UNIVERSITAT POMPEU FABRA

PhD Research Proposal

Immersive Audiovisual Production Enhancement based on 3D Audio

by

Andrés Pérez-López

supervised by
Dr. Emilia Gómez
Dr. Adán Garriga

Music Technology Group
Department of Information and Communication Technologies

September 2017

Abstract

Blind Source Separation of convolutive mixtures is a well-know problem. A common approach consists of using microphone array for exploting spatial information. Ambisonics microphones are a special case of spherical arrays, which are designed to transform the incoming signals into the spherical harmonics domain - thus, keeping an intrinsic spatial representation of the sound scene, and easing source localization estimation procedures. Some research has been carried on exploring how this potential can be used on the BSS domain, but usually under simplified conditions (limeted to First Order Ambisonics and/or horizontal plane, for example). Furthermore, the successful introduction of Deep Neural Networks for the BSS problem, which has already shown a very good performance, has still not been fully applied to the Ambisonics domain. The aim of the proposed thesis is then to investigate and apply the most relevant results on the intersection of the presented topics, in order to improve separation methods and provide the basis for a new generation of immersive content creation and manipulation.

Contents

List of Figures

List of Tables

Introduction

The present document describes the author's Industrial PhD Research Proposal, which is developed between the *Music Technology Group* of the *Pompeu Fabra University*, under the supervision of Dr. Emilia Gómez, and the technological center *Eurecat*, under the supervision of Dr. Adán Garriga.

In the present chapter we will briefly introduce the motivations and research context of the work. Chapter 2 presents a comprehensive State-of-the-Art review on the related topics. In Chapter 3 we will expose our Research Plan, based on the critical analysis of the information gathered in Chapter 2.

1.1 Motivation

Blind Source Separation for complex sound scenes is a well know problem. Plenty of methods have been proposed over last decades, covering a big range of use cases, devices and mathematical formulations. *Deep Neural Networks* have been recently started to be applied to the BSS problem in some cases, showing great results and outperforming state of the art results. In an intuitive way, the more information we can gather or estimate from the sound scene, the better the performance of the separation methods. That's why we focus on the analysis of existing proposals and possibilities of Ambisonics microphones. Motivated by the current interest on immersive media and Virtual Reality, many researchers and manufacturers have payed attention again to such devices, and in general to the Ambisonics theory. The availability of new Ambisonics recording devices makes interesting to explore their intrinsic spatial representation capabilities, which on the other hand have been only partially explored by the research community. The potential application possibilities of Blind Source Separation applied to Ambisonic sound scenes go beyond the current geometrical approaches: automatic scene description, custom source enhancement and modification, de/reverberation, automatic speaker

tracking, data compression, etc. In the musical domain, the potential possibilities for innovative immersive production, analysis and manipulation tools are as well promising.

1.2 Context

Eurecat is the result of the merging process of the main Catalan Technology Centres, a process which started in 2015 and still ongoing which counts already with the sum of capacities of seven originals centres and beyond. Eurecat is currently the leading Technology Centre in Catalonia, and the second largest private research organization in Southern Europe. Eurecat manages a turnover of 43M and 600 professionals, is involved in more than 160 RD projects and has a customer portfolio of over 1.000 companies.

The *Audio Research Group*, within the *Multimedia Division*, is one of Eurecat's R&D units. The group counts within its premises with a fully equipped 3D audio studio with state of the art technology: 3D audio and binaural recording hardware, a 3D multichannel reproduction system 25.1 and the Sfar software for 3D audio and music production. The Audio Lab also provides 3D audio solutions for different creative sectors: music, cinematic VR, videogames, audio installations (museums, theatres, festivals), and advertisement.

Scientific Background

2.1 Ambisonics

2.1.1 Introduction

Ambisonics is a sonic theory, developed in the 1970s by Gerzon [3], based on the spatial decomposition of a soundfield into a sequence of *spherical harmonics*. Spherical Harmonics conform a complete set of orthogonal functions defined in spherical coordinates around a sphere. Figure 2.1 depicts the spherical harmonics up to the 3rd degree.

