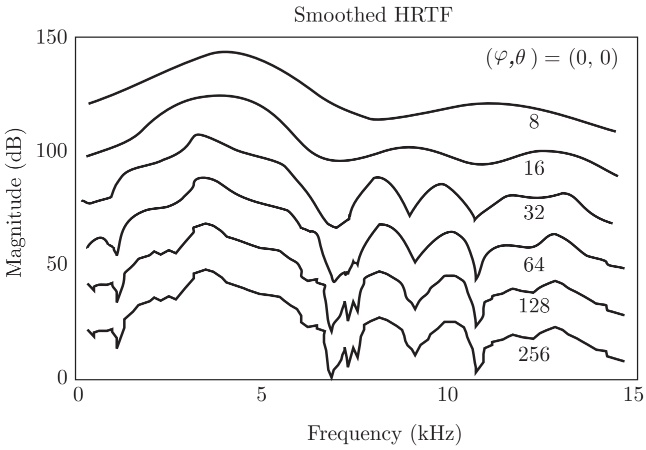


Figure 5.2. HRTF distorted with different smoothing factors.

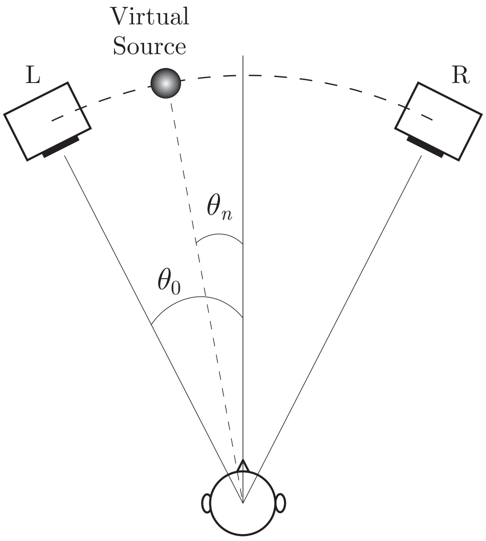


Figure 5.3. Listener in front of two louspeakers for stereo reproduction. The angle between the loudspeakers and the look direction of the listener (base angle) is θ_0. The angle of a phantom source relative to the listener is θ_n.

5.3 Sound Reproduction Systems

5.3.1 Stereophony

The physical superposition of two loudspeakers enables the building of a *phantom* source, which is understood as a substitute sound source. This effect is called "*summing localization*", which is supposed to create binaural cues very similar to the ones created by real sources [118]. There are objections to this explanation. In his "*association model*", Theile [119] argues that the superposition of the loudspeaker signals does not create localization, but rather that the signals from the two loudspeakers give two different localization stimuli that merge together into a phantom source after a complex psychoacoustic process. Leaving aside open questions regarding the nature of phantom sources, it is a fact that stereophonic panning laws and stereo microphone recording techniques have been widely used to achieve spatial localization in all kinds of sound material [106]. Today, the "stereo format" is the most common format used for the commercial distribution of sound recordings.

The practical experience and a variety of formal research works [120] state that the optimum configuration for two-loudspeaker stereo is an equilateral triangle with the listener located just to the rear of the point of the triangle as seen in Figure 5.3. Outside this sweet spot, phantom images (the apparent locations of sound sources in between the loudspeakers) become less stable, and the system is more susceptible to head rotation. If the amplitudes of the two channels are correctly controlled, it is possible to produce resultant phase and amplitude differences for continuous sounds that are very close to those experienced with natural sources, thus giving the impression of virtual or phantom images anywhere between the left and right loudspeakers. This is the basis of Blumlein's 1931 stereophonic system invention [121]. It is often assumed that the mixing coefficients used in the mix-down are related to the perceived angle θ_n of the virtual source n by the tangent panning law [122]:

$$\frac{\tan\theta_n}{\tan\theta_0} = \frac{a_{1n} - a_{2n}}{a_{1n} + a_{2n}}, \tag{5.1}$$

where θ_0 is the loudspeaker base angle, as in Figure 5.3. A complete review of stereo formats and recording techniques can be found in [123].

5.3.2 Surround Systems

Although stereo sound was a breakthrough for consumers of the 50's and 60's, it has some limitations. The difference in the left and right channels was too much emphasized in some recordings, and there were not enough mixing elements in the "phantom" center. Also, even though the sound was more realistic, the lack of ambience information, such as back reflections or other elements, left stereo sound with a "wall effect" in which everything was coming from the front. Therefore, it lacked the natural sound of back wall reflections or other acoustic elements. With the aim of improving the spatial impression of the reproduced sound in the entertainment industry, many film production companies proposed to reproduce sound tracks by using multiple audio channels. However, surround sound standards often specify little more than the channel configuration, and the way the loudspeakers should be arranged (e.g. 5.1-channel surround). This leaves the business of how to create or represent a spatial sound field entirely up to the user [101].

The development of surround sound technology began as early as before the World War II and, from the very beginning, it has been driven by the movie industry. The most known surround system is 5.1, which enables the provision of stereo effects or room ambience to accompany a primarily front-orientated sound stage. Essentially, the front three channels are intended to be used for a conventional three-channel stereo sound image, while the rear/side channels are only intended to generate supporting ambience, effects or "room impression". This is the reason why some standards use the term 3-2 stereo to denote this surround system. In the beginning of the 1990s, the 5.1 configuration, introduced in their systems by both Dolby Laboratories (*Dolby Digital* for home systems and *Dolby Digital Surround* for cinemas) and the *Digital Theater Systems* (DTS), became the "de facto" standard of loudspeaker layouts for especially home multichannel systems. Figure 5.4 shows the 3-2 format reproduction according to the ITU-R BS.775 standard. The ".1" of 5.1 refers to a dedicated *low frequency effects* (LFE) channel or sub-bass channel and it is called ".1" because of its limited bandwidth. Other layouts such as the 7.1 *Sony Dynamic Digital Sound* and the 6.1 *Dolby Digital Surround EX* are popular alternatives in cinema usage as well as in the products of gaming technology.

While two-channel stereo can be relatively easily modeled and theoretically approached in terms of localization vectors and such like, for sounds at any angle between the loudspeakers, it is more difficult to come up with such a model for the 5-channel layout, as it has unequal angles between the loudspeakers and a particularly large angle between the two rear loudspeakers [101].

5.3.3 Vector-Based Amplitude Panning

Vector Based Amplitude Panning (VBAP) is a method for positioning virtual sources to arbitrary directions using a setup of multiple loudspeakers [106]. A great advantage of VBAP reproduction is that the number of loudspeakers needed is arbitrary and they can be also positioned in an arbitrary 2-D or 3-D arrangement. VBAP is based on amplitude panning, so the same sound signal is applied to a number of loudspeakers with appropriate non-zero amplitudes. With 2-D setups, VBAP is a reformulation of the existing pairwise panning method. However, it can be generalized for 3-D loudspeaker setups as a triplet-wise panning method. A sound

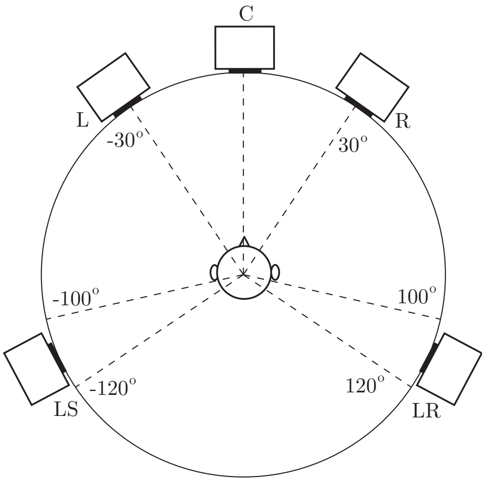


Figure 5.4. 3-2 format reproduction according to the ITU-R BS.775 standard.

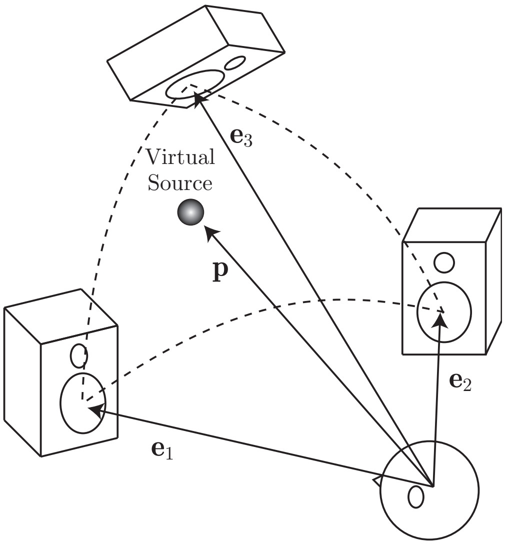


Figure 5.5. Three-dimensional panning VBAP.

signal is then applied to one, two, or three loudspeakers simultaneously. Next, we describe the derivation of this advanced panning formulation.

With loudspeaker systems that also include elevated loudspeakers, the pair-wise paradigm is not appropriate. Triplet-wise panning can be formulated for such loudspeaker configurations. The loudspeakers in a triplet form a triangle from listener's view. The listener will perceive a virtual source inside the triangle, depending on the ratios of the loudspeaker amplitudes.

In three-dimensional VBAP, a loudspeaker triplet is formulated with vectors as in Figure 5.5. The Cartesian unit-length vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ point from the listening position to the loudspeakers. The direction of the virtual source is presented with a unit-length vector $\mathbf{p}$. Vector $\mathbf{p}$ is expressed as a linear weighted sum of the loudspeaker vectors:

$$\mathbf{p} = g_1\mathbf{e}_1 + g_2\mathbf{e}_2 + g_3\mathbf{e}_3. \tag{5.2}$$

Here, g_1, g_2, and g_3 are the gain factors of the respective loudspeakers. The gain factors can be solved as

$$\mathbf{g} = \mathbf{p}^T \mathbf{L}_{123}^{-1}, \tag{5.3}$$

where $\mathbf{g} = [g_1\ g_2\ g_3]^T$ and $\mathbf{L}_{123} = [\mathbf{e}_1\ \mathbf{e}_2\ \mathbf{e}_3]$. The calculated factors are used in amplitude panning as gain factors of the signals applied to respective loudspeakers after suitable normalization, e.g. $\|\mathbf{g}\| = 1$. If more than three loudspeakers are available, a set of non-overlapping triangles are formed of the loudspeaker system before run time. There can be several virtual sources applied to one triplet and the triangularization can be performed automatically using the method presented by Pulkki in [124].

5.3.4 Sound Field Rendering

Sound field rendering or sound field synthesis methods use a large number of loudspeakers to reproduce a sound field not only at the ears of one listener, but in a larger space enclosing multiple listeners. The goal is to reproduce correctly the sound field generated by a set of virtual sources. In contrast to amplitude panning techniques, psychoacoustical effects play a minor role here, since it is assumed that the listeners respond to the synthesized sound field in the same way as to the original one.

Ambisonics is a technique that was proposed in the early 70s [125], that provides a way to encode three dimensional sound fields, usually by recording. These encoded sound fields can then be reproduced over various different speaker arrangements, which is known as Ambisonic decoding. An advantage of Ambisonics reproduction is that it is based on solid mathematics. The accuracy in the reproduction of a sound field is given by the Ambisonic order, related to the order of a spherical harmonic decomposition of the sound field. Whereas zeroth order corresponds to mono reproduction, the most common first order form is known as *B-format*. This format uses four channel encoding, corresponding to the instantaneous sound pressure and the three components of its gradient which are related to the particle velocity at a point in space. The loudspeaker signals are derived by using a linear combination of these four channels, where each signal is dependent on the actual position of the speaker in relation to the center of an imaginary sphere, the surface of which passes through all available speakers. In more advanced decoding schemes, spatial equalization is applied to signals to account for the differences in the high and low-frequency sound localization mechanisms in human hearing. Current research in Ambisonics reproduction is related to High Order Ambisonics (HOA), where more channels than in the first order B-Format are used [126].

The most popular multichannel sound system based on sound field rendering is Wave Field Synthesis, which is more extensively presented in the next section.

5.4 Wave Field Synthesis

In 1953, Snow published an overview of stereophonic techniques and discussed the acoustic curtain as the ideal stereophonic reproduction technique [127]. Figure 5.6 shows a reproduction from that article illustrating the desired and implemented stereophonic systems. It was aimed at transporting the acoustic of the recording venue to a reproduction room using microphone and loudspeaker arrays. Due to technical constraints at that time, it was not feasible to put his ideas into practice. As a compromise, they applied three-channel stereophony, accepting that the original aim of recreating the real sound field would no longer be fulfilled. Snow described this precursor of WFS in this way: "The myriad loudspeakers of the screen, acting as point sources of sound identical with the sound heard by the microphones, would project a true copy of the original sound into the listening area. The observer would then employ ordinary binaural listening, and his ears would be stimulated by sounds identical to those he would have heard coming from the original sound source".

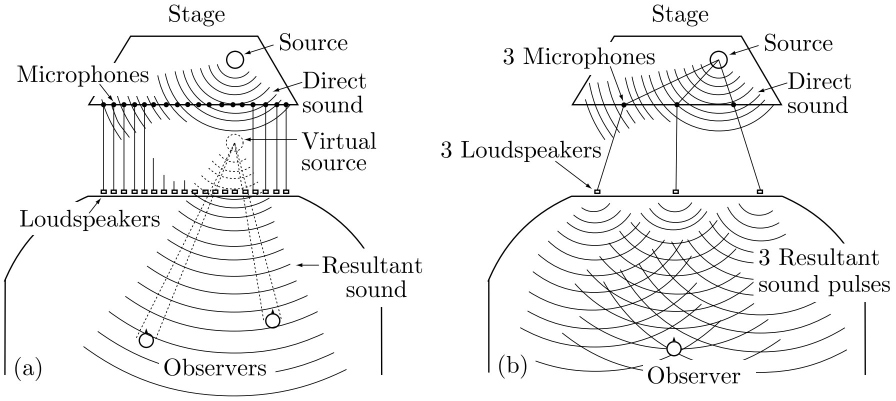


Figure 5.6. Acoustic curtain concept. (a) Ideal system. (b) Actual 3-channel stereophonic system due to early technical constrains.

The intuitive acoustic curtain concept was replaced later by a well-founded wave theory. In the late 80s, the Wave Field Synthesis (WFS) concept was introduced by the Technical University of Delft. The origin of this theory was published in "Applied Seismic Wave theory" [128] and "A holographic approach to acoustic control" [129]. The term "acoustical holography" was used, not yet called WFS, and the system was designed to be the ultimate tool for acoustical control systems in theaters. These publications introduced the physical basis of WFS by applying algorithms known from seismics to the field of acoustics. The basic work on WFS was continued by Berkhout in [130] and [131]. Since then, a number of publications have appeared to complement and improve this basic theory. The following subsections will describe the WFS concepts.

5.4.1 Kirchhoff-Helmholtz and Rayleigh Integrals

The theory of WFS is related to Huygens' principle, formulated in 1678. This principle states that each element of a wave front propagating through a particular medium may be seen as the center of an individual spherical wave. Consequently, the wave front generated by a *primary* sound source can be seen as a series of elementary, *secondary* sources. It is not very practical to position the acoustic sources on the wavefronts for synthesis. By placing the loudspeakers on an arbitrary fixed curve and by weighting and delaying the driving signals, an acoustic wavefront can be synthesized with a loudspeaker array. Figure 5.7(a) illustrates this principle.

Mathematically, the simple source formulation of the Helmholtz integral investigates the possibility that the pressure inside or outside the surface could be determined:

$$P(\mathbf{r}) = \iint_S \left[G \frac{\partial P(\mathbf{r})}{\partial \mathbf{n}} - P(\mathbf{r}) \frac{\partial G}{\partial \mathbf{n}} \right] dS, \tag{5.4}$$

where G is the free space Green function, given by

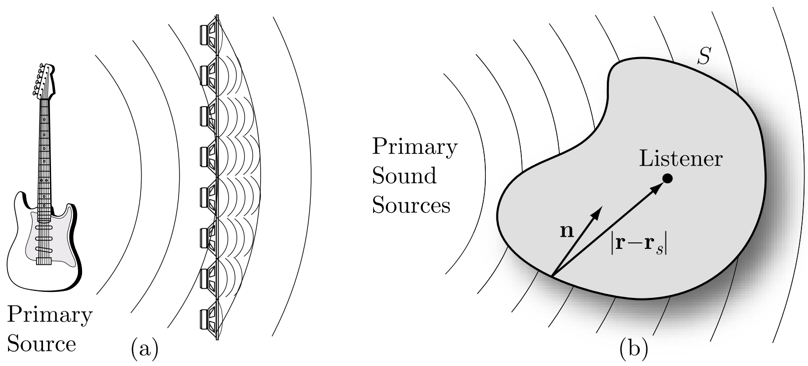


Figure 5.7. (a) Basic principle of WFS. (b) Parameters used for the Kirchhoff-Helmholtz integral.

$$G(\mathbf{r}|\mathbf{r}_s) = \frac{1}{4\pi}\frac{e^{jkR}}{R}, \tag{5.5}$$

and $R = |\mathbf{r} - \mathbf{r}_s|$. Equation (5.4) states, considering the interior problem, that the acoustic field in V generated by the events outside the surface S can be computed uniquely by replacing these events with a distribution of simple monopole surfaces over $P(\mathbf{r})G(\mathbf{r}|\mathbf{r}_s)$ and summing up their contributions over S. Thus, an arbitrary acoustical wave field can be recreated within a source-free volume V by secondary sound sources distributed on a closed boundary surface S. The latter is expressed by the so-called Kirchhoff-Helmholtz integral:

$$P(\mathbf{r},\omega) = \oint_S \frac{1}{4\pi}\left[P(\mathbf{r}_s,\omega)\frac{\partial}{\partial\mathbf{n}}\left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|}\right) - \frac{\partial P(\mathbf{r}_s,\omega)}{\partial\mathbf{n}}\left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|}\right)\right] dS, \tag{5.6}$$

where $P(\mathbf{r},\omega)$ is the Fourier transform of the sound pressure $p(\mathbf{r},t)$, k is the wave number, $\mathbf{r}$ is the coordinate vector of an observation point and $\mathbf{r}_s$ is the coordinate vector of the integrand functions on the surface S. The underlying geometry is illustrated in Figure 5.7(b). The first part of this expression represents a distribution of dipoles with the source strength given by the sound pressure, measured on the surface S. The second term represents a distribution of monopoles, whose strength is given by normal particle velocity component of a sound field, which is proportional to $\partial P/\partial\mathbf{n}$.

In practice, the Kirchhoff-Helmholtz integral states that at any listening point within the source-free volume V the sound pressure $P(\mathbf{r},\omega)$ can be calculated if both the sound pressure and its gradient are known on the surface enclosing the volume. This can be used to synthesize a wave field within the surface S by setting the appropriate pressure distribution $P(\mathbf{r}_s,\omega)$. This fact is used for WFS based sound reproduction. If the surface S degenerates to a plane $z = z_1$, separating the listening area from the primary source area, as shown in Figure 5.8(a), then Equation (5.6) can be written as the Rayleigh II Integral [131]:

$$P(\mathbf{r},\omega) = |z - z_1| \int_{S_1} P(\mathbf{r}_s,\omega)\frac{1 + jk|\mathbf{r}-\mathbf{r}_s|}{2\pi|\mathbf{r}-\mathbf{r}_s|^3}\, e^{-j|\mathbf{r}-\mathbf{r}_s|} dS_1. \tag{5.7}$$

An auditorium oriented geometry for (5.7) is shown in Figure 5.8(b), where the surfaces S_2, S_3 and S_4 are fully absorptive. The wave field in the listening area can be generated by a secondary source distribution at z_1, each secondary source represents a dipole, the source signal of which is given by the primary sound pressure at its location.

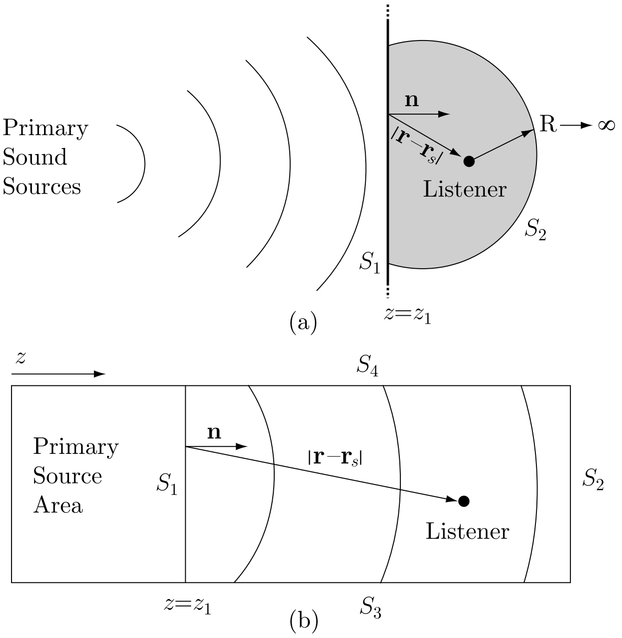


Figure 5.8. (a) Simplification of the half-space Kirchhoff-Helmholtz integral. (b) Auditory oriented geometry.

Hence, it is possible to physically synthesize the wave fronts at any listening point by reradiating the sound pressure, recorded by microphones at $z = z_1$, with loudspeakers having dipole characteristics, as indicated in Figure 5.9(a).

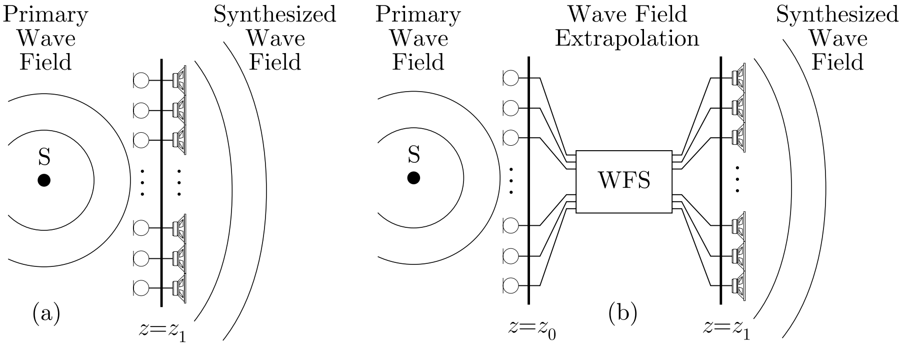


Figure 5.9. (a) Illustration of practical WFS according to Equation (5.6). (b) Generalization of the diagram in (a): wave field extrapolation prior to wave field emission.

