Acoustic Source Localization Using Microphone Arrays

Daniele Salvati

Department of Mathematics and Computer Science

University of Udine

A thesis submitted for the degree of

PhilosophiæDoctor

March 2012

Daniele Salvati
Department of Mathematics and Computer Science, University of Udine, Italy,
Doctorate in Multimedia Communication, Faculty of Education, 2009-2011, ciclo XXIV.
email: daniele.salvati@uniud.it
website: http://users.dimi.uniud.it/ ∼ daniele.salvati/

Supervisor:

Sergio Canazza
Assistant Professor, Department of Information Engineering, University of Padova, Italy.

Abstract

The sensory capacity to analyze acoustic space is a very important function of an auditory system. The need for the development of an understanding of the sound environment has attracted many researchers over the past twenty years to build sensory systems that are capable of locating acoustic sources in space. Several application areas that may potentially provide advantages in using the acoustic location have led to the development of many signal processing algorithms, which mostly consider the type of acoustic environment, the type of sounds of interest, and the aim of localization.

In this thesis, we focus first on the nature of the sounds of interest, then on the environment where the sounds are located, and finally on the type of context where it could be applied. Two fundamental characteristics of a sound are its spectral content and its evolution over time. The latter allows the classification of short sound events and continuous sounds, while the spectral content over time characterizes the noisy or harmonic (or generally pseudo-periodic) sounds. Currently, localization systems have achieved good results with regard to the location of a single source, while the case of multiple simultaneously active sources has only recently been investigated in depth by the scientific community. In particular, the localization of sounds in a multi-source case with short events has been limited to the use of systems implementing a Bayesian filter because an initialization time is required to bring in the optimum phase of work, and therefore, greater potentials are applied to the continuous sounds. Short events are typically present in real-world situations, such as in many urban and natural contexts. Thus, they may occur in outdoor environments (but are not limited to these) and are typically free-field and far-field environments. The interest in locating these types of sounds may be attractive for audio surveillance, sound monitoring and the analysis of acoustic scenes.

In this context, a new approach to solve the multi-source case, called the Incident Signal Power Comparison, is proposed. This approach is based on identifying the incident signal power of the sources of a microphone array using beamforming methods and comparing

the power spectrum between different networks of arrays network using spectral distance measurement techniques. This method solves the ambiguities caused by simultaneous multiple sources by identifying sounds to enable the spectrum power distance to minimize the error criterion.

Furthermore, the performance of one of the most widely used signal processing techniques for time delay estimation, the Generalized Cross-Correlation, is dramatically reduced in the case of harmonic sounds, or generally pseudo-periodic sounds; a framework architecture is proposed to solve the localization of these types of sounds. These sounds are related to musical instruments, and thus, the context is a near-field reverberant environment. Potential applications may include human-computer interaction systems for controlling audio processing. The system proposed is based on an adaptive parameterized generalized cross-correlation and phase transform weighting with a zero-crossing rate threshold, which includes a pre-processing Wiener filter and a post-processing Kalman filter.

Both prototypes were produced with arrays of a very small size (the minimum necessary to locate noise sources in the plan): in the far-field environment, two linear arrays each consisting of four microphones were used, and in the near-field case, a linear array of three microphones was used. The purpose was to evaluate the performance of minimal systems, which from a practical standpoint are less invasive; moreover, the computational cost is particularly advantageous for real-time applications.

In conclusion, the objective of this thesis is to present the state of the art in acoustic source localization using a microphone array and to propose two experimental real-time prototype systems in the far-field and near-field environments. The first prototype aims to investigate the use of multi-sources with microphone arrays for applications of analyzing, monitoring and surveilling acoustic scenes in real contexts in which short-duration events often occur and are related to far-field and free-field environments. The objective of the second prototype is to open new fields in human-computer interaction and musical applications by solving the problem of harmonic sound localization in moderate reverberant and noisy environments.

To Laura and Giacomo

Acknowledgements

I wish to express my sincere gratitude to my advisor, Sergio Canazza (Assistant Professor, University of Padova, Italy), for his support, guidance and feedback throughout my doctoral studies.

This work was performed at the Artificial Vision and Real Time Systems Laboratory of University of Udine. I would like to thank all the people working at the laboratory, especially Gian Luca Foresti (Professor, University of Udine, Italy) for providing me with the facilities and resources that made this research possible. Special appreciation is given to Antonio Rodà (Assistant Professor, University of Padova, Italy) for his suggestions, critical discussions and assistance in the laboratory experiment.

I thank my readers, Rudolf Rabenstein (Professor, University of Erlangen-Nrnberg, Germany) for his valuable advice and detailed comments, and Augusto Sarti (Associate Professor, Politecnico di Milano, Italy) for his helpful suggestions.

I would also like to thank all those who reviewed my papers over the last three years for their insightful comments, guiding my research down the right path.

Finally, I would like to express my deepest gratitude to my wife, Laura Cingolani, for accompanying and supporting me, and to my son, Giacomo, who was born during my doctoral studies.

Contents

1 Introduction

1.1 Acoustic Source Localization

Acoustic Source Localization (ASL) is an increasingly important aspect in a growing number of applications. The aim of an ASL system is to estimate the position of sound sources in space by analyzing the sound field with a microphone array, a set of microphones arranged to capture the spatial information of sound. From a conceptual point of view, the localization of acoustic sources is simple. However, the performance of these systems involves considerable complexity and still remains an open field of research.

Sound is a mechanical wave that propagates in an elastic medium with finite speed. If we place microphones that analyze the change in pressure at different points in space, the acoustic wave reaches these sensors at different times. This simple concept underlies the basic method for estimating the space location of an acoustic source. The complexity and the problems that we face concern the nature of sound, the propagation phenomena and the technology we use in developed localization systems. In particular, we emphasize that the acoustic sound is a signal that has a non-stationary different spectral content (depending on the physical nature of the perturbation that has produced it). Different spectral components have different capacities to spread during their propagation in air because the absorption of acoustic energy from the air is greater as the frequency increases. In addition, the propagation of

sound undergoes the phenomenon of reflection, refraction and attenuation of energy in the propagation medium. Hence, in indoor acoustic environments, we will have the problem of reverberation, which may be critical in many applications. Moreover, the speed of sound is influenced by temperature because an increase in temperature corresponds to an increase in speed. In outdoor environments, this phenomenon can cause deviations in the pressure wave. Other environmental factors can cause additional difficulties for the localization of large distances; for example, humidity affects the attenuation of energy as a function of frequency, and wind can cause noise barriers (for a review of the influence of meteorological conditions on sound propagation see [Ingard, 1953]). Another important issue involves the noise of the signal. Signals from the microphones are affected by noise produced by the environment, interference and electric/mechanical signal transduction. In addition, the presence of multiple, simultaneously active sources is a major issue that has only recently been investigated. Finally, the use of digital systems, which implies the discretization of the signal over time, leads to the discretization of the analysis space, implying that the area appears to have a non-homogeneous accuracy in the analysis of location. Estimation of the accuracy can be achieved with an increase in the number of sensors, which means an increase in the logical and physical complexity of the system.

Fields of application in which identification of the location of acoustic sources is desired include audio surveillance, teleconferencing systems, hands-free acquisition in car, system monitoring, human-machine interaction, musical control interfaces, videogames, virtual reality systems, voice recognition, fault analysis of machinery, autonomous robots, processors for digital hearing aids, high-quality recording, multi-party telecommunications, dictation systems and acoustic scene analysis.

In general, a localization system can be represented by the following steps [Tashev, 2009], shown in Figure 1.1. The acquisition block is formed by the array of microphones or an array network, and

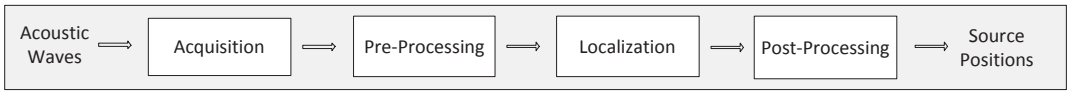

Figure 1.1: *Block diagram of the ASL system*

it is designed to capture the pressure waves in space. Pre-processing of the signals consists of noise reduction to increase the Signal to Noise Ratio (SNR), which permits a more precise estimation of the source position in particularly critical situations with low SNR. During localization, all of the signals from the network arrays are processed using algorithms that provide the position of the sources. Finally, post-processing is a fundamental and crucial step that provides increased precision of the position data

and attempts to minimize or eliminate results obtained from reflection, reverberation and error measurements. It also provides the ability to track the source in case of movement.

Multi-channel signal processing for sound localization can be divided into two categories: Time Delay Estimation (TDE) and Steered Response Power (SRP) beamforming. The first category consists of estimating the Time Difference Of Arrival (TDOA) between a microphone pair, using the classic Cross-Correlation, the Generalized Cross-Correlation (GCC) [Knapp & Carter, 1976] and the Adaptive Eigenvalue Decomposition (AED) [Benesty, 2000] based on the Blind System Identification (BSI), which focuses on the impulse responses between the source and the microphones. To improve the performance in the case of an array containing M microphones ($M > 2$), the Steered Response Power Phase Transform (SRP-PHAT) [DiBiase *et al.*, 2001] provides the sum of the GCC-PHAT from all of the microphone pairs, while the Multichannel Cross-Correlation Coefficient (MCCC) [Chen *et al.*, 2003] [Benesty *et al.*, 2004] uses the spatial prediction error to measure the correlation among multiple signals and uses the redundant information between microphones to estimate the TDOA in a more robust manner under a reverberant and noisy condition. The extension of the AED in the case of multiple microphones was proposed in Huang & Benesty [2003], and it is called Adaptive Blind Multichannel Identification (ABMCI). In contrast, the SRP is based on maximizing the power output of a beamformer. Beamforming is a combination of the delayed signals from each microphone in a manner in which an expected pattern of radiation is preferentially observed. The conventional beamformer is the Delay & Sum (DS) [Bartlett, 1948]; it consists of the synchronization of signals that steer the array in a certain direction, and it sums the signals to estimate the power of the spatial filter. The high-resolution SRP has been developed to improve the performance of the spatial filter, and the adaptive beamformer is called the Minimum Variance Distortionless Response (MVDR) due to Capon [Capon, 1969]. The MUltiple SIgnal Classification (MUSIC) algorithm is based on an eigen subspace decomposition method [Schmidt, 1979] [Schmidt, 1986], and the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) is based on subspace decomposition exploits the rotational invariance [Paulraj *et al.*, 1986] [Roy *et al.*, 1986] [Roy & Kailath, 1989].

Recently, more sophisticated algorithms have been proposed for time delay estimation that use Minimum Entropy [Benesty *et al.*, 2007] [Wen & Wan, 2011] and broadband Independent Component Analysis (ICA) [Lombard *et al.*, 2011]. In this paper, the authors demonstrate that the ICA-based methods are more robust against high background noise levels compared with the conventional GCC-PHAT approach.

A widely used approach, called the indirect method, is used to estimate source positions and consists of two steps: in the first step, a set of Time Difference Of Arrivals (TDOAs) are estimated using

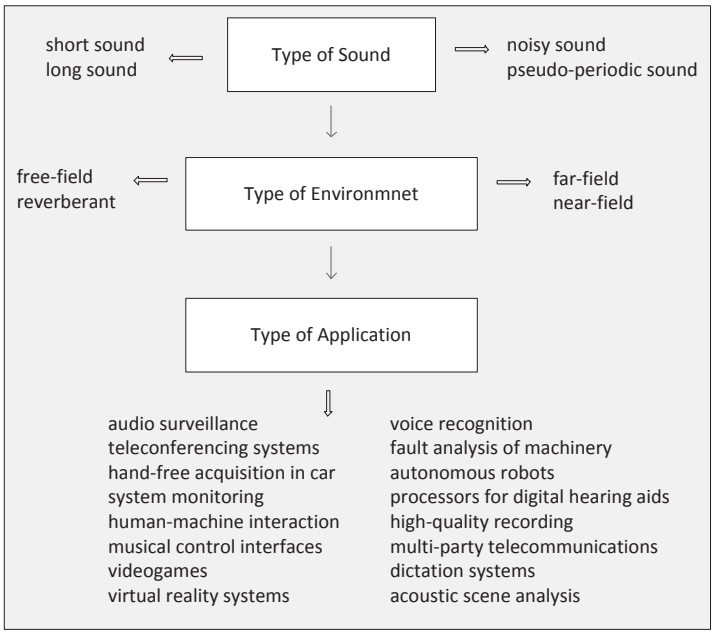

Figure 1.2: *Steps for the considered variables for localization.*

measurements across various combinations of microphones, and in the second step, when the position of the sensors and the speed of sound are known, the source positions can be estimated using geometric considerations and approximate estimators: closed-formed estimators based on a least squares solution [Schmidt, 1972] [Schau & Robinson, 1987] [Smith & Abel, 1987] [Chan & Ho, 1994] [Brandstein *et al.*, 1997] [Huang *et al.*, 2001] [Gillette & Silverman, 2008] (for an overview on closed-form estimators, see [Stoica & Li, 2006]) and iterative maximum likelihood estimators [Hahn & Tretter, 1973] [Wax & Kailath, 1983] [Stoica & Nehorai, 1990] [Segal *et al.*, 1991] [Chen *et al.*, 2002] [Georgiou & Kyriakakis, 2006] [Destino & Abreu, 2011].

However, the direct method yields an acoustic map of the area, from which the position of the sources can be estimated directly and spatial likelihood functions can be defined [Aarabi, 2003] [Omologo & DeMori, 1998] [DiBiase *et al.*, 2001] [Ward *et al.*, 2003] [Pertilä *et al.*, 2008].

Both of these procedures have been tested in many single source scenarios; however, in multiple sources cases, they require new consideration. Several works address the problem of multiple sources using a Bayesian approach based on the tracking of the sources and using Kalman filter [Strobel *et al.*, 2001a] [Strobel *et al.*, 2001b] [Bechler *et al.*, 2003] [Potamitis *et al.*, 2004] [Klee *et al.*, 2006] [Gannot & Dvorkind, 2006] [Liang *et al.*, 2008] [Seguraa *et al.*, 2008] and Particle filter [Zotkin *et al.*, 2002]

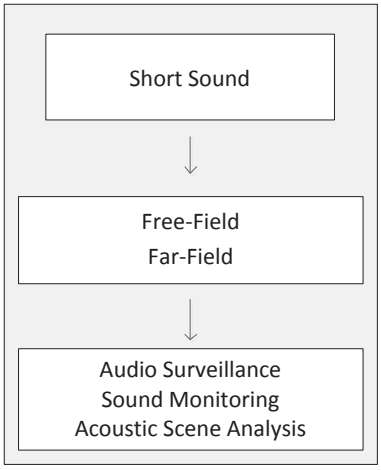

Figure 1.3: *Steps for the considered variables for ISPC localization.*

[Ward *et al.*, 2003] [Antonacci *et al.*, 2006] [Michaudy & Rouat, 2007] [Talantzis *et al.*, 2008] [Quinlan *et al.*, 2009] [Levy *et al.*, 2011].

Some studies consider an approach without tracking. In [Nishiura *et al.*, 2000] [Scheuing & Yang, 2008] [Hu & Yang, 2010] [Brutti *et al.*, 2010] [Lombard *et al.*, 2011], solutions are proposed for the multiple source problem in near-field and reverberant environments.

In applications involving very large arrays, decentralized data fusion methods provide optimal fusing of source estimation measurements by two or more localization systems. The goal of these methods is to reduce the cost of computation and communication in a distributed sensor network [Stoica *et al.*, 1995] [Liu *et al.*, 2003] [Kozick & Sadler, 2003] [Chen *et al.*, 2004] [He & Chong, 2004] [Prandi *et al.*, 2008].

This thesis focuses first on the nature of the sound of interest, then on the environment in which it is located and finally on the context in which it may be applied. The Figure 1.2 presents these variables.

In general, we can divide short event sounds from long and continuous events into time domains. In the multi-source events with short duration sounds, techniques based on Bayesian filters, and on the Kalman and Particle filters, are limited because these filters need some time to bring in the initialization phase for optimal functioning. As a solution to this problem, this thesis presents a new approach: the Incident Signal Power Comparison (ISPC). It is based on source separation and on a verification of similarity among sounds. The first step consists of source separation using beamforming techniques and estimation of the Incident Signal Power (ISP) of every source captured on the array. The second step involves the comparison of the ISP spectrum using a spectral distance measure. The ISP spectrum permits identification of sounds so that the spectrum power distance minimizes an error criterion. The

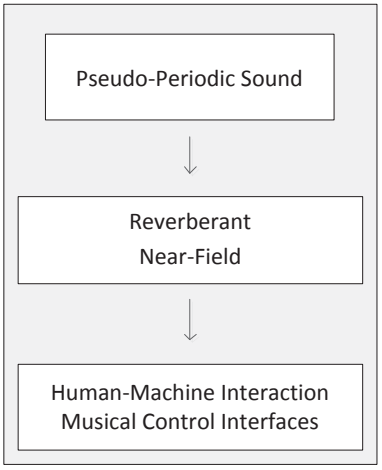

Figure 1.4: *Steps for the considered variables for pseudo-periodic source localization.*

environment is considered far-field and free-field, i.e., an outdoor environment. The location of these types of sound can be employed for audio surveillance, sound monitoring and analysis of acoustic scenes. Figure 1.3 shows the variables that are involved in the ISPC localization system.

By analyzing the spectral content, noisy sounds and pseudo-periodic or harmonic sounds can be extracted. The latter are closely related to musical instruments, so their location can usually be considered in near-field and reverberant environments. Domain applications include human-machine interaction in musical applications. The performance of one of the most widely used signal processing techniques for time delay estimation, the Generalized Cross-Correlation (GCC), is dramatically reduced in the case of harmonic sounds, or generally pseudo-periodic sounds. Thus, as a solution to this problem, this thesis proposes an adaptive parameterized GCC and Phase Transform (PHAT) weighting with a zero-crossing rate threshold, which includes a pre-processing Wiener and a post-processing Kalman filter. These interfaces, based on a novel architecture, can be used to control audio processing through the spatial movement of a sound source, such as voice, traditional musical instruments and sounding mobile devices, and opens new possibilities for applications in musical contexts such as expressive audio control performances. Figure 1.4 shows the variables that are involved in the pseudo-periodic source localization.

1.2 Organization of the Thesis

The state of the art of acoustic source localization, pre-processing for noise reduction techniques, and post-processing for localization enhancement are addressed in Chapters 2, 3, 4, and 5, respectively. In Chapter 2, the localization problem and the methods for the source position estimation are described: closed-form estimators, iterative maximum likelihood estimators, spatial likelihood functions and decentralized data fusion. The most important multi-channel signal processing techniques for sound localization are explained in Chapter 3 following a description of signal models: the TDE methods for microphone pair, multiple microphone array and the SRP beamforming approach. Pre-processing for signal enhancement is described in Chapter 4 and includes both frequency (Short-Time Spectral Attenuation) and time domain (autoregressive model and Extended Kalman filter) algorithms. The post-processing methods for localization enhancement are presented in Chapter 5 and include the Bayesian tracking, Kalman and Particle filter, and clustering methods. Finally in Chapter 6, two experimental prototypes are introduced as an innovative contribution to this thesis. In the first part, the Incident Signal Power Comparison approach is described to solve the multi-source problem in far-field and free-field environments, with particular attention to short-duration sounds. After the presentation of the prototype setup, some experimental results in real-world scenarios are presented. In the second part, a prototype for localization of pseudo-periodic sounds and some experimental results in real, moderate reverberant and noisy environments are presented. Chapter 7 presents the conclusions, with a summary of the contents of this thesis as well as proposed solutions and future potential directions for this work.

2 Source Localization

2.1 Problem Formulation

We consider a sound field in which the sources are omnidirectional and radiate the sound in spherical waves, thus neglecting the shape and size of the source. We also assume that the transmission in the air medium is homogeneous with a constant and known speed of sound. We can then define a three-dimensional Cartesian space (Figure 2.1) and the two vectors that identify the position of the acoustic source s_n and the omnidirectional microphone m_i

$$\begin{aligned} \mathbf{s}_n &= [x_n \; y_n \; z_n]^T \\ \mathbf{m}_i &= [x_i \; y_i \; z_i]^T. \end{aligned} \tag{2.1}$$

We can calculate the distance between the source s_n, and the microphone m_i and we obtain

$$r_i = ||\mathbf{s}_n - \mathbf{m}_i|| \tag{2.2}$$

where $||\cdot||$ denotes the Euclidean vector norm. The approximate speed of sound c (m/s) in dry (0% humidity) air can be calculated from the air temperature T_C (degrees Celsius)

$$c = 331.3 + 0.606T_C. \tag{2.3}$$

Then the propagation time of the sound wave from the source s_n to the microphone m_i is

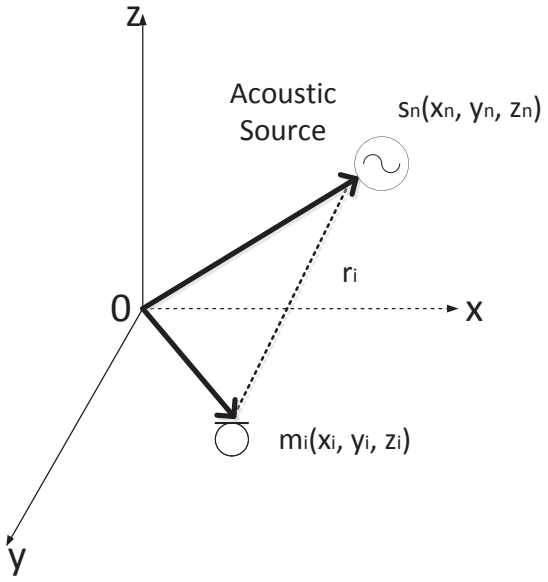

Figure 2.1: *Cartesian space and the variable of localization problem.*

$$t_i = \frac{r_i}{c}. \tag{2.4}$$

We then add the microphone m_j ($j \neq i$), to create an array of two microphones with distances $d_{ij} = ||\mathbf{m}_j - \mathbf{m}_i||$. The Time Difference Of Arrival (TDOA) of the wavefront at the two microphones becomes

$$\tau_{ij} = t_j - t_i = \frac{r_j - r_i}{c} = \frac{\delta_{ij}}{c}. \tag{2.5}$$

The difference δ_{ij} is usually termed the range difference. From a geometrical point of view, after substituting equation (2.2) in (2.5) and expanding, we have

$$\begin{aligned}\tau_{ij} = \frac{1}{c}\Big(&\sqrt{(x_n - x_j)^2 + (y_n - y_j)^2 + (z_n - z_j)^2} - \\ &\sqrt{(x_n - x_i)^2 + (y_n - y_i)^2 + (z_n - z_i)^2}\Big).\end{aligned} \tag{2.6}$$

This equation is that of a hyperboloid. It describes all of the possible points of an acoustic source that generates the same TDOA to an array of two microphones. To uniquely determine the position of the source (the three unknown coordinates), we need, at a bare minimum, a system of three equations, which describe the intersection of the three hyperboloids. The solution is obtained by adding two microphones to the array. However, this condition is not sufficient to have a unique solution because the geometry of the array is important to uniquely estimate the position of the source. To better understand this issue we

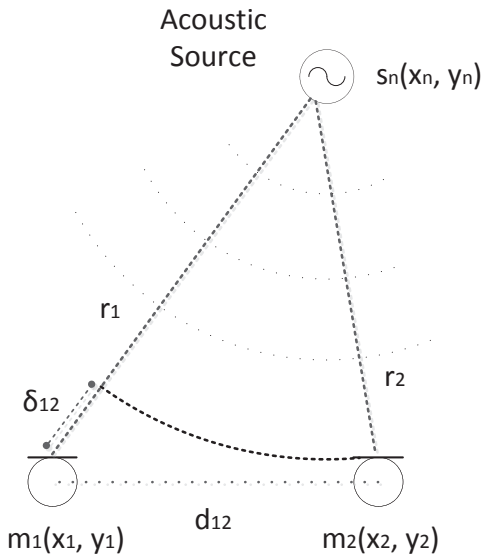

Figure 2.2: *Planar case of two microphones in a near-field environment.*

consider the case in a plane space with two microphones. Referring to Figure 2.2, we have

$$\tau_{12} = \frac{1}{c}\Big(\sqrt{(x_n - x_2)^2 + (y_n - y_2)^2} - \sqrt{(x_n - x_1)^2 + (y_n - y_1)^2}\Big). \tag{2.7}$$

Figure 2.3 shows a hyperbola that represents the same TDOA. If we add a third microphone from a distance, a uniform linear array is obtained and we have a second hyperbola (Figure 2.4)

$$\tau_{23} = \frac{1}{c}\Big(\sqrt{(x_n - x_3)^2 + (y_n - y_3)^2} - \sqrt{(x_n - x_2)^2 + (y_n - y_2)^2}\Big). \tag{2.8}$$

However, we note in Figure 2.4 that there is an ambiguity front-rear, for which there is a phantom source to the real one. To resolve this problem, we can add a fourth microphone to form a plane array. From a practical point of view, the linear array is widely used because it is suitable to analyze a half-plane, such as a room. This ambiguity is extended to the double-sided three-dimensional case. In fact, with a planar array of four microphones we can locate the source in the half-space, and we have to add a fifth microphone to obtain a three-dimensional array to investigate the entire space.

In the case of a source located away from the microphone array, we are no longer able to detect the spherical wavefront, which is then approximated by the wavefront plane. Figure 2.5 shows the far-field environment with two microphones. In this situation with an array of microphones, we are able to estimate only the Direction Of Arrival (DOA) of the source but not its distance from the array. The

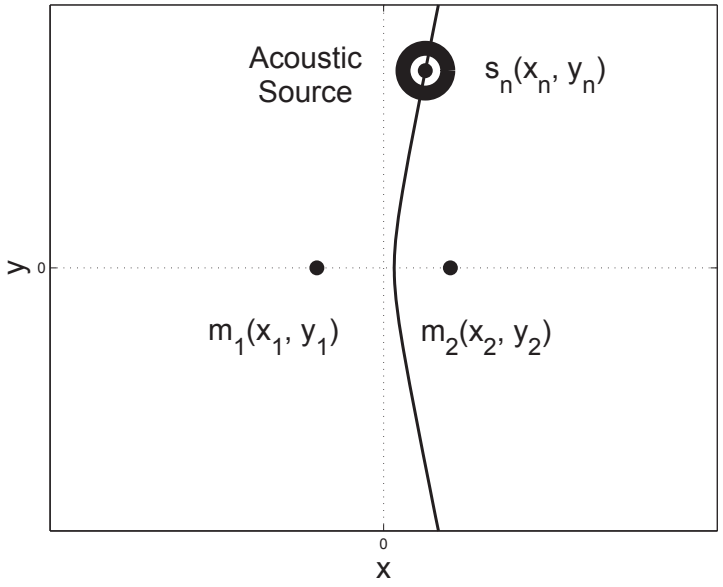

Figure 2.3: *The hyperbola that generates the same TDOA between two microphones in a near-field environment.*

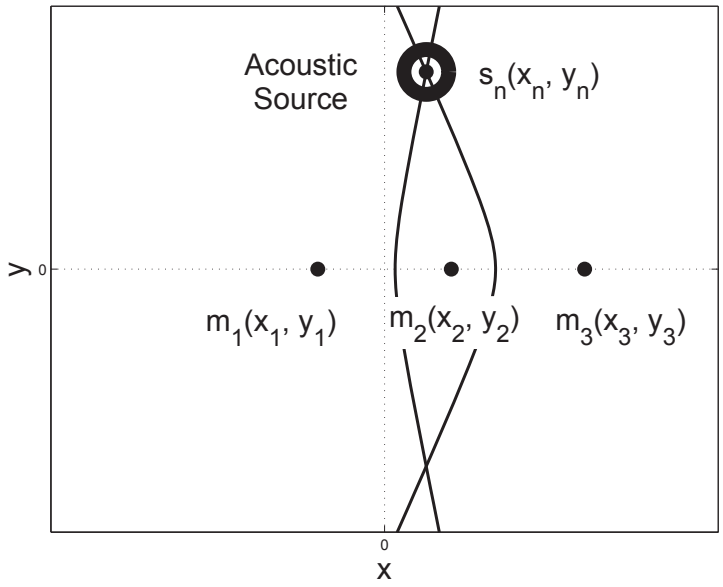

Figure 2.4: *Two hyperbolas between three microphones in a near-field environment.*

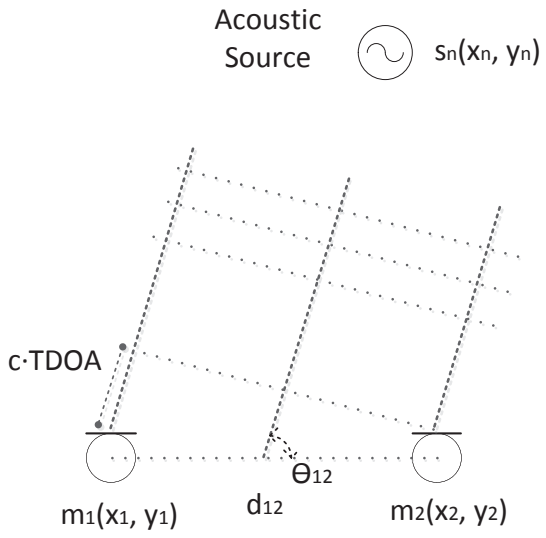

Figure 2.5: *Plan case of two microphones in far-field environment.*

relationship between DOA and TDOA is easily solved, as shown in Figure 2.5

$$\theta_{12} = \arccos\left(\frac{\tau_{12}c}{d_{12}}\right). \tag{2.9}$$

In the far-field condition the hyperboloid, the locus of points that generates the same TDOA to a microphone pair, can be approximated with the cone whose vertex is located at the midpoint between the two microphones. Therefore, regardless of the ambiguity of the front-rear, in a far-field environment, we need at least two linear arrays for the plane case and at least three plane arrays (placed so that the distance between them permits the detection of changes in the plane wave source) in the three-dimensional case. In the near-field condition, we need at least a linear array of three microphones for the plane case and a plane array of four microphones for the three-dimensional case. These conditions apply to the assumptions made in (2.1), when only one source was present.

If multiple sources are concurrently active, we need to make further considerations. In fact, in some applications, situations arise for which we cannot assign unambiguously TDOAs or DOAs to the same source. The example in Figure 2.6 shows the case of two sources with a configuration of two arrays for the location of a plane. As we can see, the combination of incorrect angles leads to an incorrect position estimation. The two DOAs calculated by the two arrays can be combined following two different configurations: 1) $\theta_{1,1} - \theta_{2,1}$, $\theta_{1,2} - \theta_{2,2}$; 2) $\theta_{1,2} - \theta_{2,1}$, $\theta_{1,1} - \theta_{2,2}$. The first configuration implies

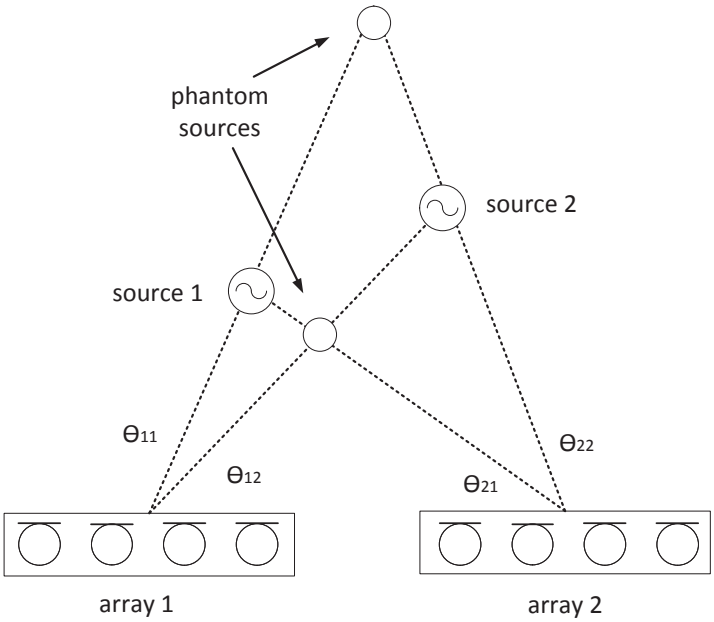

Figure 2.6: *The problem of multiple source localization*

the correct localization of the sound sources, whereas the second leads to an incorrect localization of both the sources. A solution to this problem of multiple source localization in far-field environment is proposed in section (6.2).

To locate a source, we can perform a direct accuracy estimation of the TDOAs using multi-channel signal processing with Time Delay Estimation (TDE) methods, or we can evaluate the changes that these TDOAs cause in the energy power output of a spatial filter, i.e., the so-called steered beamforming techniques.

