Sound Source Localization for Enhancement of Orchestral Music for Multi-Sensor Recordings

Xavier Lizarraga

SOUND AND MUSIC COMPUTING
MASTER THESIS-UPF / 2015

SUPERVISED BY
Julio Carabias & Jordi Janer

Music Technology Group

Universitat Pompeu Fabra Barcelona

To my family...

Acknowledgements

I would like to give thanks to Xavier Serra and Emilia Gómez for giving me the opportunity to expand my learning in this area of study. Also, an special mention to my supervisors, Julio J. Carabias-Orti and Jordi Janer for their help, their compromise and their patience. Besides, I would like to give thanks to my family for supporting me, all the classmates and all the MTG guys who has participated along the course. Thanks!

Abstract

This research is focus on the combination of *Sound Source Localization* (SSL) methods with *Sound Source Separation* (SSS) techniques inside an Orchestral music's framework. The position at each source is estimated by a 3D grid search algorithm, by means *Generalized Cross-Correlation-PHAse Transform* (GCC-PHAT) feature vectors and the distances between microphones. This approach could be useful for audio applications in real time as automatic camera steering, videogaming, autonomous robots or auralization techniques.

The main scope of this work is to learn in depth into signal processing algorithms for SSL, separation and audio enhancement with microphone arrays in a multi-track music recording. So, supplying a robust and low-complexity method for music applications as the up-mixing of the acoustic scene in other formats (mono-to-stereo, 5.1, Dolby Digital or other) or an audio rendering motor able to surfing through the acoustic scene.

Some studies are done to combine SSL with SSS, nevertheless, most of them propose experiments with speech signals [3, 4, 5]. Therefore, in this dissertation we have performed some experiments to assessing, by means of objective metrics, the combination of these processes with Orchestral music.

Resum

Aquesta investigació es centra en la combinació de mètodes per la localització i separació de fonts sonores dins d'un marc de música d'orquestra. La posició de cada font s'estima amb un algoritme de graella 3D, mitjançant vectors característics de GCC-PHAT y les distàncies entre micròfons. Aquest enfocament podria ser util per aplicacions en temps real com la conducció de càmera automàtica, videojocs, robots autònoms o tècniques d'auralització.

El principal abast d'aquest treball es aprendre en profunditat sobre algoritmes de processament del senyal per la localització , separació i realçament de fonts sonores amb un conjunt de micròfons en una gravació multicanal. Així, oferint un mètode robust de baixa complexitat per aplicacions musicals com la sobremescla de l'escena acústica en diferents formats (de mono a stereo, 5.1, Dolby Digital o altres) o un motor de representacio d'àudio capaç de navegar per l'escena acustica.

Alguns estudis s'han fet per combinar la localitzacio de fonts amb la separació, però la majoria d'ells proposen experiments amb senyals de veu [3, 4, 5]. Per això, en aquest document hem realitzat alguns experiments per evaluar, mitjançant mesures objectives, la combinació d'aquests procesos amb música d'orquestra.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

During the last decades, many researches have been designed robust *Automatic Speech Recognition* (ASR) systems with microphone arrays to indoor adverse acoustic conditions, particularly to reverberant and noisy environments, by taking advantage of source positions for the enhancement of SSS methods, in order to recognise the active speaker [1][2]. Applying a realistic approach, live music performances are played in halls with diverse acoustic conditions (studio room, indoor classical music venues and concert venues), where the *Reverberation Time* (RT_{60}) is always present and it generates complex signals, due to the superposition signals, which makes extremely difficult the localisation tasks. This research is focus on estimating the source sound position on the 3D space in order to use the location information for the improving of SSS techniques and the enhancement of Orchestral music correcting the spatial location of sound sources by means phase alignment in music scenarios. Therefore, this work aims to propose an SSL approach adapted specially for Orchestral music. However, the main scope of this work is to learn in depth into signal processing algorithms for sound SSL, SSS and audio enhancement with microphone arrays for multi-track music recording, supplying a robust and low-complexity method for music applications as down-mixing techniques to convert multi-channel signals in other formats (Stereo 2.0, 5.1, Dolby Digital or other). Or creating new mixture of an acoustic scene, changing the position of listener in order to surf inside the acoustic space. The resulting system could be useful for audio applications as automatic camera steering, automatic lighting systems, video-gaming, audio enhancement and dereverberation, acoustic scene analysis, auralisation processes, interactive systems and autonomous robot.

1.1 Motivation

On one hand, as Do & Silverman, 2010 show in [50, 51], there are particular difficulties with SSL methods when simultaneous sources are active. Today, it is available a lot of *Musical Instrument Digital Interface* (MIDI) files, which can be employed as pitch information to determine the non-overlapped frequency components of the musical sources along the time. This data can be applied to select the target to localise.

***Hypothesis 1*:** The score lets to define the regions of *Time-Frequency* (TF) representation at each source involved in an isolated fashion, avoiding overlapped frequency components. This data can be essential to define unique TF regions at each source to enhance the SSL methods when a target is defined.

On the other hand, it is demonstrated that SSS algorithms, with prior knowledge and prior defined constraints, exceed the results reached by *Blind Source Separation* (BSS) models. Some researches are pointing to use panning information and score alignments to improve the separation process [5, 41, 52, 53].

***Hypothesis 2*:** The using of sound source position in the space at each sound source, as prior knowledge can be employed to estimate mixing parameters and it will improve the outcomes for a Position-Informed Source Separation approach with regard to BSS approaches.

1.2 Research Goals

The whole framework for the project proposal offers different research lines which are in keeping with the hypotheses.

1. Employing MIDI score to define the regions of TF representation at each instrument in the audio mixture, allowing the location estimate even though sources are active simultaneously.

2. Assessment for the application of source localisation to the separation of Orchestral signals.

1.3 Methodology

In order to reach the goals proposed, it is necessary to follow a research methodology. The appropriated methodology carries:

1. Analyzing the state-of-the-art for SSL methods.

2. Assessment and deployment of Steered Response Power-PHAse Transform (SRP-PHAT). This algorithm is used for speaker localization in cocktail party situations. In our particular case, we will study the application of this method with Orchestral signals.

3. Defining a work plan for the designed experiments.

1.4 Structure of the document

This document is composed by five chapters, the chapter 2 introduces a literature review and the state of the art are introduced, reviewing SSL and SSS methods. Next, the chapter 3 explains the SRP-PHAT performance with Orchestral signals, its implementation and the experiments proposed for this approach. Then, the position informed sound source separation approach is tested by DUET and NMF models. Finally, the conclusions are listed and future research work is explained.

Chapter 2

STATE OF THE ART

2.0.1 Introduction

Hitherto, the computational acoustic scene analysis in a indoor venue with adverse acoustic conditions is still a challenging problem. During the last decades, many researches have been designed robust ASR systems with microphone arrays in adverse acoustic conditions, particularly to reverberant and noisy environments, taking the advantage of SSL for the enhancement of SSS methods, in order to recognise the active speaker [1, 2]. So, this work aims to propose an approach adapted specially for Orchestral music carried out inside of the Performances as Highly Enriched aNd Interactive Concert eXperiences (PHENICX) project.

In this chapter, the theoretical basements and a review of the relevant literature for this thesis are presented. Firstly, the differences between speech and musical signals are discussed. Secondly, a sorting of mixture models are introduced. Thirdly, the basic concepts and methods for sound source localisation are presented. Next, a specific approach of *Position-informed Sound Source Separation* (PSSS) for orchestral music and the literature review for this challenge is introduced.

2.1 Sources and Mixtures

2.1.1 Audio Sources

The temporal and spectral features are needed for separation algorithms for time-varying signals are analysed. Now, the basic aspects of the most important audio sources (speech and music) are briefly described.

- **Speech Sources**
  Speech is the oral form of human communication by using the vocal folds as excited signal. Combining this signal with the resonant mechanisms (vocal tract) and speech articulations generate a vast diversity of sounds [6]. They can imply emotions such as surprise, happiness or sadness. Speech sound can be split as a sequence of phonemes. Speech signals consist of harmonic components, with a periodic trait and an stochastic component produced by transients and noisy sounds. The harmonic part are sinusoidal partials, multiples of the *fundamental frequency* (F0) that defines *pitch*. The periodic part uses to have shorter duration than stochastic components. The pitch changes over the time in a range of 40 Hz to 150 Hz for male and to 250 Hz for female talkers.

- **Music Sources**
  In the musical context, pitch is strong related to musical tones and it is assumed to be identical to the F0. The F0 in music signals span the range from 30Hz to 4 kHz varies and they varies slower in time than speech signals. Each F0 or pitch can be discretized by *notes*, the simplest unit for music sequences. By contrast, the timbre defines the quality or color of the sound of a musical note and it is described by the relationship between harmonics and its amplitude factors of a harmonic serie. *Music Information Retrieval* (MIR) studies have shown different audio features to describe and classify music signals [7, 8]. In Western music there is usually a lot of overlapping of frequency components when musical instruments play together.

2.1.2 Music Mixtures

In the situation of real life performances or recordings, the venue where the sound is captured plays an important role, due to the captured signal at each microphone is a sum of direct signal and the reflected signal with an uncorrelated time lag.

- **Stereo Mixtures** Since some years ago, the most common format for sound recording and reproduction is stereophony (stereo 2.0) [64]. Even though, DVD and BlueRay formats make available multichannel recordings for 5.1 systems, the stereo format continue being the standard. Nowadays, most of audiovisual content broadcast still are in stereo. The stereo systems create a phantom source of a captured sound. Its position can be altered changing the gains applied in the left and right channel during mixing process, known as *amplitude panning* (AP). Through AP, a pan mapping can be built in a 2D plane, as it is mentioned in the section 2.4.2.

  Stereophony effect can be recorded on situ with a microphone pair making

use of current music stereo recording techniques as spot microphone, XY, ORTF, MS, NOS, AB, Decca Tree or Ambisonics are well defined in the literature [66].

Each technique provides diverse spatial effect in the mixture and they are combined to recreate music wavefronts. XY, MS and ORTF are stereo techniques, coincident and quasi-coincidents, quite well correlated where the polarity of source tends to be positive or defined by a direction of arrival between $45°$ and $-45°$, where the stereo image is solid. when more spaced microphones are more uncorrelated signals are captured and the spatial effect is generated by differences in time and differences in intensity. Instead with coincident and quasi coincident techniques mostly provides differences in intensity and poor response in time differences.

Amplitude panning is the most frequent virtual source positioning method. A sound signal is applied to loudspeakers with different amplitudes, formulated as

$$x_i(t) = g_i x(t) \qquad i = 1..., N \tag{2.1}$$

where $x_i(t)$ is the signal to be applied to loudspeaker i, g_i is the gain factor for the corresponding channel, N is the number of loudspeakers and t is the time variable.

However, the direction is dependent on the gain factors and in the time difference. The higher difference of time between microphones defines a time range that can be mapped as panning angle. AB microphone recording technique offers a very open stereo image depending on the distance between microphones and source. It could be more rare to define the angle of sources because some times can appear phantom images due to negative panning coefficients.

- **Multichannel Mixtures**
  There are many cases where a live performance in front of an audience is recorded and broadcast. In this situations complex techniques should be applied to analyse and compare between each pair of signals. An acoustic space mapping can be built in a 2D plane making use of M-microphones given more importance to the microphones closer to the source and with less probabilities of leakage.

2.2 Sound Source Localization

The localisation of sound sources by human beings is based on the binaural hearing system and on the analysis that the brain does continuously with the signals conveyed by the auditory nerve fibbers when any sound arrives at our ears. It is based on two different approaches of analysing the acoustic waveform, the first becomes a spectral analysis in which the comparison of sound energy across different frequency bands arriving at each ear provide the vertical dimension. It is considered a monoaural cue and it is generated by the attenuation of particular frequencies by the pinna and the outer ear. The second approach takes into account mainly difference cues,generated by the detecting and comparing the slight differences in level and time between the movement of the eardrums. It provides the horizontal dimension. The physical cues and physiological factors, as spectral cues, describe mechanisms that could support some aspects of binaural processing [9, 10, 11].

Spectral notches described in Figure 2.1 are noted as *Head Related Transfer Function* (HRTF) or the *Anatomical Transfer Function* (ATF). They are pretty important when sound comes from the back or to define the elevation angle for the sound source direction but they vary depending on the dimension and physiology (head dimensions, eardrum shape and the distance between the head and shoulders) [10]. *Interaural Intense Difference* (IID) and the *Interaural Time difference* (ITD) are important binaural cues for sound localization and physiologists have elucidated some of the the neural mechanisms involved in the processing [12]. In contrast, for each possible ITD it exists a cone of confusion, that describes an area where the sound could be located because there is the same distance and therefore the same time difference. So if we want to estimate the sound source position in a 3D-space we should to propose smart strategies able to define mainly the information supplied by HRTFs or ATFs. According to the RT_{60} and the *Signal-to-Noise Ratio* (SNR) present in the acoustic scene, the difference cues cannot appoint precisely the direction from which the sound comes in a three-dimensional space [10].

About Figure 2.1 depicts the three spatial cues for human hearing. On **A,** the spectral analysis is decoded in our brain as monoaural cue for SSL in the vertical plane. The interplay of sounds with the body, head and outer ears modify the spectrum of the sound influenced on the eardrum in a way depending on the location of the sound source in the perpendicular plane. *Head-Related Impulses Response* (HRIR) can be estimated with a *Knowles' Electronics Manikin for Auditory Research* (KEMAR) and averaging, at each elevation angle, 50 repetitions captured before. **B** depicts the difference in the arrival time of a wavefront at the two ears (Δt) are used to confine a sound source in the horizontal plane. The frequencies below 2 kHz provide information for ITD processing. Finally **C,** for

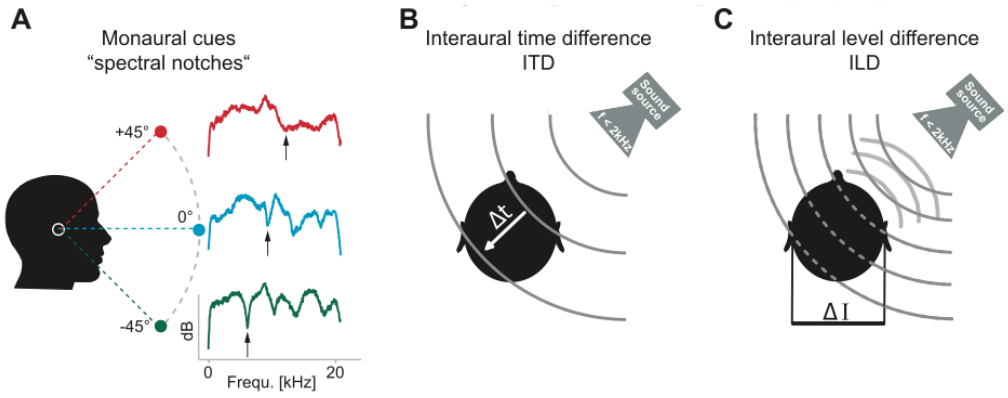

Figure 2.1: Spatial cues for binaural hearing.

higher frequencies than 2 kHz, the shadow effect of the head creates obvious differences in the intensity of sounds arriving at the two ears (ΔI), due to the sound propagation laws in the air. IID are employed for SSL in the horizontal plane. Picture published to [11].

2.2.1 Auditory inspired cross-correlation methods for source localisation

In the field of auditory analysis perception, Bergman proposed a cognitive procedure, named *Auditory Scene Analysis* (ASA), for describing acoustic complex scenes as the auditory system does analysing the mixture of sounds [13].This approach is out of fashion but the ideas that Bergman provided still are standing for computational implementations of algorithms known as *Computational Auditory Scene Analysis* (CASA), that estimates the sound separation capabilities of binaural human systems [9]. In this computational model different ideas have been proposed to describe how the time difference and distance between source and listener are computed. For instance, the Cross-Correlogram (CC) was described by the Jeffress theorem [14] but it was never applied until recent discoveries. Jeffress proposed the idea that each neural circuit, in which firing activity that arises from the same critical band of each ear, travels a long a pair of delay lines. With this mechanism the cross-correlation between each neural circuit is used to estimate the source position.

$$ccf(n, c, \beta) = \sum_{k=0}^{M-1} a_L(n-k, c) a_R(n-k-\beta, c) h(k) \tag{2.2}$$

where $a_R(n, c)$ and $a_R(n, c)$ represents the simulated auditory nerve response

at discrete time n and frequency channel c. β makes reference to the delay time candidate.

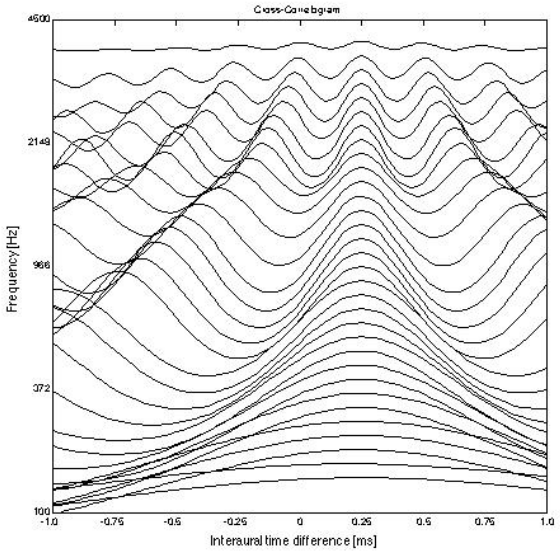

Figure 2.2: General cross-correlogram in binaural hearing (2 channels).[1]

This principle has encourage a variety of models based on CC, such as Faller and Merimaa proposed with an *Interaural Coherence* (IC) measurement, a normalised CC [15] , which may be defined for a specific frequency channel at time *t*. In [9], it is presented an extended literature related with these perception models, that demonstrate that ITD and IID easily could be emulated measuring difference cues between a pair of microphones, the *Time Difference Of Arrival* (TDOA), and they are sufficiently powerful cues to estimate the source position of binaural signals though in this case, HRTF or ATF are not involved.

Binaural CASA system approaches are simple models of source localisation that only makes use of the time lag and the distances between microphones, much useful for estimate the sound direction. Mostly, the efficient SSL methods make use of multi-sensor systems and the computation of the source position throughout the combination of the different pair of microphones with advanced signal processing techniques as CC or IC. These methods are based on the *Direction Of Arrival* (DOA) for maximising the SNR for each source. Nevertheless, it is still a challenge accomplish this task when undesired effects, as acoustic reflexions and a reduced SNR, are present in music live performances in indoor venues. The approach proposed with multi-sensor systems usually face with an *over-determined*

[1] www.casabook.org

or *even-determined* SSS problem due to the proportion between the number of mixture channels and the number of original sources.

2.2.2 Time Difference Of Arrival

The most part of microphone arrays signal processing techniques to estimate the source localisation are based on TDOA between each pair of microphones for each sound source. Different methods has been proposed to assess the sound source position by means TDOA [17, 18, 19] .

The TDE between signals from any pair of microphones can be computed by the cross-correlation function of the two signals after applying the appropriate weighted step. The time lag wherein the cross correlation function has its maximum is taken as the time delay between the two signals. The *Generalized Cross Correlation* (GCC) is a method proposed by Knapp and Carter in 1976 [20], and it is extensively used. GCC is weighted with different functions, but with the *Phase Transform* (PHAT) approach has been shown it perform well in noisy environments. However, to solve this kind of algorithms is quite difficult because it appears false peaks that may be stronger than the real solution.

