POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Master of Science in
Computer Engineering

An approach to Source Separation with Microphone Arrays based on a Plenacoustic Representation of the Sound Field

Candidate

Fabrizio D'Amelio
Student Id. number 797472

Thesis Supervisor

Prof. Fabio Antonacci

Assistant Supervisor

Dr. Lucio Bianchi

Academic Year 2014/2015

POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Laurea Magistrale in
Ingegneria Informatica

Un approccio alla Separazione di Sorgenti con Schiere di Microfoni basato sulla Rappresentazione Plenacustica del Campo Sonoro

Candidato

Fabrizio D'Amelio
Matricola 797472

Relatore

Prof. Fabio Antonacci

Correlatore

Dr. Lucio Bianchi

Anno Accademico 2014/2015

An approach to Source Separation with Microphone Arrays based on a Plenacoustic Representation of the Sound Field

Master thesis. Politecnico di Milano

This thesis has been typeset by LaTeX and the smcthesis class.

Author's email: fabrizio.damelio@mail.polimi.it

Dedicated to
...

Sommario

L'elaborazione di segnali di parlato è un ambito di ricerca di grande interesse oggigiorno. Questo è dovuto alle sue eccezionali applicazioni, volte a migliorare la qualità di molti dispositivi. I metodi di separazione di segnali di tipo parlato sono impiegati in sistemi di elaborazione del parlato. In applicazioni reali, i metodi di separazione devono far fronte alla mancanza di informazioni a priori sul numero di sorgenti e sulla loro posizione nella scena acustica. Questo problema è noto in letteratura come *Blind Source Separation* (BSS). Noi proponiamo un approccio al BSS basato su una rappresentazione plenacustica. Questo approccio codifica l'informazione della funzione plenacustica, campionata in diversi punti, in una struttura dati definita come ray-space image. Campionare la funzione plenacustica significa stimare gli pseudospettri in svariati punti lungo la *Observation Window*. Ciò può essere realizzato utilizzando un *Uniform Linear Array* (ULA) di microfoni suddiviso in sotto array. Al fine di ottenere un sistema reattivo, è stato sviluppato un algoritmo efficiente per calcolare gli pseudospettri a ogni sotto array. L'algoritmo permette di precalcolare una matrice di trasformazione per ottenere gli pseudospettri con una sola moltiplicazione di matrici. Partendo dagli pseudospettri, la wideband ray-space image è composta combinando i corrispettivi pseudospettri. Una defezione di accuratezza è stata riscontrata quando gli spettri su più frequenze sono calcolati per i segnali di parlato a causa della distribuzione energetica del segnale e dei limiti di risoluzione dell'ULA. Proponiamo un nuovo e accurato algoritmo, basato sul contenuto frequenziale dei segnali di parlato, che è anche robusto agli errori di aliasing spaziale. La ray-space image ottenuta permette di rappresentare come linee i soggetti parlanti. Di conseguenza, metodi di pattern analysis possono essere impiegati per individuare queste linee e stimare la posizione delle rispettive sorgenti nella scena acustica. In questo modo il BSS viene trasformato in un problema informed. L'approccio plenacustico, usato in precedenza, viene quindi applicato per realizzare la separazione. I filtri di beamforming *Linearly Constrained Minimum Variance* (LCMV) sono implementati per ciascun sotto array, per ottenere stime multiple da diversi punti di vista. L'approccio multi vista è essenziale al fine di ottenere risultati soddisfacenti di separazione del parlato, quando i soggetti parlanti sono allineati rispetto a un microfono dell'array. Infatti, i filtri LCMV garantiscono le migliori prestazioni in termini di soppressione delle sorgenti interferenti, ma presentano anche l'inconveniente di fallire nella separazione quando la loro differenza angolare è irrisoria. Quindi, l'infattibilità della separazione delle sorgenti in caso di sovrapposizione delle stesse è compensata dagli apporti degli altri sotto array, dai quali le sorgenti sono viste con una differenza angolare maggiore. Inoltre, la conoscenza dell'esatta posizione, nei limiti dell'errore di stima, permette di ricostruire i segnali stimati come se fossero stati acquisiti da un microfono virtuale posto vicino al soggetto parlante. I risultati ottenuti dalle simulazioni hanno confermato la validità del metodo proposto in caso di sovrapposizione delle sorgenti e in caso di differenze angolari elevate. Notiamo che le prestazioni misurate con metriche oggettive sono state confermante da *Mean Opinion Score* raccolti durante sessioni di test percettivi. I risultati ottenuti dimostrano che l'approccio plenacustico supera il metodo LCMV in prestazioni nella maggior parte dei casi sottoposti. In conclusione, risultati promettenti sono stati ottenuti in ambienti reali riverberanti.

Abstract

Speech processing is a main interest in the research scenario nowadays. This is due to its several outstanding applications, aimed at increasing the quality of many every day devices. Among the different applications, that of speech separation is crucial for enabling many services. Often in real world applications, speech separation methods have to cope with the lack of a priori information on the number of speakers and their positions in the sound scene. This challenging separation problem is known as *Blind Source Separation* (BSS) in the literature. Our work proposes an approach to BSS based on a plenacoustic representation. This approach encodes the plenacoustic function information, sampled in several points, in a data structure defined in the literature as ray-space image. The ray space image consists in a measurement of the directional components of the sound field in several points along an *Observation Window*. This can be easily done by using a *Uniform Linear Array* (ULA) of microphones subdivided into smaller sub-arrays. In order to achieve responsiveness of the system, a fast algorithm to measure the directional components of the sound field at each sub-array has been devised. The algorithm exploits the fact that the directional components of the sound field at each sub-array can be estimated through a precomputed linear transformation of the acquired signals. Starting from the directional components at each sub-array, a wideband ray-space image is obtained. An accuracy issue emerges when the wideband pseudospectra are calculated for speech signals because of speech energy distribution, and resolution limits of the ULA. We propose an accurate algorithm, based on the peculiar frequency content of speech signals, which is also robust to spatial aliasing errors. The ray-space image obtained permits to intuitively visualize the active speakers in the sound scene as lines. Consequently, well-known pattern analysis methods are employed to detect these lines and estimate the position of the related sources in the sound scene. This way the blind source separation is turned into an informed problem. The plenacoustic approach adopted to localize speakers is then applied to perform speech separation. Accordingly, *Linearly Constrained Minimum Variance* (LCMV) beamforming filters are implemented at each sub-array to extract multiple estimations, from different points of view, of the speech signals. The multiple-view approach is essential to obtain satisfactory speech separation results when speakers are aligned with respect to one microphone of the array. In fact, LCMV filters provide the best performances in terms of interference rejection, but present also the important drawback of failing in separating sources when their angular displacement is too small. Thus, the unfeasible separation of sources in case of source overlap is compensated by the other sub-array contributions, from which a larger angular displacement is attained. Furthermore, the knowledge of the exact position, up to an estimation error, allows to back-propagate the estimated signals. The final speech signal emulates a virtual microphone placed near the speaker. Results obtained with simulation sessions have confirmed the validity of the proposed method in case of source overlap and large angular displacements. Interestingly enough, the separation performances measured with objective metrics have been confirmed by *Mean Opinion Scores* collected with a campaign of perceptive tests. The results achieved also show that the plenacoustic approach outperforms the LCMV method in the majority of the situations. Finally, promising results have been obtained in real world reverberant environments.

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An increasing interest in speech processing techniques has been experienced in the last few years. Researchers and companies are striving to make smart systems capable of interacting with humans in a totally different way to the ones seen so far. It is through the improvement of these interactive aspects that a higher level of quality and experience can be achieved. In such a scenario, a plethora of possible applications have been devised. For example, automatic speech recognition [4], smart home solutions [5], teleconferencing and hands-free communication [6] and hearing aids [7] are a part of our every day interaction with speech processing technologies. These applications rely on speech separation techniques to process, hence to attenuate or enhance, a single or multiple speech signals coming from different directions to obtain an estimate of each single speech signal from a mixture. A particularly challenging scenario to resolve speech separation is when no information is assumed both on the number of speech sources acting in the sound scene, and on their positions. In addition, if sources are not static but move freely in space interacting with the surrounding environment, the speech separation task is even more complex. This problem takes the name of *blind source separation* in the literature [8].
This thesis provides an approach to blind source separation using an array of microphones. Array processing is a well-known technique in signal processing [9] that finds numerous applications from sound field acquisition and analysis [2] to its processing, by attenuating or enhancing sound waves coming from a specific direction, i.e. spatial filtering [10]. In this work, we show how the two fields are strictly tied up to achieve satisfactory speech separation results. By employing a single extended *Uniform Linear Array* (ULA), we are able to capture a sound field and extract important parameters as a first step, then, as a second step, these parameters are used to perform spatial filtering and extract speech signals singularly from the mixture acquired at microphones. Essential sound field parameters can be estimated thanks to the plenacoustic approach suggested in [2], on which our proposed source separation method relies.
The main idea behind a plenacoustic approach to sound field analysis is to measure the *Plenacoustic Function* [11] in several points in space. The plenacoustic function mathematically describes the acoustic radiance in every direction through every point in space. If we want to measure the plenacoustic function in a single point, we can do so by centering a microphone array in that location, and estimating through spatial filtering the acoustic radiance along all the look directions (pseudospectrum)

[12]. The spatial filtering technique of beamforming [13] is employed for this purpose and allows us to point a beam in space and to acquire acoustic energy irradiated from a specific *Direction Of Arrival* (DOA), while attenuating all the others. A device of this sort is called "acoustic camera". A natural extension of this concept would be that of a "plenacoustic camera", intended as a theoretical device that acquires the plenacoustic function over a spatially continuous "Observation Window" (OW) facing the acoustic scene. We are interested in implementing a device that captures the plenacoustic function over an OW based on an array of microphones. One rather straightforward way of doing so is to think of this device as an array of acoustic cameras that sample the OW. This can be easily obtained by subdividing a microphone array into smaller sub-arrays. The sampling operation in space, given by the inter-microphone distance, introduces a degradation on the information acquired called spatial aliasing. The greater the distance between microphones the greater the spatial aliasing introduced. Then, each sub array provides only an approximation of the acoustic radiance of the sound field from its point of view. What we would like to achieve is a smart parametrization of this different information acquired at sub-arrays in order to visualize the sound scene in an intuitive way.
In [2] Markovic et al. proposed a parametrization that permits to visualize as lines in an image, i.e. the ray-space image, the acoustic rays emitted by audio sources. The method was conceived to work with every kind of audio signal, not considering the specific distribution of frequency content of speech signals and the spatial aliasing error that might affect the resultant ray-space image. Since we are focusing on speech processing, we designed a wideband image reconstruction method more robust to spatial aliasing, and consistent to speech energy distribution in frequency. This method is aimed at estimating a precise wideband pseudospectrum image, starting from its several narrowband components that span the whole frequency range. Speech signals are composed by harmonic vowels, where most of the energy resides, and noise-like consonants, perceptually significant but with content at high frequencies that might be affected by spatial aliasing. In addition, low frequencies DOAs are coarsely estimated, due to their long wavelengths, while high frequencies provide a detailed information on DOAs. We need a method which considers this structure and properly weights the narrowband components of the pseudospectrum to produce its wideband equivalent. To this intent, we resorted the *spectral flatness* measure [14], which indicates whether the frequency signal resembles a white noise (flat frequency content) or a harmonic sound (spiky frequency content), to discern which frequencies consider the most in the computation of the wideband image. Spectral flatness is used jointly with the sum of energy content at each frequency, to maintain energy ratios as much unaltered as possible. Furthermore, a fast algorithm to calculate the image is desirable in order to have a responsive characterization of the time varying sound scene. In this work, we also propose a first implementation of a fast ray-space image computation that might open opportunities for further work to achieve real time performances.
The characteristic mapping of audio sources into lines of the ray-space image enables easy linear pattern analysis [15] to detect these lines and estimate the position of the correlated source. This way, we obtain precious information about the number of active speakers in the sound scene and their position in space. Given that we can compute the ray-space image adaptively in time, by framing the signals in short

time windows, we are able to track speakers movements. The aforementioned blind source separation problem has just been turned into an informed one. Since we now know sources positions and our plenacoustic camera is based on beamforming, it is natural to employ spatial filtering technique of beamforming once again to point a beam towards one of the speakers and to attenuate all the rest. A well-known beamforming method called *Linearly Constrained Minimum Variance* (LCMV) [16], allows us to constrain the filter output to present some predefined desired responses for signals coming form specific DOAs. A resolution problem emerges when two speakers draw close together, destabilizing the system that tries to attenuate, and at the same time, to enhance two different sources arriving from the same direction. This flaw has been resolved through diagonal loading [17]. A spatially white noise can be thought as a decorrelated random signal coming from every direction. On one hand, the outcome of this operation is a less effective spatial filtering, on the other hand, filters are more robust to instabilities.
Once we know the number of speakers, their position, and our system is robust to instabilities, we can proceed with performing speech separation recalling the plenacoustic approach. To this end, we perform beamforming at each sub-array to extract the relative estimation of the speaker signal. A particularly useful situation in which this multiview approach proves its validity is when speakers, free to move in the scene in front of the array, lie in front of each other, occluding themselves with respect to one microphone of the array. In fact, the standard beamforming-based approach calculates the different sources DOAs with respect to the array center. If two or more speakers lie on the line that connects the array center with themselves, the system considers those speakers as one unique speech source, hampering DOA-based separation of beamforming. With our method, instead, if a sub-array fails to separate speaker signals because of an overlapping situation, the other sub-arrays come into help, providing their contribution acquired from a different position, in which sources are seen under different angles. Thus, we are able to perform speech separation independently to angular displacement of speakers by simply weighting sub-array contributions according to the difference of DOA. Furthermore, the estimation of source positions allows us to process the acquired signals in order to back-propagate them. The perceptive result obtained is a virtual microphone that can be placed at any point around the speaker. Indeed, in the simplest case in which the surrounding environment is not taken into account, sound waves propagation consist in a simple attenuation dependent on the distance ranged. Then, sound waves at microphones are just a delayed and attenuated version of the sound waves produced at the speaker position. Knowing the speaker position permits to invert this process and simulate the waves as if they were acquired in a single desired point. Source separation issues get more intertwined if the environment in which speakers move is considered. Walls, floors and other objects in the scene under analysis reflect the sound waves produced by speakers, making rejection of undesired sources harder. If we could tighten up our beamforming system to augment the precision of acquisition towards a specific direction, we would be able to reject most of the acoustic energy coming from other directions hence guaranteeing a satisfactory degree of separation. A beamforming method based on this consideration has been built in [3], and it has been utilized in our system, when we tested it against reverberations. Indeed, a campaign of simulations and experiments in a semi-anechoic and a reverberant

room has been conducted to validate our proposed method. In addition, it has been pointed out in the literature [18], [19], that objective measures employed in simulations do not always reflect speech features at a higher level of abstraction. This is due to the coarse relationship between objective measures and the auditory system. Therefore, an assessment of separation algorithms by means of perceptive tests is usually required to validate and compare a method with others. In light of this, we conducted a campaign of perceptive tests that found a correspondence with the objective results calculated on the extracted signals, after separation processing, as in [20].

The manuscript is organized as follows. In Chapter 2 we provide an overview on the theoretical background of the signal processing techniques employed in this thesis. Once acquainted with the processing tools needed, we proceed in Chapter 3, with a detailed description of the plenacoustic approach, on which our method relies. We show how to calculate the ray-space image in a fast and robust way that appropriately manages aliasing errors and peculiar energy distribution of speech formants. Then, in Chapter 4 we focus on the speech separation filters to be employed for separating speech signals, relying on the information drawn from the ray-space image. In addition, we show how to merge the sub-array signals to virtualize a microphone at any point in space. Finally, in Chapter 5 we corroborate the method proposed with properly designed simulation results. The simulations are aimed at endorsing the quality of the method and the robustness on the localization error, angular displacement, and source overlap in several situations. The method proposed showed satisfactory results mirrored in the perceptive test and real-world scenario performances.

Chapter 2

Theoretical Background

In this chapter, we introduce the reader to spatial filtering concept, beamforming techniques, plenacoustic function and how they are applied to source separation problems, which is the final goal of this work of thesis. Although source separation has been performed with different approaches, we propose a new method based on the plenacoustic function and its representation in the "ray space". The main idea behind the plenacoustic approach is to have a responsive and intuitive representation of the sound field that permits to easily assess important parameters. Thus, the ray-space image allows us to perform source localization, tracking and extraction of other sound field parameters that are fundamental for powerful location-informed source separation filters. Both source separation filters and the ray-space image acquisition processes rely on beamforming techniques. Beamforming is a spatial filtering method which exploits the information conveyed by an array of sensors. We will discuss how beamforming methods are employed in the source separation field, providing also a thorough state-of-the-art review to picture the overall research carried out so far. In particular, *Linearly Constrained Minimum Variance* (LCMV) and *Minimum Variance Distortionless Response* (MVDR) beamforming techniques will be analyzed in depth, pointing out their weaknesses and strengths in different situations of ambient noise and source to interference parameters. We will pay particular attention on the *LCMV*, that is the beamforming technique used in our approach, since, in its theoretical formulation, gives the best results in terms of separation of sources. However, its performances decrease when ambient noise is preponderant or at least comparable with the target sources to be separated.
The whole method developed in this work of thesis is described in Chapters 3 and 4 but its insights and concepts refer to the theoretical background defined here.
The Chapter is organized as follows. In Section 2.1 we present the basics of spatial filtering (or array processing) on which our method relies. In Section 2.1.2 conventional methods of beamforming are presented. In Section 2.2 we will show how these methods are employed in source separation problems highlighting their behavior and limitations. In Section 2.3 a representation of the plenacoustic function, namely the ray-space image, is described stating what advantages brings to our problem. One important application of the ray-space image is the source localization, which is carried out by means of linear pattern analysis, thanks to its clever representation of acoustic rays as lines. Source localization is discussed in 2.4.

2.1 Spatial filtering

The methods and techniques for source separation that will be discussed later in Section 2.2 and Chapter 4, rely on spatial filtering. Although, there are studies to enhance signals over noise in single sensor signal processing literature, performances are not comparable with those obtained with spatial filtering for source separation problems. The reason is that a single sensor does not exploit spatial information given by the presence of redundant sensors. Spatial filtering or array signal processing is a specialized branch of signal processing that focuses on information conveyed by propagating waves. By cleverly combining sensors outputs, spatial filtering could address different tasks, such as enhancing the *Signal to Noise Ratio* (SNR) beyond that of a single sensor's output; as well as characterizing the field by determining the number of sources, their locations and the waveforms they are emitting; and also tracking the sources as they move in space [9]. Although spatial filtering can be applied in several fields, we will restrict our work to audio signals, and specifically to speech signals, without losing generality.
As mentioned, each sensor samples the sound field giving as output a signal $y_m(n)$ where n is the time index in samples and $m = 1, \dots, M$ with M equal to the number of sensors, is the index of the sensors within the array. Each wave varies in time and space accordingly to the direction of propagation. The direction of propagation of a wave can be estimated using an array of microphones by exploiting the different time of arrivals of the wave at each sensor. In Section 2.1.2, we discuss a technique called *beamforming* which builds upon this concept.
One possible application of array processing is estimating how energy is distributed over space, or spatial spectral estimation from a frequency domain point of view [12]. An array of sensors samples the field at different locations gathering the energy over a finite area, called *aperture*. Let us now define a first model for the output signal of the receiving sensor array. This model will be then expanded as some hypothesis will be dropped to work under different scenarios. As a first assumption, we consider the sources to be situated in the far field of the array, i.e. the wavefront may be considered planar at the sensor positions with respect to the array length, without introducing sensible errors. This assumption computationally simplifies the model, because the only parameter that characterizes the source locations is the so-called angle of arrival, or *Direction Of Arrival* (DOA). However, the method proposed in this thesis works both with the far-field and near-field assumptions, hence, we will show both models and their differences. Furthermore, we assume that both the sources and the sensors in the array are in the same 2D plane and that the sources are point emitters. In addition, it is assumed that the propagation medium is homogeneous (i.e., not dispersive) so that the waves arriving at the array can be considered to be planar. It is also assumed that the number of sources L is known. Finally, it is assumed that the sensors in the array can be modeled as linear (time–invariant) systems; and that their transfer characteristics are ideal, i.e. microphones' transfer function is equal to 1 at each frequency bin and omnidirectional, i.e. the microphones capture waves coming from every direction equally. Finally, let us assume that microphones locations are known. Under these assumptions an array is called "calibrated".
The source in Figure 2.1 generates a wave field that travels through space and is

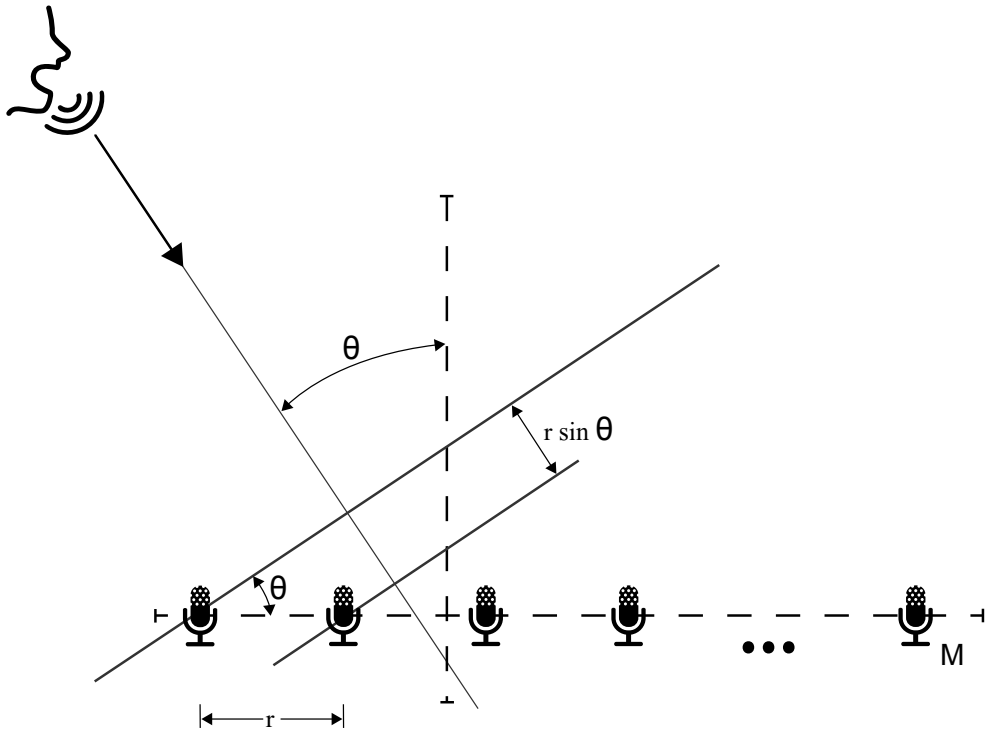


Figure 2.1. Model of sound propagation at the array, under far-field assumption

sampled, in both space and time, by the array. It is not restrictive to think the model with just one source, since once it is obtained for one source, the superposition principle guarantees its validity for multiple sources. As shown in Figure 2.1, we define a model for a *Uniform Linear Array* (ULA), i.e. its sensors are deployed with the same distance along a line. Among all the possible array configurations we analyze the ULA case because it is the one adopted in Chapter 3, and 4. Hence, supposing that a single waveform in an *anechoic environment* (no reverberations) is captured by the array, let $s(n)$ denote the value of the signal waveform as measured at some reference point, at discrete time n. It is customary to take one of the microphones as reference point. The physical signals received by the array are discrete time waveforms. Hence n is a discrete variable which takes values $n = 1, .., N$, unless otherwise stated. Let τ_m denote the time needed for the wave to travel from the reference point to sensor m ($m = 1, \dots, M$). Then the output of sensor m can be written as

$$y_m(n) = s(n - \tau_m) + e_m(n) \tag{2.1}$$

where e_m is an additive noise. The noise may enter in equation (2.1) either as "thermal noise" generated by the sensor's circuitry, as "random background radiation" impinging on the array, or in other ways. In (2.1), the "input" signal $s(n)$, as well as the delay τ_m, are unknown and the source location enters in (2.1) through τ_m. Hence, the source location problem is basically one of time-delay estimation. It is not true in the near-field case, where a source location is determined by its coordinates in space, thus, τ_m have to be assessed at each microphone depending on the relative distance. Going back to the far-field case, equation (2.1) can be reformulated in the

frequency domain by using the *Fourier Transform*, so to have $s(\omega)$ and $e_m(\omega)$ to be the frequency counterparts of $s(n)$ and $e_m(n)$ respectively, with ω denoting the angular frequency. Then we can write

$$y_m(\omega) = s(\omega)e^{-i\omega\tau_m} + e_m(\omega). \tag{2.2}$$

Assuming the signal $s(\omega)$ to be narrowband we can write the *Array Transfer Vector* (or *Array Steering Vector*) with fixed ω, so we can ignore its dependency to frequency,

$$\boldsymbol{a}(\theta) = [1 \quad e^{-i\omega\tau_2} \dots \quad e^{-i\omega\tau_{M-1}}]^T, \tag{2.3}$$

and the overall model, which is a generalization with L sources,

$$\boldsymbol{y}(n) = [\boldsymbol{a}(\theta_1) \dots \boldsymbol{a}(\theta_L)] \begin{bmatrix} s_1(n) \\ \vdots \\ s_L(n) \end{bmatrix} + e(n) \triangleq \boldsymbol{A}\boldsymbol{s}(n) + \boldsymbol{e}(n) \tag{2.4}$$

where θ_l represents the DOA of the lth source and $s_l(n)$ is the lth source signal. Also, we call

$$\boldsymbol{x}(n) = \boldsymbol{A}\boldsymbol{s}(n), \tag{2.5}$$

which represents the noiseless signal at microphones. Herein, we assumed the signal $s(n)$ to be narrowband to derive our model, this does not represent a burdensome limitation since we can reconduct a wideband signal to its frequency domain components by using the *Discrete Fourier Transform* (DFT). We call ω_k one of this components or frequency bins, then we can consider the signal at each bin as narrowband. The planar wave assumption will now be considered, as τ_m depends only on θ. Let d be the inter-microphones distance, then

$$\tau_m = (M-1)\frac{d\sin(\theta)}{c} \quad \text{for} \quad \theta \in [-90°, 90°], \tag{2.6}$$

where c is the propagation velocity of the impinging waveform inserting (2.6) into (2.3) gives

$$\boldsymbol{a}(\theta) = \left[1 \; e^{-i\omega d\sin(\theta)/c} \dots e^{-i(M-1)\omega d\sin(\theta)/c}\right]^T. \tag{2.7}$$

The restriction of θ to lie in the interval $[-90°, 90°]$ is a limitation in ULAs configurations. In fact, two emitting sources positioned at symmetric locations with respect to the array line lead to identical delays τ_m, hence sources cannot be distinguished from each other.
Since we may consider the ULA as performing a uniform spatial sampling of the wavefield along a line, d, i.e. the spatial sampling period, should be smaller than half of the signal wavelength. If the condition is met, the wavelength taken in consideration is perfectly sampled avoiding ambiguities in DOA estimation. This is utterly analogous with the *Shannon sampling theorem* in the time-frequency domain. Let λ denote the signal wavelength

$$\lambda = c/f, \quad f = \omega/2\pi \tag{2.8}$$

being f the temporal frequency of the signal. The spatial frequency is then defined as

$$\omega_s = 2\pi f_s = \omega_k \frac{d\sin(\theta)}{c}, \tag{2.9}$$

where $k = 1, \dots, K$ is the frequency bin. Finally (2.7) can be rewritten as a *Vandermonde* vector

$$\boldsymbol{a}(\theta) = \left[1 \; e^{-i\omega_s} \dots e^{-i(M-1)\omega_s}\right]^T. \tag{2.10}$$

In the ULA case (2.10) is uniquely defined, i.e. there is no "spatial aliasing", if $\omega_s \leq \pi$. This condition is equivalent to $\mid f_s \mid \leq \frac{1}{2} \Leftrightarrow d \mid \sin(\theta) \mid \leq \frac{\lambda}{2}$. In the worst case $\mid \sin(\theta) \mid = 1$ and the aliasing condition becomes

$$d \leq \frac{\lambda}{2}. \tag{2.11}$$

Let us expand the *array transfer vector* showing now the frequency dependency, in order to be suitable for wideband scenarios,

$$\boldsymbol{a}(\theta, \omega) = \left[1 \; e^{-i\omega d \sin(\theta)/c} \dots e^{-i(M-1)\omega d \sin(\theta)/c}\right]^T \tag{2.12}$$

Then, the processing is carried out on each frequency bin, in the time or frequency domain. Space, time and frequency are the three dimensions considered in the design of spatial filters.