A further step is to place the arrays of transducers used for recording and synthesizing the wave fronts in planes with different coordinates z_0 and z_1, respectively, as shown in Figure 5.9(b). Then, using Equation (5.7) again, the microphone signals should be transferred to the loudspeakers through a processor simulating the wave front propagation from z_0 to z_1 numerically. This process is the so-called *extrapolation.* In this configuration, loudspeaker positions $\mathbf{r}_n$ at $z = z_1$ act as virtual "listener positions" and thus, the driving signal for each loudspeaker at $z = z_1$, $P(\mathbf{r}_n, \omega)$, is calculated by processing the pressure signals $P(\mathbf{r}_l, \omega)$ recorded by all microphones at $z = z_0$ according to the Rayleigh II integral.

The WFS concept can be compared to optic holography: first, the optical wavefield is recorded over a plane, and later it is recreated by a distribution of light sources, placed on this plane. In sound holography, the acoustic wavefield is recorded over a plane S given by a planar microphone array. Wave field reproduction is then made by secondary sound sources, separately driven loudspeakers. Instead of an ideal continuous distribution of secondary sources, a discrete distribution is used, which leads to artifacts on the reproduction stage that will be addressed in Section 5.4.3.

First Rayleigh Integral Scheme

The first Rayleigh integral (*Rayleigh I*) states that the wave field in the listening half space can be reconstructed from the original sound field by measuring only the *particle velocity* in the measurement plane and using these measurements as source signals for a distribution of *monopoles* on the reproduction plane [132]:

$$P(\mathbf{r},\omega) = \frac{jk\rho c}{2\pi} \iint_S u_n(\mathbf{r}_s,\omega)\,\frac{e^{-j|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|}dxdy, \tag{5.8}$$

or in its discretized form:

$$P(\mathbf{r},\omega) = \frac{jk\rho c}{2\pi} \sum_n u_n(\mathbf{r}_n,\omega)\,\frac{e^{-j|\mathbf{r}-\mathbf{r}_n|}}{|\mathbf{r}-\mathbf{r}_n|}\Delta x\Delta y, \tag{5.9}$$

where the index n indicates the sampling points in the plane S. The secondary sources can be built with small monopole loudspeakers with volume velocity:

$$U_n(\omega) = u_n(\mathbf{r}_n,\omega)\Delta x\Delta y. \tag{5.10}$$

Notice that for a dynamic loudspeaker, above its resonant frequency, the volume velocity is related to the input voltage $E_n(\omega)$ at its moving coil by

$$U_n(\omega) = \frac{K_m}{j\omega}E_n(\omega), \tag{5.11}$$

where K_m is a constant which depends on the electro-mechanical properties of the loudspeaker systems. From the last two equations:

$$E_n(\omega) = \frac{j\omega}{K_m}\Delta x\Delta y u_n(\mathbf{r}_n,\omega). \tag{5.12}$$

Note that as a consequence of the spatial integration performed by the loudspeaker array, a 6 dB/oct filter is required before applying the excitation signals to the loudspeakers.

Second Rayleigh Integral Scheme

The second Rayleigh integral (*Rayleigh II*) gives a similar relation between the *sound pressure* in the measurement plane and the distribution of *dipoles* in the reproduction plane:

$$P(\mathbf{r},\omega) = \frac{jk}{2\pi} \iint_S P(\mathbf{r}_s,\omega)\frac{1+jk|\mathbf{r}-\mathbf{r}_s|}{jk|\mathbf{r}-\mathbf{r}_s|}\cos\phi\frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|}dxdy, \tag{5.13}$$

with $\cos\phi$ defined as $|z - z_s|/|\mathbf{r} - \mathbf{r}_s|$, where $|z - z_s|$ is the distance between the reproduction plane and the observation point. The discrete form is given by:

$$P(\mathbf{r}, \omega) = \frac{jk}{2\pi} \sum_n P(\mathbf{r}_n, \omega) \frac{1 + jk|\mathbf{r} - \mathbf{r}_n|}{jk|\mathbf{r} - \mathbf{r}_n|} \cos\phi_n \frac{e^{-jk|\mathbf{r}-\mathbf{r}_n|}}{|\mathbf{r} - \mathbf{r}_n|} \Delta x \Delta y. \tag{5.14}$$

In this case, the secondary sources can be built with small unbaffled loudspeakers with volume force:

$$F_n(\omega) = P(\mathbf{r}_n, \omega)\Delta x \Delta y. \tag{5.15}$$

The relation between the volume force and the input voltage above the mechanical resonance frequency is given by:

$$F_n(\omega) = K_d E_n(\omega), \tag{5.16}$$

where K_d is an electro-mechanical constant of the dipole loudspeaker systems. From the last two equations:

$$E_n(\omega) = \frac{1}{K_d} \Delta x \Delta y P(\mathbf{r}_n, \omega). \tag{5.17}$$

Notice that, with respect to the Rayleigh I result, no frequency weighting is needed in this case.

5.4.2 Derivation of the Driving Signal Function

For the derivation of the loudspeaker driving signal, the geometry shown in Figure 5.10 is considered. The pressure field of a virtual source, also known as *notional source* at $\mathbf{r}_m$ in the plane $z = z_0$, should be reconstructed in the horizontal ear plane of a listener located at $\mathbf{r}$ in the plane $z = z$, using a linear array of loudspeakers parallel to the x axis in the plane $z = z_1$. The line connecting source and listener conforms an angle θ with the z axis. Note that source, array and listener are all located in the plane $y = 0$, that is, at the same height. In practice, the loudspeaker array will often be mounted above stage and audience levels. In [131], it was shown that the array height y_1 can be neglected when this height is much smaller than the horizontal distances between source and array, and between listener and array, which is often the case.

The derivation given here is a generalized version of the one given in [131]. According to Rayleigh's theorem, the loudspeaker driving signals can be written as a weighted version of the pressure field of the notional source at the array position.

$$Q(\mathbf{r}_n, \omega) = A_n P(\mathbf{r}_n, \omega) = A_n S(\omega) \frac{e^{-jk|\mathbf{r}_n - \mathbf{r}_m|}}{\mathbf{r}_n - \mathbf{r}_m}, \tag{5.18}$$

where A_n is a weighting function, which depends on the lateral position of the nth loudspeaker $A_n = A(x_n, \omega)$. For synthesizing a spherical wavefront, its mathematical formulation must be considered:

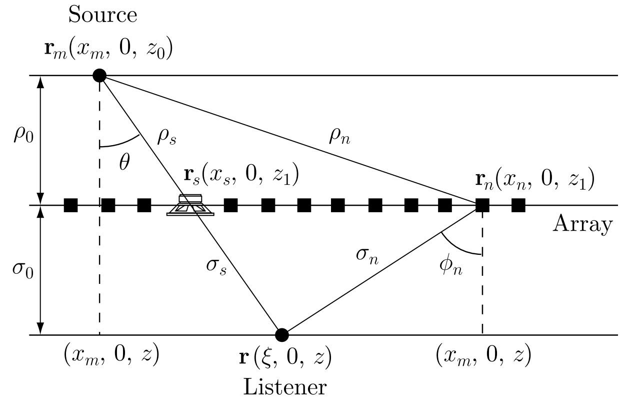


Figure 5.10. Configuration for WFS. Loudspeaker array at $z = z_1$ synthesizes wavefield of a source at $\mathbf{r}_m$ in the receiver plane at $z > z_1$.

$$P(\mathbf{r}, \omega) = S(\omega) \frac{e^{-jk|\mathbf{r}-\mathbf{r}_m|}}{\mathbf{r} - \mathbf{r}_m}, \tag{5.19}$$

where $S(\omega)$ is the spectrum of the notional source. According to the Rayleigh equation, the spherical wavefront can be synthesized as

$$P(\mathbf{r}, \omega) = \sum_{n=1}^{N} \left[Q(\mathbf{r}_n, \omega) G(\phi_n, \omega) \frac{e^{-jk|\mathbf{r}-\mathbf{r}_n|}}{|\mathbf{r} - \mathbf{r}_n|} \right] \Delta x, \tag{5.20}$$

where N is the number of loudspeakers in the array, $Q(\mathbf{r}_n, \omega)$ is the driving signal of the nth loudspeaker, ϕ_n is the angle between the main axis of the nth loudspeaker and its connection line to the listener position, and Δx is the spatial interval between the array elements [133]. Note that the only unknown elements in the synthesis operator are the driving signals of the loudspeakers. Using the geometry of Figure 5.10, the latter expression can be written as:

$$\frac{e^{-jkr}}{r} = \sum_{n=1}^{N} \left[A_n G(\phi_n, \omega) \frac{e^{-jk(\rho_n + \sigma_n)}}{\rho_n \sigma_n} \right] \Delta x. \tag{5.21}$$

This discretized integral equation can be approximated by using its stationary-phase representation [134]. Physically this approximation means that the wavefront is synthesized by all loudspeakers of the array together, but a dominant contribution is given by the loudspeaker positioned at the point of stationary phase. After substantial mathematical manipulation, the driving signal $Q(\mathbf{r}_n, \omega)$ of the nth loudspeaker can be found as:

$$Q(\mathbf{r}_n, \omega) = S(\omega) \frac{\cos\theta_n}{G(\theta_n, \omega)} \sqrt{\frac{jk}{2\pi}} \sqrt{\frac{|z - z_1|}{|z - z_0|}} \; \frac{e^{-jk|\mathbf{r}_n - \mathbf{r}_m|}}{\sqrt{|\mathbf{r}_n - \mathbf{r}_m|}}, \tag{5.22}$$

This driving function contains a cylindrical spatial divergence factor $|\mathbf{r}_n - \mathbf{r}_m|^{-1/2}$ and thus, the driving signal of the nth loudspeaker can be interpreted as a weighted version of the sound pressure field at $\mathbf{r}_n$ caused by a notional line source at $\mathbf{r}_m$. The fact that only one horizontal line in the reconstruction plane $z = z_1$ is used in the wavefront synthesis process is “compensated” for by the spatial extension of the notional source from a point to a vertical line. As shown in [133], any loudspeaker type can be used to form an array for WFS. By adapting the driving signal function according to Equation (5.22) the loudspeaker directivity characteristics are virtually transferred to the notional sources. Recently, in [135], the theory of WFS was revisited and a unified theoretical framework covering arbitrarily shaped loudspeaker arrays for two- and three-dimensional reproduction was presented.

Since the formulation of the theoretical framework of WFS, the spatial audio research community has contributed with various research projects, such as the European Carrouso [136],[137], original work in the technical literature and a number of PhD theses [138], [139], [140], [141], [142], [143], [144], [145], [146], [147].

5.4.3 Potential and Constrains of WFS

The principle characteristic of WFS is that the acoustic scene remains constant for the entire listening area, that is, the absolute setup of the acoustic scene is independent of the listening position. The relative acoustic perspective as perceived by listeners changes with their movements. This change involves a realistic change of the sound pressure level when the distance to the virtual source is varied.

The theoretical capabilities of WFS to create a quasi-realistic sound field or to recreate an existing sound field are even larger. For instance, it is possible to simulate a certain directivity characteristic of the virtual source. Furthermore, the location of the secondary sources (loudspeaker array) is no limitation for the creation of virtual sources. WFS theoretically allows the synthesis of virtual sources both in front of and behind the array. In particular, the creation of the so-called focused sources, which are sources in front of the array, makes a significant difference to conventional sound reproduction techniques, as will be shown in Section 5.4.4. From a creative point of view, WFS offers an improvement of flexibility: both direction and location stable sources can be reproduced. The design of the acoustic scene is less limited by the constraints of the reproduction technique in comparison to stereo. The simulation of a real acoustic scene is more plausible. Moreover, although arrays of conventional cone loudspeakers have a considerable visual impact on the reproduction room, there are powerful alternatives for WFS reproduction using Distributed Mode Loudspeakers (DMLs) . With this technology, several transducers can be applied to a flat panel, forming a Multiactuaor Panel (MAP) suitable for being used in advanced WFS setups [148].

However, since the aim of WFS is the creation of a true copy of a natural sound field, this high aim can only be fulfilled with certain restrictions in practice. Practical implementation of the WFS technique is based on loudspeaker arrays, which act as secondary sound sources. The distribution of these sources is not densely and infinitely continuous, but a finite set of band-limited signals will drive the individual discrete loudspeakers, which in turn, makes the array finite. These two effects limit the performance of real WFS systems:

1. *Spatial Aliasing*
   The distance between transducers Δx defines a spatial sampling frequency for a wavefield at a recording level. Then, the reconstructed wavefield will be physically correct up to the Nyquist frequency:

$$f_{\mathrm{nyq}} < \frac{c}{2\Delta x} \quad \leftrightarrow \quad \Delta x < \frac{\lambda_{\min}}{2}, \tag{5.23}$$

where λ_{min} is the smallest sound wavelength of concern. Above the Nyquist frequency, together with the correctly reconstructed wave front, additional undesired wave energy will be emitted incorrectly. A description of these aliasing artifacts can be found in [149], [150], [151].

In practice, wave field recreation without spatial aliasing artifacts is possible for frequencies less than the spatial aliasing frequency. This limit frequency is determined by the time difference between two successive loudspeaker signals interfering at the listeners' position. This time difference depends on the spatial sampling interval, that is, the loudspeaker and microphone inter-spacing. Moreover, the maximum wavelength being sampled correctly without spatial aliasing depends on the maximum angle on the microphone side. Accordingly, the maximum wavelength being received correctly without aliasing also depends on the maximal angle on the receiver side. This aliasing frequency is given more generally by

$$f_{\text{al}} = \frac{c}{2\Delta x|\sin\alpha_{\text{max}}|}, \tag{5.24}$$

where α_{max} indicates a maximum angle between an incident plane wave and a microphone array surface. If the angle α is equal to 0°, the wave front is perpendicular to the array surface and the spatial sampling interval Δx can be seen as infinite. In the worst case, when the angle α is equal to 90°, f_{al} will be equal to f_{nyq}. Assuming a loudspeaker spacing Δx=10 cm, the minimum spatial aliasing frequency is f_{al}=1.7 kHz. Regarding the standard audio bandwidth of 20 kHz, spatial aliasing seems to be a problem for practical WFS systems. Fortunately, the human auditory system is not very sensitive to these aliasing artifacts.

When virtual sound sources are recreated by means of WFS, the angle α_{max} can be set to a certain value. Radiation of plane waves at a wider angle than α_{max} is then suppressed and spatial aliasing effect can be avoided up to f_{al} frequency. The same technique can be applied for the wave-field recording, where directional microphones will capture waves radiating up to this certain angle.

2. *Truncation Effects*
In theory, the synthesis of the wave field arises from the summation of an infinite number of loudspeaker signals. In practice, however, the loudspeaker array used will always have a finite length. The finite array can be seen as a window, through which the primary or virtual source is either visible, or invisible, to the listener. Hence, an area exists which is "illuminated" by the virtual source, together with a corresponding "shadow" area [152]. Applying this analogy, diffraction waves are originated from the edges of the finite loudspeaker array. These error contributions appear as after-echoes for virtual sources and pre-echoes for focused sources. Depending on the level and time-offset at the receiver's location of the aliased contributions, it may give rise to colouration. This effect can be successfully minimized if a weighting function is applied to the signals driving the loudspeaker array. At the same time, decreasing the contribution of edge loudspeakers will reduce an area with a correctly reproduced wave field. Thus, the choice of the weighting function will depend on a trade-off between the reduction of diffraction artifacts and the size of the listening area.

5.4.4 Special Properties of WFS

Localization of Virtual Sources

Through WFS, the sound engineer has a powerful tool to design a sound scene. One of the most important properties, with respect to conventional techniques, is its outstanding capability to provide a realistic localization of virtual sources. Typical problems and constraints of a stereophonic image vanish in a WFS sound scene. In contrast to stereophony, WFS can produce a number of source stimuli, based on virtual sources and plane waves. These sources are localized on the same position throughout the entire listening area so listeners can move without losing their localization. In Figure 5.11, the arrows indicate the directions of the auditory events when virtual point sources and plane waves are reproduced.

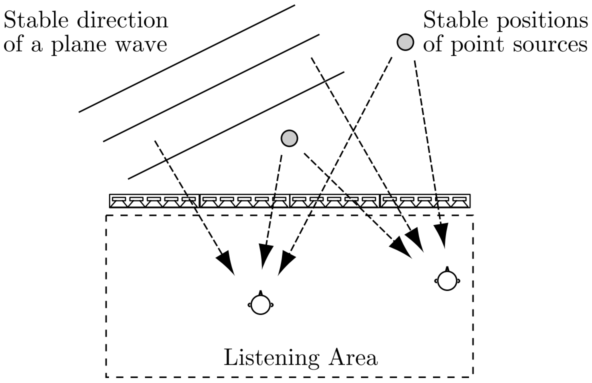


Figure 5.11. WFS is capable of reproducing both the stable positions of point sources and the stable direction of a plane wave.

WFS can enhance the localization of virtual sources and the sense of presence and envelopment through a very convincing reproduction of sound scenes. As spherical secondary sources are employed instead of cylindrical sources, some errors can be found in the variation of the pressure with the distance to the loudspeaker array. However, these errors have shown to be difficultly perceivable.

Subjective experiments on sound localization, correspondence of perceived auditory and visual source direction can be found in [153].

These properties enable the synthesis of complex sound scenes which can be experienced by the listener while moving around within the listening area. Figure 5.11 illustrates the way in which the sound image changes at different listening positions. This feature can be deliberately used by the sound engineer to compose new spatial sound design ideas [154]. Moreover, it has been shown that the enhanced resolution of the localization compared with stereophony enables the listener to easily distinguish between different virtual sources, which makes the sound scene significantly more transparent.

Typical implementations of sound field reproduction systems do not take the Doppler Effect into account. However, the Doppler Effect, both for moving virtual sound sources and inherently also for moving listeners, can be accurately reproduced in WFS [155], [156].

Virtual Sources in Front of the Loudspeaker Array

Figures 5.12(a) and (b) show the wave fronts of a point source behind the array and in front of the array, respectively, in a simulation. The concave wave fronts of Figure 5.12(a) achieve the synthesis of the signal of a virtual source behind the array. However, WFS is also capable of synthesizing a virtual source in front of the array. Therefore, the WFS array emits convex wave fronts which focus on a point that will be the "focused source", illustrated in Figure 5.12(b). Naturally, the localization will not be correct for listening positions between the focus point and the array because the sound emission of the virtual source occurs here reversely.

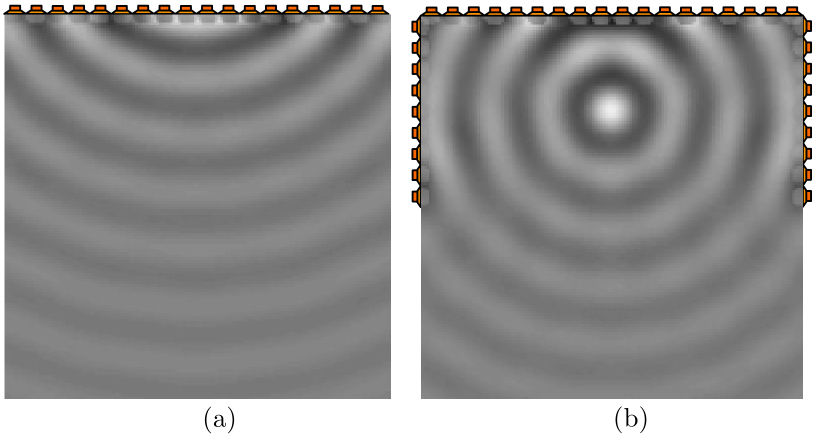


Figure 5.12. (a) Virtual source behind the array. (b) Virtual source in front of the array, also known as focused source.

For practical application, it is an enormous progress that virtual sound sources can be created in the field between the listener and the loudspeakers. Sound engineers can be offered completely new tools for spatial sound design. Moreover, the reproduction of focused sources with directional characteristics is also possible but with a limited listening area [135], [156].

5.5 Audio Up-Mixing and New Coding Schemes

5.5.1 Audio Up-Mixing

Despite the advances in spatial sound reproduction, the availability of multichannel audio recordings is still very limited nowadays. While recent movie soundtracks and some musical recordings are available in multichannel format, most music recordings are mixed in stereo. The playback of this material over a multichannel system poses a fundamental problem: stereo recordings are mixed with a very particular set up in mind, which consists of a pair of loudspeakers placed symmetrically in front of the listener. Thus, listening to this kind of material over a multichannel sound system raises the question of what signals should be sent to the additional channels. In this context, audio systems aimed at solving the up-mixing problem are widely used today, for example, Dolby Pro Logic II [157], Logic7 [158], Neo:6 [159], or CircleSurround [160].

Dolby began introducing the multichannel to stereo down-mix feature in its codecs in order to respond to the requirements of backwards compatibility. Additionally, Dolby Pro Logic II includes the up-mix from stereo back to 5.1 multichannel format. In the down-mix, the original source audio signals are encoded into two program channels, that can be played back as stereo.

The left and right stereo signals, called left-total and right-total, or Lt and Rt, are assembled by adding to the left and right multichannel signals (L and L) the center channel signal (C) as well as the corresponding surround channel signal (LS or RS), both attenuated by 3 dB. The phases of the surround channel signals are additionally shifted by 90 degrees and they are added with opposite signs. Similarly, the up-mix is carried out using the following decoding matrix:

$$\begin{bmatrix} L \\ R \\ C \\ LS \\ RS \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \mathrm{j}\frac{1}{\sqrt{\frac{3}{2}}} & -\mathrm{j}\frac{1}{\sqrt{3}} \\ \mathrm{j}\frac{1}{\sqrt{3}} & -\mathrm{j}\frac{1}{\sqrt{\frac{3}{2}}} \end{bmatrix} \cdot \begin{bmatrix} Lt \\ Rt \end{bmatrix}, \tag{5.25}$$

The LFE channel signal is derived by low-pass filtering the sum of Lt and Rt signals.The LFE channel signal is derived by low-pass filtering the sum of Lt and Rt signals.

Alternatives to simple matrix-based up-mixing systems have also been proposed. For example, Avendano and Jot developed more advanced frequency domain techniques for the up-mix of stereo recordings into multichannel audio [161]. Aiming at a natural and generic multichannel audio mix, their method takes into account both the apparent directions of individual sound sources, and the ambient sound consisting of diffuse sound, reverberation and noise. The method compares the STFT of the left and right stereo signals and identifies a set of components for the up-mix in a similar fashion as described in Section 2.6.2.