2.2.3 Mixing Models

The simplest anechoic signal model with time delay between receiving signals can be defined by:

$$x_i(t) = a_i s(t - t_i) + w_i(t) \qquad i = 1, 2..., M \tag{2.3}$$

where a_i are the amplitude activations factors, t_i are the signal arrival time delays $w_i(t)$ is additive noise signals uncorrelated with the signal of interest $s(t)$ and M denotes the number of microphones. By means the Eq. (2.3), the TDOA between a set of microphones would be defined. Considering an M microphone system, given a source signal $s(t)$ and its frequency representation $S(w)$ at location $\mathbf{q}$ the captured signal from each microphone can be modelled as

$$X_i(w, p_i) = D_i(w, c)A_i(w)U_i(w, c)S(w) + W_i(w) \tag{2.4}$$

where the first term is defined as,

$$D_i(w, c) = \frac{e^{-jwc||\mathbf{q}-p_i||}}{||\mathbf{q} - p_i||} \tag{2.5}$$

it represents the delay due to the distance to the microphone, given by c, sound propagation speed, $p_i = [x_i, y_i, z_i]$, the known vector position for each sensor

and $\mathbf{q}$ is the hypothetical spatial position of the sound source. The term $A_i(w)$ is the frequency response of the system preamplifier and the energy losses in the air. $U_m(w, c)$ is the microphone directivity, $S_i(w)$ is the source signal, $W_i(w)$ models the sum of possible noise sources and the reverberation provoked by the sources and p_i is the position for each microphone. The reverberation effect can be model as a convolutive mixture and assumed as a new noise component. So, $W_i(w) = N1(w)+N2(w)$ Also, room acoustics and the radiation pattern for each source should be model for a more realistic model, but the scope of this work is quite large to introduce these important aspects. Anyway, the same model can be rewritten for discrete signals as:

$$x_i[n] = \sum_{n=1}^{N} a_{in} s_i[n] + w_i[n] \tag{2.6}$$

where n is defined by $\hat{n}T_s$, that comes from the period defined by the sampling frequency F_s in each sample and $\hat{n}$ it is an increment defined from 1 to ∞. The same model could be expressed as [16] in the frequency domain by,

$$X_{il}[k] = \sum_{n=-N/2}^{N/2-1} w_i[n] x_i[n + lH] e^{j2\pi kn/N} \qquad l = 0, 1..., \tag{2.7}$$

in this case, X_{il} refers to a matrix with the *Short Time Fourier Transform* (STFT) output for each channel stored at each row, in this case $w_i[n]$ is the window function for each i-microphone signal to smooth the boundaries in each frame, l denotes the frame index, H is the interval defined by the hop size parameter in the STFT and k is the frequency bin index. So, in this way *Time-Frequency Analysis* (TFA) can be applied. In the real world, this approach is quite simple because reflections should be involved by the reverberation or echo effect. This signal model is named convolutive model and it is expressed as,

$$x_i[n] = \sum_{n=1}^{N} a_{in} s_n(t - \delta_{in}) + w_n[n] \qquad i = 1, 2..., M \tag{2.8}$$

where $\delta(t)$ are Kronecker deltas and the $\delta(t-\delta_{mn})$ defines a delay between source n and sensor i. The model can be expressed by matrices,

$$\mathbf{x} = \mathbf{A} \star \mathbf{s} \tag{2.9}$$

2.2.4 Adaptative Beamforming Techniques

The most important SSL algorithms are based on *Adaptative Beamforming* (AB) or *Steering Beamforming* (SB) techniques that are commonly applied to cancel

the interfering signals and steer or generate a strong beam to the desired signal though its computed weight vectors. AB can be determined by a filter-and sum process, which filters some temporal components to the microphone signals before summing them to produce a single signal. This method delays the microphone signals so that all versions of the source signal are time-aligned before they are summed. This time alignment is considered as a signal-enhancement process. The most remarkable AB methods used in SSL tasks are GCC [20, 21], SRP-PHAT [26, 3, 22, 23], *Linear Constrained Minimum Variance* (LCMV) [24], *Minimum Variance Distortionless Response* (MVDR) [25] and *Multi-microphone Position-Pitch* (M-PoPi) [27]. In this work, SRP-PHAT is proposed and tested as SSL method with musical signals for its great advantages in indoor venues with low SNRs [26]. In such case, the beamformer output is known as the steered response power (SRP), which actually is equivalent to the of the GCC of all possible microphone pairing. The advantage for this algorithm is that it only requires TDOA between each pair of microphones for maximising the SNR. When the point or the direction of scan coincides with the source location, the SRP will be maximized. To improve its performance, the PHAT is incorporated resulting in the algorithm SRP-PHAT, which is superior in facing the adverse acoustical conditions than the other AB methods [23]. The accuracy of the SRP-PHAT algorithm is limited by the time resolution of the PHAT weighted cross correlation functions. So, the filter or weighted function applied could be improved to get a more smoothed SRP space and find quickly the maximum peak of the surface. However,this algorithm is quite robust to noise and reverberant fields, although the computational cost is a huge problem because SRP function has many local extrema and all of them must be assessed. Several SRP-PHAT methods have been proposed to reduce the computational cost, such as those based on *Stochastic Region Contraction* (SRC) and *Coarse-to-Fine Region Contraction* (CFRC) [28], *Particle Filtering* (PF)[29] , *Gradient Descent* (GD) [3] or *Quadtrees and Octrees* structures [30].

2.2.5 Generalized Cross Correlation

It is assumed that there are at least two microphones which capture the signals $x_1(t)$ and $x_2(t)$. These signals include noise, reverberation, and a time lag version (2.4) of a signal whose TDOA must be estimated. The best known TDOA estimator proposed in the literature is the GCC proposed by Knapp and Carter (1976) [20]. This method consists of prefiltering sound signals from microphone pairs then calculating the cross-correlation (2.2). The GCC based sound localisation methods is computationally simple, in comparison with other SSL methods. The

main advantage are the high accuracy and low computational cost.

$$R_{ij}(\tau_{ij}) = \int_{-\infty}^{\infty} \psi_{ij}(w) X_i(w) X_j^*(w) e^{jw(\tau_i - \tau_j)} dw \tag{2.10}$$

where $\psi_{ij}(w)$ is the weighting function, sometimes called pre-filter function, X_j^* denotes the complex conjugation of *j*-microphone signal and $\tau_i - \tau_j$ is the delay due to the distance between a ij-microphone pair. Remark that there are an amount to $K = M(M-1)/2$ microphone pairs that should be process. Different weighted function has been proposed [23], because it affects the time delay estimation performance and it can be chosen following some criterion. The role of ψ_{ij} is to give a sharper and larger peak in the GCC. Some authors have proposed *Maximum Likelihood* (ML), *Roth Impulse Response* (RIR) [31] and PHAT. The PHAT weighting [32] performs very well under realistic conditions, it removes amplitude information from the signal achieving it to take particularly attention on the phase information. Therefore, it forces the process to be less sensitive to noise and chiefly reverberation. Its equation is showed below:

$$\psi_{ij}(w) = \frac{1}{|X_i(w) X_j^*(w)|} = \frac{1}{|X_i(w)||X_j(w)|} \tag{2.11}$$

considering Eq.(2.10) and Eq.(2.11) the GCC-PHAT can be simplified to reduce the computational cost to M computations [22].

$$R(\tau_i) = \int_{-\infty}^{\infty} \left| \frac{X_i(w) e^{jw\tau_i}}{|X_i(w)|} \right|^2 dw \tag{2.12}$$

2.2.6 Steered Response Power - Phase Transform

The SRP-PHAT has been considered one of the most robust SSL algorithm in reverberant environment. As we mentioned before, SRP naturally extends the GCC method for a pairwise to a multi-microphone technique.

$$P_l'(\tau_1 ... \tau_M) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \int_{-\infty}^{\infty} \psi_{ij}(w) X_i(w) X_j^*(w) e^{jw(\tau_i - \tau_j)} \tag{2.13}$$

and it can be simplified as,

$$P_l'(\tau_1 ... \tau_M) = \int_{-\infty}^{\infty} \left| \sum_{i=1}^{M} \frac{X_i(w) e^{jw\tau_i}}{|X_i(w)|} \right|^2 dw \tag{2.14}$$

The SRP at spatial grid point $\mathbf{q}= [x, y, z]^T$ in each frame is denoted as,

$$P'_l(Q) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} R_{ij}(\tau_{ij}(\mathbf{q})) \tag{2.15}$$

where $\tau_{ij}(q)$ regards to the *Inter-Microphone Time-Delay Function* (IMTDF). This function represents the hypothetical time delay of arrival for the microphone pair (i,j) consequence of a sound source located at $\vec{\mathbf{q}}$ point. It is defined below,

$$\tau_{ij}(\vec{\mathbf{q}}) = \frac{||\vec{\mathbf{q}} - p_i|| - ||\vec{\mathbf{q}} - p_j||}{c} \tag{2.16}$$

where p_i denotes the i-microphone vector position, p_j for the j-microphone position, $\vec{\mathbf{q}}$ is the spatial grid point proposed at each iteration. Therefore, the space is discretised with the spatial resolution r and it could be delimited by the known microphone positions. The estimated source location $\vec{\mathbf{q_s}}$ at l frame time is considered to be that maximising $P'_l(x)$ function over the spatial grid.

$$\vec{\mathbf{q_s}} = \arg\max_{x \in \xi} P'_l(Q) \tag{2.17}$$

Dibiase, 2000; explains how the candidate TDOAs can be defined in terms of the DOA at the array origin. If the propagation vector is designed by ζ_0, the TDOA can be computed by:

$$\tau_{ij}(\vec{\mathbf{q}}) = \frac{-\zeta_0 \cdot (||\vec{\mathbf{q}} - p_i|| - ||\vec{\mathbf{q}} - p_j||)}{c} \tag{2.18}$$

and finally, the points in the direction opposed the direction of propagation can be defined in terms of the azimuth and elevation angles, θ and ϕ, as

$$\zeta_0 = \left\{ \begin{array}{l} \cos\phi \sin\theta \\ \cos\theta \sin\theta \\ \sin\phi \end{array} \right\} \tag{2.19}$$

Casually, this definition is identical to the Ambisonic B-format[2] encoding. These angles define the estimated DOA, relative to the local origin, O. By evaluating the Equation (2.18), the far-field steered response power can be expressed in terms of θ and ϕ.

$$P_{FAR}(\theta, \phi, Q) \equiv \sum_{i=1}^{M} \sum_{j=i+1}^{M} R_{ij}(\tau_{ij}(\theta, \phi, \mathbf{q})) \tag{2.20}$$

[2]An 3D extension of MS, adding additional difference channels, height and depth.

2.2.7 Limitations and Constraints

The most important limitations of SRP-PHAT is the high computational cost and the time consumption to be applied in *real-time*. The resolution in the 3D search grid will depend on the space sampling rate and in the number of microphones considered in our process. However, in a 3D-space the computational cost is increased due to the SRP-PHAT must be calculated for each possible 2D-plane. Also, we should to test how it works with Orchestral music signals in terms of window function, the FFT size, RT_{60} and diverse microphone recording techniques. Other constraint can be defined by the directivity pattern of microphones involved in the scene, omnidirectional patterns introduce more difficulties for localisation because it generates phantom images and polarity is not clear. However, cardioid microphone closes so much the directivity beam and some sources can be not detected. Also Rui et al. mention about phase problems for cardioid microphones and SSL [57]

2.2.8 Clustering Methods for SSL

Since years ago, the improving of these kind of techniques is a challenge and a lot of literature has been dedicated to solve the data clustering in a 3D-grid space: coarse spatial grid [28], the accumulation limits as a function gradient [3], Quadtrees and Octrees [30], Bayesian approaches [29] and others. To improve and optimise the clustering AB techniques, some authors have proposed *Artificial Intelligent* (AI) techniques in order to built-up the SRP-PHAT beamforming ability. Previous researchers demonstrate that the use of metaheuristics algorithm has been increasing instead of exhaustive and exact procedures [33, 34]. In this thesis, some *Genetic Algorithms* (GA) are proposed and tested to reduce the high computational cost of the 3D grid search generated by the SRP-PHAT. Finally, this part was discarded for the scope of this document, but it should be bearing in mind as a future work.

Genetic Algorithms

GA are a class of probabilistic search based on biological evolution. These methods are useful as search method for solving problems and for modelling evolutionary systems. They are also used in optimisation and combinatorial problems. GA operates on a set of solutions and it uses a fitness function (*consonance/adaptation*) to guide the search. Also, it showed that they works quite well with multimodal functions or high-dimensionality functions. GA stands on ideas from population genetics. Firstly, a population of individual components of population are created randomly. In the simplest case, each member is a part of a chain

and it can be seen as a candidate solution for the problem concerned. Mutations among members in the population generate some individual being more fit than others. This contrasts are used to influence the selection of a new set of candidate solutions at the next time step. While selection, a new population is created by duplicating the most successful members and deleting the less prosperous ones. Albeit, the duplicates are not precise, so there is a likelihood of mutation (random flips), crossover (an interchange of similar substrings between two members) or other random changes during the copy process. By permutating the preceding set of "good members" to a new one, the mutation and crossover process generate a new set of members, or samples, that theorically have higher possibility to be better than the average. At the moment that this cycle of evaluation, selection, and genetic processing is repeated for many reproductions, the average fitness of the population usually gets better and the members in the population represent improved answer to any problem was model in the fitness function.

Distinct GAs have been proposed in the literature, and some of them already have been shown its advantages combined with SB techniques [33]. In this work, we propose to test GA for clustering the SRP-PHAT function. There are some algorithms proposed to use in this framework: *Artificial Bee Colony* (ABC) [35], *Intelligent Water Drops* (IWD) [36], *Gravitational Search Algorithm* (GSA) [37] and *Particle Swarm Optimization* (PSO) [38].

2.3 Source Separation

The BSS problem was set down in 1982 by Bernard Ans, Jeanny Hérault and Christian Jutten [39] in the study field of neural modelling, for motion decoding in vertebrates. Instantaneously, this papers attract the attention of signal processing researchers, chiefly in France and later in Europe. After the middle of 1990s, BSS techniques has been faced by many researchers, coming from various domains: signal processing, statistics, neural networks. The BSS consists of retrieving unknown sources assuming from known mixtures. The simplest mixture model can be written as,

$$x(t) = As(t) \tag{2.21}$$

where $x(t)$ are the mixtures, $x(t) = (x_1(t), \ldots, x_p(t)^T \in \mathbb{R}^P$, s(t) denote unknown sources $s(t) = (s_1(t), \ldots, s_n(t)^T \in \mathbb{R}^N$ and A is a matrix with an unknown mapping for a linear transformation from $\mathbb{R}^N$ to $\mathbb{R}^P$. There consider different mixture models, however in this thesis we only will consider the models introduced on the section 2.3.3., linear instantaneous mixtures and linear convolutive mixtures. For this problem the most important is that A be invertible $(N \leq P)$ where N and P correspond to the number of sources and the number of sensors. The solution of this linear combination can be solved with a residual distortion

that are the typical indeterminacies of BSS problem. On the other hand, if N is greater than P, the mixing process is defined as *undetermined*.

The BSS problem is ill-posed without additional assumptions. The usual assumption is the statistical independence between the sources, which is a realistic assumption and justified. Instead, other assumptions can be success for ensuring source separability [40]. It is necessary that sources satisfy some basic assumptions: statistical independence, positivity and sparsity. Along the time, different methods have been proposed to solve this BSS problem: *Indepencent Component Analysis (ICA)*, *Independent Subspace Analysis (ISA)*, *Non-Negative Sparse Coding (NNSC)*, *Positive Matrix Factorization* (PMF), *Non-Negative Matrix Factorization (NMF)* and *Bayesian approaches*.

2.3.1 Sound Source Separation

Most available acoustic signals are mixtures of different sources. The separation of real signals remains difficult to achieve. Indeed, acoustics is among the most difficult applications field of source separation. The requirement for SSS techniques get up with various real-world applications as music, hearing aids and radio broadcast. Each one of these signals are captured through different techniques, which result in different signal qualities. In recorded mixtures, the easiest way of transforming a sound scene into a digital signal is to record all the sources simultaneously using one or several microphones. This technique is used for meeting recordings and live concert broadcasts, and it may involves both, near-field and far-field recording, with diverse spatial directivities and spacings.

The source mixing or *Cocktail Party* phenomenon is generated from the simultaneous propagation of sound waves through air from sources to the microphones and the acoustic-electric conversion performed by the sensor. In normal situations, the mixing process combines an additive superposition of the contributions associated with individual sources and the contribution of each source to each microphone.

The objective of SSS techniques is to separate one or any source signals from the multichannel mixture signal, being regarding some signals as undesired noise. The problem is that this undesired noise could be time-varying effects from the same signal which provokes some distortions. The cocktail party effect is very named in this field because it describes very well the source-separation problem, it makes reference to the situation when the observed mixture signal results from some people speaking at the same time in a indoor venue and all signals are interested. Evidently, this effect show us that not only SSS techniques are necessary for simultaneous grouping and sequential grouping of content, which are quite important in the musical scenario. Different authors propose models of attention or the use of *Gestalt principles* to solve this problem [9]. Instead, some SSS propos-

als for music substitute these models by applying *Time-Frequency Masks* (TFM) [43] and score-informed techniques [41].

As we mentioned in section 2.3.1, at 1990, Bergman introduced a new cognitive approach for describing hearing auditory environments [13]. A chain of processes, named ASA, try to explain the different processes required for the human auditory system and how perceptual organisation can be decoded by the human brain. Therefore, the first studies about SSS where carried out by ICA and ASA approaches [9]. ICA finds automatically the directions of the sources in a mixture, whereby making possible extracting a target source. The handicap for this technique is that only can obtain at most N sources for an N-channel signal. The beginning of 2000s gave an noticeable advancement of BSS technique with the NMF technique [44], which does not based on directional information. It assumes that a sound mixture is composed by a number of basis elements, which contains a latent characteristic of the sources. NMF eliminates the ICA limitations, that the number of signals should be larger than or equal to the number of target sources. In the last decade, different modifications on NMF have been proposed: semi-NMF, *Non-Negative Matrix Factor Deconvolution* (NMFD), 2D-NMFD, Multi-layer NMF, Tri-NMF, PARAFAC or *CANoncal DECOMPosition* (CANDECOMP) or *Non-Negative Tensor Factorization* (NTF).

Spectral Models

The majority of literature SSS techniques are applied on the frequency domain, exactly the spectrogram $V \in \mathbb{R}^{K\text{x}L}$, where K means the number of frequency components and L the number of frames. The spectral matrix can be decomposed in two matrices, $W \in \mathbb{R}^{K\text{x}I}$, that contains the basis components and $H \in \mathbb{R}^{I\text{x}L}$ where the activation coefficients are expressed in terms of time for each base.

$$V \approx WH = \sum_{i=1}^{I} w_i^T h_i \tag{2.22}$$

In this mater thesis it is considered multidimensional data provided by a microphone array, hence V_m can be extended to tensors ${V_{kl}}^m \in \mathbb{R}^{K\text{x}L\text{x}M}$. It can be written as,

$$V \approx WH = \sum_{m=1}^{M} V_m = \sum_{m=1}^{M} w_m^T h_m = \sum_{m=1}^{M} \sum_{i=1}^{I} w_{im}^T h_{im} \tag{2.23}$$

Some algorithms make use of the positivity constraints in unmixing systems. Usually researchers introduce prior knowledge in the factorisation process, applying TFA, creating an acoustic map and considering this information for each pair of microphones to apply panning knowledge in the process.

Sound Source Separation Techniques

Different techniques are used in the literature, but we have considered three well-know techniques for this thesis: *Degenerated Unmixing Estimation Technique* (DUET) [42, 43], as a simple approach, NMF [44] and *Spatial Cues Non-Negative Tensor Factorization* (scNTF) [5], an interesting approach for multi-channel process incorporating spatial information. In DUET approach is quite easy to use and extract pan mapping or histograms for 2-channels mixture, or in this case for each pair of microphones. Instead, the most powerful technique could be to combine this information with NMF approaches, introducing a matrix with the source localisation constraints. In [5], *Mitsufuyi* et *al.* introduce a new approach to combine spatial cues and NTF, that extends the NMF model to tensors. They propose to add prior knowledge introducing a new matrix Q in the factorisation, which contains spatial cues.

- **Degenerated Unmixing Estimation Technique**
  It is the first practical approach for separation of anechoic mixtures, where some panning knowledge can be extracted. Each channel is transformed into the STFT domain, where the relative attenuation and delay values between two observations can be calculated from the ratio as,

$$R_{21}(k,l) = \frac{X_1(k,l)}{X_2(k,l)} \tag{2.24}$$

  The symmetric attenuation is estimated as,

$$\alpha(k,l) = |R_{21}(k,l)| - \frac{1}{|R_{21}(k,l)|} \tag{2.25}$$

  and the lag time can be estimated by

$$\delta(\hat{k},l) = -\frac{1}{w_k}\angle R_{21}(k,l) \tag{2.26}$$

  where w_k in the angular frequency related to k, the frequency bin index. But in this work the delay values are computed previously by the SSL algorithm. However, with this approach a set of TFM is obtained for each channel.