2.1.1 Temporal filter analogy

As previously stated, one of the tasks spatial filtering accomplishes is the estimation of sources waveforms exploiting spatial information. This can be done by focusing on a target source at a specific DOA and finds its equivalence in the time-frequency domain as focusing on a single, or a class of frequencies (notch filters, lowpass filters etc.). As Linear filtering allows us to separate frequencies, a *spatio-temporal* filtering is needed to separate signals according to their directions of propagations and their frequency content. Temporal FIR filter is defined by the relation

$$y_F(n) = \sum_{m=0}^{M-1} h_m s(n-m) \triangleq \boldsymbol{h} * \boldsymbol{y}(n) \tag{2.13}$$

where h_m are the filter weights, $s(n)$ is the input signal and $\boldsymbol{h} = [h_0 \dots h_{M-1}]^*$,$\boldsymbol{y}(n) = [s(n) \dots s(n-M+1)]^T$. Considering $\boldsymbol{a}(\omega) = [1 \; e^{-i\omega} \dots e^{-i(M-1)\omega}]$ we can rewrite the filter in time in a similar form as the spatial filter

$$y_F(n) = [\boldsymbol{h}^* \boldsymbol{a}(\omega)] \, s(n) \tag{2.14}$$

by selecting proper values of $\boldsymbol{h}$ as described in the filter design theory one can attenuate or enhance the power of $y_F(n)$ at frequency ω. The same holds true in the spatial case

$$y_F(n) = [\boldsymbol{h}^* \boldsymbol{a}(\theta)] \, s(n). \tag{2.15}$$

By selecting appropriate values of $\boldsymbol{h}$ we can enhance or attenuate signals coming from a given direction θ. This is the main idea behind spatial filtering and the so-called techniques of *beamforming*.

2.1.2 Beamforming

Early spatial filters were conceived to attenuate or enhance signals impinging on the array from a specific direction. The response of beamforming to waves impinging in the array from different DOAs takes the name of *beampattern* and is defined as

$$\boldsymbol{B}(\theta,\omega) = \mid \boldsymbol{h}(\theta,\omega)^H \boldsymbol{a}(\theta,\omega) \mid^2 \tag{2.16}$$

where $(\cdot)H$ is the Hermitian of $\boldsymbol{h}(\theta,\omega)$. Different filter designs of $\boldsymbol{h}(\theta,\omega)$ are available for beamforming depending on how the problem is modeled, and which assumptions are considered. For sake of completeness we will discuss the most important methods, pointing out in which situations are more suitable and why they are used in our work or not.
First of all, we might categorize beamforming methods in data-independent, statistically optimal, parametric, and adaptive [13]. As the name suggests, data-independent beamformers do not depend on the input signal(s) data, and the beampattern is fixed when the DOA is fixed. These beamformers are the simplest in their theoretical formulation and in their computation. In fact, no estimation of second order statistics is required, which is instead needed in the statistically optimal beamformers to assess the optimal array response. When the statistics of the array data are not known and cannot be estimated, adaptive algorithms are typically employed to determine the filter weights. Two adaptation strategies may be chosen, block adaptation, where statistics are estimated from a temporal block of array data and used in an optimum weight equation, and continuous adaptation, where the weights are adjusted as the data is sampled, so that the resulting weight vector sequence converges to the optimum solution. When the number of sensors increases up to fifty or more, convergence time and computational load might be an issue. Partially adaptive beamformers reduce the adaptive algorithm computational load at the expense of a loss in statistical optimality.

2.1.3 Statistically optimal beamforming

Several algorithms have been devised that follow the statistically optimal weights approach. Several different least-squares solution can be found depending on how the problem is stated.
Let us start with a simple least-squares solution on the desired beampattern. Considering $\boldsymbol{B}(\theta,\omega)_d$ as the desired response of our filter, we seek a solution $\boldsymbol{B}(\theta,\omega)$ which approximates the desired response. It can be carried out by means of minimization problems as in FIR filter formulations. Let us state the problem as a L_2 norm minimization in L points (θ_l,ω_l) with $1 \leq l \leq L$ if $L > M$ we obtain an overdetermined least-square problem

$$\min_h \mid \boldsymbol{A}^H \boldsymbol{h} - \boldsymbol{B}_d \mid^2 \tag{2.17}$$

where we dropped the dependency on θ and ω for conciseness, and $\boldsymbol{B}_d(\theta,\omega) = [\boldsymbol{B}_d(\theta_1,\omega_1)\dots\boldsymbol{B}_d(\theta_L,\omega_L)]^H$, provided that $\boldsymbol{A}\boldsymbol{A}^H$ is invertible, the solution to (2.17) is

$$\boldsymbol{h} = (\boldsymbol{A}\boldsymbol{A}^H)^{-1}\boldsymbol{A}\boldsymbol{B}_d \tag{2.18}$$

The *white noise gain* of a beamformer is defined as the output power due to unit variance white noise at the sensors. White noise gain is therefore computed as the squared norm of the weight vector, $\boldsymbol{h}^H\boldsymbol{h}$. If the white noise gain is large, then the accuracy by which $\boldsymbol{h}$ approximates the desired response is a questionable point, as the beamformer output will have a poor SNR due to white noise contributions. The matrix $\boldsymbol{A}$ is ill-conditioned when the numerical dimension of the space spanned by the $\boldsymbol{a}(\theta_l, \omega_l)$,$1 \leq l \leq L$ is less than M. If $\boldsymbol{A}$ is ill-conditioned, then $\boldsymbol{h}$ can have a very large norm, which is an undesirable feature. A solution to this problem is the so-called *diagonal loading* [17].
Let us know define the filter output variance as the expected value of the squared absolute output values $E\{\mid \boldsymbol{y}_F(n) \mid^2\} = \boldsymbol{h}^H\boldsymbol{\Phi}_y\boldsymbol{h}$, where $E[\,\cdot\,]$ is the expected value operator and $\boldsymbol{\Phi}_y$ is the covariance matrix of microphone signals $\boldsymbol{y}(n)$ defined as

$$\boldsymbol{\Phi}_y = E\left[\boldsymbol{y}(n)\boldsymbol{y}^H(n)\right]. \tag{2.19}$$

Another different solution can be found if a minimization on the output variance of the solution is sought. One of the most important methods is called *Linearly Constrained Minimum Variance* (LCMV) beamforming [16]. The main idea behind this method is to constrain the response of the beamformer so that signals from the directions of interest are filtered with specified gain and phase, while the output variance is minimized. Use of linear constraints is a very general approach that permits extensive control over the adapted response of the beamformer. The weights are chosen to minimize output variance and honoring the constraints. Thus, by linearly constraining the filter to satisfy $\boldsymbol{h}^H\boldsymbol{a}(\theta_l, w) = g$, where g is a complex constant, we ensure that any signal from angle θ_l and frequency ω is passed to the output with response g. The LCMV problem is then written as

$$\min_h \boldsymbol{h}^H\boldsymbol{\Phi}_y\boldsymbol{h} \quad \text{subject to} \quad \boldsymbol{a}^H(\theta,\omega)\boldsymbol{h} = g^* \tag{2.20}$$

whose solution is

$$\boldsymbol{h} = g^*\frac{\boldsymbol{\Phi}_y^{-1}\boldsymbol{a}(\theta,\omega)}{\boldsymbol{a}^H(\theta,\omega)\boldsymbol{\Phi}_y^{-1}\boldsymbol{a}(\theta,\omega)}. \tag{2.21}$$

Starting from (2.21), it is interesting to show how the other beamforming solutions are derived. If we set $g = 1$, we impose a real constraint only on the look direction which corresponds to passing undistorted the signal arriving from that direction. This corresponds to the *Minimum Variance Distortionless Response* (MVDR) beamformer [21].
The single constraint in the LCMV filter is easily generalized to multiple linear constraints to add control over the beampattern as indicated in [16]. For example, it might be useful to place a zero on a particular DOA because of an undesired source coming from that direction. Then, we build a matrix called constraint matrix

$$\boldsymbol{C} = [\boldsymbol{a}(\theta_1,\omega), \ldots, \boldsymbol{a}(\theta_P,\omega)], \tag{2.22}$$

where P is the total number of constraints. Now, what we called g is a vector of desired responses $\boldsymbol{g}$ which takes the name of response vector. Point constraints fix the beamformer response to have a specific value at a specific DOA θ and frequency ω_k. The number of points at which response can be constrained is limited to M.

In fact, each linear constraint uses one degree of freedom in the weight vector, so with L constraints there are only $M - L$ degrees of freedom available for minimizing variance [13]. If M constraints are used then there are no degrees of freedom left for power minimization and a data independent beamformer is obtained. Thus, we have a trade-off between variance minimization and beampattern control.
This kind of approach to beamforming is very versatile when dealing with source separation problems. As we discuss in Section 2.2.1, by properly tuning the response vector either optimal interference suppression, or higher noise suppression can be achieved. Further, these kind of filters can be reduced to constrained delay and sum beamformers that can be quickly calculated. In this case no second order estimation is needed, which makes the system a bit less robust but faster. For these reasons, statistically optimal filters are employed in our plenacoustic based source separation method.

2.1.4 Data-independent Delay And Sum beamforming

If we consider the signal $y(n)$ to be spatially white, i.e. sources' signal arrives at the array with equal power along DOAs, then $\mathbf{\Phi}_y = \boldsymbol{I}$ and $g = 1$ we can write the minimization problem as

$$\min_h \boldsymbol{h}^H \boldsymbol{h} \quad \text{subject to} \quad \boldsymbol{a}^H(\theta, \omega)\boldsymbol{h} = 1, \tag{2.23}$$

which assures the signal to be undistorted at the specific DOA θ. If $\boldsymbol{a}(\theta, \omega)$ is specified as in (2.12) the solution to (2.23) reduces to the *Delay And Sum* (DAS) beamformer equation

$$\boldsymbol{h} = \frac{1}{M} \cdot \boldsymbol{a}(\theta, \omega). \tag{2.24}$$

Intuitively, DAS beamformer aligns or equivalently adjust in phase signals at microphones coming from a DOA; this corresponds to the "delay" step attained by filtering by $\boldsymbol{h}$. Then, the outputs of the single sensors are summed obtaining a summation in phase for the signal coming from direction θ as shown in Figure 2.2, whilst the other signals impinging on the array from other DOAs are summed destructively. Data independent filters have the major advantage of being quickly computed in every scenario of SNR, since they are independent from the signal data at hand. This could represent also a major drawback since resolution decreases. However when an efficient implementation is required, this kind of filter is the most suitable. For this reason it is used in our method for plenacoustic image calculation.
Furthermore, if we consider a generic desired response g and $\mathbf{\Phi}_{\boldsymbol{y}} = \boldsymbol{I}$ the minimization problem assumes the form

$$\min_h \boldsymbol{h}^H \boldsymbol{h} \quad \text{subject to} \quad \boldsymbol{a}^H(\theta, \omega)\boldsymbol{h} = g^*, \tag{2.25}$$

whose solution is

$$\boldsymbol{h} = g^* \frac{\boldsymbol{a}(\theta, \omega)}{\boldsymbol{a}^H(\theta, \omega)\boldsymbol{a}(\theta, \omega)} \tag{2.26}$$

which can be considered a constrained *Delay And Sum* beamformer. This kind of filter is a sort of hybrid between statistically optimal filter LCMV and data independent filter DAS.

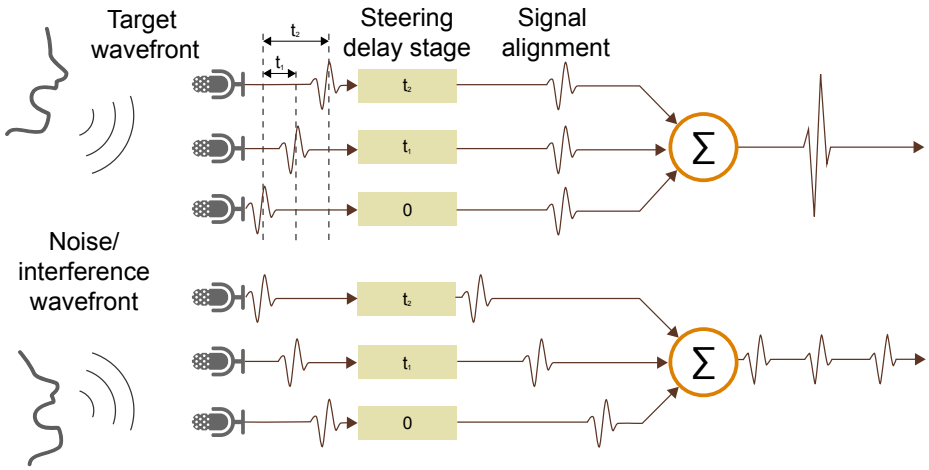


Figure 2.2. Delay And Sum beamformer with multiple active speech sources, impinging on the array from different DOAs

2.1.5 Parametric methods of beamforming

Another set of beamformers takes the name of *parametric* methods. Albeit they can be comprised in the statistically optimal beamformers set, they need a separate discussion because of the assumptions they make and the parameters these algorithms need to estimate. In (2.4) we added a noise component $e(n)$ which now we assume to be spatially white with identical variance components, hence its covariance matrix has the form $E\{e(n)e(n)^*\} = \sigma^2\boldsymbol{I}$. The signal covariance matrix is indicated with $\boldsymbol{P}$ instead, and it is assumed to be nonsingular(signals are not coherent) and uncorrelated with noise signals. Thus, the array output vector covariance matrix can be decomposed as follows

$$\boldsymbol{\Phi}_y = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^* + \sigma^2\boldsymbol{I}. \tag{2.27}$$

It is also assumed that $M > L$ so that $rank(\boldsymbol{\Phi}) = M$ while $rank(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^*) = L$. Consequently, we can determine L eigenvalues/eigenvectors associated with the signal covariance matrix and other $M-L$ ones associated with the noise components. Then, we can identify two sets of eigenvalues, their eigenvectors $\{s_1, \ldots, s_L\}$ and $\{g_1, \ldots, g_{M-L}\}$ and their relative matrices $\boldsymbol{S}$ and $\boldsymbol{G}$. It can be proved that

$$\boldsymbol{\Phi}_y\boldsymbol{G} = \sigma^2\boldsymbol{G} = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^*\boldsymbol{G} + \sigma^2\boldsymbol{G}. \tag{2.28}$$

The last equality holds because the columns of $\boldsymbol{G}$ belongs to the *null space* of $\boldsymbol{A}^*$, indicated as $\mathcal{N}(\boldsymbol{A}^*)$. Since $rank(\boldsymbol{A}) = L$, the dimension of $\mathcal{N}(\boldsymbol{A}^*)$ is equal to $M-L$

which is also the dimension of the *range space* of $\boldsymbol{G}, \mathcal{R}(\boldsymbol{G})$. By definition, we have $\boldsymbol{S}^*\boldsymbol{G} = 0$, then it leads to $\mathcal{R}(\boldsymbol{G}) = \mathcal{N}(S^*)$; hence, $\mathcal{N}(\boldsymbol{S}^*) = \mathcal{N}(\boldsymbol{A}^*)$. Since $\mathcal{R}(\boldsymbol{S})$ and $\mathcal{R}(\boldsymbol{A})$ are the orthogonal complements to $\mathcal{N}(\boldsymbol{S}^*)$ and $\mathcal{N}(\boldsymbol{A}^*)$, it follows that $\mathcal{R}(\boldsymbol{A}) = \mathcal{R}(\boldsymbol{S})$. It can be shown that $\mathcal{R}(\boldsymbol{S}) \subset \mathcal{R}(\boldsymbol{A})$. The following key result is obtained from (2.27).

$$\boldsymbol{a}^*(\theta_l, \omega)\boldsymbol{G}\boldsymbol{G}^*\boldsymbol{a}(\theta_l, \omega) = 0. \tag{2.29}$$

The *MUltiple SIgnal Classification* (MUSIC) beamformer weights are so defined to have peaks along the desired directions θ_l,

$$\boldsymbol{h} = \frac{1}{\boldsymbol{a}^*(\theta, \omega)\hat{\boldsymbol{G}}\hat{\boldsymbol{G}}^*\boldsymbol{a}(\theta, \omega)}, \tag{2.30}$$

where $\hat{\boldsymbol{G}}$ is an approximation of $\boldsymbol{G}$. An important computational complexity and statistical accuracy trade-off should be considered when this method is applied. The trade-off depends on the number of sensors employed; if $M = L + 1$ the MUSIC method is called *root* MUSIC which provides the best performance in terms of speed of calculus but also the least accuracy.
Parametric methods are not employed in the method we propose in Chapter 3 and 4 since they have high computational burden and also need noise and source signals estimations, which are not always achievable.

2.2 Source separation

The term source separation refers to the practice of extracting specific target sources from a mixture of signals captured from the sensors, and attenuating the remaining signal contributions. Source separation is widely used in numerous fields, from telecommunication systems, to speech enhancement and speech separation. In audio processing, for example, real time speaker separation for simultaneous translation, sampling of musical sounds for electronic music composition, or speech enhancement within hearing aids are possible through source separation.
As it will further discussed in this Section, there are several methods to perform source separation, one of these is through beamforming. Since the method we propose in this thesis accomplishes source separation by means of beamforming, we will discuss the state of the art of this research branch. However, it seems important to give a general look to other methods in order understand our choice of beamforming-based source separation.
Let us take a step back to observe the problem from a general point of view. The signal processing literature classifies systems that acquire certain mixtures of signals as under/over determined, instantaneous or convolutive and time varying or time invariant. As stated in the previous Section the solution to the problem strongly depends on the kind of problem to be addressed. As a first step, let us define a new model which is more general than the one stated in (2.4) in Section 2.1, since it does not assume the matrix $\boldsymbol{A}$ to be a simple delay between sensors. Thus, we drop the single far-field source in an anechoic environment assumption we can write

$$y_m(n) = \sum_{l=1}^{L} \sum_{\tau=-\infty}^{+\infty} a_{ml}(n - \tau, \tau)s_l(n - \tau), \tag{2.31}$$

where a_{ml} are the mixing filter weights, and $s_l(n)$ the signal sources. We can rearrange (2.31) in a matrix form and add a noise component, in the same way we did in (2.4),

$$\boldsymbol{y}(n) = \boldsymbol{A}\boldsymbol{s}(n) + \boldsymbol{e}(n), \tag{2.32}$$

where vector $\boldsymbol{s}(n)$ is a $L \times 1$ containing the signals values at discrete time n and $\boldsymbol{A}$ is a $M \times L$ instantaneous mixing matrix containing the mixing filter weights. Matrix $\boldsymbol{A}$ can also be seen as the collection of the impulse responses from sources $s_l(n)$, $l = 1, \dots, L$ to sensors m, $m = 1, \dots, M$. Starting from (2.32), various approaches have been devised to solve the problem. One of them is the beamforming approach seen in the previous Section 2.1.2, which exploits the intrinsic spatial information carried by the signals impinging upon the array. This approach works well with instantaneous or slightly convolutive mixtures, because, theoretically, it may not produce artifacts or signal distortions when signals arrive with a certain angle from just one direct path. In the case of convolutive mixtures, channels estimations [22], or some directivity trade-offs [3] are needed in order to extract the target signal without any contributions from other sources. In fact in reverberant environments, we can identify a direct path of the wave propagating from source $s_l(n)$ to sensor m and several secondary paths tracked by sound waves that bounce against obstacles, e.g. walls. The waves reflected by objects can be modeled as fictitious image sources that emit the wave impinging on the object towards the reflection direction. Considering all this factors the model and its solution increase in complexity. We find in the literature techniques aimed at solving this problem. The main approaches are, *Beamforming-based Source Separation*, which is the method adopted in this thesis, *Blind Source Separation* (BSS) [8], which represents data in a statistical domain, and time-frequency masking [23]. In the following paragraphs we examine the beamforming-based source separation.

2.2.1 Beamforming-based source separation

Source separation through beamforming has been successfully employed in many research works. A deep characterization of these methods is provided in the literature with several assessments of their performances in several scenarios of anechoic or reverberant environment with different ambient noise ratios. When source separation is considered, the main parameters considered in a system are the distortions introduced in the extracted signals as well as the attenuation of all the other interfering sources or/and ambient noise that is highly likely to be present. In theory, the LCMV beamformer can achieve perfect dereverberation and noise cancellation when the *acoustic transfer functions* between all sources (including interferences) and microphones are known [24].
Let us rewrite the model in the frequency domain with reference to (2.32). It is useful to derive the frequency domain version of the model because separation filters are usually computed in such domain. We can define frequency counterpart of the model in (2.4) through DFT,

$$\begin{aligned} \boldsymbol{Y}(\omega) &= \boldsymbol{A}(\omega)\boldsymbol{S}(\omega) + \boldsymbol{E}(\omega) \\ &= \boldsymbol{X}(\omega) + \boldsymbol{E}(\omega). \end{aligned} \tag{2.33}$$

where

$$\begin{aligned}
\boldsymbol{Y}(\omega) &= [\boldsymbol{y}_1(\omega)\ \boldsymbol{y}_2(\omega)\dots\ \boldsymbol{y}_M(\omega)]^T, \\
\boldsymbol{A}(\omega) &= [\boldsymbol{a}_1(\omega)\ \boldsymbol{a}_2(\omega)\dots\ \boldsymbol{a}_M(\omega)]^T, \\
\boldsymbol{a}_m(\omega) &= [a_{1,m}(\omega)\ a_{2,m}(\omega)\dots\ a_{L,m}(\omega)]^T, \\
\boldsymbol{S}(\omega) &= [\boldsymbol{s}_1(\omega)\ \boldsymbol{s}_2(\omega)\dots\ \boldsymbol{s}_L(\omega)]^T, \\
\boldsymbol{E}(\omega) &= [\boldsymbol{e}_1(\omega)\ \boldsymbol{e}_2(\omega)\dots\ \boldsymbol{e}_M(\omega)]^T, \\
\boldsymbol{X}(\omega) &= [\boldsymbol{x}_1(\omega)\ \boldsymbol{x}_2(\omega)\dots\ \boldsymbol{x}_M(\omega)]^T, \\
\boldsymbol{x}_m(\omega) &= \boldsymbol{a}_m^T(\omega)\boldsymbol{S}(\omega).
\end{aligned}$$

Using the Fourier transform relationship , the covariance matrix, firstly defined in (2.19), can be expressed as power spectral density (PSD) of the received signal at the mth sensor

$$\begin{aligned}
\phi_{y_m}(\omega) &= \phi_{x_m}(\omega) + \phi_{e_m}(\omega) \\
&= \boldsymbol{a}_m^H(\omega)\Lambda_s(\omega)\boldsymbol{a}_m(\omega) + \phi_{e_m}(\omega),
\end{aligned} \tag{2.34}$$

for $m = 1, 2, ..., M$, where $\phi_{y_m}(\omega), \phi_{x_m}, \Lambda_s(\omega) = diag[\phi_{s_1}(\omega), ..., \phi_{s_L}(\omega)]$, and $\phi_{e_m}(\omega)$ are the PSDs of the mth sensor signal, the mth sensor reverberant signal, the coherent signals, and the mth sensor noise signal, respectively. The beamforming is then performed by applying a complex weight(real weights beamformers have been conceived too [25]) to each sensor and summing across all sensors:

$$y_F(\omega) = \boldsymbol{h}^H(\omega)\boldsymbol{Y}(\omega) = \boldsymbol{h}^H(\omega)\left[\boldsymbol{X}(\omega) + \boldsymbol{E}(\omega)\right], \tag{2.35}$$

where $y_F(\omega)$ is the beamformer output, $\boldsymbol{h}(\omega) = [h_1(\omega)\ h_2(\omega)...h_M(\omega)]^T$ is the beamforming weight vector. The PSD of the beamformer output is

$$\phi_{y_F} = \boldsymbol{h}^H(\omega)\boldsymbol{\Phi}_x(\omega)\boldsymbol{h}(\omega) + \boldsymbol{h}^H(\omega)\boldsymbol{\Phi}_e(\omega)\boldsymbol{h}(\omega), \tag{2.36}$$

where

$$\boldsymbol{\Phi}_x(\omega) = E\left[\boldsymbol{X}(\omega)\boldsymbol{X}^H(\omega)\right] = \boldsymbol{A}(\omega)\Lambda_s(\omega)\boldsymbol{A}^H(\omega). \tag{2.37}$$

In (2.21) we defined the LCMV solution for filters $\boldsymbol{h}(\omega)$, in this Section instead we show how to set the L constraints $\boldsymbol{g}(\omega)$ while the remaining degrees of freedom are employed to minimize the contribution of the additive noise to the array output. It should be noted that the LCMV filter is constructed using only spatial information related to the undesired sources (given by $\boldsymbol{A}$), i.e., we do not require the PSDs of the undesired source signals. This makes the beamformer especially attractive in a scenario where the undesired sources are highly non-stationary and their spatial position is fixed or slowly time-varying. This is an essential characteristic of the LCMV beamformer since the situation just described might correspond to the scenario our system has to deal with. In fact, speech signals are non-stationary signals and we do not assume speakers position to be fixed.
As previously stated, essential parameters to evaluate source separation methods are the distortion on the desired source introduced by signal processing operations, the interference and the noise suppression. We now discuss the state-of-the-art

research for beamforming-based source separation, underlining these three aspects. In addition, particular emphasis is given to MVDR and LCMV methods, since, as previously stated, the most suitable beamformers for our system are the statistically optimal ones.

2.2.2 Constraints definition for source separation filters

Let us first determine how to properly tune the constraint set of the LCMV filter for source separation. A thorough study about the constraints matrix and response vectors definition has been carried out in [26] to find out the trade-offs that rule the LCMV beamforming performances. They derived a speech-distortion and interference-rejection constraint beamformer that is able to trade-off between speech distortion and interference-plus-noise reduction on the one hand, and undesired signal and ambient noise reductions on the other hand. In general, the aforementioned trade-offs can be realized by modifying the optimization problem in (2.20). In fact, a controlled distortion of the desired speech signal, as received by the reference microphone (for example the first microphone), can be defined by properly defining the constraint sub-matrix $\boldsymbol{C}_d, \boldsymbol{C}_u$ of the desired and undesired signals and their related response. In order to control responses solely on the desired and undesired signals, we can split matrix $\boldsymbol{S}(\omega)$ in two in (2.33), each part regarding the desired sources $\boldsymbol{S}_d(\omega)$ or the undesired $\boldsymbol{S}_u(\omega)$. We proceed in the same way for matrix $\boldsymbol{C}(\omega)$. Now the constraints are defined as

$$\boldsymbol{h}^H(\omega)\boldsymbol{C}_d(\omega) = \alpha(\omega). \tag{2.38}$$

The parameter $\alpha(\omega)$ is a complex number. The closer is the value of $\mid \alpha(\omega) \mid^2$ to one, the less the amplitude response of the desired signal is distorted. Intuitively, we force the filter to have a real response equal to 1 at those desired directions corresponding to the desired sources DOAs. Being $\alpha(\omega)$ a complex number, the phase response has to be considered as well. When the phase response of $\alpha(\omega)$ is linear the desired signal at the beamformer's output is a delayed version of the desired signal as received by the reference microphone. For other phase responses, unequal to zero, the desired signal might contain audible distortions. The same idea can be applied in order to trade-off between reduction of the undesired signal and ambient noise. Thus, we have

$$\boldsymbol{h}^H(\omega)\boldsymbol{C}_u(\omega) = \beta(\omega) \tag{2.39}$$

where $\beta(\omega)$ is a complex number. The closer the value of $\mid \beta(\omega) \mid^2$ is to zero, the more the undesired signal is reduced. Putting these constraints together we obtain

$$\boldsymbol{C}^H(\omega)\boldsymbol{h}(\omega) = \boldsymbol{g}(\omega), \tag{2.40}$$

where

$$\boldsymbol{C}(\omega) = \left[\boldsymbol{a}_d(\omega)\ \boldsymbol{a}_u(\omega)\right], \tag{2.41}$$

where $\boldsymbol{a}_d(\omega), \boldsymbol{a}_u(\omega)$ are the array transfer vectors of the desired and undesired sources. The response vector is

$$\boldsymbol{g}(\omega) = \left[\alpha(\omega)\ \beta(\omega)\right]^H. \tag{2.42}$$

A parametrized beamformer can be conceived as done in [26],

$$\boldsymbol{h}_p(\omega) = \alpha(\omega)\boldsymbol{h}_d(\omega) + \beta(\omega)\boldsymbol{h}_u(\omega), \tag{2.43}$$

where $\boldsymbol{h}_d(\omega)$ refers to the beamforming filters for the desired source(s) and $\boldsymbol{h}_u(\omega)$ refers to the undesired source(s) (a detailed explanation on how to derive this filters analytically can be found in [26]). Since the goal of LCMV filter, as originally conceived in [16], is to pass the desired source undistorted and to completely attenuate undesired/interference source, the LCMV filter is obtained from (2.43) by setting $\alpha(\omega) = 1$ and $\beta(\omega) = 0$. An estimate of the desired source $\hat{s}_d$ is then obtained. The MVDR, instead, is derived by setting $\alpha(\omega) = 1$, since we want the desired signal to pass undistorted through the beamformer filter, and $\beta(\omega)$ so to maximize the *Signal to Interference plus Noise Ratio* (SINR), so defined

$$SINR = \frac{\boldsymbol{h}^H(\omega)\boldsymbol{\Phi}_d(\omega)\boldsymbol{h}(\omega)}{\boldsymbol{h}(\omega)^H\left[\boldsymbol{\Phi}_u(\omega) + \boldsymbol{\Phi}_e(\omega)\right]\boldsymbol{h}(\omega)}, \tag{2.44}$$

where $\boldsymbol{\Phi}_d$ and $\boldsymbol{\Phi}_u$ are derived from (2.37) considering the source signals to be classified as desired and undesired we can split the PSD at sensor m as follows

$$\phi_{y_m}(\omega) = \phi_{d_m}(\omega) + \phi_{u_m}(\omega) + \phi_{e_m}(\omega). \tag{2.45}$$

In fact, the MVDR filter does not have a constraint on the interference source but only one on the desired source. The remaining degrees of freedom are employed to minimize both noise and interference. It is demonstrated that independent distortion and noise reduction control, as well as interference rejection, can be attained by setting proper values of $\alpha(\omega)$ and $\beta(\omega)$. Results in [26] in anechoic environment show that MVDR maximizes the output SINR, and in case the power of the undesired source is much larger than the power of ambient noise the performance of the LCMV and MVDR are comparable in terms of noise reduction. Total interference rejection can be achieved by LCMV filters, which generally attain better results in terms of interference suppression.