The design aim of a high quality up-mixer can be summarized with three general goals relating to spatial imagery [162]. These goals are related to modifying the listening experience of a conventional loudspeaker-pair reproduction of a musical recording:

1. To create a source image with a spatial quality similar to the original 2/0 mix.
2. So as to create a natural-sounding ambiance (reverberance) image.
3. To create a listening experience which people would prefer over the original 2/0 listening experience.

The third goal is assumed subservient to the first two; such high quality up-mixers are not intended as a special effect which reinterprets the mixing intention of the recording producer, but rather as a system to compliment these intentions in ways which are consistent with sound in the natural environment. Notice that this design criteria are very suitable for stereo-to-surround up-mixers, but they do not take into account the object-based conception of advanced reproduction systems such as WFS.

5.5.2 Spatial Audio Object Coding

The typical audio production and transmission chain consists of a set of operations that are executed in a very specific order. For example, for musical content, various audio objects (instruments, vocals, etc) are first recorded (or synthetically produced), and subsequently mixed for playback on a specific reproduction system. The mixing process is performed by an audio engineer who decides on object positioning, relative levels and effects that are employed according to aesthetical and technical objectives. In many applications, the resulting mix is transmitted using lossy compression algorithms.

This conventional chain leaves virtually no flexibility in changing the composition at the reproduction side. A similar limitation holds for multiple-talker communication systems (teleconferences).

There exists a range of applications that can benefit from user-control of various audio objects at the playback side. Examples are teleconferencing, remixing applications, on-line gamin, and karaoke functionality. Although such functionality can be obtained by transmitting all objects independently, this scenario is undesirable due to large bandwidth requirements and the fact that it is difficult to guarantee a certain aesthetical quality level, which is extremely important in the music industry.

Following the recent trend of employing parametric enhancement tools for increasing coding or spatial rendering efficiency, *Spatial Audio Object Coding* (SAOC) is one of the recent standarization activities in the MPEG audio group [163]. SAOC is a parametric multiple object coding technique that is aimed at overcoming the above drawbacks. It is designed to transmit a number N of audio objects in an audio signal that comprises K down-mix channels, where $K < N$ and K is typically one or two channels. Together with this backward compatible down-mix signal, object meta data is transmitted through a dedicated SAOC bitstream to the decoder side. Although this object meta data grow linearly with the amount of objects, the amount of bits required for coding these data in a typical scenario is negligible compared to the bit-rate required for the coded down-mix channels.

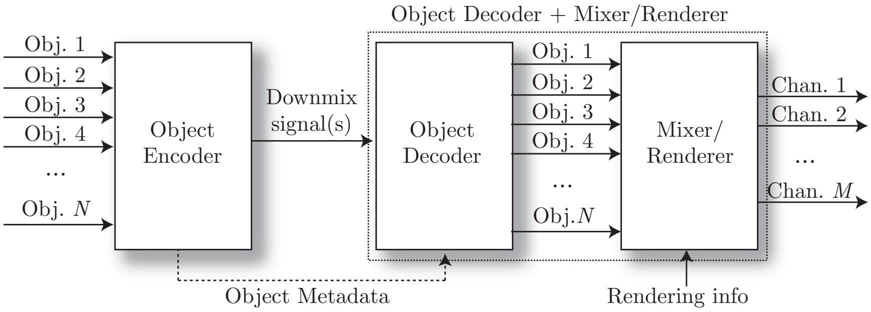


Figure 5.13. SAOC block diagram.

In a conceptual overview, as illustrated in Figure 5.13 (reproduced from [163]), the decoder side can be divided into an object decoding part decomposing the N objects and a rendering part, that allows manipulation and mixing of the original audio object signals into M output channels. For those processes, the object decoder requires the object meta data, while the renderer requires object rendering information. The decoding and rendering can be performed in an integrated fashion, avoiding a costly intermediate up-mix to N discrete audio object signals.

SAOC provides an object-based conception of sound scenes, which is one of the general aims pursued by SSS algorithms in this thesis. However, it must be clarified that the SAOC framework is just an efficient way of coding multiple source signals so that object manipulation is possible at the decoding stage. Therefore, the original source signals are needed in the SAOC scheme in order to correctly extract the required side information. The problem of having the independent source signals appears here again, thus, SSS methods may find more applications in this coding framework.

5.6 Conclusion

This chapter has presented an overview of spatial sound reproduction systems. The goal of audio reproduction has always been to accurately recreate a given sound scene in terms of source localization. Thus, this is the main characteristic of spatial audio systems. To achieve this goal, these systems are designed in a way that they provide the listeners with those spatial localization cues that are necessary for stimulating the presence of a localized sound source in the auditory system, a challenge that involves both psychoacoustical and physical issues. The chapter has described which are the main cues used by humans to localize sound sources and how basic spatial sound systems as stereo have evolved to sophisticated multichannel sound reproduction techniques based on sound field rendering. Despite the advances observed in the spatial audio field, most audio material is still intended to be reproduced over a two-channel system. This fact has motivated the design of audio up-mixers and new coding schemes, giving the listeners the possibility to experience the real advantages of spatial sound reproduction.

Sound Scene Resynthesis 6

Sound Scene Resynthesis 6

Most of the recorded material is intended for stereo reproduction and, therefore, adapting this material to WFS is a challenging issue. Sound Source Separation techniques can be used to extract the sources present in a stereo mixture, obtaining separated tracks which can be used for reconstructing spatially enhanced scenes. However, timbral distortion and inter-source residuals are limiting factors that can degrade the perceived quality. In this chapter, the quality of several acoustic scenes (music and speech) is evaluated in three different situations: stereo image, WFS scene with original sources and WFS scene with separated sources from stereo mixtures. A study on the change of perceptual attributes among the different cases has been conducted by means of listening tests, being this the first study ever reported on the combination of WFS reproduction and source separation.

6.1 Introduction

The previous chapter discussed the evolution of spatial sound reproduction systems over the last decades. Unfortunately, despite the advances made in the field, there are still many problems concerning the availability of suitable audio material for the new reproduction formats. In this context, audio up-mixers were presented as a solution for reproducing stereo recordings over surround systems. This way, the advantages of multichannel audio reproduction can be exploited when listening to stereo material. However, while these systems are focused on 5.1 reproduction schemes, few solutions have been proposed for reproducing stereo material over sound field rendering techniques such as WFS.

As WFS systems are not yet widely deployed, up-mixing processors fully designed for converting stereo recordings into synthesized scenes have rarely been discussed in the literature. The main objective of stereo-to-WFS up-mixers would be the same as those developed for five-channel up-mixing: to enhance the spatial quality of conventional stereo recordings. However, the spatial properties of WFS open a new door to go further than the conventional up-mixing scheme. The channel-based conception inherent to current up-mixers must change to an object-based conception, which is indeed a difficult problem to overcome. In this context, sound source separation (SSS) techniques provide a promising solution to deal with WFS up-mixing.

The use of SSS techniques for audio up-mixing is one of the most challenging audio oriented applications of source separation. The combination of SSS with spatial sound reproduction implies also new evaluation schemes for separation algorithms, since spatial attributes play a major role in the overall perceptual quality of sound scenes. These attributes are supposed to be related to meaningful features in spatial sound perception such as localization accuracy or the perceived source width, and provide a way to assess the perceptual changes that occur between different spatial sound reproduction systems.

The purpose of this chapter is to evaluate the subjective quality and spatial attributes of synthesized acoustic scenes in WFS when the virtual sources are generated using separated tracks from stereo mixtures. Although WFS has its own artifacts which degrade the perceived quality due to practical imperfections, the degradation caused by the separated sources used for rendering the sound field is far greater. These degradations include timbre modification and burbling artifacts, musical noise due to spectral substraction and inter-source residuals [164]. However, the masking effects involved in the rendering process usually make these artifacts less perceptible when the whole scene is being reproduced. Several spatial attributes in three different situations are studied: stereo image, WFS scene with original sources and WFS scene with separated sources using different separation methods. Listening tests are conducted in order to evaluate the changing of these attributes over each case. The results of this study were recently published in [165].

The chapter is structured as follows. In Section 6.2, the resynthesis of sound scenes is discussed, justifying the necessity of carrying out an object-based up-mixing for WFS systems. In Section 6.3, the perceptual attributes usually evaluated in other works with regard to spatial sound reproduction and perceived quality are summarized. Section 6.4 presents the experimental setup and the evaluation process followed in in this thesis, discussing the results obtained from the listening tests in Section 6.5. Finally, in Section the conclusions of this chapter are summarized.

6.2 Resynthesis of Sound Scenes

A sound scene is defined as a complex entity of acoustic stimuli that is processed by our brain, resulting in a symbolic description corresponding to a listener's perception of the different sound events and sources that are present [10]. The reproduction of sound scenes must not only be restricted to source localization. In fact, the recreation of the diffuse field in the reproduced acoustic environment has also special relevance in the perceived spatial impression of the sound scene. Fortunately, practical audio rendering systems do not need to take every reflection in a room into account in order to give a convincing impression of reverberation [166]. Although "naturalness" may be a good attribute of a reproduced sound scene, the objective of many commercial recordings is not spatial fidelity, which is true to an original sound field. In fact, in a large number of commercial releases the mixing engineer does not attempt to recreate a scene similar to a "natural" reference. Here, "*the acoustic environment implied by the recording* *engineer and producer is a form of "acoustic fiction" or "acoustic art"...*" [167]. For this reason, sound scene reproduction should not be limited to the strict emulation of reality, but should give the impression the scene was conceived for.

The *resynthesis* of sound scenes refers to the task of reproducing a sound scene from audio signals which contain sound events. As stated before, the reproduction must be aimed at emulating an original sound scene, rearranging an original scene or creating a completely new scene. The complete scene is constructed by playing the sound events involved simultaneously by using any of the techniques described in Chapter 5. Therefore, the task of resynthesizing a sound scene involves having available the signals corresponding to the sound events that constitute the scene. Note that this object-based conception of a sound scene totally agrees with the mixing philosophy of WFS [168]. However, most of the times, having independent signals for all the audio events of a sound scene is not possible for many practical reasons. Ideally, the better option would be to extract all these signals from an available mixture containing all the sound events. This can be thought as an up-mixing process where audio material for common stereo systems is adapted to be reproduced over WFS from an object-based perspective.

6.2.1 Stereo to Wave Field Synthesis Up-Mixing

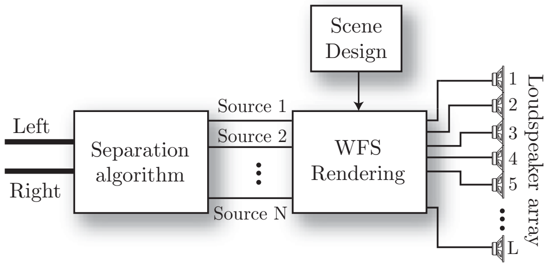


Figure 6.1. Stereo to WFS up-mixing scheme.

A commonly held assumption with regard to up-mixer design is that the sound imagery evoked must be consistent with that of a conventional two-loudspeaker sound scene, created using the same recording [162]. This criterion in the design is in fact very appropriate for home-theater oriented up-mixers, but lacks the flexibility to adapt all kinds of material to a WFS system. Section 5.4.4 described some of the special properties of WFS. To take advantage of this potential, a more complex up-mix is required. Sound scene identity between a recorded sound scene and reality should be subjected to the specifications of a given application. The up-mixing process must be thought of as a method for generating a perceptually convincing sound scene that can be parametrically controlled and be used in a wide range of multimedia applications. The *Tapestrea* framework follows this philosophy [169].

This freedom in the resynthesis of sound scenes can only be achieved if more sophisticated schemes are used. SSS algorithms are proposed in this thesis as a solution for isolating independent sound events in a stereo mixture and resynthesizing a WFS scene arbitrarily similar to the original stereo recording. A diagram of this up-mixing system is depicted in Figure 6.1: the left and right channels are the input signals of a SSS algorithm, which extracts a set of tracks corresponding to estimations of the original sources that were added in the stereo mix-down. These tracks feed the WFS rendering algorithm which drives the excitation signals corresponding to each unit of the loudspeaker array.

As explained in Chaper 2, current algorithms are far from achieving perfect separation and take many different approaches depending on the type of mixing used in the input mixtures. In fact, a severe degradation in the quality of the extracted sources in comparison to the original ones is usually found. This degradation is due to residuals from other sources in the mixtures, musical noise and burbling effects due to spectral substraction and non-linear filtering. Nevertheless, a perceptual improvement is usually found when the separated sound events are spatially mixed together to resynthesize the whole sound scene [170].

6.3 Sound Scene Evaluation

Several studies have been carried out in order to find out what listeners perceive when evaluating spatial audio signals, and which attributes have the most relevance in this context [167]. The objective measures used to evaluate audio coding systems do not take spatial properties into account. However, spatial aspects of sound certainly have a bearing on the overall quality score. This is the reason why new models and procedures are emerging for evaluating the perceived quality of spatial audio reproduction systems. The QESTRAL protocol has been specifically designed to take account of distortions in the spatial domain, allowing to perform assessments over a wide range of spatial distortions by means of an artificial listener [171].

An analysis carried out in [172] quantified the contribution of spatial fidelity to overall judgments of reproduced sound quality in a 5.1 channel surround context. The outcome of the analysis showed that spatial fidelity contributed a substantial component to the overall *basic audio quality* (BAQ) judgment, following the equation:

$$\mathrm{BAQ} = 0.80\,\mathrm{Tm} + 0.30\,\mathrm{Fr} + 0.09\,\mathrm{Sr} - 18.7 \tag{6.1}$$

where Tm, Fr and Sr are the Timbral, Frontal and Surround properties, respectively. This equation suggests that, although timbral fidelity plays the most important part of the BAQ rating, surround and especially frontal spatial fidelity are important too. Although Eq.(6.1) was derived from a 5.1 reproduction context, it is easy to surmise that the change of the spatial attributes in WFS scenes may contribute in some way to the perceived overall quality as well. An overview about the major perceptual effects for spatial reproduction and its relation to WFS is given in [173].

6.3.1 Perceptual Attributes of Spatial Sound

The following sections are centered on the spatial properties of synthesized sound scenes using source signals obtained by using SSS algorithms. The meaning and validity of spatial attributes for comparing audio reproduction systems is very important in this kind of work. According to Rumsey [174], spatial attributes that are meaningful should be identified, in the order of priority:

1. to individual subjects;
2. to a well-defined group of expert subjects forming a listening panel, and that agree upon a set of attributes to be graded;
3. to expert listeners not associated with that listening panel;

4. to independent observers or readers of the results.

Moreover, they should be unambiguous, unidimensional and also enable meaningful distinctions between the techniques or products under test. Rumsey also points out that the spatial attributes of importance are strongly dictated by the nature of the source material and the context or task in question:

- Dynamic descriptors have to be used if the scene changes with time.
- The source material should be chosen to reveal or highlight the attributes in question.
- The characteristics of subjects can influence their perception of spatial attributes, based upon their experience and education.
- Simple scenes are preferred, because they make the subjective task easier for a listener.

From the fourth issue, it can be said that simple stimuli are considered very important, as one can have a more accurate control and be clear about which subjective factors are affected. However, in this work, we are evaluating sound scenes from complex audio material, and questions about some attributes can become ambiguous. In order to address this need to evaluate complex reproduced source material, Rumsey proposes a *scene-based paradigm*, where the elements of the scene are grouped according to their function within the scene. Then, one could talk about individual sources, ensemble (a group of sources forming an entity in the scene) or environment-related attributes. *Micro* and *macro* attributes are then introduced. Micro attributes describe the features of individual elements within a scene, whereas macro attributes describe the scene as a whole, or grouping of elements within it.

Width and Aperture

The width attribute is always referred to the perceived width of a certain scene entity. Taking this scene-based evaluation of attributes into account, there are at least three different types of width attributes, listed from micro to macro [175]:

- Individual source width (ISW): Width of individual source(s) within the scene.
- Ensemble width: Overall width of a defined group of sources (may be all the sources in the scene if required).
- Environment width: Broadness of (reflective) environment within which individual sources are located.

The individual source width is related to the perceived width of isolated sources in a scene. The macro entity called ensemble is a group of sources that has a common cognitive label (orchestra, band, etc.). The ensemble width is then related to the perceived width of this entity. The ensemble aperture is defined in terms of the perceived angle of this entity. Environment width seems to be related to a perception of the reverberant sound within the reproduced space and is dependent on the ability to experience a sense of presence, but it will be not considered in this thesis. The concept of this scene-based evaluation is graphically illustrated in Figure 6.2.

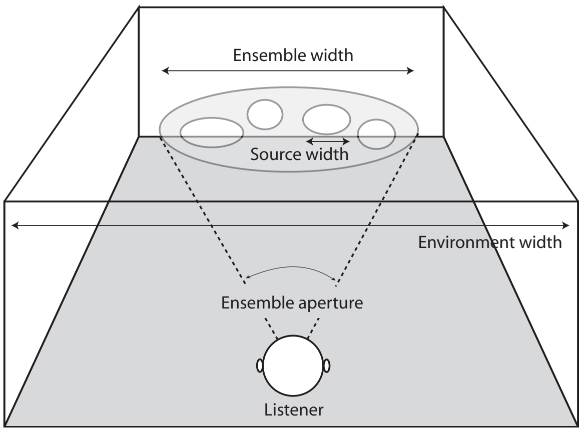


Figure 6.2. Scene-based evaluation of the width attribute.

Locatedness

The definition of *locatedness*, according to Blauert [1], is the degree to which an auditory event can be said to clearly be in a particular location. Although the ISW could be related to the locatedness of a certain source (it is easy to locate a source that has a small ISW), the relationship between these two attributes does not have to be necessarily correlated.

Localization Accuracy

Other localization measures, such as the *Mean Run standard deviation* $<\bar{S}>$ have been used in other studies [176]. This is calculated as follows. The standard deviation S is defined as the deviation of all assessments of one person and one stimulus. By averaging the standard deviations from all test items the *Run standard deviation*, $\bar{S}$, is calculated. Averaging all test subjects' Run standard deviations $\bar{S}$ results in the Mean Run standard deviation $<\bar{S}>$. This procedure may be regarded as valid if it is done with respect to a reference, a single loudspeaker, having a small source width, sharp focus and good locatedness.

Sound/Timbral Quality

The timbre or sound color, is one of the most important attributes describing a sound or sound reproduction technique. The sensitivity of humans to sound color is high [177]. The American Standards Association [178] definition for timbre is: "*Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar*". We are interested in how the perceived sound color of the different sources that compose a given sound scene changes when blindly separated sources from stereo material are used. In our case, this attribute is the one which may be most affected when separated sources are used for resynthesizing a sound scene. This coloration is mainly produced by the artifacts resulting from the separation process. An adaptation of the ITU-R BS.1534, broadly known as MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*), is proposed in [179] for a subjective evaluation of SSS algorithms. In [180], a modification of this method is also used for the subjective assessment of the coloration in WFS and stereo systems. Other modifications of the method for evaluating the sound quality of several up-mixing algorithms are used in [181]. Following the guidelines of [179], the original source images are used as reference signals and three anchors are considered to provide absolute quality ratings of interference, noise and artifacts. Interference anchors are obtained by adding a scaled version of the sum of the other source images to the original source image, and noise anchors by adding scaled white noise to the original source image. The scaling factors are defined so that the ratio between the loudness of the distortion signal alone and the loudness of the anchor equals 0.5. Artifact anchors are computed by setting randomly to zero half of the STFT points of the original source image signal (using a window length of 1024 samples and 50% overlap). Other features of MUSHRA, including the rating scale, are kept unchanged. The attributes previously described were considered for the subjective evaluation of the resynthesized sound scenes, as described in the next section.

6.4 Experiments

Three stereo mixtures were used for evaluating the spatial attributes described in the previous section after the separation and rendering of the stereo sound scene. The sound material used in the evaluation was selected in order to:

1. be representative of the potential application of SSS to WFS resynthesis: realistic upmixing scenarios (music and videoconference).

2. be short enough to avoid listener fatigue: as many attributes had to be evaluated, it was preferable to use a small but representative group of test items.

3. be representative of different types of common commercial music material: pop music with singing voice (people usually gives greater attention to the singing voice than to the rest of the instruments) and folk music (instrumental).

4. satisfy the underdetermined nature of common stereo material: all the scenes have four different sources and two mixture channels.

5. represent different disjointness conditions: the disjointness of speech is more balanced in both time and frequency, while the disjointness of music signals is more influenced by the frequency resolution [23].

In this section, we discuss the experiments carried out in order to evaluate different spatial attributes of WFS scenes constructed using different separation algorithms. Each scene is evaluated in a three stage process:

- Stereo mixing of a reference sound scene.

- Apply a source separation algorithm to the stereo mixture and objectively evaluate the separated sources as described in Section 2.8. However, as WFS accepts mono signals, only SDR, SIR and SAR considering a constant gain distortion were calculated.

- Resynthesize the sound scene in the WFS system and evaluate the spatial attributes of the scene.

6.4.1 Stage 1: Stereo Mixing of Reference Sound Scenes

The sound scenes used in the following experiments are three audio fragments (16 bit, 44.1 kHz) of music and speech with four different sources in each one. The first sound scene is a full instrumental music fragment. The second one is a pop music fragment with a singing voice and the last one is a mixture of four people talking in English and Japanese, obtained from the evaluation SASSEC campaign [67]. In each of the cases, the original sources were instantaneously mixed using the following mixing matrix:

$$\mathbf{A} = \begin{bmatrix} 0.99 & 0.89 & 0.71 & 0.32 \\ 0.14 & 0.44 & 0.71 & 0.95 \end{bmatrix} \tag{6.2}$$

where the energy preserving panning law was used (Eq.(2.2)). The mixing matrix (6.2) results in the following reference angles for the sources (Eq.(5.1), Figure 5.3):

$$\begin{aligned} s_1 &: \quad \theta_1 = 33.5^\circ \\ s_2 &: \quad \theta_2 = 15^\circ \\ s_3 &: \quad \theta_3 = 0^\circ \\ s_4 &: \quad \theta_4 = -22^\circ \end{aligned}$$

6.4.2 Stage 2: Source Separation and Objective Evaluation

The algorithms described in section 2.6 were applied to the mixtures (STFT framesize 2048 and overlap 75%). The parameters of the algorithms were tuned following the recommendations of the authors and selecting the ones that obtained best sounding. This was done by a small group of experts in an informal pre-test.