- **Non-Negative Matrix Factorization**

  The principal differences between NMF and ICA are the constraints placed on the factorizing matrices W and H. The goal of this algorithm is to define

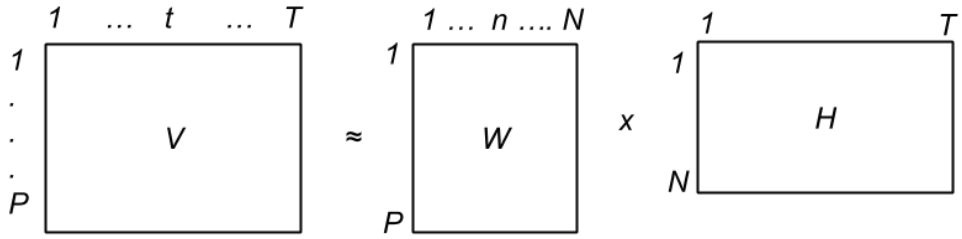

Figure 2.3: Basic NMF approach.

W and H that minimise the divergence between the spectral data V and the approximation WH.

NMF is popular for applying the multiplicative update rules in the factorisation. These rules are attractive due to they are simple and avoid the selection of an update parameter; besides,the multiplicative nature and the positivity of the magnitude spectral components guarantee that the elements don't become negative. This algorithm assumes the positivity of W and H. NMF is hold by the minimisation problem.

$$\min \mathscr{D}(V|W,H) \quad W,H \geq 0, \tag{2.27}$$

where $\mathscr{D}$ denotes the divergence or the cost function that grade the quality of approximation expressed by

$$\mathscr{D}(V|W,H) = \sum_{k=1}^{K}\sum_{l=1}^{L} d([V]_{kl}|[WH]_{kl}) \tag{2.28}$$

The most accepted cost functions are the *Euclidean* (EUC) distance,

$$d_{\mathrm{EUC}}(x|y) = \frac{1}{2}(x-y)^2 \tag{2.29}$$

the *Kullback-Leibler* (KL) divergence,

$$d_{\mathrm{KL}}(x|y) = x\log\frac{x}{y} - x + y \tag{2.30}$$

and *Itakura and Saito* (IS) divergence, which it was presented as "as a measure of goodness of fit between two spectras" [48].

$$d_{\mathrm{IS}}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1 \tag{2.31}$$

Using IS divergence has a strong connection with the Maximum Likelihood (ML) estimation of the variance parameters in the NMF model. Several estimation criteria are applied: *Expectation Maximization* (EM) and *Multiplicative Update* (MU) [58].

- **Non-Negative Tensor Factorization**

  Whereas NMF model is known as a *Two-Way Factor Model*, NTF can be considered as a *Three-Way Tensor* or *Three-Way Factor Model*, which can be considered as a set of multi-channels spectrograms. Three way arrays model are considered of this form,

$$x_{ptq} = \sum_{n} a_{pn} s_{nt} d_{qn} = \sum_{n} a_{pn} v_{tn} d_{qn} \tag{2.32}$$

  and in matrix notation can be written as,

$$X_q \approx A D_q S = A D_q V^T \tag{2.33}$$

  Actually, this model is named as the PARAFAC or CANDECOMP model. When the PARAFAC model presents non-negative constrain is called NTF. NTF can arise the number of factors and complexity. Nevertheless, in many situations it does not get up the number of factors and usually it get down the complexity. The method can be expanded to the spectral model proposed before:

$$V \approx WQH = \sum_{m=1}^{M} V_m = \sum_{m=1}^{M} w_m q_m^T h_m = \sum_{m=1}^{M} \sum_{i=1}^{I} w_{im} q_{im}^T h_{im} \tag{2.34}$$

  where $V \in \mathbb{R}^{K \text{x} L}$ denotes the spectrogram, $W \in \mathbb{R}^{K \text{x} I}$ is the matrix basis components, $H \in \mathbb{R}^{I \text{x} L}$ contains the activation coefficients and $Q \in \mathbb{R}^{I \text{x} T}$ is a matrix which can provides prior knowledge like spatial cues [5].

2.3.2 Position Informed Source Separation

In this Master thesis, the SSS system relies on AB techniques. These systems need prior knowledge, like the relative position of the microphones and sources and the time interval when the target source is active [41, 47]. Therefore, an Multi-Informed Source Separation model for Orchestral music is proposed, applying TFM, prior score information, the type of instruments evolved and the source positions computed previously with SRP-PHAT. The estimated source positions in the acoustic scene can be mapped to spatial cues. Moreover, through the

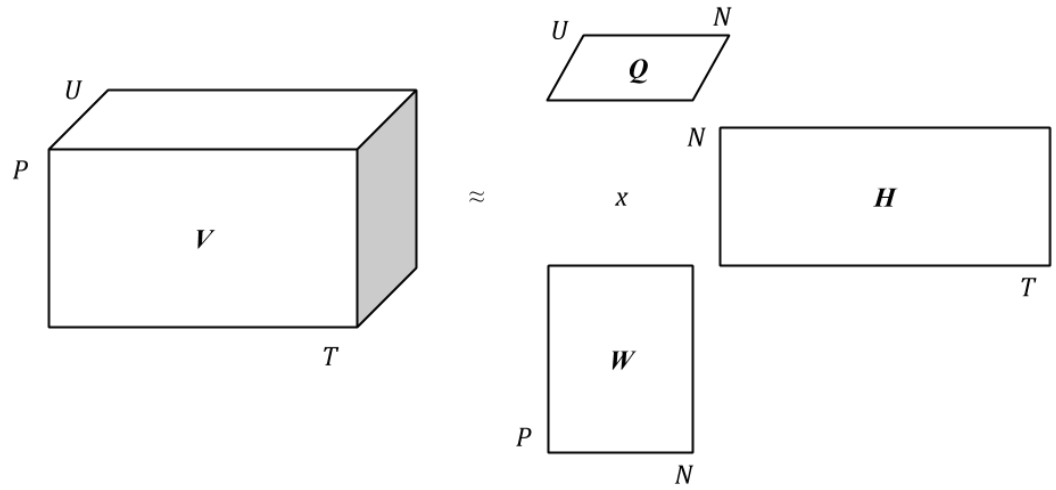

Figure 2.4: Basic NTF-PARAFAC approach.

source position and the microphone position can be built an acoustic map where panning features can be extracted for each microphone pair. Indeed, this acoustic map is essential to built new acoustic scenes applying rotations and scaling at each estimated signal. On the other hand, NMF does not produce any spatial information during the separation process and it is difficult to associate a spatial cue. By contrast, NTF model provides a chance to introduce prior knowledge in a third matrix, as [5], therefore scNTF proposals have been considered in this thesis. The idea is to introduce a Q, another matrix that describes the energy distribution of each base component on each channel, which can be thought as spatial cues. Hence both ideas might be combined to improve the separation quality and the computational cost.

Panning Information Retrieval

The sound direction has been used in Position-Informed Source Separation methods [5, 41, 46, 47]. Source separation from a microphone pair can be assumed as a stereo signal using the panning index metric introduced by Avendano [45, 46], a cross-channel metric applied in the frequency domain and directly related to the panning gained and TDOA between each channel.

- **Stereo Pan Mapping**
  In a stereo 2.0 system, the physical superposition of the waves produced by two loudspeakers provides the building of a phantom source. The position

of the phantom sources can be altered with the gains applied to each channel or with the Difference Of Arrival (DOA). However, usually *Amplitude panning* is used during the mix down process to avoid the comb filtering provoked by the *Time panning*. The panoramic parameter $\phi_n \in [0, 1]$. defines the weighting factors that multiply the signal in each of the mixture channels (α^L and α^R). Two rules are applied in this context:

- Amplitude constant law

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

- Power constant law

$$\begin{aligned} (\alpha_n^L)^2 + (\alpha_n^R)^2 &= 1 \\ \alpha_n^L &= \cos(\frac{\phi_n \pi}{2}) \\ \alpha_n^R &= \sin(\frac{\phi_n \pi}{2}) \\ \phi_n &= \arctan(\frac{\alpha_n^R}{\alpha_n^L}) \frac{2}{\pi} \end{aligned}$$

- Tangent Law

In a 2D loudspeakers setups, all actuators are on the horizontal plane. In a stereo system, two loudspeakers are placed in front of listener forming an angle phi of 135º and 45º. Angle theta denotes the perceived azimuth of the virtual sources. A panning law approximate theta from the gain factors of sources in each loudspeaker, g_L and g_R.

The direction is called panning angle or direction of arrival (DOA). Bennet et al. [65] conclude to estimate the propagation path from contralateral loudspeaker to ear with a curve line around the head with the tangent law.

$$\frac{tan\theta_T}{tan\theta_o} = \frac{g_L - g_R}{g_L + g_R} \tag{2.35}$$

where θ_T denotes the DOA for the virtual source image and θ_o is the loudspeaker base angle. g_R and g_R are defined as gain factors of the corresponding channels.

- **Spatial cues**

  In [5] is proposed other way to estimate spatial cues between two channels,

$$\vec{Q}_t = 2\tan^{-1}(\frac{q_{1,t}}{q_{0,t}}) \tag{2.36}$$

  where $\vec{Q}_t$ signify the angles between each signal in radians and q_t denotes the basis elements of a channel matrix. In the same research is proposed to expand the same idea over the spectrogram of 2ch-stereo signals, similarly as DUET proposes.

$$\vec{B}_{kl} = 2\tan^{-1}\left(\frac{X_2(k,l)}{X_1(k,l)}\right) \tag{2.37}$$

  where $X_2(k,l)$ and $X_1(k,l)$ are spectrograms bins for the left and right channels. As in DUET approach, they propose to look for the peaks in the histogram of $\vec{B}$ in order to define the mixing parameters for each source. $\vec{B}$ can be used to generate *Binary Masks* (BM) to separate frequency components with different DOA.

Chapter 3

SRP-PHAT FOR ORQUESTRAL MUSIC SOURCE LOCALIZATION

3.0.3 Introduction

In this chapter, we will show in detail how SRP-PHAT algorithm has been implemented, the experimental setup designed, to demonstrate the proposed hypotheses mentioned in the chapter 1. Following, it is reviewed the mainly aspects to apply SRP-PHAT with Orchestral signals, as the recommended TF parameters, the robustness of the algorithm with microphone positional errors, the microphone recording techniques applied, the RT_{60} and the results with concurrent notes with and without score information.

3.1 SRP-PHAT Deployment

At this section, it is explained the SRP-PHAT implementation step by step. Basically, I have started with the deployment of the most basic approach. Later, we have introduced some changes in order to improving the time consumption and to reducing the localisation error with music signals.

3.1.1 System Overview

The next scheme depicts the SRP-PHAT algorithm structure and the processes involved to estimate the sound source location. The most of procedures have been exposed in the previous section.

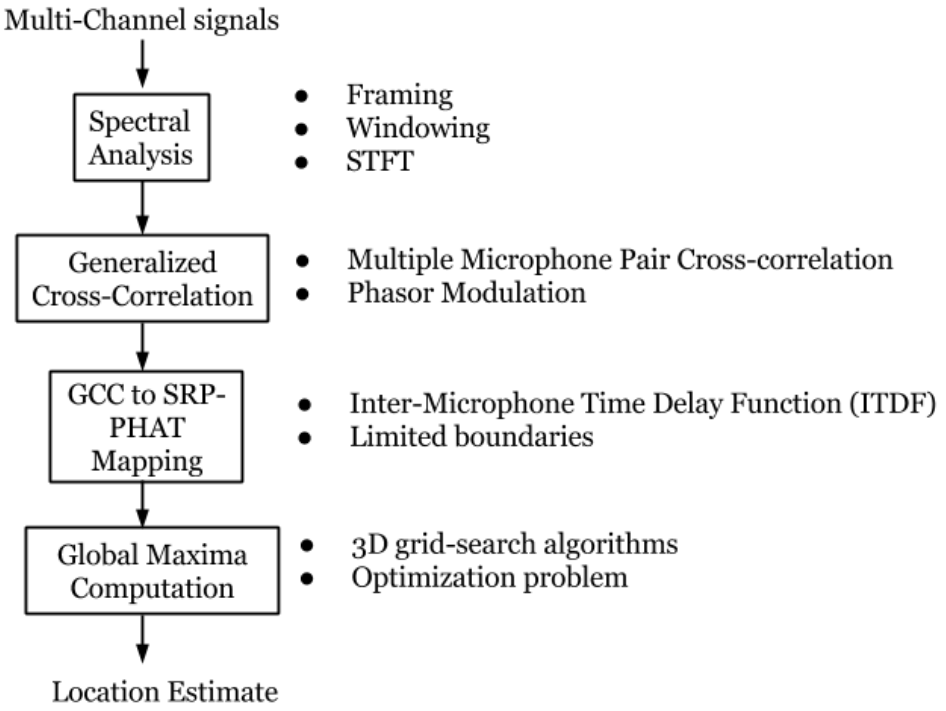

Figure 3.1: Basic SRP-PHAT structure.

Mainly, the algorithm needs a multi-channel audio signal as input, in other words one channel per microphone. The signal content depends on the location where the sensors and sources has been situated and on the acoustic conditions. Therefore, a huge amount of variables have to be defined to relate the different processes or to build a reliable ground-truth for source localisation. The goal of this algorithm is to relate the time features with the space taking in account the TDOA from all possible microphone pairs.

Time-Frequency Analysis

The basic SRP-PHAT implementation begins with a TFA that is kept in memory to execute the algorithm in different TF regions, although a pre-processing step can be applied, as it is proposed in the next sections. SRP-PHAT is based on GCC-PHAT features which is performed over the accumulated spectral representation. So, when a single source is playing the algorithm can be computed over the whole accumulated spectra. By contrast, when simultaneous sources are playing, they usually follow a music structure with different ornaments and motifs, producing time-varying dynamics and pitches in the recorded signal, generating changes in the salient sources along the time. At least, one frame is necessary

to obtain a successful localisation or reliable GCC features. The STFT lets us to set the parameters for Framing and Windowing process, as the size of *Fast Fourier Transform* (FFT), the window function type, the window size and the hop size [54]. Later, this parameters can be assessed in terms of the accuracy and efficiency of SRP-PHAT with music signals. With music signals is quite current to use a large frequency resolution, with a 4096 or 8192-window function, applying a hop size of 10-20ms, in which case, without using *Zero Padding* (ZP) in the FFT. As window function can be changed different kind of function are available (Hamming, Hanning, Blackman, Gaussian, and so on). This parameters will be tested in the next section of this chapter. In order to provide this optional analysis, STFT function[1]from FASST v1 MATLAB Toolbox[2][60] was redone to introducing hop size (H), different window functions and instantaneous frequencies (IF) as output argument. The function is available in my Github profile (`https://github.com/xaviliz`).

Generalized Cross-Correlation Estimation

The deployment of GCC-PHAT is based on the Equation 2.10. but it can be simplified on the Equation 2.12, which reduces the time computation for this feature extraction. Both versions have been implemented to test and compare the time consumption between them. Basically, GCC is computed between the signals proceeding from each microphone pair. In the Equation 2.10, GCC is computed for each microphone pair, whereas Equation 2.12 obtain the same result for the accumulation of each microphone signal. The result obtained at every microphone pair is showed in the next figure, where the highest peak represent the relation in time where the signals coincide. So, the GCC is defined by the number of microphone pairs and by the resolution of the phasor (Equation 2.10) [22]. To obtain the GCC, the accumulated frames are evaluated with a phasor modulation and the maximum value is kept in memory along the delay candidates or offset time. The higher peak represents the relation in time where the signals coincides, so the phase difference. The peak is unambiguously related with the delay between microphones and the source position. Each microphone pair obtains a GCC vector with the time resolution defined by the phasor modulator. By contrast, when more than one source are involved diverse peaks can appear on the GCC vectors.

As it is defined in [28], the phasor resolution can be defined by the maxima distance between a microphone pair and the sound velocity. To reduce the computation time and to increase the accuracy of GCC-PHAT, a band-pass filter can be applied rejecting the spectra content of frequency bins above 0.09 to 4 kHz, where mostly spectra energy in orchestra signals is concentrated. This process may be

[2]stft_multi.m

[2]`http://bass-db.gforge.inria.fr/fasst/`

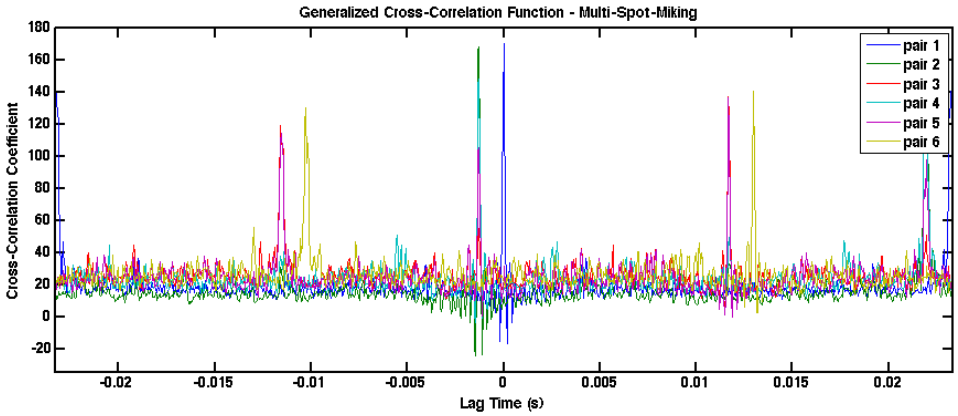

Figure 3.2: GCC-PHAT feature vectors at M microphones ($M = 4$).

done with BM processing, before to compute the GCC, applied over the accumulated tensor containing the M-spectrograms. It also implies a *DC-Removal*.

Mapping GCC

In this step, TDOA is mapped through evaluating the IMTDF (Equation 2.17) for each microphone pair on a 3D-grid space. Each point in the space is a possible microphone position and supposedly the source location will be at the position where the IMTDF becomes the global maxima. Therefore, GCC-PHAT feature vectors are evaluated by IMTDF. The room dimensions are used to limit the 3D grid, which defines a euclidean space where to evaluate the GCC-PHAT. The resulting function is normalised by the number of microphone pairs, *K*.

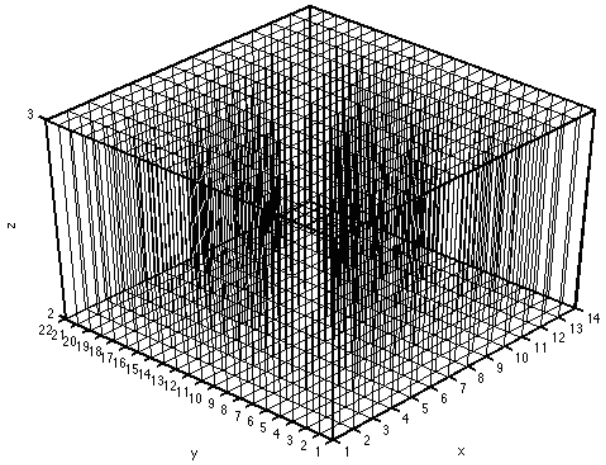

Figure 3.3: Spatial candidates in a 3D-grid.

Algorithm 1 GCC mapping algorithm

procedure GCC EVALUATION(*lsb, usb, micloc, GCC*)
  $efmax \leftarrow \max$ *(dist(micloc))* ▷ Maximum distance
  $bstart \leftarrow$ *lsb* ▷ Define rectangle boundaries
  $bend \leftarrow$ *usb*
  $R =$ *GCC* ▷ GCC feature vectors
  $M = length($*micloc*$)$ ▷ Number of microphones
  $ngrid = size($*R*$, 1)$ ▷ Number of vertex in a plane
  $np = size($*R*$, 2)$ ▷ Number of microphone pairs
  for i:=bstart(1) **to** bend(1) **step** r **do**
    for j:=bstart(2) **to** bend(2) **step** r **do**
      for k:=bstart(3) **to** bend(3) **step** r **do**
        $\hat{x} = [i, j, k]$ ▷ vertex to evaluate
        find distances between microphone locations and vertex
        dists $:= sqrt(sum(((ones($*M*$, 1) \cdot \hat{x}) -$ *micloc*$)^2))$
        calculate the difference between distances.
        diff $:= dists - dists'$
        convert distances to indexes.
        idx $:= round($*diff* $\cdot$ *(ngrid-1)*$/(2 \cdot$ *efmax*$) + ($*ngrid* $- 1)/2 + 1$
        v1 = *R*(*idx*)
        yval1 = sum(*v1*)
        pos = $\hat{x}$
  return (*yval1, pos*)

The spatial resolution *r* is the parameter that defines as much rigorous the localisation, morel time would be necessary to compute the global maxima search algorithm. Each r position involves a TDOA between microphone and source, and therefore a phase difference. Some tests to evaluating the grid resolution were made with speech signals in [3]. In this thesis, the spatial resolution was established to 0.1m, which is sufficient to determine accurately the location estimate.