2.2.3 Noise reduction and interference rejection performances

It is important to understand noise reduction capabilities of MVDR and LCMV methods with respect to interference rejection, in order to motivate and deeply understand the benefits our approach brings in in terms of interference rejection. Two important studies about these two parameters can be found in [1] and [27]. The behavior of the MVDR and LCMV are investigated in terms of output SNR and *Signal to Interference Ratio* (SIR), which are two very common indices for speech separation assessment. It is demonstrated in these two studies that LCMV filter can achieve infinite interference suppression if the response vector presents a 0 along the DOA of the interferer. By spending a degree of freedom for controlling the beampattern we assure the system to completely attenuate signals from a specific DOA. The MVDR method is showed to perform better in terms of SNR when the source displacement $\Delta\theta$, intended as the DOA difference between desired and undesired sources, decreases beyond 15°. Physically, as the interference, whose desired response is 0, moves towards the target source, whose desired response is 1,

it becomes harder for the LCMV to satisfy two contradictory constraints. The result is that the array gain is switched from zero to one for the undesired source. This fact results in instabilities because of ill-conditioning of the constraint matrix $\boldsymbol{C}$, that translate into the appearance of sidelobes in the beampattern and displacement of the maximum attenuation far from the interference. These sidelobes lead the beamformers to capture or even enhance the white noise at sensors which spans the whole space since it does not impinge on the microphones from a specific DOA. A similar behavior is not encountered in MVDR filters since they have only one constraint and they aim at minimizing the overall noise and interference contributions. Thus, MVDR filter performances are almost not affected by sources displacement. In general we have:

$$\begin{aligned} SNR[h_{MVDR}(\omega)] &\approx SNR[h_{LCMV}(\omega)] \\ SIR[h_{MVDR}(\omega)] &\leq SIR[h_{LCMV}(\omega)], \end{aligned}$$

but if the noise power is comparable to interference power MVDR outperform LCMV having

$$\begin{aligned} SNR[h_{MVDR}(\omega)] &\geq SNR[h_{LCMV}(\omega)] \\ SIR[h_{MVDR}(\omega)] &\approx SIR[h_{LCMV}(\omega)]. \end{aligned}$$

In Figure 2.3 noise suppression performances are depicted for MVDR filters (MVDR-II filter consider the noise as spatially white) and LCMV. The global noise-suppression factor is defined in [1] as the ratio of the PSD of the original noise (sensor noise and undesired sources) at the reference microphone over the PSD of the residual noise (the remaining noise after filtering). The displacement refers to the original position $\boldsymbol{p}$ and the displaced position $\tilde{\boldsymbol{p}}$ of the undesired source. The displacement of the source is given by $\Delta = \|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|$.

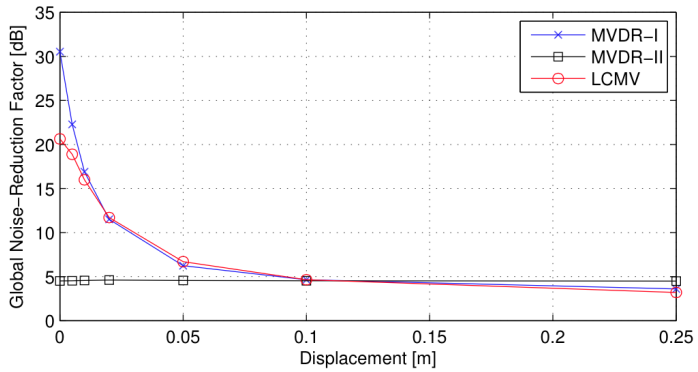
2.2.4 Source separation in reverberant environment

In reverberant environments, it is found that the amount of speech distortion obtained is higher than that one obtained in an anechoic environment and at the same time interference rejection results are poorer. This is due to wave reflections with the objects present in the environment. Attenuation of sources coming from the DOA of the direct path is attained with beamforming, but there are no constraints on the reflected undesired waves that might enter in the beam pointed towards the desired source. Thus, when beamforming is applied in reverberant environments some improvements to the system are required to accomplish robustness. The filters that we use in our model, in reverberant scenarios, are LCMV-based as proposed in [3]. The PSD defined in (2.34) can be modified to consider also the diffuse field. We recall that the diffuse field is due to late reflections and does not present a specific DOA, thus, it can be seen as a homogeneous mixture of late reflections. Hence we decompose the PSD of the signals at microphones as

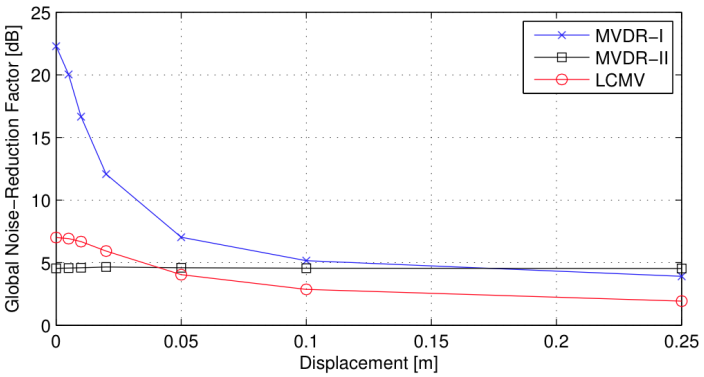
$$\boldsymbol{\Phi}_y(\omega) = \sum_{l=1}^{L} \boldsymbol{\Phi}_l(\omega) + \boldsymbol{\Phi}_d(\omega) + \boldsymbol{\Phi}_{\boldsymbol{e}}(\boldsymbol{\omega}), \tag{2.46}$$

with

$$\boldsymbol{\Phi}_d(\omega) = \phi_d(\omega)\boldsymbol{\Gamma}_d(\omega) \tag{2.47}$$



(a) Noise suppression with $SNR = -10dB$ considering the desired source as desired signal and the undesired signal as noise.



(b) Noise suppression with $SNR = 5dB$ considering the desired source as desired signal and the undesired signal as noise.

Figure 2.3. Global noise-reduction factor as a function the noise source displacement for two scenarios. Picture taken from [1].

$$\boldsymbol{\Phi}_e(\omega) = \phi_e(\omega)\boldsymbol{I}, \tag{2.48}$$

where the subscript d stands for diffuse. Although for conciseness time dependency n has been omitted in this formulation, the system is supposed to work in time frames to be responsive to changes of the sound field. Therefore, all second-order statistics refer to a time frame. The diffuse field is assumed to be spherically isotropic hence the coherence matrix $\boldsymbol{\Gamma}_d(\omega)$ has as elements $\gamma_{m_i,m_j} = sinc(\frac{\omega r_{m_i,m_j}}{c})$ and $r_{m_i,m_j} = \|d_{m_j} - d_{m_i}\|$ is the distance between microphones. The proposed informed spatial filter is obtained by minimizing the diffuse plus self-noise power subject to a constraint set, similar to (2.20). Thus, we can define the minimization problem as

$$\begin{aligned} &\min_h \boldsymbol{h}^H(\omega)\left[\boldsymbol{\Phi}_d(\omega) + \boldsymbol{\Phi}_e(\omega)\right]\boldsymbol{h}(\omega) \\ &\quad \text{subject to } \boldsymbol{h}^H(\omega)\boldsymbol{a}(\theta_l,\omega) = g, \end{aligned} \tag{2.49}$$

which corresponds to

$$\begin{aligned} &\min_{h} \boldsymbol{h}^H(\omega)\left[\Psi(\omega)\boldsymbol{\Gamma}_d(\omega)+\boldsymbol{I}\right]\boldsymbol{h}(\omega) \\ &\quad \text{subject to } \boldsymbol{h}^H(\omega)\boldsymbol{a}(\theta_l,\omega)=g, \end{aligned} \tag{2.50}$$

where $\boldsymbol{J}(\omega)=\Psi(\omega)\boldsymbol{\Gamma}_d(\omega)+\boldsymbol{I}$ and $\Psi(\omega)=\frac{\phi_d(\omega)}{\phi_e(\omega)}$ is the *Diffuse to Noise Ratio* (DNR). The minimization problem is then defined as in (2.20), thus has a similar solution to (2.21). It is important to note the beamformer so defined needs DOA estimates at each time frame to be "informed" about how to shape its beampattern. Thus, it leaves to other DOA estimation methods, like MUSIC or other beamforming methods described in Section 2.1.2 the duty of this estimate. This approach particularly suites our system because, in contrast to other methods that tries to estimates impulse responses of the direct or secondary paths between sources and microphones, or PSDs of target and undesired sources, we use a different approach that tries to assess, adaptively in time, the sound field and thus the plenacoustic function. We discuss the background theory in Section 2.3 and the method itself in Chapter 3. The remaining pending issue is how to estimate DNR. A novel estimator is developed in [3] which exhibits a sufficiently high temporal and spectral resolution to reduce both reverberation and noise. To estimate $\Psi(\omega)$ Thiergart et al. propose to use an additional spatial filter which cancels the L sources plane waves (it works with the far field assumption) such that only diffuse sound is captured. The weights of this spatial filter $\boldsymbol{h}_\Psi$ are found as in (2.26) by maximizing the white noise gain of the array. Regarding the response vector, it is set to cancel all the sources in the sound scene but pointing its beam along the DOA with the largest distance to all the other source signals DOAs. With some mathematical rearrangements we obtain

$$\Psi(\omega)=\frac{\boldsymbol{h}_\Psi^H\boldsymbol{\Phi}_y(\omega)\boldsymbol{h}_\Psi-\phi_e(\omega)\boldsymbol{h}_\Psi^H\boldsymbol{h}_\Psi}{\phi_e(\omega)\boldsymbol{h}_\Psi^H\boldsymbol{\Gamma}_d(\omega)\boldsymbol{h}_\Psi}. \tag{2.51}$$

Another important factor is the Directivity Index (DI) that represents the ratio of the total sound power in an isotropic noise filled environment, incident on an array, compared to the power actually received by the system after beamforming filtering. The DI is written as

$$DI=10\log_{10}\frac{\phi_y}{\phi_{y_F}}=10\log_{10}\frac{1}{\int_0^{\pi/2}\boldsymbol{B}(\theta,\omega)\cos(\theta)\mathrm{d}\theta}. \tag{2.52}$$

In [3] the authors show that a tradeoff between WNG and DI can be achieved. This tradeoff holds in both the case of noise power higher and lower than diffuse noise field. When noise power is higher than diffuse field power, and, viceversa, maximum directivity is achieved. This means that when noise is predominant in the sound scene, the beamformer tries to attenuate its output power with respect to all DOAs. Conversely, when diffuse field power takes over, the beamformer tries to maximize its directivity, in order to acquire the target source only along the direct path, attenuating reflected waves that might arrive from close DOAs. In order to obtain the estimated noise power required for the DNR, it is assumed that a sufficient number of silent signal frames are available. The proposed DNR estimator does not necessarily provide the lowest estimation variance in practice due to the

chosen optimization criteria, but provides unbiased results.
An alternative formulation which turns the minimization problem of the aforementioned method into a MMSE problem has been proposed in [28]. The minimization problem is equal to the one proposed in (2.17) and find its solution as seen in (2.18). It is then demonstrated that as the filter weights can be represented as a MVDR filter which is responsible for extracting the desired source undistorted, and a single-channel MMSE filter accountable for noise and interference reduction. The delicate point in the method just described is the PSDs estimation required to build this MMSE filter. The performance of the method strictly depends on the goodness of the estimations, which are not always disposable or accurate in every scenario. In fact, not only noise PSD is requested but also source signals' and undesired signals' PSDs. Under/Over estimation of these parameters leads to leakage errors, therefore, distortion of desired sources and/or insufficient suppression of undesired contributions.

2.3 Acquisition of the plenacoustic images

The *Plenacoustic Function* (PAF) [11] has been derived in order to understand and capture the spatio-temporal acoustic sound field. The concept of PAF comes from the plenoptic function introduced in [29], vastly used in computer vision and computer graphics to model light rays and their propagation. The PAF describes the acoustic radiance in every direction through every point in space. In the case of 2D geometric domain, the PAF has the form $f(x, y, \theta, \omega, n)$ with x, y indicating the position in space; θ indicating the direction; ω indicating the frequency; and n indicating time. In [11] has been conducted a thorough study on how to acquire the PAF in a discretized way, both in the free field and in a reverberant room. However, in that work the dependency on θ is not taken into account, resulting in a omnidirectional function. As explained in Section 2.1 the best way to capture signals both in time and in space is through an array of sensors disposed in a certain configuration. Using an array of sensors means placing an *Observation Window* (OW). In the 2D geometry ideal case, the OW is a line segment through which the acoustic scene is "observed". In real applications this segment is sampled by a finite array of microphones. As already stated, having a finite length array with a certain distance between sensors implies sampling limitations both in time and in space since spatial aliasing depends both on the inter-microphone spacing d, and on signal frequency ω, as showed in equation (2.9). Hence, only in the theoretical case a sound field can be totally acquired. In real cases, spatial frequency, as defined in (2.9), and temporal frequency involve aliasing problems that have to be resolved by satisfying the sampling theorems. Despite temporal frequency that may not be an issue, because of bandlimited signals or devices, spatial frequency must be taken into account to avoid aliasing and to understand the decay of the spatio-temporal sound pressure field spectrum along the spatial and temporal frequency axes. Now, being aware of the limitations in sampling a sound field, a quick and intuitive way for acquiring and visualize, hence analyze, the sound field is desirable. This is further justified by the cumbersome task of assessing a sound pressure field, generally done by gathering measurements and combining the related constraints, through a specifically developed process for the problem at hand. Moreover, we want to

restore the dependency on direction θ of the PAF, since we need estimates of DOAs of acoustic rays.
A solution to this problem was found by Markovic et al. in [2]. The authors found a way to measure the PAF by means of beamforming techniques. The authors implemented a device that captures the plenacoustic function over an OW based on an array of microphones. One rather straightforward way of doing so is to think of this device as an array of acoustic cameras that sample the OW. The unavoidable compactness of these cameras, however, causes one such device to exhibit severe resolution limitations. This means that this system cannot represent the direct acoustic counterpart of a plenoptic camera. However, in [2] a novel parameterization has been introduced for the domain of the plenacoustic function (ray space), which conveniently displays (as an image) all the elements of the acoustic scene in such a way to facilitate its analysis despite this loss of resolution. The resulting image will be here referred to as "ray-space" image, and the device for capturing it, we will call "ray-space" or "soundfield" camera. With this new parameterization, acoustic primitives such as sources and reflectors, are mapped onto rectilinear features/regions of the ray-space image, which greatly simplifies acoustic scene analysis algorithms. In fact, this allows us to approach space-time processing problems with pattern analysis tools, which are readily available in the rich literature of computer vision and multidimensional signal processing.
The plenacoustic function in [2] is thought in the 2D case as $f(x, y, \theta, \omega, n)$. In particular, they are interested in the dependency on space (x, y) and direction θ, therefore, by simplifying the notation we drop both ω and n. Under the hypothesis of validity of geometrical acoustics, expressing the soundfield as a function of the spatial/directional parameters x, y and θ, corresponds to adopting a representation based on acoustic rays.
Let us show how a compact and simple parameterization for the rays on an OW could be defined. As we are interested in defining a soundfield camera, the parameterization will be "one-sided", as it will cover only the rays that cross the OW in just one of the two possible directions. The invariance of the acoustic radiance along the direction of rays, allows us to establish an equivalence between rays and oriented lines that cross that window in the same direction. We therefore need a rule for implicitly and uniquely specifying the orientation of a line given the line parameters. Placing the observation window on the y axis between $y = -q_0$ and $y = q_0$, a relation between rays and lines can be found by writing the line equation

$$y = m_t x + q, \tag{2.53}$$

where $\mid m_t \mid < \infty$ is the angular coefficient for lines not parallel to the y axis. This line has two possible directions: one pointing towards the y axis from the positive half-space $x > 0$, and one from the opposite half-space. Conventionally, the line orientation is set towards the y axis for the x positive half-space. This allows to identify an equivalence between rays and lines, which is why the authors refer to the (m_t, q) space as the "ray-space". If the space $\mathcal{P}$ of all possible parameters (m_t, q) covers the rays that point towards the y axis from the positive half-space, the subset of such rays that only "illuminate" the OW lies within the region $\mathcal{V} = \{(m_t, q) \in \mathcal{P} : -q_0 \leq q \leq q_0\}$, which it is called "visibility region" of the OW. Given an acoustic primitive (a source, a reflector, etc.), we are interested in finding which of the "visible"

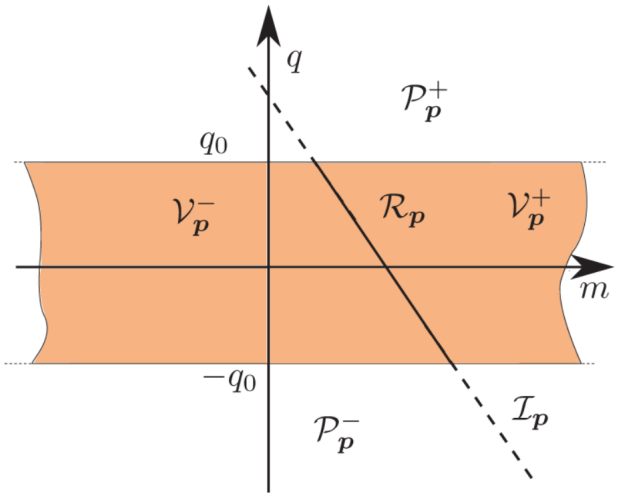


Figure 2.4. ROI $\mathcal{R}_{\overline{\boldsymbol{x}}}$ of the point $\overline{\boldsymbol{x}}$, and related regions of visibility that this ROI defines on $\mathcal{V}$. Picture taken from [2].

rays (those in $\mathcal{V}$) are coming from points of that primitive, in order to assess "what" of the radiance produced by that primitive could be picked up by the soundfield camera. This region of the ray space, referred to as the *Region Of Interest* (ROI) of the primitive, is closely related to the concept of visibility region.

In order to have a better idea of what a ray-space image is expected to look like in our case, let us begin with characterizing the ROIs of point sources. A point-like source positioned at $\boldsymbol{p} = [\overline{x}, \overline{y}]^T$ with $\overline{x} > 0$, can be equivalently thought of as the set of all the lines that pass through it. These lines identify only those rays that depart from the source and point towards the y axis. The region of the ray space describing the parameters of such lines is called the dual $\mathcal{I}_p$ of the point and is represented by the line $q = -\overline{x}m_t + \overline{y}$. The ROI of $\boldsymbol{p}$ is the set of lines that pass through both $\boldsymbol{p}$ and the OW:

$$\mathcal{R}_{\boldsymbol{p}} = \mathcal{V} \cap \mathcal{I}_{\boldsymbol{p}} = \{\boldsymbol{r} = [m_t, q]^T \in \mathcal{V} : q = -\overline{x}m_t + \overline{y}\}, \tag{2.54}$$

where $\boldsymbol{r}$ is a ray in the ray space, and $\mathcal{R}$ is the ROI of $\boldsymbol{p}$. As shown in Figure 2.4 $\mathcal{R}_{\boldsymbol{p}}$ divides $\mathcal{V}$ in the two regions $\mathcal{V}_{\boldsymbol{p}}^{+}$ and $\mathcal{V}_{\boldsymbol{p}}^{-}$. Rays in $\mathcal{V}_{\boldsymbol{p}}^{+}$ reach the OW after going around $\boldsymbol{p}$ in a clockwise fashion; while rays in $\mathcal{V}_{\boldsymbol{p}}^{-}$ fall on the OW after going around $\boldsymbol{p}$ counterclockwise. A similar definition can be given for the two half-spaces $\mathcal{P}_{\boldsymbol{p}}^{+}$ and $\mathcal{P}_{\boldsymbol{p}}^{-}$ that $\mathcal{I}_{\boldsymbol{p}}$ divides the parameter space into. Multiple sources can be managed as well, thanks to the superposition principle, as depicted in 2.5.

In order to derive the ray-space image, let us start from the classical parametrization $f(x, y, \theta)$ of the PAF and map it onto the ray space $\mathcal{P}$. This mapping is defined by $x = 0$ (the OW is on the y axis); $\theta = \arctan(m_t), -\pi/2 < \theta < \pi/2$; and $q = y$. The resulting ray-space image is therefore $p_r(m_t, q) = f(0, q, \arctan(m_t))$. This image carries information on both magnitude and phase of the acoustic radiance, therefore it is generally complex-valued. The images we need in this work of thesis are power images, such as $\boldsymbol{P}_r(m_t, q) = \mid p_r(m_t, q) \mid^2$, since we do not need the phase information but just the power distribution. The plenacoustic function in ROI $\mathcal{R}_{\boldsymbol{p}}$ can be determined using the radiance beampattern $\boldsymbol{B}_{\boldsymbol{p}}(\theta, \omega)$ of the source $\boldsymbol{p}$, which

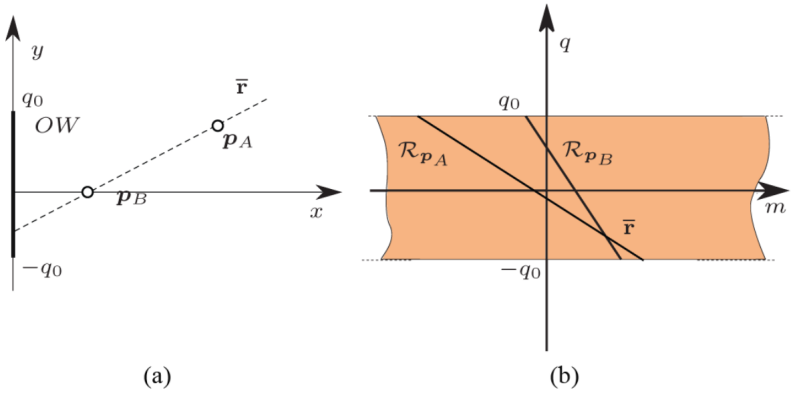


Figure 2.5. The sources $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$ in the geometric domain (a) and the corresponding ROIs (b), which generate an overlap. Picture taken from [2].

is the distribution of acoustic radiance produced by the source, as a function of the angle $\theta = \arctan(m_t)$ and frequency ω. The invariance of the acoustic radiance along the ray allows us to write

$$p_{r_{\boldsymbol{p}}}(m_t, q) = \begin{cases} \boldsymbol{B}_{\boldsymbol{p}}(\arctan(m_t)), & (m,q) \in \mathcal{R}_{\boldsymbol{p}} \\ 0, & \text{elsewhere} \end{cases}. \tag{2.55}$$

The image so calculated is idealized, i.e. no issues of limited resolution or aliasing phenomena. We notice in Figure 2.4 as the spatial extension of the OW increases, so does the thickness of the strip $\mathcal{V}$. An infinitely wide OW in fact, could ideally capture the plenacoustic function over the whole ray space $\mathcal{P}_r$.
Let us now discuss the real case when the camera is not idealized. In principle, just like in the optical domain, the soundfield camera can be thought of as an array of acoustic cameras, placed on a grid that samples the OW. If the acoustic scene is not static, we need to resort to a one-shot acquisition procedure based on a spatially extended microphone array. We discuss the ULA microphone configuration partitioned into smaller sub-arrays, since it is the one adopted in our work. One line of the ray-space image is obtained by applying beamforming to each sub-array and then mapping the output onto the ray space. For each location of the array, the angular distribution of the acoustic power is estimated through the computation of a pseudospectrum [12]. The pseudospectrum is the output power of a beamforming which points in every direction, and it shows peaks in the DOAs of the sources. We recall that this method works in the near-field hypothesis for sub-arrays but in the far field when considering the whole array. Under this condition each sub-array is able to consistently determine the directions of arrival of the sources. Thus, different sub-arrays observe the sources under different angles (i.e., from different positions). Consider the simple setup in Figure 2.6. The acoustic source is located in $\boldsymbol{p}_l$, and the microphone array is located on the y axis between $y = q_0$ and $y = -q_0$. The mth microphone in particular is in $m = [0, q_0 - 2q_0(m-1)/(M-1)]^T$. Let us consider a sub-array centered in m_i(the microphone in m_i is the reference sensor of the sub-array). The sensors in the sub-array are located at $m_i = i - \frac{W-1}{2}, \ldots, i + \frac{W-1}{2}$,

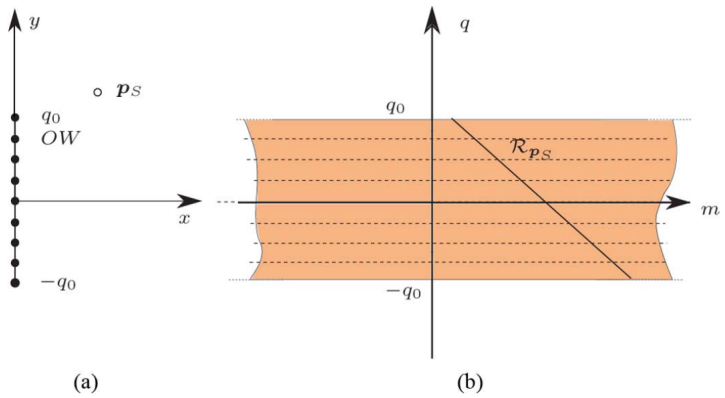


Figure 2.6. Implementation of a soundfield camera using ULA. Picture taken from [2].

where W is the odd number of microphones of the sub-array. Then we can define a beamformer at each sub-array and at each frequency ω as showed in Section 2.1.2. A wideband version of the pseudospectrum is obtained as

$$\boldsymbol{P}_i(\theta) = \Pi_{k=1}^{K} \boldsymbol{h}_i^H(\theta, \omega_k)\boldsymbol{y}(\omega_k), \quad i = \frac{W+1}{2}, \dots, M - \frac{W+1}{2}, \tag{2.56}$$

i being the index of the sub-array. Then we must map the pseudospectra onto the ray space. We recall that the pseudospectrum $\boldsymbol{P}_i(\theta)$ measures the power distribution of rays passing through the location $\boldsymbol{i}$ of the ith microphone. An acoustic ray passing through this point at an angle θ has parameters

$$\begin{aligned} m_t &= \tan(\theta) \\ q_i &= q_0 - 2q_0 \frac{i-1}{M-1}, \end{aligned} \tag{2.57}$$

therefore we can write

$$\tilde{\boldsymbol{P}}_r(m_t, q_i) = \boldsymbol{P}_i(\arctan(m_t)), \tag{2.58}$$

where $i = (W+1)/2, \dots, M - (W+1)/2$. The scanlines $q = q_i$ are the dashed ones in Figure 2.6. The real ray-space image $\tilde{\boldsymbol{P}}_r(m_t, q_i)$ that we obtain will differ from what we would obtain with an ideal camera for a twofold reason: it is sampled along q (due to limited number of sub-arrays); and it is blurred along m_t (due to limited number of sensors in each sub-array), Figure 2.7.
Starting from (2.56), in our work, we find a method to perform wideband image reconstruction that is robust to spatial aliasing artifacts and to the peculiar energy distribution of speech signals. Moreover, we wish a fast way to calculate the ray-space image so to perform frame analysis in a reasonable time. However, having the sound pressure field depicted in the ray-space image permits to draw from it, by means of linear pattern analysis, important information as source location in space. In fact exploiting the powerful Hough transformation [15] on the ray-space image one can go back to the source point location in the (x, y) space. Further, knowing the location of sources in space and other useful information drawn from the ray-space image, gives a huge hint for building source separation filters.