The results obtained from the objective evaluation of the separated sources are shown in Table 6.1. Analyzing the SDR values of the different algorithms in the three scenes, it can be said that, in general, the results are better for the folk music scene, probably because no percussive sounds are present in this mixture. Percussive sounds have a lot of energy in localized time frames (onset frames), causing a severe spectra overlap and reducing the overall disjointness of the sources. In addition, the three performance measures are dependent on the source panning position. This is especially noticeable for source s_2 in all the scenes, which has the smallest values in comparison to the results obtained for the other sources. The mean value of SDR, SIR and SAR throughout all the separated sources in the case of each separation algorithm is shown in Figure 6.3. As can be seen in the figure, SIR values are much higher than SDR and SAR for all the algorithms, reflecting the tradeoff between source isolation and high quality separation. DUET has the highest mean values of SDR and SAR, which may result in separated sources with less artifacts. However, it also has the lowest SIR, incrementing the amount of energy from other sources after the separation. The ADRess and PIW algorithms have higher SIR values, but a decrease in SAR and SDR can also be observed. The MuLeTS algorithm performs similarly to DUET, with a little improvement in SIR and a little degradation in SAR and SDR.

6.4.3 Stage 3: Resynthesis and Subjective Evaluation

Test set-up

The resynthesis of the different sound scenes was carried out using a 24 loudspeaker WFS array. This array is placed inside our recording studio, which is acoustically treated to get a

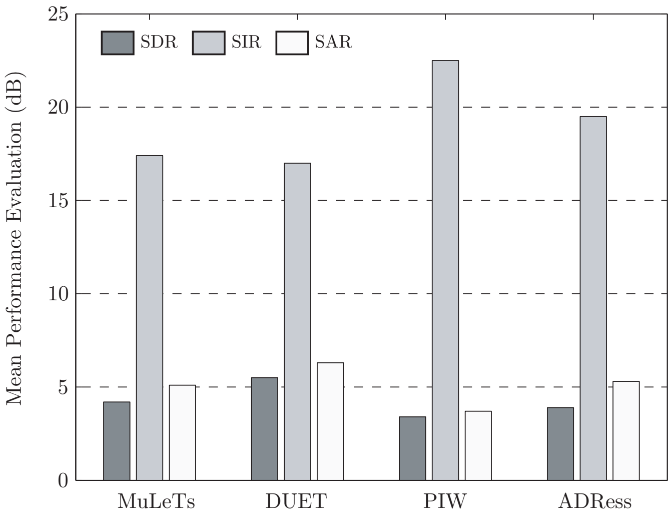


Figure 6.3. Mean objective evaluation results.

	MuLeTS			DUET			PIW			ADRess		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Scene 1: Folk												
s_1: guitar	6.4	20.0	6.6	7.7	18.6	8.1	5.8	22.1	5.9	7.7	21.1	8.0
s_2: accordion	1.0	13.8	1.2	0.8	14.2	1.1	0.6	19.1	0.7	0.7	16.1	0.9
s_3: sax	14.2	23.7	14.8	15.0	25.8	15.4	14.6	28.7	14.9	12.5	15.8	14.4
s_4: violin	10.2	32.2	10.2	12.0	32.6	12.1	5.6	41.9	5.6	7.6	46.4	7.6
Scene 2: Pop												
s_1: bass	2.6	17.6	2.8	5.7	23.5	5.8	1.0	26.3	1.0	2.2	26.3	2.2
s_2: drums	-3.1	5.3	-1.4	-5.4	2.3	-2.6	-7.0	4.0	-5.3	-6.0	3.1	-3.7
s_3: vocals	3.5	8.8	5.5	6.0	12.7	7.2	6.1	14.6	6.9	6.0	8.6	10.0
s_4: guitar	6.3	26.8	6.4	7.4	18.1	7.9	3.7	22.8	3.7	6.0	20.1	6.2
Scene 3: Speech												
s_1: speaker 1	2.0	24.0	2.0	5.3	16.2	5.8	3.0	28.0	3.0	3.9	19.1	4.1
s_2: speaker 2	-0.4	6.1	1.6	2.4	9.7	3.8	1.0	17.1	1.2	1.8	11.3	2.6
s_3: speaker 3	2.2	6.2	5.3	2.7	9.8	4.1	3.9	13.8	4.6	1.9	3.6	8.4
s_4: speaker 4	6.0	24.5	6.0	6.2	20.7	6.4	2.0	31.5	2.0	2.5	27.1	2.5

Table 6.1. Performance Evaluation Measures.

T_{60}@1000Hz < 0.25 s. The volume of this room is 96 m^3 an its floor size is 4 by 9.1 m. The background noise inside the studio was measured, obtaining SPL values below 25 dB(A). The experimental set up is depicted in Figure 6.5. Each element of the array is a two-way system, using a 5.5-in woofer and a 1-in tweeter. The loudspeaker separation of the array is 180 mm. The spatial aliasing frequency for this arrangement is about 1 kHz in the worst case. Figure 6.4 is a photograph of the system under test. No special aliasing improvement techniques were used in the rendering process. Single loudspeakers placed behind the array were used as a reference for the experiment. The loudspeaker model was the same as the one used in the array in order to keep the response unchanged. Stereo reproduction was carried out using two loudspeakers of the WFS array, marked as L and R in Figure 6.5. Following the procedure used in [180], an acoustically transparent curtain was used to hide the loudspeakers and avoid the ventriloquism effect. The subjects used a laser pointer to indicate the perceived direction of the events in a graduated marker line. Only azimuth angles were considered.

Seven different situations were subjectively evaluated: 1) Scene with real sources (loudspeakers placed in reference positions), 2) WFS scene with original sources, 3) Stereo scene resulting from amplitude panning of the original sources, 4-7) WFS scenes with separated sources using different separation algorithms.

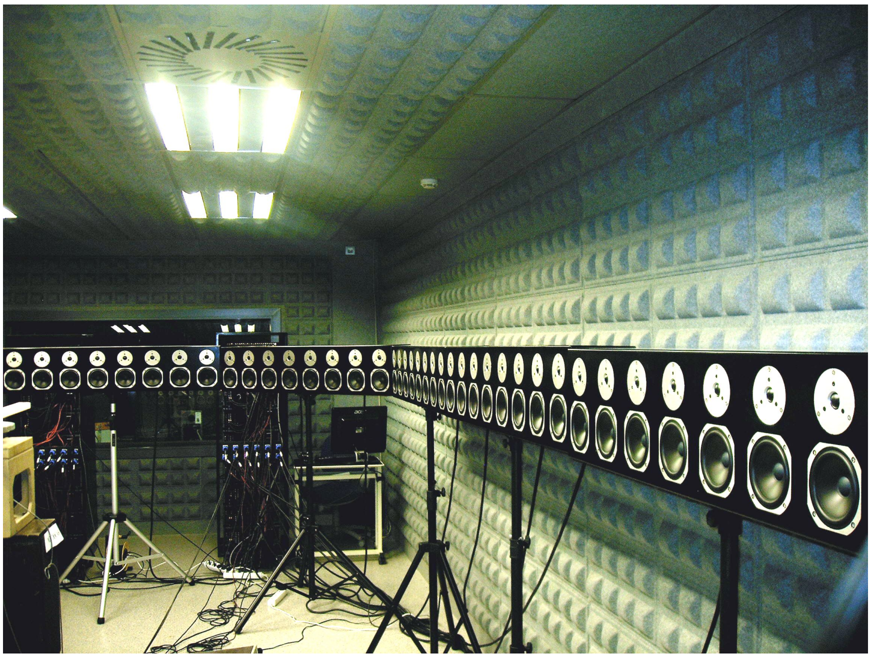


Figure 6.4. WFS array in the recording studio of the Universidad Politécnica de Valencia.

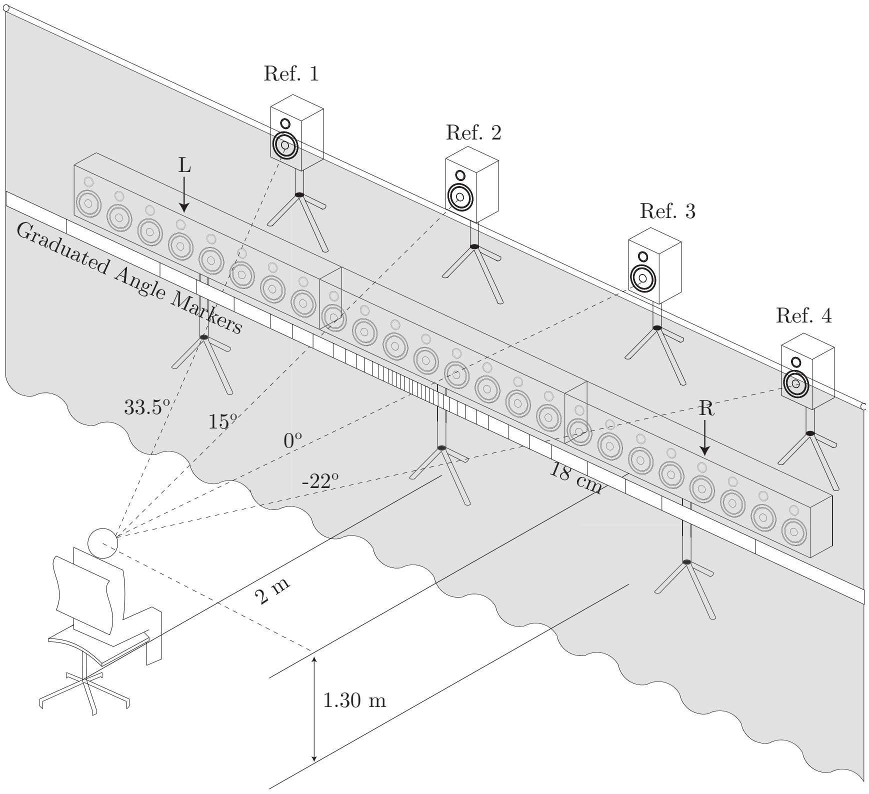


Figure 6.5. Location in the virtual sources in the WFS scenes.

The spatial arrangement of the WFS loudspeaker array and the location of the virtual sources that compose the reference mixtures is depicted in Figure 6.5.

Test procedure

The listening test is controlled by a laptop computer running a Matlab GUI. This Matlab GUI is communicated via Ethernet with the central computer which runs the WFS render software and stores all the test sound scenes. The subject is sitting in front of an acoustically transparent curtain and at the mid-point of the array, coinciding with the sweet spot of the stereo system (L and R loudspeakers in Figure 6.5). The results of the evaluation are stored automatically on the laptop computer. The complete test involves five different steps corresponding to the evaluation of the attributes described next.

Spatial attributes

The spatial attributes were evaluated as follows:

- **Locatedness**: A 5-grade scale was used, ranging from 1 (very good) to 5 (very bad). *How well can you localize the source? How well can you assign a particular direction to the perceived source?*

  1. very bad
  2. bad
  3. fair
  4. good
  5. very good

- **Localization accuracy**: The subjects were ask to point towards the perceived direction of the source. Single loudspeakers were placed in the reference directions in order to measure the deviation of the subjects' perceived directions. The reference angles were set up in the reference directions (Section 6.4.1). The mean run standard deviation was also calculated as a localization accuracy measure.

- **Source Widthness**: The scene-based paradigm was considered, evaluating source width and ensemble aperture separately. Other evaluation experiments have shown that it is difficult to train subjects in the individual source width attribute (ISW) [182][183]. In our case, although trained listeners were recruited for the listening tests, we found that they had problems in evaluating quantitatively ISW. Therefore, we preferred to evaluate the source widthness using a 5-grade scale, similarly to locatedness. Although widthness is not a recognized cue, we introduced this attribute with the aim of evaluating the difficulty found by the listeners in perceiving a defined lateral extent of the source. *How well can you assign a particular width to the perceived source?*

  1. very bad
  2. bad
  3. fair
  4. good
  5. very good

- **Ensemble Aperture**: All of the sources playing at the same time produce a single ensemble entity. The ensemble aperture was evaluated by asking the subjects to point out the perceived edges of the sound scene.
- **Sound/Timbral Quality**: A modified MUSHRA method was utilized as described in 6.3.1.

Test panel

Listening tests were conducted using a panel of 15 trained listeners, including a sound engineer, four musicians, a loudspeaker designer and people involved in audio research. Trained listeners are preferred when difficult aspects of sound have to be evaluated. However, a training session before starting the test was also carried out in order to explain to the subjects the meaning of the different spatial attributes they had to evaluate.

6.5 Results

The duration of the full test was around 1.5 hours but the listeners took breaks every half hour in order to avoid fatigue. Once the listening tests were finished, the data collected in the evaluation process were processed in order to identify changes in the listening experience due to the use of blindly separated sources. In this section we discuss the results obtained for each attribute considered.

6.5.1 Locatedness

The subjective data for the locatedness assessment are presented in Figure 6.6 (averaged over the different types of played material). We can observe these main differences between the reproduced scenes:

1. The locatedness of the real case is best. No other system achieves such a good grade, not even a WFS system using original sources not degraded by a demixing process, that is, WFS with sources recorded separately. This was also observed in [180].
2. The locatedness of the stereo scene is the worst for every reference direction compared to the locatedness obtained using blindly separated tracks, except for the ADRess algorithm. This means that resynthesized scenes using blindly separated tracks can improve the localization quality of a stereo scene, although this enhancement is mainly due to the WFS reproduction system.
3. The results are rather constant for the different separation algorithms and worse than using original sources, but there is a clear difference in the grades given for the reference directions. The sources in reference directions +15° and 33.5° are harder to perceive than the other two sources. However, this difference is also observed when original sources are used for rendering the scene, which means that these differences are not caused by the separation algorithms.
4. The small confidence intervals show that the subjects were rather consistent in their assessment of the locatedness.

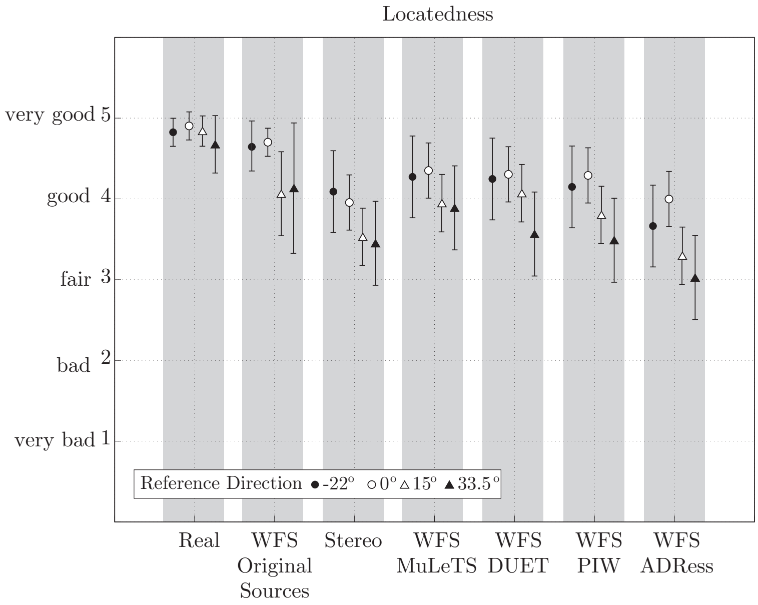


Figure 6.6. Subjective assessment of locatedness, showing mean and 95% confidence intervals of all test items of one system and reference direction.

6.5.2 Localization Accuracy

Figure 6.7 shows the results obtained for the perceived azimuth angles in the different examples. The following observations can be made:

1. In general terms, there are noticeable differences for the three examples considered, although the reference directions were the same for all of them. This is probably due to the difficulties involved in making an assessment when dynamic stimuli are considered and playing at the same time. This difficulty is also apparent when observing the long confidence intervals, which are especially long for sources with a bigger reference angle.

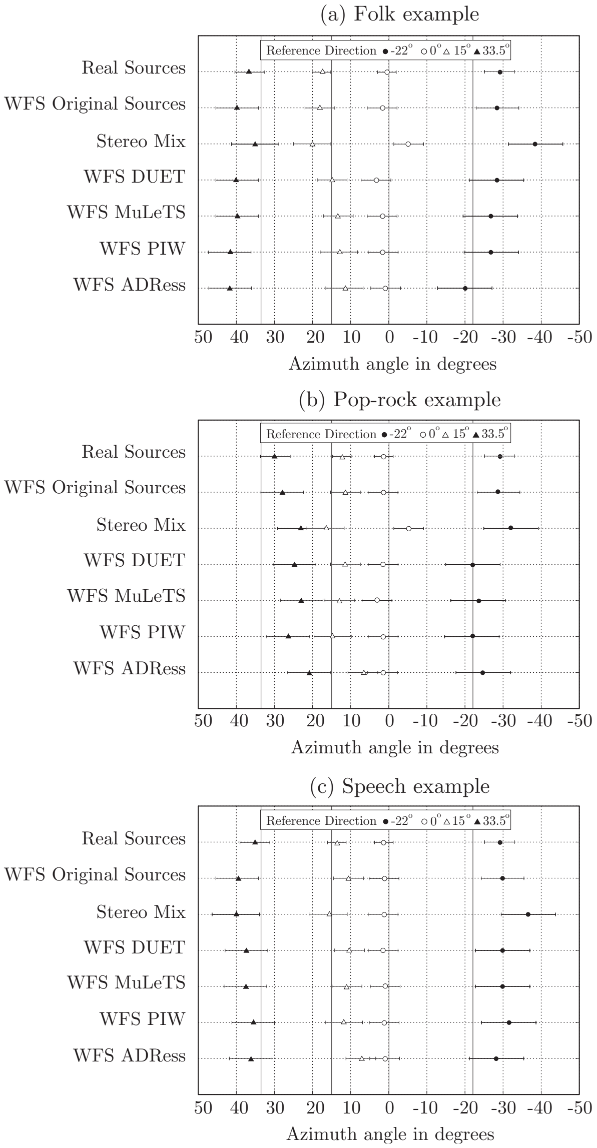
2. The angles were generally overestimated for the sources placed in reference directions -22° and 33.5°, except for the pop-rock example, were the bass (33.5°) is perceived to be placed more centrally. This deviation is produced by the source content itself and not by the separation algorithm used, because the scene composed of original sources is perceived in a similar way.

3. Sources located in the center are perceived more accurately and consistently.

4. The different separation algorithms show noticeable deviations from the real source case and WFS using original sources. These deviations may be the result of imperfections in the separation process. Nevertheless, the long confidence intervals obtained in all of the cases reveal the difficulty for the subject in accurately perceiving the direction of each source when other sources are playing simultaneously.

6.5.3 Source Widthness

The averaged widthness results are shown in Figure 6.8. The highest widthness is achieved in the real case and WFS using original sources. WFS is intended to reproduce each source as


(a) Folk example

(b) Pop-rock example

(c) Speech example

Figure 6.7. Localization accuracy: Mean azimuth angles. Bars show 95% confidence intervals.

a virtual point source, which should result in narrower sources than in the stereo case. It is interesting to observe that the source in reference direction 0 is perceived in all the cases as the narrowest source. Although no significant differences are observed between the different separation algorithms, there is a slight decrease in the grade given for these scenes and the scene composed of original sources. This means that the separation process affects the perceived width of the individual sources, making more difficult to experience a well-defined lateral extent of the source. In fact, the listeners expressed verbally an apparent widening of the sources when separated sources were used. In [174], Rumsey explains that a source perceived as having a small ISW will have a high level of locatedness, whereas more “diffuse” sources, those with a larger ISW, are more likely to appear poorly located. The results obtained in our experiments show also a relation between locatedness and widthness, confirming the correlation existent between width-related attributes and locatedness.

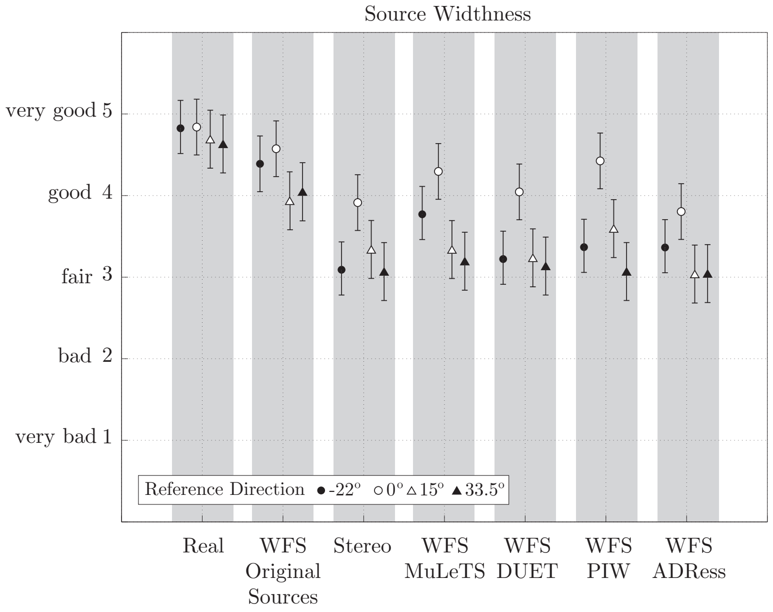


Figure 6.8. Subjective assessment of widthness, showing mean and 95% confidence intervals of all test items of one system and reference direction.

6.5.4 Ensemble Aperture

In Figure 6.9 the results obtained for the ensemble aperture are presented. The expected ensemble aperture is calculated as the angle existent between the most distant sources. In general, the stereo scene is perceived as the one which has a wider aperture. This should not be interpreted to mean that a characteristic of stereo is that it gives a wider impression than WFS. In fact, the WFS scenes were rendered with the sources placed at the angle where the stereo phantom sources were located, and therefore between the left and right loudspeakers of the stereo set up. This is why the real source system and WFS scenes have less ensemble aperture than stereo, but ideally they should have the same. However, as in the case of locatedness and source widthness, again there is a decrease in the perceived ensemble width of the scenes with separated sources in comparison to the same scene using original sound sources. The ensemble width is rather constant for all the separation algorithms, although it is a bit smaller in the case of the ADRess algorithm.

6.5.5 Sound/Timbral Quality

The mean results obtained using the modified MUSHRA procedure are shown in Figure 6.10. The real scene using natural sources was used as the reference scene. Non of the subjects had to be rejected in the post-screening process, as all of them were able to detect the hidden reference and anchors. In terms of overall sound quality, stereo was given the worst score in comparison to the rest of the WFS scenes. In this case, we can see that there is a noticeable difference between separation algorithms. The DUET algorithm has the best score, followed by the MuLeTS algorithm. These two algorithms are close to the score obtained by the WFS scene with original sound sources. This means that very good quality can be achieved using separation algorithms instead of original sound sources. The PIW and ADRess algorithms are also similar and near the good range, which is still better than the stereo case.