Global Maxima Computation

In this step there is offered the possibility to apply *Optimization Methods* (OM) or AI processes to reduce the time consumption in the search grid algorithm. Some researchers have applied GD or SRC. In this case, however, we have evaluated the function in each vertex to know how SRP-PHAT function can be handle. So, we create a list with a sorting in descending order in terms of SRP-PHAT scoring with all the vertices and we select the first S-vertices, where S denotes the number of sources. The chosen vertices should correspond with the source location.

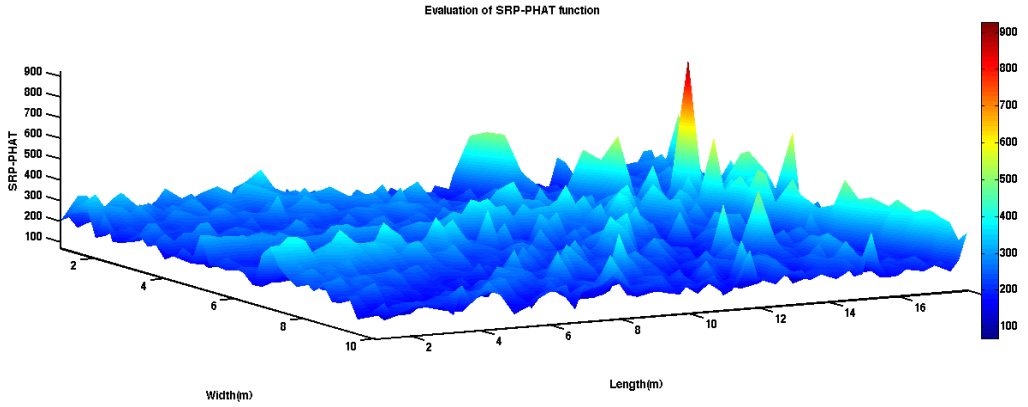

Figure 3.4: Evaluation of SRP-PHAT function for a plane with $r = 0.1$m.

SRP-PHAT function has spurious peaks around the global maxima when the spatial resolution is shorter than 1m and some false location can appear. Therefore, we decide to accumulate the SRP-PHAT values in each 1m^2, what it reduces the localisation error in the process and avoids fake source locations, created by phantom images in the stereo pair. The descending list can be used to evaluate the procedure with the ground truth.

Algorithm 2 Global maxima computation

1: **procedure** GLOBAL MAXIMA SRP-PHAT(yval, pos, S)
2: **relate** spatial grid position with the yval index
3: **accumulate** yval with a resolution of 1m
4: **generate** descent list with yval
5: **choose** the first S values of list as location estimates **return**

3.2 SRP-PHAT for ORCHESTRA SIGNALS

This section explains the methodology and experiments designed to the assessment of SRP-PHAT performance with orchestral signals. It is tested the parameters that can affect the source location estimate, as the type and size of the windowing function, the current microphone recording techniques used for Orchestra recordings, the effect of microphone's directivity pattern, the SRP-PHAT's robustness with the microphone location error, the adverse acoustic conditions in terms of *SNR* and RT_{60} and the SRP-PHAT's application on acoustic scenes with a single-source or simultaneous sources playing concurrent notes.

At the last point, it is suggested two approaches to employing SRP-PHAT with Orchestral signals:

1. The use of score information to identify the isolated frames where the orchestra sources are playing in order to compute the localisation.

2. The use of score information combined with harmonicity assumption of orchestra sources in order to generate a binary mask for each source to define the isolated TF zones, where to apply the localisation procedure. So, avoiding frames with overlapped frequency components proceeding from diverse orchestra sources.

3.2.1 Experimental Setup

Initially, diverse acoustic conditions are simulated with RoomSIM [49]. Over the same acoustic scene some experiments can determine the SRP-PHAT performance with Orchestral music and how much robust it is in a realistic environment. The dataset of acoustic scenes used in this thesis are generated virtually with a room simulator and an anechoic dataset of Orchestral instruments, we have taken in account the most realistic conditions.

Databases Used - RoomSIM

The defined method is applied to music extracts from a quartet from the 10 J.S. Bach four-part chorales dataset [52]. This dataset includes ten Orchestral pieces and the quartet is performed by bassoon, clarinet, tenor saxophone and violin signals, recorded in isolation while the musicians listened to the others by headphones. Individual melodic lines are mixed in each piece. But only the first composition is evaluated in this testing[3]. These signals can be considered as a *point*

[3]virtual XY stereo recording generated with RoomSIM at the acoustic scene 2, available in: https://drive.google.com/file/d/0B2Jk6yXhTDbnV01OUGZQTW9oVHc/view

source given that radiation law for music instruments [55] are not covered in this research.

By contrast, RoomSIM lets us build virtual acoustic scenes making use of these *point source* signals, it is a toolbox for MATLAB to create the acoustic conditions of a shoe box room and so simulating the most real conditions. It applies a ray tracing model to estimate the *Impulse response* (IR) for each microphone to virtualize the recording process with signals that you can provide. It provides convolution process to generate the signals virtually captured in each microphone taking in account Room Acoustics principles and *Inverse Square Law* (ISL) [56]. RoomSIM has a lot of functionalities and provide a lot of tools, though the manual it is quite short. Therefore, it needs to be defined a bunch of parameters to generate the virtual signal recorded in each microphone. The process is split in two phases:

- **Acoustic scene design:** This phase needs the following parameters to create an IR for each signal captured by each sensor applying ISL and the lag-time delay due to the distance between sources and microphone. In the Table 3.1, we point out all the parameters to define an acoustic scene. Combining microphone engineering parameters, music recording techniques are set (close spot microphone, stereo pair recordings like XY, ORTF, NOS, AB and the Decca Tree).

Acoustic Scene Design	
Scene Parameters	Nomenclature
sensor position	$[m_x, m_y, m_z]$
microphone spacing	$[d_1, d_2, \ldots, d_s]$
azimuth angle	$[\theta_1, \theta_2, \ldots, \theta_s]$
elevation angle	$[\phi_1, \phi_2, \ldots, \phi_s]$
room dimensions	$[L_x, L_y, L_z]$
absorption coefficient[4]	ρ_k
directivity pattern	*omnidirectional, cardioid...*
offset mic/source angle	$[\theta, \phi, \Omega]$

Table 3.1: Parameters of acoustic scene design.

- **Cocktail Party process:** It is applied this procedure in order to simulate the signals captured at each sensor for the designed *virtual-acoustic scene* when more than one orchestra source is involved. Thereafter, the *point sources*

[4] k defines the index of frequency band.

signals must be convolved with every IR provided by the previous process, to generate a multi-channel recording from the acoustic scene. This process is quite trivial but it could be decomposed in two steps:

- providing audio files for each source to compute the convolution and the sum of signals to simulate the recorded signal in each microphone.
- saving accumulated signal in a directory

By means RoomSIM and applying this methodology, we have generated two acoustic scenes:

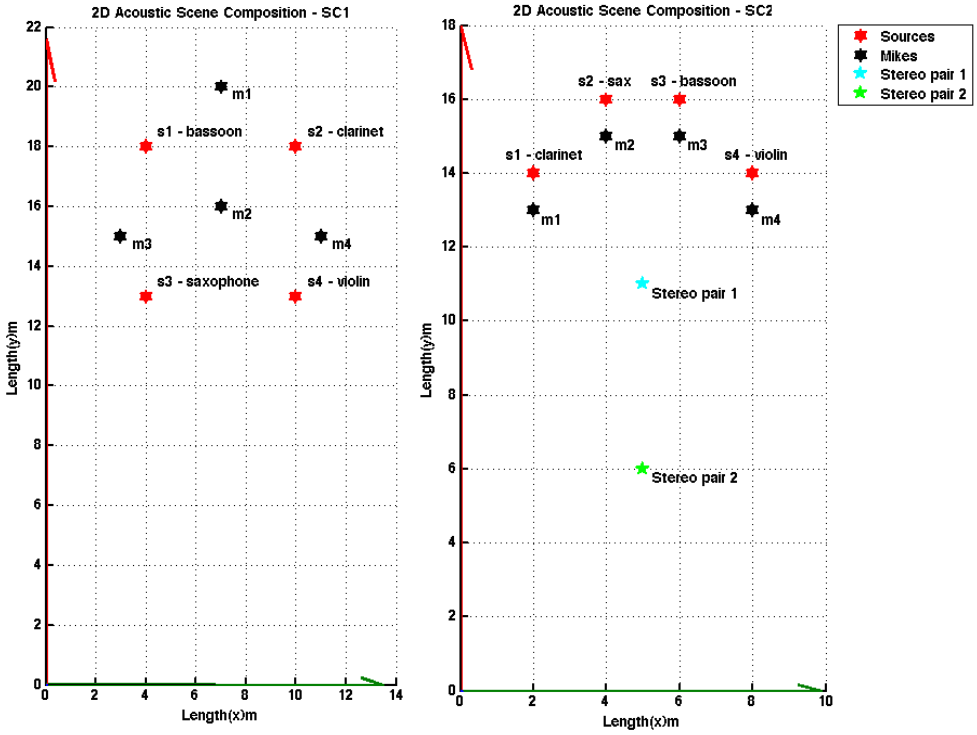

Figure 3.5: Acoustic scene layouts.

Acoustic Scene Design 1 (SC1): The following acoustic scene layout is simulated with 4 spaced omnidirectional microphones, see Figure 3.5 side left, in an anechoic room with a volume of 14x22x3m^3. In the simulated concert chamber, bassoon, clarinet, tenor saxophone and violin point sources are distributed. Whereas, a 4-microphone array is situated in random positions to evaluate the SRP-PHAT's performance with musical signals and how the main parameters affect source localization process (Table 3.2). All spatial parameters are defined in meters and at this scene, usually microphones have omnidirectional directivity.

Acoustic Scene 1		
Mic	Mic location	Directivity
m_1	$[7, 20, 2]$	*omni*
m_2	$[7, 16, 2]$	*omni*
m_3	$[3, 15, 2]$	*omni*
m_4	$[11, 15, 2]$	*omni*
Source	Source Location	Instrumentation
s_1	$[4, 18, 2]$	*bassoon*
s_2	$[10, 18, 2]$	*clarinet*
s_3	$[4, 13, 2]$	*tenor saxophone*
s_4	$[10, 13, 2]$	*violin*
Room Dimensions		$[14, 22, 3]$
Sound velocity (m/s)		340
Microphone Spacing		*variable*
RT_{60}		*variable*

Table 3.2: General parameters for acoustic scene design 1.

Acoustic Scene Design 2 (SC2): The following acoustic scene layout is simulated with 4 spot microphones[5], see Figure 3.5 side right, in a room with a volume of 10x18x3m^3. In this case, the scene simulates a close spot miking recording with 1 meter of microphone spacing with the closer source, and they are on axis. The spot microphones are set at different distance, 0.25, 0.5 and 1 meter from the source, in order to judge how it affects the localisation procedure. This scene emulates a more realistic microphone layout for music. With this layout, we have applied distinct microphone directivity pattern and different RT_{60} (0, 0.1, 0.25 and 0.5s). Also, when RT_{60} is present, we have defined other two vertexes in the space, where a stereo pair microphone are located, [5,11,1] and [5,6,1]. These pairs can be combined with the spot microphones to evaluate its contribution in SSL tasks. Distinct stereo pair recording techniques are tried out in these locations: AB, XY, ORTF, NOS and Decca Tree.

With these settings saved in a folder with a bunch of .mat files, because it can be kept with M-IRs per each microphone introduced in the scene. So, we can also create sets of audio files containing all the combinations possible, all duets, trios and single-source set with the same acoustic scene layout and conditions in order to evaluate the location estimate when different sources are active and how they interact. Mic locations, source location and sound velocity can define the TDOA

[5]http://www.dpamicrophones.com/en/Mic-University/Application-Guide.aspx

Acoustic Scene 2		
Mic	Mic location	Directivity
m_1	$[2, 11, 1]$	*omni, cardioid*
m_2	$[4, 13, 1]$	*omni, cardioid*
m_3	$[6, 13, 1]$	*omni, cardioid*
m_4	$[8, 11, 1]$	*omni, cardioid*
Stereo pair	Pair location	Techniques
p_1	$[5, 11, 1]$	*XY, NOS, ORTF, MS, Decca Tree*
p_2	$[5, 6, 1]$	*XY, NOS, ORTF, MS, Decca Tree*
Source	Source Location	Instrumentation
s_1	$[2, 12, 1]$	*clarinet*
s_2	$[4, 14, 1]$	*tenor saxophone*
s_3	$[6, 14, 1]$	*bassoon*
s_4	$[8, 12, 1]$	*violin*
Room Dimensions		$[10, 18, 5]$
Sound velocity (m/s)		340
Microphone Spacing		1
RT_{60}		*variable*

Table 3.3: General parameters for acoustic scene design 2.

in (s) for each source in a microphone pair and distance between microphones can define the DOA in (rad).

Evaluation Metrics

The following scheme shows the basic process designed to evaluate SRP-PHAT algorithm.

We use the predefined location for each source, as ground truth location, to measure the *Root Mean Square Error* (RMSE) with the location estimate. When simultaneous sources are active, the location estimate have less probabilities to fail, but the wave interference between sources worsen the results as we will see in the next section. The experiments designed in [50] show a metric called *Correct Localized Frames %*. It is introduced to evaluate the localisation process at each frame. It considers the location estimate have an error if it is off by more than 0.1m in any dimension. We have used the same metric but the error threshold is defined by more than 0.1m in the euclidean distance, being more restrictive. The ground truth location was defined at each frame considering a source is active

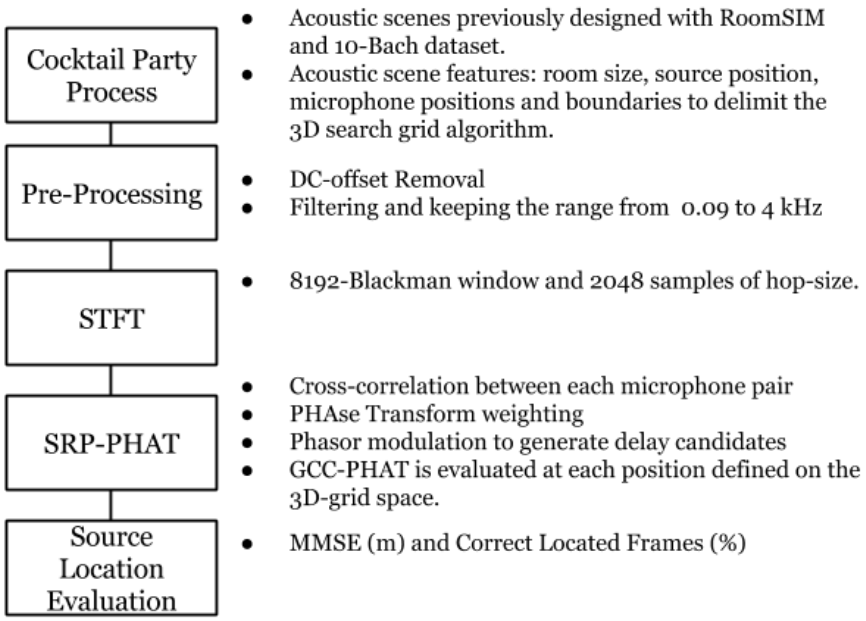

Figure 3.6: The scheme for SRP-PHAT evaluation.

if its close microphone channel's energy was greater than 30% of the dynamic range for that channel. So, this condition oblige to find out the distances between microphones and sources before to compute the localisation, which is affordable because we know all the positions. Then, automatically, the closer microphone can be assigned to each source.

When a single-source set is evaluated it is enough taking the maximum of the sorting in descendent order of SRP-PHAT to evaluate the position. Instead, when simultaneous sources are present in the acoustic scene, we should select more than one candidates. SRP-PHAT function has spurious peaks around the global maxima when the spatial resolution is shorter than 1m and some false locations can appear. Therefore, to accumulate the SRP-PHAT values in each $1m^2$ in a colormap, can reduce the localisation error in the process and so avoiding fake source locations. The descending list can be created after this process to evaluate the procedure with the ground truth, applying MMSE with the between the ground truth and the estimated locations.

$$RMSE = \sqrt{\frac{1}{SN}\sum_{j=1}^{S}\sum_{i=1}^{N}(\hat{p}_{ij} - p_{ij})^2} \tag{3.1}$$

where *s* denotes the number of sources involved in the scene and n the real source location of each source, the minimum RMSE for one candidate position

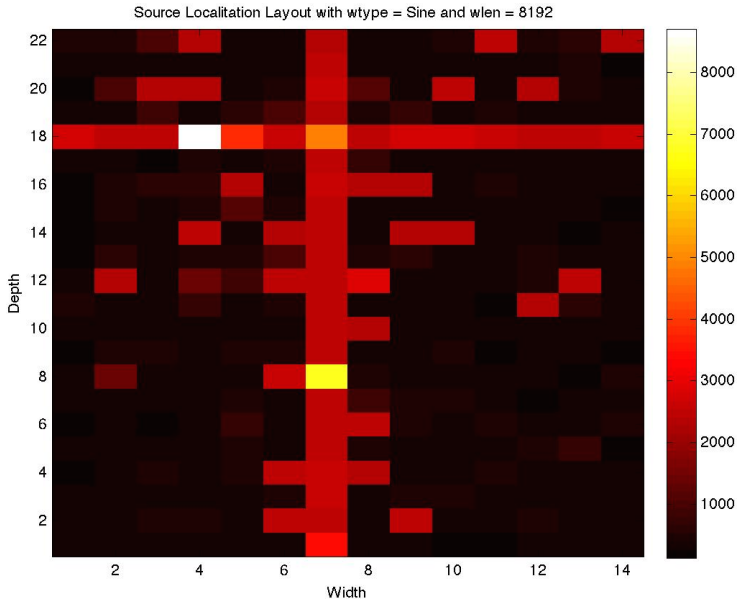

Figure 3.7: Accumulated SRP-PHAT function in a colormap with $r = 0.5$.

can be assumed as the estimated position by the system. When simultaneous sources are active to build a list with the lowest errors it is quite useful in the evaluation step.

3.2.2 Evaluation

The first experiment proposed in this section defines a overall performance for the different SRP-PHAT approaches.

1. Experiment 1 - SRP-PHAT vs TF Analysis

In the first experiment, it is used the acoustic scene design 1 to evaluate the localisation with a single source playing. So, the 4 different sets are defined with different RT_{60} (0, 0.1, 0.25 and 0.5s), respectively for s_1, s_3, s_4 and s_2. We have performed this testing with our SRP-PHAT implementation, using a MATLAB *Executable* (MEX) file, compiled in C, to reduce the computation time of GCC-PHAT. We evaluated over all frames of a 5 second segment of recorded Orchestra music, on selected section where all the instruments are playing. By default, the following test were done with N-points Blackman window, advancing with a hop of N/4 samples and with a sampling rate of 44.1kHz, where N is the number of samples of window function.

1. **FFT size**

In this test all the parameters are fixed less the FFT size and the hop size. The Figure 3.8 depicts that at least 8192-samples are necessary to obtain enough frequency resolution to ensure a proper source localization for lowest pitches. Anyway, the results demonstrate how the percentage of CLF for s_1, which corresponds with the bassoon, is so good with a low resolution and s_2 (clarinet) precises a higher frequency resolution to get a good percentage. On the other hand, saxophone and violin (s_2 and s_3) have a similar percentage and error. So, it may be said that the process is not robust to the timbre features but particularly it coincides also with the increment of RT_{60}. Hence, this acoustic parameter also reduces the efficiency and the accuracy of this algorithm when low frequency resolution is applied in the *TF analysis*. The best results were reached when the hopsize fulfil the next condition,

$$H = N/8 \tag{3.2}$$

where H denotes the hop size and N the window size.