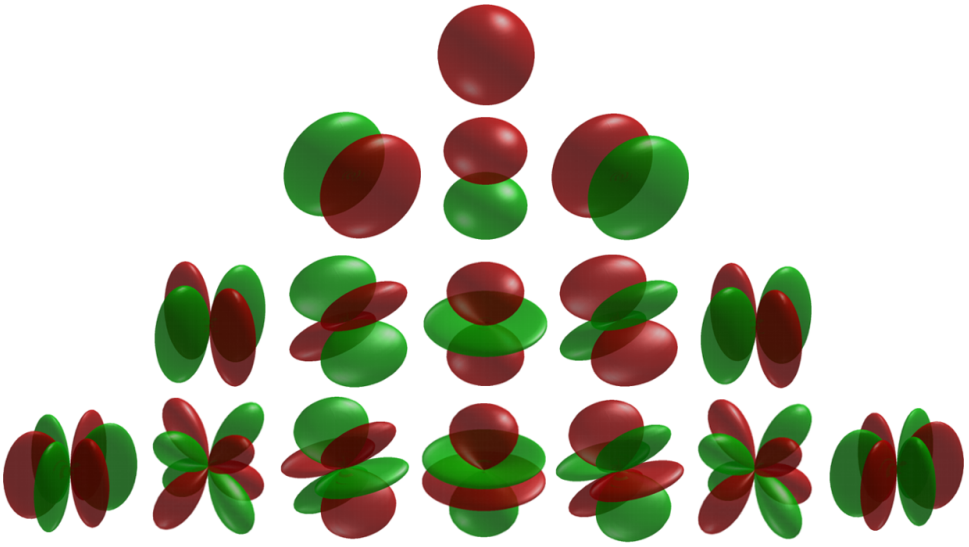

FIGURE 2.1: Spherical harmonics (from [1])

In an intuitive way, the Ambisonics decomposition might be compared with the Fourier Transform. In the latter, a signal is decomposed by the weighted infinite summation of a set of basis functions, and provides an alternative representation of the signal in the Frequency Domain (Figure 2.2). The spherical harmonics decomposition provides a similar approximation, transforming a soundfield into the *spherical harmonics domain*.

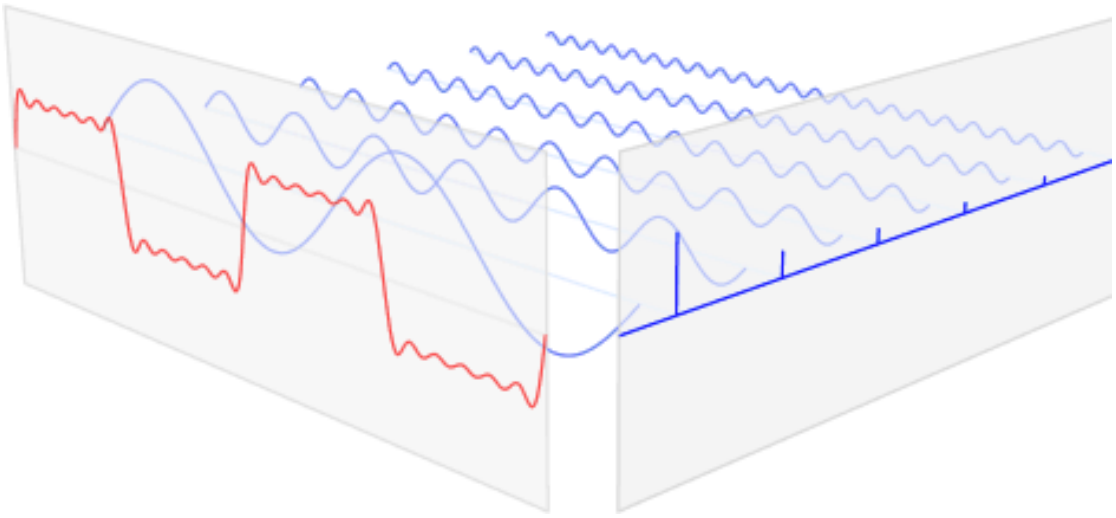

FIGURE 2.2: Dual Time-Frequency representation of the Fourier Transform (from [2])