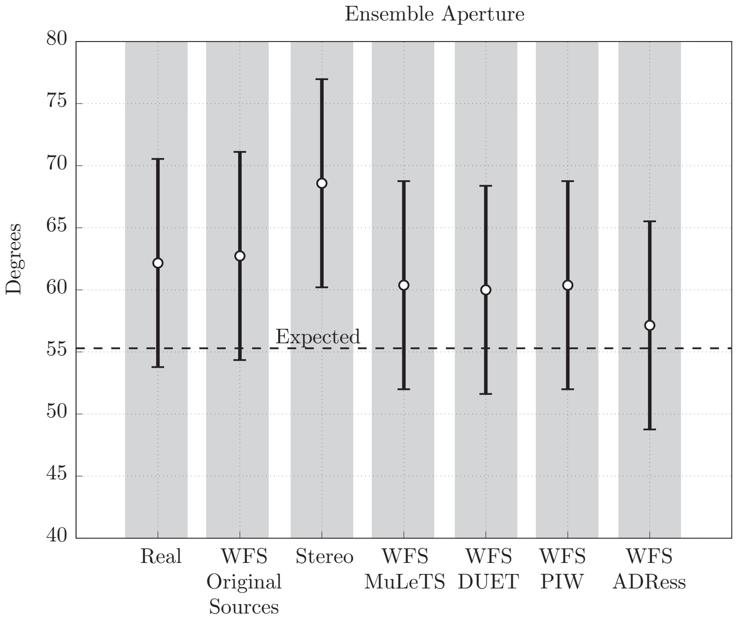


Figure 6.9. Subjective assessment of ensemble width, showing mean and 95% confidence intervals of all test items of one system and reference direction.

In Helmut Wittek's thesis [184], a set of experiments aimed at evaluating sound colour properties in different spatial audio systems was performed. The reproduction systems were artificially synthesized by means of a HRTF-based acoustic system including head-tracking that allowed to simulate arbitrary WFS and stereo setups. The spatial aliasing effect was shown to have an influence on the perceived coloration, while stereo reproduction seemed to be less affected due to other psicoacoustical factors involved in the perception of phantom sources.

It must be clarified that the results found in this thesis should not be interpreted as being in conflict with those evaluated in [184], since both experiments are very different in nature. Although a modified MUSHRA was also performed in Wittek's experiments, simple stimuli (pink bursts) were considered in order to establish a controlled experimental environment, since these were regarded to be most sensitive to changes in the sound colour. Moreover, the anchors were constructed as several sine-ripple spectral distortions with different ripple depth. On the other hand, the experiments presented by the author in this chapter are intended to evaluate timbral distortion in the reproduction of complex sound scenes, where many factors have an influence on the perceived colouration: sound scene material, source arrangement, inter-source effects, musical preferences of the subjects, etc. In addition, anchors were selected to be in concordance with the expected distortions due to source separation. Thus, the results here reported may be understood as an evaluation of the overall timbre impression with complex scene reproduction, which needs for sure further investigation.

6.5.6 Significance analysis with ANOVA

In order to strengthen the conclusions drawn from the previous results, a further analysis was carried out by means of the *Analysis of Variance* (ANOVA) test. So far, we have examined the means of a set of attributes using several systems under test, showing that some differences can be found among them.

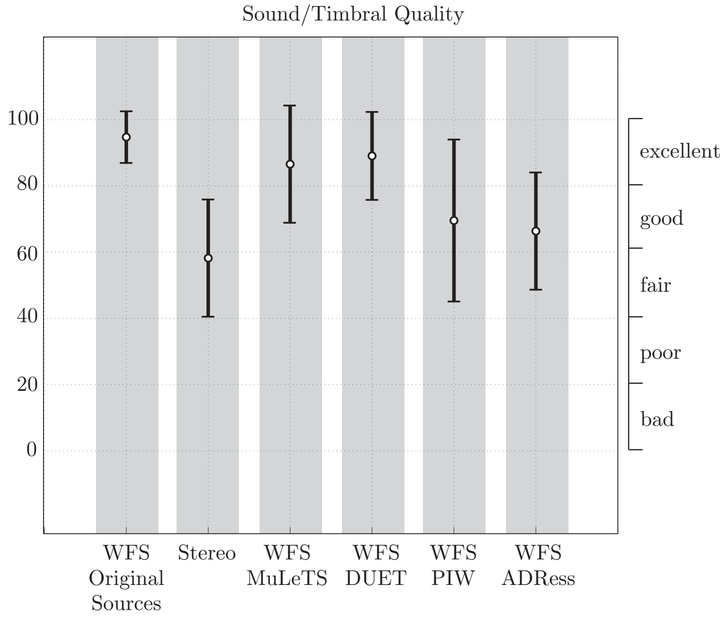


Figure 6.10. Subjective assessment of perceived sound/timbral quality, showing mean and 95% confidence intervals of all test items of one system and reference direction.

The ANOVA procedure gives a statistical test to reject the *null hypothesis*[1], concluding with a high degree of confidence that the aggregate mean differences among the groups of measures stem from something more than mere chance coincidence.

With the above aim, a one-way ANOVA for repeated measures is carried out. This assumes that there is one independent variable, i.e. a variable capable of influencing another (dependent) variable, and also that the subjects have been measured under all the tested systems.

Basically, the test consists in the computation of a ratio, F, which is defined for the general case as

$$F = \frac{MS_{\mathrm{bg}}}{MS_{\mathrm{wg}}}, \tag{6.3}$$

where MS_{bg}, is a variance estimate pertaining to the particular fact whose significance wants to be assessed (e.g., those differences explained by the different groups (systems)). Similarly, MS_{wg} is a variance estimate reflecting the amount of pure random variability present in the situation.

The factor probability P of not being significant (and so the differences observed in the means are mere coincidence) is given by the $F(df_{\mathrm{bg}}, df_{\mathrm{wg}})$ distribution, named after the English statistician Sir Ronald Fisher. Two degrees of freedom $df_{\mathrm{bg}} = k_g$ and $df_{\mathrm{wg}} = k_g(N_g - 1)$ define this distribution, where k_g and N_g are the number of groups and the number of subjects per group.

When repeated measures are taken on the subjects, the part of the variability within the total array of data that derives from individual differences among the subjects must be ignored, and MS_{wg} is replaced by MS_{error}. These estimates are computed from the sum of squared deviates between-groups SS_{bg}, within-groups SS_{wg} and from the subjects' deviates SS_{subjects}.

[1] In our context, the null hypothesis would be "changing the system does not produce any modification in a perceived spatial attribute".

The calculation of all these quantities is out of the scope of this section, but the interested reader can refer to [185] for a detailed description of different types of ANOVA.

ANOVA results

Table 6.2 shows the ANOVA parameters obtained for each spatial attribute averaged over the different audio programmes. The P probabilities of localization accuracy and ensemble aperture are considerably high, which means that it cannot be concluded that the changes in the means observed in these attributes are due to using different systems. Therefore, the tested source separation methods seem to have little influence regarding these attributes. On the other hand, the rest of attributes have very low values of P, thus, being significant at levels beyond 99% (locatedness and widthness) and 95% (timbral).

Recall once again that this term “significant” always has an *If/Then* logical structure embedded within it, and that the center-point of the structure is always the null hypothesis. For our case the structure is this: If the null hypothesis were true, i.e. if the differences among the means were occasioned by nothing more than random variability, then the likelihood of ending up with a larger F ratio would be less than 1% ($P < 0.01$) (or 5% in the case of colouration). As a consequence we can accordingly reject the null hypothesis, provisionally concluding that the observed effects have been produced by the use of blindly separated sources.

6.5.7 Joint Analysis

In Figure 6.11 the mean values obtained for all of the stimuli and listeners are represented jointly in order to show the overall performance of each system/separation method considered. Source locatedness and widthness are represented in a 5 grade scale and the sound quality conserves the MUSHRA scale (0 to 100). The localization accuracy shows the mean run standard deviation $<\bar{S}>$ in degrees, calculated as explained in Section 6.3.1. The ensemble aperture is represented as the deviation from the expected aperture value (in degrees as well). This figure is intended to summarize all the results in a single graph, although some considerations should be taken when evaluating the displayed values. There are some attributes that are clearly related to the perceived basic audio quality, such as localization accuracy or timbral fidelity. However, the relation of widthness and ensemble aperture to the perceived audio quality is not totally clear. As Rumsey [101] pointed out, concert hall experiments have shown that listeners prefer larger amounts of *Auditory Source Width* (ASW) (without differentiating between ensemble or individual source width). With regards to reproduced sound, it is unclear whether the same preference for larger ASW exists. In addition, the scales showed for each vertex are not the same, although they have been selected to be distant to the center in the better case. Therefore, although the area enclosed by each polygon can be loosely related to the perceived audio quality, these results should not be interpreted in a strict way.

It can be observed that WFS reproduction using original sources has the greatest values in terms of sound/timbral quality, source locatedness and source widthness. All of the separation algorithms produced scenes with smaller values with regard to these attributes, but it is interesting to notice how localization accuracy is slightly improved using blindly separated tracks. Nevertheless, all of the values remain close to the ones obtained using original sources and the spatial quality improves significantly in comparison to stereo reproduction. The DUET and MuLeTS algorithms achieved the best results, both being very similar. The PIW and ADRess

	Factor: Locatedness				
	SS	*df*	*MS*	*F*	*P*
Between groups	17.09	6	2.85	16.31	< 0.01
Within groups	15.84	98			
Error	14.68	84	0.17		
Subjects	1.17	14			
Total	32.94	104			

	Factor: Localization Accuracy				
	SS	*df*	*MS*	*F*	*P*
Between groups	138.2	6	23.04	1.35	0.24
Within groups	1709	98			
Error	1430	84	17.03		
Subjects	278,3	14			
Total	1847	104			

	Factor: Widthness				
	SS	*df*	*MS*	*F*	*P*
Between groups	24.59	6	4.10	50.69	< 0.01
Within groups	7.87	98			
Error	6.79	84	0.08		
Subjects	1.07	14			
Total	32.46	104			

	Factor: Ensemble Aperture				
	SS	*df*	*MS*	*F*	*P*
Between groups	10204	6	170.06	0.98	0.44
Within groups	18254	98			
Error	14615	84	170.99		
Subjects	3638	14			
Total	19274	104			

	Factor: Sound/Timbral Quality				
	SS	*df*	*MS*	*F*	*P*
Between groups	14802	5	2960.40	2.97	0.02
Within groups	76806	84			
Error	69742	70	996.31		
Subjects	7064	14			
Total	91608	89			

Table 6.2. ANOVA results.

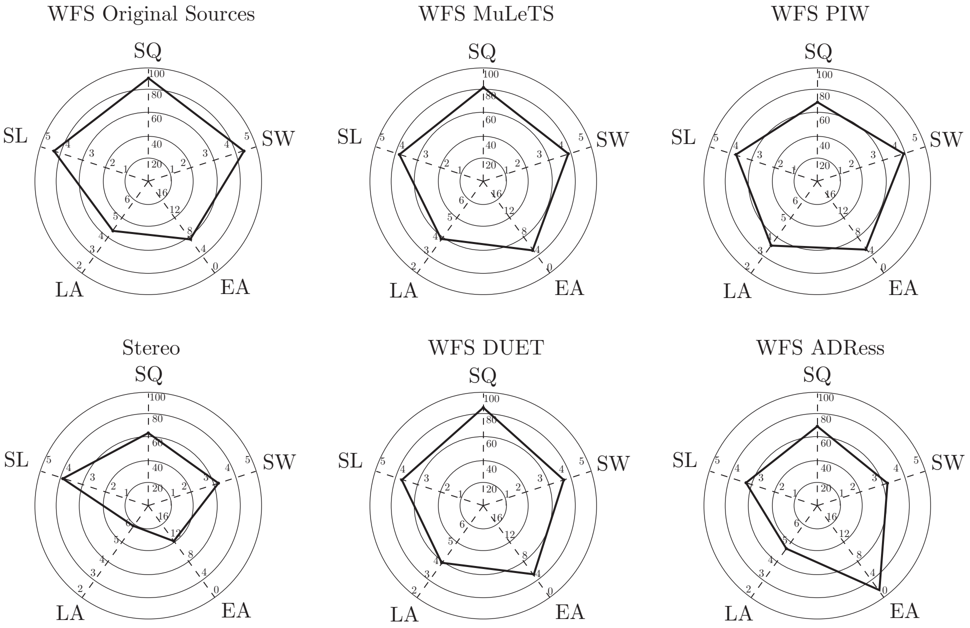


Figure 6.11. Overall performance of the systems. SQ = Sound Quality, SL = Source Locatedness, LA = Localization Accuracy, EA = Ensemble Aperture, SW = Source Widthness.

methods had acceptable results, but not as good as the other two separation algorithms.

6.5.8 Objective vs. Subjective Measures Discussion

Although it is very difficult to establish a complete correlation between every spatial attribute and the separation algorithm used, some interesting relationships have been observed in the experiments previously described:

- Localization accuracy seems to be related to the SIR. The more isolation achieved in the separation, the more localization accuracy perceived in the resynthesis. This is especially observable in the case of the PIW algorithm results. From Figures 6.3 and 6.11, it can be seen that the PIW algorithm is the one which has best SIR and localization accuracy.

- The modified MUSHRA results show a high correspondence between the sound quality of the scene and the SAR of the sources. The DUET algorithm obtained the best sound quality score and also the best SAR in the objective evaluation.

From the above preliminary observations we can state that, in general terms, localization attributes may be affected by the grade of interference rejection obtained in the separation process. Obviously, artifacts introduced by the separation algorithm affect the sound quality of the scene, but the separation is still very good in the context of multisource scene resynthesis.

6.6 Conclusions

In this chapter, SSS techniques have been proposed, employed and evaluated in the context of stereo to Wave Field Synthesis up-mixing for the first time. Although separation algorithms are far from giving high fidelity audio signals, the masking effects involved in the listening to spatially remixed scenes perceptually improve the quality of the WFS virtual sources. Listening tests were carried out in order to find out which spatial attributes were more affected in the reproduction of complex sound material, including music and speech. The outcome of the analysis carried out in this chapter showed that some differences in the spatial properties of the resynthesized scenes were detected, specially those related to the timbral quality and width of the sources. Nevertheless, the aim of constructing spatially enhanced scenes from stereo material has been achieved, confirming that SSS methods can be considered as a potential solution to WFS up-mixing.

7 Localization and Enhanced Reproduction

7 Localization and Enhanced Reproduction

Small microphone arrays provide many advantages in their application to mobile devices. Their enhanced acoustic properties can be exploited in many speech processing systems, such as hands-free devices, videoconferencing or hearing aids. In this chapter, several techniques based on small microphone arrays and time-frequency processing are presented: a multiple source localization method based on model fitting, a selective amplitude panning technique for enhanced stereo reproduction and a three-microphone array for automatic binaural synthesis. Although source separation is not considered for these applications, the proposed processing is also based on the sparsity and disjointness achieved by time-frequency representations.

7.1 Introduction

The first part of this thesis was centered on the separation of audio sources from stereo mixtures, which was later shown to be a solution for the reproduction of spatially enhanced scenes by means of Wave Field Synthesis (WFS). However, as seen in the previous chapters, extracting more sources than sensors from a stereo mixture is a difficult task. This underdetermined problem is usually faced by working with a sparse representation of the signals, which is commonly provided by a time frequency transformation, such as the STFT (Section 2.5.4). Additionally, the interaural differences between the sensors can be exploited to perform separation of both instantaneous (Section 3.2) and real (Section 3.3) mixtures.

The framework used in stereo SSS is not necessarily related to the extraction of sources from mixtures. In fact, using the frontend provided by the STFT and the interaural differences found between microphone pairs, it is possible to develop many interesting audio applications were the goal is not to obtain a separate signal for each source.

This chapter presents a set of developments based on the underdetermined SSS framework. Section 7.2 introduces a model-fitting approach for the localization of multiple sound sources.

To this end, the processing used in Section 3.3 is followed, obtaining a spatial representation of the signals in environments with moderate reverberation. Instead of separating the histogram into angular sections, a *Laplacian Mixture Model* (LMM) is fitted to the data by means of the *Expectation-Maximization* (EM) algorithm, resulting in accurate localization of the sources. Section 7.3 describes an enhanced stereo reproduction system based on the same described framework [186]. In this case, a selective amplitude panning is applied to each time-frequency point according to its DOA. This results in an impressive spatialization effect when listening to the synthesized time-domain signals through a conventional stereo reproduction system. Similarly, Section 7.4 extends the method for working with localized sources in the whole azimuth plane using binaural synthesis reproduction. In this case, three microphones are used, which enables to resolve the front/back ambiguity in DOA estimation.

7.2 Multiple Source Localization

Microphone arrays have been intensively studied in the last years due to their enhanced acoustic properties. One of the most active research lines in multichannel signal processing is acoustic source localization. In fact, estimating the direction of arrival of multiple sound sources in a real scenario is a very difficult task. Algorithms for acoustic source localization are often classified into direct approaches and indirect approaches [187]. Indirect approaches estimate the *time delay of arrival* (TDOA) between various microphone pairs and then, based on the array geometry, estimate the source positions by optimization techniques. On the other hand, direct approaches compute a cost function over a set of candidate locations and take the most likely source positions [188].

Cross-correlation-based methods in the time domain have been widely applied in DOA estimation [189]. However, the poor resolution achieved by time-domain methods led to frequency domain and subspace approaches [190]. Techniques based on the steered response power (SRP) are often chosen when more than two microphones are available [32]. In the case of multiple source localization using only two microphones or an acoustic mannequin, DOA estimation is usually performed via binaural localization cues [191]. When a source is not located directly in front of the array, sound arrives slightly earlier in time at the microphone that is physically closer to the source, and with somewhat greater energy. This fact produces the interaural time difference (ITD) and the interaural level difference (ILD) between the two sensors. DOA estimation methods based on binaural models, such as the Jeffress or equalization-cancelation (EC) models, have shown to successfully estimate locations of two sources in anechoic environments [192]. The DUET separation technique [193], which is also based on channel differences, can be used for estimating with high accuracy the TDOA of several sources in the time-frequency domain assuming that only one source is active in each point. This algorithm accumulates the power of each time-frequency bin in a histogram where the most likely source positions can be identified as strong peaks when the mixture is approximately anechoic. Unfortunately, peaky regions are spread out and overlap with one another when reverberation appears and its performance is severely degraded. In order to deal with this problem, Gaussian Mixture Models (GMM) for azimuthal histograms have been proposed to increase the robustness of DOA estimates against reverberation [194][195]. However, the sparse nature of speech signals in the time-frequency domain makes supergaussian distributions more appropriate to model these histograms, specially when only a set of reliable DOA estimates pre-selected from their short-time coherence is used.

In this section, we present a method to estimate the DOA of multiple simultaneous sources using two omnidirectional microphones under the framework of Section 3.3. Instead of applying multi-level thresholding for separating the sources, the distribution of DOA estimates is modeled as a mixture of Laplacian functions and the EM algorithm is used to find the parameters of this model. The real DOA of the sources are considered to be the centers of the Laplacian functions that best explain the observed distribution.

7.2.1 Laplacian Mixture Model

Speech signals can be considered as having a sparse distribution in the STFT domain. There is a number of models that can be used to represent sparsity. One common probabilistic model is the Laplacian density function, which is given by:

$$\mathcal{L}_p(\theta, \beta, \gamma) = \beta e^{-2\beta|\theta-\gamma|}, \tag{7.1}$$

where γ is the mean of the distribution of the random variable θ and $\beta > 0$ controls the "width" or approximate standard deviation. A LMM is defined as follows:

$$p(\theta) = \sum_{n=1}^{N} \alpha_n \mathcal{L}_p(\theta, \beta_n, \gamma_n) = \sum_{n=1}^{N} \alpha_n \beta_n e^{-2\beta_n|\theta-\gamma_n|}, \tag{7.2}$$

where α_n, γ_n and β_n represent the weight, mean and width of each Laplacian, respectively, and all weights should sum up to one, i.e., $\sum_{n=1}^{N} \alpha_n = 1$.

LMMs have also been successfully applied in the separation of instantaneous audio mixtures when there are more sources than sensors [196]. However, the purpose of this section is to use this model to find the DOA of the sources in a real scenario, taking profit of the coherence-based selection of Section 3.3.2, which enhances the peaky shape of the observed distribution under real reverberant conditions. The tendency of the data to cluster along DOA values is a consequence of speech sparsity in the time-frequency domain, and pre-selection allows to emphasize this tendency (see Figure 3.16). Similarly to common GMM approaches, the EM algorithm [197] can be employed to train a LMM over a training set (batch-EM) or even adapt the parameters of the LMM in real time (Online-LMM). Next, we present the updates for the batch-approach.

Batch-EM

We assume K_s training samples for θ_i obtained from the pre-selected DOA estimates (see Sections 3.3.1 and 3.3.2):

$$\theta_i = D(k, r)|_{(k,r)\in\mathcal{S}}, \tag{7.3}$$

where $\mathcal{S} = \{(k, r)\big|\Phi(k, r) > \Phi_T\}$ is the set of time-frequency bins above the defined coherence threshold (Φ_T), and $D(k, r)$ is computed from Eq.(3.56).