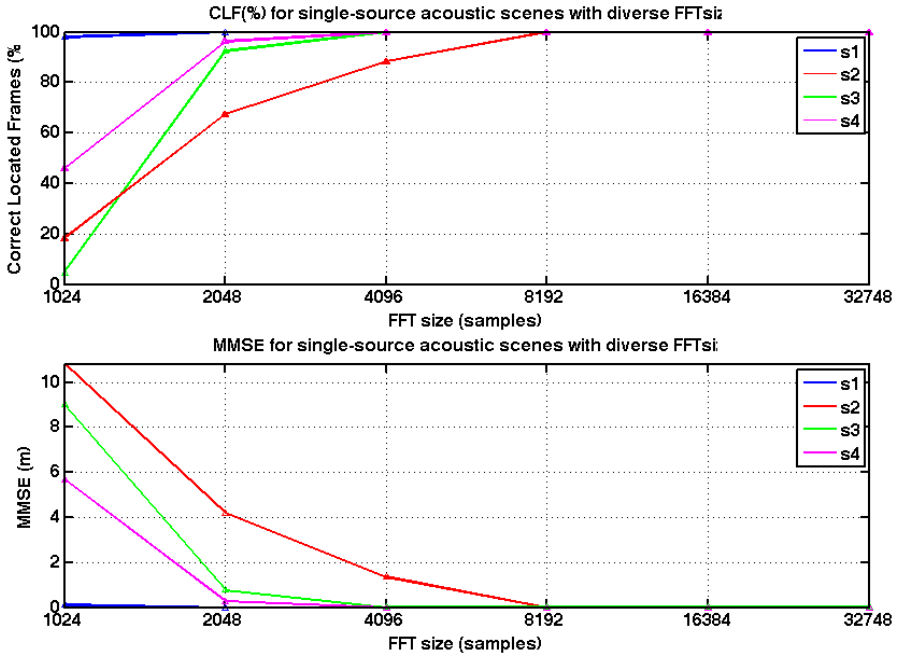

Figure 3.8: Results for single-acoustic scenes with different RT_{60} evaluating the localisation procedure by means CLF and RMSE with diverse FFT size.

2. **Window Function**

In this testing, the same conditions are applied but now, FFT size and hop size are fixed to 8192 samples and 2048 samples, and the window function is changed to test the influence of the weighting function in the localisation process. The evaluation was done with single-source and simultaneous sources settings in the same acoustic scene design 1. The single source settings were tested with omnidirectional microphones and with different RT_{60}. However, the simultaneous source settings were tested with omnidirectional and cardioid microphones to check the difference with a more realistic approach.

- **Single-source**

Acoustic Scene	**s1**		**s2**		**s3**		**s4**	
RT_{60} *(s)*	0		0.5		0.1		0.25	
Source	Bassoon		Clarinet		Saxophone		Violin	
Metrics	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**
Rectangular	100	0	20	10.82	100	0	60	3.54
Sine	100	0	100	0	100	0	100	0
Blackman	100	0	100	0	100	0	100	0
Kaiser	100	0	20	10.82	100	0	60	3.54
Bartlett	100	0	100	0	100	0	100	0
Triangular	100	0	100	0	100	0	100	0
Gaussian	100	0	100	0	100	0	100	0
Flat Top	100	0	20	5.4	100	0	100	1.2
Blackman H.	100	0	100	0	100	0	100	0
Hamming	100	0	100	0	100	0	100	0
Hanning	100	0	100	0	100	0	100	0

Table 3.4: Single-source scenes with unlike RT_{60} and window functions.

As the results show in the table above, in an anechoic model (s_1), all the window functions can works without introduce errors. But when RT_{60} increases, the *CLF* is reduced and the localisation error increases, as in the case of *Rectangular*, *Kaiser* and *Flat Top*. These functions are the worst choice to realise SRP-PHAT. Other remarkable point, it is the error, it is stable when a short reverberation time are present (0.1 and 0.25s). These levels are quite low in comparison with the music hall (1 to 2s), whereas it is close to the current level of RT_{60} in a studio room.

- **Simultaneous-sources**

  - Omnidirectional Directivity

  With this sort of pattern directivity, all the signals arriving from any angle are recorded, with different signal polarity. In this way, it is quite easy to generate *phantom images* of sources, which will be translated to uncorrelated signals in the stereo projection of each omnidirectional microphone pair. It lets increase the number of detected sources because omnidirectional pattern has a wide beam of capture and in this way more phase information about the acoustic scene. However, other difficulties can appear in source separation process due to the interference between the different *phantom images* with negative mixing parameters.

Acoustic Scene	**0**		**1**		**2**	
RT_{60} *(s)*	0		0.25		0.5	
Source	Quartet		Quartet		Quartet	
Metrics	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**
Rectangular	20	3.88	0	4.9	0	5
Sine	100	0	60	1	0	5
Blackman	**100**	0	**100**	0	**20**	**4**
Kaiser	20	3.88	0	4.9	0	5
Bartlett	**100**	0	**100**	0	**20**	**4**
Triangular	100	0	60	1	0	5
Gaussian	100	0	60	1	0	5
Flat Top	40	3.06	0	5.1	0	5
Blackman H.	100	0	60	1	0	5
Hamming	100	0	60	1	0	5
Hanning	100	0	60	1	0	5

Table 3.5: Simultaneous-source scenes with unlike RT_{60} and window functions.

When simultaneous sources are playing at the same time, like quartets, the localisation is pretty difficult due to the interference of sound waves. However, when RT_{60} is fixed to 0.25s, there are two special cases which offer better results than the other window functions. These are *Blackman* and *Bartlett* window function. With the worst adverse conditions, 0.5s, the other window functions don't localise none source, while they are able to localise sources in some frame. So, I recommend to use *Blackman* or *Bartlett* windows in the next testings or in approaches based on SRP-PHAT algorithm.

- Cardioid Directivity

In Table 3.4, for single-sources scenes, we see how the window function affects the location estimate, when anechoic acoustic scenes are tested (s1), any window function works well. Conversely, when RT_{60} increases the localisation error raises, as for instance, *Rectangular*, *Kaiser* and *Flat Top*. Anyway, if the results are compared with Table 3.6, cardioid microphones provides higher CLF % and reduced error than omnidirectional microphones. This effect is due to the cardioid microphones don't capture so much phantom images,because it covers more or less an area of 90° on each frequency component, so they can be capture a narrow area in the wavefield and leakage is reduced. So, the errors are reduced and the efficiency increases. Again *Blackman* and *Bartlett* are the best scored window function when Orchestral signals are involved in adverse acoustic conditions.

textbfAcoustic Scene	**0C**		**1C**		**2C**	
RT_{60} *(s)*	0		0.25		0.5	
Source	Quartet		Quartet		Quartet	
Metrics	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**	**CLF(%)**	**RMSE(m)**
Rectangular	20	3.88	0	4.88	0	5
Sine	100	0	100	0	20	4
Blackman	**100**	0	**100**	0	**20**	**2.75**
Kaiser	20	3.88	0	4.88	0	5
Bartlett	**100**	0	**100**	0	**33**	**1**
Triangular	100	0	60	1	20	4
Gaussian	100	0	60	1	20	4
Flat Top	20	3.84	20	3.96	0	5
Blackman H.	100	0	60	1	0	4.86
Hamming	100	0	60	1	20	4
Hanning	100	0	60	1	20	4

Table 3.6: Simultaneous-source scenes with unlike RT_{60} and window functions.

3. ***SNR***

In some music performances, noise is other source pretty common. It can be generated by the audience, air conditioner systems, humming or electric parasites. Therefore, it is interested to know how these sources can affect the SSL process and which levels are required for a proper functioning.

At this test, we have used the conditions than in the previous test, applying the optimised parameters previously estimated, 8192 samples of FFT size, 2048 samples of hop size and weighting each frame with *Bartlett* window function. In order to generate different situations to testing noisy environments with constant amplitude, a white noise source has been added at all the audio channels keeping 30, 20, 10 and 5 dB of SNR between the Orchestral signals and the added noise source. In order to know the amplitude of noise component in each audio channel, we have used the difference between the maximum peak value at each channel and the SNR ratio.

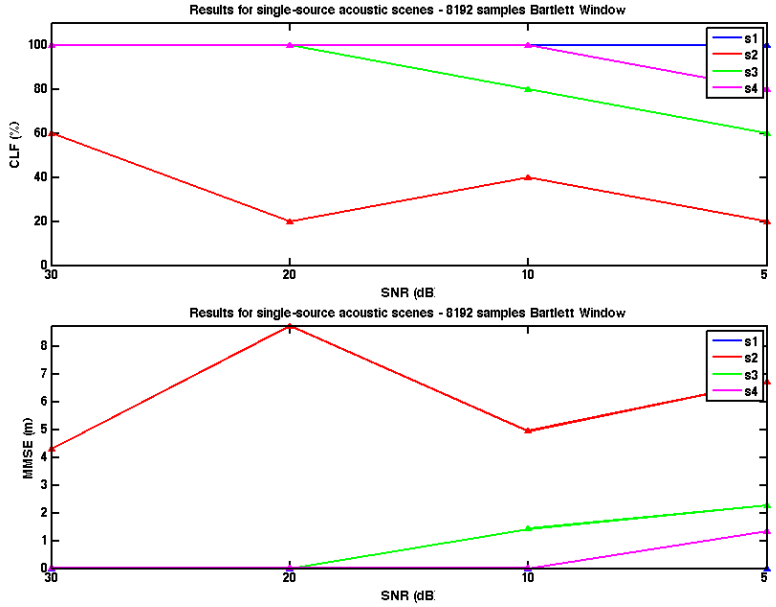

Figure 3.9: Results for single-source acoustic scenes with different RT_{60} evaluating the localisation procedure through CLF and RMSE with diverse SNR.

As the above figure depicts, when reverberation time is lower than 0.25s, the algorithm is quite robust to the lowest levels of SNR for single-source. With higher levels than 0.5s, noise sources can generate large localisation errors, reaching more than 8 meters of deviation with SNR of 20dB. In the signal model proposed for this thesis (Equation 2.4) assumes that reverberation components may be considered as a noise source composed by delayed harmonic traces of all sound sources. Hence, if RT_{60} and SNR decrease the process can discriminate quite well the estimation. When SNR and RT_{60} are below 10dB and 0.25s, respectively, the system works with a 80% of efficiency and with deviations of 1m. This results were obtained with omnidirectional microphones.

The following figure shows the results for scenes with simultaneous-sources, with omnidirectional and cardioid microphones. In this situations, the results for CLF(%) and MMSE are similar to the single-source's results. Nevertheless, it shows the difference of outcome between omni and carded directivity patterns. On one hand, the cardioid directivity pattern reduces the error and increases the efficiency when SNR is reduced as far as 10dB, as it is shown by the comparison between the pink and the green line (0.25s). On the other hand, with 5dB-SNR levels which is ratio quite unlikely in a realistic environment, cardioid directivity pattern are not the best scored by the metrics (black line).

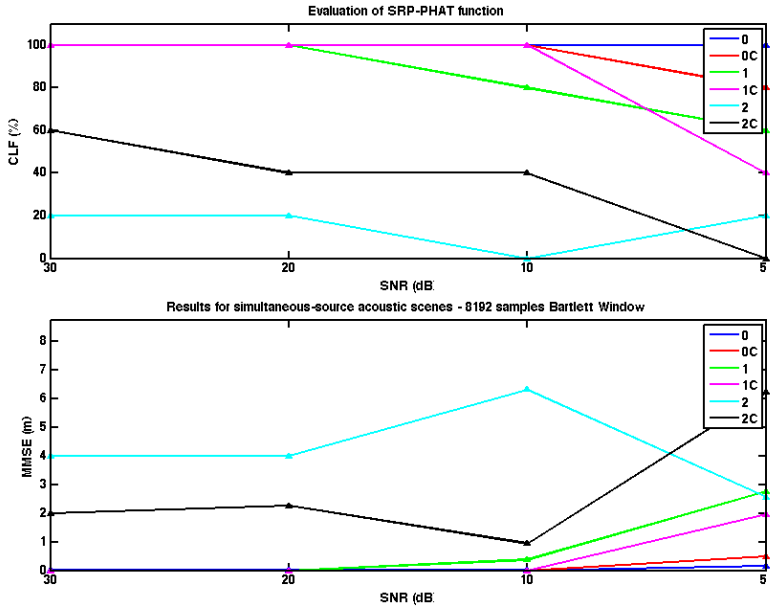

Figure 3.10: Results for simultaneous-source acoustic scenes with different RT_{60} evaluating the localisation procedure through CLF and RMSE with diverse SNR.

2. Experiment 2 - SRP-PHAT vs Microphone deviation

In a realistic environment, to know the precise microphone location in the 3D space is not a trivial challenge, and the microphone locations are vital parameters for SRP-PHAT algorithm. Furthermore, all these tests are done within a virtual domain. It is pretty relevant to know how the deviation of microphone location affects the process, to define a realistic experiment. In our framework to add controlled deviations in the positions is quite easy. So, we propose an experiment with the previous settings of acoustic scenes, applying random deviations at each microphone location. In order to achieve this proposal, 10-random trials were

launched with different ranges of deviation (mm, cm and dm) and later the average is computed for the results of each trial distribution over the same frame. Omnidirectional and cardioid directivity pattern are tested separately to show the gap between both.

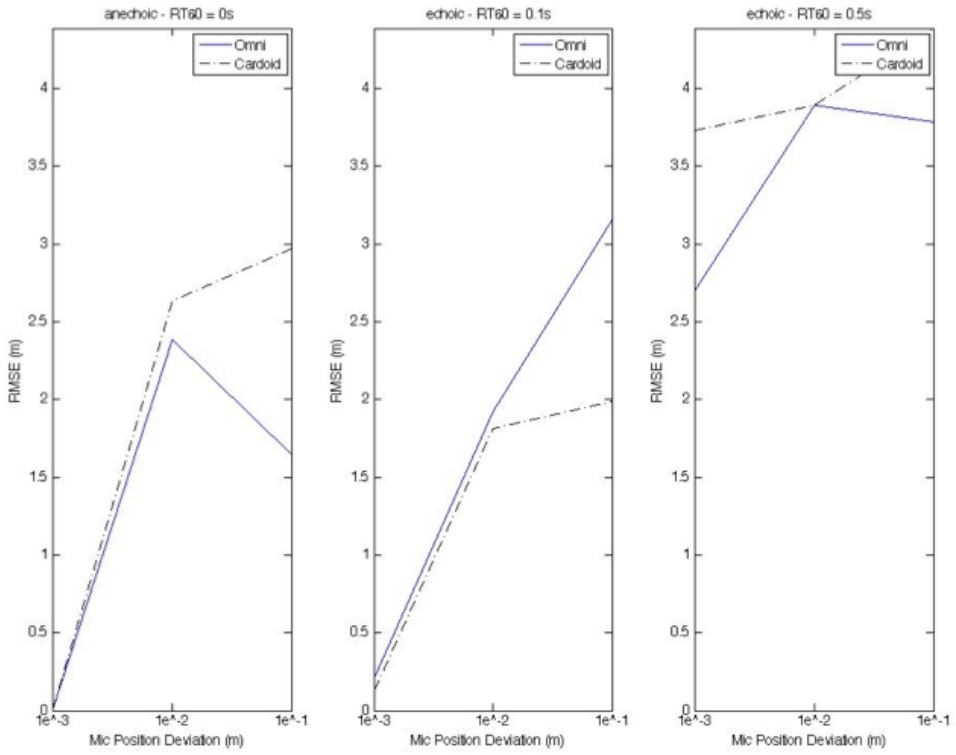

Figure 3.11: Results for relative deviations in the microphone locations with different scenarios.

Again, the localisation error of the proposed SSL system is biased by the reverberation time. With a few millimetres of deviations on the microphone location don't bias so much the original result with anechoic mixture models. However, when reverberation time increases, the deviations raises the localisation error. Besides, deviations of centimetres introduce at least 3 meters of error. By contrast, longer deviations can reduce or keep the error introduced. In the most realistic situation, on the right side, with 0.5s of reverberation time (convolution mixture approach), the original error introduced by the system is so high and deviations in the microphone locations arise the error but generating a low variance than with an anechoic mixture approach. Obviously, the algorithm is not robust to these deviations and it could be easily introduced in a realistic experiment. As it was explained in the previous chapter, the algorithm uses phase data provided by the evaluation of GCC at each microphone pair.

Next, GCC are evaluated with the distances between the proposed grid positions and the microphone locations. If the microphones are displaced, the location estimate will be moved. However, with this knowledge, we may think in new procedures to compensate this errors at the future or to design a fixed microphone array installation as it is proposed in [28].

dev (m)	*Bassoon*	*Clarinet*	*Saxophone*	*Violin*
0	0	0	0.1291	0.1291
0.01	1.3712	0.5052	1.5654	1.7094
0.02	1.8658	0.5351	2.3303	2.5264
0.03	0.0781	0.0926	3.1351	3.1352
0.04	1.8922	0.5136	2.2719	3.0111
0.05	3.7030	3.7030	1.5924	1.5924
0.06	1.8983	0.5266	2.2706	3.0112
0.07	0.0745	0.0884	3.1352	3.1352
0.08	1.8660	0.5192	2.3371	2.5529
0.09	1.4146	0.5487	1.6465	1.7185
0.10	0.0577	0.0577	0.9239	0.1357

Table 3.7: RMSE for deviations of microphone locations at single-source scenes.

The above table depicts the mean of RMSE for each source when each less than 10 centimetres are considered. The test was performed for this deviation in all axis, shaping a square around the original position. Any short deviation can introduce large and random errors, the pattern obtained doesn't show clear conclusions.

3. Experiment 3 - SRP-PHAT vs Microphone Recording Techniques

In this subsection, we propose to evaluate the current microphone recording techniques, well-known for music recording and introduced in the Chapter 2. First at all, it is shown how the microphone recordings can be combined to change the shape of SRP-PHAT function. For instance, close recordings with multi-spotting microphone provides accurate TDOAs in each microphone pair, but introduce more phantom images. However, stereo coincident microphone techniques as XY, offers poor TDOA features and the phantom images are reduced, hence localisation is done by areas. Furthermore, non-coincident stereo recording techniques, as AB with a meter of distance, define the localisation through lines. The results on the table demonstrate that combining both approaches, close-spot miking and stereo pair techniques, the localisation error can be reduced.

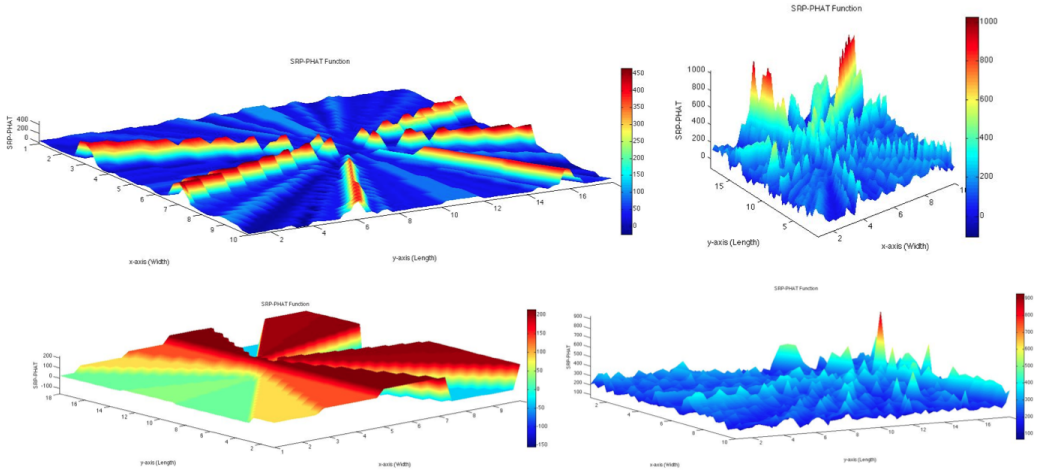

Figure 3.12: SPR-PHAT for current music recording techniques.

Again, different directivity patterns are proposed in the close microphones, but cardioid microphones are applied for stereo pair techniques and omnidirectional pattern in Decca Tree setups. In addition, manifold distances between the close microphones and each source are evaluated. Stereo pairs techniques and Decca Tree are located 4 meters away of spot microphone array (far field).

In Figure 3.12, SRP-PHAT shapes are shown: AB stereo pair technique (non-coincident microphones) provides a rough shapes (top left); Decca Tree recording technique is applied. The differences in time between involving three omnidirectional microphones provide a function with narrower peaks (top right). XY microphone pair (coincident technique) selects wider areas and the accuracy is reduced (down left). Multi-spot miking techniques, with 1 metre of spacing for each source. In this case, the most salience instrument in the frequency magnitude is located. The estimated position is quite accurate (down right). Apparently, each microphone recording technique provide different and complex SPR-PHAT shapes and they can be combined to improve the source localization process with music signals.