Another important question to highlight is the phenomena of spatial aliasing (for a comprehensive dissertation, please refer to [Dmochowski *et al.*, 2009]). If we provide the distance between the microphones d and the wavelength λ of a sinusoidal wave, we find that if $\lambda/2 \leq d$, we have an ambiguity in the time delay estimation. Hence, the distance between the microphones determines the minimum frequency beyond which spatial aliasing can occur. Thus, a sound that does not contain spectral components below the minimum frequency cannot be uniquely localized. The condition in which spatial aliasing does not occur is

$$f_{max} \leq \frac{c}{2d}. \tag{2.10}$$

In general, for an array consisting of N microphones there are $N - 1$ independent TDOAs and

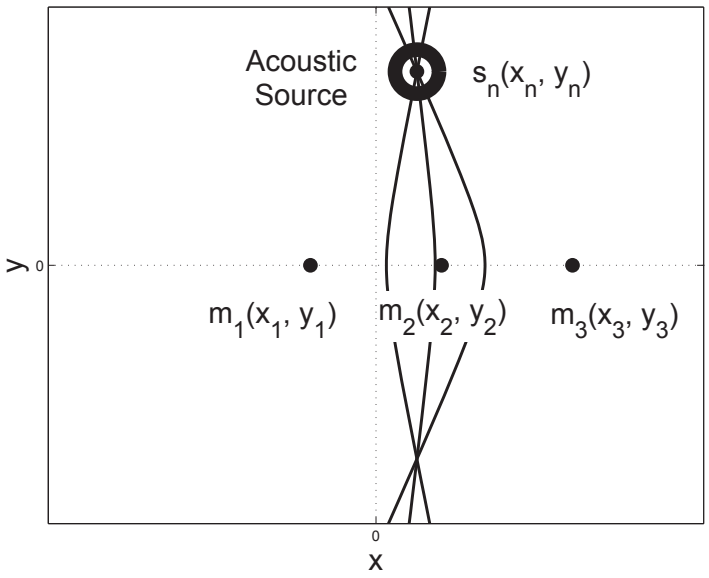

Figure 2.7: *Three hyperbola between three microphones in near-field environment.*

$N(N-1)/2$ total TDOAs. In fact, considering again the planar case of three microphones, in addition to τ_{12} and τ_{12}, we can estimate the time delay between microphones m_1 and m_3 (Figure 2.7), which follows the relationship

$$\tau_{12} + \tau_{23} = \tau_{13}. \tag{2.11}$$

Hence, in an ASL system, the use of multiple microphones permits the collocation of redundant information that can be used to achieve a more robust and accurate estimation of the source. In a real-world context, these estimations of TDOAs will be affected by error, and the system equation (2.6) is not solvable in the closed-form. We must provide methods that allow us to choose a solution that is as close as possible to the real value of the position by minimizing the measurement errors.

2.2 Source Localization

The literature offers two basic approaches to this solution. The first one (the indirect approach) is based on solving equations (2.6) and minimizing the error through the use of closed-form estimators and iterative maximum likelihood estimators. The second one (the direct approach) involves the search space by constructing a spatial analysis map and estimating, for each possible point of interest, the values that maximize a specific function that provides a coherent value from the entire system of arrays.

The spatial acoustic map is represented by spatial likelihood functions, which relate the space position with TDOAs functions.

Another important class of algorithms is the decentralized data fusion. In applications involving very large arrays, decentralized data fusion methods provide optimal fusing of source estimation measurements by two or more localization systems. The goal of these methods is to reduce the cost of computation and communication in a distributed sensor network.

It is often important to know the theoretical performance of an estimator. The Cramr-Rao Lower Bound (CRLB) is a theoretical lower bound of the variance that we can utilize for the unbiased estimation. It is useful to indicate the performance bounds of a particular algorithm. Thus, the estimator techniques are often compared with the CRLB to verify the performance. The CRBL was derived based on the accuracy of the estimator of the source location by [Bangs & Schultheiss, 1973].

2.2.1 Closed-Form Estimators

2.2.1.1 Plane Intersection (PI)

The Plane Intersection (PI) estimator is based on the consideration that the TDOAs of three sensors whose positions are known provide a plane of possible source locations in three-dimensional space [Schmidt, 1972]. Different planes obtained from different sensor triplets are intersected to find the source position. The equation (2.5) can be multiplied by $(r_i + r_j)$, and we obtain

$$(r_i + r_j)\delta_{ij} = (r_j - r_i)(r_i + r_j) = r_j^2 - r_i^2. \tag{2.12}$$

Substituting equation (2.2) in (2.12) and expanding, referring to the microphones m_i, m_j and m_k, we have

$$\begin{aligned} r_i + r_j &= \frac{2x_n(x_i - x_j) + 2y_n(y_i - y_j) + 2z_n(z_i - z_j) + a_j^2 - a_i^2}{\delta_{ij}} \\ r_j + r_k &= \frac{2x_n(x_j - x_k) + 2y_n(y_j - y_k) + 2z_n(z_j - z_k) + a_k^2 - a_j^2}{\delta_{jk}} \\ r_k + r_i &= \frac{2x_n(x_k - x_i) + 2y_n(y_k - y_i) + 2z_n(z_k - z_i) + a_i^2 - a_k^2}{\delta_{ki}} \end{aligned} \tag{2.13}$$

where $a_m^2 = x_m^2 + y_m^2 + z_m^2$ $(m = i, j, k)$.

The equation of the plane, considering the three equations (2.13), can be written as

$$A_{ijk}x + B_{ijk}y + C_{ijk}z = D_{ijk} \tag{2.14}$$

where

$$A_{ijk} = (x_i\delta_{jk} + x_j\delta_{ki} + x_k\delta_{ij})$$
$$B_{ijk} = (y_i\delta_{jk} + y_j\delta_{ki} + y_k\delta_{ij})$$
$$C_{ijk} = (z_i\delta_{jk} + z_j\delta_{ki} + z_k\delta_{ij})$$
$$D_{ijk} = \frac{1}{2}(\delta_{ij}\delta_{jk}\delta_{ik} + a_i^2\delta_{jk} + a_j^2\delta_{ki} + a_k^2\delta_{ij}).$$

Considering M microphones, the sensor triplet combinations that we can obtain are $M!/6(M-3)$, and we note that the solution requires that $M > 3$. In fact, we know that in three dimensions, at least four (non-coplanar) sensors are required to provide a solution. The set of $M!/6(M-3)$ equations can be written in matrix notation as

$$\mathbf{LP} = \mathbf{D} \tag{2.15}$$

where

$$\mathbf{L} = \begin{bmatrix} A_{123} & B_{123} & \dots & C_{123} \\ \vdots & \vdots & \ddots & \vdots \\ A_{ijk} & B_{ijk} & \dots & C_{ijk} \end{bmatrix}$$
$$\mathbf{P} = [x\ y\ z]^T$$
$$\mathbf{D} = [D_{123} \dots D_{ijk}]^T.$$

The solution of the linear equations (2.15) is obtained using a linear Least Square (LS) estimation. The source position estimation $\widehat{\mathbf{s}}_n$ is calculated by solving the quadratic minimization problem

$$\widehat{\mathbf{s}}_n = \underset{\mathbf{P}}{\operatorname{argmin}} ||\mathbf{D} - \mathbf{LP}||^2. \tag{2.16}$$

This minimization problem has a solution given by the normal equation

$$\widehat{\mathbf{s}}_n = \mathbf{P}^+\mathbf{D} \tag{2.17}$$

where $\mathbf{P}^+ = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T$ is the pseudo-inverse of $\mathbf{P}$. In the case of two-dimensional space, the problem is solved using the intersection of lines.

2.2.1.2 Spherical Intersection (SX)

The Spherical Intersection (SX) estimator is based on the range error (the equation error) [Schau & Robinson, 1987]. It assumes that the source position is given by the intersection of spheres, which define surfaces of constant distance from a single sensor. The point of intersection of two hyperboloids can move considerably for a relatively small change in eccentricity of one of the hyperboloids. This

is not true for intersecting spheres when the radius of one of the spheres is changed. Thus, the SX estimator aims to reduce the numerical difficulties associated with intersecting hyperboloids.

We first map the spatial Cartesian origin to an arbitrary sensor, which considers the reference one m_1, according to $r_1 = r_s = ||\mathbf{s}_n||$. In this manner, we can write the following set of $M-1$ equations in matrix notation by considering the geometrical relationships and introducing the so-called error equation, which assumes that the TDOAs are typically non-precisely measured

$$\boldsymbol{\epsilon} = \boldsymbol{\Lambda} - 2r_s\boldsymbol{\Delta} - 2\mathbf{M}\mathbf{s}_n \tag{2.18}$$

where $\boldsymbol{\epsilon}$ contains the $M-1$ equation errors that has to be minimized and

$$\boldsymbol{\Lambda} = \begin{bmatrix} r_2^2 - \delta_{21}^2 \\ r_3^2 - \delta_{31}^2 \\ \vdots \\ r_M^2 - \delta_{M1}^2 \end{bmatrix}$$
$$\boldsymbol{\Delta} = [\delta_{21}\ \delta_{31}\ \ldots\ \delta_{M1}]^T$$
$$\mathbf{M} = \begin{bmatrix} x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ \vdots & \vdots & \vdots \\ x_M & y_M & z_M \end{bmatrix}.$$

The LS solution for $\mathbf{s}_n$ given r_s is

$$\widehat{\mathbf{s}}_n = \frac{1}{2}\mathbf{M}^+(\boldsymbol{\Lambda} - 2r_s\boldsymbol{\Delta}). \tag{2.19}$$

The source range r_s is unknown. To solve this problem, the SX estimator proposes to substitute the LS solution (2.19) for $\mathbf{s}_n$, given r_s into the quadratic equation constraint

$$r_s^2 = \mathbf{s}_n^T\mathbf{s}_n. \tag{2.20}$$

Substituting equation (2.19) into equation (2.20) yields after expansion

$$ar_s^2 + br_s + c = 0 \tag{2.21}$$

where

$$a = 4 - 4\boldsymbol{\Delta}^T(\mathbf{M}^+)^T\mathbf{M}^+\boldsymbol{\Delta}$$
$$b = 2\boldsymbol{\Delta}^T(\mathbf{M}^+)^T\mathbf{M}^+\boldsymbol{\Lambda} + 2\boldsymbol{\Lambda}^T(\mathbf{M}^+)^T\mathbf{M}^+\boldsymbol{\Delta}$$
$$c = -\boldsymbol{\Lambda}^T(\mathbf{M}^+)^T\mathbf{M}^+\boldsymbol{\Lambda}.$$

The two solutions to the quadratic equation (2.21) are

$$\widehat{r}_s = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{2.22}$$

The positive and real root is taken as an estimate of the distance between the source and the reference microphone (the origin of the Cartesian coordinate system). This value $\widehat{r}_s$ is then substituted in to (2.19) to obtain the source location estimation $\widehat{\mathbf{s}}_n$. We need to consider that the SX method requires the estimation of r_s. Thus, if the solution of the quadratic equation is not positive and real, the source localization solution does not exist. Moreover, if the solutions are both real and positive, the estimation is not unique.

2.2.1.3 Spherical Interpolation (SI)

The Spherical Interpolation (SI) [Smith & Abel, 1987] aims to solve the LS problem of equation (2.19) in a more robust and accurate manner with respect to the SX estimator. The basic idea of this closed-form solution is to substitute (2.19) into (2.18) and to minimize the new equation error obtained, for this case, with respect to r_s. After the substitution, the new equation error becomes

$$\boldsymbol{\epsilon}' = (\mathbf{I} - \mathbf{M}\mathbf{M}^{+})(\boldsymbol{\Lambda} - 2r_s\boldsymbol{\Delta}) = (\mathbf{I} - \mathbf{P}_s)(\boldsymbol{\Lambda} - 2r_s\boldsymbol{\Delta}) \tag{2.23}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{P}_s$ is an idempotent ($\mathbf{P}_s^2 = \mathbf{P}_s$) projection matrix. The orthogonal projection matrix $\mathbf{P}_s^{\perp}$ removes components in the space spanned by the columns of $\mathbf{M}$ and it is defined by

$$\mathbf{P}_s^{\perp} = \mathbf{I} - \mathbf{P}_s = \mathbf{I} - \mathbf{M}\mathbf{M}^{+}. \tag{2.24}$$

By considering the symmetric $\mathbf{P}_s^{\perp} = \mathbf{P}_s^{\perp T}$, to determine the weighted LS solution in which the weighting matrix is $\mathbf{P}_s^{\perp}\mathbf{V}\mathbf{P}_s^{\perp}$ and $\mathbf{V}$ is a positive-definite matrix, we write

$$\boldsymbol{\epsilon}'^{T}\mathbf{V}\boldsymbol{\epsilon}' = \mathbf{P}_s^{\perp}(\boldsymbol{\Lambda} - 2r_s\boldsymbol{\Delta})^{T}\mathbf{V}\mathbf{P}_s^{\perp}(\boldsymbol{\Lambda} - 2r_s\boldsymbol{\Delta}) \tag{2.25}$$

and the solution is given by

$$\widehat{r}_s = \frac{\boldsymbol{\Delta}^{T}\mathbf{P}_s^{\perp}\mathbf{V}\mathbf{P}_s^{\perp}\boldsymbol{\Lambda}}{2\boldsymbol{\Delta}^{T}\mathbf{P}_s^{\perp}\mathbf{V}\mathbf{P}_s^{\perp}\boldsymbol{\Delta}}. \tag{2.26}$$

Substituting this solution into (2.19) allows us to estimate the source position $\widehat{\mathbf{s}}_n$.

In [Stoica & Li, 2006], the authors clarify and streamline the SI method, introducing the Unconstrained Least Squares (ULS) estimator and demonstrating that is identical to the SI estimator. The solution is simplest than basic SI and it is obtained without the unnecessarily complicated second stage: the substitution of (2.19) into (2.18). The ULS criterion is written from (2.18) as

$$\widehat{\mathbf{y}}(\mathbf{s}_n) = \underset{\mathbf{y}(\mathbf{s}_n)}{\operatorname{argmin}} ||\boldsymbol{\Lambda} - \boldsymbol{\Phi}\mathbf{y}(\mathbf{s}_n)||^2 \tag{2.27}$$

where $\mathbf{\Phi} = [2\mathbf{\Delta} \quad 2\mathbf{M}]$ and $\mathbf{y}(\mathbf{s}_n) = [r_s, \mathbf{s}_n^T]^T$. This unconstrained minimization, not considering the dependence $r_s = ||\mathbf{s}_n||$, is

$$\widehat{\mathbf{y}}(\mathbf{s}_n) = \mathbf{\Phi}^+ \mathbf{\Lambda}. \tag{2.28}$$

The corresponding ULS estimate $\widehat{\mathbf{s}}_n$ is given by

$$\widehat{\mathbf{s}}_n = [0 \quad \mathbf{I}]\widehat{\mathbf{y}}(\mathbf{s}_n) \tag{2.29}$$

where 0 is a column vector of zeros and $\mathbf{I}$ is a 3×3 identity matrix.

2.2.1.4 Hyperbolic Intersection (HI)

The Hyperbolic Intersection (HI) method resolves the nonlinear intersection of hyperbolic curves by dividing the procedure into two linear LS steps [Chan & Ho, 1994]. By introducing an intermediate variable, the nonlinear equations relating TDOA estimates and source position can be transformed into a set of equations which are linear in the unknown parameters and the intermediate variable. It assumes an arbitrary microphone, which is the reference $\mathbf{m}_1 = [x_1\, y_1\, z_1]^T$, and it does not require the placement of the referenced microphone at the origin of the Cartesian system. We introduce the vector p defined as

$$\mathbf{p} = [\mathbf{s}_n^T, r_s]^T. \tag{2.30}$$

We can express the (2.18) by considering the (2.30) and we obtain

$$\boldsymbol{\epsilon} = \mathbf{h} - \mathbf{G}\mathbf{p} \tag{2.31}$$

where

$$\mathbf{h} = \frac{1}{2}\begin{bmatrix} \delta_{21}^2 - (x_2 + y_2 + z_2)^2 + (x_1 + y_1 + z_1)^2 \\ \delta_{31}^2 - (x_3 + y_3 + z_3)^2 + (x_1 + y_1 + z_1)^2 \\ \vdots \\ \delta_{M1}^2 - (x_M + y_M + z_M)^2 + (x_1 + y_1 + z_1)^2 \end{bmatrix}$$

$$\mathbf{G} = -\begin{bmatrix} x_2 & y_2 & z_2 & \delta_{21} \\ x_2 & y_3 & z_3 & \delta_{31} \\ \vdots & \vdots & \vdots & \vdots \\ x_M & y_M & z_M & \delta_{M1} \end{bmatrix}.$$

In the first step, the HI estimator assumes that no relationship exists between the variables $\mathbf{s}_n$ and r_s, and, consequently, (2.31) can be solved with the generalized LS solution

$$\widehat{\mathbf{p}} = (\mathbf{G}^T \mathbf{\Psi}^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{\Psi}^{-1} \mathbf{h} \tag{2.32}$$

where $\mathbf{\Psi}$ is the covariance matrix of $\boldsymbol{\epsilon}$. Assuming $\boldsymbol{\epsilon}$ is a Gaussian random vector, the evaluation of $\mathbf{\Psi}$ becomes

$$\mathbf{\Psi} = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = c^2\mathbf{BQB} \tag{2.33}$$

where $E[\cdot]$ denotes mathematical expectation, $\mathbf{B} = \text{diag}(r_2, r_3, \ldots, r_M)$ and $\mathbf{Q} = E[\mathbf{\Delta}\mathbf{\Delta}^T]$ is the covariance matrix of the vector $\mathbf{\Delta} = [\delta_{21}\ \delta_{31}\ \ldots\ \delta_{M1}]^T$, which contains the range differences. In the second step of the HI algorithm, it has been assumed that the correct relationship of dependence is

$$r_i = ||\mathbf{s}_n - \mathbf{m}_i||. \tag{2.34}$$

To include this condition, the HI considers that TDOAs are affected by small noise and that the vector $\mathbf{p}$ is a random vector with its mean centered at the real value and covariance matrix, defined by

$$\text{cov}(\mathbf{p}) = (\mathbf{G}^T\mathbf{\Psi}^{-1}\mathbf{G})^{-1}. \tag{2.35}$$

The error equations can then be written as

$$\boldsymbol{\epsilon}' = \mathbf{h}' - \mathbf{G}'\mathbf{p}' \tag{2.36}$$

where

$$\mathbf{h}' = \begin{bmatrix} (\widehat{\mathbf{p}}(1) - x)^2 \\ (\widehat{\mathbf{p}}(2) - y)^2 \\ (\widehat{\mathbf{p}}(3) - z)^2 \\ (\widehat{\mathbf{p}}(4)^2 \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\mathbf{p}' = \begin{bmatrix} (x_n - x_1)^2 \\ (y_n - y_1)^2 \\ (z_n - z_1)^2 \end{bmatrix}.$$

The generalized LS solution of (2.36) is

$$\widehat{\mathbf{p}}' = (\mathbf{G}'^T\mathbf{\Psi}'^{-1}\mathbf{G}')^{-1}\mathbf{G}'^T\mathbf{\Psi}'^{-1}\mathbf{h}' \tag{2.37}$$

where the covariance matrix $\mathbf{\Psi}'$, considering $\mathbf{B}' = \text{diag}\{(x_n - x_1), (y_n - y_1), (z_n - z_1), r_1\}$, becomes

$$\mathbf{\Psi}' = E\{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}'^T\} = 4\mathbf{B}'\text{cov}(\mathbf{p})\mathbf{B}'. \tag{2.38}$$

The final estimated position is obtained from $\widehat{\mathbf{p}}'$, selecting the solution that lies in the region of interest

$$\widehat{\mathbf{s}}'_\mathbf{n} = \pm\sqrt{\widehat{\mathbf{p}}'} + \mathbf{m}_i. \tag{2.39}$$

2.2.1.5 Linear Intersection (LI)

The Linear Intersection (LI) technique [Brandstein *et al.*, 1997] assumes far-field acoustical propagation. Therefore, the hyperboloid is well-approximated by a cone with its vertex at the midpoint between the sensor and the axis of symmetry along the line that combines the sensors. The array is limited to four elements, configured in two centered orthogonal pairs. By considering half-space, this array permits an estimate of the DOA of the source. The DOA is represented by a bearing line pointing to the source. Given a network of $M/2$ microphone pairs in which there are S sub-arrays arranged in mutually-orthogonal and mutually-bisecting sensor quadruples, the bearing line of the generic i_{th} sub-array can be written as

$$\mathbf{l}_i = r_i \mathbf{a}_i + \mathbf{m}_i \quad i = 1, \ldots, S \tag{2.40}$$

where $\mathbf{a}_i$ represents the rotated direction cosine vector and r_i is the range of a point on the line from the local origin (the common midpoint of the two microphones in the i_{th} sub-array) at microphone $\mathbf{m}_i$.

The approach proposed by the LI estimator calculates a number of potential source locations from the points of closest intersection for all pairs of bearing lines and uses a weighted average of these locations to generate a final source position estimation. The shortest distance between the two lines $\mathbf{l}_i$ and $\mathbf{l}_j$ is measured along a line that is parallel to their common normal and is given by

$$\varsigma_{ij} = \frac{|(\mathbf{a}_i \times \mathbf{a}_j)(\mathbf{m}_i - \mathbf{m}_j)|}{|\mathbf{a}_i \times \mathbf{a}_j|} \tag{2.41}$$

where $\times$ denotes vector product. The locale ranges are found by subtracting the two bearing vectors and solving the following over-constrained matrix equation

$$\mathbf{l}_j - \mathbf{l}_i = \varsigma_{ij}(\mathbf{a}_i \times \mathbf{a}_j) = r_j \mathbf{a}_j + \mathbf{m}_i - r_i \mathbf{a}_i - \mathbf{m}_i. \tag{2.42}$$

The potential source locations $\widehat{\mathbf{s}}_{\mathbf{ij}} = \mathbf{l}_i$ and $\widehat{\mathbf{s}}_{\mathbf{ji}} = \mathbf{l}_j$ are calculated by substituting the ranges r_i and r_i from (2.42) into (2.40). The final location estimate is then calculated as the weighted average of the potential source locations

$$\widehat{\mathbf{s}}_{\mathbf{n}} = \frac{\sum_{i=1}^{S} \sum_{j=1(j\neq i)}^{S} W_{ij} \widehat{\mathbf{s}}_{\mathbf{ij}}}{\sum_{i=1}^{S} \sum_{j=1(j\neq i)}^{S} W_{ij}} \tag{2.43}$$

where W_{ij} is the weight associated with the potential source location, by assuming that the TDOAs are independent and normally distributed, with the mean given by the estimate itself

$$W_{ij} = \prod_{k=1}^{M/2} P(\widehat{\tau}_k, \tau_k, \sigma_k^2) \tag{2.44}$$

where $P(x, m, \sigma^2)$ is the value of a Gaussian distribution with mean m and variance σ^2 evaluated at x

$$P(x, m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(\frac{-(x-m)^2}{2\sigma^2}\Big). \tag{2.45}$$

2.2.1.6 Linear Correction (LC)

Linear Correction (LC) is a constrained LS method [Huang *et al.*, 2001]. It introduces one supplemental variable, the source range r_s, in addition to the source location coordinates. LC constructs a linear error function with vector (2.30), rewriting the spherical error equation (2.18). The follow error equation in which the first microphone is regarded as the reference and is placed at the origin of the coordinate system, is

$$\boldsymbol{\epsilon} = \mathbf{H}\mathbf{p} - \mathbf{b} \tag{2.46}$$

where $\mathbf{H} = -\mathbf{G}$ from (2.31), and $\mathbf{b} = -\mathbf{h}$ from (2.31) considering the placement of the referenced microphone at the origin of the Cartesian system

$$\mathbf{b} = \frac{1}{2}\begin{bmatrix} r_2^2 - \delta_{21}^2 \\ r_3^2 - \delta_{31}^2 \\ \vdots \\ r_M^2 - \delta_{M1}^2 \end{bmatrix}.$$

The dependence among the elements of $\mathbf{p}$ can be described as a quadratic constraint

$$\mathbf{p}^T\boldsymbol{\Sigma}\mathbf{p} = 0 \tag{2.47}$$

where $\boldsymbol{\Sigma} = \text{diag}(1, 1, 1, -1)$ is a diagonal and orthonormal matrix.

The technique of Lagrange multipliers is used to solve the constrained LS problem (2.46), and the source location is determined by minimizing the Lagrangian model

$$\mathcal{L}(\mathbf{p}, \lambda) = (\mathbf{H}\mathbf{p} - \mathbf{b})^T(\mathbf{H}\mathbf{p} - \mathbf{b}) + \lambda\mathbf{p}^T\boldsymbol{\Sigma}\mathbf{p} \tag{2.48}$$

where λ is the Lagrange multiplier. The conditions for minimizing (2.48) can be obtained by taking the gradient of $\mathcal{L}(\mathbf{p}, \lambda)$ with respect to $\mathbf{p}$ and equating the result to zero

$$\frac{\partial\mathcal{L}(\mathbf{p}, \lambda)}{\partial\mathbf{p}} = 2(\mathbf{H}^T\mathbf{H} + \lambda\boldsymbol{\Sigma})\mathbf{p} - 2\mathbf{H}^T\mathbf{b} = 0. \tag{2.49}$$

Resolving this equation, we obtain the estimated vector

$$\widehat{\mathbf{p}} = (\mathbf{H}^T\mathbf{H} + \lambda\boldsymbol{\Sigma})^{-1}\mathbf{H}^T\mathbf{b}. \tag{2.50}$$

The Lagrange multiplier λ is obtained by solving the constraint equation, by substituting (2.50) into (2.47)

$$f(\lambda) = (\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{H}^T\mathbf{b})^T(\boldsymbol{\Xi} + \lambda\mathbf{I})^{-2}\mathbf{U}^T\mathbf{H}^T\mathbf{b} \tag{2.51}$$

where $\mathbf{U}$ and $\boldsymbol{\Xi}$ are the results of the eigenvalue decomposition of

$$\mathbf{H}\mathbf{H}^T\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Xi}\mathbf{U}^{-1}. \tag{2.52}$$

The matrix $\mathbf{U}$ is the matrix in which each column is the eigenvector and $\mathbf{\Xi}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues of $\mathbf{H}\mathbf{H}^T\mathbf{\Sigma}$. The Lagrange multiplier is found using iterative methods, searching for the root of the function (2.51) around zero. Typically, four interactions are used to obtain an appropriate value of λ. However, the solution for λ is not unique, so the LC estimator requires two steps [Huang *et al.*, 2001].

In [Stoica & Li, 2006], the authors propose to call the algorithm described the Constrained Least Squares (CLS) method because the two-step LC estimator is identical to the CLS estimator.

However, in [Huang *et al.*, 2001], the CLS estimator is used to correct the solution given by the ULS method. The first step aims to resolve an unconstrained global LS problem by considering that the relationship between x_n, y_n, z_n and r_s is mutually independent. Then, the LS criterion is given by

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{H}\mathbf{p} - \mathbf{b})^T\mathbf{H}\mathbf{p} - \mathbf{b} \tag{2.53}$$

and the LS solution minimizing (2.53) for $\mathbf{p}$ is

$$\widehat{\mathbf{p}}_\mathbf{1} = \mathbf{H}^+\mathbf{b}. \tag{2.54}$$

The solution $\widehat{\mathbf{p}}_\mathbf{1}$ is the same as the spherical interpolation (SI) estimate, but with less computational complexity.

The second step aims to correct the $\widehat{\mathbf{p}}_\mathbf{1}$ with a better estimation $\widehat{\mathbf{p}}_\mathbf{2}$ using the constraint (2.47), resolved by Lagrangian multipliers. The final solution of LC estimator becomes

$$\widehat{\mathbf{p}}_\mathbf{2} = (\mathbf{I} + \lambda(\mathbf{H}^T\mathbf{H})^{-1})^{-1}\widehat{\mathbf{p}}_\mathbf{1}. \tag{2.55}$$

2.2.1.7 Gillette-Silverman (GS)

The method proposed in [Gillette & Silverman, 2008] is a linear closed-form algorithm that considers more microphones as references. The method is very simple and uses a minimum of five microphones in three dimensions. Considering M microphones and m_1 as the reference, according to $r_1 = r_s$, the GS algorithm is based on setting the following equation

$$r_m^2 - r_1^2 = ||\mathbf{s_n} - \mathbf{r_m}||^2 - ||\mathbf{s_n} - \mathbf{r_1}||^2. \tag{2.56}$$

Inserting equation $r_m = \delta_{m1} + r_1$ and expanding it, we obtain a linear equation

$$\delta_{m1}r_1 - (x_m - x_1)x_n - (y_m - y_1)y_n - (z_m - z_1)z_n = w_{m1} \tag{2.57}$$

where $w_{m1} = 1/2(\delta_{m1}^2 - x_m^2 + x_1^2 - y_m^2 + y_1^2 - z_m^2 + z_1^2)$. In matrix notation we can write

$$\mathbf{\Gamma}\mathbf{p} = \mathbf{w} \tag{2.58}$$

where

$$\mathbf{\Gamma} = \begin{bmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 & \delta_{21} \\ x_1 - x_3 & y_1 - y_3 & z_1 - z_3 & \delta_{31} \\ \vdots & \vdots & \vdots & \vdots \\ x_1 - x_M & y_1 - y_M & z_1 - z_M & \delta_{M1} \end{bmatrix}$$

$$\mathbf{g} = [x_n \; y_n \; z_n \; r_1]^T$$

$$\mathbf{w} = [w_{21} \; w_{31} \cdots w_{M1}]^T.$$

The linear LS solution is given by

$$\widehat{\mathbf{g}} = \mathbf{w}\mathbf{\Gamma}^{+}. \tag{2.59}$$

We note that if the matrix $\mathbf{\Gamma}$ is singular, it is not possible to solve the linear system. This is the case for a line of microphones with uniform spacing. Therefore, this method requires an array of microphones with random spacing, in which the matrix is virtually always nonsingular. To improve the performance and to use redundancy information between microphones, the GS method can be generalized with M_r microphones as references. For example, if the two reference microphones are m_1 and m_6 in an array of six microphones, we obtain the follow matrix

$$\mathbf{\Gamma} = \begin{bmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 & \delta_{21} & 0 \\ x_1 - x_3 & y_1 - y_3 & z_1 - z_3 & \delta_{31} & 0 \\ x_1 - x_4 & y_1 - y_4 & z_1 - z_4 & \delta_{41} & 0 \\ x_1 - x_5 & y_1 - y_5 & z_1 - z_5 & \delta_{51} & 0 \\ x_6 - x_2 & y_6 - y_2 & z_6 - z_2 & 0 & \delta_{26} \\ x_6 - x_3 & y_6 - y_3 & z_6 - z_3 & 0 & \delta_{36} \\ x_6 - x_4 & y_6 - y_4 & z_6 - z_4 & 0 & \delta_{46} \\ x_6 - x_5 & y_6 - y_5 & z_6 - z_5 & 0 & \delta_{56} \end{bmatrix}$$

$$\mathbf{g} = [x_n \; y_n \; z_n \; r_1 \; r_6]^T$$

$$\mathbf{w} = [w_{21} \; w_{31} \; w_{41} \; w_{51} \; w_{26} \; w_{36} \; w_{46} \; w_{56}]^T.$$

Hence, GS does not consider the TDOAs between the referenced microphones.

2.2.2 Iterative Maximum Likelihood Estimators

The maximum likelihood estimators are a class of algorithms that require only iterative methods. This approach involves a considerable computational cost, however, with the advantage of accuracy estimation. In contrast to equation (2.18), the range differences are modeled by considering the measurement errors

$$\epsilon_{ij} = \delta_{ij} + r_i - r_j. \tag{2.60}$$

Given the set of D ranges difference estimate $\widehat{\boldsymbol{\delta}} = [\widehat{\delta}_1\ \widehat{\delta}_2 \dots \widehat{\delta}_D]^T$, errors $\boldsymbol{\epsilon} = [\epsilon_1\ \epsilon_2 \dots \epsilon_D]^T$ can be modeled as a multivariate Gaussian random variable $\boldsymbol{\epsilon} \sim N(0; \mathbf{Q})$ with zero mean, and the related covariance matrix $\mathbf{Q}$ independent from $\mathbf{s}_n$, can be defined using the Probability Density Function (PDF) as

$$p(\widehat{\boldsymbol{\delta}}, \mathbf{s}_n) = \frac{\exp\Big(-\frac{1}{2}(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^T \mathbf{Q}^{-1} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})\Big)}{\sqrt{(2\pi)^D \det(\mathbf{Q})}}. \tag{2.61}$$

The principle of maximum likelihood aims to find the distribution that yields the highest possible probability of the likelihood function $L(\mathbf{s}_n, \boldsymbol{\delta})$

$$\widehat{\mathbf{s}}_\mathbf{n} = \underset{\mathbf{s}_n}{\operatorname{argmax}} L(\mathbf{s}_n, \boldsymbol{\delta}) \tag{2.62}$$

where $L(\mathbf{s}_n, \boldsymbol{\delta}) = p(\widehat{\boldsymbol{\delta}}, \mathbf{s}_n)$. The log likelihood function is the logarithm of the likelihood function. Because the logarithm is a monotonic, strictly increasing function, the maximum of the log likelihood is precisely equivalent to the maximum of the likelihood, or the minimum of the negative log likelihood

$$\widehat{\mathbf{s}}_\mathbf{n} = \underset{\mathbf{s}_n}{\operatorname{argmin}} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta})^T \mathbf{Q}^{-1} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}). \tag{2.63}$$

Direct estimation of the minimizer is generally not practical. To solve this minimization problem, iterative techniques can be used. This approach requires a computational cost that could be a problem for real-time application. Different maximum likelihood estimators approaches are proposed in [Hahn & Tretter, 1973] [Wax & Kailath, 1983] [Stoica & Nehorai, 1990] [Segal *et al.*, 1991] [Chen *et al.*, 2002] [Georgiou & Kyriakakis, 2006] [Destino & Abreu, 2011].