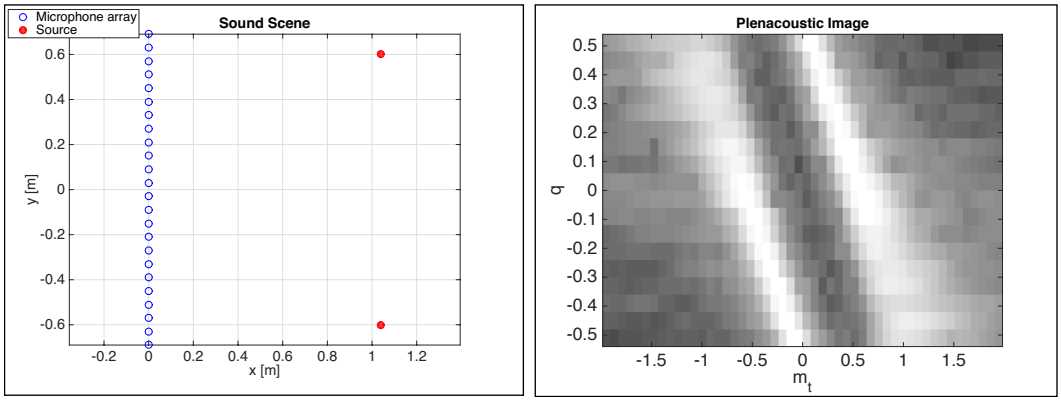


Figure 2.7. Ray-space Image for two angularly displaced sources. Speech sources are placed at distance 1 [m] impinging on the array center with DOAs 30° and −30°. The array has $M = 24$ and $W = 7$.

2.4 Source localization

Once the ray-space representation of the plenacoustic function is acquired we can exploit the information that it carries to source localization purposes. One of the biggest advantages it carries is that sources are represented by linear patterns across DOAs which can be easily clustered together. In fact, we can find local maxima at each row q_i of the ray-space image, see Figure 2.7, i.e. at each sub-array DOA estimation, and assign them to sources, trying to eliminate spurious peaks using the Hough transform [15]. The Hough transform, in fact, detects collinear local maxima and finds the parameters of the related lines, which are related to the source location. A sufficient accuracy on sources estimation would require a prohibitively large grid of the Hough map which leads also to bigger computational complexities. The Hough transform is here used only to find a first approximation of the source locations, which allows us to assign the peaks to the corresponding sources. Better estimates of the source locations can then be obtained through linear regression over measurements of the same source. Thus, we first obtain approximate coordinates $(\overline{x}_l, \overline{y}_l)$ of the L sources $\boldsymbol{p}_l, l = 1, \dots, L$ and determine the set I_l of indices that identify the ray-space image rows where the source considered is visible. Then, we define the set of source l associated maxima in each row

$$\mathcal{L}_l = \left\{ (m_{t_i}, q_i) : \frac{|m_{t_i}\overline{x}_l - \overline{y}_l + q_i|}{\sqrt{1+m_t^2}} < \epsilon, i \in I_l \right\} \tag{2.59}$$

where ϵ is a properly tuned threshold. The number of sources can be estimated through Hough transform, by simply setting a threshold on the number of associated maxima to l, considering also the number of lines present in the ray-space image. The system is then suitable for blind source localization, or, if the number of sources is known in advance localization of those sources can be performed. Now, we show how to apply least-squares technique to refine location estimates. Let us consider an acoustic source in $\boldsymbol{p}_l = [x_l, y_l]^T$, from (2.53) we know that all rays departing from $\boldsymbol{p}_l$ must satisfy the constraint $m_t x_l - y_l + q = 0$, which can be rewritten as $\boldsymbol{w}^T \boldsymbol{p}_l = -q$, where $\boldsymbol{w} = [m_t, -1]^T$. For each set of maxima $\mathcal{L}_l$ we can therefore define the system

of equations

$$\begin{cases} \boldsymbol{w}_{i_1}\boldsymbol{p}_l = -q_{i_1} \\ \vdots \\ \boldsymbol{w}_{i_{N(l)}}\boldsymbol{p}_l = -q_{i_{N(l)}} \end{cases} \tag{2.60}$$

where the subscripts $i_1, \dots, i_{N(l)}$ are the indices in I_l. Equation (2.60) can be written in the matrix form as

$$\boldsymbol{W}\boldsymbol{p}_l = \boldsymbol{q}, \tag{2.61}$$

where $\boldsymbol{W} = [\boldsymbol{w}_{i_1} \dots \boldsymbol{w}_{i_{N(l)}}]^T$ and $\boldsymbol{q} = [-q_{i_1} \dots - q_{i_{N(l)}}]^T$. We find $\boldsymbol{p}_l$ using least-squares, i.e.,

$$\hat{\boldsymbol{p}}_l = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{q}. \tag{2.62}$$

The localization procedure is repeated for all the sets $\mathcal{L}_l$. As we can see, source localization and, in particular, the problem of disambiguating measurements and matching them with sources is here turned into a pattern analysis problem performed on an image. The fact that the patterns are rectilinear, turns the localization algorithm into that of solving a system of linear equations, which is quite a desirable feature. Furthermore, it showed in [2] that for ULA configuration a limit on DOA such that $-\pi/3 \leq \theta \leq \pi/3$ is imposed for localization purposes. This is due to resolution limitations on m_t. A major outcome of this approach is the solution of localization in case of mutual occlusion of sources with respect to the array reference microphone. In fact, it has been until now an important limitation of the single array methods.

2.5 Conclusions

In this chapter we introduced the reader to the spatial filtering field providing a model adopted throughout this manuscript and explaining how it is employed in the sound field acquisition. Besides, sensor arrays allow to estimate many spatio-temporal information, e.g. source demixing, acoustic transfer functions, and to perform noise reduction and source separation through beamforming.
For these reasons, we described beamforming method in depth. We tried to give a complete and, at the same time profound, glance at the state-of-the-art of beamforming methods, underlining their performances in various ideal and real scenarios. We showed how to design the optimal beampattern for the problem at hand by means of several beamforming techniques such as DAS, LCMV, MVDR, GSC, MMSE. All these methods have been discussed showing the minimization problems they solve and their model assumptions.
In addition, we showed how beamforming is tightly linked to source separation and sound field acquisition, because of its spatio-temporal nature. In fact, beamforming is employed in many state-of-the-art source separation methods. We focused on this problem, in line with a rising interest, both in research and in commercial applications, bestowed to speech processing and separation. Giving a particular focus on the two most used methods, namely MVDR and LCMV, we showed their behavior in several circumstances, with particular attention to their interference rejection and noise suppression performances. Although, LCMV achieve better interference rejection, an important lack of robustness emerged with the LCMV

method when the angular displacement of the sources is diminished down to or less than 15°. This weakness is reflected both in SNR and in SIR, two common metrics to evaluate source separation tasks. This gap is not fulfilled neither from MVDR, even though in this case performs better, nor from other state of the art methods. In this thesis we aim at resolving this issue by approaching the problem from a plenacoustic stand point.

Using the plenacoustic function representation in the ray-space, we are able to represent in a intuitive way point sources acting in the sound scene. However, when dealing with speech sources the method presented in equation (2.56) lacks of accuracy, since it does not take into account neither the energy distribution of speech sources, nor the possible aliasing errors that might affect the pseudospectrum. Moreover, in order to perform source separation adaptively at each time frame, we need a smart algorithm to avoid burdensome computations at each sub-array. In Chapter 3 we show a way to solve these problems.

Once gained the ray-space image, in a fast and accurate way, we can easily perform linear pattern analysis, given that point sources are represented as lines in the ray space, to obtain an estimate of source locations. Source location estimation allows us to turn the blind speech separation into an informed speech separation problem, thus, allowing us to point filter beams towards the desired source. Then, knowing (x, y) coordinates of speech sources, enables us to reconstruct the targeted speech signal at any position in space by means of a appropriate sub-array signals fusion, as explained in 4. In addition, we are able to perform source separation even when angular displacement of sources diminishes beyond 15° thanks to the multiview plenacoustic approach.

Chapter 3

Efficient and Accurate computation of the Plenacoustic Image

Acquisition of a sound field is an essential step in order to study sound behavior in an enclosure or free field. If the sound energy distribution is totally known, theoretically signal processing techniques would allow to process the field both on the analysis side and on the rendering side perfectly. Due to physical limits we saw in the preceding Chapter 2 in Section 2.3 that just an acquisition approximation is achievable through microphone array processing. The coarseness of this approximation is heavily affected by physical characteristics of the device used to capture and analyze the sound field. We showed in Chapter 2 that a suitable candidate for this task is array processing with microphones. In order to acquire an adequate approximation, particular attention must be paid when waveforms in the field are wideband, which impose a large observation window to process low frequency waves and a close spatial sampling for a complete and precise (without aliasing artifacts) acquisition. This is due to the spatial frequency concept introduced in Section 2.1. If resolution limits are taken into account, microphone array methods represent a valid, versatile and efficient option for capturing sound fields and process them. This is especially true because, generally, the same task is done by gathering measurements and combining the related constraints, through a specifically developed process for the problem at hand. Specifically, we seek an efficient and accurate calculus of the sound field mapped onto the ray space. It allows us to apply linear pattern analysis to extract essential information on the field itself, like acoustic primitives positions, in a reasonable time. Thus, our work represents a first step towards a possible real time implementation of a sound field analysis for tracking purposes. No restrictions are imposed on the sources behavior, which can move in space in front of the observation window, and still be resolved in a block-adaptive way (via time frames). The information gained at this step is essential for other successive applications like, as we will see in Chapter 4, source separation. A time adaptive characterization of the sound field is essential in beamforming-based source separation techniques. In fact, through plenacoustic image analysis we propose a possible approach to transform a blind source separation into an informed beamforming separation problem. Hence, we

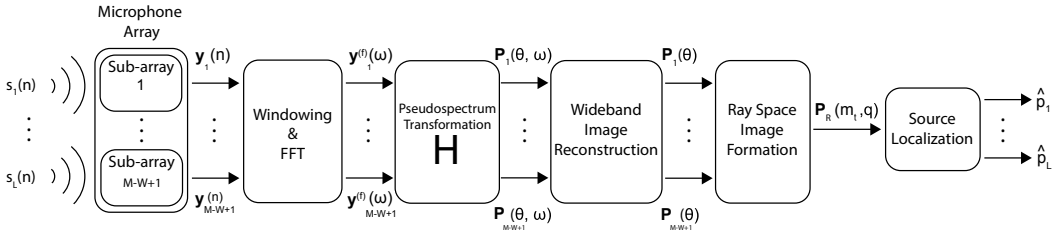


Figure 3.1. Efficient computation of the plenacoustic image and sources localization block diagram.

ward off from often needed acoustic paths transfer functions estimation or second order statistics of desired and undesired sources that are not always available and often make the system rigid for the estimation conditions they require.
We will focus on the 2D case with special attention to the ULA configuration. This will grant optimal calculus performances, as we designed a specific ULA configuration fast algorithm to compute ray-space images and because among the possible configurations the ULA utilizes the less number of sensors, and consequently it manages less amount of data, to uniformly sample space. On the other hand, ULA lacks of optimal spatial resolution. For example a quincunx configuration would guarantee better resolution. However, we compensate to breaches of the spatial sampling theorem by performing an accurate frequency weighted mean to calculate the plenacoustic image. Basing our discussion on the models we described in the preceding Chapter, we follow our dissertation in the following way: in Section 3.1 we will discuss the efficient acquisition of a plenacoustic image. We will also describe in Section 3.1.1 how to build a robust plenacoustic image in the ray space when wideband speech signals are processed with all the related issues. Successively, we apply localization methods discussed in 2.4 on the accurately calculated ray-space image to extract as much accurate estimation on sources positions.

3.1 Efficient computation of the plenacoustic image

The goal of the methodology we are going to describe consists in achieving a rapid characterization of the sound field of interest, in order to allow adaptive localization of sources in time. A block diagram of the system we are going to discuss in this Section is represented in Figure 3.1. As we affirmed in Section 2.3 a valid method is proposed in [2]. This method relies on the beamforming technique calculated by an acoustic camera (i.e. a sub-array of microphones) which samples the observation window. Such device is called plenacoustic camera. Then, a plenacoustic camera consists of multiple acoustic cameras with different array centers. The various acoustic cameras composed by W (odd) microphones are disposed along the overall linear array comprising M microphones, as described in Section 2.3. In order to have a sufficient spatial sampling frequency it is better to fix a certain overlap between acoustic cameras. We can take as reference microphone either the first sensor, or better, the central sensor of the array for symmetry reasons. In both cases the number of sampling points is equal to $M - W + 1$. This further sampling on the

observation window allows to construct both an intuitive representation of the sound field in the ray space, localization of sources, and consequently to obtain meaningful improvements in the source separation problem. We see that if the overall array extends in the range $y \in [q_0, -q_0]$, the q axis in the ray-space image is sampled at the same points $q_i = q_0 - 2q_0 \frac{i-1}{M-1}$. We indicate with i the sub-array indices.
As we previously mentioned, we seek a responsive computation of the ray-space image to extract sources location esteem. We aim at a fast computation because we want to track and separate sources even in dynamic scenarios where speakers move in space. In order to track possible moving sources, we perform a time windowing operation on signals acquired by microphones. In Figure 3.1 we indicate microphone signals frame in frequency domain as $\boldsymbol{x}_i^{(f)}$. Since our system works in signal frames, and for notational simplicity, we ignore the frame index just defined in successive mathematical derivations. Next, we discuss how to calculate the beamforming filter $\boldsymbol{h}_i(\theta, \omega)$ at each acoustic camera and the pseudospectrum $\boldsymbol{P}_i(\theta)$ that measures the power distribution of rays passing through the location of the ith microphone. Hence, we exploit different points of view corresponding to the acoustic camera centers to assess acoustic rays distribution, this way avoiding being restricted to just one sampling point in space. As we have illustrated in Section 2.1.2 there are several ways to determine a beamforming filter. Among the data-independent we derived the DAS beamformer, a well-known simple and fast method. Due to its data-independent nature, it can be computed independently from the signal data to be processed. In addition, given that we want to direct a beam at each direction in the visibility region of the observation window, we can precompute a matrix $\boldsymbol{H}$ containing all beamformer filters for the acoustic camera. This matrix is fixed and each row has a *Vandermonde* form, if the first microphone at each sub-array is taken as reference. If we take a closer look at $\boldsymbol{H}$ matrix we see that each column forms a basis function in the spatial domain, according to DOAs. Completely alike to *Fourier Transform*, that calculates frequency components of a time domain signal, we compute spatial components of a frequency domain signal, opening to fast implementation possibilities like it has been done for the time-frequency domain with the well-known *Fast Fourier Transform*.
In Section 2.1 we defined the array model we are going to use now to derive the transformation matrix $\boldsymbol{H}$. Considering that we are working on sub-arrays composed by a small number of microphones, usually between 3 and 7, thus, the sub-array extent is not reduced, we can assume a far-field scenario. Consequently, we can define matrix $\boldsymbol{A}(\omega)$ as in (2.33) for a sub-array, considering its reference microphone. Then, we consider $\boldsymbol{a}(\theta, \omega)$ as in (2.7). Then we consider DAS filter as determined in (2.24), where the number of microphones is W. Since we want to point a beam at each direction, we consider θ_j with $-\frac{\pi}{2} \leq j \leq \frac{\pi}{2}$. For simplicity, we derive the transformation matrix for the ith sub-array, but in accordance with (2.35), we can write with a little abuse of notation

$$\boldsymbol{y}_i(\omega) = \begin{bmatrix} y_i(\omega) \\ \vdots \\ y_{i+W-1}(\omega) \end{bmatrix} = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega) + \boldsymbol{E}(\omega), \tag{3.1}$$

Notice that for notation conciseness and to better visualize the *Vandermonde* structure of matrix $\boldsymbol{H}$, we are taking as reference the first microphone of each

sub-array. Thus, we can write

$$\boldsymbol{P}_i(\theta,\omega) = \begin{bmatrix} \boldsymbol{h}^H(\theta_1,\omega)\boldsymbol{y}_i(\omega) \\ \boldsymbol{h}^H(\theta_2,\omega)\boldsymbol{y}_i(\omega) \\ \vdots \\ \boldsymbol{h}^H(\theta_J,\omega)\boldsymbol{y}_i(\omega) \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}^H(\theta_1,\omega) \\ \boldsymbol{h}^H(\theta_2,\omega) \\ \vdots \\ \boldsymbol{h}^H(\theta_J,\omega) \end{bmatrix} \boldsymbol{y}_i(\omega) \tag{3.2}$$

Hence, we define:

$$\boldsymbol{H}(\theta,\omega) = \begin{bmatrix} \boldsymbol{h}^H(\theta_1,\omega) \\ \boldsymbol{h}^H(\theta_2,\omega) \\ \vdots \\ \boldsymbol{h}^H(\theta_J,\omega) \end{bmatrix} = \begin{bmatrix} 1 & e^{j\omega d\sin(\theta_1)/c} & \dots & e^{j\omega d\sin(\theta_1)(W-1)/c} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j\omega d\sin(\theta_J)/c} & \dots & e^{j\omega d\sin(\theta_J)(W-1)/c} \end{bmatrix}, \tag{3.3}$$

where j is the imaginary unit. We recall that $\boldsymbol{y}_i(\omega)$ refers to a frame of microphone signals in the frequency domain. It is also important to note that in order to obtain a uniform spacing between bases in equation (3.3), we have to choose the directions of arrival θ for which $\sin(\theta)$ is uniform. Keeping in mind that we want to calculate the plenacoustic image in a efficient way, we can rearrange microphone signals in a matrix that has as columns the ith sub-array microphone signals at frequency ω. Proceeding in this manner, we obtain

$$\boldsymbol{P}(\theta,\omega) = \begin{bmatrix} \boldsymbol{P}_1^T(\theta,\omega) \\ \vdots \\ \boldsymbol{P}_{M-W+1}^T(\theta,\omega) \end{bmatrix} = \Big(\boldsymbol{H}(\theta,\omega)\left[\boldsymbol{y}_1(\omega)\ \dots\ \boldsymbol{y}_{M-W+1}(\omega)\right]\Big)^T. \tag{3.4}$$

If we take the density of the spectral energy, $|\ \boldsymbol{P}(\theta,\omega)\ |^2$, so to discard the phase information, we obtain all the pseudospectra for frequency bin ω, obtained with just one matrix multiplication. Hence, we have to calculate $(JW)\times(M-W-1)$ multiplications and $J\times(M-W+1)\times(W-1)$ additions. Phase information is not required at this step because we just need energy distribution across DOAs in order to understand how source waves impinge on the sub-array at each frequency bin ω, as depicted in Figure 3.2. In fact, the output of filtering operation between signals at microphones and DAS beamformer previously defined, exhibit peaks at directions θ corresponding to the direction of arrival of sources, as showed in Figure 2.2.

3.1.1 Wideband image reconstruction

The pseudospectrum represents the energy distribution of the sound field with respect to DOAs and frequency [12]. In order to move towards a compact and intuitive representation of the sound field, i.e. the ray-space image, we need to perform a merging operation on frequencies. It means reducing a wideband DOA estimation problem into a frequency independent problem. From the wideband information, we want to extract precise DOAs values at which acoustic rays arrive at the sub-array. The method must be robust with respect to aliasing errors, which occurs at frequencies beyond the spatial limit, and it must maintain energy ratios between sources to better characterize the sound field. Once we gained this information we proceed to map it onto the (m_t, q) space, namely the ray space as we showed in

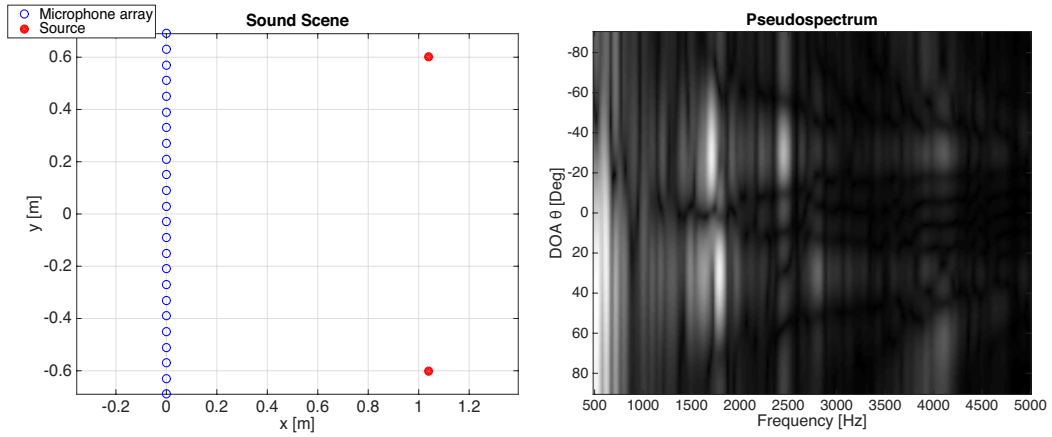


Figure 3.2. Sound scene and relative pseudospectrum calculated between 500-5000 [Hz] at 9th sub-array. Speech sources placed at 30° and −30° with respect to the array center and the array has $M = 24$ and $W = 7$.

Section 2.3.
We term the procedure of reducing frequency dimension as *wideband image reconstruction*. This procedure is important when signals are wideband (speech signals in our case) and a unique and precise estimation of DOAs is required. We need to calculate the frequency-dependent pseudospectrum before extrapolating frequency independent DOAs, because each frequency carries different information and energy contribution. In fact, it can be seen in Figure 3.2 that most of the energy is carried at low frequencies by the pseudospectrum, presenting large and smeared peaks, while precise peaks are exposed at high frequencies. This operation has been carried out in the literature [2], by means of different average methods. In [30] the authors propose to use a *geometric times harmonic mean* across frequencies in comparison with other averaging methods like *geometric mean* adopted in [2], *harmonic mean* and *arithmetic mean*. It can be easily seen, from the results proposed in [30], that the width of the main lobe of the geometric times harmonic mean is much narrower than those of the other methods and has no noticeable side-lobe structure. The other two methods, i.e. harmonic mean and arithmetic mean, on the other hand, do exhibit small side-lobe structures. Since the geometric mean is based upon the product operation, the lower frequencies eliminate any side-lobes, while the higher frequencies narrow the beamwidth and hence give better resolution. However, using these rather simple averaging methods does not always fit to the problem under analysis. The necessity of treating frequencies equally, might be required because of resolution issues and different values of SNR at each frequency. In fact, averaging effectiveness decreases when the SNR at each frequency bin varies, since the DOA estimate at some frequencies may be affected by large errors, and the final frequency data combination may be inaccurate. In [31] the authors underline this issue, proposing an averaging method called *Normalized Arithmetic Mean*. It aims at mitigating the effect of incorrect response power estimation due to the variations of the SNR at each frequency. The goal is to obtain a power pseudospectrum in which each frequency gives the same contribution to the final result. This is achieved by implementing a normalization on power pseudospectrum values across DOAs and at each frequency

bin. Thus, the power values are restricted to be in the range $[0, 1]$. In fact, a generic high-value element $P(\theta_j, \omega)$ affected by an estimation error, has a larger impact on arithmetic mean than on geometric mean, due to the natural logarithm in the latter. On the other hand, a correctly estimated low-value element has a marginal impact on arithmetic mean, while it provides a substantial negative contribution to geometric mean, which is null if $P(\theta_j, \omega) = 0$. Normalized arithmetic mean has the advantage of balancing the contribution of the two cases, thus increasing in general the robustness of the fusion in terms of peaks detection, when the values of the power response span different numerical ranges.
Such method cannot be applied in our case for many reasons. First of all, we are working with speech signals that exhibit a peculiar behavior in frequency due to speech formants. Hence, rather than flattening all the contributions in the range $[0, 1]$ we compress data values in a wider range applying a properly tuned power law transformation

$$P_c(\theta_l, \omega_k) = P(\theta_l, \omega_k)^{\gamma}. \tag{3.5}$$

Tuning of γ can be done according to the variance of $\boldsymbol{P}(\theta, \omega)$. The different energy contributions at different frequencies are so taken into account to maintain energy ratios between the two sources almost unaltered and still be robust to SNR variations as previously described. Furthermore, we need to properly consider information at high frequencies. DOAs are accurately resolved at high frequencies, in contrast with low frequencies where high energy fundamentals formants are present, but coarse spatial resolution is achievable. This is due to the limited extension of the sub-array. In the ideal case, where no spatial aliasing is present in the pseudospectrum acquisition, no additional formulation would be required, but since we are working with a microphone array with a fixed distance d between sensors, aliasing is a concern that might alter DOAs estimation. Now, we introduce two frequency weighting methods employed in our DOA estimation. The first weight is called *Inverse Spectral Flatness* which is based on the *Spectral Flatness* (SF), firstly introduced in [14]. SF measure is employed in several applications of audio signal processing, from voice activity detection [32], to linear prediction analysis of speech [14]. It measures the "whiteness" of signals, knowing that a Gaussian-distributed, temporally white signal evinces a flat spectrum. The SF measure is defined as the geometric mean over the arithmetic mean, which applied in our case to power pseudospectrum values:

$$\mathcal{F}(\omega_k) = \frac{(\Pi_{j=1}^{J} P_c(\theta_j, \omega_k))^{1/J}}{\frac{1}{J}\sum_j P_c(\theta_j, \omega_k)}, \quad -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2} \tag{3.6}$$

where $P(\theta_j, \omega_k)$ is the power pseudospectrum at DOA j and frequency ω_k. Since our goal is to assign low weights to unresolved low frequencies we apply the inverse of the measure previously defined, so to have $\mathcal{F}^{-1}(\omega) \approx 0$ if the DOA estimation is not resolved at that frequency bin and $\mathcal{F}^{-1}(\omega) \approx 1$ if the DOA estimation shows many peaks resembling a white noise. The weights so applied have the drawback of enhancing those high frequencies beyond the spatial aliasing limit, presenting spurious peak leaking at greatest DOAs. Being aware of this issue, we consider also the norm values of frequency bins, so defined

$$\|P(\theta, \omega_k)\|_1 = \sum_{j=1}^{J} P(\theta_j, \omega_k). \tag{3.7}$$

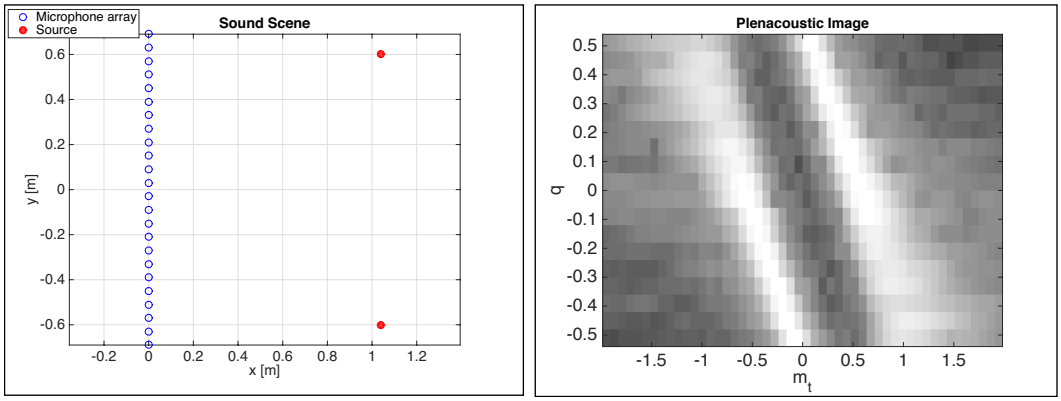


Figure 3.3. Ray-space Image for two angularly displaced sources. Speech sources are placed at distance 1 [m] impinging on the array center with DOAs 30° and −30°. The array has $M = 24$ and $W = 7$.