In Table 3.8, it is noted that the best results with anechoic mixture approach DecaTree is the best option, though it introduces the highest location error. It is effective but not very accurate. Conversely, when distance between source and sensor for the close spot microphones is reduced cardioid pattern responses more accurately than omnidirectional one. By contrast, when RT_{60} increases the efficiency (CLF) is reduced drastically, but if stereo pair techniques are combined with the spot miking approach, the algorithm offers an appropiate CLF score (Multi-spot card. + AB (1m)), keeping the error so short. So, it demonstrates that the combination of a cardioid spot microphone array and stereo pair micro-

Recording Technique	Distance (m)	RT_{60}	CLF (%)	MMSE (m)
Multi-spot omni.	1	0	90.7407	$1.731e^{-15}$
Multi-spot card.	1	0	96.2963	$7.0029e^{-16}$
Multi-spot card.	0.5	0	96.2963	$4.5263e^{-16}$
Multi-spot hyper-card.	1	0	96.2963	$4.5263e^{-16}$
Multi-spot card.	0.5	0.5	33.9286	$4.4409e^{-3}$
DecaTree (omni)	**4**	**0**	**100**	**0.4056**
Multi-spot omni.	1	0.1	90.9091	$1.3944e^{-15}$
Multi-spot card.	1	0.1	92.7273	$4.615e^{-16}$
Multi-spot card.	**0.5**	**0.1**	**96.3636**	**$4.5247e^{-16}$**
Multi-spot hyper-card.	1	0.1	94.5455	$4.5263e^{-16}$
DecaTree (omni)	4	0.1	80	0.35
Multi-spot omni.	1	0.5	7.1429	$1.9984e^{-15}$
DecaTree (omni)	4	0.5	5.3571	0.40
Multi-spot card.	1	0.5	23.2143	$4.4409e^{-16}$
Multi-spot card.	0.25	0.5	37.5	$6.5409e^{-5}$
Multi-spot hyper-card.	0.5	0.5	30.3571	$1.2449e^{-1}$
Multi-spot card. + ORTF	1	0	100	$1.102e^{-11}$
Multi-spot card. + XY	1	0	100	$6.5791e^{-16}$
Multi-spot card. + XY	0.25	0.5	57.1429	$5.4123e^{-2}$
Multi-spot card. + XY +ORTF	0.25	0	63.6364	$7.357e^{-3}$
Multi-spot card. + ORTF	0.25	0.5	58.9286	$4.4409e^{-16}$
Multi-spot card. + AB (1m)	**0.25**	**0.5**	**80.3571**	**0.0022**
Multi-spot card. + DecaTree	0.25	0.5	42.8571	0.4

Table 3.8: Simultaneous-source scenes with unlike RT_{60} and window functions.

phone located some meters away is a good strategy to success SSL with this sort of music signals.

4. Experiment 4 - SRP-PHAT vs Concurrent Notes

In this testing, we want to depicts how harmonic model and MIDI/score can be combined with SRP-PHAT to isolate the notes or segments of music discarding overlapped frequency bins at a simple harmonic approach taking in account all the sources a long the time. So, we need the MIDI files with the piano roll played by each instrument. In [53], MIDI files for Bach10 dataset and other datasets are available. The files are very well aligned with the harmonic components in order to avoid the transients at the beginning and endings of each music event.

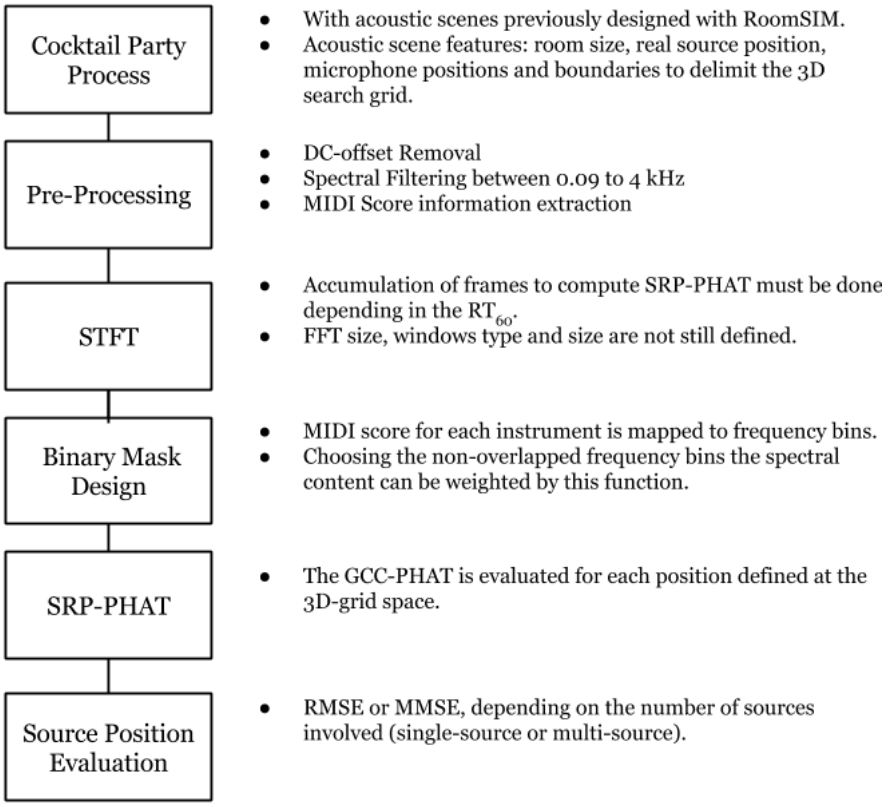

Figure 3.13: The scheme for position informed SSL strategies with music signals.

Aligned MIDI score offers temporal and spectral information to the separation algorithms, they may help to improve the source localisation process generating a binary mask in order to discard all the unneeded spectral information, leaving active only the frequency bins when there is non-overlapped spectral content. This step provides a masked spectrogram where the non-overlapped bins are hold and the rest are zeros (Zeroing). It should remove the contribution of the sources to obtain the location estimate from the target selected.

The TF representation is computed with 8192-points (185.8ms) Blackman window, advancing with a hop of 1024 samples (23.2ms) and a sampling rate of 44.1kHz. We evaluated over all frames of a 5 second excerpt of recorded Orchestral music (215 frames), on selected section where all the instruments are playing and without control what was being played by any musical instrument. We also created scene settings of audio files containing all duets, trios and single-sources with the same acoustic scene layout and conditions in order to evaluate the location estimate when different sources are active.

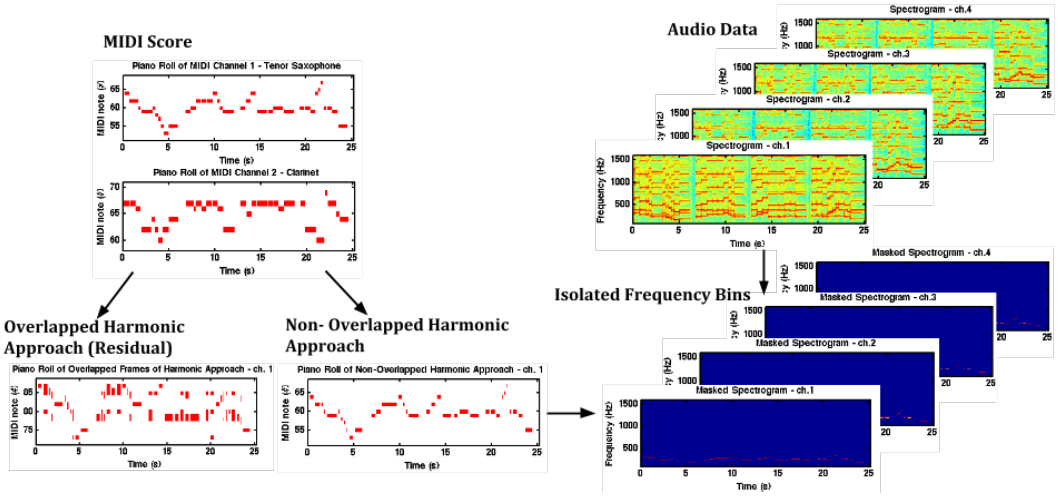

Figure 3.14: The procedure used to isolate the non-overlapped frequency bins.

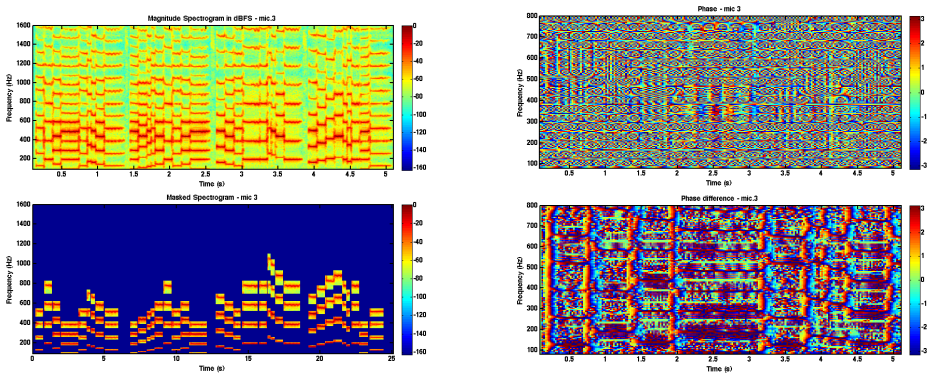

Figure 3.15: Magnitude, phase spectra, the masked magnitude spectra and the instantaneous frequencies.

- **Assumptions**

  The methodology stages for this testing are quite similar to the last one. Just before SRP-PHAT, Binary Masks (BM) at each instrument are designed, checking the harmonic approach for the rest of MIDI scores and avoiding the overlapping data. In the testing a 4 harmonic model is assumed. Applying this mask over the spectrogram the selected source is located. Diverse BM are generated at each scene, as many as sources are present in the scene.

The system lets us pick out the number of harmonics to remove no target contributions.

First, single-source scenes with orchestral signals are assessed and later simultaneous-sources are evaluated with orchestral signals. CLF metric is not necessary this time. In the next SSL experiments, we apply SRP-PHAT over the accumulated TF representation no each frame. Instead we can use the *Correct Localised Source* (CLS) to indicate the target source located. When binary masks are used, there is a target to localise. It may return the other sources or none source, when high errors appears.

- **Results with single-source scenes and score information**

  With this procedure, the location estimates are more accurate and phantom image effect is reduced . In this case, only RMSE evaluation is done because the SRP-PHAT is computed over the accumulated TF representation of first 5 seconds-excerpt without binary mask and with binary mask at each source. The evaluation is done over the accumulated spectra because when binary mask is applied, SRP-PHAT only must be computed in some frames and some frequency bins, which reduce drastically the computation time of GCC-PHAT. Otherwise, the evaluation will be biased by the frames which the algorithm is not applied.

Acoustic Scene	**s1**		**s2**		**s3**		**s4**	
Source	Bassoon		Clarinet		Saxophone		Violin	
Metrics	**RMSE(m)**	**Avg. Time(s)**	**RMSE(m)**	**Avg. Time(s)**	**RMSE(m)**	**Avg. Time(s)**	**RMSE(m)**	**Avg. Time(s)**
SRP-PHAT	0	25.4711	0	25.4297	5	25.6799	4	25.6088
BM-SRP-PHAT	0	6.5548	0.1	6.8886	1	6.8036	0.7	7.0719

Table 3.9: Single-source scenes with binary masks and SRP-PHAT.

The results show how the computational time consumed by the algorithm is considerably reduced almost 20s and the localisation error is reduced avoiding phantom images in some situations. Moreover, the SRP-PHAT results show that the accumulated approach can introduce large errors with saxophone and violin examples. But, it is to be noted that the error introduced in saxophone and violin and how it is reduced when interferences are discarded by BM. Nevertheless, clarinet seems to be deviated by the BM and bassoon seems to be transparent to this process. The deviation in saxophone and clarinet when mask is not applied can be due to the deviation on the upper harmonics. With 4 harmonics BM, this deviation is removed. Instead,

with clarinet where the harmonic deviation is shorter may be some problem with the MIDI file or the upper harmonic let to accurate the localisation.

- **Results with simultaneous-source scenes and score information**

  The following tables contains the results obtained for duets,trios and quartets on the acoustic scene 1. Each source are labeled by instrumentation (Duet/Target) to see the differences by timbre. $\hat{p}$ gives the estimated position for that target. CLS indicates the localised source, very interested to know which is the most salient source in phase.

Duet/Target	$\hat{p}$	*CLS*	*RMSE* (m)	Avg. Time (s)
Bassoon	[10,18]	Clarinet	0	6.5520
Clarinet	[10,18]	Clarinet	0	6.7184
Bassoon	[4,13]	Saxophone	0	6.7749
Saxophone	[4,13]	Saxophone	0	7.0613
Bassoon	[ ]	None	3.46	6.5702
Violin	[ ]	None	1.87	6.7401
Clarinet	[10,18]	Clarinet	0	6.6827
Saxophone	[10,18]	Clarinet	0	6.6680
Clarinet	[ ]	None	1.8708	6.7880
Violin	[10,18]	Clarinet	0	6.7293
Saxophone	[ ]	None	1.8708	6.7048
Violin	[4,10]	Saxophone	2.7386	6.8300
		Avg.	1.8708	6.7350

Table 3.10: Overall performance for six duets.

In the first duet, BM only works for clarinet, bassoon is never localised as target, but the localisation is quite accurate. By the way, the average of computational time is equal when the number of sources increase, though in this situations the probability to success highly is reduced. The computation time only is affected by the number of microphones, the spatial resolution, the phasor resolution and by reducing the TF representation to analyse.In the second duet, bassoon and tenor saxophone, again one target is always lost. In thins case saxophone seems to have richer response with the approach. Moreover, in the third case, the algorithm returns none correct location, the interference between the two sources confuses to the process and the error increases. With clarinet and saxophone, again only one source is salient when BM is applied.

The clarinet seems to overlap saxophone and bassoon, but saxophone overlaps bassoon. But when clarinet and violin are combined the result is quite rare, because clarinet is note detected as target but it is detected when violin is the target. It may be a error on the MIDI file mapping to pitch. But again clarinet covers violin. In the last pair, saxophone is only located when violin is the target. The last three results are difficult to understands, but it seems like SRP-PHAT have more difficulties to get a well localisation when different timbres are combined. When bassoon, saxophone and violin are involved the process has more probabilities to fail.

Trio/Target	$\hat{p}$	*CLS*	*RMSE* (m)	Avg. Time (s)
Bassoon	[4,13]	Saxophone	0	6.5155
Clarinet	[10,18]	Clarinet	0	6.6121
Saxophone	[10,18]	Clarinet	0	6.6325
Bassoon	[9,13]	Violin	0.7071	6.8375
Clarinet	[12,9]	None	3.1623	6.6150
Violin	[13,21]	None	3.0000	6.6713
Bassoon	[12,9]	None	3.1623	6.5294
Saxophone	[1,1]	None	8.7464	6.6213
Violin	[10,13]	Violin	0	6.7272
Clarinet	[12,9]	None	3.1623	6.6453
Saxophone	[11,18]	None	3.6056	6.7366
Violin	[12,7]	None	4.4721	6.6945
		Avg.	2.5015	6.6532

Table 3.11: Overall performance for four trios.

Quartet/Target	$\hat{p}$	*CLS*	*RMSE* (m)	Avg. Time (s)
Bassoon	[13,9]	None	3.5355	6.5155
Clarinet	[12,9]	None	3.1623	6.6121
Saxophone	[1,1]	None	8.7414	6.6325
Violin	[10,13]	Violin	0	6.6713
		Avg.	3.8611	6.7988

Table 3.12: Overall performance for four trios.

The last table show the results of evaluation for trios and quartet. It shows what we explained before, when more different signals are involved the process doesn't works with an accumulated approach. Phase spectra is quite

sensible to any change and a binary mask may be is a technique so aggressive and a smoothing function can be needed to success. It is curious how the algorithm works quite well with some combinations and not in others. Again, clarinet, violin and saxophone are located in some situation. Instead, bassoon is never detected in simultaneous source scene.

Chapter 4

LOCALIZATION INFORMED SOUND SOURCE SEPARATION

4.0.3 Introduction

SSL provides enough information to define the mixing parameters for each source in each microphone pair. The location of a microphone pair and the angles between sources and sensors can generate a variety of possibilities. However, knowing the microphone location and the source location can afford new fashions in separation tasks. In this master thesis, DUET and NMF techniques are tested to show the possibilities offered by a position informed SSS approach.

4.1 DUET approach

DUET algorithm is well defined in the literature, it can be easily computed through the magnitude spectra or phase features (Section 2.4). It is based on comparing the differences in intensity and in time for each frequency bin, with a mixing parameter estimator (spatial cues or tangent law). Generally, a sound source image is defined in stereophony with two parameters, q_L and q_R. Also, it can be defined by an angle (DOA). This parameters are quite relative for all the frequencies, it can be used to generate a binary mask to define the contribution of each source in the spectrogram given in radians and defined by $X_L(k)$ and $X_R(k)$. We decide to use the mixing parameters estimator introduced by Roebel [5] which is defined in the Equation 2.37. Histogram of mixing parameter distribution can helps to demarcate an estimated source. The point is to defining a region for each source and how to estimate this region. Usually, in the direction where the source is located appears a peak, but the distribution varies depending on: the timbre of the source, the source location regarded to the microphone pair and the mixture model pro-

cessed. So, the mixing parameter estimator can generate binary masks to separate sources.

On the other side, we assume the number of sources and its location are known by the system. So, it can be utilised with the sound velocity and the spacing between microphones to define the DOA of each source. Position information may be quite useful in DUET, especially to cluster source components in the DOA distribution's histogram. We propose to cluster sources with 1D-Gaussian Mixture Model (GMM) using EM to estimate the ML. Initialising the same number of components as sources are involved on the scene, with the mean of gaussian components defined by the estimated mixing parameters, previously computed by distances between sensors and sources. The strategy proposed is tested for duet scenarios (2 sources) with anechoic mixture model, generated in the dataset to test SRP-PHAT with concurrent notes (Experiment 5). Finally, the process is objective evaluated with BSS Eval[1][59], a MATLAB Toolbox to measure the performance of sound SSS algorithms, comparing separated sources with the corresponding microphone captured signal in the same scene for single-source setttings.

4.1.1 DUET implementation

This section explains the strategy designed to face this problem and the corresponding experiments. The experiments have been done assuming the simplest situations (anechoic mixtures). In this approach, the parameters of the acoustic scene (Table 3.2) provide position information, allowing to know the mixing parameters for each source at each microphone pair (prior knowledge).

Position Information - DOA Estimation

First at all, a list with the microphone identifier at each pair is filled. The order is vital to avoid issues with the reference system when mixing parameters are computed. This list is quite useful in the evaluation step to define the correct ground truth signal. As I said, acoustic scene parameters are very important in this step. We assume the system knows the scene layout. So, as DUET is computed by microphone pair, first at all, distances between sources and both microphones are computed. Subtracting them, TDOA is estimated for that source at each microphone pair. The time difference can be positive or negative, it depends in the system reference created by the microphone location. So, it is important to keep an order to sort microphone pairs by ascendent x axis location. In this way, the microphone situated in the left always will be the left channel and we will avoid flips in the stereo image. Then, microphone spacing defines the DOA, mapping to

[1] `http://bass-db.gforge.inria.fr/bss_eval/`

$90°$ range for tangent law and $180°$ for spatial cues. Depending on the estimator applied, the mixing parameters can be negative, which it is interested to know if signals are in anti-phase, correlated or uncorrelated. Later, estimated mixing parameters will be used to initialise the GMM algorithm in the clustering procedure.

Other main point is the directivity pattern of microphones, omnidirectional patterns capture around its location, based on sound pressure variations. So, the sources located behind the microphone, introduce a lot of leakage in the spot microphones and phantom image can appears because for this sensors positive or negative polarity are trivial. Instead, the cardioid microphones have a thinner beam of capture and so, leakage is minimised. Besides, they function with pressure gradient response.

Time Frequency Analysis

Next, TFA is computed at each microphone pair to estimate the DOA distribution in the frequency domain through magnitude spectra. Therefore, STFT is applied with a 4092-*Sine* window function, with hop size of half window.

DUET histogram

Every microphone pair can be assumed as stereo signal, so, an simple *Mid-Side* signal analysis can help to decide what stereo signals should be separated by the process. Then, MS signals are computed and the Root Mean Squared (RMS) is computed for each one of them. Through a MS ratio between the energy in S and the energy in M and a threshold defined at 0.75, close stereo images are discard and open ones are accepted to processing, which have more probabilities to achieve decent outcomes. If the stereo pair passes the MS evaluation, then DOA is estimated by Equation 2.37 and the histogram can build.