2.2.3 Spatial Likelihood Functions

The approach based on the spatial likelihood function (as has termed by [Aarabi, 2003]) searches for the estimate position directly by scanning the whole region of interest. Given the vector $\mathbf{s} = [x\ y\ z]^T$ of space, we can define a spatial function likelihood $\mathcal{S}[\mathbf{s}, p(\tau)]$ that links a point in space with the function $p(\tau)$ depending on the TDOA between microphones (time delay estimation or the output of a steered response beamforming). Hence, the position of the source is estimated by picking the maximum peak

$$\widehat{\mathbf{s}}_\mathbf{n} = \underset{\mathbf{s}}{\operatorname{argmax}} \mathcal{S}[\mathbf{s}, p(\tau)]. \tag{2.64}$$

Several methods are proposed to define the spatial function. Considering a network of R arrays, we can obtain the R function $p(\tau_r)$. The Global Coherence Field (GCF) [Omologo & DeMori, 1998], more

often referred to as the SRP-PHAT [DiBiase *et al.*, 2001], is based on applying a coherence sum. In general, it uses a pair of microphone arrays but is not limited to one pair. The GCF results

$$\mathfrak{S}[\mathbf{s}, p(\tau)]_{sum} = \sum_{r=1}^{R} p(\tau_{r,\mathbf{s}}). \tag{2.65}$$

The GCF is the optimal solution for a single source; when multiple sources are active the estimation of the second peak is a problem, because the multiple intersections of the hyperboloid causes more peaks to appear (see figure 2.6) due to constructive interferences. A solution in the case of multiple sources in near-field environment can be found here [Brutti *et al.*, 2010].

The method for the multiplication can be found in [Ward *et al.*, 2003]. The product is used to reduce the ghost source due to multiple peaks. Whereas the summation represents a sum of a set of equations, the multiplication represents the intersection of the sets.

$$\mathfrak{S}[\mathbf{s}, p(\tau)]_{mult} = \prod_{r=1}^{R} p(\tau_{r,\mathbf{s}}). \tag{2.66}$$

In [Pertilä *et al.*, 2008], the authors proposed the use of the Hamacher T-norm, which is close to multiplication and uses the intersection of sets. The generalized multiplication becomes

$$h(a, b, \gamma) = \frac{ab}{\gamma + (1 - \gamma)(a + b - ab)}. \tag{2.67}$$

When $\gamma = 1$ we have multiplication. The spatial function becomes

$$\mathfrak{S}[\mathbf{s}, p(\tau)]_{t-norm} = h(h(p(\tau_{1,\mathbf{s}}), p(\tau_{2,\mathbf{s}}), \gamma), \ldots, \tau_{R,\mathbf{s}}, \gamma). \tag{2.68}$$

In [Pertilä *et al.*, 2008], the results indicate that the intersection methods provide the best results under different SNRs and reverberation conditions when using a Particle filter.

2.2.4 Decentralized Data Fusion

In the applications involving very large arrays (a distributed sensor network), the previously described methods for the source location have a significant increase in computation and communication costs. Therefore, the data processing may need to be decentralized. Decentralized data fusion methods provide optimal fusing of source estimation measurements by two or more localization systems.

In the case of the far-field environment, two methods for decentralized array processing based on the maximum likelihood estimation of the DOA are proposed in [Stoica *et al.*, 1995]. In [Liu *et al.*, 2003], the authors present an approach based on collaborative signal processing, focusing on a vehicle tracking application using Bayesian filtering. The goal of collaboration is to select embedded sensors

to participate in estimation and to extract useful information with minimal resource usage. The aim of work proposed in [Kozick & Sadler, 2003] is to reduce the communication bandwidth with a central processing node, by modeling the wavefronts with perfect spatial coherence over individual arrays and frequency-selective coherence between distinct arrays, and modeling the sensor signals as Gaussian random processes. In [Chen *et al.*, 2004], a decentralized dynamic clustering algorithm for source tracking in wireless sensor networks is proposed. The problem of source tracking is also addressed in [He & Chong, 2004]. The method is based on the combination of Particle filtering for belief-state estimation and sampling-based Q-value approximation for lookahead, to determine which sensors to activate over time to trade off tracking performance with sensor usage costs.

In the near-field environment, a two-step decentralized data fusion solution is described in [Prandi *et al.*, 2008]. It solves the source localization by subdividing the problem between two or more groups of acoustic sensors. In the first step, each array provides an estimate of the source position by measuring the TDOAs between each microphone pair in the array. In the second step, these estimates are optimally fused taking into account the geometry of the arrays to provide the final source position estimation.

2.3 Summary

The localization problem and the methods for the source position estimation have been presented. Two approaches to the localization solution can be used. The indirect approach is based on solving equations of the intersection of hyperboloids, minimizing the error through the use of closed-form estimators and iterative maximum likelihood estimators. The direct approach involves the search space by constructing a spatial analysis map (represented by spatial likelihood functions) and estimating the values that maximize a specific function; this function provides a coherent value from the entire system of arrays. The most successful algorithms for indirect closed-form methods are the ULS estimator (SI estimator) and the CLS estimator (LC estimator), whereas other methods require a special configuration of the array: four elements configured in two centered orthogonal pairs for the LI estimator, and an array of microphones with random spacing for the GS estimator. The indirect iterative maximum likelihood estimators have the advantage of accuracy estimation, but this approach involves a considerable computational cost. On the other side, the most widely used direct approach is the GCF, more often referred to as the SRP-PHAT. It is based on applying a coherence sum to construct the spatial analysis map. A brief description of the decentralized data fusion for the solution of the localization problem involving a large array, generally referred to as distributed sensor network, closed this chapter.

3

Signal Processing for Sound Localization

3.1 Signal Model

We assume N acoustic sources and R arrays, each composed of M microphones, and consider the omnidirectional characteristics of both the sources and the microphones. We will refer to the model of discrete-time obtained by performing a sampling operation on the continuous-time signal $x(t)$ with a uniform sampling period T_s. A discrete-time signal is expressed by

$$x(kT_S) = x(k/f_s) \quad k = 0, 1, \ldots \tag{3.1}$$

where k is the sample time index and f_s is the sampling frequency. As usual, we will allow the sample period T_s to remain implicit and refer to it simply as $x(k)$. In the frequency domain the signal $x(k)$ is obtained in sampling frequency by the Discrete Fourier Transform (DFT)

$$X(f) = \sum_{k=0}^{L-1} x(k) e^{\frac{-2\pi j f k}{L}} \quad f = 0, 1, \ldots, L-1 \tag{3.2}$$

where f is the frequency index and L is the number of samples of the observation time. The Inverse DFT (IDFT) has the form

$$x(k) = \frac{1}{L} \sum_{f=0}^{L-1} X(f) e^{\frac{2\pi j f k}{L}} \quad k = 0, 1, \ldots, L-1. \tag{3.3}$$

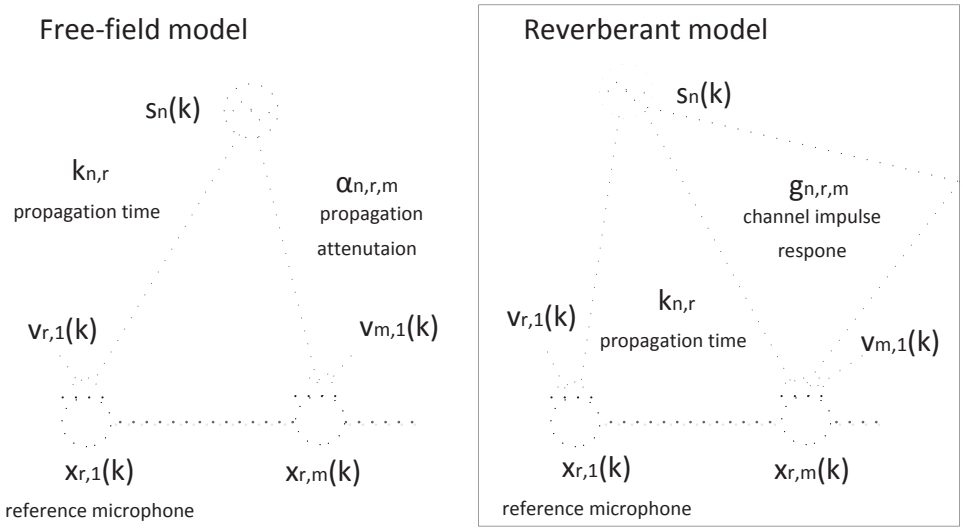

Figure 3.1: *Free-field and reverberant signal model.*

The discrete-time signal received by the m_{th} microphone of r_{th} the array can be modeled as

$$x_{r,m}(k) = \sum_{n=1}^{N} \alpha_{n,r,m} s_n(k - k_{n,r} - \tau_{n,r,m}) + v_{r,m}(k) \tag{3.4}$$

where $\alpha_{n,r,m}$ is the attenuation of the sound propagation (inversely proportional to the distance from source n to microphone m of array r), $s_n(k)$ are the unknown uncorrelated source signals, $k_{n,r}$ is the propagation time from the unknown source n to the reference sensor of array r, $\tau_{n,r,m}$ is the TDOA of the n_{th} signal between the m_{th} microphone and the reference of the r_{th} array, and $v_{r,m}(k)$ is the additive noise signal at the m_{th} sensor, assumed to be uncorrelated with not only all of the source signals but also with the noise observed at the other sensors.

This model, which contains only the direct paths, is appropriate to describe the free-field environment; however, in indoor space, we must introduce a variable that is able to describe the reverberant field and all of the reflections that have been added to the direct sound that reaches the microphones. Introducing the channel impulse response $g_{n,r,m}$ from the source n to microphone m of r, we can express the reverberant model as

$$x_{r,m}(k) = \sum_{n=1}^{N} g_{n,r,m} * s_n(k - k_{n,r} - \tau_{n,r,m}) + v_{r,m}(k) \tag{3.5}$$

where $*$ denotes convolution.

3.2 Time Delay Estimation Methods for Microphone Pair

3.2.1 Cross-Correlation (CC)

The Cross-Correlation (CC) is a measure of similarity of two signals. Given two wide-sense stationary signals x_1 and x_2, the CC is defined as

$$R_{x_1x_2}(p) = E[x_1(k)x_2(k+p)] \tag{3.6}$$

where $E[\cdot]$ denotes mathematical expectation. The relative time delay τ is obtained using an estimation of the maximum peak detection in the cross-correlation function

$$\widehat{\tau} = \underset{p}{\operatorname{argmax}} R_{x_1x_2}(p). \tag{3.7}$$

In practice, because of the finite observation time and the non-stationary of acoustic source, the CC is calculated at time k on observed samples of length k+L by its time-averaged estimate; therefore, the CC in digital implementation becomes

$$R_{x_1x_2}(p) = \begin{cases} \frac{1}{L-p}\sum_{l=0}^{L-1} x_1(k+l)x_2(k+l+p), & p = 0, \ldots, \tau_{max} \\ \frac{1}{L-p}\sum_{l=0}^{L-1} x_2(k+l)x_1(k+l+p), & p = -\tau_{max}, \ldots, 1 \end{cases} \tag{3.8}$$

where τ_{max} is the maximum TDOA of the microphone pair, and depends on the distance between microphones

$$\tau_{max} = \frac{d_{12}}{c}. \tag{3.9}$$

To normalize the CC, the Pearson Correlation Coefficient (PCC) can be used. The PCC is a measure of the correlation between two signals, giving a value between +1 and -1, inclusively. The PCC is defined as the covariance of the two variables divided by the product of their standard deviations

$$\rho_{x_1x_2} = \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}} = \frac{E[(x_1 - E[x_1])(x_2 - E[x_2])]}{\sigma_{x_1}\sigma_{x_2}}. \tag{3.10}$$

The performance of the CC is often degraded by many factors such as signal self-correlation and reverberation. Therefore, its application in a real-world context is not appropriate.

3.2.2 Generalized Cross-Correlation (GCC)

The Generalized Cross-Correlation (GCC) [Knapp & Carter, 1976] is the classic method to estimate the relative time delay associated with acoustic signals received by a pair of microphones in a moderately reverberant and noisy environment. The GCC basically consists of a cross-correlation followed by a

filter that aims to reduce the performance degradation caused by additive noise and multi-path channel effects. The GCC in the frequency domain is

$$R_{x_1x_2}^{GCC}(k) = \frac{1}{L}\sum_{f=0}^{L-1}\Psi(f)S_{x_1x_2}(f)e^{\frac{2\pi jfk}{L}} \tag{3.11}$$

where $\Psi(f)$ is the frequency domain general weighting function, and the cross-spectrum of the two signals is defined as

$$S_{x_1x_2}(f) = E[X_1(f)X_2^*(f)] \tag{3.12}$$

where $X_1(f)$ and $X_2(f)$ are the DFT of the signals and * denotes the complex conjugate. GCC is used for minimizing the influence of moderate uncorrelated noise and moderate multipath interference, maximizing the peak in correspondence of the time delay.

The relative time delay τ is obtained using an estimation of the maximum peak detection in the filter cross-correlation function

$$\widehat{\tau}^{GCC} = \underset{k}{\operatorname{argmax}} R_{x_1x_2}^{GCC}(k). \tag{3.13}$$

The CC is computed when $\Psi(f)_{CC} = 1$. The CC is estimated using the DFT and the IDFT, which can be efficiently implemented with the Fast Fourier Transform (FFT).

The most used and effective weighting function is the Phase Transform (PHAT) [Knapp & Carter, 1976]. It places equal importance on each frequency by dividing the spectrum by its magnitude. The PHAT normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC

$$\Psi_{PHAT}(f) = \frac{1}{|S_{x_1x_2}(f)|}. \tag{3.14}$$

Other weighting functions proposed by Knapp and Carter [Knapp & Carter, 1976] are the Roth Impulse Response (RIR), the Smoothed Coherence Transform (SCOT), the Hannan & Thomson (HT) and the Eckart. The RIR weighting is calculated according to the SNR value of one signal

$$\Psi_{RIR}(f) = \frac{1}{S_{x_1x_1}(f)}. \tag{3.15}$$

On the contrary, the SCOT filter assigns weighting according to the SNR of both signals

$$\Psi_{SCOT}(f) = \frac{1}{\sqrt{S_{x_1x_1}(f)S_{x_2x_2}(f)}}. \tag{3.16}$$

The HT weighting function, also known as the Maximum Likelihood (ML), uses the magnitude square coherence function $|\gamma_{x_1x_2}|^2$ between the signals

$$\Psi_{HT}(f) = \frac{|\gamma_{x_1x_2}|^2}{|S_{x_1x_2}(f)|(1-|\gamma_{x_1x_2}|^2)} \tag{3.17}$$

where

$$|\gamma_{x_1x_2}|^2 = \frac{|S_{x_1x_2}(f)|^2}{S_{x_1x_1}(f)S_{x_2x_2}(f)}.$$

The Eckart filter requires knowledge or estimation of the signal and noise spectrum

$$\Psi_{Eckart}(f) = \frac{S_{s_1s_2}(f)}{S_{v_1v_1}(f)S_{v_2v_2}(f)}. \tag{3.18}$$

However, the GCC is effective improving the time delay estimation between a microphone pair, especially in reverberant and noisy environments [Ianniello, 1982] [Champagne *et al.*, 1996] [Omologo & Svaizer, 1997]. A comparison of the GCC with different weighting is reported in Figure 3.2. These methods still tend to break down when room reverberation is high. The GCC methods are computationally efficient, and their use is optimal for monitoring situations that require an estimate for real-time systems. In addition, it is important to note that the GCC performance is dramatically reduced in the case of harmonic sounds, or generally pseudo-periodic sounds. In fact, the GCC has less capability to reduce the deleterious effects of noise and reverberation when it is applied to a pseudo-periodic sound. An accurate analysis of the PHAT performance for a broadband and narrowband signal can be found in [Donohue *et al.*, 2007]. The results of this work highlight the ability of the PHAT to enhance the detection performance for single or multiple targets in noisy and reverberant environments, when the signal covers most of the spectral range.

3.2.3 Adaptive Eigenvalue Decomposition (AED)

Adaptive Eigenvalue Decomposition (AED) [Benesty, 2000] is a BSI estimation method for time delay between a microphone pair based on a reverberant model (3.5) using eigenvalue decomposition. AED assumes that the system (room) is linear and time invariant; therefore, neglecting the influence of noise we can write

$$x_1 * g_2 = x_2 * g_1. \tag{3.19}$$

The vectors of the signal samples at the microphone outputs and the impulse response vectors of length L can be expressed as

$$\mathbf{x}_i(k) = [x_i(k), x_i(k-1), \ldots, x_i(k-L+1)]^T \tag{3.20}$$

and

$$\mathbf{g}_i = [g_{i,0}, g_{i,1}, \ldots, g_{i,L-1}]^T. \tag{3.21}$$

Substituting the vectors, the equation (3.19) becomes

$$\mathbf{x}_1^T\mathbf{g}_2 - \mathbf{x}_2^T\mathbf{g}_1 = 0. \tag{3.22}$$

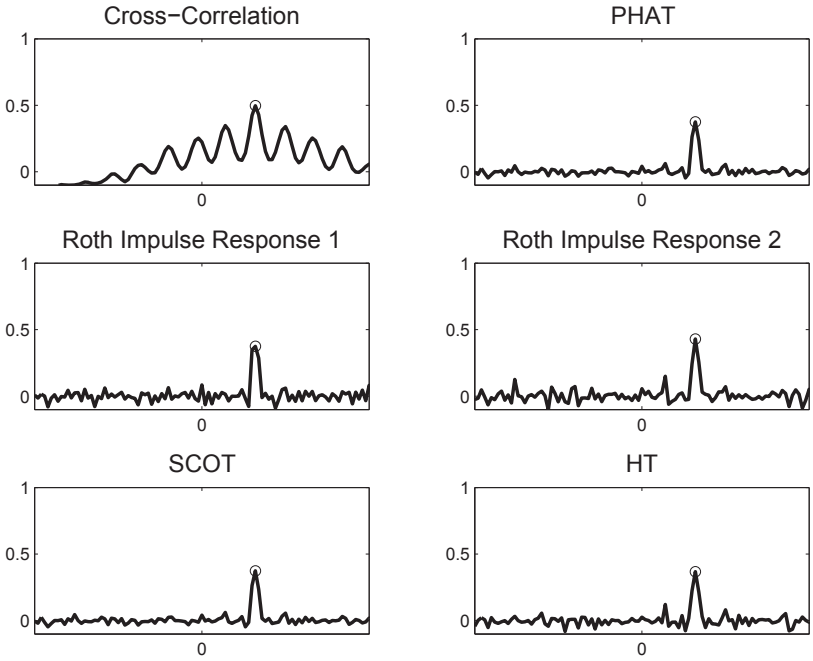

Figure 3.2: *GCC simulation with a female voice sound and SNR = 20 dB.*

Then, introducing the correlation matrix, we can define the following equation in matrix notation from (3.22)

$$\mathbf{R}\mathbf{u} = 0 \tag{3.23}$$

where $\mathbf{R}$ is the correlation matrix and $\mathbf{u}$ is a $2M \times 1$ vector formed by the juxtaposition of the two impulse responses

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{x_1x_1} & \mathbf{R}_{x_1x_2} \\ \mathbf{R}_{x_2x_1} & \mathbf{R}_{x_2x_2} \end{bmatrix}$$
$$\mathbf{u} = [\mathbf{g}_2 \ -\mathbf{g}_1]^T$$

and $\mathbf{R}_{x_ix_j} = E[\mathbf{x}_i(k)\mathbf{x}_j^T(k)]$. The equation (3.23) implies that $\mathbf{u}$ is the eigenvector corresponding to the zero-valued eigenvalue of $\mathbf{R}$. If the two impulse responses have no common zeros and the autocorrelation matrix of $s(n)$ is full rank, there is only a single zero-valued eigenvalue.

In practice, only estimate of the sample correlation matrix is available, and, therefore, instead of the zero-valued eigenvalue, we search for the minimum eigenvalue of $\mathbf{R}$ using the iterative method. The minimization of the vector $\mathbf{R}\mathbf{u}$ can be written as the minimization of $\mathbf{u}^T\mathbf{R}\mathbf{u}$, and imposing a constraint on $||\mathbf{u}||^2 = \mathbf{u}^T\mathbf{u} = 1$. The constrained minimization problem is

$$\widehat{\mathbf{u}} = \underset{\mathbf{u}}{\operatorname{argmax}} \mathbf{u}^T\mathbf{R}\mathbf{u}. \tag{3.24}$$

If the smallest eigenvalue is equal to zero, which is the case here, the error signal can be written in the following simplified form

$$e(k) = \mathbf{u}^T \mathbf{x} \tag{3.25}$$

where $\mathbf{x} = [\mathbf{x}_1^T \ \mathbf{x}_2^T]^T$. Note that minimizing the mean square value of $e(k)$ is equivalent to solving the above eigenvalue problem. Taking the gradient of $e(k)$ with respect to $\mathbf{u}(k)$, we obtain the gradient-descent constrained Least Mean Square (LMS) algorithm

$$\mathbf{u}(k+1) = \frac{\mathbf{u}(k) - \mu e(k)\mathbf{x}(k)}{||\mathbf{u}(k) - \mu e(k)\mathbf{x}(k)||} \tag{3.26}$$

where μ the is a positive constant adaptation step. Finally, the time delay estimation is

$$\widehat{\tau}_{12}^{AED} = \underset{l}{\text{argmax}}\widehat{g}_{1,l} - \underset{l}{\text{argmax}}\widehat{g}_{2,l}. \tag{3.27}$$

To speed up the convergence and to achieve efficient implementation, a Normalized MultiChannel Frequency domain LMS (NMCFLMS) can be used [Cho & Park, 2009].

The AED algorithm is valid only for a single source and if there is no noise and if spatiotemporally white noise is present. In [Doclo & Moonen, 2003], the authors extend the AED algorithm to noisy and reverberant acoustic environments, by deriving an adaptive stochastic gradient algorithm for the generalized eigenvalue decomposition. Under observation, AED may suffer from whitening effects with temporally correlated natural sounds. An improved AED method, proposed in [Cho & Park, 2009], imposes sparse priors on the responses to reduce the temporal whitening and provide a more accurate and robust time delay estimation. In [Buchner *et al.*, 2007], the authors present a system based on TRINICON, a general framework for broadband adaptive multiple-input-multiple-output signal processing. It is shown that the optimization criteria used for BSI allow a generalization of the AED algorithm for several simultaneously sources.

3.3 Time Delay Estimation Methods for Multiple Microphones

3.3.1 Steered Response Power Phase Transform (SRP-PHAT)

The Steered Response Power Phase Transform (SRP-PHAT) [DiBiase *et al.*, 2001] is based on the concept of adding several time delay estimation functions from the microphone pairs. It consists of calculating the GCC-PHAT function between pairs of microphones and using the GCF (2.65) to construct a spatial analysis map to improve the localization performance. Given R microphone pairs, the SRP-PHAT can be expressed

$$\mathcal{S}[\mathbf{s}, R^{GCC}(k)] = \sum_{r=1}^{R} R_{r,\mathbf{s}}^{GCC}(k) \tag{3.28}$$

where $\mathfrak{S}[\mathbf{s}, R^{GCC}(k)]$ is the spatial function likelihood that links a point in space $\mathbf{s} = [x\ y\ z]^T$ with the the GCC-PHAT $R^{GCC}_{r,\mathbf{s}}(k)$ of the r_{th} pair. The position of the source is estimated by picking the maximum peak

$$\widehat{\mathbf{s}}_{\mathbf{n}} = \underset{\mathbf{s}}{\operatorname{argmax}} \mathfrak{S}[\mathbf{s}, R^{GCC}(k)]. \tag{3.29}$$

We note that SRP-PHAT, which uses the sum of the GCCs of the microphone pairs, is equivalent to using a steered response filter and sum beamforming with PHAT weighting. In fact, the SRP of a 2-element array is equivalent to the GCC of those two microphones (see equation 3.59) [DiBiase *et al.*, 2001].

The SRP-PHAT algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments [Silverman *et al.*, 2005]. This algorithm enhances the performance of localization with a network of large arrays. However, the computational cost of the method is very high. To reduce the processing time of search algorithms, improvements have been suggested [Zotkin & Duraiswami, 2004] [Dmochowski *et al.*, 2007] [Cho *et al.*, 2009] [Cobos *et al.*, 2011].

3.3.2 Multichannel Cross-Correlation Coefficient (MCCC)

The Multichannel Cross-Correlation Coefficient (MCCC) algorithm is a spatial correlation-based method, which takes advantage of the redundant information provided by multiple sensors [Chen *et al.*, 2003] [Benesty *et al.*, 2004]. The idea is to use the spatial prediction (or interpolation) error to measure the correlation among multiple signals. Considering an array, a single source and neglecting the noise terms, we can write the signal model (3.4) as

$$x_m(k + \tau_m) = \alpha_m s(k - t). \tag{3.30}$$

In this way, we know that $x_1(k)$ is aligned with $x_m(k + \tau_m)$, and the new signal vector can be written

$$\mathbf{x}(k, p) = [x_1(k)\ x_2(k + \tau_m) \dots x_n(k + (M - 1)\tau_m)]^T \tag{3.31}$$

where p is a dummy variable for the hypothesized TDOA τ. The spatial correlation matrix of M microphones array is

$$\mathbf{R}(p) = \begin{bmatrix} \sigma^2_{x_1} & R_{x_1x_2}(p) & \dots & R_{x_1x_M}(p) \\ R_{x_2x_1}(p) & \sigma^2_{x_2} & \dots & R_{x_2x_M}(p) \\ \vdots & \vdots & \ddots & \vdots \\ R_{x_Mx_1}(p) & R_{x_Mx_2}(p) & \ddots & \sigma^2_{x_M} \end{bmatrix} \tag{3.32}$$

where $\sigma_{x_i}^2 = E(x_i)^2$ is the variance of signal x_i and $R_{x_i x_j}(p)$ is the cross-correlation between x_i and x_j. The spatial correlation matrix can be factored as

$$\mathbf{R}(p) = \mathbf{\Sigma}\widetilde{\mathbf{R}}(p)\mathbf{\Sigma} \tag{3.33}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{x_1} & 0 & \dots & 0) \\ 0 & \sigma_{x_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ddots & \sigma_{x_M} \end{bmatrix}$$

and

$$\widetilde{\mathbf{R}}(p) = \begin{bmatrix} 1 & \rho_{x_1 x_2}(p) & \dots & \rho_{x_1 x_M}(p) \\ \rho_{x_2 x_1}(p) & 1 & \dots & \rho_{x_2 x_M}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{x_M x_1}(p) & \rho_{x_M x_2}(p) & \dots & 1 \end{bmatrix}$$

is a symmetric matrix, and

$$\rho_{x_i x_j}(p) = \frac{r_{x_i x_j}(p)}{\sigma_{x_i}\sigma_{x_j}}$$

is the PCC between the i_{th} and j_{th} aligned microphone signals. The MCCC algorithm can be used to estimate the TDOA between the first two microphone signals as

$$\widehat{\tau}^{\mathrm{MCCC}} = \underset{p}{\operatorname{argmin}} \det[\widetilde{\mathbf{R}}(p)]. \tag{3.34}$$

Because the matrix $\widetilde{\mathbf{R}}(p)$ is symmetric and positive semi-definite, and its diagonal elements are all equal to one, we have

$$0 \leq det[\widetilde{\mathbf{R}}(p)] \leq 1. \tag{3.35}$$

The use of a filtered cross-correlation function can be used to improve the performance. Using the PHAT filter, we can write the new spatial correlation matrix for MCCC-PHAT

$$\widetilde{\mathbf{R}}^{\mathrm{PHAT}}(p) = \begin{bmatrix} 1 & R_{x_1 x_2}^{\mathrm{GCC}}(p) & \dots & R_{x_1 x_N}^{\mathrm{GCC}}(p) \\ R_{x_2 x_1}^{\mathrm{GCC}}(p) & 1 & \dots & R_{x_2 x_N}^{\mathrm{GCC}}(p) \\ \vdots & \vdots & \ddots & \vdots \\ R_{x_M x_1}^{\mathrm{GCC}}[p] & R_{x_M x_2}^{\mathrm{GCC}}(p) & \ddots & 1 \end{bmatrix}. \tag{3.36}$$

3.3.3 Adaptive Blind Multichannel Identification (ABMCI)

The extension of AED in the case of multiple microphones, as was proposed in Huang & Benesty [2003], is called Adaptive Blind Multichannel Identification (ABMCI). From (3.22), we can write the relationship between a generic pair of microphones i and j as

$$\mathbf{x}_i^T \mathbf{g}_j - \mathbf{x}_j^T \mathbf{g}_i = 0 \quad i, j = 1, \dots, M. \tag{3.37}$$

The error signal can be written

$$e_{ij} = \frac{\mathbf{x}_i^T \mathbf{g}_j - \mathbf{x}_j^T \mathbf{g}_i}{||\mathbf{g}||} \quad i,j = 1,\ldots,M \tag{3.38}$$

where $\mathbf{g} = [\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_M]^T$, and the multichannel LMS solution becomes

$$\mathbf{g}(k+1) = \mathbf{g}(k) - \mu \frac{\partial J(k+1)}{\partial \mathbf{g}(k)} \tag{3.39}$$

where $J(k+1)$ is the cost function defined by

$$J(k+1) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} e_{ij}(k+1)^2. \tag{3.40}$$

A simplified solution of this multichannel LMS problem proposed is

$$\mathbf{g}(k+1) = \frac{\mathbf{g}(k) - 2\mu[\mathbf{R}_+(k+1)\mathbf{g}(k) - J(k+1)\mathbf{g}(k)]}{||\mathbf{g}(k) - 2\mu[\mathbf{R}_+(k+1)\mathbf{g}(k) - J(k+1)\mathbf{g}(k)]||} \tag{3.41}$$

where

$$\mathbf{R}_+(k) = \begin{bmatrix} \sum_{m\neq 1} \mathbf{R}_{x_M x_M} & -\mathbf{R}_{x_2 x_1} & \cdots & -\mathbf{R}_{x_M x_1} \\ -\mathbf{R}_{x_1 x_2} & \sum_{m\neq 2} \mathbf{R}_{x_M x_M} & \cdots & -\mathbf{R}_{x_M x_2} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{R}_{x_1 x_M} & -\mathbf{R}_{x_2 x_M} & \cdots & \sum_{m\neq M} \mathbf{R}_{x_M x_M} \end{bmatrix}$$

and $\mathbf{R}_{x_i x_j} = \mathbf{x}_i \mathbf{x}_j^T$. Finally, the TDOA between two microphones is

$$\widehat{\tau}_{ij}^{ABMCI} = \underset{l}{\operatorname{argmax}} \widehat{g}_{i,l} - \underset{l}{\operatorname{argmax}} \widehat{g}_{j,l} \quad i,j = 1,\ldots,M. \tag{3.42}$$

3.4 Steered Beamforming Techniques

3.4.1 Steered Response Power (SRP)

The Steered Response Power (SRP) is based on maximizing the power output of a beamformer. Beamforming can be seen as a combination of the delayed signals from each microphone in such a way that an expected pattern of radiation is preferentially observed. The conventional beamformer is the Delay & Sum (DS), a natural extension of the classical Fourier-based spectral analysis to sensor data by Bartlett [Bartlett, 1948]. In general, the DS output y at time k is:

$$y(k) = \frac{1}{M} \sum_{m=1}^{M} a_m x_m(k) \tag{3.43}$$

where x_m is the received signal at microphones m and a_m is the steering value to delay the signal m. In the frequency domain, the DS output in matrix notation becomes

$$Y(f) = \mathbf{A}(f)\mathbf{X}(f) \tag{3.44}$$

where $\mathbf{X} = [X_1(f)\ X_2(f) \dots X_M(f)]^T$, $Y(f)$ and $X_m(f)$ are the DFT of the signals, and $\mathbf{A}(f) = [A_1(f)\ A_2(f) \dots A_M(f)]^T$ is the steering vector. The Power Spectral Density (PSD) of the output beamformer is given by

$$P = E[|Y(f)|^2] = \mathbf{A}(f)^H E[\mathbf{X}(f)\mathbf{X}(f)^H]\mathbf{A}(f) = \mathbf{A}(f)^H \mathbf{\Phi}(f)\mathbf{A}(f) \tag{3.45}$$

where $\mathbf{\Phi}(f)$ is the cross-spectral density matrix, which is square $M \times M$ and symmetric. The superscript H represents the Hermitian (complex conjugate) transpose. Then, the PSD is the sum of the frequency bin power

$$P_{DS} = \sum_{f=0}^{F} \mathbf{A}(f)^H \mathbf{\Phi}(f)\mathbf{A}(f) \tag{3.46}$$

where F is the max frequency bin. The values θ corresponding to the peak of the PSD allow for an estimation of the DOA of the source in the case of far-field environment

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} P_{DS}(\theta). \tag{3.47}$$

In the case of near-field environment the maximum value of the PSD corresponds to the position of source

$$\widehat{\mathbf{s}}_n = \underset{\mathbf{s}}{\operatorname{argmax}} P_{DS}(\mathbf{s}). \tag{3.48}$$

The beam pattern, representing the gain of the beamformer, is written as the magnitude of the steering vector

$$A_{bp}(f) = |\mathbf{A}(f)|. \tag{3.49}$$

In the case of a uniform linear array and far-field environment and assuming an angle range of: (-90°, +90°) $(-\pi/2 < \phi < \pi/2)$ (where zero is in front of the array and the microphone reference is the first from left), the beam pattern on direction ϕ becomes

$$A_{bp}(\phi, f) = \left| \frac{1}{M} \sum_{m=1}^{M} e^{\frac{-j2\pi f(m-1)d(\sin(\phi)-\sin(\theta))}{cL}} \right|. \tag{3.50}$$

Figure 3.3 shows the beam pattern for an equispaced linear array of four and sixteen microphones, with a microphone distance of d = 25 cm and a desired direction of $\theta = 0°$. The beam in the desired direction with the highest amplitude is named the mainlobe, and all the others are called sidelobes. The sidelobes represent the gain pattern for noise and competing sources along the undesired directions. The beamforming techniques aim to make the sidelobes as low as possible so that the signals coming from other directions are attenuated as much as possible. For this reason, to improve the beamforming performance, some filter methods have been developed.