If Laplacian distributions are considered, the log-likelihood of these training samples takes the form:

$$I(\alpha_n, \beta_n, \gamma_n) = \sum_{i=1}^{K_s} \log \sum_{n=1}^{N} \alpha_n \mathcal{L}_p(\theta_i, \beta_n, \gamma_n). \tag{7.4}$$

The conditional expectation of the log-likelihood can be simplified as:

$$J(\alpha_n, \beta_n, \gamma_n) = \sum_{i=1}^{K_s} \sum_{n=1}^{N} (\log \alpha_n + \log \beta_n - 2\beta_n|\theta_i - \gamma_n|)p(n|\theta_i), \tag{7.5}$$

where $p(n|\theta_i)$ represents the probability of sample θ_i belonging to the nth Laplacian. The updates for $p(n|\theta_i)$ and α_n are given by [198]:

$$p(n,|\theta_i) = \frac{\alpha_n \mathcal{L}_p(\theta_i, \beta_n, \gamma_n)}{\sum_{n=1}^{N} \alpha_n \mathcal{L}_p(\theta_i, \beta_n, \gamma_n)}, \tag{7.6}$$

$$\alpha_n^+ \leftarrow \frac{1}{K_s} \sum_{i=1}^{K_s} p(n|\theta_i), \tag{7.7}$$

where $^+$ stands for the updated value in the current iteration. Setting $\partial J(\alpha_n, \beta_n, \gamma_n)/\partial\gamma_n = 0$ and $J(\alpha_n, \beta_n, \gamma_n)/\partial\beta_n = 0$, we can obtain the updates for γ_n and β_n, respectively:

$$\gamma_n^+ \leftarrow \frac{\sum_{i=1}^{K_s} \frac{\theta_i}{|\theta_i - \gamma_n|} p(n|\theta_i)}{\sum_{i=1}^{K_s} \frac{1}{|\theta_i - \gamma_n|} p(n|\theta_i)} \tag{7.8}$$

$$\beta_n^+ \leftarrow \frac{\sum_{i=1}^{K_s} p(n|\theta_i)}{2\sum_{i=1}^{K_s} |\theta_i - \gamma_n| p(n|\theta_i)}. \tag{7.9}$$

The above update rules are iterated until convergence ($p(n,|\theta_i)^+ - p(n,|\theta_i) < \epsilon$), where ϵ is small value. The pre-selection described in the previous section can emphasize the Laplacian structure of the observed distribution. This enhancement will increase the convergence speed of the EM algorithm, leading to more accurate DOA estimations.

Initialization

Some initialization issues should be considered. First of all, the number of sources N needs to be known *a priori*. In addition, the clusters found by the algorithm depend on the initial values of the Laplacian centers. However, as the solution space is known in advance ($\gamma_n \in [-1, 1]$) and we assume that the sources are relatively well separated in azimuth, it is convenient to initialize γ_n in equal intervals in the azimuth plane $(0°, 180°)$. This way, it is easier for the algorithm to cover most source positions in the mixture.

7.2.2 Experiments

In this section, we evaluate the proposed approach using real recordings obtained from the public data used in the *First Stereo Audio Source Separation Evaluation Campaign* [67]. In our experiments, we will discuss the accuracy of the described method in estimating the DOA of an increasing number of speakers, the localization accuracy for several types of mixtures (male,female and music) and the robustness of the method against reverberation.

LMM versus GMM

In order to test the validity of the proposed approach and for comparison purposes, several models are considered: GMM without pre-selection, GMM with pre-selection, LMM without pre-selection and LMM with pre-selection. The input signals are male speech mixtures sampled at 16 kHz and have a duration of 10 s. The room where the signals where captured had a reverberation time of 250 ms and the microphone spacing was set to 5 cm. TF analysis was carried out using half overlapping Hanning windows with duration 512 samples. The coherence function used was Mohan's coherence test (Section 3.3.2) and the number of time frames averaged for the computation of the covariance matrix $\hat{\mathbf{R}}(k, r)$ was $C_r = 10$.

N	Real DOA	GMM	Iter.	LMM	Iter.
1	(75°)	(75°)	31	(75°)	26
2	(75°,105°)	(80°,103°)	135	(78°,103°)	35
3	(75°,105°,45°)	(78°,102°,51°)	371	(77°,102°,47°)	200
4	(75°,105°,45°,140°)	(74°,103°, 37°, 110°)	500	(73°,103°,40°,144°)	121

Table 7.1. Estimated DOA for several speakers and iterations until convergence without Pre-selection.

N	Real DOA	GMM+Pre	Iter.	LMM+Pre	Iter
1	(75°)	(75°)	29	(75°)	25
2	(75°,105°)	(81°,103°)	100	(75°,104°)	24
3	(75°,105°,45°)	(80°, 103°, 50°)	210	(75°,105°,43°)	92
4	(75°,105°,45°,140°)	(79°,103°,45°,108°)	349	(74°,103°,44°,140°)	103

Table 7.2. Estimated DOA for several speakers and iterations until convergence with Pre-selection.

Table 7.1 and Table 7.2 show the results obtained for each configuration without and with using pre-selection, including the number of iterations completed until convergence. The results given for the GMM models are the best obtained for different initialization values, as GMM models showed to be much more dependent on initial values than LMM models. In general, it can be observed that LMM models perform better than GMM models, both in terms of DOA estimation accuracy and convergence. Note that the number of iterations is considerably reduced when pre-selection is applied (Table 7.2), especially when the number of sources is high. This is a consequence of the enhancement of the Laplacian structure of the distribution, which leads to a faster convergence of the EM algorithm. Figure 7.1(a) shows the fitted model and the observed distribution for the 4 sources case using a LMM with pre-selection. The corresponding GMM model is shown in Figure 7.1(b). The centers of the different curves are the estimated DOAs.

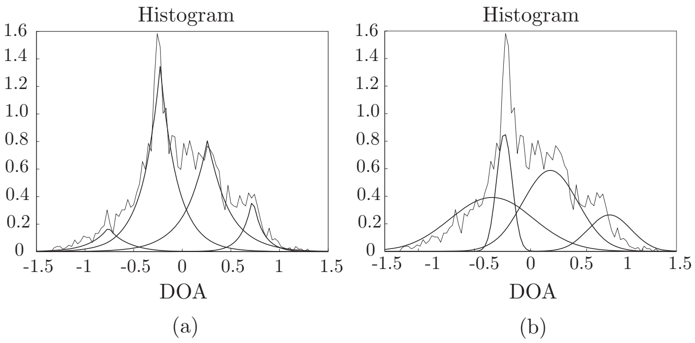


Figure 7.1. Fitted models for a mixture of 4 sources. (a) LMM with Pre-selection. (b) GMM with Pre-selection.

Localization Accuracy and Robustness

The accuracy of the proposed method is tested by using a setup which is similar to the one described above. In this case, several types of mixtures are tested: male speech mixtures, female speech mixtures, music mixtures with drums, and music mixtures without drums. Three sources are present in each mixture. With the aim of showing the robustness of the method against reverberation, three impulse responses corresponding to different degrees of reverberation have been simulated using the Matlab simulation tool *Roomsim* [199]. Table 7.3 shows the results obtained for each type of mixture and reverberation time (RT_{60}). As it can be seen in the table, the sound type has not a significant impact on accuracy. However, reverberation affects negatively the accuracy of the estimation. Nevertheless, although estimation is worse under very reverberant scenarios, the accuracy achieved by the method is still good for most practical applications.

Mixture	$\mathrm{RT}_{60} = 0$ ms	$\mathrm{RT}_{60} = 250$ ms	$\mathrm{RT}_{60} = 800$ ms
Female	0°	1.29°	4.24°
Male	0°	1.15°	4.04°
Music w/d	0°	1.63°	6.60°
Music n/d	0°	1.29°	5.25°

Table 7.3. Localization Accuracy (root mean square error).

7.3 Stereo Enhancement With Selective Amplitude Panning

Since the widespread introduction of stereophonic sound reproduction in the 1950s, multiple methods have been proposed to allow manipulation of the spatial characteristics of sound recordings [200]. *Stereo enhancement* refers to processing stereophonic music or sound in such a way as to add spaciousness to the stereo sound field. The purpose of stereo enhancement is to widen the stereo sound field, thereby immersing the listener in a cleaner, richer sound experience, significantly improving the quality, depth and feel of the sound played. From a practical point of view, stereo enhancement is intended to spread the stereo field into a 180 degree arc in front of the listener.

Stereo enhancement can be applied to listening situations with loudspeakers as well as headphones. Although the field of stereo enhancement has relatively few instances of scientific literature as compared to source positioning using binaural synthesis, there is a a huge amount of patents related to enhancement of stereophonic recordings.

Many enhancement schemes make use of the channel difference signal ($L^{ch} - R^{ch}$) and/or the sum signal ($L^{ch} + R^{ch}$) in order to emphasize the difference between the left and right signals [201][202][203]. For example, if the L^{ch} and R^{ch} signals contain a substantial common component, it is possible to express $L^{ch} = M^c + L^o$, and $R^{ch} = M^c + R^o$, where M^c is the common signal and L^o and R^o are the left-only and right-only components. In this situation $L^{ch} - R^{ch} = L^o - R^o$, so adding $L^{ch} - R^{ch}$ to L^{ch} gives $M^c + 2L^o - R^o$, which boosts the proportion of L^o in the composite left signal. Similarly, subtracting $L^{ch} - R^{ch}$ from R^{ch} performs the same operation on the right channel. Furthermore, the presence of the inverted components ($-R^o$ in the left output and $-L^o$ in the right output) also serves to give a broadened spatial impression to the resulting stereo sound field [200].

While these type of inventions have been widely used in the home entertainment industry to provide wider stereo scenes from conventional stereo mix-downs, there are no stereo enhancement methods aimed at improving the listening experience for recordings captured by small mobile devices, such as small digital cameras, PDAs, iPods or mobile phones.

This section describes a simple method for enhancing the stereo scene by post-processing recordings obtained with two small omnidirectional microphones. The technique next described is based again in the time-frequency processing of Section 3.3. A simple amplitude panning applied over each time-frequency point allows to obtain a highly spatialized sound scene in the reproduction of the processed signals.

7.3.1 Time-Frequency Selective Panning

Consider the STFT of the two channels from the stereo input, $X_1(k,r)$ and $X_2(k,r)$. The cosine of the DOA can be estimated by using Eq.(3.56). Firstly, values out of the range $[-1,1]$ are restricted to be within this range:

$$\bar{D}(k,r) = \begin{cases} \text{sign}(D(k,r)) & \text{if} \quad |D(k,r)| \geq 1 \\ D(k,r) & \text{otherwise} \end{cases} \quad \forall (k,r), \tag{7.10}$$

where sign() is the sign of a real number.

Then, a simple mapping to the range of the panoramic parameter ϕ_n is applied:

$$\bar{\phi}(k,r) = \frac{\bar{D}(k,r)+1}{2}, \quad \forall (k,r). \tag{7.11}$$

Note that, with this transformation, each time-frequency point is assigned a panning parameter depending on its DOA. Time-frequency points coming from 180° $(D(k,r) = -1)$ will be assigned a top left panning $(\bar{\phi} = 0)$, while points coming from 0° $(D(k,r) = 1)$, will be top right panned $(\bar{\phi} = 0)$. The selective panning is then applied on each time-frequency point accordingly:

$$X_1'(k,r) = \cos\left(\frac{\bar{\phi}(k,r)\pi}{2}\right) X_1(k,r), \quad \forall (k,r), \tag{7.12}$$

$$X_2'(k,r) = \sin\left(\frac{\bar{\phi}(k,r)\pi}{2}\right) X_2(k,r), \quad \forall (k,r). \tag{7.13}$$

The enhanced stereo signals, $x_1'(t)$ and $x_2'(t)$, are recovered by applying the inverse STFT operator. Note that the panning preserves the azimuth order of the estimated DOA, thus, the original stereo image is preserved.

Aperture Modification

The cosine of the DOA $D(k,r)$ is an estimation of the real azimuth position of the dominant sound source at time-frequency point (k,r). However, it is possible to change the degree of broadness of the stereo aperture by applying a transformation to the values of $\bar{D}(k,r)$. With this purpose, a transformation parameter P_D is introduced, which relates the original estimated $\bar{D}(k,r)$ with a modified one:

$$\bar{D}_m(k,r) = \text{sign}(\bar{D}(k,r)) \frac{|\bar{D}(k,r)| + \frac{1}{P_D}|\bar{D}(k,r)|}{2|\bar{D}| + \frac{1}{P_D} - 1}. \tag{7.14}$$

The P_D parameter is defined so that, if it is positive, the stereo aperture is increased, while if it is negative, the stereo aperture is decreased. A value $P_D = 0$ indicates that no transformation

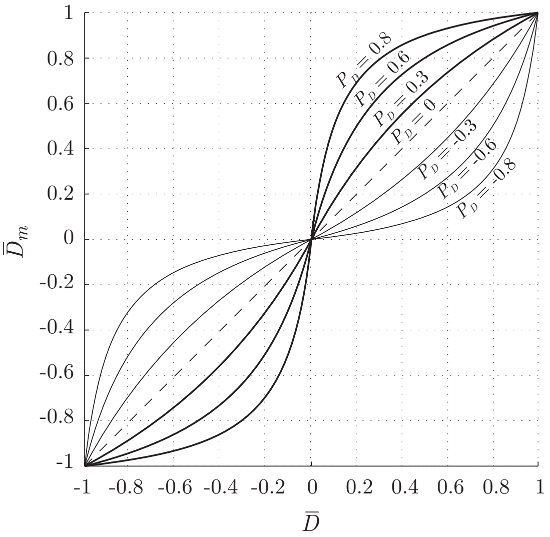


Figure 7.2. Modified values of $D(k,r)$ for different aperture parameters P_D.

is applied. Figure 7.2 shows how the original estimations $\bar{D}(k,r)$ are transformed into the modified ones $\bar{D}_m(k,r)$ using different values of P_D. The rest of the processing is identical, following with the computation of $\bar{\phi}(k,r)$ (Eq.(7.11)) and the selective panning of Eq.(7.12) and Eq.(7.13).

7.4 Enhanced Reproduction using Binaural Synthesis

The previous section proposed a stereo enhancement technique which can be used for reproduction over any two-channel system. This section is aimed at proposing a similar enhancer technique for binaural sound synthesis using headphones. The two-channel DOA estimation, which can not resolve the front/back ambiguity, is extended to full azimuth (0° to 360°) by using three omnidirectional microphones. In the synthesis stage, a selective HRTF filtering is used to add spaciousness to any of the mono signals recorded by the microphones.

7.4.1 System Geometry

The geometry used for deriving the relationships between the phase differences of each sub-array and the DOA of the source with respect to the center of the array is depicted in Figure 7.3. As we are covering the full azimuth range with this array, the estimation of the cosine of the DOA is not sufficient, thus, the sine of the DOA is needed to resolve the correct angular quadrant:

$$\cos(\theta^{360})(k,r) = \frac{\cos(\theta_{31})(k,r) - \cos(\theta_{21})(k,r)}{2\cos(\pi/3)}, \quad \forall(k,r) \tag{7.15}$$

$$\sin(\theta^{360})(k,r) = -\cos(\theta_{23}), \quad \forall(k,r) \tag{7.16}$$

where θ^{360} is the DOA angle of the source with respect to the array center and the between-sensor relationships are given by

$$\cos(\theta_{ij})(k,r) = \frac{c}{\omega_k d}\left(\angle X_i(k,r) - \angle X_j(k,r)\right), \quad \forall(k,r) \tag{7.17}$$

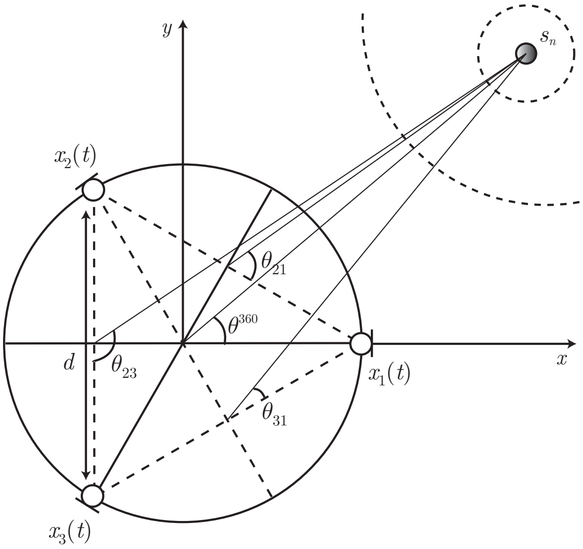


Figure 7.3. Three-microphone array geometry for DOA estimation.

The result of this processing are two time-frequency DOA maps that represent the estimates of $\cos(\theta^{360})$ and $\sin(\theta^{360})$ in each frequency and time-frame. A two-dimensional amplitude-weighted histogram of both values enables to have an intuitive representation of the energy distribution in the horizontal plane. Figure 7.4 shows the 2D weighted histograms (normalized) of a mixture of 4 speech sources with directions 15°, 75°, 170° and 260°. Two different room conditions are simulated in order to see the effect of reverberation in the energy distribution. Note how in the anechoic case (a), the sources appear as localized energy zones corresponding to their real DOAs. The diffuseness added by room reflections can be clearly appreciated in (b), where the energy, although clustered around the real DOA angles, has been highly spread.

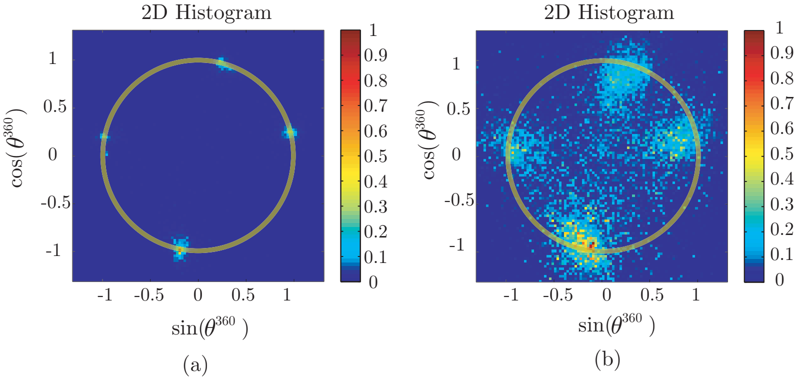


Figure 7.4. Three-microphone array geometry for DOA estimation. (a) $\rho = 0$ (anechoic). (b) $\rho = 0.5$.

It is important to notice that a perfectly estimated direction will show an agreement between their cosine and sine values, i.e., $\cos^2(\theta^{360})^2(k,r) + \sin^2(\theta^{360})(k,r) = 1$. Therefore, perfect estimations will lie on the unit circumference (as most of the points in the anechoic case). In contrast, points suffering spectral overlap between the sources or corrupted by reverberation, will be outside or inside the unit circumference. This property can be used for designing proper interaural coherence measures that quantify the diffuseness of the recorded sound.

7.4.2 Full-Azimuth DOA Map

From the two time-frequency maps of cosine and sine estimations, $\cos(\theta^{360})(k,r)$ and $\sin(\theta^{360})(k,r)$, a single full-azimuth DOA map is computed as follows:

$$D^{FA}(k,r) = \arctan^{360}\left(\cos(\theta^{360})(k,r), \sin(\theta^{360})(k,r)\right), \quad \forall(k,r) \tag{7.18}$$

where the operator $\arctan^{360}$ is the quadrant-resolving arctangent:

$$\arctan^{360}(\sin\theta, \cos\theta) \begin{cases} \arctan\theta & \text{if} \quad \sin\theta \geq 0, \cos\theta \geq 0 \\ \arctan\theta + \pi & \text{if} \quad \sin\theta \geq 0, \cos\theta < 0 \\ \arctan\theta + \pi & \text{if} \quad \sin\theta < 0, \cos\theta < 0 \\ \arctan\theta + 2\pi & \text{if} \quad \sin\theta < 0, \cos\theta \geq 0 \end{cases} \tag{7.19}$$

The angular distribution of the energy can be represented by an amplitude-weighted histogram (Section 3.2.2). The full-azimuth DOA map $\mathbf{D}^{FA}(k,r)$ for the example mixture is shown in Figure 7.5(a). The angular energy distribution can be observed in (b). Note that this histogram can be also used by the MuLeTS algorithm for performing separation in the full azimuth plane.

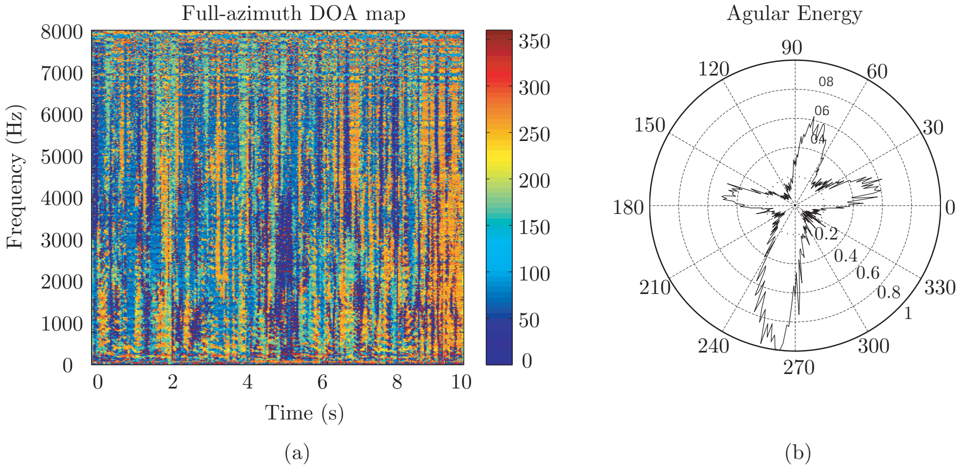


Figure 7.5. Full-azimuth DOA map and angular energy for $\rho = 0.5$. (a) Full-azimuth DOA map. (b) Polar representation of the angular energy

7.4.3 Binaural Synthesis

The binaural synthesis stage next described works similarly to the time-frequency selective amplitude panning of Section 7.3.1. In this case, instead of using the simple panning expressions for applying a selective gain to each time-frequency point, a complex HRTF look-up table is defined from any measured HRTF database. Therefore, a $2 \times K \times N_\theta$ matrix $\mathbf{HRTF}(k,\theta)$, is constructed with the left and right K-point HRTFs at the N_θ measured azimuth angles (with zero elevation).

Once the HRTF matrix is available, one of the input signals $X_m(k,r)$ is selectively filtered according to the value of the calculated full-azimuth DOA map:

$$X_1'(k,r) = X_m(k,r)\text{HRTF}(1,k,D^{FA}(k,r)), \quad \forall(k,r), \tag{7.20}$$

$$X_2'(k,r) = X_m(k,r)\mathrm{HRTF}(2,k,D^{FA}(k,r)), \quad \forall(k,r). \tag{7.21}$$

Note that in the above equations, it is assumed that a HRTF is available for any calculated DOA angle $D^{FA}(k,r)$. This is not usually the case, since HRTFs are usually measured at discrete azimuth angles. Two solutions to this problem are possible: to carry out an interpolation of the available HRTFs or filtering with the available HRTF closest to the estimated DOA angle. The author, in a real-time implementation of the described method, used the latter option with a 5° azimuth resolution HRTF database. Although a subjective evaluation of the method has not been conducted yet, no artifacts were apparently produced without interpolation.

7.5 Conclusion

In this chapter, three developments related to the framework of stereo underdetermined SSS have been presented: a method for localizing multiple sources in adverse environments, a stereo enhancement technique for improved spatial reproduction and a binaural synthesis method using a three-microphone array.