Clustering DOA regions

The distribution of DOA is quite irregular and to define regions manually can be easy. But DOA can be clustered automatically by GMM. It is applied with a constrained initialisation, locating the gaussian components where the position information has been determined in the first step. If GMM is not initialised with this data, the results are random and different at each trial. The number of sources involved in the scene defines the number of gaussian components. Making use of *Mathematical Empirical Rules* (MER) [61], regions can be determined by the next interval, $[mu_{g_s} - std_{g_s}, mu_{g_s} + std_{g_s}]$ (Figure 4.2), where *mu* is the mean of the data distribution and *std* is the *standard deviation* error estimated for each

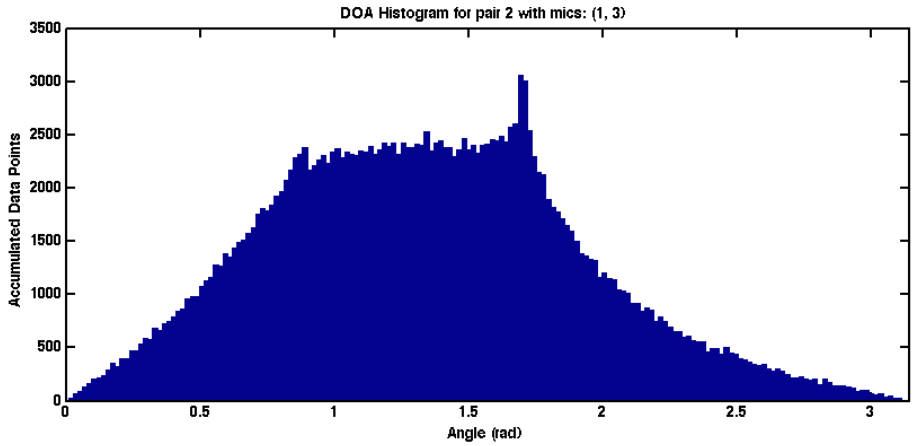

Figure 4.1: Histogram for estimated DOA in microphone pair 3 (SC1).

gaussian by GMM. This interval contains the 67% of a normal distribution. GMM implementation has a input argument to define convergence as stop condition. During the experiments convergence was fixed to 0.01. Assuming that one part of the signal will be lost due to the region definition, sources may be well separated. Now, there are two possibilities, overlapped or non-overlapped regions. Non-overlapped regions are the best situation to obtain good results, but it depends on the scene layout. Instead, overlapped regions are more critical, in terms to define a DOA region in the next step. In those cases, we have defined a criteria to choose the best region candidate.

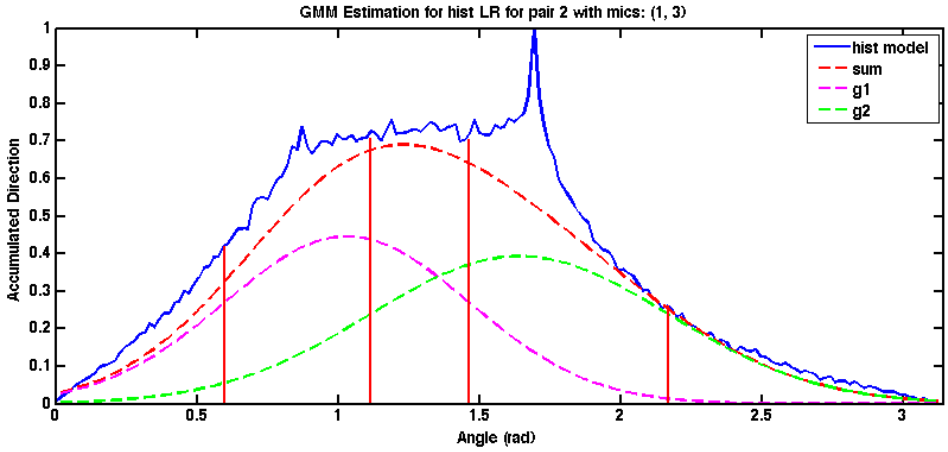

Figure 4.2: Histogram for estimated DOA in microphone pair 2 (SC1).

Separation criteria for overlapped gaussian components

The criteria applied is simple, in Figure 4.2, an example for overlapped regions is shown (pair 2). Regions are defined by the vertical red lines and they define the 67% of the normal distribution at each gaussian component. So, if regions are overlapped, the gaussian component with higher weight and less *std* is chosen as a possible candidate. So, it will be easy to split regions and the separation process should be more efficient. Automatically, the other gaussian component would be discarded as a bad candidate. This criteria makes that the channel selected by the system corresponds with the closer microphone to the target source. On the other hand, it was chosen for a duets examples, as greater the number of sources, more constrained should be the criteria to define the regions in DOA domain.

Defining Binary Masks

Now, DOA regions are defined for each source, so we need to create BM to separate each source contribution in terms of DOA. Therefore, logical statements can define BM, with ones, for the frequency bins with DOA inside this region, and zeros ,if DOA for the frequency bins are out this region. Obviously, BM are quite sharp with overlapped components but we assume it introduce few artefacts in the separated signals. Next, BM mask is applied as weighted function over the TF representation of target signal and the *Inverse Short Time Fourier Transform* (*i*STFT) is computed to obtain the separated signal.

4.1.2 Experimental Setup

Dataset

In these experiments, the dataset is the same than in the Section 3.2.1. The acoustic scene design 1 is used with a virtual anechoic mixture. In the scene, there are 4 omnidirectional microphones and the sources are distributed shaping a triangle (Figure 3.5). The distribution in the scene is not so realistic for a music live performance but we start with simplest situations (duets and trios). The separated audio files will be evaluated with BSS Eval.

Evaluation Metrics

BSS Eval use different metrics to evaluate the separation process. We will take attention to SDR, SIR, SAR, and ISR. How they are computed are very well explained and defined in the following documents [62, 63, 59].

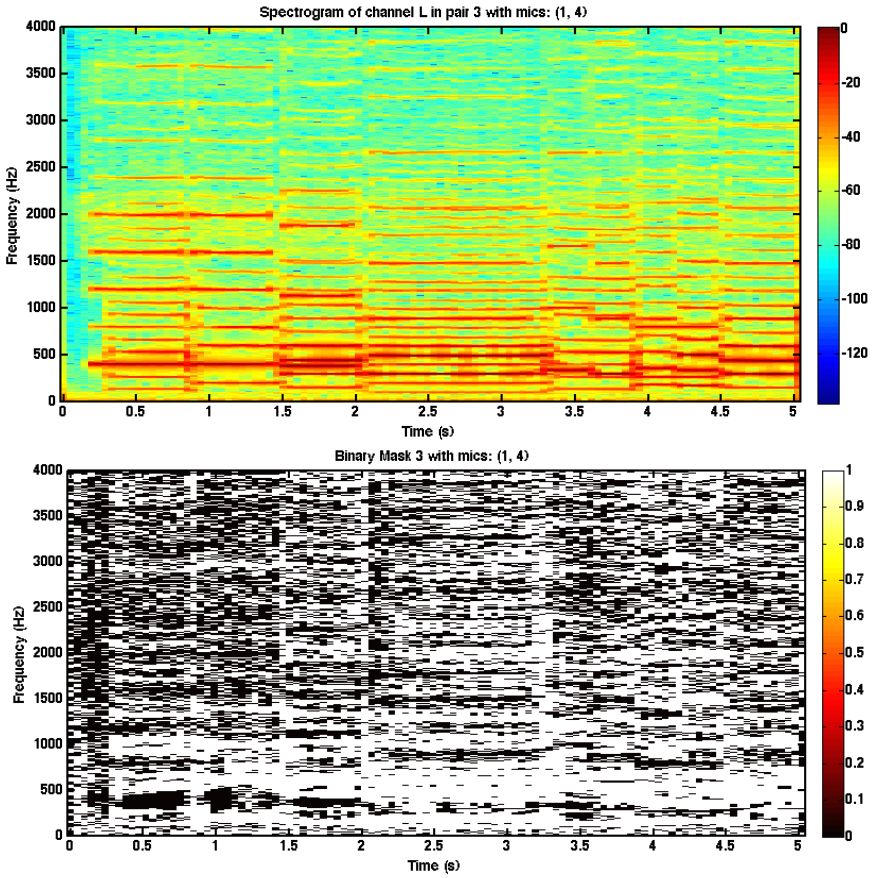

Figure 4.3: Spectrogram and binary mask for bassoon in stereo pair 2 (SC1).

4.1.3 Evaluation

Experiment 1 - Anechoic Mixtures

In this experiment, we compare the score evaluation for Position Informed DUET and Blind DUET. The position information is used to initialise the GMM. In the blind model gaussian components are initialised randomly.

- Position Informed DUET Approach

1. Duet approach

   The process discards the pair 1 due to most of the signal is concentrated in the Mid signal. Both sources has the same TDOA, so to design a mask is quite difficult in these cases. In the other pairs the process works quite well.

Obviously, it functions better with clarinet in terms SDR, where the separation achieves a very good target source. Instead the results shows how bassoon has a lower score. It can be due to frequency resolution, it should be tested in related future. Also, I would like to mention that the process automatically takes the target signals for different pairs but the selected microphone are always the same (m_3 and m_4).

pair	mics	MS rate	sel. mic	Target	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)
1	None	0	None	Discarded	0	0	0	0
2	m_1, m_3	1.6149	m_3	Bassoon	11.1057	16.4229	16.6605	14.1291
3	m_1, m_4	1.8457	m_4	Clarinet	16.2107	29.5946	23.1866	17.4909
4	m_2, m_3	1.6149	m_3	Bassoon	11.1057	16.4229	16.6605	14.1291
5	m_2, m_4	1.8457	m_4	Clarinet	16.2107	29.5946	23.1866	17.4909
6	m_3, m_4	0.8505	m_3	Bassoon	10.2590	31.3990	30.7460	10.4255
	m_3, m_4	0.8505	m_4	Clarinet	7.7138	21.6279	22.6956	8.5756

Table 4.1: Overall performance for duets applying the position informed DUET approach.

- Blind DUET Approach

1. Duet approach

   When GMM are initialised randomly, the components don't fit very well with the histogram, because the sharp shape of function confuse the pattern recognition technique. The microphone chosen are again the same because it is done before by the MS evaluation of stereo signal. Moreover, the mask is not appropriate for a good separation. Bassoon is the only case where the approach works at the same way with 11dB of SDR. Also, the results in the pair 6 are kept. hose table show how DOA can improves DUET processes when the mixture of signals is more complex and DOAs are overlapped.

2. Trio approach

   When more than 2 sources are in the scenario, the approach proposed doesn't works properly. The constraints introduced as MER to define DOA region should be changed and more strictly defined. The space is the same but DOA distribution is more complex when the number of sources increases. At the future work, weighted histogram by magnitude should be applied

pair	mics	MS rate	sel. mic	Target	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)
1	None	0	None	Discarded	0	0	0	0
2	m_1, m_3	1.6149	m_3	Bassoon	5.4029	7.1255	9.5400	15.9713
3	m_1, m_4	1.8457	m_4	Clarinet	8.8485	16.7607	11.3154	12.8727
4	m_2, m_3	1.6149	m_3	Bassoon	11.0888	16.1316	14.7024	15.9443
5	m_2, m_4	1.8457	m_4	Clarinet	9.5956	17.8265	12.6631	12.8727
6	m_3, m_4	0.8505	m_3	Bassoon	10.0860	34.6925	28.0381	10.2697
	m_3, m_4	0.8505	m_4	Clarinet	7.4451	21.7617	28.5875	8.1466

Table 4.2: Overall performance for duets applying the blind DUET approach.

and evaluated. On the other hand, we are using omnidirectional microphones and the instrumentation and microphone distribution is quite weird. So, when other instruments are introduced, the superposition of gaussian components in DOA domain is higher and new strategies should be introduced.

4.2 NMF Approach

This trending method for sound source separation is tested making use of FASST v1 [58, 60] and the possibility to introduce mixing parameters fixed or estimated by the own system (blind). FASST offers different separation models: NMF, GMM, *Hidden Markov Model* (HMM), *Gaussian Scaled Mixture Model* (GSMM) and *Scaled Hidden Markov Model* (SHMM). It is based on the maximisation of likelihood over the covariance matrix of the stereo mixture. The toolbox provides a few scripts as example, which are quite useful to design experiments with position informed as prior knowledge.

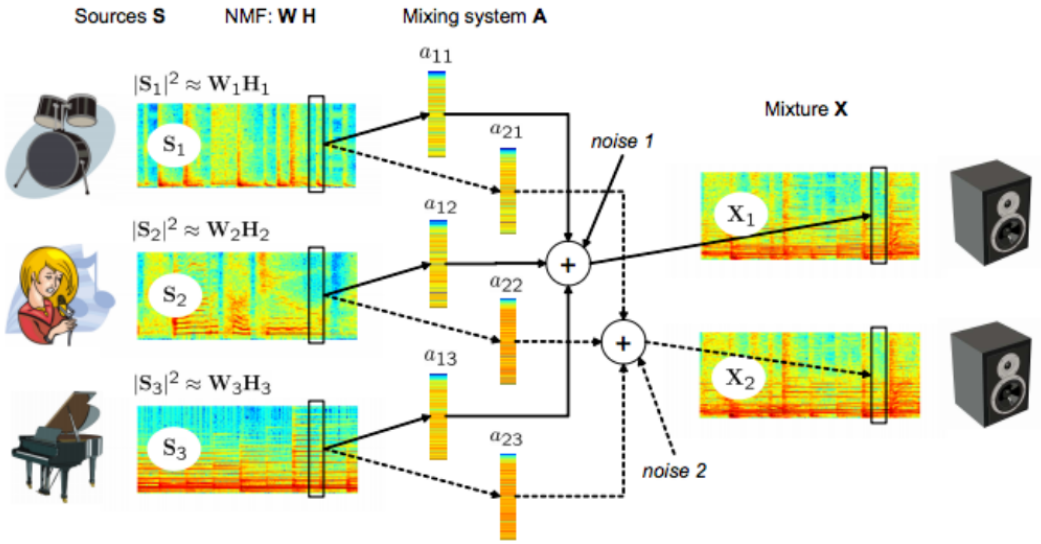

Figure 4.4: Multichannel NMF problem [60].

4.2.1 NMF Implementation

So, there is a MATLAB script in the toolbox[2], it is designed to separate anechoic mixtures or virtual anechoic mixtures. Therefore, it can be used to separate signals in each microphone pair of the acoustic scenes designed for the SSL experiments.

Again, the DOA estimator is used to initialise the mixing parameters, so we use the same method to estimate the mixing parameters than we have presented for the DUET approach. But, there is a problem with the mixing parameters, TL and NMF. When signals are uncorrelated in the stereo image, or which is the

[1] `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.330.2508&rep=rep1&type=pdf`

same the DOA angle in the TL domain is higher than $90°$, in those situations, the mixing parameters are negative and maybe NMF model is not able to handle a proper estimation due to the positivity constraint.

In the system, first at all, some initial parameters are defined: directories to load and write audio data, the name of audio file, TF parameters, the number of sources and the number of NMF components. Later, an stereo audio file is loaded to compute the TF representation of signals with the stereo covariance matrix. The model works with a defined structure by the toolbox makers. It is generated to initialise the estimation. In this point, mixing parameters are initialised randomly. So, later you can redefine the mixing parameters if you have this information. Next, the parameters for SSS are estimated and spatial source images can be saved in a new wave file.

4.2.2 Experimental Setup

The acoustic scenes designed for SSL experiments is again our dataset (SC1 and SC2). The stereo images formed by the microphone pairs present at the scene can be processed as a stereo image. But in this case, the DOA's angle can be quite large and it provokes phase problems with mixing parameters estimated with the TL. In this occasion, the virtual anechoic mixtures are generated with two sources to test how the system works. On the other hand, SC2 has set stereo pair recordings applying coincident and quasi coincident techniques as XY, ORTF and NOS. They should contain a big proportion of correlated signals if we compare with AB or the microphone pairs that we use as spot microphones.

So, the position information in SSS is evaluated against the random initialisation (blind approach). FASST lets to fix the mixing parameters or being estimated automatically by the system. The performance comparison can be measured in terms of computational time and with objective metrics with BSS_Eval MATLAB toolbox. So, two NMF models are tested: piNMF (position informed NMF) and rNMF (NMF with mixing parameters estimated by the model). With piNMF, mixing parameters are previously estimated by means of TDOA, computing the DOA with the microphone and source location; and they are fixed by the FASST parameter, mix_str.spat_comps.frdm_prior. On the other hand, in rNMF the mixing parameters are estimated by the model. In the experiments, 200 iterations are defined to ensure a good separation. In the process, a 2048 samples window size is applied with half-overlapped window (1024 samples). The number of sources should be defined depending on the sources contained in the mixture. 4 components are used on the NMF model.

[2]EXAMPLE_ssep_Mult_NMF_inst.m

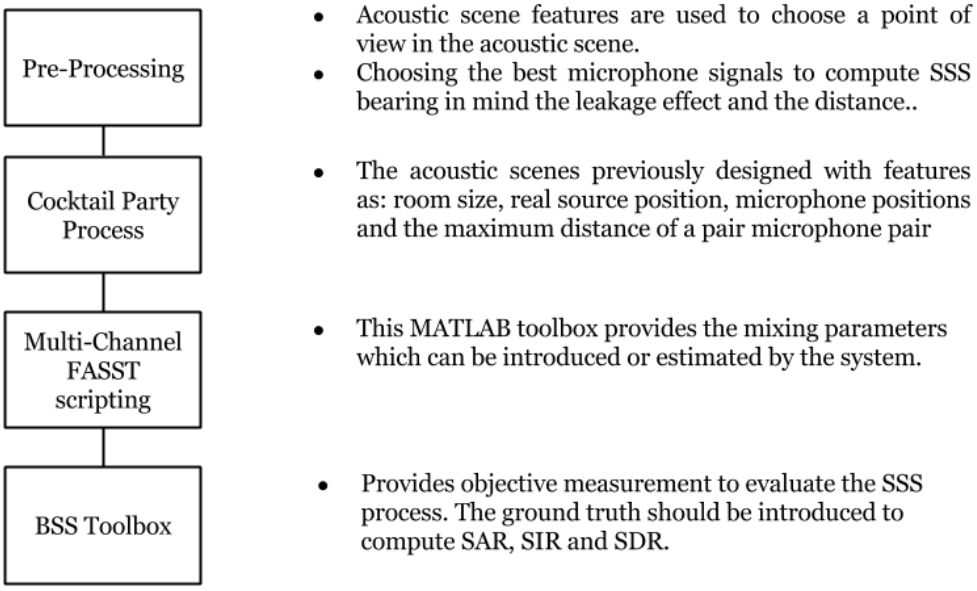

Figure 4.5: System overview for NMF evaluation.

4.2.3 Evaluation

Experiment 1- Virtual anechoic mixtures

In this case, a virtual anechoic mixture was created with the original dataset (10-Bach Chorales). Clarinet and bassoon stereo image were located at $45°$ and $-45°$, which are defined by the following mixing parameters, $[10]$ and $[01]$. The stereophonic mixture was created by MATLAB scripting.

model	mix params	source	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)	Avg. Time
piNMF	Fixed	clarinet	21.4373	21.5222	41.4993	40.7679	1035
	Fixed	bassoon	18.6703	18.7651	37.1188	38.2680	1035
rNMF	Random	clarinet	21.4514	21.5366	41.4966	40.7802	989.5830
	Random	bassoon	18.6689	18.7634	37.1110	38.3015	989.5830

Table 4.3: Overall performance for NMF models with virtual anechoic mixture.

With this NMF model implementation the computational time doesn't change when mixing parameters are fixed or free. Instead it is a vital parameters to get an improved separation when this parameter is fixed and previously estimated with the position information. It seems that FASST latex more when the mixing parameters are fixed (30s).

However, when the estimated mixing parameters are well defined and fixed, the results don't show a notable improvement, if it is compared with rNMF. By contrast, clarinet obtains a better results in terms of SDR than bassoon. Probably, if the frequency resolution is increased it would improve, due to bassoon has a lower *f0*.