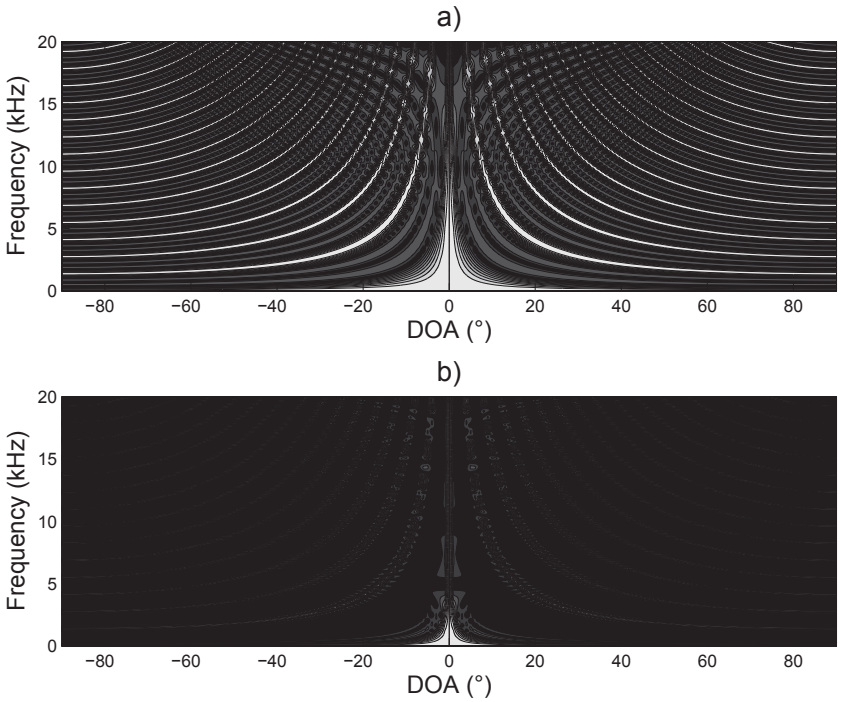

Figure 3.3: *Beam pattern of SRP for a uniform ($d = 25$ cm) linear array: a) four microphones, b) sixteen microphones.*

3.4.2 Filter Steered Response Power (FSRP)

The Filter Steered Response Power (FSRP) aims to improve the performance of the conventional beamformer. There are numerous modifications proposed by researchers; however, in general, the FSRP can be written in the frequency domain as

$$Y(f) = \mathbf{W}(f)\mathbf{X}(f) \tag{3.51}$$

where $\mathbf{W}(f)$ is the vector of the beamformer weights for steering and filtering the data. Typically, the independent weight data are the classic windowing and optimum-approximation approaches. The Dolph-Chebyshev (DC) [Dolph, 1946] window is analytically derived from the minimum and maximum approximation property of the Chebyshev polynomials and it minimizes the sidelobes level with a equal attenuation

$$P_{DC}(f) = [\mathbf{h} \cdot \mathbf{A}(f)]^H \mathbf{\Phi}(\mathbf{f})[\mathbf{h} \cdot \mathbf{A}(f)] \tag{3.52}$$

where $\mathbf{h} = [h_1, h_2, \ldots, h_M]^T$ is DC windowing of length M and $\cdot$ denotes element-by-element multiplication. The beam pattern of the DC beamformer in shown in Figure 3.4. The data independent

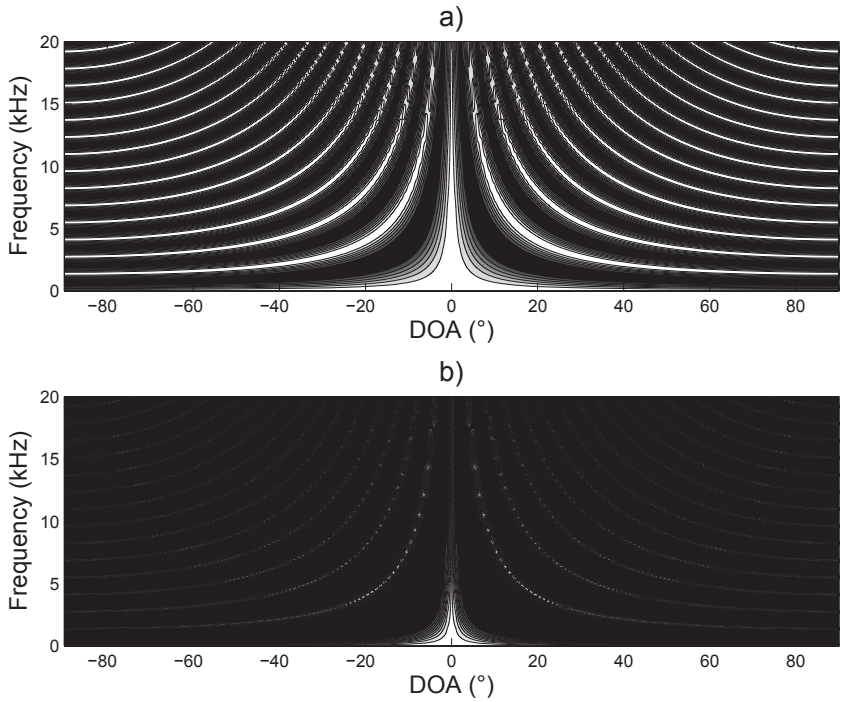

Figure 3.4: *Beam pattern of Dolph-Chebyshev windowing SRP for a uniform ($d = 25$ cm) linear array: a) four microphones, b) sixteen microphones.*

beamforming techniques can fully take advantage of the array geometry and source location information to optimize their beam pattern. In contrast, the adaptive beamforming uses characteristics of the source and the noise signals to improve the beam pattern.

The most widely used adaptive beamformer is the Minimum Variance Distortionless Response (MVDR) due to Capon [Capon, 1969]. The MVDR beamformer is based on resolving the minimization problem

$$\min \mathbf{W}(f)^H \mathbf{\Phi}(f) \mathbf{W}(f) \quad \text{subject to} \quad \mathbf{W}(f)\mathbf{A}(f) = 1. \tag{3.53}$$

In this way, the aim is to minimize the noise and sources coming from different directions, while keeping a fixed gain on the desired direction $\mathbf{A}$. Solving (3.53) using the method of Lagrange multipliers, we can write

$$\mathbf{W}_{MVDR}(f) = \frac{\mathbf{A}(f)^H \mathbf{\Phi}(f) \mathbf{A}(f)}{\mathbf{A}(f)^H \mathbf{A}(f)}. \tag{3.54}$$

Consequently, the power of the beamformer output becomes

$$P_{MVDR}(f) = \frac{1}{\mathbf{A}(f)^H \mathbf{\Phi}(f)^{-1} \mathbf{A}(f)}. \tag{3.55}$$

In practical applications, the inverse of the cross-spectral density matrix can be calculated using the Moore-Penrose pseudoinverse, i.e. using the singular value decomposition

$$\mathbf{\Phi}^{+} = \mathbf{V}\mathbf{S}^{+}\mathbf{U}^{H}. \tag{3.56}$$

where the singular value decomposition is $\mathbf{\Phi} = \mathbf{U}\mathbf{S}\mathbf{V}^{H}$ and the pseudoinverse $\mathbf{S}^{+}$ is obtained by replacing every nonzero diagonal entry by its reciprocal.

If the cross-spectral density matrix is ill-conditioned, the spatial spectrum may not exist. Therefore, a Diagonal Loading (DL) [Cox *et al.*, 1987] [Carlson, 1988] method is adopted to calculate the inverse matrix. The spatial spectrum function becomes

$$P_{MVDR-DL}(f) = \frac{1}{\mathbf{A}(f)^{H}(\mathbf{\Phi}(f) + \mu\mathbf{I})^{-1}\mathbf{A}(f)} \tag{3.57}$$

where $\mathbf{I}$ is the identity matrix and μ is the loading level

$$\mu = \frac{1}{L}\text{trace}(\mathbf{\Phi})\Delta \tag{3.58}$$

where Δ is the normalized loading constant. Typically, the values are: $\Delta = 0.1$, $\Delta = 1$, $\Delta = 10$ [Huang *et al.*, 2011].

To conclude this section, the SRP-PHAT, described in (3.3.1) as the time delay estimation method because it uses a summation of the GCC-PHAT of all the microphone pairs of the array network, can be formalized here in terms of the FSRP [DiBiase *et al.*, 2001] using the following equation

$$P_{SRP-PHAT}(f) = \mathbf{A}(f)^{H}(\mathbf{\Phi}(f) \div |\mathbf{\Phi}(f)|)\mathbf{A}(f) \tag{3.59}$$

where $\div$ denotes element-by-element division.

3.4.3 Multiple Signal Classification (MUSIC)

The MUltiple SIgnal Classification (MUSIC) algorithm is a high-resolution beamforming technique developed for a narrowband signal. It is based on an eigen subspace decomposition method, and it is dependent on the correlation matrix of the data [Schmidt, 1979] [Schmidt, 1986]. From the free-field model (3.4), we consider $\alpha = 1$, and then in the frequency domain for the frequency f the signal model for the generic array r of M microphones and the source n becomes

$$X_m(f) = S(f)e^{-j2\pi f(k_1+\tau_m)} + V_m(f) \tag{3.60}$$

where $X_m(f)$, $S(f)$, and $V_m(f)$ are, respectively, the DFT of $x_m(k)$, $s(k)$ and $v_m(k)$, and k_1 is the propagation time from the unknown source n to the reference sensor of array r. We define the vectors

of M signals $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{V}$; hence, we have

$$\mathbf{X} = \mathbf{A}_m\mathbf{S} + \mathbf{V} \tag{3.61}$$

where

$$\mathbf{A}_m = \left[e^{-j2\pi f(k_1+\tau_1)}, e^{-j2\pi f(k_1+\tau_2)} \dots e^{-j2\pi f(k_1+\tau_M)}\right]^T.$$

Computing the cross-spectral density matrix $\mathbf{\Phi}$ and considering Gaussian white noise with the same variance σ_v^2 at each microphone, we obtain

$$\mathbf{\Phi} = E[\mathbf{X}\mathbf{X}^H] = \mathbf{A}_m\mathbf{\Phi}\mathbf{A}_m^H + \sigma_v^2\mathbf{I}. \tag{3.62}$$

Therefore, if we perform the eigenvalue decomposition of the cross-spectral density matrix, we have

$$\mathbf{\Phi} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H \tag{3.63}$$

where $\mathbf{U}$ is the square $M \times M$ matrix whose i_{th} column is the eigenvector $\mathbf{q}_i$ of $\mathbf{\Phi}$ and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. If we assume that N is the number of sources, the first N eigenvectors must span the signal-plus-noise subspace, whereas the $M - N$ eigenvectors q_i ($N < i < M$) span the noise-only subspace. Accordingly, we have

$$\mathbf{\Phi}\mathbf{q}_i = \sigma_v^2\mathbf{q}_i. \tag{3.64}$$

We also know that

$$\mathbf{\Phi}\mathbf{q}_i = (\mathbf{A}_m\mathbf{\Phi}_{ss}\mathbf{A}_m^H + \sigma_v^2\mathbf{I})\mathbf{q}_i. \tag{3.65}$$

Combining equations (3.64) and (3.65), we find that

$$\mathbf{A}_m\mathbf{\Phi}_{ss}\mathbf{A}_m^H\mathbf{q}_i = 0 \tag{3.66}$$

which is equivalent to

$$\mathbf{A}_m^H\mathbf{q}_i = 0. \tag{3.67}$$

Hence, the eigenvectors of the noise-only subspace corresponding to the zero eigenvalue are orthogonal to all N signal steering vectors, which leads us to define the spatial pseudo-spectrum

$$J = \frac{1}{\mathbf{A}^H\mathbf{G}\mathbf{G}^H\mathbf{A}} \tag{3.68}$$

where $\mathbf{G}$ is the $M \times (M - N)$ matrix containing the eigenvectors corresponding to the noise-only subspace and $\mathbf{A}$ is a steering vector towards candidate source location

$$\mathbf{A} = \left[e^{\frac{j2\pi f\tau_1}{L}}, e^{\frac{j2\pi f\tau_2}{L}}, \dots e^{\frac{j2\pi f\tau_M}{L}}\right]^T \tag{3.69}$$

where L is the number of samples of the observation time.

MUSIC was originally developed for narrowband signals; however, in the case of broadband signals, the pseudo-spectrum is calculated for each frequency, and we compute the incoherent average output power [Wax & Kailath, 1984]

$$J_{MUSIC} = \sum_{f=0}^{L-1} \frac{1}{\mathbf{A}(f)^H \mathbf{G}(f) \mathbf{G}(f)^H \mathbf{A}(f)}. \tag{3.70}$$

Finally, the values corresponding to the N peaks of the pseudo-spectrum allow for an estimation of the sources. In far-field condition, the values corresponding to the DOAs are

$$\widehat{\theta} = \arg(\underset{\theta}{\text{local}})\max J_{MUSIC}(\theta). \tag{3.71}$$

Another method to compute the MUSIC for a broadband signal is by using geometric and harmonic averaging [Azimi-Sadjadi *et al.*, 2008], which can be also be used in other SRP methods, and by using a coherent combination of the spatial signal spaces of the temporally narrow-band decomposition [Wang & Kaveh, 1985] [Yoon *et al.*, 2006].

An improvement of MUSIC is the Root-Music [Barabell, 1983] [Rao & Hari, 1989]. It performs the DOAs estimation for far-field environment using the roots of a polynomial formed from the noise subspace, with the advantage of directly estimating the DOAs, without the steered step and the maximum peak search on the pseudo-spectrum output.

3.4.4 Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT)

ESPRIT stands for Estimation of Signal Parameters via Rotational Invariance Techniques [Paulraj *et al.*, 1986] [Roy *et al.*, 1986] [Roy & Kailath, 1989] and is another subspace method for DOA estimation. The goal of the ESPRIT technique is to exploit the rotational invariance in the signal subspace created by the translational invariance structure of two sub-arrays. Consider a uniform linear array with microphone distance d and assuming $N \leq M$ sources, the two sub-arrays are displaced by distance d: the sub-array$_1$ is composed by microphones (1,2,...,M-1) and the sub-array$_2$ by (2,3,...,M). The signals induced on each of the arrays are given by the signal output of two sub-arrays can be written as follows

$$\begin{aligned} \mathbf{x}_1(k) &= \mathbf{as}(k) + \mathbf{v}_1(k) \\ \mathbf{x}_2(k) &= \mathbf{a}\omega\mathbf{s}(k) + \mathbf{v}_2(k) \end{aligned} \tag{3.72}$$

where $\mathbf{a}$ is the steered vector and ω is the rotation operator, a diagonal $M \times M$ matrix of the phase delays between the two sub-arrays. In the frequency domain, we have

$$\begin{aligned} \mathbf{X}_1(f) &= \mathbf{AS}(f) + \mathbf{V}_1(f) \\ \mathbf{X}_2(f) &= \mathbf{A\Omega S}(f) + \mathbf{V}_2(f) \end{aligned} \tag{3.73}$$

where $\mathbf{A}$ and $\mathbf{\Omega}$ are the DFT of $\mathbf{a}$ and ω. From the eigenvalue decomposition of the spectral correlation matrix (3.63), we can define the vector $\mathbf{E}_s$, which contains the N eigenvectors that span the signal-plus-noise subspace. The subspaces of the eigenvectors are related by a unique nonsingular transformation matrix $\mathbf{T}$ such as

$$\mathbf{E}_s = \mathbf{AT} \tag{3.74}$$

where $\mathbf{A} = [\mathbf{A} \quad \mathbf{A\Omega}]^T$. The structure of two-subarray implies that $\mathbf{E}_s$ can be decompose into

$$[\mathbf{E}_{x_1} \ \mathbf{E}_{x_2}]^T = [\mathbf{AT} \ \mathbf{A\Omega T}]^T. \tag{3.75}$$

Because the sub-array$_1$ and the sub-array$_2$ are translationally related, the subspaces of eigenvectors are related by a unique nonsingular transformation matrix $\mathbf{\Psi}$ such that

$$\mathbf{E}_{x_1}\mathbf{\Psi} = \mathbf{E}_{x_2}. \tag{3.76}$$

Finally, we can write

$$\mathbf{T\Psi T}^{-1} = \mathbf{\Omega}. \tag{3.77}$$

This is the basic equation of the ESPRIT method. Hence, the eigenvalues of $\mathbf{\Psi}$ must be equal to the diagonal elements of $\mathbf{\Omega}$, and the columns of $\mathbf{T}$ are the eigenvectors of $\mathbf{\Psi}$. The goal is to estimate the matrix $\mathbf{\Psi}$ to obtain the eigenvalues $(\psi_1, \psi_2 \ldots, \psi_N)$ and to calculate the angles θ corresponding to the DOAs of the N sources

$$\widehat{\theta}(f) = \text{asin}\Big(\frac{\arg(\psi_n)cL}{2\pi f \mathbf{\Delta}}\Big) \tag{3.78}$$

where $\mathbf{\Delta}$ is the translation displacement vector of the microphones. The total LS criterion can be applied to solve equation (3.76) and to find the matrix $\mathbf{\Psi}$. A performance analysis of the total least squares ESPRIT algorithm can be found in [Ottersten *et al.*, 1991].

ESPRIT was originally developed for the DOA estimation of narrowband emitter signals. The problem of bearing estimation of a single wideband source is addressed in [Khan & Tufail, 2009]. In [Teutsch & Kellermann, 2005] [Sun *et al.*, 2011], a novel method, called EB-ESPRIT, is proposed. It consists of calculating the spherical harmonics (eigenbeams) for representation of the acoustic wavefield. The obtained wavefield representation is then used to serve as a basis for high-resolution ESPRIT.

3.5 Summary

This chapter has reviewed the most important multi-channel signal processing techniques for sound localization. The GCC-PHAT is the classic and the most effective method to estimate the TDOA of microphone pair because it reduces the problem of additive noise, self-correlation and multi-path channel effects of a moderate reverberation. The SRP-PHAT (or GCF-PHAT, referring to the sum of different microphone pairs) and the MCCC-PHAT provide TDE methods when an array contains M microphones, using redundant information between microphones to estimate the TDOA in a more robust manner under a reverberant and noisy conditions. However, the PHAT weighting has problems in a high reverberation environment. Thus, the BSI methods focus on the impulse responses between the sources and the microphones to solve the TDOA estimation in highly reverberant environments. AED is the classic BSI method used to estimate the TDOA of a microphone pair for a single source (the ABMCI is an extension in the case of multiple microphones), and TRINICON is a generalization of the AED algorithm for several simultaneous sources.

The SRP approach is based on maximizing the power output of a beamformer. The beamforming techniques aim to make the sidelobes as low as possible so that the signals coming from other directions are attenuated as much as possible. For this reason, to improve the beamforming performance, multiple filter methods have been developed. The MVDR is the most widely used adaptive filter beamformer. Another approach, developed for narrowband signal processing, is based on subspace decomposition, which includes the MUSIC and ESPRIT algorithms. These high-resolution beamforming techniques perform eigenvalue decomposition to analyze the signal-plus-noise subspace and the noise-only subspace to estimate the position of several sources.

To conclude, it is important to note that the performance of PHAT weighting is dramatically reduced in the case of pseudo-periodic sounds. A framework for the solution of this problem is proposed in section (6.3).

4

Pre-Processing for Signal Enhancement

4.1 Signal Enhancement

Single-channel noise reduction techniques provide a useful tool to improve the quality of a signal affected by noise. The SNR of a microphone signal is an important parameter that affects the capability to locate several acoustic sources. However, some of the multi-channel signal processing methods described in chapter (3) have the capability of reducing the effect of noise, especially uncorrelated noise: PHAT weighting, SRP-MVDR, and high-resolution beamforming. Nevertheless, there may be some situations where it is helpful to perform pre-processing for signal enhancement with the goal of reducing the presence of uncorrelated and correlated noise [Salvati & Canazza, 2009]. In section (6.3), experimentation illustrates how noise reduction can improve accuracy when locating a pseudo-periodic source in a moderate noisy environment. This chapter is structured into two sections called Frequency Domain Methods and Time Domain Methods. Later versions of the Extended Kalman filter use an autoregressive (AR) model representing the signal. In chapter (5), the Kalman filter will be revisited when the classical formulation of state-space methods is applied to navigation problems.

4.2 Frequency Domain Methods

Frequency domain methods, which are based on Short-Time Spectral Attenuation (STSA), have been proposed by Boll in [Boll, 1979]. The method suppresses stationary noise along the entire signal by subtracting the spectral noise bias calculated during activity in the absence of the signal of interest. These de-noise systems consist of two important components: a noise estimation method and a suppression rule. These techniques employ a signal analysis through the Short-Time Fourier Transform (STFT) (which is calculated on windowed sections of the signal as it changes over time) and can be considered as a non-stationary adaptation of the Wiener filter [Wiener, 1949] in the frequency domain. In particular, STSA consists of applying the short-time spectrum of the noise to a time-varying suppression, and it does not require the definition of a model for the audio signal. If we consider the useful signal $s(k)$ as a stationary aleatory process to which some noise $v(k)$ is added (uncorrelated with $x(k)$) to produce the degraded signal $x(k)$, the relation that connects the respective power spectral densities is

$$P_x(f) = P_s(f) + P_v(f). \tag{4.1}$$

If we are to succeed in retrieving an adequate estimate of $P_v(f)$, during the silence intervals of the signal $x(k)$, and during the signal portions of $P_x(f)$, we can expect to obtain an estimate of the spectrum of $s(k)$ by subtracting $P_v(f)$ from $P_x(f)$. The initial assumption of being stationary can be considered locally satisfied because short temporal windows are employed. Note that the use of a short-time signal analysis is equivalent to the use of a filter bank. First each channel (that is, the output of each filter) is appropriately attenuated, after which it is possible to proceed with the synthesis of the de-noising signal. The time-varying attenuation applied to each channel is calculated through a determined suppression rule, which has the purpose if producing an estimate (for each channel) of the noise power. Each particular STSA technique is characterized by the implementation of the filter bank and of the suppression rule.

If we denote the STFT of the $x(k)$ noisy signal with $X(k, f)$, where k represents the temporal index and f represents the frequency index (with $f = 0, 1, ..., L - 1$, L representing the number of STFT channels), the result of the suppressing rule application can be interpreted as the application of a $G(k, f)$ gain to each value $X(k, f)$ of the STFT of the noisy signal. This gain corresponds to a signal attenuation and is takes a value between 0 and 1. In most of the suppression rules, $G(k, f)$ only depends on the noisy signal power level (measured at the same point) and on the estimate of the noisy power at the f frequency

$$\widehat{P}_v(f) = E[|V(k, f)|^2] \tag{4.2}$$

(which does not depend on the temporal index k due to the presumed stationary noise). At this point, a relative signal can be defined

$$Q(k,f) = \frac{|X(k,f)|^2}{\widehat{P}_v(f)} \tag{4.3}$$

which, starting from the hypothesis that the $v(k)$ noise is not correlated with the $x(k)$ signal, we deduce should be greater than 1

$$E[Q(k,f)] = 1 + \frac{E[|S(k,f)|^2]}{\widehat{P}_v(f)}. \tag{4.4}$$

A typical suppression rule is based on the Wiener filter [Wiener, 1949] and can be formulated as follows

$$G(k,f)_{WIENER} = \frac{|X(k,f)|^2 - \widehat{P}_v(f)}{|X(k,f)|^2}. \tag{4.5}$$

Typically, a mistake made by this procedure in retrieving the original sound spectrum has an audible effect, because the difference between the spectral densities can yield a negative result at certain frequencies. If we decide to arbitrarily force the negative results to zero, a disturbance will occur in the final signal that is composed, constituted numerous random frequency pseudo-sinusoids that start and finish in a rapid succession, generating what is known in the literature as musical noise.

To improve the performance of the STSA and to reduce the musical noise, different methods have been developed. In [Ephraim & Malah, 1984] [Ephraim & Malah, 1985], the authors proposed a technique that utilizes a minimum mean-square error STSA estimator and the mean-square error of the log-spectra. Later in [Cappe, 1994], a study presented the noise suppression technique proposed by Ephraim and Malah and demonstrated how the artifact is actually eliminated without bringing distortion to the signal even if the noise is only poorly stationary. The Ephraim and Malah Suppression Rule (EMSR) can be written as

$$G(k,f)_{EMSR} = \frac{\sqrt{\pi}}{2}\sqrt{\frac{Y_{prio}(k,f)}{(1+Y_{post}(k,f))(1+Y_{prio}(k,f))}}e^{-\frac{\beta}{2}}\left[(1+\beta)I_0(\frac{\beta}{2}) + \beta I_1(\frac{\beta}{2})\right] \tag{4.6}$$

where

$$\beta = \frac{Y_{prio}(k,f)(1+Y_{post}(k,f))}{(1+Y_{prio}(k,f))}$$

$$Y_{post}(k,f) = \frac{|X(k,f)|^2}{\widehat{P}_v(f)} - 1$$

$$Y_{prio}(k,f) = \begin{cases} (1-\alpha)Y_{post}(k,f) + \alpha\frac{|G(k-1,f)X(k-1,f)|^2}{\widehat{P}_v(f)} & Y_{post}(k,f) > 0 \\ \alpha\frac{|G(k-1,f)X(k-1,f)|^2}{\widehat{P}_v(f)} & Y_{post}(k,f) \leq 0 \end{cases}$$

with I_0 and I_1 being the Bessel modified functions of zero and one order, respectively. The α parameter controls the balance between the current frame information and that of the preceding one. By varying this parameter, the filter smoothing effect can be regulated.

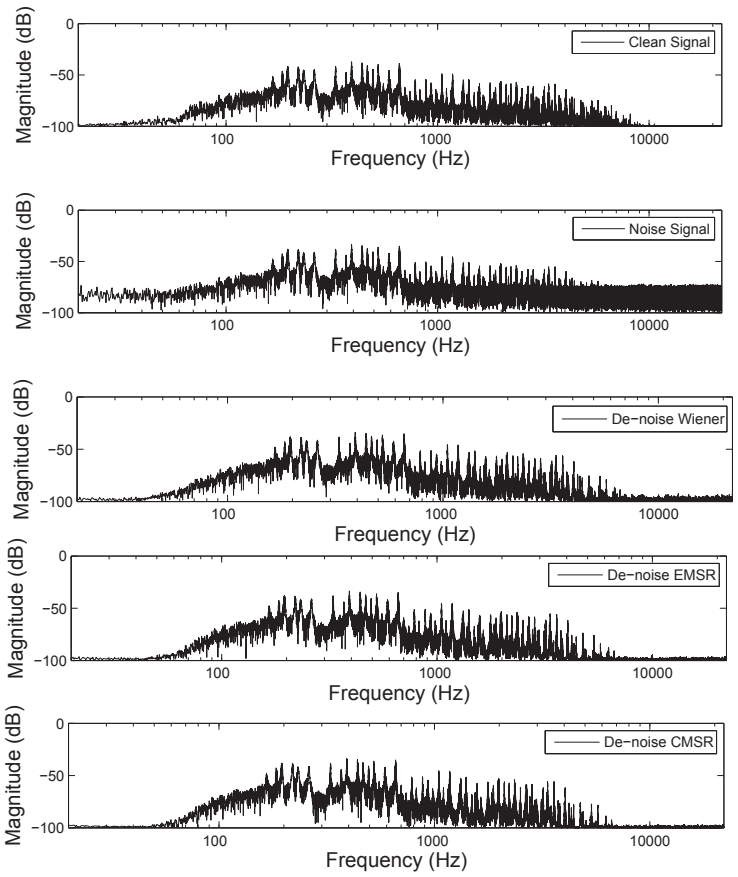

Figure 4.1: *Simulation of STSA noise reduction with additive Gaussian white noise, SNR = 15 dB.*

Two evolutions of the STSA were proposed in [Canazza *et al.*, 2001] [Bari *et al.*, 2001] [Canazza, 2007]. The first is an improvement of the Ephraim and Malah suppression rule and is called the Canazza-Mian Suppression Rule (CMSR). This rule is based on the idea of using a punctual suppression without memory in the case of a null estimate of $Y_{post}(k, f)$. The following condition on α is the basis of the CMSR

$$\alpha = \begin{cases} 0.98 & Y_{post}(k, f) > 0 \\ 0 & Y_{post}(k, f) \leq 0 \end{cases} \tag{4.7}$$

The second improvement is based on a perceptive filter and involved, developing a psychoacoustic model from the works of [Beerends & Stemerdink, 1992] [Beerends & Stemerdink, 1994] [Tsoukalas *et al.*, 1997a] [Tsoukalas *et al.*, 1997b]. The algorithm consists of transforming the Hertz scale to the Bark scale through the estimation of the relative signal excitation at each critical band, computing the

outer-to-inner ear transformation to obtain the time spreading (which is an operation with memory of the preceding frame), and finally computing the frequency spreading. Once the psychoacoustic model is obtained, the STSA can be applied in the Bark domain.

4.3 Time Domain Methods

Methods based on the time domain approach refer to an autoregressive (AR) model representing the signal. An initial Bayesian approach can be found in [Godsill & Rayner, 1995], in which the authors derive the a posteriori probability for the location of bursts of noise additively superimposed on a Gaussian AR process. In [Niedzwiecki & Cisowski, 1996], the nonlinear filtering problem is solved using the theory of the Extended Kalman Filter (EKF). Later, Monte Carlo filtering and smoothing were studied using the Rao-Blackwellized particle filter [Fong *et al.*, 2002], which improves the standard particle filter and the EKF. In [Ning *et al.*, 2006], the perceptually constrained Kalman filter was developed. This filter consists of minimizing the estimation error variance of the EKF under the constraint that the energy of the estimation error is smaller than a masking threshold of human auditory systems. The EKF is improved using four new procedures in [Canazza *et al.*, 2010], which consist of a stability test, a bootstrap procedure, a variable-forgetting factor and forward/backward filtering. Because the EKF approach is sensitive to parameter setting, the work also describes a three-pass procedure detailing the choice of the parameter values for each step.

Here, a brief notation of the EKF for broadband and impulsive noise removal is described, referring to [Canazza *et al.*, 2010]. For the basic details of the Kalman filter and EKF, see [Kalman, 1960] and [Schmidt, 1966]. We first define the state vector $\mathbf{x}(k)$ of $(q+p)$ dimension $(q > p)$, by combining the unknown AR model coefficient $a_i(k)$ and the signal vector $s(k)$

$$\mathbf{x}(k) = [s(k), s(k-1), \ldots, s(k-p), \ldots, s(k-q+1), a_1(k), a_2(k), \ldots, a_p(k)]^T. \tag{4.8}$$

The state-space signal model can be written by the following state equation

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{W}(k)\mathbf{x}(k) + \mathbf{u}(k) \\ x(k+1) &= \mathbf{c}^T\mathbf{x}(k) + \xi(k) \end{aligned} \tag{4.9}$$

where $\mathbf{u}(k)$ is the vector containing the Gaussian zero-mean white noise error of the AR model and the

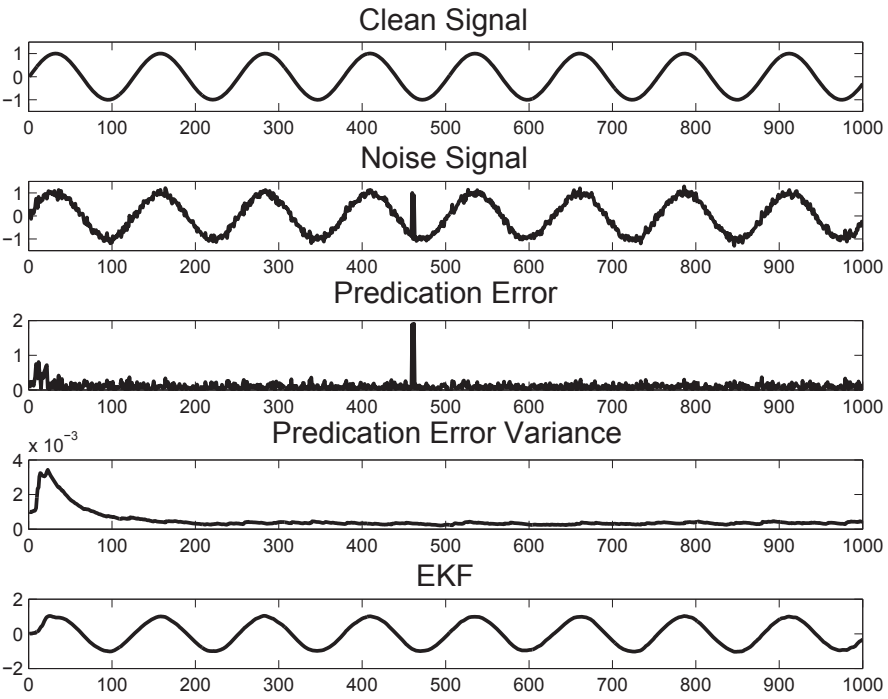

Figure 4.2: *Simulation of EKF noise reduction with additive Gaussian white noise and impulsive noise.*

random walk model, $\xi(k)$ is the sum of broadband noise and impulsive noise, and

$$\mathbf{W}(k) = \begin{bmatrix} a_1(k) & a_2(k) & \dots & a_p(k) & \dots & 0_{q-1} & 0_q & 0_{q+1} & \dots & 0_{q+p} \\ 1 & 0 & \dots & & & \vdots & \vdots & \vdots & & \vdots \\ & \ddots & & & & & & & & \\ 0_p & & \ddots & & & & & & & \\ \vdots & & & \ddots & & \vdots & \vdots & \vdots & & \\ 0_{q-1} & \dots & & & \ddots & 0 & 0 & 0 & \dots & \\ 0_q & \dots & & & & 1 & 0 & 0 & \dots & \\ 0_{q+1} & \dots & & & \dots & 0 & 0 & 1 & & \\ \vdots & & & & & \vdots & \vdots & & \ddots & \\ 0_{q+p} & \dots & & & & & & & & 1 \end{bmatrix}$$

$\mathbf{c} = [1, 0, \dots, 0_{q+p}]^T$.