This measure will give us high values if that frequency carries high energy information with it, e.g. it is a formant or one of the harmonics of the voiced frame of the signal, whereas it will assume low values for high frequencies without great energy content, where aliasing errors resides. By multiplying the two weights together with the power pseudospectrum, we obtain a frequency-wise weighted pseudospectrum, where, intuitively, the frequency bins that show both an important energy content and well defined peaks are enhanced. Conversely, all the bins that do not show both characteristics are mitigated. Then, we define the weighted pseudospectrum as

$$P_{\text{sf}}(\theta_j, \omega_k) = P_c(\theta_j, \omega_k) \cdot \mathcal{F}(\omega_k)^{-1} \tag{3.8}$$

$$P_{\text{sfn}}(\theta_j, \omega_k) = P_{\text{sf}}(\theta_j, \omega_k) \cdot \|P_{\text{sf}}(\theta, \omega_k)\|. \tag{3.9}$$

The wideband DOA values estimation can now be reduced to a unique frequency independent estimation. Hence, we apply a *geometric per harmonic mean* as proposed in [30] to the weighted pseudospectrum, obtaining in this way the DOA values estimation for a sub-array of our system,

$$\boldsymbol{P}_i(\theta) = (\Pi_{k=1}^{K} P_{\text{sfn}}(\theta_j, \omega_k))^{1/K} \cdot \frac{(K)}{\sum_{k=1}^{K} \frac{1}{P_{\text{sfn}}(\theta_j, \omega_k)}}. \tag{3.10}$$

The image so computed presents sharper and more precise peaks with respect to previous methods which leads to a more accurate, consistent and robust information that can be extracted from the plenacoustic function representation. The ray-space plenacoustic image is then formed by each sub-array contribution $\boldsymbol{P}_i(\theta)$ and shows the acoustic rays distribution impinging on the array with a certain angular coefficient m_t and intercept q. Once the ray-space image is calculated, source location estimation can be carried out. Details of how to calculate such estimation have been given in Section 2.4. The location in space of the sources acting in the sound scene is of paramount importance to inform source separation LCMV filters. It is important to remark the localization technique we adopt does not need a priori information on the number of sources acting in the sound scene. Hence, the main advantage of computing a plenacoustic image for source separation purposes is that a blind

source separation problem can be turned into an informed problem. In fact, knowing the position of speakers at each time frame, which is usually approximately $0.02[s]$, gives us a precious spatial information to direct our separation filters adaptively in time. Furthermore, we can easily calculate acoustic rays distribution according to DOAs in order to determine the angular displacement of sources from the sub-array point of view (i.e. the sub-array center). This parameter will be used to coherently merge sub-array separated signal versions. We will discuss this particular weighting method in Chapter 4.

3.2 Conclusions

The key for a good source separation is knowing sources position in time. We showed in this Chapter how to achieve this knowledge adopting a time frame based computation of the ray-space image. A transformation matrix has been defined to quickly calculate wideband pseudospectra at each sub-array. Then, we showed how to reduce pseudospectrum frequency dependency to obtain acoustic rays distribution at each sub-array. We proposed a new method to perform this operation. Our method has been developed to be more robust to spatial aliasing errors than previously proposed works, to better maintain energy ratios between sources in ray-space representation. Finally, it provides more precise estimation of rays distribution along DOAs which compose each row of the ray-space image. Thus, ray-space image is formed by combining sub-array results. Thanks to this soundfield representation, we are able to easily assess sources position. As seen in Section 2.4, localization procedure is robust to sources overlapping disposition with respect to the OW. This allows us to consistently inform our LCMV separation filters.
In addition, we can intuitively see on the ray-space image sources disposition in space and calculate their acoustic ray direction difference. This parameter is derived in details in the next Chapter since it plays an essential role in our source separation method. It permits to coherently weight the separated signals estimation given by sub-arrays. The goal of this operation is to enhance those signals provided by sub-arrays from which sources are seen with a larger angular displacement. In fact, the array extension allows to have many sub-arrays with centers located at different positions along the OW which permits a multiview approach to source separation. In Chapter 4 speech separation is discussed, showing how source separation is possible with a single microphone array even in source overlap situations.

Chapter 4

Robust Speech Separation based on the Plenacoustic Image

Source separation problem has gained much interest lately, due to its numerous possible applications. Various approaches to solve the problem have been proposed and many possible ways have been explored which brought to significant results. Nevertheless, modeling the problem mathematically is not straightforward, especially when several sources interact with each other and with the environment. Reflections, ambient noise, tracking, occlusions and other related issues must be taken into account. In order to have a flexible system, we devised a system robust to all the problems mentioned. In fact, it is robust to reverberant environments where reflections of sound waves might affect the system performances; it is robust to ambient noise since it maximizes the white noise gain of separation filters in situations of high ambient noise; it is capable of tracking sources positions adaptively in time, thanks to the information provided by the plenacoustic function representation in the ray space, computed in Chapter 3. Finally, it resolves sources mutual occlusion with respect to the OW thanks to the plenacoustic approach to source separation. The latter issue is of particular interest in this Chapter explaining how such result is achievable with just one extended microphone array. Sources overlap/occlusion problem has recently been resolved in literature with distributed arrays. The drawback of this method is that it requires to deploy in space more than one array of sensors, placed in specific positions, in order to acquire signals having different DOAs in any possible source position configuration.
Herein, we propose a method to achieve satisfactory speech separation results in any case of sources disposition in space with just one extended ULA. Intuitively, we are able to place a virtual microphone close to each speaker, even though we use just one microphone array relatively distant from the speaker itself. We discuss our plenacoustic-image-based approach which relies on the plenacoustic camera that samples the observation window and performs source separation at each sampling point. A single sampling point of the OW could be seen as a ”point of view” through which the sound scene is observed. Then separation operation is done by LCMV filters at each ”point of view”. We know from Section 2.2.2 that LCMV beamforming

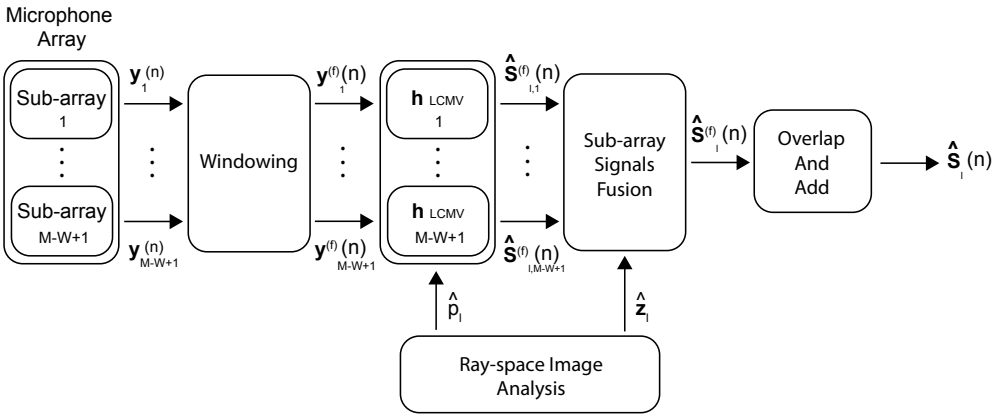


Figure 4.1. Block diagram of speech separation based on plenacoustic image

method needs constraints to be defined in order to extract the desired source(s) and attenuate the interferer(s). Once gained an estimate of sources position, as we saw in Chapter 3, and 2, we are able to inform our plenacoustic camera at each sub-array, thus, to properly set the constraints of each sub-array LCMV filter. It has been shown in Section 2.2.3 that LCMV filters suffers of instabilities and exhibit poor separation performance when the angular displacement of the desired and undesired sources, with respect to the sub-array center, is small. For this reason, we need to appropriately weight and merge sub-array reconstructed signals. A definitive evaluation of speech signal is therefore obtained and it represents an approximation the speech signal as if it was picked up at speaker's position.
Finally, in order to gain robustness against every possible real scenario, separation filters have been designed to work both in anechoic conditions and in reverberant environments.
In Chapter 5, we evaluate the performance of our speech separation method in terms of accuracy of separation.

4.1 Informed and robust speech separation filters

A consistent and time responsive estimate of the source location is a desired feature to perform source separation. Indeed, knowing speakers position in time, so their movements in space, allows us to set our separation spatial filters appropriately. As described in Section 3.1.1, a block-adaptive processing is adopted in our method. This technique permits to exploit time-variant parameters extracted at the previous processing step in Chapter 3. One of these parameters is the number of active speakers at each time frame. It has been shown how this number L can be drawn from ray-space image analysis. Thoroughly alike to ray-space image computation algorithm, we proceed in separating sources with a plenacoustic methodology, thus considering the plenacoustic camera and the sub-arrays that compose it. In Figure 4.1 a block diagram of the algorithm adopted for source separation is presented.
Let us consider a scenario in which L speakers are present in a time frame, whose position has been estimated $\hat{\boldsymbol{p}}_l = [\overline{x}_l, \overline{y}_l], l = 1, \ldots, L$. The number of speakers is

influent in terms of estimation accuracy, because, as we stated in 2.1.2 and 2.2.1, we need to have $L < W$ where W is the number of sensors our sub-array is made up to attain acceptable estimation results and avoiding ill conditioning of constraint matrix. We can adaptively modify the number W of sensors in each sub-array to satisfy this condition. The ray-space image is a precious resource at this stage, since we can easily estimate the number of active speakers from the number of lines, thus, the number of linear patterns showed in the image. By augmenting W we reduce the OW and consequently the precision on localization, since the Hough transform grid is reduced as well. Another important outcome of increasing the number W is

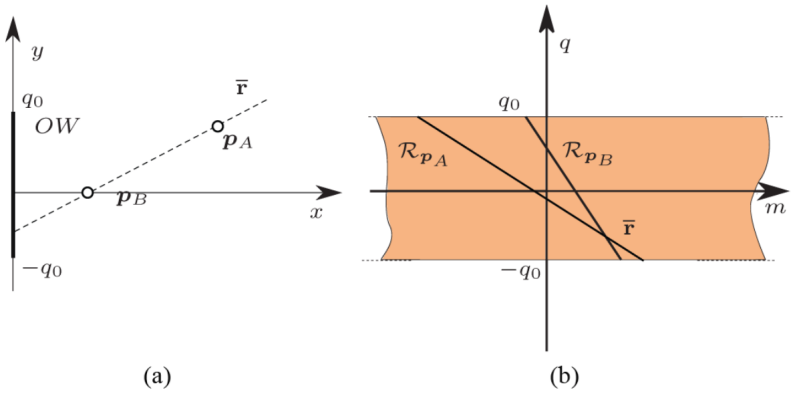


Figure 4.2. The sources $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$ in the geometric domain (a) and the corresponding ROIs (b), which generate an overlap. Picture taken from [2].

that the near-field assumption imposed at sub-arrays might not be verified anymore. This situation is showed in Figure 4.3. Once again the ray-space image comes in help and allows for fast and intuitive assessments of the sound field scenario. In fact, when two sources are in overlap with respect to a sub-array center, a peculiar disposition of rays is experienced as shown in Figure 4.2.
Next, let us assume the sensor number W has been properly tuned for the scenario at the frame under analysis, we proceed in specifying the separation filters. As already mentioned, we choose to comply with the LCMV design. This choice has a twofold reason: LCMV attains superior interference rejection among statistically optimal filters [27], it is flexible to data-independent (2.26) and robust data-dependent implementations (2.50) for reverberant environments. An important issue in LCMV filters is the constraint set $\boldsymbol{C}$ and response vector $\boldsymbol{g}$ definition as seen in Section 2.2.2. If we consider to constrain the direct path from the speaker l to the sensors of the sub-array i, $m = [i - \frac{W-1}{2}, \dots, i + \frac{W+1}{2}]$, matrix $\boldsymbol{C}_{l,i}$ will assume the form

$$\boldsymbol{C}_{l,i}(\theta,\omega) = [\boldsymbol{a}(\theta,\omega)_{l,1} \quad \dots \quad \boldsymbol{a}(\theta,\omega)_{l,W}]\,, \tag{4.1}$$

where $\boldsymbol{a}_{l,i}(\theta,\omega)$ has the form of $\boldsymbol{a}(\theta,\omega)$ that has been specified in (2.12) where the first microphone of the array was taken as reference. Nevertheless, its form can be easily derived also in the case when the central microphone is taken as reference. The constraint matrix specified in (4.1) has to be determined for each speaker $l = 1, \dots, L$. Then, we consider $\boldsymbol{a}(\theta,\omega)_{l,i}$ as the array transfer vector for the ith

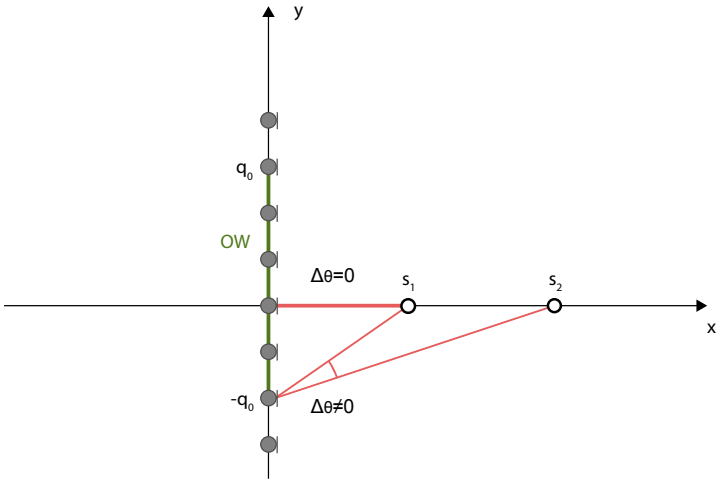
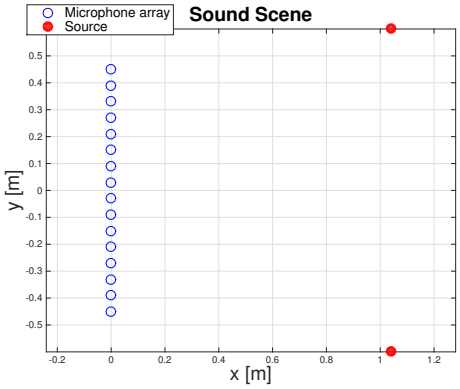


Figure 4.3. Angular displacement at different sub-array centers considering an ULA configuration with $M = 8$ and $W = 3$.

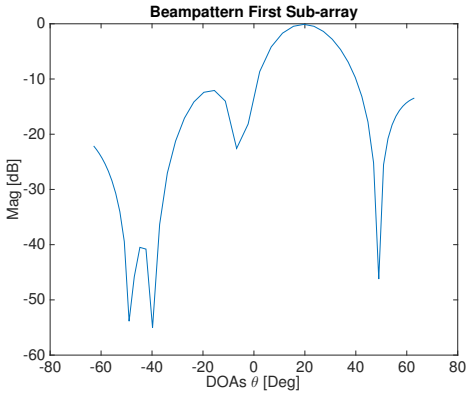
sub-array, for the source l, that has a DOA equal to θ_l. We recall that the DOAs θ_l with $l = 1, \ldots, L$ of the speech sources have been estimated from the ray-space image analysis as described in Chapter 3. Then the response vector is set according to the desired source that we want to extract. In fact, our method focuses its separation filters on one speaker at a time, aiming at extracting from the mixture acquired at microphones, one desired speaker and attenuating all the others. The practice of setting the separation filters constraint matrix and its response vector, according to a previous estimation of sources position, is called "informing" separation filters. Repeating the whole procedure described for all the speakers acting in the sound scene at each time frame enables us to obtain an estimation for each speech source. Consequently, we set the response vector

$$\boldsymbol{g}_l = [0 \ldots 1 \ldots 0]^T . \tag{4.2}$$

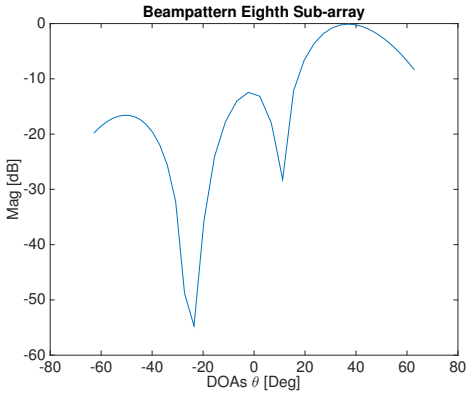
The value 1 is placed at index l corresponding to the lth desired source whereas all the other sources are set to 0. We fix this desired responses because we aim at extracting undistorted the desired speech source meanwhile maximizing the interference rejection rather than the interference plus noise attenuation. However, if a high value of noise is experienced at microphone signals, a data-dependent LCMV filter can be employed as in (2.21) so to use the remaining $W - L$ degrees of freedom to reduce noise power. Then, our separation filters for sub-array i and speech source



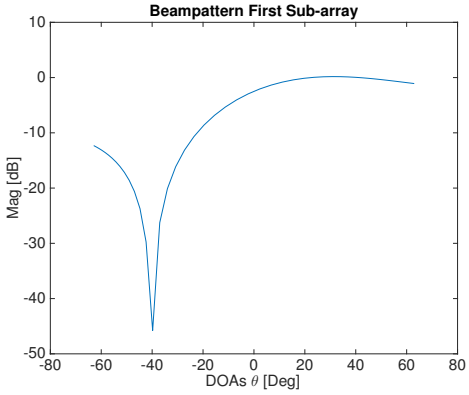
(a) Sound scene with two sources with $\Delta\theta = 60°$ with respect to the ULA center composed by $M = 16$ microphones.



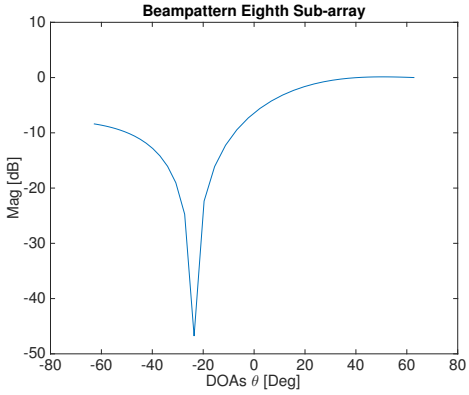
(b) First sub-array beampattern at frequency 2000 Hz.



(c) Eighth sub-array beampattern at frequency 2000 Hz.



(d) First sub-array beampattern at frequency 600 Hz.



(e) Eighth sub-array beampattern at frequency 600 Hz.

Figure 4.4. Sub-array beampatterns of LCMV separation filters targeting speech source at $[1.03, 0.6]$.

l, would become

$$\boldsymbol{h}_{l,i}(\omega_k) = \frac{\boldsymbol{\Phi}_{y_i}(\omega_k)^{-1}\boldsymbol{C}_{l,i}(\theta,\omega_k)}{\boldsymbol{C}_{l,i}^H(\theta,\omega_k)\boldsymbol{\Phi}_{y_i}^{-1}(\omega_k)\boldsymbol{C}_{l,i}(\theta,\omega_k)}\boldsymbol{g}_l. \tag{4.3}$$

Otherwise, it is convenient to use a fast data independent filter similarly to equation (2.26),

$$\boldsymbol{h}_{l,i}(\omega_k) = \frac{\boldsymbol{C}_{l,i}(\theta,\omega_k)}{\boldsymbol{C}_{l,i}^H(\theta,\omega_k)\boldsymbol{C}_{l,i}(\theta,\omega_k)}\boldsymbol{g}_l. \tag{4.4}$$

Particular attention should be paid to matrices inversions in both cases. Diagonal loading application should be considered in order to avoid instability of these filters. Instability is due to spatial aliasing experienced at frequencies higher than the spatial limit. If diagonal loading is not applied to matrix inversions, performances of the filters may rapidly degrade for increasing DOAs, as discussed in Section 2.1. An example of the beampattern obtained with this method is presented in Figure 4.4. Another LCMV separation filter design was presented in Section 2.2.4 in equation (2.50). That problem finds a similar solution to (4.3), specifically

$$\boldsymbol{h}_{l,i}(\omega_k) = \frac{\boldsymbol{J}(\omega_k)^{-1}\boldsymbol{C}_{l,i}(\theta,\omega_k)}{\boldsymbol{C}_{l,i}^H(\theta,\omega_k)\boldsymbol{J}^{-1}(\omega_k)\boldsymbol{C}_{l,i}(\theta,\omega_k)}\boldsymbol{g}_l. \tag{4.5}$$

The latter extracts the target source as the other LCMV filters but it also tries

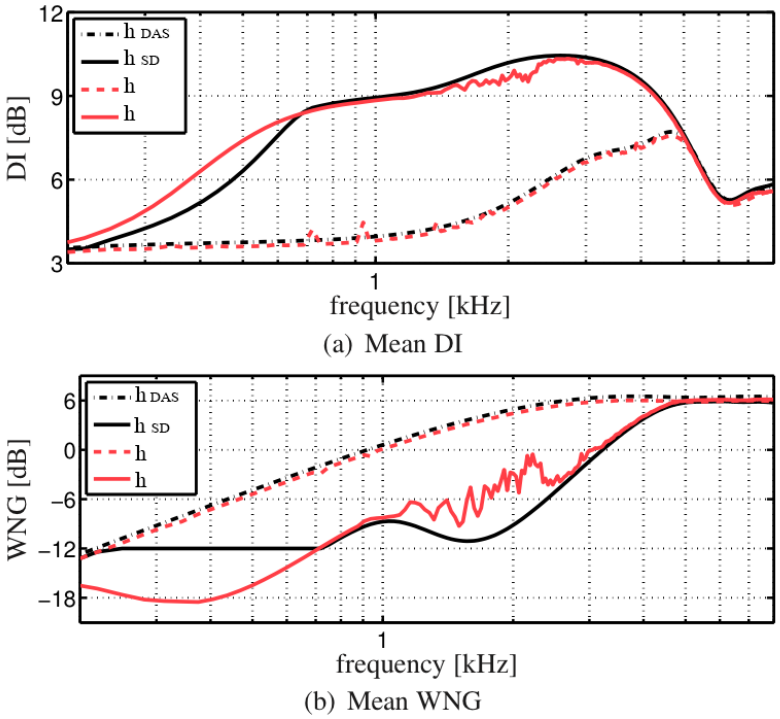


Figure 4.5. DI and WNG of the spatial filters $\boldsymbol{h}$, DAS, SD. For h_d, the minimum WNG was set to -12 dB to make the spatial filter robust against the microphone self-noise. Picture taken from [3].

to minimize the DNR, achieving better performances in reverberant environments. This is due to the behavior of the filter (4.5) which combines the benefits of two beamformer filters, DAS (2.24) and Super-Directive beamformer (SD) [33]. The first one tries to minimize the spatially white noise, while the latter tries to minimize the diffuse sound power at the filter's output. We recall that $\boldsymbol{J}(\omega) = \Psi(\omega)\boldsymbol{\Gamma}_d(\omega) + \boldsymbol{I}$

and $\Psi(\omega) = \frac{\phi_d(\omega)}{\phi_e(\omega)}$. Thus, when the DNR $\Psi(\omega)$ assumes high values, i.e. speakers are active, the directivity index (DI), defined in (2.52), is maximized; vice versa, the during silent parts, a maximum WNG is provided leading to a minimal self-noise amplification, i.e., high robustness. In Figure 4.5 this peculiar behavior is shown, where dashed lines represent filter's behavior during silence and solid lines during speech activity.
Filtering the microphone signals, with the most suitable filter $\boldsymbol{h}_{l,i}(\omega)$ for the problem at hand, at ith sub-array to extract one of the speech sources l is then performed as depicted in Figure 4.1 with the following notation:

$$\begin{aligned} \hat{s}_{l,i}^{(f)}(\omega_k) &= \boldsymbol{h}_{l,i}^H(\omega_k)\boldsymbol{y}_i^{(f)}(\omega_k), \\ &\text{with} \\ \boldsymbol{h}_{l,i}(\omega_k) &= \left[h_{i-\frac{W-1}{2}}(\omega_k), \dots, h_{i+\frac{W+1}{2}}(\omega_k)\right]^T, \\ \boldsymbol{y}_i^{(f)}(\omega_k) &= \left[y_{i-\frac{W-1}{2}}(\omega_k), \dots, y_{i+\frac{W+1}{2}}(\omega_k)\right]^T, \end{aligned} \tag{4.6}$$

where $\hat{s}_{l,i}^{(f)}(\omega_k)$ is the estimation of targeted lth speech source at frequency ω_k, and $\boldsymbol{h}_{l,i}(\omega_k)$ contains the filter weights of the separation filter for frequency bin ω_k. The equivalent operation can be performed in the time domain by convolving each sub-array sensor filter with each sub-array microphone signal and then summing contributions together. This operation is done at each sub-array, producing as output W different versions of the same time frame of the separated speech signal.

4.2 Fusion of signals extracted at sub-arrays

As it can be drawn from Figure 4.1, each sub-array filters microphone signals trying to extract the targeted source from the mixture. Since we want as final output of the system a single estimation of the targeted speech source, we need to devise a method to combine the W different versions produced at sub-arrays. Two goals are followed at this processing block: we wish we could place a virtual microphone close to the targeted speaker, and we wish we could do it for every speakers location.
As a first step, we resolve the first issue by determining a filter $\tilde{\boldsymbol{w}}(\omega)$ that aims at inverting the direct path of the wave propagation, from source l to sensor m, to compensate its attenuated and delayed version acquired at sensors. In fact, the direct path of a wave could be modeled as an attenuation of amplitude of the wave as it propagates in space. The signal acquired by microphones would then be an attenuated and delayed version of the original speech signal. Thus, we exploit the Green function [34] defined as follows,

$$\boldsymbol{w}_{l,m}(\omega) = \frac{1}{4\pi r_{l,m}} e^{-j\omega\tau_{l,m}}, \tag{4.7}$$

where $r_{l,m}$ is the distance between speaker l and sensor m and $\tau_{l,m}$ corresponds to the time delay that it requires to the wave emitted by the lth speaker to arrive at sensor m. Once $\boldsymbol{w}_{l,m}(\omega)$ has been defined we can easily find a filter $\tilde{\boldsymbol{w}}_{l,m}(\omega)$ for which it holds

$$\boldsymbol{w}_{l,m}(\omega) \cdot \tilde{\boldsymbol{w}}_{l,m}(\omega) = 1. \tag{4.8}$$

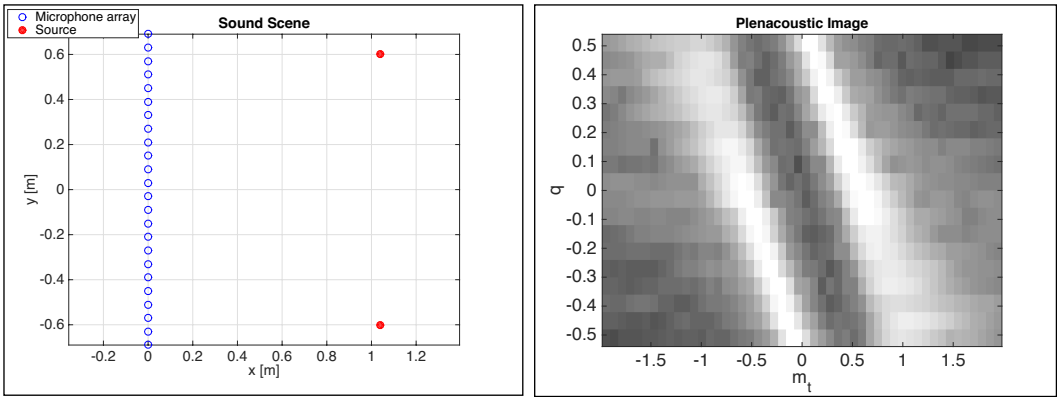


Figure 4.6. Sound scene and its relative ray-space image. Speech sources are placed in front of the array center at 0.5 [m] and 1.8 [m]. The microphone array is composed by $M = 24$ and $W = 7$.

Next, we cope with the second issue. The main idea behind the weights we are going to determine relies on the fact that, conversely to classic spatial filtering approaches, we have several points of view at our disposal each capturing the sound field. It might happen that sources are disposed in such a way that one source hides the other with respect to some of these points of view, as showed in Figure 4.3. We showed in Section 2.2.3 that LCMV separation filters of those sub-arrays (whose array center corresponds to a point of view of the sound field) for which sources are in an overlapping situation have poor performances. Then, we want to attenuate, or even discard, those contributions coming from these sub-arrays and enhance those ones for which separation filters provide satisfactory results. In order to coherently weight every contribution coming from each sub-array, we resort the ray-space image (i.e. the plenacoustic function representation in the ray space). It clearly shows overlap situations as lines intersections. In fact, sources lines arrive with the same DOA at OW sampling point q_i while they have different DOAs at other sampling points $q_{i'} \neq q_i$. The result is a line intersection in the ray-space image, showed in Figure 4.6. It can be seen that an intersection in the ray-space image corresponds to a $\Delta m_{t_i} = 0$ at ith sub-array, where $\Delta m_{t_i} = m_{t_i'} - m_{t_i''}$, $(m_{t_i'} \in \mathcal{L}_{l',i}, m_{t_i''}) \in \mathcal{L}_{l'',i}$, $m_{t_i'} \neq m_{t_i''}$ and $l' \neq l''$. The set $\mathcal{L}_{l,i}$ has been defined in Section 2.4 and it represents the set of maximum values for a row i of the ray-space image for source l. Thus, we take the index value $m_{t_i'}$ of the line corresponding to source l' and we calculate the difference, i.e. Δm_{t_i}, with the index $m_{t_i''}$ of the line corresponding to another source l''. Then, this measure is easily employed to weight sub-array contributions. Let us consider the two sources case, then the weight vector is

$$\boldsymbol{z} = \left[\Delta m_{t_1}, \ldots, \Delta m_{t_{M-W+1}}\right], \tag{4.9}$$

where $\Delta m_{t_i} = m_{t_u} - m_{t_v}$, $(m_{t_u}, m_{t_v}) \in \mathcal{L}_i$, $u \neq v$. In the general case where L speakers are active in the scene we obtain a weight matrix

$$\boldsymbol{Z} = \left[\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{L-1}\right]^T. \tag{4.10}$$

Then, we obtain the lth speaker estimation as a weighted average with respect to weight $\boldsymbol{z}_l$

$$\hat{s}_l^{(f)}(\omega_k) = \frac{\sum_{i=1}^{M-W+1} \tilde{w}_{l,i}(\omega_k)\hat{s}_{l,i}^{(f)}(\omega_k)Z_{l,i}}{\sum_{i=1}^{M-W+1} Z_{l,i}}. \tag{4.11}$$

Finally, framed signals of the lth speech source in time $\hat{s}_l^{(f)}(n)$ must be combined to obtain the entire signal. A well-known method to perform this operation in signal processing is the *overlap and add* [35]. It consists in summing together successive time frames of the processed signal with a certain overlap. The overlap rate must be chosen in accordance with the kind of window applied on the original source signal in order to have a perfect reconstruction. Usually Hanning windows with 50% overlap are employed in speech signal processing. The final output of this overlap and add, and thus of the whole system is $\hat{s}_l(n)$ that represents an approximation of the entire speech signal of one of the speakers acting in the sound scene.