Experiment 2 - Anechoic mixtures with a microphone pair (SC1)

The last experiment is repeated for the microphone pairs on SC1. The results for all pairs are quite meaningless in terms of separation. However we can determine that when the mixing parameters are fixed the model don't reduce the computational time, instead it is always increased. However, the objective metrics are better when piNMF is applied. It seems that the delay time between sources and microphones makes that these stereo pairs were considered as a convolutive mixture when this code is for anechoic mixtures.

model	mix params	source	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)	Avg. Time
piNMF	Fixed	bassoon	0.6134	3.3652	1.1952	11.0545	196.6583
	Fixed	clarinet	0.8058	5.6609	1.7993	13.3925	196.6583
rNMF	Random	bassoon	0.0366	0.0628	6.2658	−2.4928	194.7765
	Random	clarinet	−0.9717	19.4802	−0.4361	26.0294	194.7765

Table 4.4: Overall performance for NMF models with anechoic mixture (pair1).

model	mix params	source	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)	Avg. Time
piNMF	Fixed	bassoon	−0.4788	−0.1426	9.0642	62.1552	229.2802
	Fixed	clarinet	−3.2707	8.0323	−2.0564	56.6767	229.2802
rNMF	Random	bassoon	−1.8304	−0.3945	−3.4070	10.7028	219.1634
	Random	clarinet	−4.3702	−1.5263	−0.1251	6.3105	219.1634

Table 4.5: Overall performance for NMF models with anechoic mixture (pair2).

Experiment 3 - Anechoic mixtures recorded by XY stereo pair technique (SC2)

Finally, we test the one microphone stereo technique, well-known to provides a correlated stereo signal, where ITD are very short. It is tested with 2 sources and 4 sources at the SC2, and the stereo pair is located at point p_1 defined in Table 3.3. This table shows that the separation when coincident microphone technique are applied in the scene separation can be achieved with a minimum of quality.

model	mix params	est. mix params	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)	Avg. Time
piNMF	Fixed	bassoon	3.8376	16.1335	14.1706	54.5408	994.9887
	Fixed	clarinet	5.2285	5.4381	3.4287	60.1900	994.9887
rNMF	Random	bassoon	12.2993	21.6625	12.8718	47.6658	1016.4
	Random	clarinet	13.6914	14.1810	23.3126	48.1116	1016.4

Table 4.6: Overall performance for NMF models with anechoic XY mixture (2sources).

However, the best score is for rNMF, so it means that the mixing parameters introduced don't works well. It can be provoked because the mixing parameters are estimated in other scale different to which we are using. FASST uses tangent law but when the mixing parameters are automatically estimated, they are normalised to 2, instead to 1. In the experiment with virtual anechoic mixtures, the mixture tested was generated with the tangent law to estimate the gain factors. and in that case, the gain factors estimated by the system coincides. At future work, more time should be dedicated to prove this point.

model	mix params	est. mix params	SDR (dB)	ISR (dB)	SIR (dB)	SAR (dB)	Avg. Time
piNMF	Fixed	clarinet	0.4287	0.4624	5.5660	6.5495	2482.3
	Fixed	saxophone	−2.1970	8.8933	−4.4698	17.8684	2482.3
	Fixed	bassoon	−3.2646	7.6890	−6.5404	16.1183	2482.3
	Fixed	violin	0.3054	−0.1426	5.6051	7.0913	2482.3
rNMF	Random	clarinet	9.2803	13.1659	12.4424	16.1242	2570.3
	Random	saxophone	2.1868	12.8184	2.4840	13.1320	2570.3
	Random	bassoon	−3.3009	4.2692	−9.3668	15.7817	2570.3
	Random	violin	0.1543	0.2414	10.7549	8.3273	2570.3

Table 4.7: Overall performance for NMF models with anechoic XY mixture (4 sources).

Finally, when more sources are present more computation time is consumed by the algorithm to estimate the sources. However the results are quite bad, the number of iterations is not enough to let the model to find the best results. In the future, more testings should be done in this line to determine more firm conclusions.

Chapter 5

CONCLUSION AND FUTURE WORK

Along this work, there have appeared a lot of parameters, which are quite messy, but reality or realistic approaches involves these difficulties, any variation can affects the process. So, when more strict the experiments are more relevant they may be the contribution. For me, it was the main conclusion. But if we are concentrated about the conclusion of this work I may point the most relevance in a list:

1. In SRP-PHAT, the TF representation needs at least a 8192 samples-FFT and a hop size defined by N/8, where N is the window size. Also, the sort of window function chosen affects the localisation. For anechoic models (s_1), all the window functions can works without introduce errors. But when RT_{60} increases, the *CLF* is reduced and the localisation error increases.

2. It is quite easy to generate *phantom images* of sources, which will be translated to uncorrelated signals in the stereo projection of each microphone pair. It lets increase the number of detected sources because omnidirectional pattern has a wide beam of capture and in this way more phase information about the acoustic scene.

3. When simultaneous sources are playing at the same time, like quartets, the localisation is pretty difficult due to the interference of sound waves. However, when RT_{60} is fixed to 0.25s, there are two special cases which offer better results than the other window functions. These are *Blackman* and *Bartlett* window function.

4. SRP-PHAT with orchestral music signals works well in anechoic spaces and with reverberation. However, when simultaneous sources are playing,

it becomes a great mesh in the phase data which is very distorted by the contribution of secondary lobes of window function in the TFA.

5. SRP-PHAT works different with different timbres. For instance with violin works very well however for bassoon there are more obstacles to achieve a good location estimate. It may be due to the spectral distribution of bassoon filter model, where the third harmonic contains the most of energy. Also, the bandwidth of frequency bins in low frequencies introduce deviations. Besides, tenor saxophone always introduce a short deviation. So, harmonic approach can works with some instruments but it should be improved with more knowledge or instrument models to weight in different way each harmonic contribution.

6. The cardioid directivity pattern reduces the error and increases the efficiency when SNR is reduced as far as 10dB, as it is shown by the comparison between the pink and the green line (0.25s). On the other hand, with 5dB-SNR levels which is ratio quite unlikely in a realistic environment, cardioid directivity pattern are not the best scored by the metrics.

7. The algorithm is not robust to these deviations and it could be easily introduced in a realistic experiment. We should think in new procedures to compensate this errors or to design a fixed microphone array installation as it is proposed in [28].

8. Each microphone recording technique provide different and complex SPR-PHAT shapes and they can be combined to improve the source localization process with music signals. Combining close-spot miking and stereo pair techniques, the localisation error can be drastically reduced.

9. The results show how the computational time consumed by the algorithm is considerably reduced almost 20s and the localisation error is reduced avoiding phantom images in some situations. Moreover, the SRP-PHAT results show that the accumulated approach can introduce large errors with saxophone and violin examples.

10. Harmonic BM approach only works for clarinet, bassoon is never localised as target, although the localisation is quite accurate. By the way, the average of computational time is equal when the number of sources increase, though in this situations the probability to success highly is reduced. The

computation time only is affected by the number of microphones, the spatial resolution, the phasor resolution and by reducing the TF representation to analyse.

11. When more different signals are involved the process doesn't works with an accumulated approach. Phase spectra is quite sensible to any change and a binary mask may be is a technique so aggressive and a smoothing function can be needed to success.

12. DUET approach with virtual anechoic mixtures functions better with clarinet in terms SDR, where the separation achieves a very good target source. Instead the results shows how bassoon has a lower score. On the other hand, when more than 2 sources are in the scenario, the approach proposed doesn't works properly. The constraints introduced as MER to define DOA region should be changed and more strictly defined. The space is the same but DOA distribution is more complex when the number of sources increases.

13. NMF model from FASST v1 works well with virtual anechoic mixtures and anechoic mixtures when the mixing parameters estimated by the DOA angle are positive. With negative values the model doesn't provide a well separated signals.

14. With the NMF model used, the computational time doesn't change when mixing parameters are fixed or free. Instead it is a vital parameters to get an improved separation when this parameter is fixed and previously estimated with the position information. It seems that FASST latex more when the mixing parameters are fixed (30s). However, when the estimated mixing parameters are well defined and fixed, the results don't show a notable improvement, if it is compared with rNMF. By contrast, clarinet obtains a better results in terms of SDR than bassoon. Probably, if the frequency resolution is increased it would improve, due to bassoon has a lower *f0*. However we can determine that when the mixing parameters are fixed the model don't reduce the computational time, instead it is always increased. However, the objective metrics are better when piNMF is applied. It seems that the delay time between sources and microphones makes that these stereo pairs were considered as a convolutive mixture.

15. However, when correlated signals are contained in the microphone pair, as when coincident stereo pair techniques are applied, the system is able to separate signal components. This determines that for SSL uncorrelated signals are essential for SSL methods instead for a well performance of FASST v1 correlated stereo signals are vital.The separation when coincident microphone technique are applied in the scene separation can be achieved with a

minimum of quality. However, the best score is for rNMF, so it means that the mixing parameters introduced don?t works well. It can be provoked because the mixing parameters are estimated in other scale different to which we are using.

16. Finally, when more sources are present more computation time is consumed by the sound separation algorithm to estimate the sources. In the future, more testings should be done in this line to determine more firm conclusions.

While this work have advanced more possibilities it offers to define interested future works. As it is mentioned in the Section 2.3, a number of techniques have been proposed, but nobody has proposed to use GA with SRP-PHAT. So, there is the possibility to evaluate the GCC-PHAT during the computation applying GA. So far, no method are applied to reduce this problem but it is a possible future work to make. It is evaluated the whole grid space and the global maxima defines the estimated source position.

Other research line may be to define new experiments with musical signals and to learn in depth own the different timbres affects the model. By contrast, to generate a large dataset of scenes with larger RT_60 to evaluate the algorithm at music hall applications.

On the concurrent notes approach with SRP-PHAT, weighted masks should be tested. On the other hand, the piNMF may be tested with different microphone techniques and combinations of microphones. Personally, I think there is a lot to explore in this field and also to define a framework or toolbox to apply piNTF. However, the most ambitious future work would be to repeat the same experiments in a realistic approach and with a fixed microphone array.

Most of MATLAB code and audio files tested are available in my GitHub profile: `https://github.com/xaviliz`

Bibliography

[1] Cobos, M. (2009). *Application of Sound Source Separation Methods to Advanced Spatial Audio Systems*. Universidad Politécnica de Valencia.

[2] Marti, A. (2013). *Multichannel Audio Processing for Speaker Localization, Separation and Enhancement*. Universidad Politécnica de Valencia.

[3] Marti, A., Cobos, M., Lopez, J. J., & Escolano, J. (2013). *A steered response power iterative method for high-accuracy acoustic source localization*. The Journal of the Acoustical Society of America, 134(4), 2627-2630.

[4] Do, H., & Silverman, H. F. (2011, October). *A robust sound-source separation algorithm for an adverse environment that combines MVDR-PHAT with the CASA framework*. In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on (pp. 273-276). IEEE.

[5] Mitsufuji, Y., & Roebel, A. (2013, May). *Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior knowledge*. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 71-75). IEEE.

[6] Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition (Vol. 14)*. Englewood Cliffs: PTR Prentice Hall.

[7] Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*.

[8] Schedl, M., Gómez, E., & Urbano, J. (2014). *Music Information Retrieval: Recent Developments and Applications*. Foundations and Trends in Information Retrieval, 8(2-3), 127-261.

[9] Brown, G. J., & Cooke, M. (1994). *Computational auditory scene analysis.* Computer Speech & Language, 8(4), 297-336.

[10] Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT press.

[11] Grothe, B., Pecka, M., & McAlpine, D. (2010). *Mechanisms of sound localization in mammals*. Physiological Reviews, 90(3), 983-1012.

[12] Moore, D. R. (1991). *Anatomy and physiology of binaural hearing*. International Journal of Audiology, 30(3), 125-134.

[13] Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

[14] Jeffress, LA. (1948). *A place theory of sound localization*. J Comp Physiol Psychol. 41:35-39.

[15] Faller, C., & Merimaa, J. (2004). *Source localization in complex listening situations: Selection of binaural cues based on interaural coherence*. The Journal of the Acoustical Society of America, 116(5), 3075-3089.

[16] Serra, X. (1989). *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.*.

[17] Pourmohammad, A., & Ahadi, S. M. (2011, July). *TDE-ILD-HRTF-Based 3D entire-space sound source localization using only three microphones and source counting.* In Electrical Engineering and Informatics (ICEEI), 2011 International Conference on (pp. 1-6). IEEE.

[18] 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, 19(6), 1490-1503.

[19] Khaddour, H.. (2011). *A comparison of algorithms of sound source localization based on time delay estimation*. Electrotechnic magazine: elektrorevue, 2, 31-37.

[20] Knapp, C., & Carter, G. C. (1976). *The generalized correlation method for estimation of time delay*. Acoustics, Speech and Signal Processing, IEEE Transactions on, 24(4), 320-327.

[21] Azaria, M., & Hertz, D. (1984). *Time delay estimation by generalized cross correlation methods*. Acoustics, Speech and Signal Processing, IEEE Transactions on, 32(2), 280-285.

[22] Zhang, C., Zhang, Z., & Florêncio, D. (2007, April). *Maximum likelihood sound source localization for multiple directional microphones*. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 1, pp. I-125). IEEE.

[23] Zhang, C., Florêncio, D., & Zhang, Z. (2008, March). *Why does PHAT work well in lownoise, reverberative environments?*. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (pp. 2565-2568). IEEE.

[24] Van Veen, B. D., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). *Localization of brain electrical activity via linearly constrained minimum variance spatial filtering*. Biomedical Engineering, IEEE Transactions on, 44(9), 867-880.

[25] Krolik, J., & Eizenman, E. (1988, April). *Minimum variance spectral estimation for broadband source location using steered covariance matrices*. In Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on (pp. 2841-2844). IEEE.

[26] DiBiase, J. H. (2000). *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. (Doctoral dissertation, Brown University).

[27] Kepesi, M., Ottowitz, L., & Habib, T. (2008, May). *Joint position-pitch estimation for multiple speaker scenarios*. In Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008 (pp. 85-88). IEEE.

[28] Do, H., Silverman, H. F., & Yu, Y. (2007, April). *A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array*. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on (Vol. 1, pp. I-121). IEEE.

[29] Ward, D. B., Lehmann, E. A., & Williamson, R. C. (2003). *Particle filtering algorithms for tracking an acoustic source in a reverberant environment*. Speech and Audio Processing, IEEE Transactions on, 11(6), 826-836.

[30] Zotkin, D. N., & Duraiswami, R. (2004). *Accelerated speech source localization via a hierarchical search of steered response power*. Speech and Audio Processing, IEEE Transactions on, 12(5), 499-508.

[31] Roth, P. R. (1971). *Effective measurements using digital signal analysis*. Spectrum, IEEE, 8(4), 62-70.

[32] Carter, G. C., Nuttall, A. H., & Cable, P. (1973). *The smoothed coherence transform*. Proceedings of the IEEE, 61(10), 1497-1498.

[33] Darzi, S., Sieh Kiong, T., Tariqul Islam, M., Ismail, M., Kibria, S., & Salem, B. (2014). *Null Steering of Adaptive Beamforming Using Linear Constraint Minimum Variance Assisted by Particle Swarm Optimization, Dynamic Mutated Artificial Immune System, and Gravitational Search Algorithm*. The Scientific World Journal, 2014.

[34] Sieh Kiong, T., Ismail, M., & Hassan, A. (2006). *WCDMA forward link capacity improvement by using adaptive antenna with genetic algorithm assisted MDPC beamforming technique*. Journal of Applied Sciences, 6, 1766-1773.

[35] Karaboga, D., & Basturk, B. (2007). *A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm*. Journal of global optimization, 39(3), 459-471.

[36] Shah-Hosseini, H. (2012). *An approach to continuous optimization by the intelligent water drops algorithm*. Procedia-Social and Behavioral Sciences, 32, 224-229.

[37] Rashedi, E., Nezamabadi-Pour, H., & Saryazdi, S. (2009). *GSA: a gravitational search algorithm*. Information sciences, 179(13), 2232-2248.

[38] Blackwell, T., & Branke, J. (2004). *Multi-swarm optimization in dynamic environments*. In Applications of evolutionary computing (pp. 489-500). Springer Berlin Heidelberg.

[39] Hérault, J., Jutten, C., & Ans, B. (1985). *Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé*. In 10° Colloque sur le traitement du signal et des images, FRA, 1985. GRETSI, Groupe d?Etudes du Traitement du Signal et des Images.

[40] Comon, P., & Jutten, C. (2010). *Handbook of blind source separation*. Elsevier, 1, 35-48.

[41] Carabias-Orti, J. J., Cobos, M., Vera-Candeas, P., & Rodríguez-Serrano, F. J. (2013). *Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings*. EURASIP Journal on Advances in Signal Processing, 2013(1), 1-16.

[42] Jourjine, A., Rickard, S., & Yilmaz, O. (2000). *Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures*. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on (Vol. 5, pp. 2985-2988). IEEE.

[43] Yilmaz, O., & Rickard, S. (2004). *Blind separation of speech mixtures via time-frequency masking*. Signal Processing, IEEE transactions on, 52(7), 1830-1847.

[44] Lee, D. D., & Seung, H. S. (2001). *Algorithms for non-negative matrix factorization*. In Advances in neural information processing systems (pp. 556-562).

[45] Avendano, C., & Jot, J. M. (2002, May). *Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix*. In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on (Vol. 2, pp. II-1957). IEEE.

[46] Avendano, C. (2003, October). *Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications*. In Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on. (pp. 55-58). IEEE.

[47] Cobos, M., & López, J. J. (2008). *Stereo audio source separation based on time-frequency masking and multilevel thresholding*. Digital Signal Processing, 18(6), 960-976.

[48] Itakura, F., & Saito, S. (1968, August). *Analysis synthesis telephony based on the maximum likelihood method*. In Proceedings of the 6th International Congress on Acoustics (Vol. 17, pp. C17-C20). pp. C17?C20.

[49] Campbell, D. R., Palomaki, K. J., & Brown, G. (2005). *A MATLAB simulation of "shoebox" room acoustics for use in research and teaching*. Computing and Information Systems, 9(3), 48.

[50] Do, H., & Silverman, H. F. (2010, March). *SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data*. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (pp. 125-128). IEEE.

[51] Do, H., & Silverman, H. F. (2008, March). *A method for locating multiple sources from a frame of a large-aperture microphone array data without tracking*. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (pp. 301-304). IEEE.

[52] Duan, Z., & Pardo, B. (2011). *Soundprism: An online system for score-informed source separation of music audio*. Selected Topics in Signal Processing, IEEE Journal of, 5(6), 1205-1215.

[53] Fritsch, J., & Plumbley, M. D. (2013, May). *Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis*. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 888-891). IEEE.

[54] Oppenheim, A. V., Willsky, A. S., & Nawab, S. H. (1998). *Signals and Systems*. Pearson Educación.

[55] Fletcher, N. H., & Rossing, T. (2012). *The physics of musical instruments*. Springer Science & Business Media.

[56] Howard, D. M., & Angus, J. (2009). *Acoustics and psychoacoustics*. Taylor & Francis.

[57] Rui, Y., Florencio, D., Lam, W., & Su, J. (2005, March). *Sound source localization for circular arrays of directional microphones*. In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on (Vol. 3, pp. iii-93). IEEE.

[58] Ozerov, A., & Févotte, C. (2010). *Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation*. Audio, Speech, and Language Processing, IEEE Transactions on, 18(3), 550-563.

[59] Vincent, E., Sawada, H., Bofill, P., Makino, S., & Rosca, J. P. (2007). *First stereo audio source separation evaluation campaign: data, algorithms and results*. In Independent Component Analysis and Signal Separation (pp. 552-559). Springer Berlin Heidelberg.

[60] Ozerov, A., Vincent, E., & Bimbot, F. (2012). *A general flexible framework for the handling of prior information in audio source separation*. Audio, Speech, and Language Processing, IEEE Transactions on, 20(4), 1118-1133.

[61] Stewart, J. Q. (1947). *Empirical mathematical rules concerning the distribution and equilibrium of population*. Geographical Review, 461-485.

[62] Vincent, E., Gribonval, R., & Févotte, C. (2006). *Performance measurement in blind audio source separation*. Audio, Speech, and Language Processing, IEEE Transactions on, 14(4), 1462-1469.

[63] Févotte, C., Gribonval, R., & Vincent, E. (2005). *BSS EVAL toolbox user guide–Revision 2.0*.

[64] Everest, F. A., & Pohlmann, K. C. (2001). *The master handbook of acoustics (Vol. 4)*. New York: McGraw-Hill.

[65] Bennett, J. C., Barker, K., & Edeko, F. O. (1985). *A new approach to the assessment of stereophonic sound system performance*. Journal of the Audio Engineering Society, 33(5), 314-321.

[66] Eargle, J. (2012). *The Microphone Book: From mono to stereo to surround-a guide to microphone design and application*. CRC Press.