The EKF equations, which optimally suppress the disturbing noise provided, for the prediction step become

$$\begin{aligned} \widehat{\mathbf{x}}(k|k-1) &= \mathbf{W}(k-1)\widehat{\mathbf{x}}(k-1|k-1) \\ \mathbf{\Sigma}(k|k-1) &= \mathbf{F}(k-1)\mathbf{\Sigma}(k-1|k-1)\mathbf{F}^T(k-1) + \mathbf{\Omega} \end{aligned} \tag{4.10}$$

where $\mathbf{F}(k)$ denotes the state transition matrix of the linearized system

$$\mathbf{F}(k) = \begin{bmatrix} a_1(k) & a_2(k) & \dots & a_p(k) & \dots & 0_{q-1} & 0_q & s(k) & \dots & s(k-p) \\ 1 & 0 & \dots & & & \vdots & \vdots & 0 & \dots & 0 \\ & \ddots & & & & & & \vdots & & \vdots \\ 0_p & & \ddots & & & & & & & \\ \vdots & & & \ddots & & \vdots & \vdots & \vdots & & \\ 0_{q-1} & \dots & & & \ddots & 0 & 0 & 0 & \dots & \\ 0_q & \dots & & & & 1 & 0 & 0 & \dots & \\ 0_{q+1} & \dots & & & \dots & 0 & 0 & 1 & & \\ \vdots & & & & & \vdots & \vdots & & \ddots & \\ 0_{q+p} & \dots & & & & & & & & 1 \end{bmatrix}$$

$$\mathbf{\Omega} = \begin{bmatrix} 1 & 0 & \dots & 0_{q+p} \\ 0 & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0_{q+p} & \dots & \dots & 0_{q+p} \end{bmatrix}.$$

The equation for the update step is

$$\begin{aligned} \widehat{\mathbf{x}}(k|k) &= \widehat{\mathbf{x}}(k|k-1) + \mathbf{L}(k)\epsilon(k) \\ \mathbf{\Sigma}(k|k) &= (\mathbf{I}(q+p) - \mathbf{L}(k)\mathbf{c}^T)\mathbf{\Sigma}(k|k-1) \end{aligned} \tag{4.11}$$

where $\mathbf{\Sigma}(k|k)$ is the state estimation error covariance and $\mathbf{\Sigma}(k|k-1)$ is the state prediction error covariance. The predication error $\epsilon(k)$ is defined as

$$\epsilon(k) = x(k) - \mathbf{c}^T\widehat{\mathbf{x}}(k|k-1). \tag{4.12}$$

The Kalman gain $\mathbf{L}(k)$ depends on the presence of impulsive noise, which in the positive case is $\mathbf{L}(k) = 0$ and otherwise is

$$\mathbf{L}(k) = \frac{\mathbf{\Sigma}(k|k-1)\mathbf{c}}{\mathbf{c}^T\mathbf{\Sigma}(k|k-1)\mathbf{c} + \eta(k)}. \tag{4.13}$$

More insights into how to verify the presence of impulsive noise and the optimal choice of parameters for the filter (q, p, etc.) can be found in [Canazza *et al.*, 2010]. The de-noise sample of the EKF at each step of time k is

$$\widehat{s}(k) = \mathbf{q}\widehat{\mathbf{x}}(k|k) \tag{4.14}$$

where $\mathbf{q} = [0, \dots, 1_q, \dots, 0q+p]$.

4.4 Summary

This chapter has presented frequency and time domain noise reduction techniques for the single-channel enhancement. Frequency domain methods require that the operator has some prior information to carry out the reduction; only an estimate of the noise present is necessary (noise print) because it is assumed to be stationary along the entire signal. Any further needed information is automatically calculated by the algorithm through the analysis of the signal characteristics. Algorithms in the time domain, which use signal models, require that a priori information is employed to estimate the probable distribution of the sound events, the excitation signal and the filter coefficients. Therefore, the algorithm carries out (a posteriori information) the signal tracking. The models, which can be applied to different signal typologies, have little a priori information. It is therefore necessary to update the model from time to time, according to the signal being examined.

Therefore, if very little a priori information is available in addition to the context of the acoustic source localization (required for real-time applications), then frequency methods are the best option; these are based on STSA.

5

Post-Processing for Localization Enhancement

5.1 Localization Enhancement

The goal of post-processing for localization enhancement is to improve of the quality of the positional estimates. Post-processing is a fundamental and crucial step that provides increased precision for position data and attempts to minimize or eliminate results obtained from reflection, reverberation and error measurements. There are two main approaches, namely the Bayesian filter and clustering. The Bayesian filter, i.e., Kalman and Particle filter, also has the advantage that it solves the problem of multiple source localization, and moreover, it provides the ability to track the source in case of movement. Clustering can be an alternative when sporadic and concurrent events are present because the tracking of short events may be particularly difficult with classical filtering approaches.

5.2 Kalman Filter (KF)

The Kalman filter (KF) [Kalman, 1960] is the optimal recursive Bayesian filter for linear systems observed in the presence of Gaussian noise. We consider that the state of the sound localization could be summarized by three position variables (x, y, z), and velocities (v_x, v_y, v_z). These variables are the

elements of the state vector $\mathbf{x}(k)$

$$\mathbf{x}(k) = [x, y, z, v_x, v_y, v_z]^T. \tag{5.1}$$

The process model relates the state at a previous time $k-1$ with the current state at time k, so we can write

$$\mathbf{x}(k) = \mathbf{F}\mathbf{x}(k-1) + \mathbf{w}(k) \tag{5.2}$$

where $\mathbf{F}$ is the transfer matrix and $\mathbf{w}(k-1)$ is the process noise associated with random events or forces that directly affect the actual state of the system. We assume that the components of $\mathbf{w}(k-1)$ have a Gaussian distribution with zero mean normal distribution and covariance matrix $\mathbf{Q}(k)$, $\mathbf{w}(k-1) \sim N(0, \mathbf{Q}(k))$. Considering the dynamical motion, if we measure the system to be at position x with some velocity v at time k, then at time $k+dk$ we would expect the system to be located at position $x + v \cdot dk$, suggesting that the correct form for $\mathbf{F}$ is

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & dk & 0 & 0 \\ 0 & 1 & 0 & 0 & dk & 0 \\ 0 & 0 & 1 & 0 & 0 & dk \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5.3}$$

At time k, an observation $\mathbf{z}(k)$ of the true state $\mathbf{x}(k)$ is made according to the measurement model

$$\mathbf{z}(k) = \mathbf{H}\mathbf{x}(k) + \mathbf{v}(k) \tag{5.4}$$

where $\mathbf{H}$ is the observation model that maps the true state into the observed space and $\mathbf{v}(k)$ is the observation noise that is assumed to be the zero mean Gaussian white noise with covariance $\mathbf{R}(k)$,$\mathbf{v}(k) \sim N(0, \mathbf{R}(k))$. We only measure the position variables. Hence, we obtain

$$\mathbf{z}(k) = \begin{bmatrix} \widehat{x} \\ \widehat{y} \\ \widehat{z} \end{bmatrix} \tag{5.5}$$

and then we obtain

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{5.6}$$

The filter equations can be divided into a prediction and a correction step. The prediction step projects forward the current state and covariance to obtain an a priori estimate. Afterwards, the correction step

uses a new measurement to obtain an improved a posteriori estimate. In the prediction step, the time update equations are

$$\begin{aligned}\widehat{\mathbf{x}}(k|k-1) &= \mathbf{F}(k)\widehat{\mathbf{x}}(k-1|k-1)\\ \mathbf{P}(k|k-1) &= \mathbf{F}(k)\mathbf{P}(k-1|k-1)\mathbf{F}^T + \mathbf{Q}(k-1)\end{aligned} \tag{5.7}$$

where $\mathbf{P}(k)$ denotes the error covariance matrix. The posterior Probability Density Function (PDF) is calculated in the correction step, in which the measurement update equations are

$$\begin{aligned}\widehat{\mathbf{x}}(k|k) &= \widehat{\mathbf{x}}(k|k-1) + \mathbf{K}(k)(\mathbf{z}(k) - \mathbf{H}(k)\widehat{\mathbf{x}}(k|k-1))\\ \mathbf{P}(k|k) &= (\mathbf{I} - \mathbf{K}(k)\mathbf{H})\mathbf{P}(k|k-1)\end{aligned} \tag{5.8}$$

where $\mathbf{I}$ is the identity matrix and the so-called Kalman gain matrix is

$$\mathbf{K}(k) = \mathbf{P}(k-1|k-1)\mathbf{H}^T(\mathbf{H}(k)\mathbf{P}(t-1|t-1)\mathbf{H}^T + \mathbf{R}(k))^{-1}. \tag{5.9}$$

This formulation requires that the dynamic of the system be linear. However, the specific problem is nonlinear. To accommodate the nonlinear state transition and observation models, the Extended Kalman Filter (EKF) [Schmidt, 1966] implements a local linearization of models. Thus, we need to compute new values for $\mathbf{F}$ at every time step based on the state $\mathbf{x}$ to approximate the real update. In the EKF, the state transition and observation models need not be linear functions of the state but may instead be differentiable functions

$$\begin{aligned}\mathbf{x}(k) &= f[\mathbf{x}(k-1)] + \mathbf{w}(k)\\ \mathbf{z}(k) &= h[\mathbf{x}(k)] + \mathbf{v}(k)\end{aligned} \tag{5.10}$$

where the nonlinear system dynamic model $f[\mathbf{x}(k-1)]$ and $h[\mathbf{x}(k)]$ are assumed known.

Many studies are related to tracking an acoustic source with EKF. In [Strobel *et al.*, 2001a] and [Strobel *et al.*, 2001b], a joint audio-video signal processing based on a decentralized Kalman filter structure was modified so that different sensor measurement models could be incorporated. In [Bechler *et al.*, 2003], a method for tracking a single speaker is presented, in which an adaptive Kalman filter is used to obtain a smoothed trajectory of the speakers movement. Alternatively, the solution for tracking multiple moving speakers using multiple microphone arrays is given by [Potamitis *et al.*, 2004]. In [Klee *et al.*, 2006], the proposed algorithm is applied a Kalman filter to directly update the speakers position estimate based on the observed TDOAs and not to smooth the position estimate so that the closed-form approximation is eliminated. The use of the temporal information with KF to improve the tracking performance is proposed in [Gannot & Dvorkind, 2006]. A multiple model Kalman filter approach was covered by [Liang *et al.*, 2008] and [Seguraa *et al.*, 2008] for multi speaker tracking in a reverberant and noisy room.

5.3 Particle Filter (PF)

The Particle Filter (PF) is a technique for implementing a recursive Bayesian filter using Monte Carlo simulations [Gordon *et al.*, 1993] [Carpenter *et al.*, 1999] [Orton & Fitzgerald, 2002] [Gustafsson *et al.*, 2002] [Arulampalam *et al.*, 2002]. Application of multiple sources can be found in [Hue *et al.*, 2002] [Larocque *et al.*, 2002]. A PF supports nonlinear and non-Gaussian state space models and uses the representation of the posterior PDF as a set of random samples with associated weights. It estimates the new state-space by processing these samples and weights and is thus a numerical approximation of the Bayesian filter.

Let us define each sample of the state vector, which is referred to a particle, $[\mathbf{x}_{0:k}^i w_k^i]\, i = 1, \ldots, N_s$, by a random measure that characterizes the posterior PDF with associated weights w_k^i. The weights are normalized such that $\sum_{i=1}^{N_s} w_k^i = 1$. The Sequential Importance Resampling (SIR) is the original particle filter proposed in [Gordon *et al.*, 1993]. After the initialization of particle $x_0(i) \sim p(x_0)$, where $p(x_0)$ is the initial distribution, the SIR method considers the evaluation of the posterior PDF as

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \propto p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k-1}) \tag{5.11}$$

where the likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ is called the importance function and is calculated from the state vector equation using the known measurement of noise density (5.2). The PF approximates $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k-1})$ with samples, according to

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k-1}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{5.12}$$

where δ is the delta-Dirac function. Then, the weights update becomes

$$w_k^i = w_{k-1}^i p(\mathbf{z}_k|\mathbf{x}_k) \quad i = 1, \ldots, N_s \tag{5.13}$$

and normalizes to

$$w_k^i = \frac{w_k^i}{\sum_{i=1}^{N_s} w_k^i}. \tag{5.14}$$

Usually, the estimation of the state vector can be approximated as

$$\widehat{\mathbf{x}}_k = \sum_{i=1}^{N_s} w_k^i \mathbf{x}_k^i. \tag{5.15}$$

Now the prediction can be computed from the model vector (5.2). A common problem is the degeneracy phenomenon, in which after a few iterations some particles have negligible weight. The resampling step

eliminates particles with small importance weights when the effective number of samples is less than a threshold $N_r < N_s$

$$N_{eff} = \frac{1}{\sum_{i=1}^{N_s}(w_k^i)^2} < N_r. \tag{5.16}$$

The use of a PF to track acoustic sources can be found in [Zotkin *et al.*, 2002], in which a particle-filter based tracking framework for performing audio-video fusion is described for tracking people in a videoconferencing environment using multiple cameras and multiple microphone arrays. In [Ward *et al.*, 2003], the use of a PF for acoustic localization in a reverberant environment is reported with the capability to accurately track a moving source in a moderately reverberant room (with a measured reverberation time of $0.39s$). An extension to multi-source reverberant environments was proposed by [Antonacci *et al.*, 2006]. A method based on a frequency-domain implementation of a steered beamformer along with a Particle filter-based tracking algorithm and the application of an autonomous robot are presented in [Michaudy & Rouat, 2007]. A PF using important sampling was developed in [Lehmann & Williamson, 2006], and a PF integrated with voice activity detection is presented in[Lehmann & Johansson, 2007]. The authors of [Talantzis *et al.*, 2008] propose a system based on a PF and an information theoretical framework to provide accurate acoustic source location under reverberant conditions, and a fusion to a video system is applied to improve the performance. In [Quinlan *et al.*, 2009], a method for tracking intermittent speaking was proposed. Recently, [Levy *et al.*, 2011] addresses the extended PF scheme by adapting the multiple-hypothesis model, which is associated with the multi-path of reverberation, to track a single acoustic source in reverberant environments.

5.4 Clustering

In contrast to the Bayesian approach, the goal of clustering is to enhance the localization estimation, eliminating the incorrect data, by grouping the instantaneous location estimates that are close in time and space in a cluster. Because tracking a sporadic and concurrent source is a difficult problem with the KF and PF approaches, because in general the filter requires an initialization phase to enter an optimal state of work, clustering aims to solve these sporadic event problems with an intermediary task between instantaneous localization and continuous source tracking.

The core of clustering is the evaluation of the distance measure between data points, in our case the coordinates of sources over time, to identify the groups of similarity and dissimilarity, namely dense clusters that indicate the position of the source and disperse clusters containing few points that are associated with erroneous measurements.

Different algorithms are proposed to analyze the distance between points to appropriately cluster the data. We mention those that are used in acoustic source localization applications: the k-means clustering, the fuzzy C-means clustering and the Gaussian mixture models clustering.

The k-means clustering [MacQueen, 1967] defines k centroids, one for each cluster, after which each data point is associated with the nearest centroid and the centroids are re-calculated as the centers of the clusters. These iterative steps aim to minimize the squared error function

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} ||x_k^i - v_i||^2 \tag{5.17}$$

where n is the number of data points, c is the number of clusters, x_k is the k_{th} data point and v_i is the centroid i. K-means clustering is used in the context of the automatic annotation of speakers [Ajmera *et al.*, 2004] and intelligent robot auditory systems [Lee & Choi, 2010].

The fuzzy C-means clustering [Dunn, 1973] is an iterative optimization algorithm that minimizes the function

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m ||x_k - v_i||^2 \tag{5.18}$$

where the constant m is greater than 1 and μ_{ik}^m is the degree of membership of the k_{th} data point into the i_{th} cluster, and it is defined as follows

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{||x_k - v_i||}{||x_k - v_j||} \right)^{2/(m-1)}}. \tag{5.19}$$

In [Claudio & Parisi, 2001], a fuzzy C-means with efficient methods for estimating the cluster center [Chiu, 1994] is used to solve the localization problem in a multi-source and reverberant environment. To increase the algorithms robustness against sound reflections, the authors propose a weighted fuzzy C-means [Khnea *et al.*, 2009], based on local DOA statistics near the sound onsets because these regions are less affected by reverberation.

Finally, we mention the Gaussian mixture models clustering, which is applied in multiple moving speakers environments [Lathoud & Odobez, 2007], in which each cluster is mathematically represented by a parametric Gaussian distribution and the entire data set is therefore modeled by a mixture of these distributions. Expectation-Maximization (EM) algorithms are used to determine the mixture [Dempster *et al.*, 1977].

5.5 Summary

The Kalman filter, the Particle filter and clustering for localization enhancement have been discussed. The KF is a statistically optimal estimate for linear systems observed in the presence of Gaussian noise,

and EKF accommodates the nonlinear state transition and observation models to navigation problems. In contrast, the PF is a numerical approximation of the Bayesian filter and supports nonlinear and non-Gaussian state space models. The Bayesian filter can fail during the initialization phase of the filter, when the sources have an unpredictable trajectory (e.g., in the case of rapid changes of the velocity vector), and when two sources have intersecting trajectories.

The Kalman and Particle filters are used to solve the problem of multiple source localization by the tracking of sources. However, the tracking of sporadic and concurrent events (i.e., short sound events) may be particularly difficult. A solution to this problem of multiple source localization is proposed in section (6.2). Clustering can be used to improve the quality of the position measures in the case of short duration events.

6 Experimental Prototypes

6.1 Introduction

In this chapter two experimental prototypes are introduced as an innovative contribution to this thesis. In the first part, the Incident Signal Power Comparison approach is described to solve the multi-source problem in far-field and free-field environments, with particular attention to short-duration sounds. After the presentation of the prototype setup, some experimental results in real-world scenarios are presented. In the second part, a prototype for localization of pseudo-periodic sounds and some experimental results in real, moderate reverberant and noisy environments are presented.

In particular, the far-field application section describes a prototype system for multiple source localization in a public area. This application is of interest for audio surveillance, sound monitoring and analysis of acoustic scenes. Very small sized arrays are used, namely two linear arrays each consisting of four microphones because a real application of such systems would require that the public spaces are not invaded in an excessive way; therefore, there might not be enough space to install the arrays. The major problem encountered in an application of this type is the detection of a significant number of short-duration events by the localization system. For this reason, finding the limitations of applying the Bayesian filters is necessary. A new algorithm to solve the problem of multi-source localization is presented.

Conversely, the near-field application section describes a prototype system for pseudo-periodic source localization that can be used in musical human-computer interaction. In recent years, musical interfaces are often used to allow the performer to enhance the expressive control on the sounds generated by their acoustic instruments in a live electronics context. These systems are based on electric field, optical and video camera sensors. In general, those types of systems have considerable complexity and problems may arise in some situations. For example, the performer has to wear sensors or devices, which can be a hindrance to his/her movements. In camera-based systems, there may be problems with the low or uncontrollable lighting in the concert hall. Thus, a digital musical interface based on sound source localization is used to allow a performer to plan and conduct the expressivity of a performance by controlling an audio processing module in real-time through spatial movement of a sound source. The proposed interface has the advantage of being completely noninvasive (no need for markers, sensors or wires on the performance), and no dedicated hardware is required. Hence, a novel framework for the localization of pseudo-periodic sounds in moderate reverberant and noisy environments is proposed.

6.2 Far-Field Application

In this section, the prototype system for the localization of multiple acoustic sources in real world noisy environment is described. The first part covers the Incident Signal Power Comparison algorithm to solve the ambiguity (see Figure 2.6) to correctly link the DOAs from different arrays to the same source. After the description of the system setup for experimental evaluation, simulation work is presented with DOA estimation methods including SRP, SRP with Dolph-Chebyshev windowing, SRP-PHAT, MVDR, MVDR-DL, MUSIC, Root-Music, and ESPRIT. Finally, the ISPC performance in the real world is reported using different beamforming techniques and spectral difference estimations.

6.2.1 Incident Signal Power Comparison (ISPC)

Incident Signal Power Comparison (ISPC) [Salvati *et al.*, 2010] [Salvati *et al.*, 2011c] combines the DOAs from different arrays by considering the similarity criterion among sources. To check for this similarity, we can estimate for each array the ISP referring to all estimated DOAs using beamforming techniques. Once the ISPs are obtained, we can define an efficient error criterion for comparing the different possible combinations of the ISPs using a spectral distance measure. In summary, the steps of the algorithm are 1) TDOAs and DOAs estimation, 2) source separation using beamforming techniques, 3) ISP comparison using spectral distance measurement, 4) verification of the most consistent target

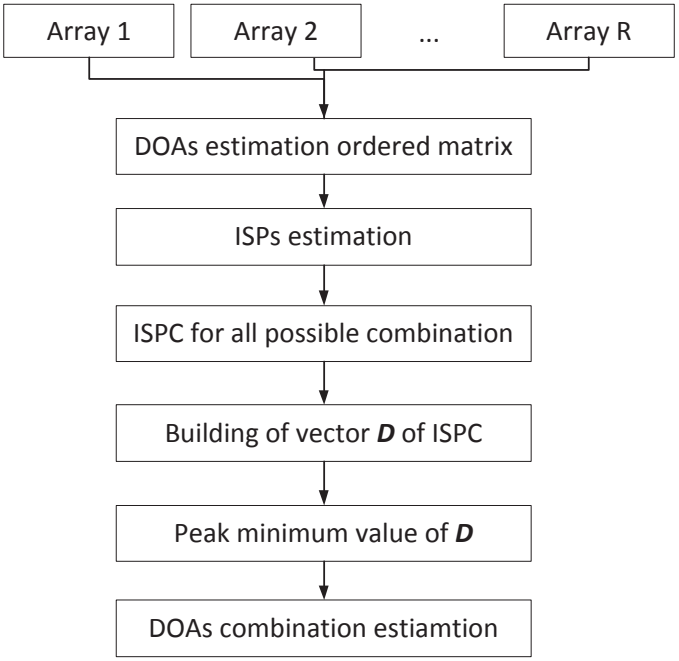

Figure 6.1: *The steps for the ISPC algorithm.*

combinations minimizing an error criterion, and 5) localization of sources. Figure 6.1 illustrates the ISPC steps. We start by defining the vector $\mathbf{\Theta_n}$ for each source n and considering the signal model 3.4

$$\mathbf{\Theta_n} = [\theta_{1,n}, \theta_{2,n}, \ldots, \theta_{R,n}]^T \tag{6.1}$$

which contains the DOAs of the acoustic source n by each array. In the case of N sources and R arrays, we can write the matrix $R \times N$, which contains all DOAs as

$$\mathbf{\Theta} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \ldots & \theta_{1,N} \\ \theta_{2,1} & \theta_{2,2} & \ldots & \theta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{R,1} & \theta_{R,2} & \ddots & \theta_{R,N} \end{bmatrix}. \tag{6.2}$$

The estimated DOAs angles, obtained for each array r are written with the following vector

$$\underline{\widehat{\mathbf{\Theta}}_\mathbf{r}} = [\theta_{r1}, \theta_{r2}, \ldots, \theta_{rN}]^T \tag{6.3}$$

where we consider the angle values in ascending order ($\theta_{r1} < \theta_{r2} < \theta_{r3}$, etc.). Next, the estimated ordered matrix $\underline{\widehat{\mathbf{\Theta}}}$ is defined as

$$\underline{\widehat{\mathbf{\Theta}}} = \begin{bmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1N} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{R1} & \theta_{R2} & \ddots & \theta_{RN} \end{bmatrix}. \tag{6.4}$$

The position of the n_{th} source can be calculated by combining the DOAs estimated by the R arrays for that source. The next problem is to correctly assign those R DOAs values to the n_{th} source. In general, the goal is to get the matrix $\mathbf{\Theta}$ to properly order the values of equation (6.4). Considering θ_{ri} as the i_{th} DOA of array r, the assignment of the correct value of the angle for the unknown sources can be ambiguous (see Figure 2.6), namely the exact position of the elements in the matrix of equation (6.3) cannot be uniquely determined

$$\theta_{ri} \rightarrow \theta_{r,n}. \tag{6.5}$$

There are $(N!)^{(R-1)}$ possible combinations of the DOAs of matrix (6.4). However, after obtaining an estimation of the DOAs, we can compute the estimation of the ISP for each DOA of arrays.

The ISP is the power spectral density of the output beamformer that is steered to a specified direction DOA. From equation (3.51), the ISP corresponding to a generic θ_{ri} and a frequency f can be written as a FSRP beamforming

$$ISP_{ri}(f) = E[|Y(f)|^2] = \mathbf{W}(f,\theta_{ri})^H \mathbf{\Phi}(f) \mathbf{W}(f,\theta_{ri}) \tag{6.6}$$

where $\mathbf{W}(f,\theta_{ri})$ is the vector of the beamformer weights for filtering the data and steering the beamformer to θ_{ri}, $\mathbf{\Phi}(f)$ is the cross-spectral density matrix and $Y(f)$ is the output signal of beamformer. The four beamforming methods that are used for comparing experimental results are the SRP, SRP-DC, MVDR, and MVDR-DL. Hence, the ISP corresponding to SRP can be written from equation (3.46) as

$$ISP_{ri}^{SRP} = \sum_{f=F_{min}}^{F_{max}} \mathbf{A}(f,\theta_{ri})^H \mathbf{\Phi}(f) \mathbf{A}(f,\theta_{ri}) \tag{6.7}$$

where $\mathbf{W}(f,\theta_{ri}) = \mathbf{A}(f,\theta_{ri})$ is the steering vector corresponding to direction θ_{ri}, F_{min} and F_{max} are the values of considered frequency bin range. The SRP-DC is obtained from equation (3.52) introducing the Dolph-Chebyshev window $\mathbf{h}$

$$ISP_{ri}^{SRP-DC} = \sum_{f=F_{min}}^{F_{max}} [\mathbf{h} \cdot \mathbf{A}(f,\theta_{ri})]^H \mathbf{\Phi}(\mathbf{f}) [\mathbf{h} \cdot \mathbf{A}(f,\theta_{ri})]. \tag{6.8}$$

The adaptive MVDR beamforming is based on minimization problem of equation (3.53, and the ISP from (3.55) is

$$ISP_{ri}^{MVDR} = \sum_{f=F_{min}}^{F_{max}} \frac{1}{\mathbf{A}(f,\theta_{ri})^H \mathbf{\Phi}(f)^{-1} \mathbf{A}(f,\theta_{ri})}. \tag{6.9}$$

Consider the loading level μ defined in equation (3.58), the ISP of the modified MVDR, which uses the DL to estimate the pseudoinverse cross-correlation spectral matrix, is

$$ISP_{ri}^{MVDR-DL} = \sum_{f=F_{min}}^{F_{max}} \frac{1}{\mathbf{A}(f,\theta_{ri})^H (\mathbf{\Phi}(f) + \mu\mathbf{I})^{-1} \mathbf{A}(f,\theta_{ri})}. \tag{6.10}$$

Therefore, we can define the matrix $\mathbf{P}$ containing all the ISPs related to the matrix (6.4)

$$\mathbf{P} = [ISP_{11}, ISP_{12}, \ldots, ISP_{1N}, ISP_{21}, \ldots, ISP_{2N}, ISP_{R1}, \ldots, ISP_{RN}]^T \tag{6.11}$$

which has a dimension of $RN \times (F_{max} - F_{min})$.

To compare the ISPs of different arrays (i.e., ISPC), spectral distance functions are used. Distance measures produce measurements of the dissimilarity of two sound spectra. We define the spectral distance estimation between the ISP$_{ri}$ and the ISP$_{bj}$ of two DOAs of different arrays as

$$E_{ribj} = \frac{1}{L} \sum_{f=F_{min}}^{F_{max}} |S[ISP_{ri}(f), ISP_{bj}(f)]| \tag{6.12}$$

where r and b are the index labels of the array, $r \neq b$, i and j are the index labels for the ordered DOAs of array, L is the number of samples for the observation time and $S[ISP_{ri}(f), ISP_{bj}(f)]$ is the Spectral Distance Functions (SDF) to measure the dissimilarity of spectra. We consider the four most common SDF to verify how our system performance varies as a function of different equations. A classic spectral estimation method is Linear Prediction (LP) Makhoul [1975], for which we insert a negative one to standardize the minimum to zero as all functions

$$E_{ribj}^{LP} = \frac{1}{L} \sum_{f=F_{min}}^{F_{max}} \Big| \frac{ISP_{ri}(f)}{ISP_{bj}(f)} - 1 \Big|. \tag{6.13}$$

The other functions are the Itakura-Saito (IS) distance measure McAulay [1984]

$$E_{ribj}^{IS} = \frac{1}{L} \sum_{f=F_{min}}^{F_{max}} \Big| \frac{ISP_{ri}(f)}{ISP_{bj}(f)} - \log \frac{ISP_{ri}(f)}{ISP_{bj}(f)} - 1 \Big| \tag{6.14}$$

the Root Mean Square (RMS) log Pfeifer [1974]

$$E_{ribj}^{RMS} = \frac{1}{L} \sum_{f=F_{min}}^{F_{max}} \Big| \log \frac{ISP_{ri}(f)}{ISP_{bj}(f)} \Big|^2 \tag{6.15}$$

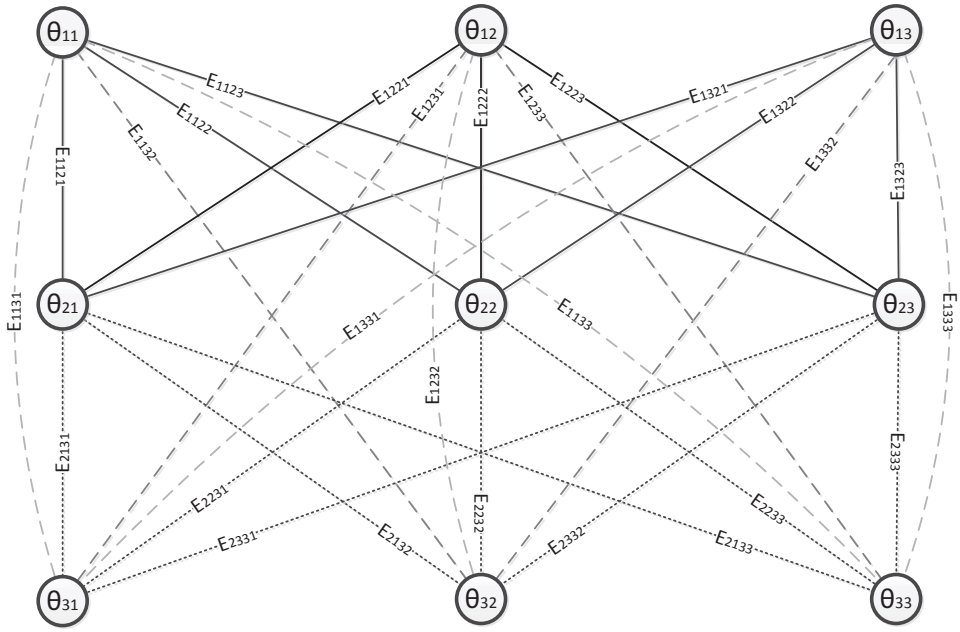

Figure 6.2: *Graphic representation of DOAs and spectral distance estimations.*

and the COSH measure Gray & Markel [1976]

$$E_{ribj}^{COSH} = \frac{1}{L}\sum_{f=F_{min}}^{F_{max}} \left| \frac{ISP_{ri}(f)}{ISP_{bj}(f)} - \log\frac{ISP_{ri}(f)}{ISP_{bj}(f)} + \frac{ISP_{bj}(f)}{ISP_{ri}(f)} - \log\frac{ISP_{bj}(f)}{ISP_{ri}(f)} - 2 \right|. \tag{6.16}$$

We now represent the sorted matrix of the DOAs using the graph theory to better understand the verification of the most consistent target combination minimizing an error criterion. Then, we can express the matrix (6.4) and all of its combinations as being composed of nodes and edges, connecting pairs of vertices. An example of three arrays and three sources is shown in Figure 6.2. Each row of the graph contains the ordered DOAs of an array: $\underline{\widehat{\mathbf{\Theta}}_{\mathbf{1}}} = [\theta_{11}, \theta_{12}, \theta_{13}]^T$, $\underline{\widehat{\mathbf{\Theta}}_{\mathbf{2}}} = [\theta_{21}, \theta_{22}, \theta_{23}]^T$, and $\underline{\widehat{\mathbf{\Theta}}_{\mathbf{3}}} = [\theta_{31}, \theta_{32}, \ldots, \theta_{3N}]^T$. Each DOA is a node of graph and the edges represent the possible connections between nodes with the values E_{ribj}, estimating spectral distance. The combination of incorrect angles leads to an incorrect position estimation (see Figure 2.6). Thus, if we represent a combination of angles as a sum of values of the edges that connect the nodes, we expect that the minimum value of different sums corresponds to the correct combination. To calculate the possible combinations of angles between

the arrays, it is helpful to introduce a matrix labeling the angles (6.4)

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & \dots & N \\ N+1 & N+2 & \dots & 2N \\ \vdots & \vdots & \ddots & \vdots \\ (R-1)N+1 & (R-1)N+2 & \dots & RN \end{bmatrix}. \tag{6.17}$$

The matrix label $\mathbf{B}$ associates the position of the angles referring to the sorted matrix $\underline{\widehat{\mathbf{\Theta}}}$. Estimating the minimum error of the ISPC, we can obtain the matrix $\widehat{\mathbf{\Theta}}$ with the correct position of the angles, in which each column contains the DOAs of the source n.