4.3 Conclusions

Turning a blind source separation, i.e. no a priori knowledge is assumed on the position and the number of sources present in the scene, into an informed separation problem is an important achievement for separation efficiency. Even more appealing is knowing the exact source location in time. One outcome of locating sources in time is the possibility of virtualizing a microphone to be adaptively placed anywhere in space. If a device of this sort is applied to speech separation then it would be possible to track speakers acting in the sound scene and pick up their voice singularly. We built a robust and efficient system capable of doing what described.
In Chapter 3 we showed how to efficiently represent a sound field with ray-space images, on which localization can be easily applied by means of linear pattern analysis. Then we based our speech separation algorithm on the knowledge acquired by ray-space image analysis. Herein, we showed how to exploit a plenacoustic camera to perform informed speech separation in anechoic and reverberant environments. A major achievement, attained by using a plenacoustic camera, is the capability of separating speech sources whatever position they assume in space in front of the camera. Such result has never been achieved using only one ULA. In fact, for low values of angular displacement of sources spatial filtering fails in the attempt to separate sources. We overcome this limitation through the ray-space image that shows the ray distribution of sources with respect to the sub-arrays composing the plenacoustic camera. Then, we can weight sub-array contributions to consistently succeed in source separation. In addition, knowing the position of sources, up to an estimation error, allows us to define a filter to invert the direct path of sources wave propagation and recover the signal as it was emitted at speakers positions.
In Chapter 5 we evaluate performances and robustness of our method. We take as reference standard metrics to evaluate speech separation methods and test our system to estimation errors, angular displacement of sources and different sources overlap situations. In order to validate simulative results, we also test our method with real data acquired in real semi-anechoic and reverberant environments. Finally, we show a campaign of perceptive tests on people to explore the relation of our objective results with people opinion.

Chapter 5

Results

Speech separation has been resolved in this thesis with a new method based on the plenacoustic representation in the ray-space. The advantages this approach brings are numerous. First of all, we have discussed as source localization can be performed on the ray-space image by applying linear pattern analysis. The more the location estimation is accurate the more the overall performances of the separation system increase. The reason for this peculiar behavior resides in the spatial filtering techniques we adopted. In fact, source separation by spatial filtering requires estimates of DOAs of sources, thus, the more accurate are the DOAs the better we inform our speech separation filters. As presented in Section 2.2.1, attenuation of undesired sources can be achieved by constraining the output of the separation filter to have a specified response along certain DOAs. It was also shown, in the aforementioned Section, that LCMV filtering method outperforms all the other spatial filters in terms of interference rejection but fails when angular displacement of sources diminishes beneath a limit. In Chapter 4 we illustrated as the method we propose overcomes this limit by exploiting the plenacoustic camera advantages and the information extracted from the ray-space image. Herein, several simulations have been carried out to confirm the outstanding behavior of our method in sources overlap situations. Furthermore, a characterization of the system to angular displacement of speech sources is provided to show that our system is not significantly affected by source angular displacement. Both simulation campaigns showed impressive results demonstrating the quality of the system and the validity of the remarkable approach it embraces.
In order to validate results obtained with simulations, real data has been acquired in different reverberant environments. In particular our method has been tested in semi-anechoic and reverberant scenarios. The outstanding simulation results have been confirmed by experimental data in semi-anechoic environments whereas comparable results have been obtained in a highly reverberant environment.
The results we extracted have been given in terms of *Source to Interference Ratio* (SIR) and *Source to Distortion Ratio* (SDR), which are two metrics of a set of metrics proposed by [20] based on energy ratios. We decided to used SIR and SDR because are the most significant in our application.
However, it has been established in the literature [18] that objective metrics do not always reflect accurately perceptive evaluation of quality, intelligibility and other speech related characteristics, well known in speech enhancement [36] and speech

separation [20]. For this reason, we conducted a perceptive test evaluation campaign on 27 individuals. A relationship between objective metrics and objective *Mean Opinion Score* (MOS) has been traced, validating once again our approach to source separation with respect to classic techniques based on beamforming.

5.1 Evaluation metrics

Evaluation of separated speech signals is not a trivial topic. Several studies and evaluation campaigns have been carried out to determine the best metrics for speech separation. Often these metrics differ from classic signal processing metrics as the well-known SNR because of the different nature of the problem. A simple signal noise ratio is not enough to model all the different possible distortions that might affect a speech signal during its processing. In addition, we would like to have a measure of the separation degree attained, i.e. a measure that considers the interference rejection, since our goal is achieving an adequate speech separation result.
Vincent et al. [20] proposed a set of metrics to evaluate audio blind source separation algorithms. The authors compared the metrics they conceived with state-of-the-art measures demonstrating that these metrics better represent both objective distortion and interference phenomena, as well as subjective opinions on speech assessment. The key result resides in the way metrics are derived.
First of all, a set $\mathcal{S}$ of allowed distortions on the estimated signals has to be specified. Allowed distortions do depend on the application the system is aimed at. In our case, we consider a time-invariant gain as conceded distortion, which is the most common distortion and the less annoying for the listener. Time-invariant gain distortion just tweaks the amplitude of the waveform constantly in time but phase or frequency content remains unaltered. In order to model the main distortion phenomena in speech separation, different performance measures are computed for each estimated source $\hat{s}_l$ by comparing it to a given true source s_l. The computation of the criteria involves two successive steps. In a first step, $\hat{s}_l$ is decomposed as

$$\hat{s}_l = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}, \tag{5.1}$$

where $s_{target} = f(s_l)$ is a version of s_l modified by an allowed distortion $f \in \mathcal{S}$, and where e_{interf}, e_{noise} and e_{artif} are, respectively, the interferences, noise, and artifacts error terms. These four terms should represent the part of $\hat{s}_l$ perceived as coming from the desired source s_l, from other undesired sources $s_{l'}$ with $l' \neq l$, from sensor noises e_m, and from other causes (like forbidden distortions of the source and/or "burbling" artifacts).
In the second step energy ratios are computed to evaluate the relative amount of each of these four terms on the whole signal duration. The aforementioned decomposition is obtained by means of orthogonal projections on the subspaces spanned by original and estimated signals. Let us denote $\Pi\{s_1, \ldots, s_L\}$ the orthogonal projector onto the subspaces spanned by vectors $s_1, \ldots, s_L$. The projector is a $T \times T$ matrix where

T is the length of these vectors. Three orthogonal projectors defined as

$$\begin{aligned} Q_{s_l} &:= \Pi\{s_l\} \\ Q_{\boldsymbol{s}} &:= \Pi\{(s_{l'})_{1\leq l'\leq L}\} \\ Q_{\boldsymbol{s},\boldsymbol{e}} &:= \Pi\{(s_{l'})_{1\leq l'\leq L}, (e_m)_{1\leq m\leq M}\}, \end{aligned} \tag{5.2}$$

then $\hat{s}_l$ is decomposed in four terms

$$\begin{aligned} s_{\text{target}} &:= Q_{s_l}\hat{s}_l \\ e_{\text{interf}} &:= Q_{\boldsymbol{s}}\hat{s}_l - Q_{s_l}\hat{s}_l \\ e_{\text{noise}} &:= Q_{\boldsymbol{s},\boldsymbol{e}}\hat{s}_l - Q_{\boldsymbol{s}}\hat{s}_l \\ e_{\text{artif}} &:= \hat{s}_l - Q_{\boldsymbol{s},\boldsymbol{e}}\hat{s}_l. \end{aligned} \tag{5.3}$$

The computation of s_{target} is straightforward since it involves only a simple inner product: $s_{\text{target}} = \langle \hat{s}_l, s_l\rangle s_l/\|s_j\|^2$. The computation of e_{interf} is a bit more complex. If the sources are mutually orthogonal, then $e_{\text{interf}} = \sum_{l'\neq l}\langle \hat{s}_l, s_{j'}\rangle s_{l'}/\|s_{l'}\|^2$. Otherwise, if we use a vector $\boldsymbol{u}$ of coefficients such that $Q_{\boldsymbol{s}}\hat{s}_l = \sum_{l'=1}^{L}\overline{u}_{l'}s_{l'} = \boldsymbol{u}^H\boldsymbol{s}$ where $\overline{u}$ means the complex conjugate of u and $(\cdot)^H$ the Hermitian operator. Then, $\boldsymbol{c} = \boldsymbol{D}_{\boldsymbol{ss}}^{-1}[\langle \hat{s}_l, s_1\rangle, \ldots, \langle \hat{s}_l, s_L\rangle]^H$, where $\boldsymbol{D}_{\boldsymbol{ss}}$ is the Gram matrix of the sources defined by $(\boldsymbol{D}_{\boldsymbol{ss}})_{ll'} = \langle s_l, s_{l'}\rangle$. The computation of $Q_{\boldsymbol{s},\boldsymbol{e}}$ proceeds in a similar fashion; however, most of the time noise signals can be assumed to be mutually orthogonal and orthogonal to each source, so that $Q_{\boldsymbol{s},\boldsymbol{e}}\hat{s}_l \approx Q_{\boldsymbol{s}}\hat{s}_l + \sum_{m=1}^{M}\langle \hat{s}_l, e_m\rangle e_m/\|e_m\|^2$. Next, referring to the decomposition of $\hat{s}_l$ in (5.2) and (5.3), the energy ratios expressed in decibels are defined to provide a numerical measure to speech separation quality. Firstly, we define the source-to-distortion ratio

$$\text{SDR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}, \tag{5.4}$$

the source-to-interferences ratio

$$\text{SIR} = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}, \tag{5.5}$$

the sources-to-noise ratio

$$\text{SNR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{noise}}\|^2}, \tag{5.6}$$

and the source-to-artifacts ratio

$$\text{SAR} = 10\log_{10}\frac{\|s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\|^2}{\|e_{\text{artif}}\|^2}. \tag{5.7}$$

The four measures are inspired by the usual definition of the SNR, with some modifications. For instance, the definition of the SNR involving the term $s_{\text{target}} + e_{\text{interf}}$ at the numerator aims at making it independent of the SIR. Indeed, consider the case of an instantaneous mixture of two source, i.e. sources are linearly combined (see Section 2.2), where $\hat{s}_1 = \epsilon s_1 + s_2 + e_{\text{noise}}$ with $\|\epsilon s_1\| \ll \|s_2\|$, $\|e_{\text{noise}}\| \approx \|\epsilon s_1\|$ and ϵ an arbitrary coefficient. Then $\hat{s}_1$ is perceived as dominated by the interfering signal,

with the noise energy making an insignificant contribution. This is consistent with $SIR \approx -\infty$ and $SNR \approx +\infty$ using the definitions given before (5.4)-(5.7). An SNR defined by $10\log_{10}(\|s_{\text{target}}\|^2/\|e_{\text{noise}}\|^2)$ would give $SNR \approx 0$ instead. Similarly, the SAR is independent of the SIR and the SNR since the numerator in (5.7) includes the interferences and noise terms as well.
The measures so defined are applied throughout our simulations and experiments sessions as objective measures. In Section 5.3.2 we will also explore the relationship between these objective measures and the subjective rates that 27 individuals gave to separated signals in perceptive test session.

5.2 Simulation setup

In order to assess the validity of our method, as well as its robustness against parameter estimation errors and particular challenging speech sources dispositions, we designed specific simulation sessions. A sound scene is simulated in every session using MATLAB-r2014b software. The 2D sound scene is composed by an ULA of $M = 16$ omnidirectional microphones disposed along the y axis with center on the origin. The number of microphones M is fixed for all simulations, whereas the number of microphones forming the sub-arrays, $W(\text{odd})$ with central microphone as reference sensor, is $W = 7$ in Section 5.2.1 and in Section 5.2.2 and $W = 5$ in Section 5.2.3. This is due to the fact that a larger OW, i.e. an higher number of sub-arrays comprised in the overall array, is advantageous in situations of source overlap. We recall that our method works in time frames of signals and a overlapping positioning of sources can be easily identified on the ray-space image, thus, the number W of microphones can be modified accordingly. The distance between microphones in the array is fixed, $d = 0.06$[m], thus, giving a total array length of 0.90 [m]. Also, aliasing errors are experienced for frequencies higher than 2.858 [kHz]. In fact, from Section 2.1 we know that the condition $d < \lambda/2$ must hold to avoid spatial aliasing, consequently, with some mathematical rearrangements $f < \frac{c}{2d}$ (we indicate frequency values with f) to totally avoid spatial aliasing. The sound propagation speed c has been approximated to be constant in dry(0% humidity) air (approximately a homogeneous medium), then $c = 331.3[m/s]\sqrt{(1 + temp/273.15[K])}$, where $temp$ is the temperature expressed in celsius degrees ($^\circ C$) and the value 331.3 represent sound speed at 0°. In our case, $temp = 20$ because it approximates common scenarios temperature.
Directions of arrival are taken such that $-71.56^\circ \leq \theta \leq 71.56^\circ$ and $\tan(\theta)$ is uniformly sampled in order to have a uniform axis m_t in the ray space. A uniform axis m_t is advisable for sources localization and visualization purposes. In fact, linear regression and least-squares minimization to find a precise estimate of source locations work in linear domain (see Section 2.4).
Regarding the speech signals utilized in simulations, we used two sources a female and a male speech source taken from the dataset "EBU SQAM" *https* : *//tech.ebu.ch/publications/sqamcd*. These two speech sources have been recorded with sampling frequency $F_s = 44100$ [Hz] in an anechoic environment. We needed speech sources in anechoic environment since the simulations try to characterize and validate the system behavior in ideal conditions. Since we perform signal processing also in the frequency domain, we want to avoid temporal aliasing performing DFT in

$K = 1024$ points with $-F_s/2 \leq f_k \leq F_s/2$ and $\omega_k = 2\pi f_k$. Then, in order to avoid frequencies, either seriously affected by spatial aliasing errors, or with wavelength, λ, too large with respect to the array length, a bandpass filter is applied to signals before being processed. The cut-off frequencies of this filter are set to 500 [Hz] and 5000 [Hz], which is a reasonable choice since it manages to process most of the formants in a female and male speech, hence, without deteriorating too much the intelligibility of voice.
Further, we take time frames of 1024 samples which corresponds to approximately 0.023 [s] of the speech signal. Speech signals in such time frame can be considered quasi-stationary to obtain coherent estimations of second order statistics. Time frames are obtained by applying an Hanning window with 50% overlap, which grants perfect reconstruction after processing, when signals are singularly overlapped and added together.
We simulated propagation of waves between speakers and sensors by modeling the direct path between them. Thus, a simple filter to delay and attenuate signals has been implemented, according to Green's function (4.7). Then, ray-space image has been calculated as described in 3. Regarding speech separation filters we employed the constrained DAS showed in (4.4), giving that no sensor noise is injected in the system and no reverberations are modeled. A properly tuned diagonal loading [17], is applied to matrix $\boldsymbol{C}^H(\theta,\omega)\boldsymbol{C}(\theta,\omega)$ because it might present an unstable behavior when the aforementioned matrix is inverted, as in (4.4). This instability is due to ill conditioning of the matrix $\boldsymbol{C}^H(\theta,\omega)\boldsymbol{C}(\theta,\omega)$ that is experienced when it is not full rank, thus, when any row is obtained as linear combination of another. This phenomenon is manifested when vectors $\boldsymbol{a}(\theta,\omega)_l, \boldsymbol{a}(\theta,\omega)_{l'}$ with $l \neq l^{'}$ have the same DOA θ, which implies the two vectors to be equal. The same phenomenon might manifest also in the opposite situation, when the two DOAs are different, but, because of infringements on the spatial sampling they are mis-interpreted as equal or approximately equal.

5.2.1 Impact of source localization error on separation accuracy

As stated in Chapter 3 and in Chapter 4, no a priori knowledge on the number of sources and on their position is assumed. This kind of approach to source separation is called blind source separation. Then, performances of blind source separation methods to speech separation are affected by errors on source position estimation. This is especially true in our case, where we apply informed beamforming filters to perform separation. As we discussed in Section 4.1, we steer filter beams along the DOA of the desired source while we try to completely reject the undesired speech source. Inevitably, errors on DOA estimation leads to imperfect speech separation. Hence, it seems important to evaluate the impact of localization errors on the speech separation performances.
In order to evaluate the relation between localization errors and performance degradation, we designed a simulation session with a sufficient statistical relevance. In fact, we know that speech separation performances depend on the direction along which the error is verified, because both relative DOA and sources position, that may lead to overlap, is affected. Thus, we injected 50 realizations of a controlled

Gaussian noise on the estimated source location with random direction. We executed 25 simulation sessions with different noise variance σ^2. Specifically variance has been uniformly increased in $V = 25$ points in the range $[0, 0.0544]$, so to have $3\sigma_V = 0.6997$, where σ represents the standard deviation. The reason for that maximum value on three times the standard deviation is that we wanted to fully characterize the behavior of our system on source position estimation. In fact, an error of $r = \|\boldsymbol{p}_l - \hat{\boldsymbol{p}}_l\|^2 = 0.6997$, where $\boldsymbol{p}_l$ is the real position of the lth source and $\hat{\boldsymbol{p}}_l$ is its estimation, represents half the distance between real position of sources. Since two errors with the same variance are injected on the estimated positions, it approximately leads to overlap situations for some realizations. Therefore, sources in our simulations are placed at $\boldsymbol{p}_A = [0.7071, 07071], \boldsymbol{p}_B = [0.7071, -0.7071]$ which corresponds to an angle of 45° from the array center.

Finally, SDR and SIR are computed at the end of each simulation over the whole separated and source signals with an injected error equal to one realization value for a Gaussian noise with a certain variance σ_v^2. The procedure has been repeated for each realization, 50 in total, for each variance step, 25 in total. We consider only the SDR and SIR since they represent the two most important measures, among those defined in Section 5.1, for what concerns separation degree of speech sources and quality of the extracted signal.

In Figure 5.1a, we show the sound scene adopted, while in Figure 5.1b and 5.1c the SDR and SIR values with respect to the 25 standard deviations of the error on source localization (the standard deviation is calculated over 50 realizations of Gaussian noise with a certain variance). Source A refers to the female speech signal placed at $\boldsymbol{p}_A$, while source B refers to the male speech signal placed at $\boldsymbol{p}_B$. In Figure 5.2, a statistical analysis of the results obtained at each simulation for each error variance. The blue boxes represent the SDR or SIR values among the 25th and the 75th percentile, the red line indicates the median, the dashed line all the other values but outliers which are indicated with red crosses. As expected the boxes are progressively increasing, since the error variance is augmented at each session. We also note that the SIR values are more sensible to localization error, given the higher presence of outlier with respect to SDR.

As depicted in Figure 5.1, localization error do affect SDR and SIR final results. We can see that both curves decrease as the standard deviation error increases. This is due to the fact that we employ a beamforming method for separating sources, in particular the LCMV method which imposes desired responses on the DOA of the target source and the interferer. Thus, if the DOAs does not correspond to real DOAs of sources the LCMV fails in achieving satisfactory results.

Furthermore, we note that the initial bias between the two sources is due to the different total energy that the two signals carry (the energy ratio of source A over source B is equal to -7.11[dB]). This energy imbalance is reflected also in overall performances of the system which more easily rejects source A with respect to source B. In addition, source B is a male voice that has most of its energy towards low frequencies. Low frequencies are less directive than high frequencies because of the greater wavelengths λ they have. Thus, errors on the estimation of DOA of the undesired source with an important content of energy in low frequencies leads to energy leakage of the undesired source in the filtering operation output. This is validated by beampattern shapes of the filters we derived in Section 4.1. In Figure

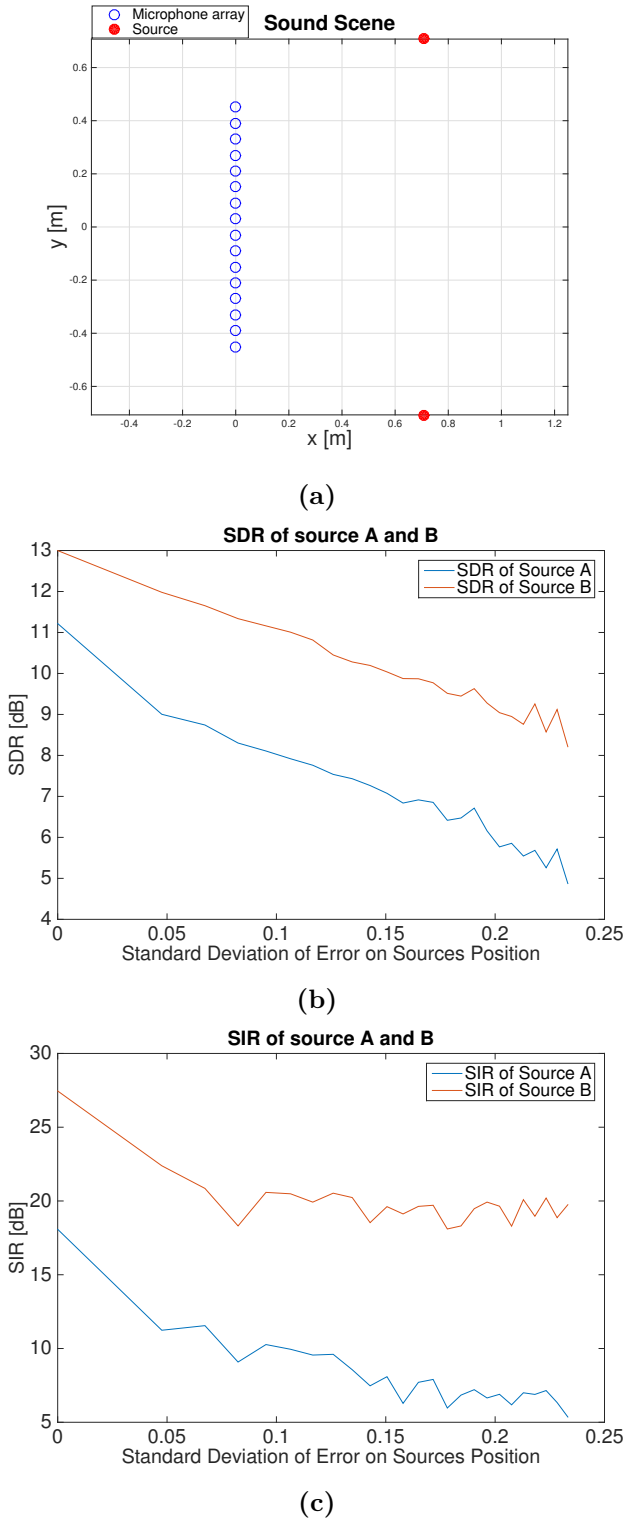


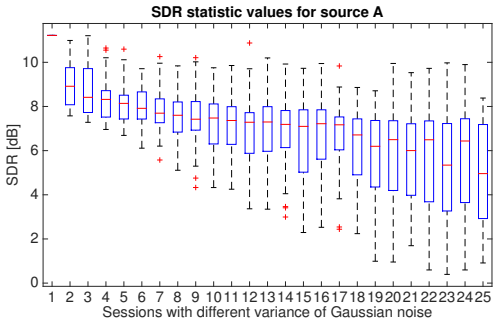
Figure 5.1. SDR and SIR metrics for two speech sources with increasing error on the estimated position of sources.

5.4 two examples of filter beampatterns exhibit a sharp notch along the estimated DOA of the undesired source while the near DOAs are barely attenuated (about 10 [dB]), leading to poor rejection if the undesired source DOA is not correctly estimated. For this reason, SIR parameter has the peculiar behavior shown in Figure 5.1c. Beyond a certain standard deviation error $\sigma \approx 0.08$[m], SIR parameter for source B could be considered stable at value the $SIR_B \approx 20$[dB]. Regarding source A, SIR parameter curve reduces its steepness gradually and can be considered stable beyond $\sigma \approx 0.15$.
SDR metrics, instead, constantly decrease as the standard deviation increases, as shown in Figure 5.1b. The reason for this trend is found in the definition of SDR in equation (5.4). SDR metric considers also spurious artifacts that appears when the estimated DOAs are almost identical.

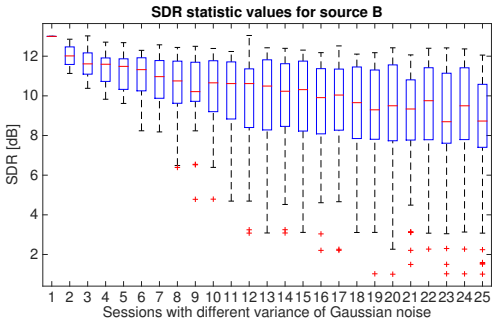
5.2.2 Separation accuracy for angularly separated sources

Source separation based on beamforming techniques, and especially LCMV method, does not show an equivalent behavior for every angular displacement $\Delta\theta = \theta_A - \theta_B$ between sources (see Section 2.2.3 and Section 4.1). Furthermore, in case that the spatial sampling condition, in (2.11), is not respected, instabilities of the separation filters for large angular displacements are experienced. We solved this problem with a properly tuned diagonal loading.
Herein, we show the results in terms of SDR and SIR we obtained from simulations of speech separation with sources placed with 25 uniformly spaced angular displacements $\Delta\theta$. Speech sources have been placed at $\boldsymbol{p}_A = [1, 0.2679]$ and at $\boldsymbol{p}_B = [1, -0.2679]$ so to have a initial $\Delta\theta$ with respect to the array center of 30°. Then, the speech sources have been progressively shifted in opposite directions along the axis $x = 1$, until they achieved a $\Delta\theta = 120°$ with respect to the array center.
In Figure 5.3 we show the behavior of the system in terms of SDR and SIR with respect to $\Delta\theta$. As expected, the more the angular displacement increases the better the sources are separated, achieving high results both in SDR and in SIR. Preventing the system from instabilities, we have made it robust to significant angular displacements. In addition, resultant metrics show a crescent trend due to the reduced effects of sidelobes in the beampattern as the difference in the DOAs of the desired and interferer sources increases.
Four examples of filters beampatterns are shown in Figure 5.4 in which we consider source A at $\boldsymbol{p}_1$ as the desired source. Figure 5.4a and 5.4b show the beampatterns of the first sub-array at $f_k = 600$[Hz] and $f_k = 2000$[Hz], when sources are positioned with $\Delta\theta = 30°$, Figure 5.4c and 5.4d show the same sub-array beampatterns for the case of $\Delta\theta = 90°$. The beampatterns exhibit a blunt peak towards the desired DOA, whereas the undesired DOA is attenuated with a precise notch. If the undesired is close to the desired source, the notch is placed in correspondence of a main sidelobe, which could cause energy leakage, due to precision limitation of the frequency bin and thus on the filter length (our filters are defined in the frequency domain). When the sources are angularly spaced enough, the undesired DOA is far from the main lobe and the area near it is approximately attenuated as well.
Furthermore, we see that the filters better attenuate low frequencies when the sources are positioned with large $\Delta\theta$. This characteristic behavior could be reconducted to

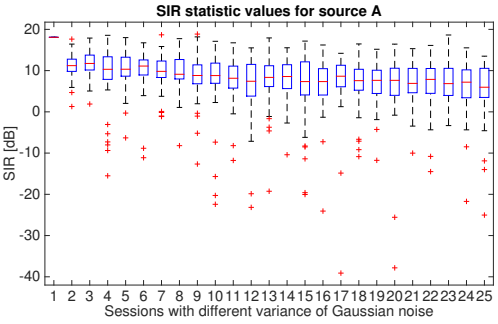
the fact that the more the angular displacement increases the higher is the diagonal loading injected in the constraint matrix at high frequencies to contrast the higher ill conditioning due to spatial aliasing. In accordance to the spatial frequency definition in equation (2.9), the more the signal frequency increases beyond the spatial aliasing condition the more is the mis-interpretation of the DOA. The outcome is a coarser interferer rejection. We recall that source B is a male voice and it contains most of its energy in the low frequencies, while source A is a female voice which contains most of its energy in the high frequencies. If we analyze Figure 5.4, we can see that high frequencies are better rejected when sources have lower $\Delta\theta$, while the opposite behavior is experienced when $\Delta\theta$ is large. However, we can also notice that the curves are almost stable with a difference of initial and final value of SDR and SIR of about 4 [dB] for source A, and 2 [dB] of SDR for source B. Consequently, we can consider the separation performance almost independent to angular displacement.