The source localization approach consists of three main stages, two of them common to the separation method of Section 3.3: time-frequency processing of the input signals for preliminary DOA estimates, coherence-based pre-selection of reliable points and LMM fitting of the observed distribution via the EM algorithm. The suitability of the model has been demonstrated by comparing the results obtained for a range of simultaneous speakers using public data from the the audio community. The experiments carried out confirm that the LMM outperforms the GMM in DOA estimation, both in terms of localization accuracy and convergence.

The patented stereo enhancement technique of Section 7.3 has also many similarities with the previous approach. In this case, the DOA map obtained from the analysis of phase differences in the time-frequency domain, is used to apply a selective amplitude panning to the input mixture channels. The stereo effect perceived by stereo recordings with two closely spaced omnidirectional microphones is very poor, thus, the described technique provides a solution for getting wider stereo images without using directional microphones. The applicability of the proposed technique is very high, since this simple processing adds a spatial quality improvement for mobile audio recording devices.

Finally, another application of time-frequency processing for small microphone arrays was provided in Section 7.4. The DOA estimation principle of the previous chapters was extended to the 3-microphone case, resolving the front/back ambiguity to cover the full azimuth plane. An improved spatial reproduction of the recording was obtained by applying a selective HRTF time-frequency filtering in accordance to the estimated directions. Although no formal subjective tests have been conducted to evaluate its performance, the results after implementing a real-time version of the technique confirm its viability to design immersive communication systems with this technology.

8 Conclusions

Conclusions 8

The overall aim of this research was to deepen into Sound Source Separation (SSS) methods and evaluate their potential application to advanced spatial audio reproduction systems. The motivation of this research came from the necessity of designing new up-mixing schemes suitable for the use of advanced reproduction formats based on sound field rendering, especially Wave Field Synthesis (WFS).

This chapter will summarize the findings of this research work, revisiting the research objectives given in the introductory chapter. First, Section 8.1 will review the contents of this study, outlining the main conclusions that were extracted from each chapter. The contributions of this thesis to the sound source separation and spatial sound reproduction fields will be listed in Section 8.2. Recommendations for future research will be also discussed. Additionally, the final section contains a list of published works occurred during the course of candidature for the degree.

8.1 Summary

As introduced in Chapter 1, the work presented throughout this thesis can be categorized within two areas: sound source separation and spatial sound.

The first part of this dissertation, related to the SSS field, presented the underdetermined source separation problem as a challenging one. The main difficulty resided in the fact that, even if the mixing process is perfectly known, the problem is not invertible, and recovering the sources without error is not an easy task. In the case of stereo mixtures, which is the most common format used to store audio material, sources usually outnumber mixture channels. Therefore, underdetermined SSS methods are needed to obtain an estimation of the original sound signals that originated the mix-down.

To this end, Chapter 3 presented two SSS techniques that addressed the underdetermined problem using a well-known segmentation technique from image processing: multi-level thresholding. In Section 3.2, the separation of instantaneous stereo mixtures was tackled by applying the proposed clustering to a weighted histogram of ILD values in the STFT domain and employing time-frequency masking. Similarly, Section 3.3 considered the separation of real sound mixtures using a two-microphone array. In this case, DOA estimates in the time-frequency domain were used as the clustering feature. In order to increase the robustness against reverberation, a coherence-based selection of reliable data points and an amplitude-weighted histogram were included. The evaluation results showed that the proposed processing can efficiently separate the sources in both cases, even when they have been mixed very close together.

The sources extracted by SSS algorithms are usually distorted versions of the original ones, and it is common to find residuals from other sources in the final separated tracks. These residuals affect the perceived quality of the signals and, as seen in Chapter 6, they have also an influence on the perceived spatial quality in spatial sound remixing. To deal with this problem, Chapter 4 proposed two post-processing techniques aimed at detecting and eliminating these residuals in the STFT domain. The first one, based on normalized energy masking, identified and suppressed these residuals using an Energy-Normalized Source to Interference Ratio. The second one, addressed the problem by detecting and reassigning small isolated clusters of active points in the separation masks. With this purpose, a likelihood criterion in the time-frequency neighborhood of these clusters was proposed. The results of the experiments carried out to evaluate the improvements achieved by these techniques showed that the isolation of the sources can be considerably increased, without introducing many artifacts after the processing.

The application of SSS to spatial audio, specifically to WFS, was covered in Chapter 5. SSS methods were proposed, employed, and evaluated in the context of stereo to WFS up-mixing. Although the sources extracted by stereo SSS algorithms are not high-fidelity audio signals, the masking effects that appear when listening to spatially remixed scenes perceptually improve the quality of the WFS sound scenes. Experimental work was carried out to find out which spatial attributes were more affected in the reproduction of complex sound material extracted with SSS algorithms. The results of the listening tests revealed that some differences in the spatial properties of the resynthesized scenes are produced, specially those related to the timbral quality and width of the sources.

Finally, Chapter 7 presented some developments related to the time-frequency processing framework used in this thesis. These developments addressed the localization of multiple speakers in teleconferencing applications and the enhanced reproduction of stereo recordings in mobile devices by using either selective amplitude panning or binaural sound synthesis. These applications showed the potential of applying time-frequency processing techniques for the development of enhanced sound reproduction systems.

8.2 Contribution to Knowledge

This section is devoted to how this research work has contributed to the field of SSS and WFS reproduction. These contributions are listed as follows:

- A novel approach for the separation of underdetermined instantaneous mixtures has been presented (Section 3.2). As many other separation algorithms, this approach performs separation by analyzing level differences in the STFT domain. However, while most approaches need to estimate the mixing matrix before proceeding to the estimation of the sources (Section 2.5.5), the proposed method separates directly the sources in a more unsupervised manner, without the need for estimating the mixing parameters. The key resides in the application of multi-level thresholding to a weighted histogram of observed level-difference values (Section 3.2.3). Since the histogram shape determines the final thresholds, a weighting function can be used to enhance the presence of the sources in it (Section 3.2.2). The sources are then recovered with time-frequency masking applying the masks defined by the computed thresholds. A set of experiments were used to demonstrate that this type of clustering is robust and efficient, and provides high separation performance with a potential to real-time applications (Section 3.2.7). The algorithm was a participant of the evaluation campaign SiSEC 2008, which obtained the lowest computing time in the reported results.

- An extension of the above method for the application of multi-level thresholding to convolutive mixtures has also been presented (Section 3.3). In order to deal with mixtures captured by two close microphones, the Direction-Of-Arrival (DOA) estimated in each time-frequency point is used as the clustering feature for computing the histogram. Although the basic generative model used is the anechoic model, some stages are included in the processing to increase the robustness of the method against reverberation: a coherence-based selection of reliable time-frequency points (Section 3.3.2) and the use of amplitude weighting in the formation of the histogram, which emphasizes the contribution of correct DOA estimates (Section 3.3.3). The experiments carried out with different mixtures, source proximities and degrees of reverberation show again the suitability of the method to perform separation of speech mixtures in adverse environments (Section 3.3.5). The results of SiSEC 2008 for underdetermined convolutive mixtures recorded by close microphones confirm also the validity of this approach.

- A post-processing technique used for eliminating residuals from separated sources was introduced in Section 4.2. Although many different approaches and algorithms can be found in the source separation literature, few works are aimed at improving the quality of the recovered sources after separation. A powerful and easy technique was introduced to eliminate interference time-frequency points by means of binary masks by means of an Energy-Normalized Signal to Interference Ratio (ENSIR) (Section 4.2.1). Moreover, the method can be used with any separation algorithm. A maximum SIR gain of 6.5 dB was obtained in average using separated sources from 20 different algorithms (Section 4.2.3), confirming the validity of the proposed approach to reduce the amount of interference found in the separated signals.

- Estimated binary masks have a great amount of isolated small clusters of non-zero points. In contrast, ideal binary masks show uniform clustered zones and, thus, musical noise and other kind of artifacts are less perceptible. With the purpose of reducing the effect of scattered non-zero clusters of binary time-frequency masks, another post-processing technique was proposed in Section 4.3. Connected points were first labeled with their cluster size and, afterwards, they were either eliminated or reassigned to other source mask. An average SIR improvement of 2.5 dB was found by suppressing small clusters composed of 10 or less points. If the points are reassigned to other masks, the improvements in SIR are lower and SAR decreases substantially, especially when the spanned likelihood is greater (Section 4.3.3). Therefore, new reassignment criteria should be explored in the future to find out how to obtain better improvements regarding artifact reduction.

- Source separation algorithms were used and evaluated in the context of WFS up-mixing for the first time. Several aspects of spatial sound reproduction were considered in the listening tests: locatedness, localization accuracy, source widthness, ensemble aperture and timbral quality (Section 6.4). The evaluation of resynthesized scenes with separated sources is quite challenging: simple stimuli can not be used and evaluating complex sound material where different sound events are played simultaneously is a difficult task. Expert listeners from the audio research field, sound engineers and musicians, were trained in the above attributes, leading to some meaningful results: while the sound quality of the resynthesized scenes was (in mean) evaluated as excellent (DUET and MuLeTS) and good (PIW and ADRess), some effects in the spatial descriptors were detected. Listeners found more difficult to perceive a defined width in the separated sources than in the original sources, which is in concordance with the lower assessed locatedness (Section 6.5).

- A multiple source localization method based on the processing of Section 3.3 and a Laplacian Mixture Model (LMM). The suitability of the model has been demonstrated by comparing the results obtained for a range of simultaneous speakers using public data from the audio community. The results confirmed that the LMM outperforms the Gaussian Mixture Model (GMM) in DOA estimation, both in terms of localization accuracy and convergence (Section 7.2.2).

- An application derived from the time-frequency processing framework used in underdetermined SSS was presented in Section 7.3. This application is intended to spatialize the listening experience in recordings using a pair of closely spaced omnidirectional microphones. The enhanced stereo image provided by the described method produces a substantial spatial quality improvement in the reproduction of stereo recordings for small mobile devices, such as digital cameras, PDAs or mobile phones. This technique relies on a selective time-frequency processing amplitude panning, so the effect can be perceived using both conventional loudspeaker systems and headphones.

- Two-microphone arrays are unable to process spatial information in 360° due to the front/back ambiguity that appears for sources producing the same time difference. A three-microphone array has been proposed to extend the processing of Section 3.3.1 with the aim of covering the full azimuth plane. Moreover, another application related to the spatialization of the recordings by means of headphone reproduction based on binaural synthesis has been proposed. Similarly to the above selective panning enhancement, the time-frequency points of one of the input channels are filtered with HRTFs according to its azimuth direction. The applicability of the technique and its viability for practical systems has been demonstrated by means of a real-time prototype.

8.3 Further Work

Following the investigations described in this thesis, the main lines of research that remain open are listed below:

- This thesis was focused on the separation of stereo recordings using STFT representations. It has been shown that the use of frequency-warped representations improves the quality of the separated sources both with music and speech material, especially in terms of artifacts. Hence, it would be interesting to evaluate the advantages introduced by these types of representations when applied to spatial sound resynthesis.

- The use of separated sources for scene resynthesis was mostly motivated by the object-based conception of sound field rendering techniques, such as WFS. That is why the evaluation has been conducted using this sound reproduction technique. However, other systems like VBAP, Ambisonics, Surround Systems and Stereo can also benefit from the advantages of SSS. Research focused on evaluating the perceived quality using other spatial sound reproduction systems would provide more insight about the spatial properties of separated sources with different systems.

- The MuLeTS algorithm assumed the number of classes (sources) in the mixture to be known beforehand. Some methods can be found in the literature to automatically select an optimum number of thresholds. Further research would consider to include these methods in the algorithm and evaluate their performance.

- Although Section 4.3 showed that the treatment of small isolated clusters of points in binary masks provides some advantages, the followed reassignment criteria was not very satisfactory. It would be interesting to conduct research with the aim of finding more effective criteria in the reassignment, such as considering different types of neighborhoods or pitch-based criteria.

- A subjective evaluation of the developments for enhanced sound reproduction (Section 7.3 and Section 7.4) would be useful to know how different listeners perceive the sound images provided by the enhanced systems. Moreover, it would interesting to extend the method for more microphones in order to estimate directions in the 3D space.

8.4 List of Publications

The following presents a list of published work produced during the course of candidature for the degree. The author of this thesis is the primary author of all the publications and none of these have previously formed a part of another thesis. These publications are listed below:

Refereed ISI Journals

- **M. Cobos** and J. J. Lopez, "Resynthesis of Sound Scenes on Wave-Field Synthesis from Stereo Mixtures using Sound Source Separation Algorithms," *Journal of the Audio Engineering Society*, Vol. 57, Issue 3, pp.91-110, March 2009.

- **M. Cobos** and J. J. Lopez, "Improving Isolation of Blindly Separated Sources using Time-Frequency Masking," *IEEE Signal Processing Letters*, Vol. 15, pp.617-620, 2008.

- **M. Cobos** and J. J. Lopez, "Stereo Audio Source Separation based on Time-Frequency Masking and Multilevel Thresholding," *Digital Signal Processing*, Vol. 18, no.6, pp.960-976, November 2008.

- **M. Cobos** and J. J. Lopez, "Two-Microphone Separation of Multiple Speakers Based on Interclass Variance Maximization" *Journal of the Acoustical Society of America*, submitted, 2009.

- **M. Cobos** and J. J. Lopez, "Multiple Speaker Localization Based On A Laplacian Mixture Model" *Digital Signal Processing*, submitted, 2009.

Peer-reviewed Magazines

- **M. Cobos** and J. J. Lopez, "Present and Future of Audio Signal Processing," in *IEEE Potentials Magazine*, to appear in 2009.

Peer-reviewed non-ISI Journals

- **M. Cobos** and J. J. Lopez, "Multi-Speaker Localization, Separation and Resynthesis for Next Generation Videoconferencing," *iTeam Research Journal*, to appear in 2009.

- **M. Cobos** and J. J. Lopez, "Técnicas de Separación de Fuentes Sonoras Aplicadas a la Resíntesis de Escenas Acústicas 3D," *Iteckne: Ciencia e Innovación en Ingeniería*, vol.7, pp.18-24, July 2007.

Book Chapters

- **M. Cobos** and J. J. Lopez, "Musical Source Separation: Algorithms and Applications," in "Music: Automatic Composition, Interpretation and Effects", *Nova Science Publishers*, New York, 2009 (in press)

.

Papers in International Conferences

- **M. Cobos** and J. J. Lopez, "Small Microphone Array for Multiple Speaker Detection and Separation," in *Proceedings of the AES 126th Convention*, Munich, Germany, May 2009.
- **M. Cobos**, J. J. Lopez and Jan O. Hinz, "A Source Reassignment Technique for Time-Frequency Masking Audio Separation," in *Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control (IWAENC 2008)*, Seattle WA., USA, September 2008.
- **M. Cobos** and J. J. Lopez, "Singing Voice Separation from Stereo Recordings Combining Panning Information and Pitch Tracking," in *Proceedings of the AES 124th Convention*, Amsterdam, Netherlands, May 2008.
- **M. Cobos**, J. J. Lopez, A. Gonzalez and J. Escolano, "Stereo to Wave-Field Synthesis Music Up-Mixing: An Objective and Subjective Evaluation," in *Proceedings of the IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP 2008)*, St. Julians, Malta, March 2008.
- **M. Cobos** and J. J. Lopez, "On the Application of Sound Source Separation to Wave-Field Synthesis," in *Proceedings of the AES 122nd Convention*, Vienna, Austria, May 2007.

Papers in National Conferences

- **M. Cobos**, J. J. Lopez, "Técnicas de Separación de Audio Estéreo Aplicadas a la Resíntesis de Escenas Sonoras," in *Proceedings of the XXIII Simposium Nacional de la Unión Científica Internacional de Radio*, Madrid, September 2008.
- J. J. Lopez and **M. Cobos**, "Nuevas Tendencias en Investigación de Audio," in *Proceedings of the XXIII Simposium Nacional de la Unión Científica Internacional de Radio*, Madrid, September 2008.
- **M. Cobos**, J. J. Lopez, A. Gonzalez and A. Cebrian, "Comparación de Técnicas de Separación de Voz en Sistemas de Videoconferencia Avanzados," in *Proceedings of the XVII Jornadas Telecom I+D*, Valencia, October 2007.

Patents

- Method and Apparatus for Stereo Enhancement in Stereo Audio Recordings with Application to Mobile Devices.
  - *Number*: P200802379.
  - *Holder*: Universidad Politécnica de Valencia.
  - *Date*: 31/07/08.

Others

- **The Beatles Live**!: Invited Demonstration in the event "*20 Years of Wave Field Synthesis*", held in the *124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, 2008.

- **SiSEC 2008**!: Participant of the "*Signal Separation Evaluation Campaign 2008*", with on-line published results available at http://sisec.wiki.irisa.fr.

Bibliography

[1] J. Blauert, *Spatial Hearing.* MIT Press, Cambridge, 1997.

[2] W. B. Snow, "Basic principles of stereophonic sound," *Journal of the Society of Motion Picture and Television Engineers*, vol. 61, no. 11, pp. 567–589, 1953.

[3] A. J. Berkhout, "A holographic approach to acoustic control," *Journal of the Audio Engineering Society*, vol. 36, pp. 977–995, 1988.

[4] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal of the Acoustical Society of America*, vol. 93, pp. 2764–2778, 1992.

[5] M. M. Boone, E. N. G. Verheijen, and P. F. van Tol, "Spatial sound field reproduction by wave field synthesis," *Journal of the Audio Engineering Society*, vol. 43, no. 12, pp. 1003–1012, 1995.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, 2001.

[7] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley, 2006.

[8] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[9] A. W. Bronkhorst, "The cocktail party phenomenom: A review of research on speech intellibility in multiple-talker conditions," *Acustica*, vol. 86, pp. 117–128, 2000.

[10] D. P. W. Ellis, "Hard problems in computational auditory scene analysis," Posted to the AUDITORY email list, August 1995.

[11] A. S. Bregman, *Auditory Scene Analysis.* MIT Press: Cambridge, MA, 1990.

[12] E. D. Scheirer, "Music-listening systems," Ph.D. dissertation, Massachussetts Institute of Technology, 2000.

[13] J. Hérault, C. Jutten, and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé," in *Proceedings of Colloque GRETSI*, Nice, France, 1985, pp. 1017–1022.

[14] J. F. Cardoso, "Blind signal separation: Statistical principles," in *Proccedings of the IEEE*, vol. 86, no. 10. IEEE Computer Society Press, October 1998, pp. 2009–2025.

[15] P. O'Grady, B. Pearlmutter, and S. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology (IJIST)*, vol. 15, no. 1, pp. 18–33, 2005.

[16] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[17] J. Woodruff, "Remixing stereo music with score-informed source separation," in *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria, British Columbia, October 2006.

[18] F. Canadas-Quesada, J. J. Carabias-Orti, R. Mata-Campos, N. Ruiz-Reyes, and P. Vera-Candeas, "Polyphonic piano transcription based on spectral separation," in *Proceedings of the 124th Convention of the Audio Engineering Society*, 2008.

[19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788791, 1999.

[20] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1135–1150, September 2004.

[21] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1475–1487, 2007.

[22] M. Wu, D. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 229–241, 2002.

[23] J. J. Burred, "From sparse models to timbre learning: New methods for musical source separation," Ph.D. dissertation, Technical University of Berlin, 2008.

[24] E. Vincent, M. Jafari, S. Abdallah, M. Plumbey, and M. Davies, "Blind audio source separation," Queen Mary, University of London, Tech Report C4DM-TR-05-01, November 2005.

[25] K. N. Stevens, *Acoustic Phonetics.* MIT Press, 2000.

[26] D. E. Hall, *Musical Acoustics.* Brooks Cole, 2001.

[27] H. Kuttruff, *Room acoustics.* Taylor & Francis, October 2000.

[28] M. Vinyes, J. Bonada, and A. Loscos, "Demixing commercial music productions via human-assisted time-frequency masking," in *Audio Engineering Society 120th Convention*, Paris, France, May 2006.

[29] M. Cobos and J. J. Lopez, "Singing voice separation from stereo recordings combining panning information and pitch tracking," in *Proceedings of the 124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, May 2008.

[30] B. Bartlett, *Stereo Microphone Techniques.* Focal Press, 1991.

[31] I. T. Jolliffe, *Principal Component Analysis.* New York: Springer, 2002.

[32] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications.* Springer-Verlag, 2001, ch. Robust Localization in Reverberant Rooms, pp. 157–180.

[33] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[34] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition," University of New Mexico, Tech. Rep., 1999.

[35] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. dissertation, Tampere University of Technology, 2006.

[36] S. A. Abdallah, "Towards music perception by redundancy reduction and unsupervised learning in probabilistic models," Ph.D. dissertation, Department of Electronic Engineering, King's College London, 2002.

[37] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, p. 23532362, 2001.

[38] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[39] F. Theis and E. Lang, "Formalization of the two-step approach to overcomplete BSS," in *Proceedings of Signal and Image Procesing (SIP)*, Kauai, USA, 2002.

[40] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, 2006.

[41] M. R. DeWeese, M. Wehr, and A. M. Zador, "Binary spiking in auditory cortex," *Journal of Neuroscience*, vol. 23, pp. 7940–7949, 2003.

[42] S. Chen and D. Donoho, "Basis pursuit," in *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, 1994.

[43] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[44] M. M. Goodwin, "Adaptive signal models: Theory, algorithms and audio applications," Ph.D. dissertation, University of California, 1997.

[45] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications.* Prentice Hall, 2006.

[46] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals.* New York: IEEE Press, 2000.

[47] A. H. Nuttal, "Some windows with very good sidelobe behavior," *IEEE Transactions on Acoustics and Speech*, vol. 29, pp. 84–91, 1981.

[48] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, pp. 51–83, 1978.

[49] J. Karvanen and A. Cichocki, "Measuring sparseness of noisy signals," in *Proceedings of the International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, Nara, Japan, April 2003.

[50] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[51] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[52] P. Vera-Candeas, N. Ruiz-Reyes, M. Rosa-Zurera, F. Lopez-Ferreras, and J. Curpian-Alonso, "New matching pursuit based sinusoidal modelling method for audio coding," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 151, pp. 21–28, 2004.

[53] L. Vielva, D. Erdogmus, and J. C. Príncipe, "Underdetermined blind source separation using a probabilistic source sparsity model," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, USA, 2001.

[54] H. Asari, B. A. Pearlmuter, and A. M. Zador, "Sparse representations for the cocktail party problem," *Journal of Neuroscience*, vol. 26, pp. 7477–7490, 2006.