Furthermore, we can represent the graph representation of angles and ISPCs as the adjacency matrix $\mathbf{A}$, which is an $RN \times RN$ matrix of ISPC values. The entry in row ($a_{ri} = 1, \dots, RN$) and column ($a_{bj} = 1, \dots, RN$) is defined as an ISPC $E_{a_{ri}a_{bj}}$ if there is an edge connecting vertex a_{ri} and vertex a_{bj} in the graph, or is defined as zero otherwise. The relationships between the row and column of the adjacency matrix $\mathbf{A}$ and the label matrix $\mathbf{B}$ are $B(r,i) = a_{ri}$ and $B(b,j) = a_{bj}$; thus, the relationship between DOAs and ISPC can expressed by following equation

$$\mathbf{A}(B(r,i), B(b,j)) = E_{ribj}. \tag{6.18}$$

The symmetric adjacency matrix results

$$\mathbf{A} = \begin{bmatrix} 0 & \dots & 0 & E_{1121} & \dots & E_{112N} & \dots & E_{11R1} & \dots & E_{11RN} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & E_{1N21} & \dots & E_{1N2N} & \dots & E_{1NR1} & \dots & E_{1NRN} \\ E_{2111} & \dots & E_{211N} & 0 & \dots & 0 & \dots & E_{21R1} & \dots & E_{21RN} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ E_{2N11} & \dots & E_{2N1N} & 0 & \dots & 0 & \dots & E_{2NR1} & \dots & E_{2NRN} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ E_{R111} & \dots & E_{R11N} & E_{R121} & \dots & E_{1121} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ E_{RN11} & \dots & E_{RN1N} & E_{RN21} & \dots & E_{RN2N} & \dots & 0 & \dots & 0 \end{bmatrix}. \tag{6.19}$$

We have $N^2R(R-1)/2$ spectral distance measures. These spectral distance function values are weights of the edges of the graph. An example of three arrays and three sources is presented in Figure 6.2; in this example, we have 27 total ISPCs (3 ISPCs for each source).

For each source, identified by R nodes, we have $R(R-1)/2$ edges; then, the number of edges for a combination of angles is $Q = NR(R-1)/2$. We can define the spectral distance estimation of the generic combination z as the sum of the weights of all the edges

$$d_z = \sum_{q=1}^{Q} E_q \tag{6.20}$$

where q contains all the information for the index r, i, b and j, and the following explains how to calculate it. To calculate all possible combinations of angles, we can work on the label matrix $\mathbf{B}$. Considering that the first row of $\mathbf{B}$ related to the first array remains unchanged, we compute the combinations in two steps. In first step, the permutations of the labels of $\mathbf{B}$ for each row $r = 2, 3, \ldots, R$ are calculated. The number of permutations for each row is $N_p = N!$. Next, we define the permutation matrix T

$$\mathbf{T} = [\mathbf{p}_{11}, \mathbf{p}_{12}, \ldots, \mathbf{p}_{1N_p}, \mathbf{p}_{21}, \mathbf{p}_{22}, \ldots, \mathbf{p}_{2N_p}, \ldots, \mathbf{p}_{R1}, \ldots, \mathbf{p}_{RN_p}]^T \tag{6.21}$$

where $\mathbf{p}_{rh}$ is the vector of permutation h, which contains the N DOAs label, of row r. The matrix $\mathbf{T}$ has a dimension of $N \times RN_p$. In second step, the combinations of row indices of matrix $\mathbf{T}$ give the $(N!)^{(R-1)}$ possible combinations. We consider the combinations of $R-1$ groups, each one composed by N_p permutation elements, assuming that one member (the index row of matrix $\mathbf{T}$) from each of the $R-1$ groups is used in each combination and assuming that the order is not a distinguishing factor. The possible combinations are $N_p^{(R-1)} = N!^{(R-1)}$. Hence, a combination label matrix $\mathbf{C}$ of $RN \times N!^{(R-1)}$ dimension is used to store the angle label of all combinations

$$\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{N!^{(R-1)}}]^T \tag{6.22}$$

where $\mathbf{c}_z$ is the vector, which contains the RN DOA labels, of combination z. The values of $\mathbf{C}$ are used to calculate the spectral distance estimation of all combination using the equation (6.20), and accordingly, we define the ISPC vector of all combinations

$$\mathbf{d} = \begin{bmatrix} d_1 & d_2 \ldots d_{(N!)^{R-1}} \end{bmatrix}. \tag{6.23}$$

Finally, the index of the minimum value of the vector $\mathbf{d}$ identifies the target combination

$$\widehat{z} = \underset{index}{\operatorname{argmin}} \mathbf{d} \tag{6.24}$$

and the DOAs matrix $\widehat{\Theta}$ is estimated by ordering the label matrix $\mathbf{B}$ with the combination $\mathbf{C}(\widehat{z})$.

The overall procedure of ISPC method is summarized in the following steps:

1. DOA estimation and creation of the matrix $\underline{\widehat{\mathbf{\Theta}}}$ (6.4). In practice, the matrix does not always present all the DOA values; in these cases, the missing value of array r can be replaced with another DOA value of array r.

2. Building of the label matrix $\mathbf{B}$ (6.17) and calculation of ISPs and the matrix $\mathbf{P}$ (6.11).

3. Estimation of the ISPC between all ISP pairs of arrays and creation of the adjacency matrix $\mathbf{A}$ (6.19).

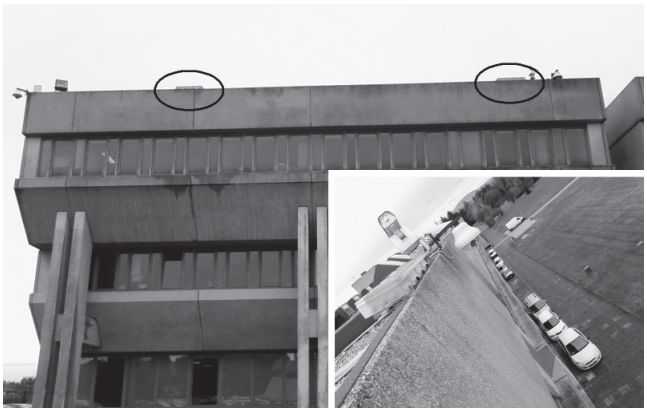

Figure 6.3: *The prototype installed on the roof of the University building. The two arrays are encircled.*

4. Calculation of the permutations matrix $\mathbf{T}$ (6.21) and the all DOA combination matrix $\mathbf{C}$ (6.22).

5. Calculation of the vector $\mathbf{d}$ (6.23) that contains the spectral distance estimation d_z with ($z = 1, \ldots, N!^{(r-1)}$).

6. Finding the minimum value of $\mathbf{d}$ (6.24) and using the index value $\widehat{z}$ in the matrix $\mathbf{C}$ to properly order the matrix $\underline{\widehat{\mathbf{\Theta}}}$ and estimate the matrix $\widehat{\mathbf{\Theta}}$.

6.2.2 System Setup

The prototype for two-dimensional localization has been installed on the roof of the building that houses the Computer Science Department in Udine (see Figure 6.3). The prototype includes two linear arrays, each composed of four omnidirectional microphones. The arrays are located 11.4 m apart at a height of 12.1 m above the plane. The sample rate of the digital system is 48 kHz, and the microphone distance is 25 cm. The prototype consists of two parallel processing lines, corresponding to the left and right arrays (Figure 6.4). The first processing step is the TDOAs estimation and is followed by the DOAs estimation (if we use TDE methods) or direct DOAs estimation (if we use SRP based methods). In the second step, the two-dimensional coordinates of the source can be estimated by combining the DOAs at the left and right arrays. If more than one source is identified, a beamformer and a spectral distance comparison provide a guide to solve the problem of associating the DOAs of the left array with those of the right array.

The assumed DOA range is -90° +90°, where zero is in front of the array and the microphone reference is the first from left. The calculation of the two-dimensional position of the source is a simple triangulation problem. However, we must consider that the two arrays are not coincident with the plane

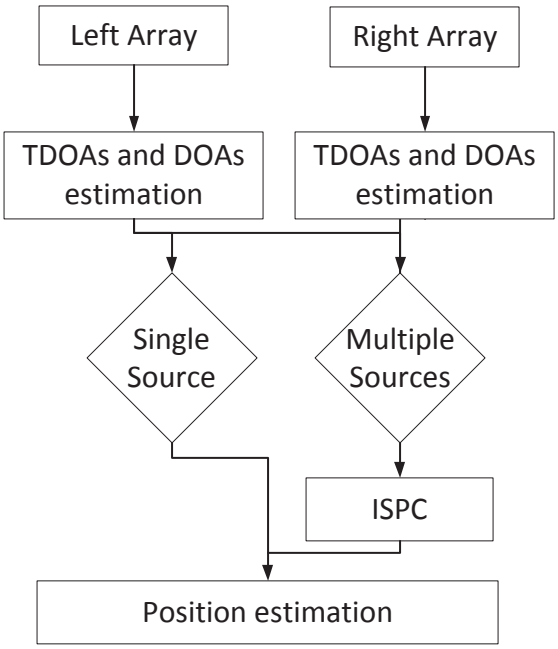

Figure 6.4: *The block diagram of the processor showing the data flow of all of the tasks of the experimental far-field prototype.*

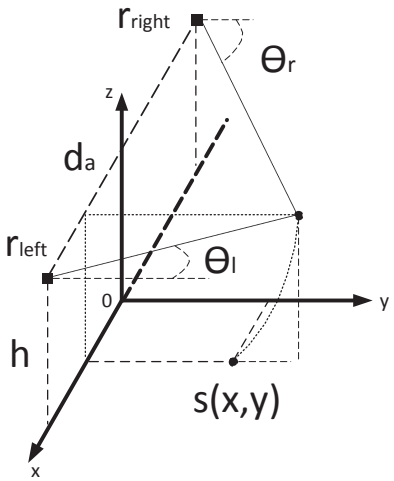

Figure 6.5: *The two-dimensional position of the source of the experimental far-field prototype.*

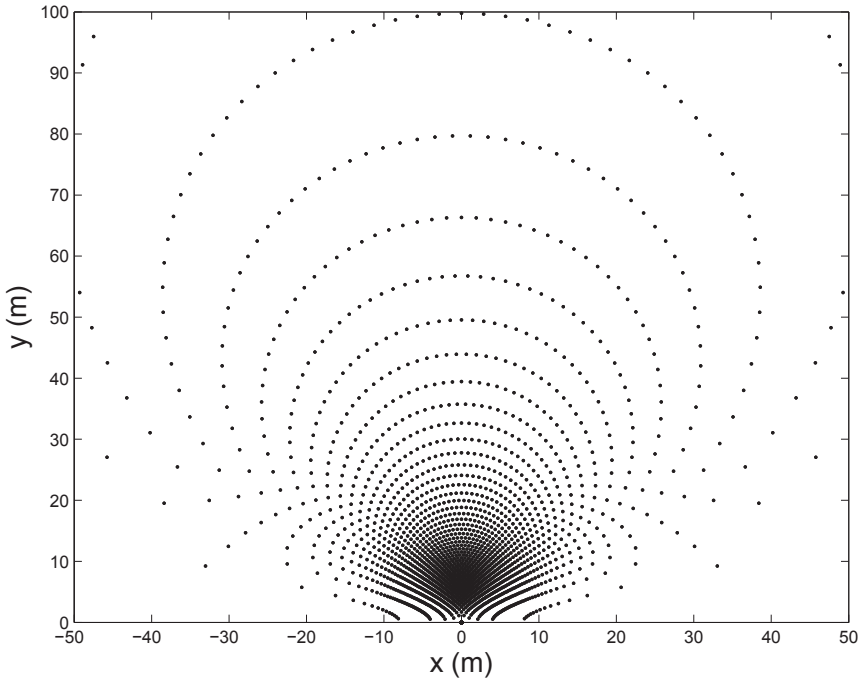

Figure 6.6: *The x-y sample space position of the plane of interest.*

of interest but are placed at a certain height. We must consider that the possible points identified by the DOA are located on a cone surface whose vertex is placed in the array and whose axis is the straight line joining the two arrays.

Every array represents a cone: the intersection of the two cones is represented by a circumference. The intersection point between the circumference and the plane of interest is the estimation of the source distance from arrays (see Figure 6.5). Hence, we consider d_a to be the distance of the arrays, h to be the height of arrays above the plane of interest, and θ_r and θ_l to be the DOA estimated on the right and left array

$$x = \frac{d_a}{2}\Big(\frac{\tan\theta_l + \tan\theta_r}{\tan\theta_l - \tan\theta_r}\Big) \tag{6.25}$$

$$y = \sqrt{\Big(\frac{d_a}{\tan\theta_l - \tan\theta_r}\Big)^2 - h^2}. \tag{6.26}$$

The spatial resolution of the system depends on the distance between the microphones, the distance between the arrays and the sample frequency of digital system. However, in this case, Figure 6.6 shows the possible xy coordinates of the prototype. The spatial resolution tends to decrease with an increasing distance from the arrays and an increasing angle from the axis perpendicular to the array.

6.2.3 Comparison of DOA Estimation Methods

DOA estimation is a crucial step for ASL systems. Our goal is to obtain a complete matrix 6.4 of all values for each source and then, as accurately as possible, a DOA estimate. To evaluate the algorithms under different SNR conditions, a simulation was carried out using three sound sources: a motor car, a female voice, and a shot gun. The methods evaluated were SRP, SRP with Dolph-Chebyshev windowing, SRP-PHAT, MVDR, MVDR-DL, MUSIC, MCCC with PHAT filter, Root-Music, and ESPRIT. For the latter two methods, which directly provide the value of the angle for each narrowband frequency, the average value of the DOA was estimated for the frequency range between 20 Hz and the aliasing spatial limit (which, in this case, is 675 Hz). To reduce the incorrect values that tend to augment the error estimation, k-means clustering [MacQueen, 1967] was performed and the highest-density cluster was considered. The SRP-PHAT was calculated using equation (3.59); thus, DOA estimation is also used here rather than building an acoustic map. The performance evaluation for a linear array of four microphones was performed, and the Percentage Success Rate (PSR), the mean value of DOA estimate and the Root Mean Square Error (RMSE) were recorded. Assuming a DOA angle θ and its estimation $\widehat{\theta}$, the PSR is defined as

$$PSR = 100\Big(\frac{\text{number of correct estimations}(\widehat{\theta}_h = \theta)}{\text{total number H of DOA estimations}}\Big) \tag{6.27}$$

where H is the total number of window frames and $\widehat{\theta}_h$ is the DOA estimation in the window frame h. The mean is

$$Mean = \frac{\sum_{h=1}^{H} \widehat{\theta}_h}{H} \tag{6.28}$$

and the RMSE is

$$RMSE = \sqrt{\frac{\sum_{h=1}^{H} (\theta - \widehat{\theta}_h)^2}{H}}. \tag{6.29}$$

For all tests, a DOA of $\theta = 23^\circ$ was assumed. Tables 6.1, 6.2 and 6.3 show the results for the motor car, female voice and gun shot sound, respectively.

The results indicate that for the sound of the car engine, MCCC is the most accurate approach up to a SNR of 0 dB with PSR=57.98 % and RMSE=0.98, and with a maximum error of 1 degree; for the sound of the voice, MVDR-DL had the best performance. For the shot gun, up to a SNR of 5 dB, MCCC is very accurate, and when we have a SNR of zero dB, the RMSE reaches a maximum value of 7.7; in contrast, for MVDR-DL, a value of 0.77 was obtained. In general, with a linear array of four microphones, MVDR-DL has the best performance for a low SNR, while for SNR $>$ 5 dB MCCC has the highest PSR. Figures 6.7 and 6.7 represent frames of analysis for the motor car sound at 20 dB and at 5 dB. The peaks represent the estimated direction of arrival of the source at the microphone array. As

	PSR (%)								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	100	100	94,52	36,99	100	98,63	100	54,79	64,38
20	100	98,63	71,23	20,55	100	89,04	100	52,05	53,42
10	65,75	56,16	52,05	32,88	80,82	71,23	100	36,99	39,73
5	49,32	46,58	41,10	23,29	54,79	58,90	87,67	28,77	31,51
0	46,58	38,36	31,51	19,18	54,79	41,10	57,53	19,18	13,70
	RMSE								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	0	0	0,23	1,00	0	0,12	0	0,67	0,60
20	0	0,12	0,54	2,27	0	0,33	0	0,69	0,68
10	0,59	0,66	0,69	2,56	0,44	0,54	0	0,91	0,94
5	0,71	0,73	0,96	2,97	0,67	0,64	0,61	1,33	1,45
0	0,86	1,30	1,27	4,85	0,67	0,86	0,98	2,57	2,88
	Mean								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	23	23	23,05	23,45	23	23,01	23	23,45	23,36
20	23	23,01	23,29	23,01	23	23,11	23	23,48	23,44
10	23,34	23,44	23,48	23,34	23,19	23,29	23	23,53	23,51
5	23,48	23,51	23,56	23,55	23,45	23,41	23,12	23,41	22,89
0	23,30	23,30	23,37	24,19	23,45	23,22	23,11	23,44	23,19

Table 6.1: Comparison of the DOA estimation for the car motor sound.

we can see, the SRP and SRP-DC have a very wide main lobe; and, thus, the resolution of localization in this case is very low, and its application is not suitable for multi-source cases. MVDR is difficult to use in practice, as we note in the case of the gun-shot, which is an impulsive sound, that the values are completely incorrect.

SRP-PHAT, MVDR-DL, MUSIC and MCCC have a clearly visible peak with good resolution. Root-Music and ESPRIT are especially useful in applications of narrowband signals, and their performance degrades with a low SNR. In addition, for multiple sources, we need to estimate the number of active sources by analyzing the matrix eigenvector. This procedure is also necessary for the MUSIC method. The number of sources that we can estimate is also limited by the number of microphones in the array. Hence, SRP-PHAT, MCCC and MVDR-DL are suitable for DOA estimation with a small size array. From the computational point of view the SRP-PHAT (especially calculated as the sum of the GCC-PHAT microphone pairs) and MCCC have a lower demand for computation and are therefore suitable in the case of real-time applications. MVDR-DL requires a greater computational cost, as the power is calculated for each steered direction and for each frequency bin considered.

	PSR (%)								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	98,90	98,90	70,33	56,04	98,90	91,21	100	64,84	70,33
20	87,91	76,92	43,96	30,77	95,60	86,81	91,21	62,64	67,03
10	52,75	57,14	41,76	38,46	61,54	54,95	64,84	46,15	46,15
5	38,46	26,37	18,68	29,67	45,05	37,36	49,45	37,36	37,36
0	38,46	30,77	23,08	21,98	53,85	41,76	32,97	21,98	20,88
	RMSE								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	0,10	0,10	0,54	10,09	0,10	0,30	0	0,59	0,54
20	0,35	0,48	0,90	1,10	0,21	0,36	0,47	0,66	0,70
10	0,98	1,13	1,48	5,81	0,65	0,67	1,79	7,08	0,92
5	1,92	2,79	10,19	2,34	0,88	0,92	7,51	25,52	24,46
0	3,26	3,50	11,12	10,21	1,48	2,87	13,81	40,43	42,07
	Mean								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	23,01	23,01	23,30	22,34	23,01	23,09	23,00	23,35	23,30
20	23,12	23,23	23,48	23,51	23,04	23,13	23,04	23,40	23,33
10	23,44	23,40	23,76	23,02	23,33	23,45	23,34	24,05	23,41
5	23,75	23,71	23,81	23,49	23,55	23,60	22,96	20,08	19,27
0	22,67	22,30	20,35	22,97	23,20	22,92	19,73	15,56	16,45

Table 6.2: Comparison of the DOA estimation for the female voice sound.

	PSR (%)								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	100	100	81,48	3,70	100	88,89	100	62,96	40,74
20	100	100	62,96	18,52	100	96,30	100	51,85	44,44
10	62,96	59,26	22,22	14,81	74,07	77,78	96,30	29,63	29,63
5	66,67	59,26	14,81	18,52	70,37	74,07	77,78	25,93	25,93
0	70,37	70,37	48,15	14,81	74,07	48,15	51,85	7,41	14,81
	RMSE								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	0	0	0,43	63,15	0	0,33	0	0,61	0,77
20	0	0	0,61	54,53	0	0,19	0	1,07	1,07
10	0,61	0,64	1	32,36	0,51	0,47	0,19	13,12	12,60
5	0,58	0,64	1,04	51,24	0,54	0,51	0,47	43,62	28,77
0	0,86	0,84	0,92	43,95	0,77	1,39	7,70	46,63	53,04
	Mean								
SNR	SRP	SRP-DC	SRP-PHAT	MVDR	MVDR-DL	MUSIC	MCCC	Root-Music	ESPRIT
30	23	23	23,19	-27,41	23	23,11	23	23,37	23,59
20	23	23	23,37	-15,22	23	23,04	23	23,26	23,48
10	23,37	23,41	23,63	7,59	23,26	23,22	22,96	25,15	25,04
5	23,19	22,96	23,56	-9,52	23,30	23,26	22,78	18,11	35,59
0	23,30	22,89	22,74	-2	23,22	23	20,93	23	21,22

Table 6.3: Comparison of the DOA estimation for the gun shot sound.

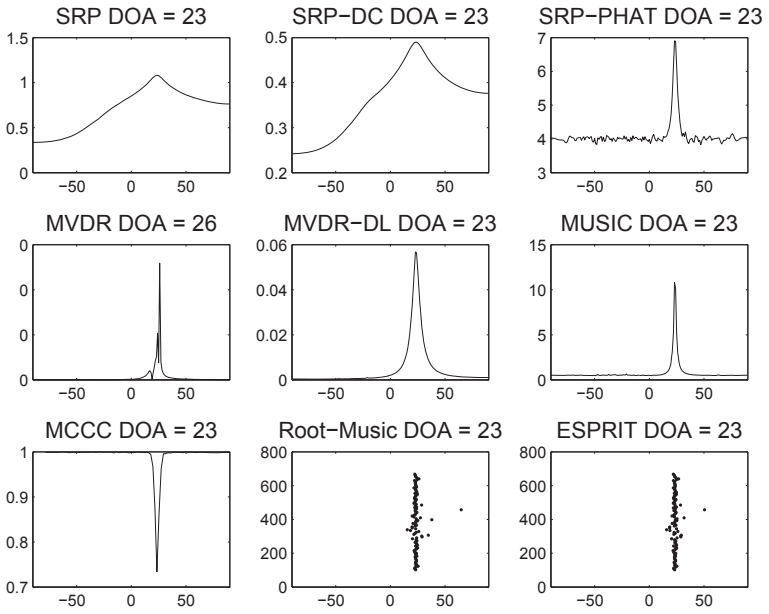

Figure 6.7: *Frame analysis of the DOA method comparison, SNR = 20 dB.*

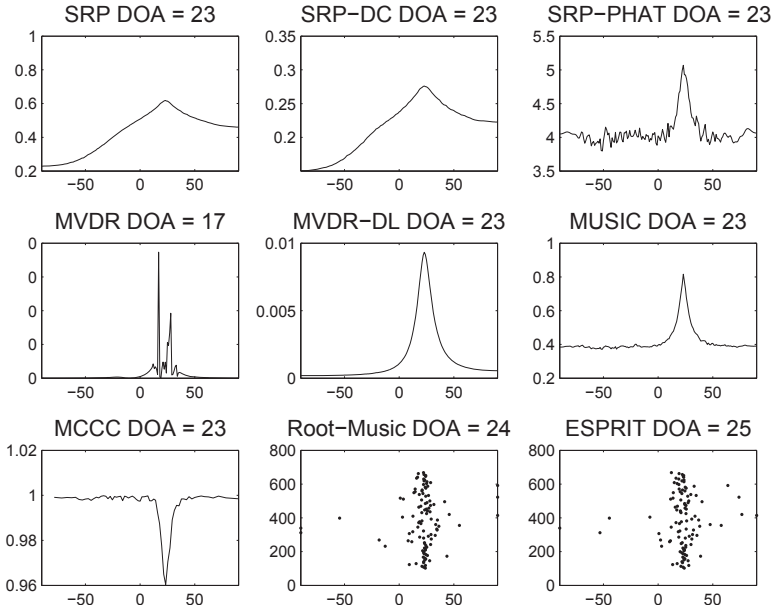

Figure 6.8: *Frame analysis of the DOA method comparison, SNR = 5 dB.*

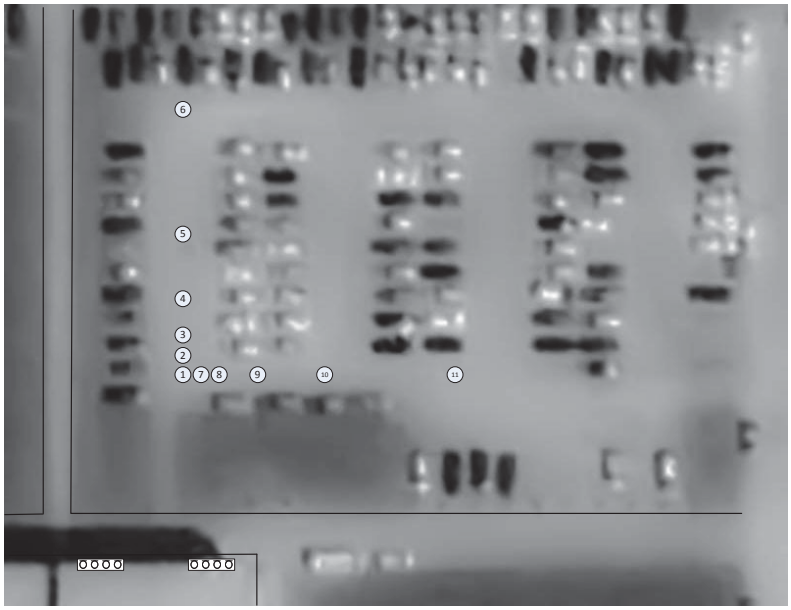

Figure 6.9: *Map of the study area indicating the position of arrays and sources.*

6.2.4 Experimental Results with ISPC

Experiments were conducted that consider the area of analysis of 60×90 m shown in Figure 6.9, i.e., the parking lots of the University. Eleven zones of acoustic source positioning are considered. The sources used are a human voice (S1), a hammer striking an iron bar (S2) and a motor car (S3). The hammer striking an iron bar is the short event sound.

Two types of experiments were performed. The first type used sounds with different spectral content, named test P. The second type, however, used sounds with similar spectral content, named test C. Test P is composed of eight parts (P1, P2, ..., P8), each one with three sources placed in different positions (see table 6.5). The zero of the xy axes reference is located in the middle of the distance between the two arrays.

In various parts of test P, the sources were positioned at increasing distances along the y axis (P1 - P4) and the x axis (P5 - P8): e.g., in part P1, we placed S1 in 1, S2 in 2, and S3 in 3. The Table 6.4 shows the xy coordinates of the points and the position estimated by the microphone array prototype, which reported the mean value of the x and y coordinates and the RMSE of the estimation.

MCCC is used in the DOAs estimation. Tables 6.6 and 6.7 summarize the results, comparing the localization success rate (as a percentage) with different beamforming algorithms (SRP, SRP with Dolph-Chebyshev windowing, MVDR, and MVDR-DL) and spectral distance functions (LP, IS, RMS, COSH).

Source Label	x	y	Mean x	Mean y	RMSE x	RMSE y
1	1,5	20	1,6	19,8	0,19	0,95
2	1,5	23	1,6	23,5	0,18	1,04
3	1,5	26	1,53	27,8	0,24	2,06
4	1,5	32	1,49	30,05	0,16	2,32
5	1,5	38	1,48	38,21	0,18	1,74
6	1,5	52	1,44	51,8	0,06	0,2
7	4,5	20	4,57	21,2	0,23	1,9
8	7,5	20	6,5	21,15	1,07	3,45
9	10,5	20	11	22,25	0,5	2,25
10	20	20	18,5	23,27	1,8	4,03
11	30	20	27,4	17,06	7,3	12,95

Table 6.4: Position referring to Figure 6.9 and the mean value estimation and RMSE.

The localization success rate is the ratio between the number of correct combinations and the Number Of Ambiguities (NOA) for that part of the test. The NOA is the number of frames in which we have ambiguity to properly associate the DOAs to the sources, i.e., the associations are incorrect in practice. The audio signal frame was divided into 17,5 ms overlapping and a Hanning-windowed with a length of 140 ms. The parking area, where the tests were conducted, is a public area. Thus, we must to consider that there are other sources in the acoustic scene: other sounds of cars that are moving in the parking area and in the nearby streets. Tables 6.8 and 6.9 summarize the results of all tests, labeled T. The three Frequency Range (FR) for the spectral distance estimation are 20-675 Hz, 20-2000 Hz and 20-8000 Hz. The frequency value of 675 takes into account the spatial aliasing limit, which, in our case, is $f = c/(2d) = 337/(2 \cdot 0.25) = 675$ Hz. The phenomena of spatial aliasing implies that the main lobe of the beamformer has a set of identical copies, called grating lobes. The appearance of grating lobes is a function of both microphone spacing and incident frequency. When fully visible, a grating lobe is equal in amplitude to the main lobe of the array. This fact reduces the array response, and, therefore, by defining the spatial sampling requirement and removing the grating lobes, we obtain a greater efficiency in the ISPC.

In test C, two car sounds were used. The test was performed by placing two car sources in 1 and 7, as shown in Figure 6.9. In Figure 6.10, the total localization success rate of the complex acoustic scene (T) is compared with the results of the case of two cars to test whether our algorithm works even with similar spectral content sounds. We note that the accuracy decreases, especially with regard to the RMS and COSH functions, and this result highlights the limitation of the proposed approach in the case of spectrally similar sources.

The best results were obtained with the RMS log spectral distance function and FR=[20-675] Hz. MVDR-DL has the greatest capacity for location with 90.9 %, SRP-DC with 88.4 % and SRP with 88.2%.

Test Label	S1	S2	S2
P1	1	2	3
P2	1	3	4
P3	1	4	5
P4	1	5	6
P5	1	7	8
P6	1	8	9
P7	1	9	10
P8	1	10	11

Table 6.5: The position of the sources of the eight tests (P).

	(Hz)	Localization Success Rate (%)								
	FR	SPR-LP	SPR-IS	SPR-RMS	SPR-COSH	SPR-DC-LP	SPR-DC-IS	SPR-DC-RMS	SPR-DC-COSH	NOA
P1	20-675	33,6	49,6	85,8	82,3	26,5	39,8	86,7	77,0	113
	20-2000	59,3	62,8	77,9	68,1	57,5	57,5	71,7	66,4	
	20-8000	56,6	57,5	76,1	63,7	52,2	52,2	71,7	60,2	
P2	20-675	58,3	58,3	75,0	41,7	75,0	41,7	66,7	41,7	12
	20-2000	58,3	66,7	41,7	41,7	50,0	33,3	33,3	25,0	
	20-8000	66,7	66,7	66,7	33,3	66,7	33,3	66,7	58,3	
P3	20-675	24,3	43,2	89,2	83,8	29,7	40,5	91,9	91,9	37
	20-2000	59,5	59,5	86,5	81,1	48,6	48,6	75,7	75,7	
	20-8000	70,3	70,3	78,4	67,6	64,9	62,2	73,0	59,5	
P4	20-675	28,9	42,2	93,3	86,7	26,7	44,4	88,9	73,3	45
	20-2000	40,0	46,7	73,3	57,8	53,3	57,8	71,1	57,8	
	20-8000	77,8	77,8	64,4	60,0	62,2	62,2	73,3	53,3	
P5	20-675	36,8	52,6	73,7	73,7	31,6	57,9	78,9	68,4	19
	20-2000	42,1	47,4	73,7	52,6	31,6	31,6	78,9	63,2	
	20-8000	73,7	73,7	73,7	73,7	57,9	57,9	73,7	63,2	
P6	20-675	30,8	45,1	90,2	81,2	32,3	41,4	91,7	80,5	133
	20-2000	39,8	42,9	86,5	68,4	43,6	45,9	76,7	65,4	
	20-8000	48,9	48,9	71,4	57,1	48,9	48,9	72,9	69,2	
P7	20-675	29,5	43,2	79,5	81,8	25,0	31,8	79,5	79,5	44
	20-2000	34,1	36,4	90,9	72,7	22,7	22,7	70,5	65,9	
	20-8000	50,0	50,0	65,9	65,9	34,1	34,1	61,4	59,1	
P8	20-675	29,2	51,4	95,8	88,9	11,1	40,3	94,4	83,3	72
	20-2000	59,7	61,1	62,5	55,6	55,6	55,6	69,4	68,1	
	20-8000	65,3	65,3	81,9	62,5	59,7	61,1	84,7	65,3	

Table 6.6: Results of the eight tests (P) with SRP and SRP-DC.