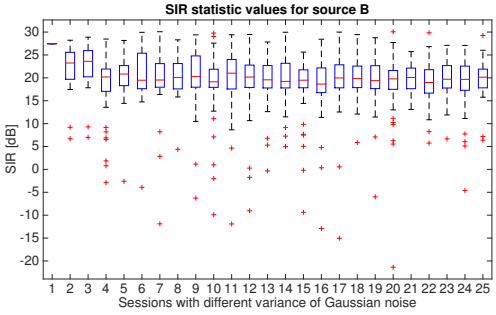
(a) Statistical analysis of SDR values for source A.



(b) Statistical analysis of SDR values for source B.

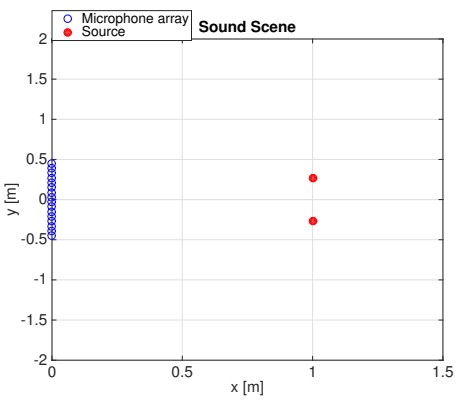


(c) Statistical analysis of SIR values for source A.

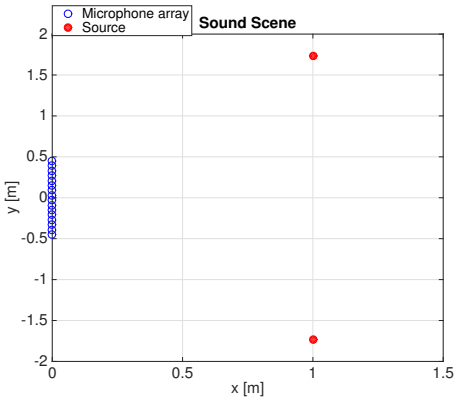


(d) Statistical analysis of SIR values for source B.

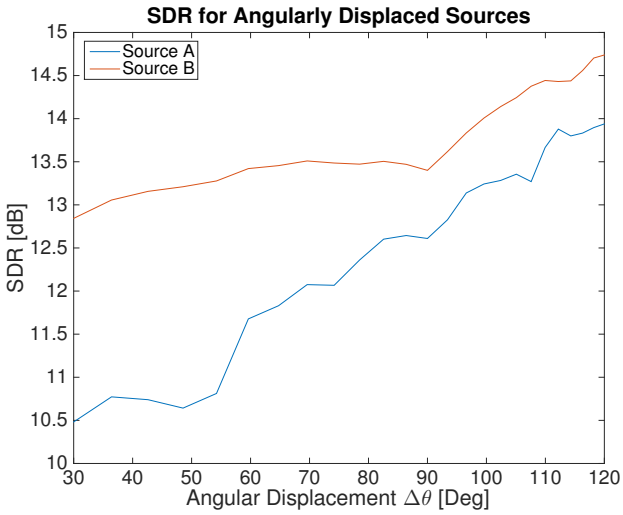
Figure 5.2. Statistical analysis of SDR and SIR metrics computed on sources A and B using the proposed method. The red bar indicated the median, the blue box comprise the all the values between the 25th and the 75th percentile, the black dashed line indicates the other values but the outliers which are indicated with red crosses.



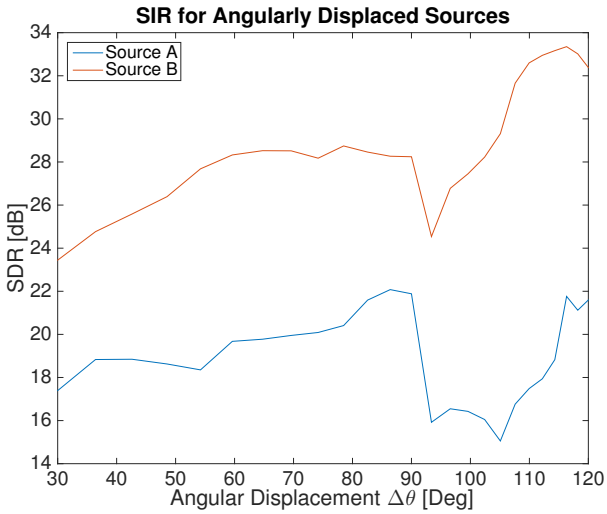
(a) Initial sound scene.



(b) Final sound scene.

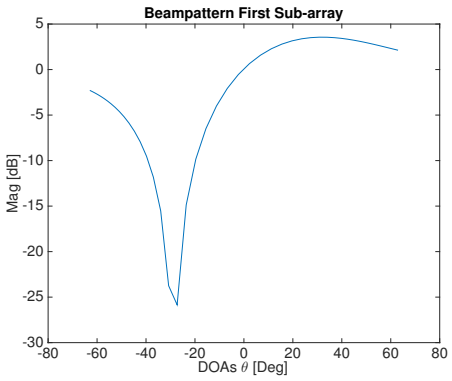


(c)

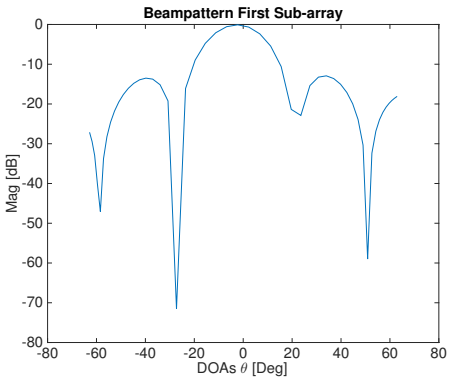


(d)

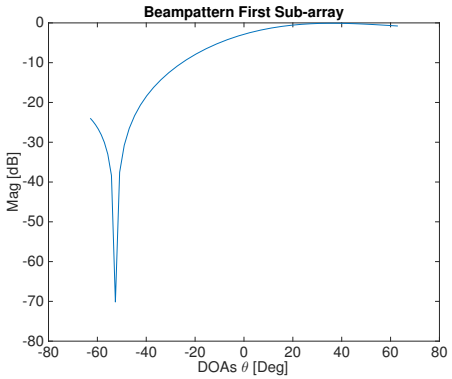
Figure 5.3. SDR and SIR metrics with respect to $\Delta\theta$.



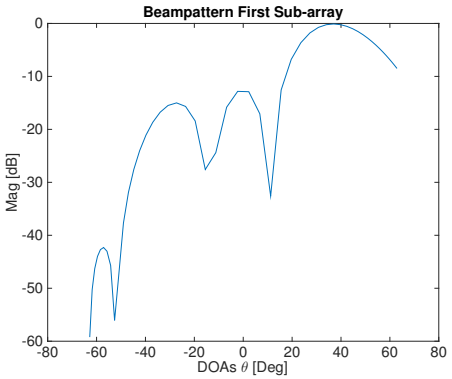
(a) Beampattern of the first sub-array at frequency 600 Hz for $\Delta\theta = 30°$



(b) Beampattern of the first sub-array at frequency 2000 Hz for $\Delta\theta = 30°$



(c) Beampattern of the first sub-array at frequency 600 Hz for $\Delta\theta = 90°$



(d) Beampattern of the first sub-array at frequency 2000 Hz for $\Delta\theta = 90°$

Figure 5.4. Beampatterns of separation filters for two different angular displacements $\Delta\theta$ of sources

5.2.3 Separation accuracy for overlapped sources

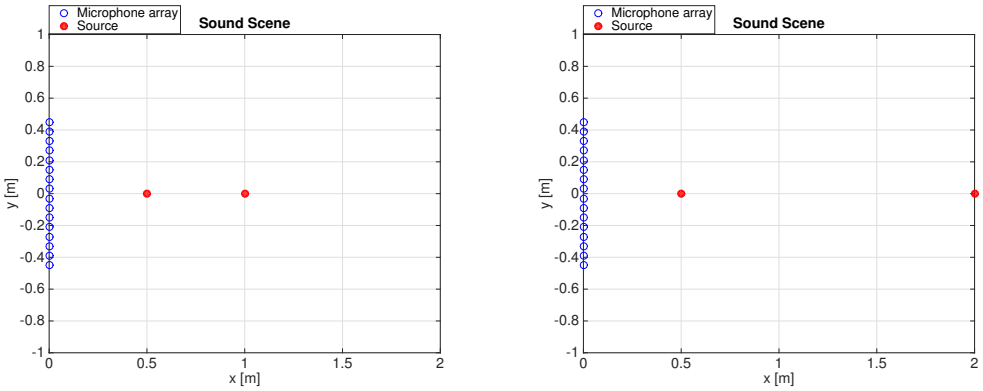
One of the greatest advantages that our method of separation brings is a satisfactory separation also in case of source aligned with the array center. In order to validate this quality of the system, we conducted a simulation session. The same setup configuration described in Section 5.2 is employed in these simulations with the only difference that $W = 5$. We set $W = 5$ because we have seen in Section 4.1 that smaller sub-arrays in situations of source overlap grant better results. Initially, the two sources (female source A and male source B) are positioned in front of the array with position $\boldsymbol{p}_A = [0.5, 0]$ and $\boldsymbol{p}_B = [1, 0]$ as shown in Figure 5.5a, subsequently, source B is moved away along the positive sense of the x axis in 15 uniformly spaced points. The final placement of source B is at $\boldsymbol{p}_B = [2, 0]$ as depicted in Figure 5.5b. In the same fashion of previous simulations, we present SDR and SIR metrics for the different source positioning.

An important improvement in regards of SDR of approximately 9 dB for source A, and 4 dB for source B, is shown if the source distance is increased up to approximately 1 meter, Figure 5.5c. This result can be reconducted to the fact that when the sources are too close to each other, as in the initial configuration, the difference of DOAs $\Delta\theta$ between them with respect to sub-array centers is too small, even for the most external sub-arrays. Due to its higher energy content and different spectral density, source B overwhelms source A.

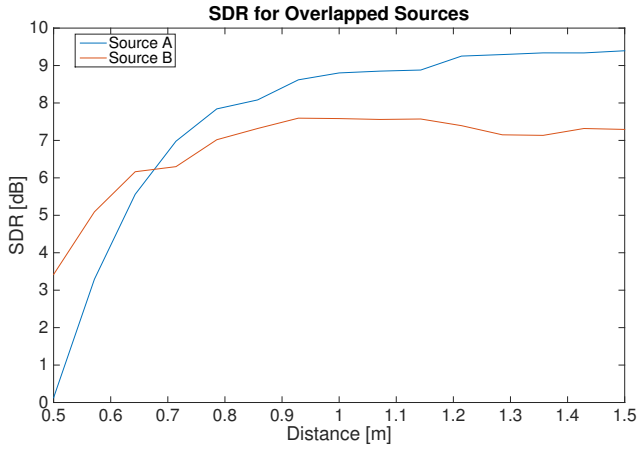
The same considerations made in Section 5.2.1 and 5.2.2 regarding the beampattern behavior of separation filters can be made also in this case. Artifacts are introduced by the filtering operation when low values of $\Delta\theta$ are assumed, therefore decreasing the SDR. When source B is moved further away from the array, the energy imbalance is diminished because of natural propagation attenuation, until we end up with the opposite situation. Since $\Delta\theta$ with respect to the most external sub-arrays does not change sensibly (it assumes values in the range $[13.2, 20.6]$[deg]), the reason for the changes in SDR has to be reconducted to the different distances of the two sources with respect to the array. Furthermore, we can approximately identify a lower threshold of $\Delta\theta = 15.8°$ in correspondence of a distance between sources of 0.7[m]. If this threshold is exceeded a substantial deterioration of performances is experienced, as showed in Figure 5.5c and 5.5d. Recalling that we had already defined a threshold of $\Delta\theta = 15°$ in Section 2.2.3, taken from the study by Souden et al. [27], we find a confirmation of that value in our work.

However, the two trends are quite similar in terms of SDR and they achieve satisfactory results of about 7 and 10 dB when sources have maximum distance. Of course, a more extended array would lead to even better results.

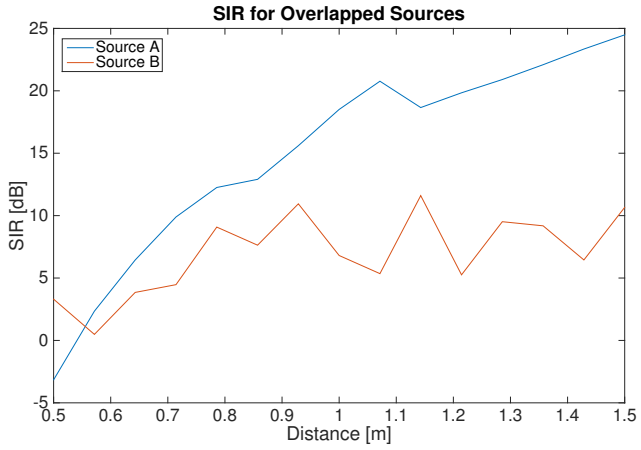
Results showed in Figure 5.5d, concerning SIR parameter, represent a slightly different behavior for source B respect to distance, whereas a tremendous improvement is experienced for source A. In fact, source A which is fixed in front of the array at $\boldsymbol{p}_A = [0.5, 0]$ gains 30 dB of SIR when source B is at the maximum distance value (1.5[m]). This fact can be reconducted to the important difference of distances of sources with respect to the microphone array.



(a) Initial sound scene. **(b)** Final sound scene.



(c)



(d)

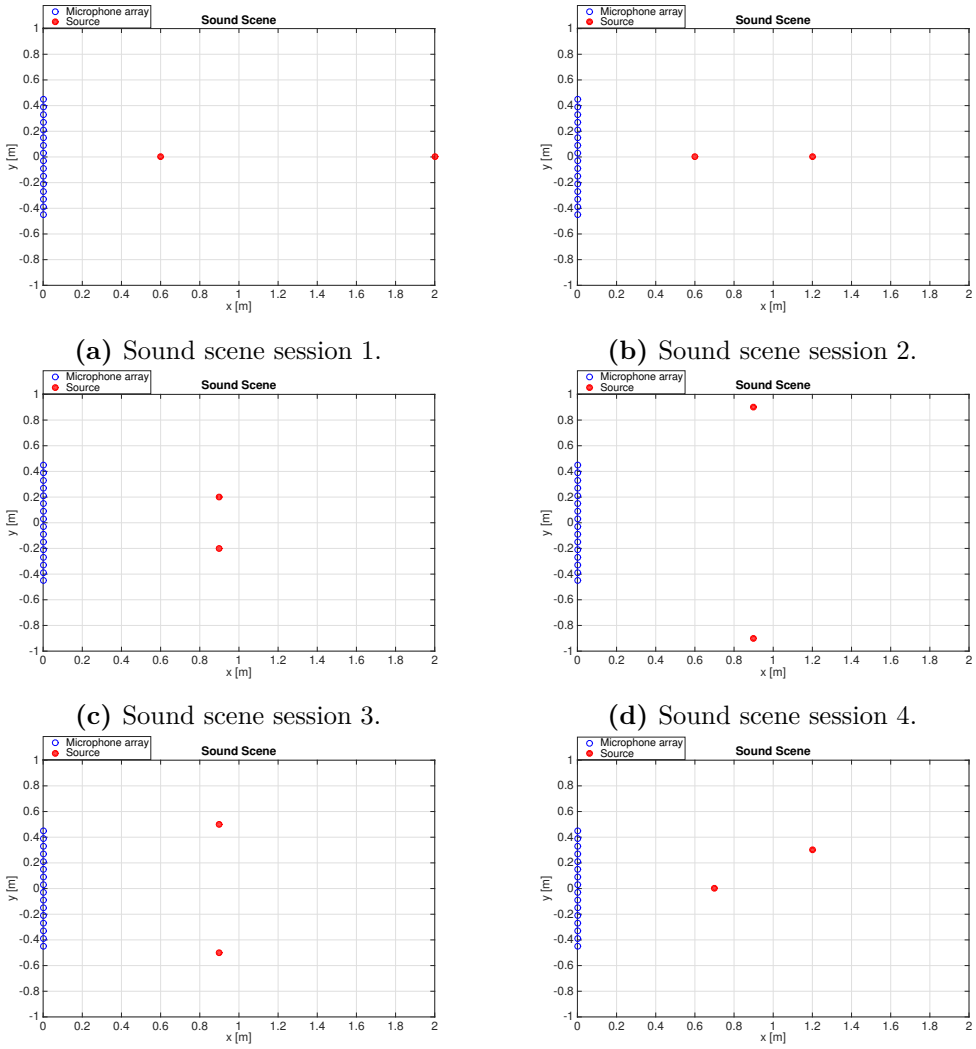
Figure 5.5. SDR and SIR metrics with respect to the distance between the two speech sources.

5.3 Perceptive Tests

An evaluation task on speech separation algorithms should always take into account perceptive tests to thoroughly assess differences and improvements among different possible methods. The reason is that objective metrics based on energy ratios, as SDR and SIR, do not always have a direct counterpart in perceptive metrics such as intelligibility. In fact, metrics can be arranged on an axis of abstraction, from those that measure the most concrete, literal properties of signals, through to those concerned with much higher-level, derived properties in the information extracted from the signals.
The energy-based measures requires that the system being measured reconstructs actual waveforms corresponding to individual sources in a mixture, and that the pre-mixture waveforms of those sources (the "ideal" outputs) are available. Depending on the application, pre-mixture waveforms are not always available representing a possible limitation of these measures. Another limit is that distortions such as fixed phase/time delays or nonuniform gains across frequency which can have only a small effect on the perceived quality of a reconstructed sound, can have a large negative effect on SDR. The common unit of measurement, energy, has in general only an indirect relationship to perceived quality. The same amount of energy will have a widely-varying impact on perceived quality depending on where and how it is placed in time-frequency; this is particularly significant in the case of speech, where most of the energy is below 500 Hz, yet very little intelligibility is lost when this energy is filtered out.
Although there have been some attempts to replace formal listening test (PEAQ, PESQ), there is no substitute for formal listening tests in which subjects rate the perceived quality of various algorithms applied to the same material. For these reasons we designed a perceptive test campaign to explore and understand the relationships between objective measurements of speech separation and to compare our results with a classic approach (i.e. one extended ULA with no sub-arrays) of speech separation based on the LCMV design.

5.3.1 Setup

As a first step, we identified six different sound scene configurations on which simulate speech separation algorithms and then gather people evaluation on the extracted signals. These sound scenes are depicted in Figure 5.6. The choice of the sound scene has been made to be somewhat balanced: on one side there are three configurations (Sessions 1-2-6) for which the proposed method outperforms the classic LCMV approach, on the other side, there are other three configurations (Sessions 3-4-5) for which the proposed method and the LCMV classic approach have similar results. Detailed specifications of the sound scenes are summarized in table 5.1. For these sound scenes, we conducted simulations with the setup described in Section 5.2 with $W = 7$ to obtain SIR metric values.
The obtained values are then compared with *Mean Opinion Score* (MOS) [37], i.e., the values on a predefined scale (0-5) that subjects assign to their opinion of the performance of the speech separation system used to extract the estimated speech sources A and B. In order to collect MOS data with a statistical relevance, listening



(a) Sound scene session 1.

(b) Sound scene session 2.

(c) Sound scene session 3.

(d) Sound scene session 4.

(e) Sound scene session 5.

(f) Sound scene session 6.

Figure 5.6. Simulation sound scenes for perceptive tests.

tests have been conducted on 27 candidates. The chosen candidates have been randomly selected in order to avoid an evaluation bias due to a specific cultural background, or age. A simple graphical interface to conduct the listening tests has been designed in which the audio signals of the six sessions, as the sessions themselves, are randomly proposed to the listener to receive an evaluation score. An example of the graphical interface is depicted in Figure 5.7. The same listening hardware, i.e. headphones AKG k171 MkII, have been provided to the candidates and the listening volume has been kept unchanged throughout all the tests. Finally, the tests took place in the semi-anechoic room located in the Sound and Music Computing Laboratory, Como Campus, Politecnico di Milano, to have a controlled environment without external noises that might affect the final results.

Table 5.1. Simulation setup for perceptive tests.

Session	Sources Position	$\Delta\theta$
1	$\boldsymbol{p}_A = [0.6, 0.0]$,$\boldsymbol{p}_B = [2.0, 0.0]$	$0°$
2	$\boldsymbol{p}_A = [0.6, 0.0]$,$\boldsymbol{p}_B = [1.2, 0.0]$	$0°$
3	$\boldsymbol{p}_A = [0.9, 0.2]$,$\boldsymbol{p}_B = [0.9, -0.2]$	$30°$
4	$\boldsymbol{p}_A = [0.9, 0.9]$,$\boldsymbol{p}_B = [0.9, -0.9]$	$90°$
5	$\boldsymbol{p}_A = [0.9, 0.5]$,$\boldsymbol{p}_B = [0.9, -0.5]$	$60°$
6	$\boldsymbol{p}_A = [0.7, 0.0]$,$\boldsymbol{p}_B = [1.2, 0.3]$	$15°$

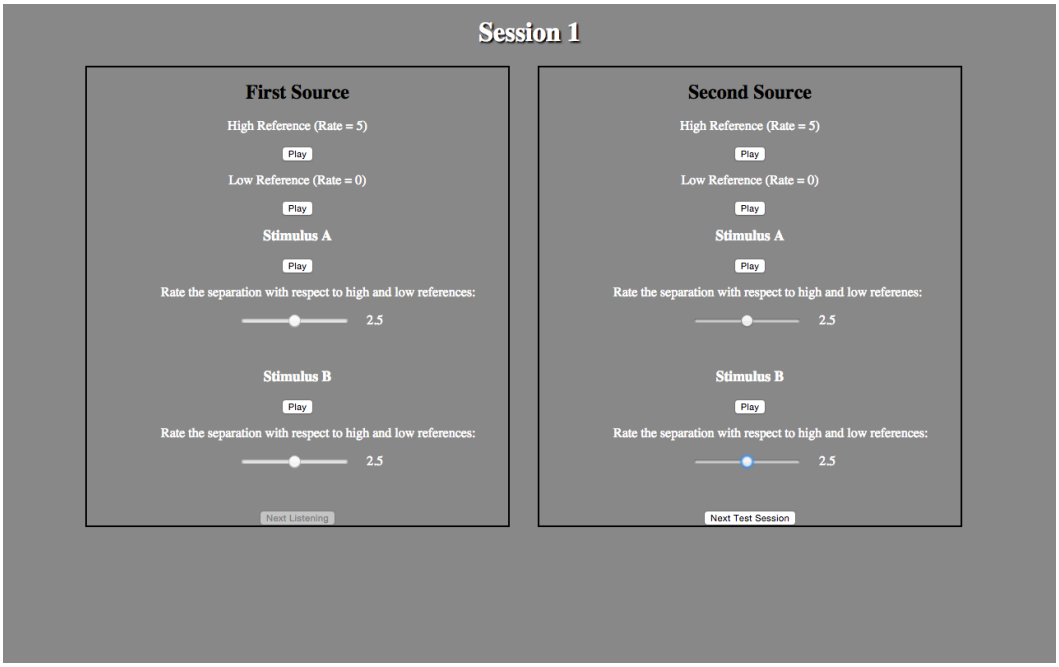
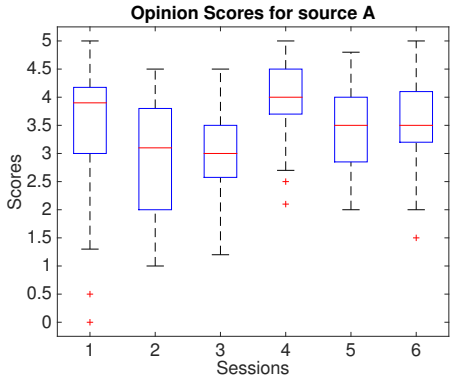


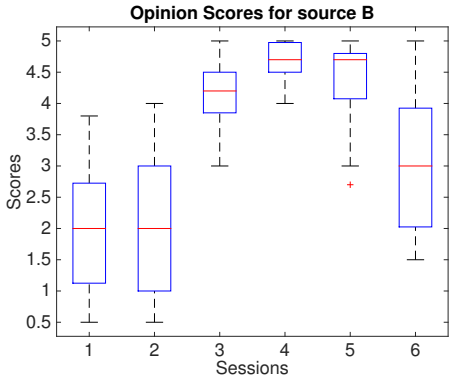
Figure 5.7. Graphical User Interface of the listening tests.

5.3.2 Perceptive tests results

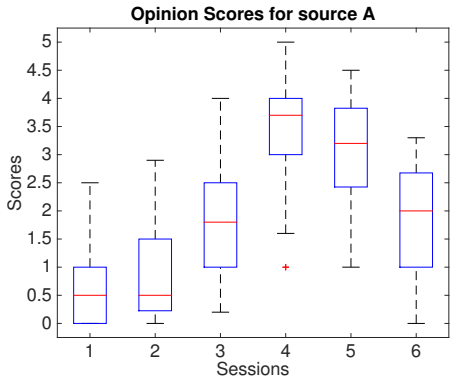
The results obtained with perceptive tests are in general exposed to a high variance due to the subjective nature of the evaluation and to the random choice of the candidates. Discarding subjects whose response exhibits an offset with respect to the average greater than three times the standard deviation of the scores assigned to a specific track of a specific session, is a solution to the aforementioned problem and more significant mean values are obtained as consequence. The statistical results obtained after this regularization are shown in Figure 5.8. In general, the scores assigned to each audio track do not vary significantly, making the results reliable. At a first glance, we can see that the proposed method performs better than the LCMV method, especially in sessions 1-2-6, which exhibit overlap situations of sources in front of the array. The LCMV performances in those cases are significantly unsatisfactory, while the method proposed achieves SIR values of approximately 22 dB, 12 dB and 20 dB for source A, and 14 dB, 9 dB, 20dB for source B, as depicted in Figure 5.9. Interestingly enough, similar trends are mirrored in the MOSs in



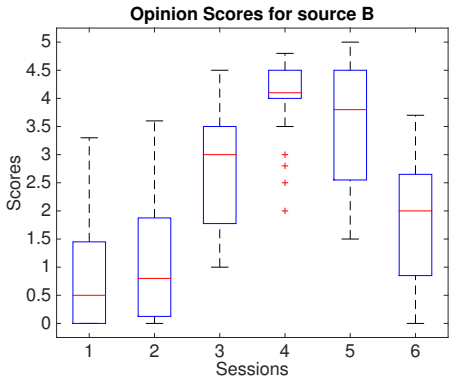
(a) Statistical analysis of the subjective scores assigned to source A with the proposed method.



(b) Statistical analysis of the subjective scores assigned to source B with the proposed method.



(c) Statistical analysis of the subjective scores assigned to source A with LCMV method.

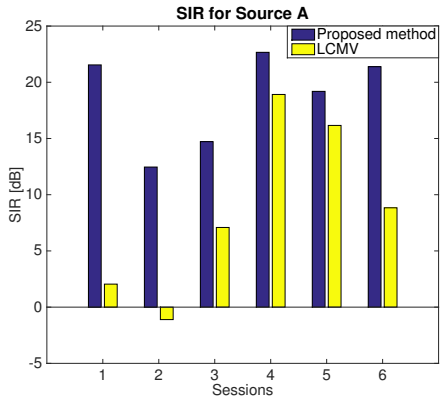


(d) Statistical analysis of the subjective scores assigned to source B with LCMV method.