[55] D. Wang, "Time frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[56] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing (ICICS 2005)*, Bangkok, Thailand, December 2005.

[57] S. Rickard and O. Yilmaz, "On the w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, May 2002, pp. 529–532.

[58] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 2005.

[59] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, September 2006.

[60] M. Küne, R. Togneri, and S. Nordholm, *Speech Recognition, Technologies and Applications.* Vienna, Austria: I-Tech, 2008, ch. Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition, pp. 61–80.

[61] S. Rickard, *Blind Speech Separation.* Springer, 2007, ch. 8: The DUET Algorithm.

[62] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2003.

[63] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discrimination and resynthesis," in *7th Conference on Digital Audio Effects (DAFX 04)*, 2004.

[64] T. Virtanen, "Sound source separation in monaural music signals," Ph.D. dissertation, Tampere University of Technology, Finland, November 2006.

[65] J. J. Burred and T. Sikora, "Monaural source separation from musical mixtures based on time-frequency timbre models," in *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, September 2007.

[66] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[67] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, London, UK, September 2007.

[68] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *International Conference on Independent Component Analysis and Signal Separation (ICA 2009)*, Paraty, Brazil, March 2009.

[69] L. G. Shapiro and G. C. Stockman, *Computer Vision.* Prentice Hall, 2001.

[70] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '05)*, Eindhoven, 2005, p. 117120.

[71] B. A. Zibulevsky, P. Pearlmutter, P. Bofill, and P. Kisilev, *Independent Component Anal: Principles and Practice.* Cambridge, 2001, ch. Blind Source Separation by Sparse Decomposition.

[72] S. Rickard, T. Melia, and C. Fearon, "DESPRIT - histogram based blind source separation of more sources than sensors using subspace methods," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, October 2005.

[73] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using DESPRIT," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 90–90, 2007.

[74] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear anechoic mixture," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 745–748.

[75] T. Abak, U. Baris, and B. Sankur, "The performance of thresholding algorithms for optical character recognition," in *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'97)*, 1997, pp. 697–700.

[76] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1191–1201, 1995.

[77] B. Bhanu, "Automatic target recognition: state of the art survey," *IEEE Transactions on Aerospace Electronic Systems*, vol. AES-22, pp. 364–379, 1986.

[78] M. Sezgin and B. Sankur, "Comparison of thresholding methods for non-destructive testing applications," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2001, pp. 764–767.

[79] M. E. Sieracki, S. E. Reichenbach, and K. L. Webb, "Evaluation of automated threshold selection methods for accurately sizing microscopic fluorescent cells by image analysis," *Applied and Environmental Microbiology*, vol. 55, no. 11, p. 27622772, 1989.

[80] J. Fan, J. Yu, G. Fujita, T. Onoye, L. Wu, and I. Shirakawa, "Spatiotemporal segmentation for compact video representation," *ignal processing. Image communication*, vol. 16, no. 6, pp. 553–566, 2001.

[81] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Transactions on System Man Cybernetics*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[82] P. K. Sahoo, S. Soltani, and A. K. C. Wong, "A survey of thresholding techniques," *Computer Vision, Graphics, and Image Processing*, vol. 41, pp. 233–260, 1988.

[83] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–165, 2004.

[84] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[85] P. Liao, T. Chen, and P. Chung, "A fast algorithm for multilevel thresholding," *Journal of Information Science and Engineering*, vol. 17, pp. 713–717, 2001.

[86] B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, September 2008.

[87] P. Y. Yin and L. H. Chen, "A fast iterative scheme for multilevel thresholding methods," *Signal Processing*, vol. 60, pp. 305–313, 1997.

[88] M. Cobos and J. J. Lopez, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," *Digital Signal Processing*, vol. 18, no. 6, pp. 960–976, 2008.

[89] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[90] C. Faller, “Parametric coding of spatial audio,” Ph.D. dissertation, École Polytechynique Fédérale de Lausanne, 2004.

[91] C. Faller and J. Merimaa, “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence,” *Journal of the Acoustical Society of America*, vol. 116, pp. 3075–3089, 2004.

[92] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, “Localization of multiple acoustic sources with small arrays using a coherence test,” *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[93] C. Avendano and J. M. Jot, “Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’02)*, Orlando, Florida, USA, May 2002.

[94] M. Cobos and J. J. Lopez, “Improving isolation of blindly separated sources using time-frequency masking,” *IEEE Signal Processing Letters*, vol. 15, pp. 617–620, 2008.

[95] ——, “A source reassignment technique for time-frequency masking audio separation,” in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, USA, September 2008.

[96] P. Bofill and E. Monte, “Underdetermined convoluted source reconstruction using lp and socp, and a neural approximator of the optimizer,” in *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2006)*, J. Rosca, D. Erdogmus, J. C. Principe, and S. Haykin, Eds., vol. 3889. Springer, Heidelberg, 2006, pp. 569–576.

[97] E. Vincent, “Complex nonconvex l_p norm minimization for underdetermined source separation,” in *Proceedings of the 7th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2007)*, M. E. Davies, J. C. J., A. S. Abdallah, and M. D. Plumbey, Eds. London, UK: Springer, September 2007, pp. 430–437.

[98] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[99] Z. Goh, K. C. Tan, and B. T. G. Tan, “Postprocessing method for supressing musical noise generated by spectral subtraction,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 287–292, 1998.

[100] R. M. Haralick and G. S. Linda, *Computer and Robot Vision.* Addison-Wesley, 1992, vol. I, pp. 28–48.

[101] F. Rumsey, *Spatial Audio.* Focal Press, 2001.

[102] J. Escolano, “Contributions to discrete-time domain methods in room acoustic simulations,” Ph.D. dissertation, Universidad Politécnica de Valencia, 2008.

[103] C. I. Cheng and G. H. Wakefield, “Introduction to head-related transfer functions (HRTF’s): Representations of HRTF’s in time, frequency, and space,” in *Proceedings of the 107th Convention of the Audio Engineering Society*, New York, USA, 1999.

[104] J. J. Lopez, “Reproduction of 3d-sound using local control techniques,” Ph.D. dissertation, Universidad Politécnica de Valencia, 1999.

[105] R. Rabenstein and S. Spors, *Springer Handbook of Speech Processing.* Springer, 2008, ch. Sound Field Reproduction, pp. 1095–1113.

[106] V. Pulkki, “Spatial sound generation and perception by amplitude panning techniques,” Helsinki University of Technology, Helsinki, Finland, Tech. Rep., 2001.

[107] M. A. Gerzon, “Ambisonics in multichannel broadcasting and video,” *Journal of the Acoustical Society of America*, vol. 33, pp. 859–871, 1985.

[108] L. L. Beranek, *Acoustics.* McGraw-Hill, 1954.

[109] H. D. Associates, *The Measure of Man and Woman.* Whitney Library of Design, 1993.

[110] J. W. S. Rayleigh, *The Theory of Sound (2nd Edition).* New York: Dover Publications, 1945.

[111] V. R. Algazi, C. Avendano, and R. O. Duda, “Elevation localization and head-related transfer function analysis analysis at low frequencies,” *Journal of the Acoustical Society of America*, vol. 109, pp. 1110–1122, 2001.

[112] R. Gilkey and T. R. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments.* Lawrence Erlbaum, 1997, ch. Spectral shape cues for sound localization, pp. 77–97.

[113] H. Wallach, “The role of head movements and vestibular and visual cues in sound localization,” *Journal of the Experimental Psycology*, vol. 40, pp. 339–368, 1940.

[114] D. S. Brungart, “Near-field auditory localization,” Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[115] W. R. Thurlow and P. S. Runge, “Effects of induced head movements on localization of direct sound,” *Journal of the Acoustical Society America*, vol. 42, pp. 480–487, 1967.

[116] H. Moller, M. F. Sorensen, C. B. Jensen, and D. Hammershoi, “Binaural technique: do we need individual recordings?” *Journal of the Audio Engineering Society*, vol. 44, pp. 451–468, 1996.

[117] A. Kulkarni and H. S. Colburn, “Role of spectral detail in sound source localization,” *Letters to Nature*, vol. 396, pp. 747–749, 1998.

[118] A. D. Blumlein, “Improvements in and relating to sound transmission, sound recording and sound reproduction systems,” Brit. Patent 394,325.

[119] G. Theile, “On the naturalness of two-channel stereo sound,” *Journal of the Audio Engineering Society*, vol. 39, pp. 761–767, October 1991.

[120] J. M. Eargle, Ed., *AES Anthology: Stereophonic Techniques.* New York: Publications of the Audio Engineering Society, 1986.

[121] A. D. Blumlein, "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems," British Patent 394,325, 1933.

[122] B. Bernfeld, "Attempts for better understanding of the directional stereophonic listening mechanism," in *Proceedings of the 44th Convention of the Audio Engineering Society*, Rotterdam, The Netherlands, 1973.

[123] C. Hugonnet and P. Walder, *Stereophonic Sound Recording.* John Wiley and Sons Ltd., 1995.

[124] V. Pulkki and T. Lokki, "Creating auditory displays to multiple loudspeakers using VBAP: A case study with DIVA project," in *International Conference on Auditory Display (ICAD)*, Glasgow, England, 1998.

[125] M. A. Gerzon, "With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, pp. 2–10, 1973.

[126] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of high order Ambisonics and Wave Field Synthesis for holophonic sound imaging," in *114th Conv. Audio Eng. Soc.*, Amsterdam, The Netherlands, Mar. 2003.

[127] W. Snow, "Basic principles of stereophonic sound," *J. SMPTE*, vol. 61, no. 2, pp. 922–940, Mar. 1953.

[128] A. Berkhout, *Applied Seismic Wave Theory.* Amsterdam: Elsevier Science, 1987.

[129] ——, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, Dec. 1988.

[130] A. Berkhout, D. de Vries, and P. Vogel, "Wave front synthesis: a new direction in electroacoustics," in *93th Conv. Audio Eng. Soc.*, no. 3379, San Francisco, USA, Mar. 1992.

[131] ——, "Acoustic control by Wave Field Synthesis," *J. Acoust. Soc. America*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[132] M. Boone and E. N. G. Verheijen, "Multichannel sound reproduction based on Wave Field Synthesis," in *95th Conv. Audio Eng. Soc.*, no. 3719, New York, USA, Oct. 1993.

[133] D. de Vries, "Sound reinforcement by Wave Field Synthesis: Adaptation of the synthesis operator to the loudspeaker directivity characteristics," *J. Audio Eng. Soc.*, vol. 44, no. 12, pp. 1120–1131, Dec. 1996.

[134] N. Bleistein, *Mathematical Methods for Wave Phenomena.* Academic, 1984.

[135] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of Wave Field Synthesis revisited," in *124th Conv. Audio Eng. Soc.*, no. 7358, Amsterdam, The Netherlands, May 2008.

[136] Carrouso, "Creating, Assessing and Rendering in Real Time of High Quality Audio-Visual Environments in MPEG-4 Context, (European Union, Information Society Technologies," http://cordis.europa.eu/ist/ka3/iaf/projects/carrouso.htm, 2001, last seen on may 2008.

[137] S. Brix, T. Sporer, and J. Plogsties, “Carrouso - an european approach to 3d-audio,” in *110th Conv. Audio Eng. Soc.*, no. 5314, Amsterdam, The Netherlands, May 2001.

[138] P. Vogel, “Application of Wave Field Synthesis in room acoustics,” Ph.D. dissertation, Delft University of Technology, The Netherlands, 1993.

[139] E. Start, “Direct sound enhancement by Wave Field Synthesis,” Ph.D. dissertation, Delft University of Technology, The Netherlands, 1997.

[140] E. Verheijen, “Sound reproduction by Wave Field Synthesis,” Ph.D. dissertation, Delft University of Technology, The Netherlands, 1997.

[141] R. Nicol, “Restitution sonore spatialisé sur une zone étendue,” Ph.D. dissertation, Université du Maine, France, 1999.

[142] E. Hulsebos, “Auralization using Wave Field Synthesis,” Ph.D. dissertation, Delft University of Technology, The Netherlands, 2004.

[143] E. Corteel, “Caractérisation et extensions de la Wave Field Synthesis en conditions réelles,” Ph.D. dissertation, Université Paris VI, France, 2006.

[144] S. Spors, “Active listening room compensation for spatial sound reproduction systems,” Ph.D. dissertation, University of Erlangen-Nürnberg, Germany, 2007.

[145] H. Wittek, “Perceptual differences between Wave Field Synthesis and stereophony,” Ph.D. dissertation, University of Surrey, United Kingdom, 2007.

[146] P.-A. Gauthier, “Synthèse de champs sonores adaptative,” Ph.D. dissertation, Université de Sherbrooke, Canada, 2007.

[147] M. Baalman, “On Wave Field Synthesis and electro-acoustic music, with particular focus on the reproduction of arbitrarily shaped sound sources,” Ph.D. dissertation, Technische Universität Berlin, Germany, 2008.

[148] B. Pueo, “Analysis and enhancements of multiactuator panels for wave field synthesis reproduction,” Ph.D. dissertation, Universidad Politécnica de Valencia, 2008.

[149] S. Spors and R. Rabenstein, “Spatial aliasing artifacts produced by linear and circular loudspeaker arrays used for Wave Field Synthesis,” in *120th Conv. Audio Eng. Soc.*, no. 6711, Paris, France, May 2006.

[150] E. Corteel, “Equalization in an extended area using multichannel inversion and wave field synthesis,” *J. Audio Eng. Soc.*, vol. 54, no. 12, pp. 1140–1161, Dec. 2006.

[151] S. Spors, “Investigation of spatial aliasing artifacts of Wave Field Synthesis in the temporal domain,” in *Proceedings of the Annual German Conference on Acoustics (DAGA)*, Mar. 2008.

[152] H. Wittek, “Perception of spatially synthesized sound fields,” 2003.

[153] W. de Brujin and M. Boone, “Application of Wave Field Synthesis in life-size videoconferencing,” in *114th Conv. Audio Eng. Soc.*, no. 5801, Amsterdam, The Netherlands, Mar. 2003.

[154] G. Thiele, H. Wittek, and M. Reisinger, “Potential Wavefield Synthesis applications in the multichannel stereophonic world,” in *Proceedings of the AES 24th Int. Conf. on Multichannel Audio*, Banff, Canada, Jun. 2003.

[155] A. Franck, A. Gräfe, T. Korn, and M. Strauß, “Reproduction of moving virtual sound sources by Wave Field Synthesis: An analysis of artifacts,” in *32nd Int. Conf. on Multichannel Audio: DSP for Loudspeakers*, Hillerød, Denmark, Sep. 2007.

[156] J. Ahrens and S. Spors, “Reproduction of moving virtual sound sources with special attention to the doppler effect,” in *124th Conv. Audio Eng. Soc.*, Amsterdam, The Netherlands, May 2008.

[157] M. Dressler, “Dolby Surround Pro Logic II decoder principles of operation,” Dolby Laboratories Information, 2000.

[158] D. Griesinger, “Multichannel matrix surround decoders for two-eared listeners,” in *Proceedings of the 101st Convention of the Audio Engineering Society*, Los Angeles, CA, USA, 1996.

[159] “NEO:6 - an overview of DTS NEO:6 multichannel,” available at: http://www.dts.com/media/uploads/pdfs/DTS

[160] “SRS circle surround,” available online at: http://www.srslabs.com/ss-technologies920.asp.

[161] C. Avendano and J.-M. Jot, “Frequency domain techniques for stereo to multichannel upmix,” in *Proceedings of the AES 22nd Conference on Virtual, Synthetic and Entertainment Audio*, 2002, pp. 121–130.

[162] J. Usher, “Design criteria for high quality upmixers,” in *AES 28th International Conference*, Pitea, Sweden, June 2006.

[163] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, “Spatial audio object coding (SAOC) - the upcoming MPEG standard on parametric object based audio coding,” in *Proceedings of the 124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, 2008.

[164] M. Cobos and J. J. Lopez, “On the application of sound source separation to wave-field synthesis,” in *Proceedings of the 122nd Convention of the Audio Engineering Society*, Vienna, Austria, May 2007.

[165] ——, “Resynthesis of sound scenes in wave-field synthesis from stereo mixtures using sound source separation algorithms,” *Journal of the Audio Engineering Society*, vol. 57, no. 3, March 2009.

[166] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, “Creating interactive virtual acoustic environment,” *Journal of the Audio Engineering Society*, vol. 35, pp. 307–316, September 1999.

[167] F. Rumsey, “Spatial audio and sensory evaluation techniques - context, history and aims,” in *Spatial audio and sensory evaluation techniques conference*, Guilford, UK, 2006.

[168] R. Pellegrini and C. Kuhn, “Wave field synthesis: Mixing and mastering tools for digital audio workstations,” in *Proceedings of the 116th Convention of the Audio Engineering Society*, Berlin, Germany, 2004.

[169] A. Misra, P. R. Cook, and G. Wang, “Musical tapestry: Re-composing natural sounds,” in *International Computer Music Conference 2006*, Tulane, New Orleans, USA, 2006.

[170] M. Cobos, J. J. Lopez, A. Gonzalez, and J. Escolano, “Stereo to wave-field synthesis music up-mixing: An objective and subjective evaluation,” in *IEEE International Symposium on Communications, Control and Signal Processing*, St. Julians, Malta, March 2008.

[171] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, and S. George, “QESTRAL (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener,” in *AES 125th Convention*, San Francisco, CA., October 2008.

[172] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality,” *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 968–977, August 2005.

[173] T. Sporer, “Spatial perception and quality - perceptual aspects of wave field syntehsis,” in *1st DEGA Symposium, Workshop Wave Field Synthesis*, Ilmenau, September 2007.

[174] F. Rumsey, “Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm,” *Journal of the Audio Engineering Society*, vol. 50, no. 9, pp. 651–666, 2002.

[175] R. Kassier, T. Brookes, and F. Rumsey, “A simplified scene-based paradigm for use in spatial audio listener training applications,” in *Audio Engineering Society 117th Convention*, San Francisco, CA, USA, 2004.

[176] W. M. Hartmann, “Localization of sound in rooms,” *Journal of the Acoustical Society of America*, vol. 75, no. 5, pp. 1380–1391, 1983.

[177] G. Bloothooft and R. Plomp, “The timbre of sung vowels,” *Journal of the Acoustical Society of America*, vol. 84, pp. 847–860, 1988.

[178] ASA, “American standard acoustical terminology,” Definition 12.9, Timbre 45, Acoustical Society of America, New York, 1960.

[179] E. Vincent, M. G. Jafari, and M. D. Plumbey, “Preliminary guidelines for subjective evaluation of audio source separation algorithms,” in *ICA Researtch Network Workshop*, University of Liverpool, September 2006.

[180] H. Wittek, F. Rumsey, and G. Theile, “Perceptual enhancement of wavefield synthesis by stereophonic means,” *Journal of the Audio Engineering Society*, vol. 55, no. 9, pp. 723–751, September 2007.

[181] S. Bube, C. Fabris, T. Hohberger, A. Köhler, J. Liebetran, T. Sporer, and A. Walther, “Perceptual evaluation of algorithms for blind up-mix,” in *Audio Engineering Society 121st Convention*, San Francisco, CA, USA, 2006.

[182] T. Neher T. Brookes and F. Rumsey, "Training of listeners for the evaluation of spatial sound reproduction," in *Audio Engineering Society 112th Convention*, Munich, Germany, May 2002.

[183] N. Ford, F. Rumsey, and T. Nind, "Evaluating spatial attributes of reproduced audio events using a graphical assessment language - understanding differences in listener depictions," in *AES 24th International Conference on Multichannel Audio*, Banff, Alberta, Canada, June 2003.

[184] H. Wittek, "Perceptual differences between wavefield synthesis and stereophony," Ph.D. dissertation, School of Arts, Communication and Humanities, University of Surrey, October 2007.

[185] R. Lowry, *Concepts and Applications of Inferential Statistics.* online at http://faculty.vassar.edu/lowry/webtext.html, 1999.

[186] M. Cobos and J. J. Lopez, "Method and apparatus for stereo enhancement in stereo audio recordings with application to mobile devices," European Patent P200 802 379, 2008.

[187] N. Madhu and R. Martin, *Advances in Digital Speech Transmission.* Wiley, 2008, ch. Acoustic Source Localization with Microphone Arrays, pp. 135–166.

[188] M. T. Dang and S. Nam, "A new cost function for direction-of-arrival estimation of multiple sound sources using two microphones," in *Proceedings of the International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC 2008)*, Seattle, WA., September 2008.

[189] F. A. Reed, P. L. Feintuch, and N. J. Bershad, "Time delay estimation using the LMS adaptive filter-static behaviour," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 561–571, June 1981.

[190] R. Schmidt, "Multiple emitter location and signal parameters estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, March 1986.

[191] R. M. Stern, G. J. Brown, and W. D., *Computational Auditory Scene Analysis.* Wiley Interscience, 2006, ch. Binaural Sound Localization, pp. 147–178.

[192] C. Liu, B. C. Wheeler, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1888–1905, 2000.

[193] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[194] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," in *Proceedins of the 9th Int. Conference on Digital Audio Effects (DAFx-06)*, Montreal, Canada, September 2006, pp. 233–238.

[195] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *Proceedings of the International Workshop for Acoustic Echo Cancellation and Noise Control (IWAENC 2008)*, Seattle, WA., September 2008.

[196] N. Mitianoudis and T. Stathaki, “Overcomplete source separation using laplacian mixture model,” *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 277–280, 2005.

[197] A. P. Dempster, N. Laird, and D. Rubin, “Maximum likelihood for incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Ser. B*, vol. 39, pp. 1–38, 1977.

[198] N. Mitianoudis and T. Stathaki, “Batch and online underdetermined source separation using laplacian mixture models,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1818–1832, 2007.

[199] D. R. Campbell, “Roomsim: a MATLAB simulation shoebox room acoustics,” 2007, http://media.paisley.ac.uk/ campbell/Roomsim.

[200] R. C. Maher, “Old and new techniques for artificial stereophonic image enhancement,” in *Proceedings of the 101st Convention of the Audio Engineering Society*, Los Angeles, CA, USA, 1996.

[201] A. Klayman, “Stereo enhancement system,” U.S. Patent 4,748,669, 1988.

[202] G. J. Barton, “Signal enhancement method for stereo system,” U.S. Patent 4,910,778, 1990.

[203] S. W. Desper, “Automatic stereophonic manipulation system and apparatus for image enhancement,” U.S. Patent 5,412,731, 1995.