	(Hz)	Localization Success Rate (%)								
	FR	MVDR-LP	MVDR-IS	MVDR-RMS	MVDR-COSH	MVDR-DL-LP	MVDR-DL-IS	MVDR-DL-RMS	MVDR-DL-COSH	NOA
P1	20-675	36,3	37,2	70,8	41,6	39,8	65,5	88,5	87,6	113
	20-2000	53,1	54,9	61,9	54,9	53,1	73,5	77,9	77,9	
	20-8000	52,2	52,2	63,7	46,9	59,3	69,0	79,6	77,0	
P2	20-675	58,3	50,0	50,0	58,3	50,0	75,0	75,0	41,7	12
	20-2000	41,7	41,7	41,7	41,7	83,3	75,0	50,0	41,7	
	20-8000	83,3	83,3	66,7	75,0	83,3	41,7	58,3	50,0	
P3	20-675	21,6	32,4	83,8	40,5	35,1	78,4	89,2	91,9	37
	20-2000	45,9	40,5	67,6	45,9	43,2	83,8	81,1	81,1	
	20-8000	51,4	51,4	59,5	62,2	35,1	56,8	51,4	59,5	
P4	20-675	37,8	33,3	68,9	53,3	44,4	44,4	97,8	95,6	45
	20-2000	42,2	40,0	66,7	62,2	40,0	55,6	82,2	77,8	
	20-8000	48,9	48,9	73,3	55,6	55,6	66,7	71,1	71,1	
P5	20-675	52,6	57,9	68,4	63,2	36,8	26,3	78,9	73,7	19
	20-2000	63,2	63,2	63,2	57,9	68,4	52,6	73,7	78,9	
	20-8000	57,9	57,9	78,9	47,4	52,6	52,6	73,7	68,4	
P6	20-675	45,9	48,1	74,4	55,6	63,2	91,0	94,0	93,2	133
	20-2000	51,9	52,6	72,9	54,9	53,4	83,5	91,0	84,2	
	20-8000	47,4	47,4	67,7	59,4	45,9	70,7	74,4	72,2	
P7	20-675	56,8	50,0	61,4	54,5	50,0	84,1	84,1	84,1	44
	20-2000	56,8	50,0	63,6	59,1	38,6	54,5	70,5	52,3	
	20-8000	47,7	47,7	59,1	38,6	52,3	47,7	47,7	56,8	
P8	20-675	40,3	41,7	81,9	56,9	47,2	52,8	95,8	88,9	72
	20-2000	50,0	56,9	73,6	48,6	27,8	75,0	70,8	72,2	
	20-8000	55,6	55,6	63,9	51,4	48,6	62,5	66,7	73,6	

Table 6.7: Results of the eight tests (P) with MVDR and MVDR-DL.

	(Hz)	Localization Success Rate (%)								
	FR	SPR-LP	SPR-IS	SPR-RMS	SPR-COSH	SPR-DC-LP	SPR-DC-IS	SPR-DC-RMS	SPR-DC-COSH	NOA
T	20-675	31,4	47,2	88,2	82,1	27,4	40,8	88,4	78,7	475
	20-2000	49,1	52,2	78,3	65,5	47,8	48,4	72,2	65,1	
	20-8000	59,2	59,4	73,5	61,5	53,3	52,4	73,3	62,7	

Table 6.8: Summary of the results of all tests (P) with SRP and SRP-DC.

	(Hz)	Localization Success Rate (%)								
	FR	MVDR-LP	MVDR-IS	MVDR-RMS	MVDR-COSH	MVDR-DL-LP	MVDR-DL-IS	MVDR-DL-RMS	MVDR-DL-COSH	NOA
T	20-675	41,7	42,5	72,8	51,4	48,6	70,1	90,9	88,4	475
	20-2000	51,2	51,6	67,4	54,1	47,4	73,1	79,6	75,8	
	20-8000	51,6	51,6	65,7	53,1	51,4	64,0	69,5	70,3	

Table 6.9: Summary of the results of all tests (P) with MVDR and MVDR-DL.

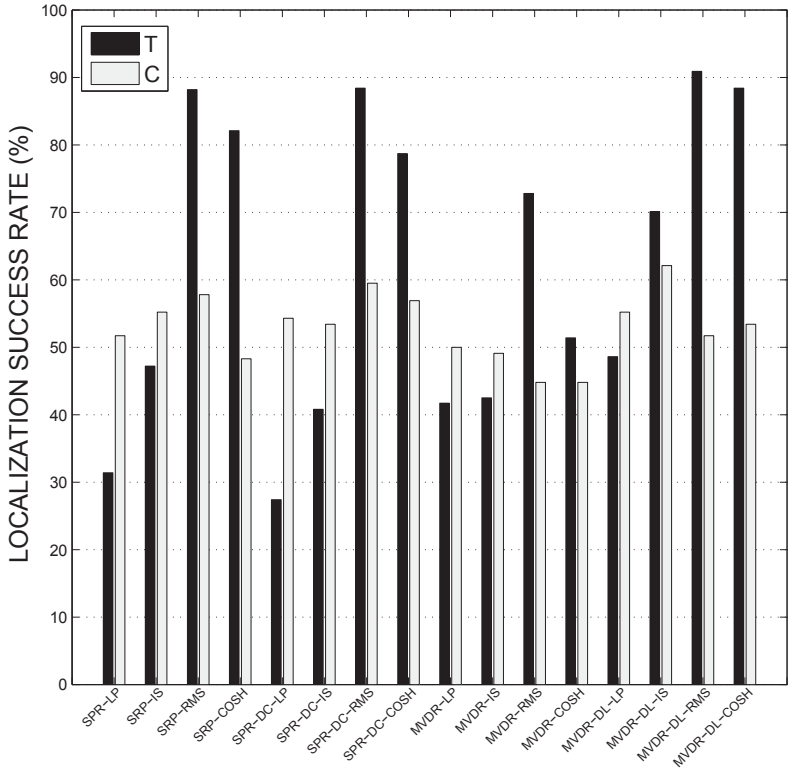

Figure 6.10: *Comparison of the summary results T with the result C of the car-car test with NOA=116. FR=[20, 675] Hz.*

6.2.5 Summary

The novel Incident Signal Power Comparison algorithm used to solve the ambiguous problem of correctly linking the DOAs from different arrays to the same source was presented. Experimental results have shown that this approach can be a solution for multi-source localization that requires a frame-to-frame analysis. This approach is particularly advantageous in identifying sounds of short duration that can be difficult to determine using a traditional Bayesian filter. The limits of this approach were presented in the case of two sources with a similar spectral content. For localization enhancement, clustering can be used to improve source position estimates. However, we must emphasize that ISPC systems can integrate with Bayesian filtering, and it can be helpful in cases that require detailed analysis over time as well as in cases in which the Bayesian filter can fail.

6.3 Near-Field Application

A framework for the localization of pseudo-periodic sounds in moderate reverberant and noisy environments is now described. This framework consists of an adaptive parameterized GCC-PHAT with a zero-crossing rate threshold, a pre-processing with a Wiener filter, and post-processing with a Kalman filter. The novel architecture can be used as a digital musical interface [Salvati *et al.*, 2011b] [Salvati *et al.*, 2011a], which allows a performer to plan and conduct the expressivity of a performance by controlling an audio processing module in real-time through the spatial movement of a sound source (i.e., voice, traditional musical instruments, and sounding mobile devices).

6.3.1 System Architecture for Pseudo-Periodic Sound Localization

The architecture consists of combining signal processing algorithms for robust sound localization. The array system is composed of three supercardioid microphones arranged in an uniform linear placement. In this way, we can localize a sound source in a plane (three microphones are the bare minimum). Signal processing algorithms estimate the sound source position in a horizontal plane by providing its Cartesian coordinates.

With musical sounds that are mainly harmonics, the GCC-PHAT approach does not work well because the PHAT filter normalizes the GCC according to the spectrum magnitude. The problem is the presence of noise; in fact, under ideal conditions we would be able to estimate the TDOA. Figure 6.13 shows a simulation with a sinusoidal wave of 300 Hz; under the condition of SNR=100 dB the peak is well defined, while when SNR=50 dB, the TDOA is impossible to estimate. We note that CC improves the performance of the TDOA estimation; however, CC suffers from the effects of moderate reverb and

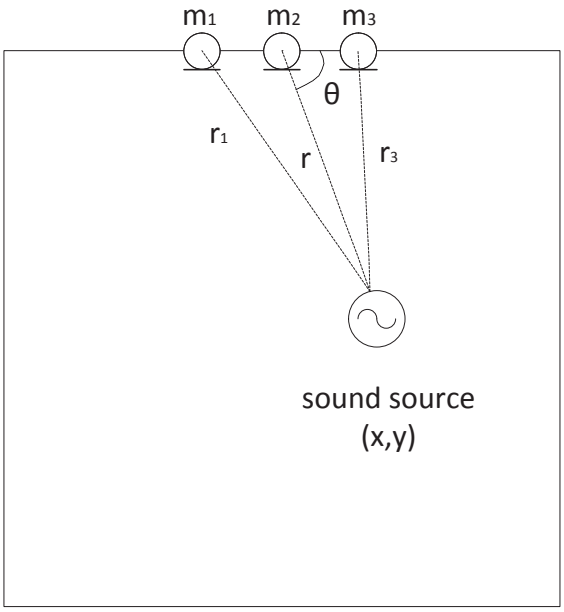

Figure 6.11: *The xy plane of interest.*

auto-correlation, so it is in practice inappropriate for localization. Thus, when noise and reverberation are present, new considerations are required to estimate the TDOA for pseudo-periodic signals. The proposal is to use a parameterized GCC-PHAT that weighs the contribution of the PHAT filtering, depending on the threshold of the ZCR parameters. In this way, we balance the CC with the improved filter of PHAT. This operation allows for the identification of source, but other filters are required to obtain a usable stable value of the position in real applications.

Therefore, a de-noise algorithm based on a Wiener filter (see equation (4.5)) is used to improve the SNR of the signals. When the maximum peak detection does not observe any source, an average estimation of noise is computed (a noise print), which will be subtracted from all three signals before the TDOA estimation task.

Then, starting from the estimated TDOA between microphones $\widehat{\tau}_{12}$ and $\widehat{\tau}_{23}$, it is possible to calculate the coordinates of the source using geometric constraints. In a near-field environment, we have

$$\begin{aligned} \widehat{x} &= r\cos(\theta) \\ \widehat{y} &= r\sin(\theta) \end{aligned} \tag{6.30}$$

where the axis origin is placed in microphone 2, r is the distance between the sound source and micro-

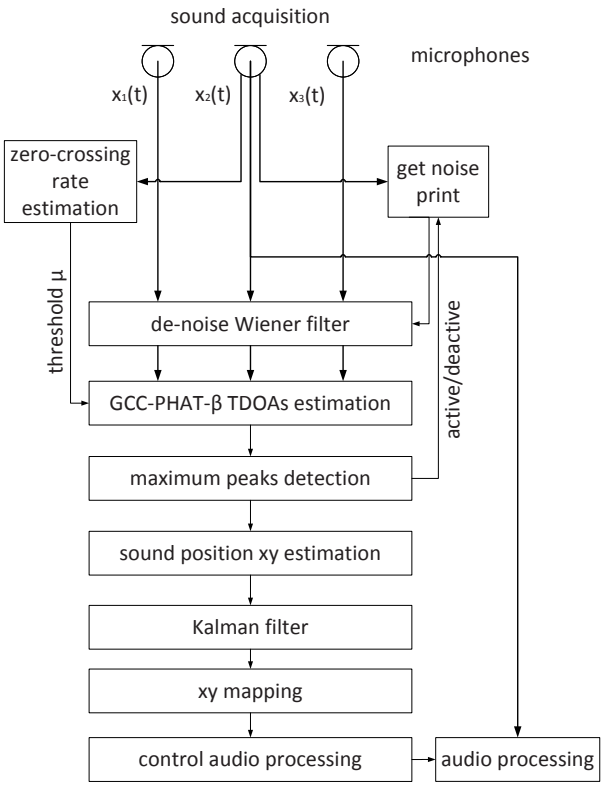

Figure 6.12: *Block diagram of interface.*

phone 2, and θ is the angle between r and the x axis

$$\theta = \arccos\Big(\frac{c(\widehat{\tau}_{12} + \widehat{\tau}_{23})(\widehat{\tau}_{12}\widehat{\tau}_{23}c^2 - d^2)}{d(2d^2 - c^2(\widehat{\tau}_{12}^2 + \widehat{\tau}_{23}^2))}\Big) \tag{6.31}$$

$$r = \frac{\widehat{\tau}_{12}^2 c^2 - d^2}{2(\widehat{\tau}_{12}c + d\cos\theta)} \tag{6.32}$$

where c is speed of sound and d is the distance between microphones.

Finally, a second filter provides a more accurate estimate and tracking of the source position if there is movement using the Kalman theory (see section (5.2)). The Kalman filter is also able to provide an estimate of the position of the source, if the TDOA estimation task misses the target in a frame of analysis. Figure 6.12 summarizes the system architecture.

6.3.2 Adaptive Parameterized GCC-PHAT with Zero-Crossing Rate Threshold

The PHAT weighting can be generalized to parametrically control the level of influence from the magnitude spectrum [Donohue *et al.*, 2007]. This transformation will be referred to as the PHAT-β and

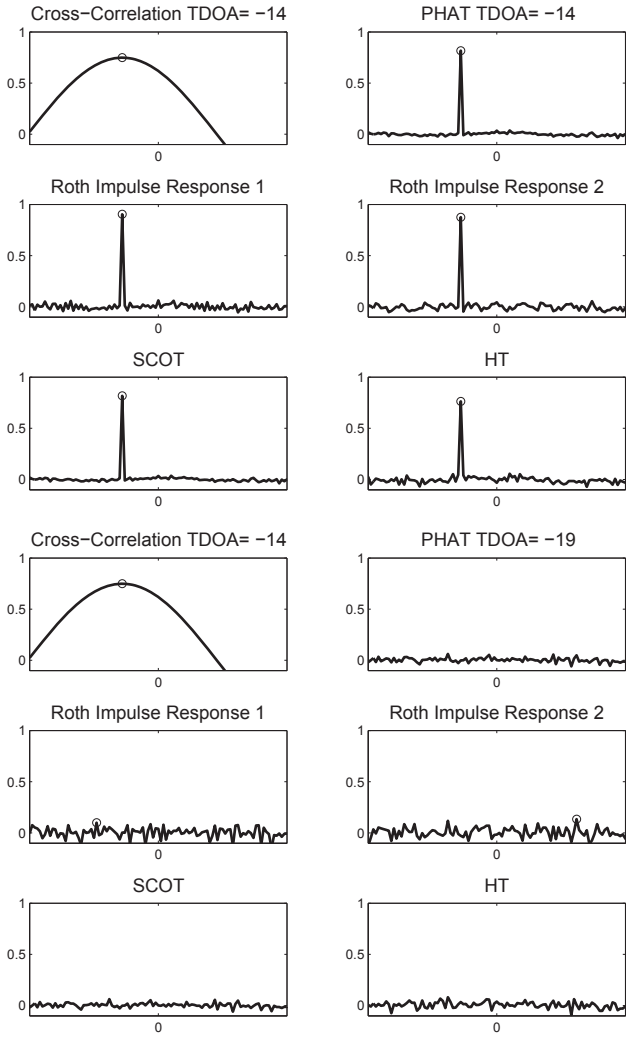

Figure 6.13: *GCC of sinusoidal wave of 300 Hz, SNR = 100 dB (figures on top) and SNR = 50 dB (figures on bottom)*

defined as

$$\Psi_{\mathrm{PHAT}-\beta}(f) = \frac{1}{|S_{x_1x_2}(f)|^{\beta}} \tag{6.33}$$

where β varies between 0 and 1. When $\beta = 1$, equation (6.33) becomes the conventional PHAT and the modulus of the Fourier transform becomes 1 for all frequencies; when $\beta = 0$, the PHAT has no effect on the original signal, and we have the cross-correlation function.

Therefore, in the case of harmonic sounds, we can use an intermediate value of β so that we can detect the peak to estimate the time delay between signals, and can have a system, at least in part,

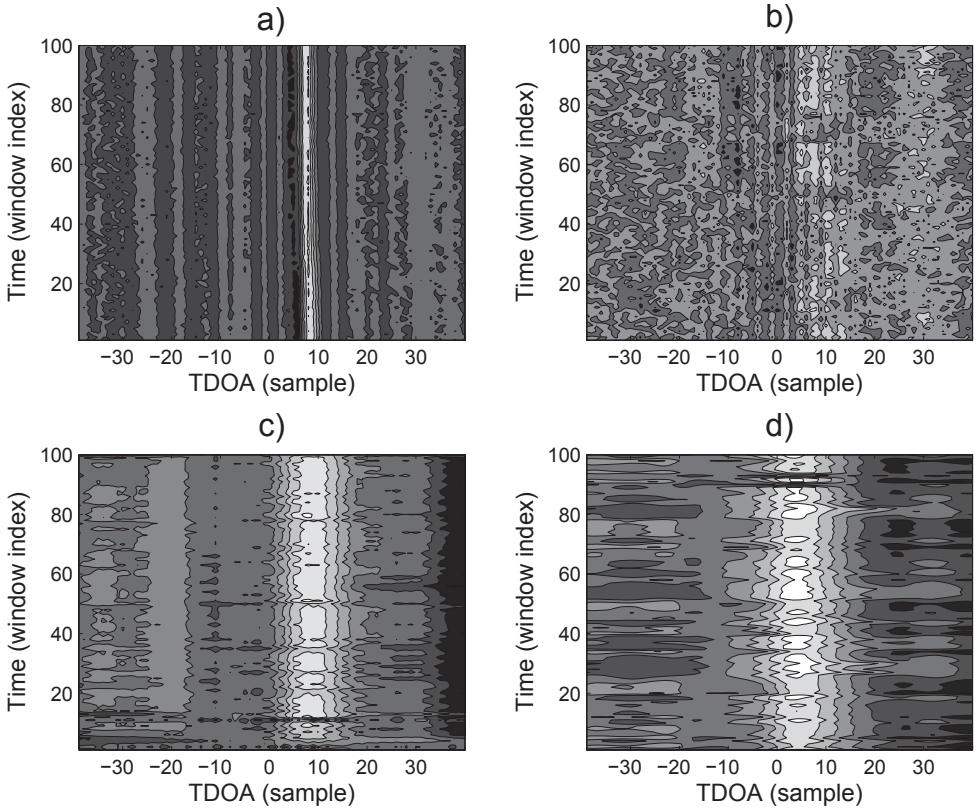

Figure 6.14: *Comparison of the parameterized PHAT-β TDOA estimation performance. a) White noise played on mobile device, $\beta = 1$. b) Flute, $\beta = 1$. c) Flute, $\beta = 0.65$. d) Flute, $\beta = 0.65$ and de-noise Wiener filter.*

that exploits the benefits of PHAT filtering to improve performance in moderately reverberant and noisy environments. To adapt the value of β, we can use the ZCR to determinate if the sound source is periodic. ZCR is a very useful audio feature and is defined as the number of times that the audio waveform crosses the zero axis

$$ZCR(k) = \frac{1}{2L} \sum_{i=1}^{L} |sgn(x(k+i)) - sgn(x(k+i-1))| \tag{6.34}$$

where sgn(x) is the sign function.

Then, we can express the adaptive parameterized GCC-PHAT, identifying by experimental tests a suitable threshold μ such as

$$\begin{cases} \beta = 1, \; if \; ZCR \geq \mu \\ \beta < 1, \; if \; ZCR < \mu \end{cases} \tag{6.35}$$

Test Label	Mean	SD
a	6.96	0.2
b	9.1	8.97
c	7.26	0.8
d	3.2	0.7

Table 6.10: The value of the mean and RMSE related to the experiments in Figure 6.14.

6.3.3 Experimental Results

Some experimental results related to the localization performance of the interface in a real scenario are presented. To verify and validate the approach to the localization of pseudo-periodic sounds, a comparison of three types of sources is used: white noise played on a mobile device, a flute played by a musician and a human voice. The interface works with sampling rate of 96 kHz, a Hanning analysis window of 42 ms and a time window for the estimation of the average noise (noise print) of 4.2 s, which is applied when the localization task does not estimate any source. Three microphones with a supercardioid pickup pattern are used; unidirectional sensors are the most frequently used microphones to acquire sound signals in electroacoustic music. It is important to note that the classic microphone for array processing is the omnidirectional polar pattern, but its use is not appropriate in this context because of possible interference with loudspeakers during application in a live performance. However, as we shall see, the use of directional microphones allows the localization of an acoustic source in the small area of interest. The working area is located in a square with 1 meter sides. The axis origin coincides with the position of microphone 2 (m2), the x axis can vary between -50 cm and 50 cm, and the y axis can vary between 0 and 100 cm. The distance between microphones is d = 15 cm.

The experiments were conducted in a rectangular room of 3.5×4.5 m, in a moderately reverberant (RT60 = 0.35 s) and noisy environment. Figure 6.14 shows a comparison of the parameterized PHAT-β TDOA estimation performance. Four tests with different parameters of interface configuration were performed. The TDOA estimation between microphone 2 and 3 was considered. All sound sources are approximately located in the center of study area, (a) (5,52) cm, (b) (4,51) cm, (c) (5,53) cm, (d) (3,51) cm. Table 6.10 summarizes the results, reporting the TDOA mean value and Standard Deviation (SD).

In the first test (a), a continuous white noise signal was played using a mobile device with $\beta = 1$ interface configuration. In this way we verified the complete efficiency offered by the PHAT filter to optimize the TDOA estimation, reducing the degradation effects due to noise and reverberation. We can see in Figure 6.14 how the maximum peak detection is clearly visible (white line). We can also see the effects of multipath reverberation represented by the other parallel gray lines. The value of the TDOA

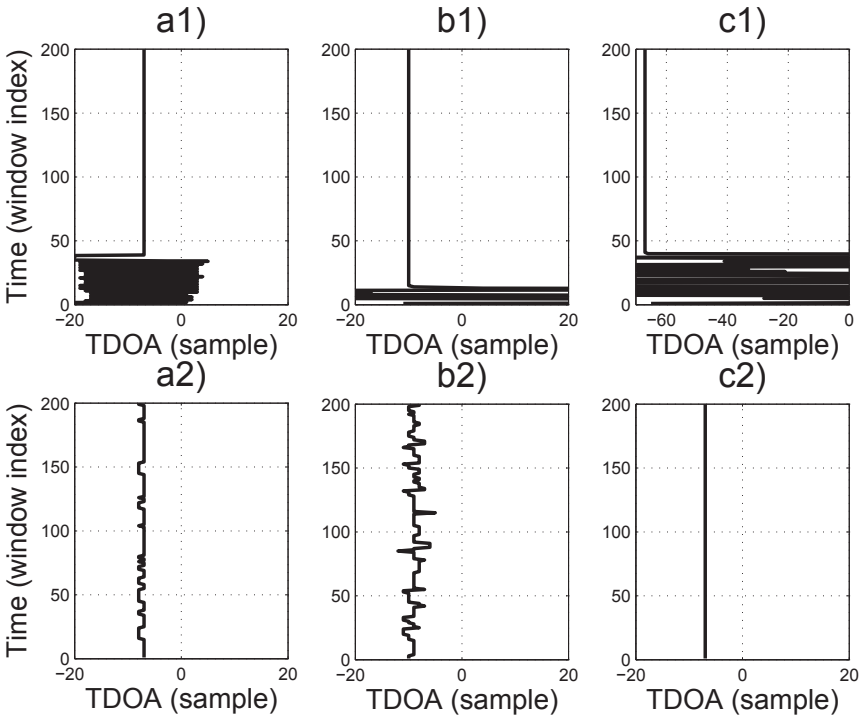

Figure 6.15: *Comparison of TDOA estimation between the AED with sparse priors and the parameterized PHAT-β. a1) White noise - AED; a2) White noise - CGC-PHAT-β; b1) Human voice - AED; b2) Human voice - CGC-PHAT-β; c1) Flute - AED; c2) Flute - CGC-PHAT-β.*

estimation is $\widehat{\tau}_{23} = 7$ (sample). The SD of the TDOA maximum peak during the entire reproduction of sound is 0.2. The TDOA estimation is extremely accurate. In test (b), a flute was considered again with $\beta = 1$ parameter. As expected, the source is not detected ($SD = 8.97$). Subsequently, in test (c) a flute was examined with $\beta = 0.65$ setting. The source is detected as shown in Figure 6.14. The value of the TDOA is $\tau_{23} = 7$ (sample), and the SD results in a value of $SD = 0.8$. In the last test (d), a flute was used with $\beta = 0.65$ and the de-noise Wiener filter task. The value of TDOA is $\tau_{23} = 3$ (sample), the SD results in a value of $SD = 0.7$. Hence, in this case, a lower value of SD indicates less swinging of the TDOA than the average value, which is the correct location of the source.

Therefore, the parameterized PHAT-β allows the TDOA estimation of harmonic sounds, and the de-noise component can improve the accuracy. However, the comparison with test (a), whose robust and well-defined result we aim to obtain, does not yet yield satisfactory results. A parameterization of PHAT with a value of $\beta = 0.65$, according to Donohue *et al.* [2007], is a good compromise between filtering and detection.

Then, a comparison performance of the BSI technique with the CGC-PHAT-β is presented. The

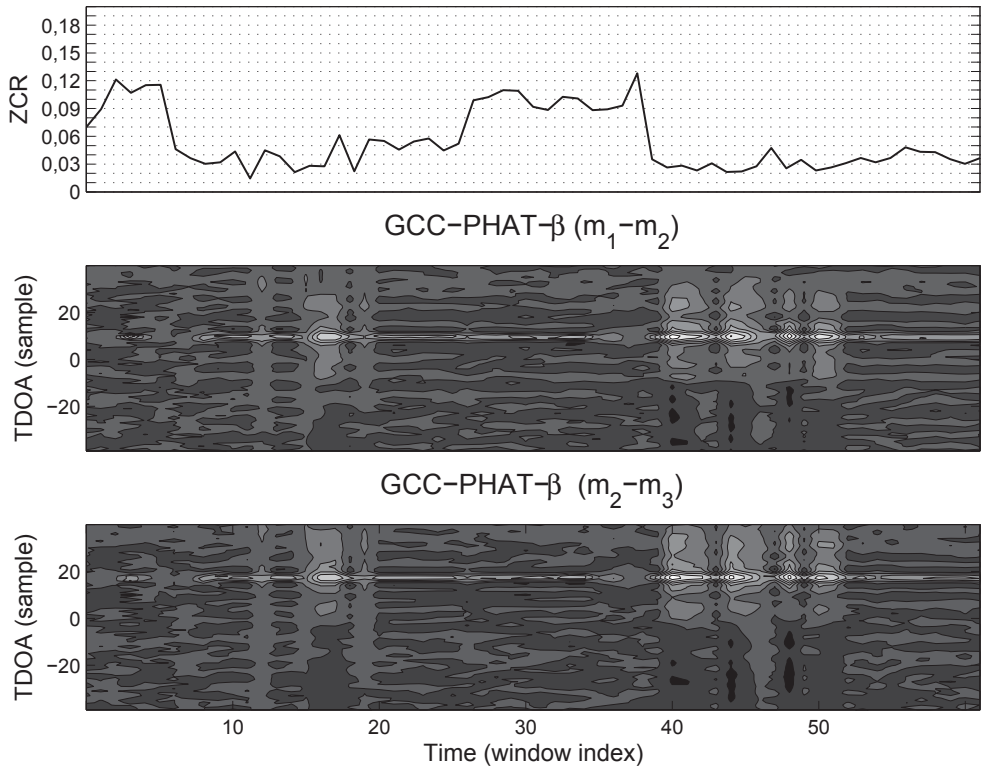

Figure 6.16: *Human voice located at (-20,70) cm. ZCR and parameterized GCC-PHAT-β with threshold value of $\mu = 0.03$; $(m_1 - m_2)$ refers to TDOA estimation between microphones 1 and 2, whereas $(m_2 - m_3)$ is between 2 and 3.*

AED algorithm with sparse priors [Cho & Park, 2009] is used to verify the ability to locate a pseudo-periodic sound. Figure 6.15 shows the results of three tests, comparing the AED with the maximum peak of CGC-PHAT-β. All tests were performed with a mobile sound device placed very close (with a distance of 20 cm from the array) to a pair of microphones. A continuous white noise was used for the test (a1,a2), a female voice was used for the test (b1,b2) and a continuous flute note (G5) was used for the test (c1,c2). We note in Figure 6.15 the performance of the AED algorithm (a1), (b1) and (c1), after the convergence of the NMCFLMS filter, to estimate that the TDOA is correct for white noise (see (a1), $\tau_{12} = -7$ (sample)) and human voice source (see (b1), $\tau_{12} = -10$ (sample))(see (a2) and (b2) for comparison with CGC-PHAT-β, $\beta = 1$). In contrast, for the harmonic sound (case (c1,c2)) the AED converges to an incorrect value of the TDOA (c1), while for the CGC-PHAT-β ((c2), $\tau_{12} = -7$ (sample), $\beta = 0.65$) performs an accurate estimation. Hence, BSI methods (such as the PHAT filter) present difficulties in working with pseudo-periodic sounds.

Before considering the experiments with the Kalman filter, the results of a test with a human voice

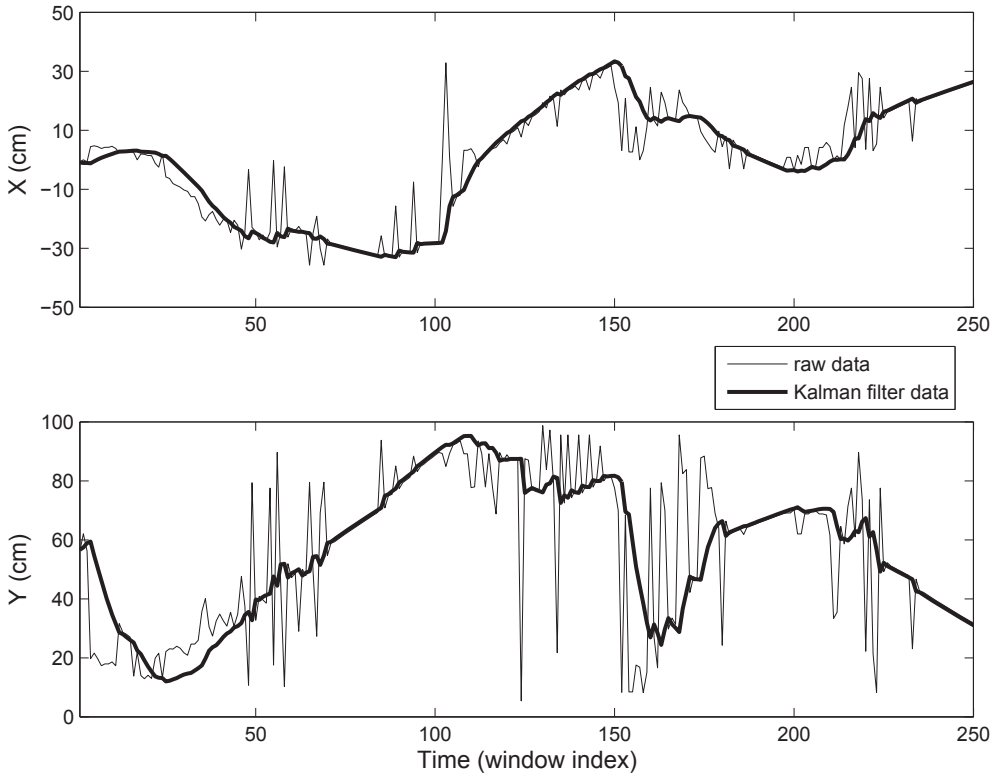

Figure 6.17: *Flute performance moving within the study area; X and Y position of the Kalman filtering data (black lines) and raw data (gray lines).*

are showed, ranging from harmonic to noise sounds to verify the threshold value of a zero-crossing rate for the activation of PHAT-β. The human voice source is located at (-20,70) cm. Figure 6.16 shows the results of the ZCR and adaptive parameterized GCC-PHAT-β with threshold value of $\mu = 0.03$, $\beta = 0.65$ when ZCR $< \mu$, and the de-noise Wiener filter is active. This value μ is enough to achieve an adequate adaptation of the GCC. Still in Figure 6.16, we can note that when the sound becomes harmonic and when we have partially filtered GCC with PHAT, the TDOA peak tends to widen, reducing its robustness, but still allowing the estimation of the source position.

Finally, the last test on the localization performance shows the effectiveness of the Kalman filter in making the xy coordinates more accurate and usable in the interface. Once again, a flute was used moving within the mapped area. The threshold value of ZCR is $\mu = 0.03$, $\beta = 0.65$, and the de-noise task is active. As seen in Figure 6.17, the black lines, which represent the data after Kalman filtering, are reported to have less stability problems due to reverberation. In fact, the estimated raw data (gray lines) present very high swinging values, which would make the interface inappropriate to control the

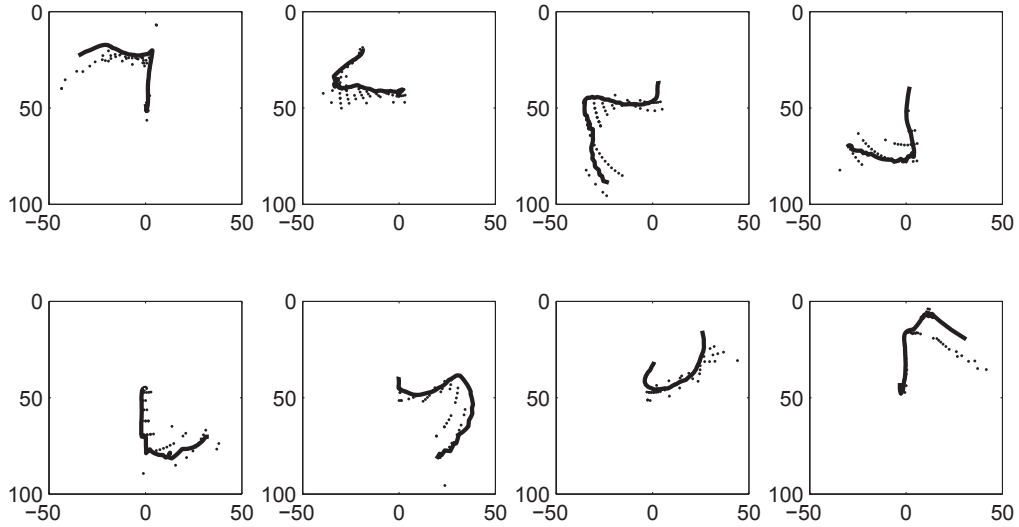

Figure 6.18: *Acoustic source localization performance. The human voice moves in different directions (the dots are the raw data), the x and y axes are in cm.*

processing parameters. Figure 6.18 shows the results of the performance related to the two-dimensional movement of the sound source. The test is composed of eight parts. In each part, the sound source, still a human voice, is moved from the center of the active area along a different direction each time. The working area, which presents good resolution localization, is included in a square with 1 meter sides.