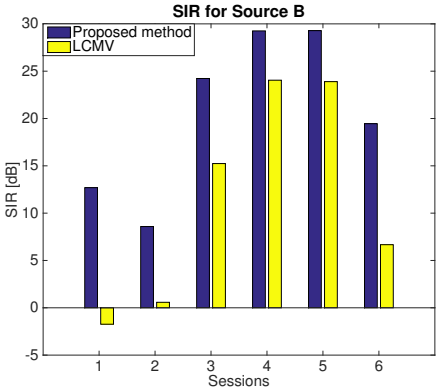
Figure 5.8. Statistical analysis of the subjective scores assigned to sources A and B using the proposed method and the LCMV method. The red bar indicated the median, the blue box comprise the all the values between the 25th and the 75th percentile, the black dashed line indicates the other values but the outliers which are indicated with red crosses.

Figure 5.9c and 5.9d. Source A shows an almost equal MOS of approximately 3.5 for each session, which is a satisfactory result since it means source positioning does not perceptively affect separation performances on source A. Regarding source B, both SIR values and MOSs attain lower values than source A in sessions 1-2 but performs better in 3-4-5. This is in accordance to the results showed in Figure 5.3d, where source B attains higher values for angularly separated sources, and in source overlap situations in front of the array center, Figure 5.5d, where SIR of source B is never more than 15 dB. In addition, SIR results both for source A and B are higher for session 1 than for session 2, where the speech sources are closer to each other, as reflected in Figure 5.5d. However, MOS values do not show this difference, possibly because there is no significant perceivable difference between the two extracted signals obtained with the proposed method. The reason for this result is that sources, also in the initial positions, do not significantly overstep the distance

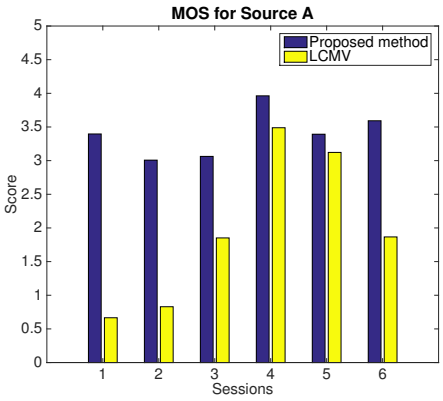
limit between sources defined in Section 5.2.3.



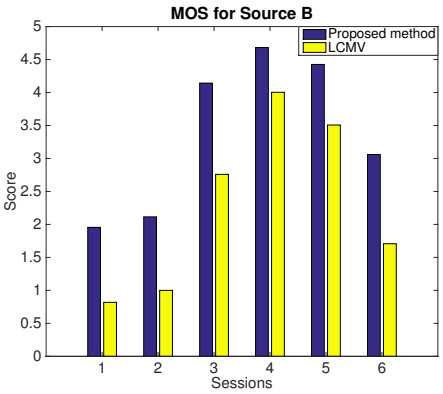
(a) SIR values extracted for source A with the proposed method and with the LCMV.



(b) SIR values extracted for source B with the proposed method and with the LCMV.



(c) Mean opinion scores assigned to source A with the proposed method and the LCMV.

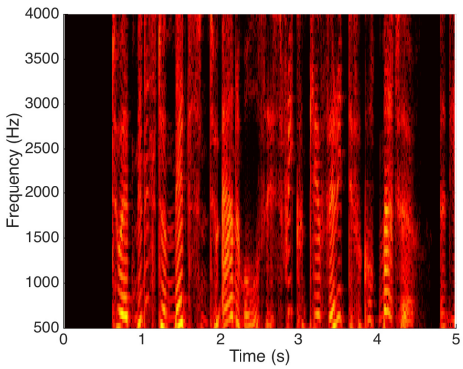


(d) Mean opinion scores assigned to source B with the proposed method and the LCMV.

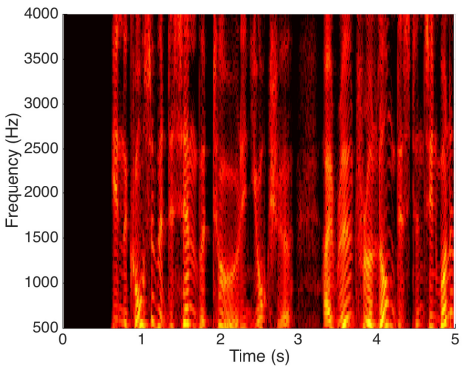
Figure 5.9. Comparison between SIR values computed on the extracted signals obtained with the proposed method and the LCMV and MOS.

5.4 Experimental setup

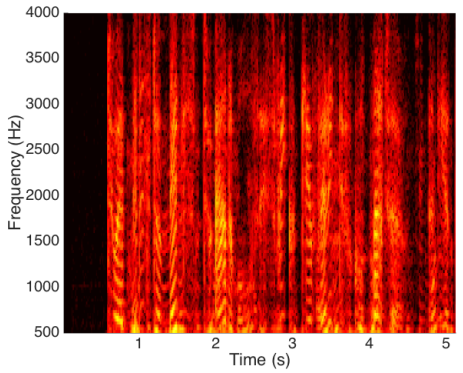
In order to meaningfully corroborate the results obtained with simulations, we conducted two experiments, one in a semi-anechoic room with dimensions $4.6 \times 4.3 \times 2.6$[m] and $T_{60} \approx 50$[ms], the other in a reverberant environment (room dimensions $\approx 5 \times 6 \times 3$, the room is not acoustically treated). Two speakers EMPIRE M2 mounted on a support at approximately 1.5[m] from the ground have been used to reproduce the speech signals mentioned in Section 5.2. The sources have been located at $\boldsymbol{p}_A = [1, 0.5]$[m] and $\boldsymbol{p}_B = [1, -0.5]$[m], and the array of microphones along the y axis extending from $y = 0.45$[m] to $y = -0.45$[m]. We used an array of $M = 16$ STM32 MEMS microphones with $SNR = 63$[dB] with an STM32F407 low power high performance 32-bit microcontroller to control acquisitions. The sampling frequency has been set to 16000 [Hz], thus the time frame length for signal processing to 512 samples which corresponds to 0.032[s]. The number of microphones in each sub-array has been set to $W = 7$. The remaining parameters have been kept equal to those described in Section 5.2.
Regarding the second experiment, performed in a reverberant environment, we assumed available a time of 4[s] of silent signal to estimate the noise power ϕ_e needed to compute the speech separation filters, equation (4.5). We assumed also that noise is stationary throughout the duration of the experiment.



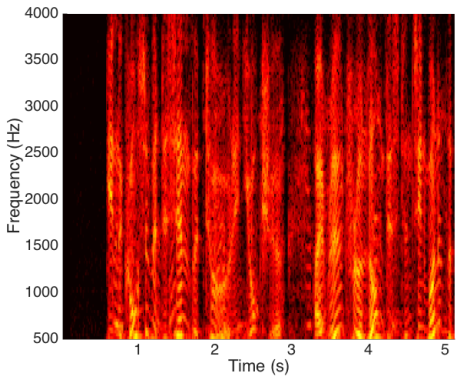
(a) Spectrogram of source A.



(b) Spectrogram of source B.



(c) Spectrogram of extracted source A.

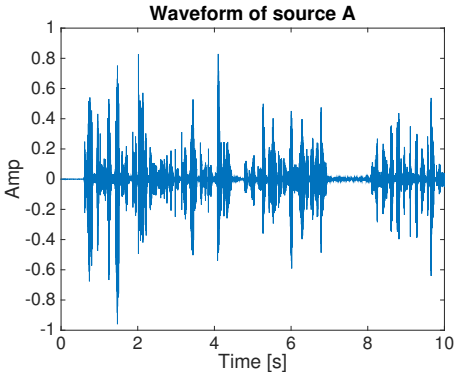


(d) Spectrogram of extracted source B.

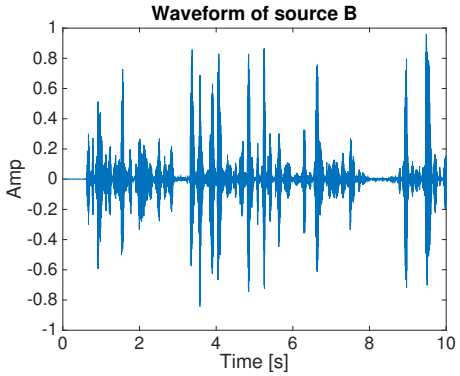
Figure 5.10. Spectrograms of extracted and original source signals.

5.4.1 Separation accuracy for angularly separated sources in semi-anechoic environment

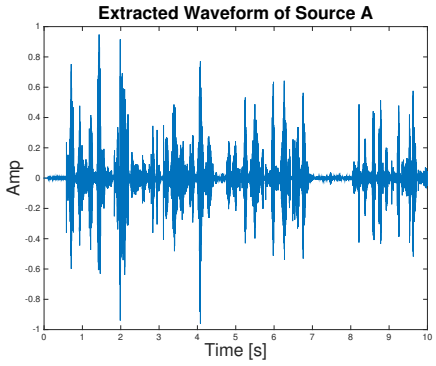
In order to evaluate the results obtained in this experiment, we show the waveforms of the original and extracted signals to understand their time behavior, Figure 5.11. A characterization also in the time-frequency domain is also helpful to easily realize if frequency distortions occurred at any time frame. For this reason, we provide in Figure 5.10 the spectrograms of source and extracted signals. The spectrogram represent the frequency content of the signal at each time frame. In Figure 5.11 we can notice that a perfect separation has not been achieved, since interferer contributions and noise are visible in the silent frames of the extracted signal in comparison to the original source signal. However, the two different waveforms are still well discernible meaning that a good grade of separation has been achieved.



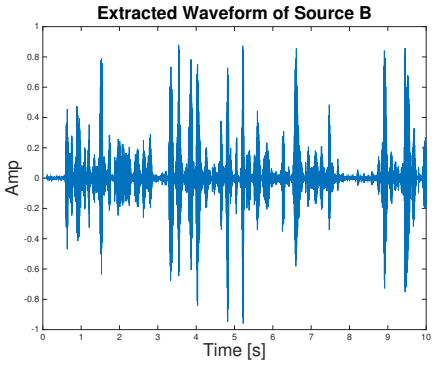
(a) Waveform source A.



(b) Waveform source B.



(c) Waveform extracted source A.



(d) Waveform extracted source B.

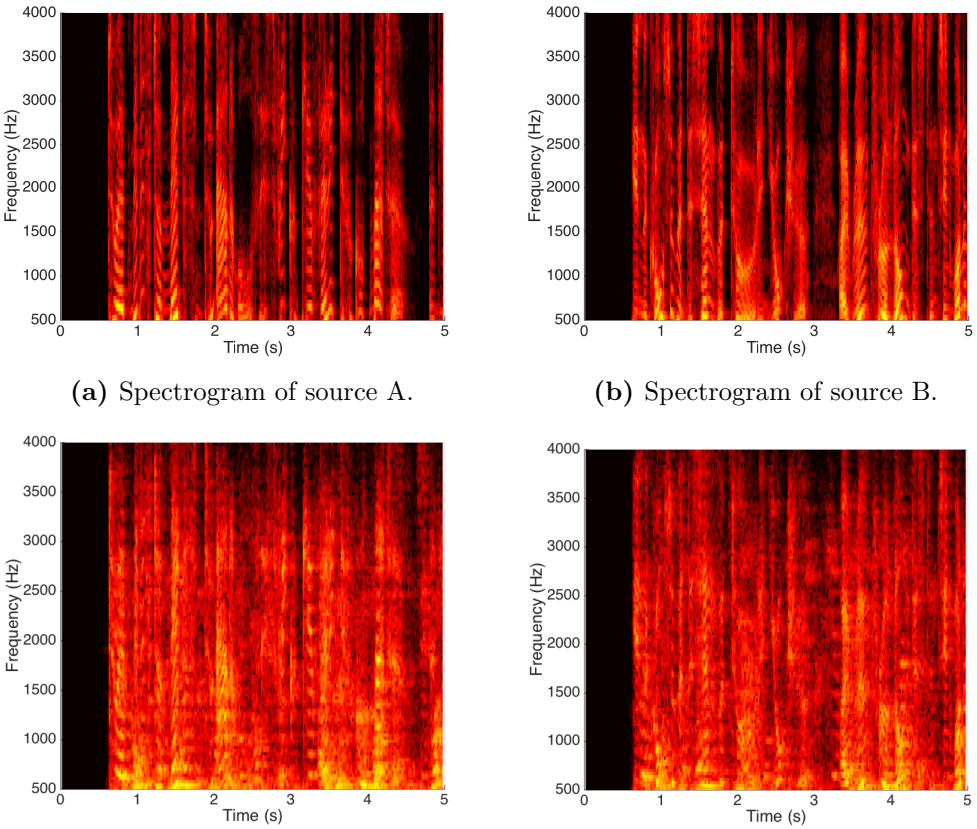
Figure 5.11. Extracted and source signal waveforms.

Some differences can be recognized also in Figure 5.10, where we show the first five seconds of the signal in the time axis for visualization purposes and frequencies between 500 and 4000 [Hz] in the frequency axis because not much frequency content of the signals is present at higher frequencies. This is due to the bandpass filter (cut-off frequencies 500 and 5000) applied to the acquired signals, moreover, we have to consider the frequency response of the microphones and the sampling frequency equal to 16000 [Hz]. An example of dissimilarity, concerning the spectrogram of the extracted and source signal, is at 2.5 [s] where it is evident that the frequency content

has been roughly approximated. Nonetheless, the spectrograms are comparable in almost every time-frequency point, confirming that a satisfactory separation has been achieved.

5.4.2 Separation accuracy for angularly separated sources in reverberant environment

The results of the experiments conducted in a reverberant room, illustrated in Figure 5.12, show that speech separation has still been achieved, even though with coarser results. This is due to the reverberations present in the experimental environment. Reverberations are present in an ambient because of reflections of sound waves in any surface and they can be categorized in early reflections and late reflections. We consider early reflections the first orders of wave reflections, while late reflections the high reflection orders, which produce the diffuse field. Early reflections affect the performance of our method the most, since they can be modeled as image sources, i.e. fictitious sources that emit a sound wave with the same direction of the reflected wave. In addition, being early reflections the less attenuated, they might preponderantly enter in the beam pointed towards a speech source.
However, if we analyze Figure 5.12, frequency content at each frame is still well recognizable, especially for low frequencies which contain the most significant amount of energy. Intelligibility is heavily correlated with accuracy on high frequencies estimation which are still recognizable in our results. In Figure 5.13 we show the waveforms of the two source signals and the extracted signals. These results validate what stated earlier also in the time domain.



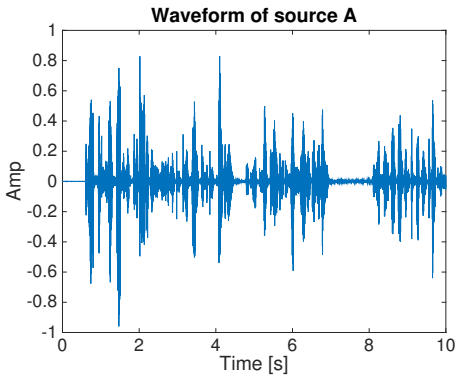
(a) Spectrogram of source A.

(b) Spectrogram of source B.

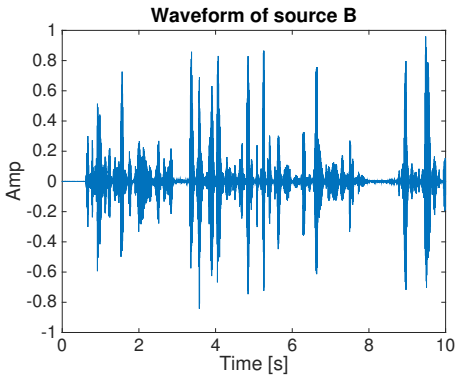
(c) Spectrogram of extracted source A.

(d) Spectrogram of extracted source B.

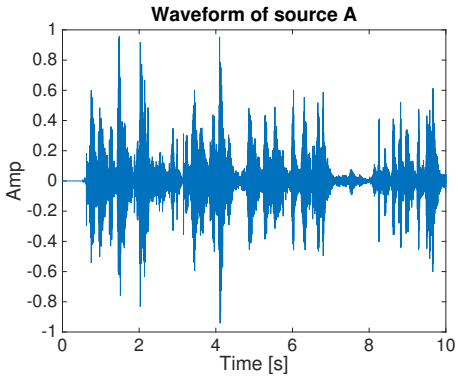
Figure 5.12. Spectrograms of source and extracted signals in reverberant environment.



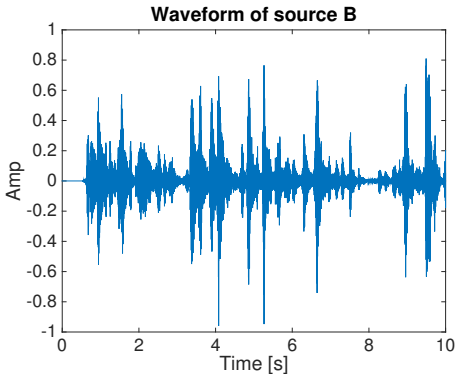
(a) Waveform of source A.



(b) Waveform of source B.



(c) Waveform of extracted source A.



(d) Waveform of extracted source B.

Figure 5.13. Waveform of source and extracted signals in reverberant environment.

5.5 Conclusions

A thorough simulation validation of the proposed method of speech separation has been provided in this Chapter. In particular, we have demonstrated that the angular displacement of speech sources relatively affects our system. Thus speech separation does produce satisfactory results in almost every case, given that the distance between the two sources does not assume values lower than a certain threshold. We identified this distance threshold to be approximately 0.7 [m] with the configuration adopted during simulations. In addition, we showed that satisfactory separation results are achieved also with source positioned in overlap with respect to a sub-array center. Interestingly enough, our method outperforms the LCMV method in this situations achieving higher SIR and SDR results, which are the two objective metrics taken into account in this work. The first one indicates the separation degree of the sources, the other indicates the distortion introduced during the processing step. However, it is known in the literature that such metrics based on energy ratios does not always represent the subjective opinions of human beings. A campaign of perceptive tests has been conducted for this reason. The results obtained further validated the simulative results, showing a correlation between SIR values and MOS assigned to the audio tracks proposed to the listening tests candidates.
Since our method is also based on source localization to inform source separation filters, it seemed important to track the behavior of the system when a source localization error is injected in the source position estimation. We found out that SIR decreases quite fast, diminishing its values of approximately 10 [dB] if the error on the position estimation has a standard deviation value of 0.1. However, SIR values are seen to be almost stable for standard deviation values higher than 0.1, suggesting that after a certain threshold the localization error does not affect performances significantly, despite the SDR values keep decreasing as the standard deviation increase.
Finally, we tested our method also in real scenarios: a semi-anechoic and a reverberant room. We obtained satisfactory results for the semi-anechoic scenario, in line with simulation results. The waveforms of the speech signals are still well distinguishable and also the frequency content is almost unchanged. A slightly inferior result have been achieved in the reverberant room. Even though the environment was highly challenging (the room has no acoustic treatments) the results obtained are satisfactory. Thanks to the reverberations-robust filter employed in this case the waveforms as well as the frequency contents of the signals have not been deteriorated too much, keeping a sufficient grade of separation.
In line with the results obtained, we provided a method to turn a blind source separation into an informed separation problem also in reverberant environment, solving the source overlap problem encountered in the literature of speech separation using a single extended ULA.

Chapter 6

Conclusions and Future Work

In this thesis, we addressed the problem of blind source separation, providing a new approach based on a plenacoustic representation of the sound field. The plenacoustic representation adopted maps the plenacoustic function, sampled in several points by means of sub-array beamforming, into the ray-space image. We saw that sampling the plenacoustic function in one point means computing the pseudospectrum in that point. To this end, a fast data-independent transformation matrix that brings properly rearranged microphone signals into pseudospectra, can be precomputed. Then, a robust wideband image reconstruction that builds upon speech frequency structure can be calculated to obtain the ray-space image. This representation displays acoustic primitives (sources, reflectors etc.) as lines allowing us to employ linear pattern analysis methods to detect these lines and estimate the position of the related sources. Therefore, the advantage that this approach brings is a fast and robust computation of the ray-space representation of the sound field and its parameters, as source positions and source angular displacements. This extracted information turns the blind source separation into an informed problem on which beamforming method can be employed again. By directing an LCMV beamformer from each sub-array to enhance the desired signal and reject the undesired ones, we are able to estimate a separated version of each speech source for each sub-array. The greatest advantage of this multiview approach is that higher separation performances can be achieved when sources overlap with respect to a microphone of the array. This is true because the angular displacement of sources affects the separation performances. Thus, by properly weighting each sub-array contribution with respect to the angular displacement, we can always maximize the separation performances, exploiting the sub-arrays at which sources appear maximally angularly separated. Another consequence of localization is that we can back-propagate the speech signals extracted.

A simulation session has been designed to explore the performance attainable with a certain sound scene setup in case of source overlap with respect to the array center. The two sources have been positioned at an initial minimum distance, then, the furthest source has been moved away along the line connecting the array center with the sources. Interestingly enough, a distance limit has been identified, beyond that the performances, measured as SDR and SIR, rapidly decrease. Otherwise, separation results are satisfactory, showing SIR values over 10 [dB] for the rear source and 20 [dB] for the front one. The rapid decrease of performances has to be

reconducted to a minimum angular displacement required by the LCMV filter to grant adequate separation results. The same performances cannot be attained with a speech separation LCMV method because sources are always seen with angular displacement value equal to zero.
The satisfactory speech separation results obtained, when sources overlap with respect to the array center, have been confirmed by the MOSs collected with a perceptive test campaign. We proposed a listening test to 27 people that rated both the performances of our approach and those obtained with a classic LCMV method in terms of separation degree. The results acquired clearly show that the plenacoustic approach for source separation with an ULA configuration outperforms the LCMV method and resolves the source overlap problem.
In addition, a simulation session has been designed to test the robustness of our method against angular displacement. Two speech sources have been initially positioned in front of the array with a minimum angular displacement, then, they have been progressively moved away, in opposite directions, to increase their angular displacement. The results proved the robustness against large angular displacements, where spatial aliasing affects the separation filters the most. This issue has been resolved with an appropriately tuned diagonal loading applied to the LCVM filters computed at each sub-array. Indeed, SDR values are almost stable at approximately 13 [dB], whereas, SIR values are approximately 20 [dB] for source A and 28 [dB] for source B.
The performances obtained applying our approach, in a situation of angularly displaced sources in real semi-anechoic and reverberant environments, have been tested as well. Promising results have been obtained by comparing the spectrograms calculated in both situations, confirming the goodness of the approach also in reverberant real world scenarios.
In our work we did not consider the possibility to map reflectors into the ray-space image, as proposed in [2], which provides useful information that can be exploited for speech separation purposes. In fact, knowing the position of reflectors can help with the choice of appropriate constraints and desired responses for the LCMV filter to attenuate undesired reflected sound waves. Another possible improvement of the system would consist in a precise localization algorithm that does not produce errors greater than 0.15[m] to guarantee always satisfactory results and an acceptable quality of experience in a real world scenario. Furthermore, when the minimum angular displacement, required by the LCMV filter, is exceeded at sub-arrays, an MVDR filter could be used to replace them and assure higher performances. Although higher SIR values are not guaranteed, the MVDR filter reduces interference plus noise better than the LCMV in these cases, thus, it could boost up SDR performances. An outstanding result would be represented by a real time implementation of the whole system, which is feasible because of the relatively fast beamforming technique. This approach is in contrast to what has been done in the literature, where blind source separation has always been addressed with computationally complex statistical methods.

Bibliography

[1] E.A.P. Habets, J. Benesty, S. Gannot, P.A. Naylor, and I. Cohen. On the application of the lcmv beamformer to speech enhancement. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, pages 141–144, Oct 2009.

[2] D. Markovic, F. Antonacci, A. Sarti, and S. Tubaro. Soundfield imaging in the ray space. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(12):2493–2505, Dec 2013.

[3] O. Thiergart and E.A.P. Habets. An informed lcmv filter based on multiple instantaneous direction-of-arrival estimates. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 659–663, May 2013.

[4] D. Vlaj, M. Kos, M. Grasic, and Z. Kacic. Influence of hangover and hangbefore criteria on automatic speech recognition. In *Systems, Signals and Image Processing, 2009. IWSSIP 2009. 16th International Conference on*, pages 1–4, June 2009.

[5] A. Fleury, N. Noury, M. Vacher, H. Glasson, and J.-F. Seri. Sound and speech detection and classification in a health smart home. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 4644–4647, Aug 2008.

[6] J.P. Dmochowski, Zicheng Liu, and P.A. Chou. Blind source separation in a distributed microphone meeting environment for improved teleconferencing. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 89–92, March 2008.

[7] J. Thiemann, M. Muller, and S. Van De Par. A binaural hearing aid speech enhancement method maintaining spatial awareness for the user. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 321–325, Sept 2014.

[8] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno. Blind separation and dereverberation of speech mixtures by joint optimization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):69–84, Jan 2011.

[9] Don H. Johnson and Dan E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Simon & Schuster, 1992.

[10] Maja Taseska and Emanuel A.P. Habets. A subspace-based perspective on spatial filtering performance with distributed and co-located microphone arrays. In *Speech Communication; 11. ITG Symposium; Proceedings of*, pages 1–4, Sept 2014.

[11] T. Ajdler, L. Sbaiz, and M. Vetterli. The plenacoustic function and its sampling. *Signal Processing, IEEE Transactions on*, 54(10):3790–3804, Oct 2006.

[12] Petre Stoica and Randolph L. Moses. *Spectral Analysis of Signals*. Prentice Hall, Upper Saddle River, NJ, 2005.

[13] B.D. Van Veen and K.M. Buckley. Beamforming: a versatile approach to spatial filtering. *ASSP Magazine, IEEE*, 5(2):4–24, April 1988.

[14] Jr. Gray, A. and J. Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 22(3):207–217, Jun 1974.

[15] H.P.V. C. Method and means for recognizing complex patterns, December 18 1962. US Patent 3,069,654.

[16] III Frost, O.L. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, Aug 1972.

[17] B.D. Carlson. Covariance matrix estimation errors and diagonal loading in adaptive arrays. *Aerospace and Electronic Systems, IEEE Transactions on*, 24(4):397–401, Jul 1988.

[18] M.I. Mandel, S. Bressler, B. Shinn-Cunningham, and D.P.W. Ellis. Evaluating source separation algorithms with reverberant speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1872–1883, Sept 2010.

[19] T.S. Gunawan and E. Ambikairajah. Subjective evaluation of speech enhancement algorithms using itu-t p.835 standard. In *Communication systems, 2006. ICCS 2006. 10th IEEE Singapore International Conference on*, pages 1–5, Oct 2006.

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

[21] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, Aug 1969.

[22] S. Markovich, S. Gannot, and I. Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1071–1086, Aug 2009.

[23] V.G. Reju, Soo Ngee Koh, and I.Y. Soon. Underdetermined convolutive blind source separation via time-frequency masking. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(1):101–116, Jan 2010.

[24] J. Benesty, Jingdong Chen, Yiteng Huang, and J. Dmochowski. On microphone-array beamforming from a mimo acoustic signal processing perspective. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1053–1065, March 2007.

[25] V. Tourbabin, M. Agmon, B. Rafaely, and J. Tabrikian. Optimal real-weighted beamforming with application to linear and spherical arrays. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9):2575–2585, Nov 2012.

[26] E.A.P. Habets, J. Benesty, and P.A. Naylor. A speech distortion and interference rejection constraint beamformer. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):854–867, March 2012.

[27] M. Souden, J. Benesty, and S. Affes. On optimal beamforming for noise reduction and interference rejection. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, pages 109–112, Oct 2009.

[28] O. Thiergart, M. Taseska, and E.A.P. Habets. An informed mmse filter based on multiple instantaneous direction-of-arrival estimates. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*, pages 1–5, Sept 2013.

[29] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.

[30] Mahmood R. Azimi-Sadjadi, Ali Pezeshki, Louis L. Scharf, and Myron E. Hohil. Wideband doa estimation algorithms for multiple target detection and tracking using unattended acoustic sensors, 2004.

[31] D. Salvati, C. Drioli, and G.L. Foresti. Incoherent frequency fusion for broadband steered response power algorithms in noisy environments. *Signal Processing Letters, IEEE*, 21(5):581–585, May 2014.

[32] Yanna Ma and Akinori Nishihara. Efficient voice activity detection algorithm using long-term spectral flatness measure. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1), 2013.

[33] H. Cox, R.M. Zeskind, and M.M. Owen. Robust adaptive beamforming. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(10):1365–1376, Oct 1987.

[34] L. Romoli, P. Peretti, S. Cecchi, L. Palestini, and F. Piazza. Real-time implementation of wave field synthesis for sound reproduction systems. In *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, pages 430–433, Nov 2008.

[35] M. Vetterli, J. Kovačević, and V.K. Goyal. *Foundations of Signal Processing*. Cambridge University Press, 2014.

[36] Yi Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):229–238, Jan 2008.

[37] International Telecommunication Union. ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology. Technical report, July 2006.