In conclusion, the architecture system was implemented by developing a Max/MSP external object, named asl∼, in order to validate the interface in real-world music application. The object receives incoming audio signals acquired by three microphones and, as output, provides the position of the sound source. The object performs all of the signal processing techniques described in the previous sections. Moreover, a simple Max/MSP patch (Figure 6.19) has been developed to control an audio processor in real-time. As mentioned, the xy values have been used to directly control the parameters of an audio effect. Different VST plug-ins, such as reverb, delay effects and sound spatialisation, are used to demonstrate the usability of the microphone array based interface in musical applications.

6.3.4 Summary

The framework for the localization of pseudo-periodic sounds consists of an adaptive parameterized GCC-PHAT with a zero-crossing rate threshold, a pre-processing with a Wiener filter, and a post-processing with a Kalman filter. Some experimental results have demonstrated the ability of the GCC-PHAT-β algorithm to estimate the TDOA from a microphone pair of harmonic sounds. Moreover, the use of the STSA Wiener filter can be helpful to improve accuracy. However, the use of position data in a

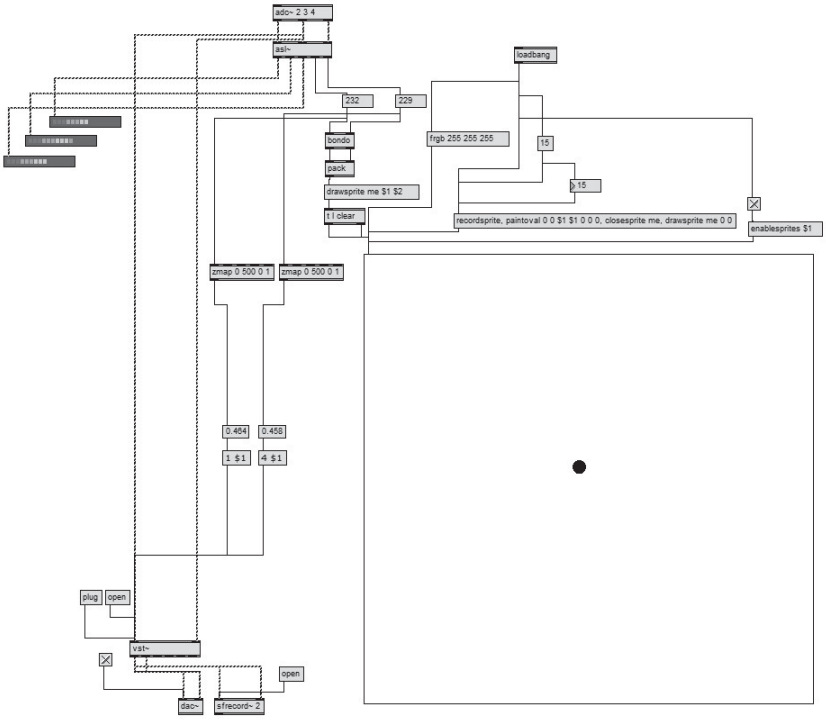

Figure 6.19: *The Max/MSP interface with the external object asl∼.*

musical human-machine interface requires post-processing to provide increased precision. The Kalman filter can be used to track the source and to reduce measurement errors and multi-path channel effects of reverberation. Finally, an interesting result illustrating the limits of the AED algorithm, a method based on BSI, was shown, to estimate the TDOA in the case of pseudo-periodic sounds.

7 Conclusions

7.1 Summary

The work presented in this thesis is a study on the localization of acoustic sources in space with the use of microphone arrays. The first part focused on the state of the art of localization, describing the problem from a geometrical point of view in relation to distributed microphones in space and introducing the techniques to estimate the position based on closed-form estimators, iterative maximum likelihood estimators and spatial likelihood functions. Signal processing techniques used for localization were also presented: the Time Delay Estimation methods and the Steered Response Power beamforming algorithms. Enhancement methodologies for localization were also mentioned, including pre-processing noise reduction, description of the frequency and time domain algorithms, and post-processing that aims to improve the accuracy of the position estimate of the source with a Kalman filter, Particle filter and clustering approach.

The second part covered the scientific contribution of this thesis in the context of two issues that are technical limitations to the state of the art. The first relates to the case of a multi-source sound of short duration events, with the aim of providing a solution in applications such as audio surveillance, sound monitoring and analysis of acoustic scenes. A far-field prototype - consisting of two arrays each with four microphones - and the Incident Signal Power Comparison approach was presented. The second is-

sue describes a prototype solution for the localization of pseudo-periodic sound sources, traditionally related to areas of music, which implements a linear array of three microphones and an architecture based on the adaptive parameterized Generalized Cross-Correlation and Phase Transform (PHAT) weighting with a Zero-Crossing Rate threshold, a Wiener filter to improve the Signal to Noise Ratio, and a Kalman filter to improve the robustness and accuracy of the position estimation. The proposed interface opens possibilities for human-machine interaction and new forms of musical expressive control.

7.2 Considerations

Many studies address the location of human voice, which involves a large number of applications. Extending the nature of the sound of interest, we can see how the innovations proposed in this thesis can provide solutions to new problems that arise.

The experimental results with the ISPC have shown that this approach can be a solution for multi-source localization that requires a frame-to-frame analysis. This approach is particularly advantageous in identifying sounds of short duration that can be difficult to accomplish using a traditional Bayesian filter. The prototype used for the experiments exhibited the best performance using ISPC with the RMS log spectral distance function and the high-resolution beamforming technique of MVDR-DL, although both SRP and SRP-DC still have a minor localization success rate. The limits of this approach were presented in the case of two sources with a similar spectral content. The experiment involved two motor cars. We must emphasize that ISPC systems can integrate with Bayesian filtering, and can be helpful in cases that require detailed analysis over time as well as, in cases in which the Bayesian filter can fail: 1) during the initialization phase of the filter, 2) when the sources have an unpredictable trajectory (e.g., in the case of rapid changes of the velocity vector), and 3) when two sources have intersecting trajectories. Therefore, the success rate of 90.9 for the MVDR-DL array with a small array size is an important result, which promises an improved performance with arrays of larger size.

The near-field prototype for the localization of pseudo-periodic sound is commonly used in a controlled environment with moderate reverb and noise, and with different sound sources in the case of single source. The experiments have demonstrated the ability of the proposed architecture to locate harmonic sounds in a reverberant environment with a RT60 of 0.35 s. The capability of interface has been verified for use in real-time audio control in a stable way by testing with small movements of the sound source (of the order of tens of centimeters). The proposed interface has the advantage being completely non-invasive (no need for markers, sensors or wires on the performance) and requires no dedicated hardware. However, its real application during a performance still requires new investigations.

7.3 Future Work

As we have seen, ISPC is based on minimizing the difference error of the spectral power output of the signals using a spatial filter. The main limitation, which has been observed in experiments (the percentage of correct associations of DOAs is greatly reduced), may be addressed by the integration of additional sound features with the spectrum difference comparison. This approach would allow robust results to be obtained even with similar sound sources.

The use of microphone-array-based interfaces for real-time musical control application requires two types of further investigations: the capability to locate and track the source in the condition of competing sounds (other musicians and performers nearby, the presence of the return due to a sound amplification system), and a validation in higher reverberant environments.

Finally, this research has focused on the use of small-sized arrays. However, we were able to obtain interesting localization results, both in a near-field and far-field environment. Nevertheless, it would be interesting to evaluate the performance of large array networks, especially to evaluate the ISPC approach, which is expected to provide a better performance in the separation of sources with beam-forming.

References

AARABI, P. (2003). The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Applied Signal Processing*, vol. 2003, n. 4, pp. 338–347. 4, 26

AJMERA, J., LATHOUD, G. & MCCOWAN, L. (2004). Clustering and segmenting speakers and their locations in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 605–608. 60

ANTONACCI, F., RIVA, D., SAIU, D., SARTI, A., TAGLIASACCHI, M. & TUBARO, S. (2006). Tracking multiple acoustic sources using particle filtering. In *Proceedings of the European Signal Processing Conference*, pp. 1–4. 5, 59

ARULAMPALAM, M.S., MASKELL, S., GORDON, N. & CLAPP, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, vol. 50, n. 2, pp. 174–188. 58

AZIMI-SADJADI, M.R., PEZESHKI, A. & ROSEVEARE, N. (2008). Wideband DOA estimation algorithms for multiple moving sources using unattended acoustic sensors. *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, n. 4, pp. 1585–1599. 44

BANGS, W.J. & SCHULTHEISS, P.M. (1973). Space-time processing for optimal parameter estimation. In *Signal Processing*, pp. 577–590, Academic Press. 16

BARABELL, A. (1983). Improving the resolution performance of eigenstructure-based direction-finding algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 336–339. 44

BARI, A., CANAZZA, S., DE POLI, G. & MIAN, G.A. (2001). Toward a methodology for the restoration of electroacoustic music. *Journal of New Music Research*, vol. 30, n. 4, pp. 351–363. 50

REFERENCES

BARTLETT, M.S. (1948). Smoothing periodograms from time-series with continuous spectra. *Nature*, vol. 161, pp. 686–687. 3, 38

BECHLER, D., GRIMM, M. & KROSCHEL, K. (2003). Speaker tracking with a microphone array using kalman filtering. *Advances in Radio Science*, vol. 1, pp. 113–117. 4, 57

BEERENDS, J.G. & STEMERDINK, J.A. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, vol. 40, n. 12, pp. 963–978. 50

BEERENDS, J.G. & STEMERDINK, J.A. (1994). A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, vol. 42, n. 3, pp. 115–123. 50

BENESTY, J. (2000). Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America*, vol. 107, n. 1, pp. 384–391. 3, 33

BENESTY, J., CHEN, C.J. & HUANG, Y. (2004). Time-delay estimation via linear interpolation and cross correlation. *IEEE Transactions on Speech and Audio Processing*, vol. 12, n. 5, pp. 509–519. 3, 36

BENESTY, J., HUANG, Y. & CHEN, J. (2007). Time delay estimation via minimum entropy. *IEEE Signal Processing Letters*, vol. 14, n. 3, pp. 157–160. 3

BOLL, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, n. 2, pp. 113–120. 48

BRANDSTEIN, M., ADCOCK, J.E. & SILVERMAN, H.F. (1997). A closed-form location estimator for use with room environment microphone arrays. *IEEE Transactions on Speech and Audio Processing*, vol. 5, n. 1, pp. 45–50. 4, 22

BRUTTI, A., OMOLOGO, M. & SVAIZER, P. (2010). Multiple source localization based on acousticmap de-emphasis. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–17. 5, 27

BUCHNER, H., AICHNER, R. & KELLERMANN, W. (2007). *Blind Speech Separation*, chap. TRINICON-based Blind System Identification with Application to Multiple-Source Localization and Separation. Springer. 35

CANAZZA, S. (2007). *Noise and Representation Systems: A Comparison Among Audio Restoration Algorithms*. Lulu Press. 50

CANAZZA, S., CORADDU, G., DE POLI, G. & MIAN, G.A. (2001). Objective and subjective comparison of audio restoration methods. *Journal of New Music Research*, vol. 30, n. 1, pp. 93–102. 50

CANAZZA, S., DE POLI, G. & MIAN, G.A. (2010). Restoration of audio documents by means of extended kalman filter. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, n. 6, pp. 1107–1115. 51, 53

CAPON, J. (1969). High resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, vol. 57, n. 8, pp. 1408–1418. 3, 41

CAPPE, O. (1994). Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, vol. 2, n. 2, pp. 345–349. 49

CARLSON, B. (1988). Covariance matrix estimation errors and diagonal loading in adaptive arrays. *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, n. 4, pp. 397–401. 42

CARPENTER, J., P.CLIFFORD & FEARNHEAD, P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings Radar, Sonar and Navigation*, vol. 146, n. 1, pp. 2–7. 58

CHAMPAGNE, B., BERDARD, S. & STEPHENNE, A. (1996). Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, vol. 4, n. 2, pp. 148–152. 33

CHAN, Y.T. & HO, K.C. (1994). A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, vol. 43, n. 8, pp. 1905–1915. 4, 20

CHEN, J., BENESTY, J. & HUANG, Y. (2003). Robust time delay estimation exploiting redundancy among multiple microphones. *IEEE Transactions on Speech and Audio Processing*, vol. 11, n. 6, pp. 549–557. 3, 36

CHEN, J.C., HUDSON, R.E. & YAO, K. (2002). Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field. *IEEE Transactions on Signal Processing*, vol. 50, n. 8, pp. 1843–1854. 4, 26

CHEN, W.P., HOU, J.C. & SHA, L. (2004). Dynamic clustering for acoustic target tracking in wireless sensor networks. *IEEE Transactions on Mobile Computing*, vol. 3, n. 3, pp. 258–271. 5, 28

CHIU, S. (1994). Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems*, vol. 2, n. 3, pp. 267–278. 60

REFERENCES

CHO, J. & PARK, H. (2009). Imposition of sparse priors in adaptive time delay estimation for speaker localization in reverberant environments. *IEEE Signal Processing Letters*, vol. 16, n. 3, pp. 180–183. 35, 89

CHO, Y., YOOK, D. & KIM, S.C.H. (2009). Sound source localization for robot auditory systems. *IEEE Transactions on Consumer Electronics*, vol. 55, n. 3, pp. 1663–1668. 36

CLAUDIO, E.D. & PARISI, R. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chap. Multi-source localization strategies. Springer. 60

COBOS, M., MARTI, A. & LOPEZ, J.J. (2011). A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling. *IEEE Signal Processing Letters*, vol. 18, n. 1, pp. 71–74. 36

COX, H., ZESKIND, R. & OWEN, M. (1987). Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, n. 10, pp. 1365–1376. 42

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Serie B*, vol. 39, n. 1, pp. 1–38. 60

DESTINO, G. & ABREU, G. (2011). On the maximum likelihood approach for source and network localization. *IEEE Transactions on Signal Processing*, vol. 59, n. 10, pp. 4954–4970. 4, 26

DIBIASE, J.H., SILVERMAN, H.F. & BRANDSTEIN, M.S. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chap. Robust localization in reverberant rooms. Springer. 3, 4, 27, 35, 36, 42

DMOCHOWSKI, J., BENESTY, J. & AFFS, S. (2009). On spatial aliasing in microphone arrays. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 57, n. 4, pp. 1383–1395. 14

DMOCHOWSKI, J.P., BENESTY, J. & AFFES, S. (2007). A generalized steered response power method for computationally viable source localization. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, n. 6, pp. 2510–2526. 36

DOCLO, S. & MOONEN, M. (2003). Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal on Applied Signal Processing*, vol. 2003, n. 11, pp. 1110–1124. 35

DOLPH, C. (1946). A current distribution for broadside arrays which optimizes the relationship between beamwidth and sidelobe level. *Proceedings of the IRE*, vol. 34, n. 6, pp. 335–348. 40

DONOHUE, K.D., HANNEMANN, J. & DIETZ, H.G. (2007). Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments. *Signal Processing*, vol. 87, n. 7, pp. 1677–1691. 33, 84, 88

DUNN, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, vol. 3, n. 3, pp. 32–57. 60

EPHRAIM, Y. & MALAH, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, n. 6, pp. 1109–1121. 49

EPHRAIM, Y. & MALAH, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, n. 2, pp. 443–445. 49

FONG, W., GODSILL, S.J., DOUCET, A. & WEST, M. (2002). Monte carlo smoothing with application to audio signal enhancement. *IEEE Transactions on Signal Processing*, vol. 50, n. 2, pp. 438–449. 51

GANNOT, S. & DVORKIND, T.G. (2006). Microphone array speaker localizers using spatial-temporal information. *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–17. 4, 57

GEORGIOU, P.G. & KYRIAKAKIS, C. (2006). Robust maximum likelihood source localization: The case for sub-gaussian versus gaussian. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n. 4, pp. 1470–1480. 4, 26

GILLETTE, M.D. & SILVERMAN, H.F. (2008). A linear closed-form algorithm for source localization from time-differences of arrival. *IEEE Signal Processing Letters*, vol. 5, pp. 1–4. 4, 24

GODSILL, S.J. & RAYNER, P.J.W. (1995). A bayesian approach to the restoration of degraded audio signals. *IEEE Transactions on Speech and Audio Processing*, vol. 3, n. 4, pp. 267–278. 51

GORDON, N.J., SALMOND, D.J. & SMITH, A.F.M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings-F Radar and Signal Processing*, vol. 140, n. 2, pp. 107–113. 58

GRAY, J.A.H. & MARKEL, J.D. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, n. 5, pp. 380–391. 68

REFERENCES

GUSTAFSSON, F., GUNNARSSON, F., BERGMAN, N., FORSSELL, U., JANSSON, J., KARLSSON, R. & NORDLUND, P.J. (2002). Particle filters for positioning, navigation, and tracking. *IEEE Transactions on Signal Processing*, vol. 50, n. 2, pp. 425–437. 58

HAHN, W. & TRETTER, S. (1973). Optimum processing for delay-vector estimation in passive signal arrays. *IEEE Transactions on Information Theory*, vol. 19, n. 5, pp. 608–614. 4, 26

HE, Y. & CHONG, K.P. (2004). Sensor scheduling for target tracking in sensor networks. In *Proceedings of the IEEE Conference on Decision and Control*, vol. 1, pp. 743–748. 5, 28

HU, J.S. & YANG, C.H. (2010). Estimation of sound source number and directions under a multisource reverberant environment. *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–14. 5

HUANG, Q., ZHONG, Q. & ZHUANG, Q. (2011). Source localization with minimum variance distortionless response for spherical microphone arrays. *Journal of Shanghai University*, vol. 15, n. 1, pp. 21–25. 42

HUANG, Y. & BENESTY, J. (2003). *Adaptive Signal Processing: Applications to Real-World Problems*, chap. Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization. Springer. 3, 37

HUANG, Y., BENEST, J., ELKO, G.W. & MERSEREAU, R.M. (2001). Real-time passive source localization: a practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, vol. 9, n. 8, pp. 943–956. 4, 23, 24

HUE, C., CADRE, J.P.L. & PEREZ, P. (2002). Tracking multiple objects with particle filtering. *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, n. 3, pp. 791–812. 58

IANNIELLO, J.P. (1982). Time delay estimation via cross-correlation in the presence of large estimation errors. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, n. 6, pp. 998–1003. 33

INGARD, U. (1953). A review of the influence of meteorological conditions on sound propagation. *Journal of the Acoustical Society of America*, vol. 25, n. 3, pp. 405–4011. 2

KALMAN, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, vol. 82, pp. 35–45. 51, 55

KHAN, M.F. & TUFAIL, M. (2009). Performance analysis of ESPRIT algorithm on a single broadband signal. In *Proceedings of the International Conference on Emerging Technologies*, pp. 310–314. 45

KHNEA, M., TOGNERI, R. & NORDHOLM, S. (2009). Robust source localization in reverberant environments based on weighted fuzzy clustering. *IEEE Signal Processing Letters*, vol. 16, n. 2, pp. 85–88. 60

KLEE, U., GEHRIG, T. & MCDONOUGH, J. (2006). Kalman filters for time delay of arrival-based source localization. *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15. 4, 57

KNAPP, C. & CARTER, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, n. 4, pp. 320–327. 3, 31, 32

KOZICK, R.J. & SADLER, B.M. (2003). Source localization with distributed sensor arrays and partial spatial coherence. *IEEE Transactions on Signal Processing*, vol. 52, n. 3, pp. 601–616. 5, 28

LAROCQUE, J.R., REILLY, J.P. & NG, W. (2002). Particle filters for tracking an unknown number of sources. *IEEE Transactions on Signal Processing*, vol. 50, n. 2, pp. 2926–2937. 58

LATHOUD, G. & ODOBEZ, J.M. (2007). Short-term spatiotemporal clustering applied to multiple moving speakers. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, n. 5, pp. 1696–1710. 60

LEE, B. & CHOI, J. (2010). Multi-source sound localization using the competitive k-means clustering. In *Proceedings of the IEEE Conference on Emerging Technologies and Factory Automation*, pp. 1–7. 60

LEHMANN, E.A. & JOHANSSON, A.M. (2007). Particle filter with integrated voice activity detection for acoustic source tracking. *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–11. 59

LEHMANN, E.A. & WILLIAMSON, R.C. (2006). Particle filter design using importance sampling for acoustic source localisation and tracking in reverberant environments. *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–9. 59

LEVY, A., GANNOT, S. & HABETS, E.A.P. (2011). Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n. 6, pp. 1540–1555. 5, 59

REFERENCES

LIANG, Z., MA, X. & DAI, X. (2008). Robust tracking of moving sound source using multiple model kalman filter. *Applied Acoustics*, vol. 69, n. 12, pp. 1350–1355. 4, 57

LIU, J., REICH, J. & ZHAO, F. (2003). Collaborative in-network processing for target tracking. *EURASIP Journal on Applied Signal Processing*, vol. 2003, n. 4, pp. 378–391. 5, 27

LOMBARD, A., ZHENG, Y., BUCHNER, H. & KELLERMANN, W. (2011). TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, n. 6, pp. 1490–1503. 3, 5

MACQUEEN, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297. 60, 74

MAKHOUL, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, vol. 63, n. 4, pp. 561–580. 67

MCAULAY, R.J. (1984). Maximum likelihood spectral estimation and its application to narrow-band speech coding. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, n. 2, pp. 243–251. 67

MICHAUDY, J.M.V.F. & ROUAT, J. (2007). Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, vol. 55, n. 3, pp. 216–228. 5, 59

NIEDZWIECKI, M. & CISOWSKI, K. (1996). Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals. *IEEE Transactions on Signal Processing*, vol. 44, n. 3, pp. 528–537. 51

NING, M., BOUCHARD, M. & GOUBRAN, R.A. (2006). Speech enhancement using a masking threshold constrained kalman filter and its heuristic implementations. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n. 1, pp. 19–32. 51

NISHIURA, T., YAMADA, T., NAKAMURA, S. & SHIKANO, K. (2000). Localization of multiple sound sources based on a CSP analysis with a microphone array. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1053–1056. 5

OMOLOGO, M. & DEMORI, P.S.R. (1998). *Spoken Dialogue with Computers*, chap. Acoustic Transduction. Academic Press. 4, 26

OMOLOGO, M. & SVAIZER, P. (1997). Use of the crosspower-spectrum phase in acoustic event location. *IEEE Transactions on Speech and Audio Processing*, vol. 5, n. 3, pp. 288–292. 33

ORTON, M. & FITZGERALD, W. (2002). A bayesian approach to tracking multiple targets using sensor arrays and particle filters. *IEEE Transactions on Signal Processing*, vol. 50, n. 2, pp. 216–223. 58

OTTERSTEN, B., VIBERGA, M. & KAILATH, T. (1991). Performance analysis of the total least squares ESPRIT algorithm. *IEEE Transactions on Signal Processing*, vol. 39, n. 5, pp. 1122–1135. 45

PAULRAJ, A., ROY, R. & KAILATH, T. (1986). A subspace rotation approach to signal parameter estimation. *Proceedings of the IEEE*, vol. 74, n. 7, pp. 1044–1046. 3, 44

PERTILÄ, P., KORHONEN, T. & VISA, A. (2008). Measurement combination for acoustic source localization in a room environment. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–14. 4, 27

PFEIFER, L.L. (1974). Inverse filter for speaker identification. Tech. rep., RADCTR-74-214, Speech Communications Research Lab Inc Santa Barbara California. 67

POTAMITIS, I., CHEN, H. & TREMOULIS, G. (2004). Tracking of multiple moving speakers with multiple microphone arrays. *IEEE Transactions on Speech and Audio Processing*, vol. 12, n. 5, pp. 520–529. 4, 57

PRANDI, G., VALENZISE, G., TAGLIASACCHI, M., ANTONACCI, F., SARTI, A. & TUBARO, S. (2008). Acoustic source localization by fusing distributed microphone arrays measurements. In *Proceedings of the EURASIP European Signal Processing Conference*. 5, 28

QUINLAN, A., KAWAMOTO, M., MATSUSAKA, Y., ASOH, H. & ASANO, F. (2009). Tracking intermittently speaking multiple speakers using a particle filter. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–11. 5, 59

RAO, B.D. & HARI, K.V.S. (1989). Performance analysis of root-music. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, n. 12, pp. 1939–1949. 44

ROY, R. & KAILATH, T. (1989). ESPRIT- estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, n. 7, pp. 984–995. 3, 44

REFERENCES

ROY, R., PAULRAJ, A. & KAILATH, T. (1986). ESPRIT - a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, n. 5, pp. 1340–1342. 3, 44

SALVATI, D. & CANAZZA, S. (2009). Improvement of acoustic localization using a short time spectral attenuation with a novel suppression rule. In *Proceedings of the International Conference on Digital Audio Effects*, pp. 150–156. 47

SALVATI, D., RODÀ, A., CANAZZA, S. & FORESTI, G.L. (2010). A real-time system for multiple acoustic sources localization based on ISP comparison. In *Proceedings of the 13th International Conference on Digital Audio Effects*, pp. 201–208. 64

SALVATI, D., CANAZZA, S. & RODÀ, A. (2011a). A sound localization based interface for real-time control of audio processing. In *Proceedings of the 14th International Conference on Digital Audio Effects*, pp. 177–184. 82

SALVATI, D., CANAZZA, S. & RODÀ, A. (2011b). Sound spatialization control by means of acoustic source localization system. In *Proceedings of the 8th Sound and Music Computing Conference*, pp. 284–289. 82

SALVATI, D., RODÀ, A., CANAZZA, S. & FORESTI, G.L. (2011c). Multiple acoustic sources localization using incident signal power comparison. In *Proceedings of the 8th IEEE International Conference on Advanced Video and Signal based Surveillance*, pp. 77–82. 64

SCHAU, H.C. & ROBINSON, A.Z. (1987). Passive source localization employing intersecting spherical surfaces from time-of-arrival differences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, n. 8, pp. 1223–1225. 4, 17

SCHEUING, J. & YANG, B. (2008). Disambiguation of TDOA estimation for multiple sources in reverberant environments. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, n. 8, pp. 1479–1489. 5

SCHMIDT, R.O. (1972). A new approach to geometry of range difference location. *IEEE Transactions on Aerospace and Electronic Systems*, vol. 8, n. 6, pp. 821–835. 4, 16

SCHMIDT, R.O. (1979). Multiple emitter location and signal parameter estimation. In *Proceedings of the RADC Spectrum Estimation Workshop*, pp. 243–258. 3, 42

SCHMIDT, R.O. (1986). Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, vol. 34, n. 3, pp. 276–280. 3, 42

SCHMIDT, S. (1966). Applications of state-space methods to navigation problems. *Advances in Control Systems*, vol. 3, pp. 293–340. 51, 57

SEGAL, M., WEINSTEIN, E. & MUSICUS, B.R. (1991). Estimate-maximize algorithms for multichannel time delay and signal estimation. *IEEE Transactions on Signal Processing*, vol. 39, n. 1, pp. 1–16. 4, 26

SEGURAA, C., ABAD, A., HERNANDO, J. & NADEU, C. (2008). Multispeaker localization and tracking in intelligent environments. *Lecture Notes in Computer Science*, vol. 4625, pp. 82–90. 4, 57

SILVERMAN, H.F., YU, Y., SACHAR, J.M. & PATTERSON, W.R.I. (2005). Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, vol. 13, n. 4, pp. 593–606. 36

SMITH, J.O. & ABEL, J.S. (1987). Closed-form least-squares source location estimation from range-difference measurements. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, n. 12, pp. 1661–1669. 4, 19

STOICA, P. & LI, J. (2006). Source localization from range-difference measurements. *IEEE Signal Processing Magazine*, vol. 23, n. 3, pp. 63–66. 4, 19, 24

STOICA, P. & NEHORAI, A. (1990). MUSIC, maximum likelihood, and cramer-rao bound: further results and comparisons. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, n. 12, pp. 2140–2150. 4, 26

STOICA, P., NEHORAI, A. & SDERSTRM, T. (1995). Decentralized array processing using the MODE algorithm. *Circuits, Systems, and Signal Processing*, vol. 14, n. 1, pp. 17–38. 5, 27

STROBEL, N., SPORS, S. & RABENSTEIN, R. (2001a). Joint audio-video object localization and tracking. *IEEE Signal Processing Magazine*, vol. 18, n. 1, pp. 22–31. 4, 57

STROBEL, N., SPORS, S. & RABENSTEIN, R. (2001b). *Microphone Arrays: Signal Processing Techniques and Applications*, chap. Joint audio-video signal processing for object localization and tracking. Springer. 4, 57

REFERENCES

SUN, H., TEUTSCH, H., MABANDE, E. & KELLERMANN, W. (2011). Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 117–120. 45

TALANTZIS, F., PNEVMATIKAKIS, A. & CONSTANTINIDES, A.G. (2008). Audiovisual active speaker tracking in cluttered indoors environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, n. 3, pp. 799 –807. 5, 59

TASHEV, I.J. (2009). *Sound Capture and Processing: Practical Approaches*. Wiley. 2

TEUTSCH, H. & KELLERMANN, W. (2005). EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. iii/89– iii/92 Vol. 3. 45

TSOUKALAS, D.E., MOURJOPOULOS, J.N. & KOKKINAKIS, G. (1997a). Perceptual filters for audio signal enhancement. *Journal of the Audio Engineering Society*, vol. 45, n. 1/2, pp. 22–36. 50

TSOUKALAS, D.E., MOURJOPOULOS, J.N. & KOKKINAKIS, G. (1997b). Speech enhancement based on audible noise suppression. *IEEE Transactions on Speech and Audio Processing*, vol. 5, n. 6, pp. 497–514. 50

WANG, H. & KAVEH, M. (1985). Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, n. 4, pp. 823–831. 44

WARD, D.B., LEHMANN, E.A. & WILLIAMSON, R.C. (2003). Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on Speech and Audio Processing*, vol. 11, n. 6, pp. 826–836. 4, 5, 27, 59

WAX, M. & KAILATH, T. (1983). Optimum localization of multiple sources by passive arrays. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, n. 5, pp. 1210–1217. 4, 26

WAX, M. & KAILATH, T. (1984). Spatio-temporal spectral analysis by eigenstructure methods. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, n. 4, pp. 817–827. 44

WEN, F. & WAN, Q. (2011). Robust time delay estimation for speech signals using information theory: A comparison study. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, pp. 1–10. 3

WIENER, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications*. Cambridge, MIT Press, Massachusetts. 48, 49

YOON, Y., KAPLAN, L.M. & MCCLELLAN, J.H. (2006). TOPS: New DOA estimator for wideband signals. *IEEE Transactions on Signal Processing*, vol. 54, n. 6, pp. 791–802. 44

ZOTKIN, D.N. & DURAISWAMI, R. (2004). Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Speech and Audio Processing*, vol. 12, n. 5, pp. 499–508. 36

ZOTKIN, D.N., DURAISWAMI, R. & DAVIS, L.S. (2002). Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, vol. 11, n. 1, pp. 1154–1164. 4, 59

List of Figures

LIST OF FIGURES

List of Tables

List of Abbreviations

ABMCI	Adaptive Blind MultiChannel Identification
AED	Adaptive Eigenvalue Decomposition
ASL	Acoustic Source Localization
BSI	Blind System Identification
CC	Cross-Correlation
CLS	Constrained Least Squares
CMSR	Canazza-Mian Suppression Rule
CRLB	Cramr-Rao Lower Bound
DC	Dolph-Chebyshev
DFT	Discrete Fourier Transform
DL	Diagonal Loading
DOA	Direction Of Arrival
DS	Delay & Sum
EKF	Extended Kalman Filter
EMSR	Ephraim-Malah Suppression Rule
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Techniques
FFT	Fast Fourier Transform
FSRP	Filter Steered Response Power
GCC	Generalizated Cross-Correlation
GCF	Global Coherence Field
GS	Gillette-Silverman
HI	Hyperbolic Intersection
HT	Hannan & Thomson
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform

LIST OF ABBREVIATIONS

IS	Itakura-Saito
ISP	Incident Signal Power
ISPC	Incident Signal Power Comparison
KF	Kalman Filter
LC	Linear Correction
LI	Linear Intersection
LMS	Least Mean Square
LP	Linear Prediction
LS	Least Squares
MCCC	Multichannel Cross-Correlation Coefficient
ML	Maximum Likelihood
MUSIC	MUltiple SIgnal Classification
MVDR	Minimum Variance Distortionless Response
NMCFLMS	Normalized MultiChannel Frequency domain Least Mean Square
PCC	Pearson Correlation Coefficient
PDF	Probability Density Function
PF	Particle Filter
PHAT	PHAse Transform
PI	Plane Intersection
PSD	Power Spectral Density
RIR	Roth Impulse Response
RMS	Root Mean Square
SCOT	Smoothed COherence Transform
SDF	Spectral Distance Functions
SI	Spherical Interpolation
SIR	Sequential Importance Resampling
SNR	Signal to Noise Ratio
SRP	Steered Response Power
STFT	Short-Time Fourier Transform
STSA	Short-Time Spectral Attenuation
SX	Spherical Intersection

TDE	Time Delay Estimation
TDOA	Time Difference Of Arrival
ULS	Unconstrained Least Squares
ZCR	Zero-Crossing Rate