In practical scenarios, the degree of the spherical harmonics summation must be truncated, and the so called *Ambisonics order* or L represents the maximum degree for a given decomposition. The bigger L, the better the spatial resolution of the sound scene, at the expense of a bigger bandwidth. The relationship between L and the number of spherical harmonic functions N (which is equivalent to the number of channels of the audio file) follows:

$$N = (L+1)^2 \tag{2.1}$$

Due to historical reasons, emphasis is usually placed on the distinction between First Order Ambisonics, as first described by Gerzon [3], and the so called Higher Order Ambisonics, which were deeply studied by Daniel's PhD Thesis [4]. The term *B-Format* usually makes reference to a file which contains FOA audio, but it is sometimes also applied to HOA files.

Ambisonics audio might be obtained by two different means. One possibility consist of syntetically create the sound scene from the individual sources, by computing analytically the *Ambisonics coefficients* (in fact, the projection into the spherical harmonics basis) given the source positions. The second way is to use spherical microphone arrays (usually called *Ambisonics microphone*), which provide Ambisonics audio though digital signal processing and beamforming techniques. In Subsection 2.1.3 we will briefly review some of the existing Ambisonics microphones.

2.1.2 Spatial Audio Delivery

It is commonly agreed that there are three types of approaches or *formats* for spatial audio delivery:

- *Scene-based:* Ambisonics decomposition of the soundfield
- *Object-based*: Sound and position metadata are stored separately for each source.
- *Channel-based*: Classical stereophony (stereo, 5.1, etc), with each channel dedicated to a specific listening position.

Table 2.1 shows some of the benefits and drawbacks for each approach. Please refer to [5] for a more detailed comparison.

Format	**Pros**	**Cons**
Scene-based	• Fixed amount of channels • Layout-independent • Allows for spatial transformations • Standard file format (B-Format) • Available microphones • "Includes" room information (reverb)	• Needs decoder for listening • N grows exponentially for higher spatial resolution (Eq. 2.1)
Object-based	• Very flexible • Provides all information about sources • Layout-independent	• Bandwidth increasing with number of sources • Needs non-audio information (metadata) • No standard file format • Does not include room information • Needs rendering stage for listening
Channel-based	• No rendering stage needed for listening • *De facto* spatial audio delivery standard	• Imposed number and position of channels

TABLE 2.1: Comparison of spatial audio formats

Among the benefits of the scene-based approach, there are two which explains the increasing popularity of Ambisonics. First, it provides an *intermediate* representation, in the sense that it might be used to reconstruct the sound scene for any speaker layout and, with appropriate processing, for binaural reproduction [6].
Second, the mathematical formulation of the spherical harmonics provides deterministic methods for spatial transformations of the sound scene, including rotation around the axes [7]. It is possible then, from a B-Format recording, to transform the scene to binaural, and to rotate the scene according to the listener's head movements with a head-tracking device. It is proved that head-tracking improves the *immersivity* of the sound scene, in terms of source localization and externalization [8].

Therefore, it is easy to understand the support of VR business to Ambisonics. In fact, Ambisonics is currently the standard option for spatial audio support on major VR/360 audiovisual content providers, like Youtube [9] or Facebook [10].

2.1.3 Ambisonics Recording

This situation has lead, as well, to a renewed interest in Ambisonics microphones. Together with well-stablished products, as the *SoundField MKII* [1], the *Tetramic* and the *Eigenmike*, a new set of microphones has been released in recent years. Examples of them are the Sennheiser's *Ambeo*, with a strong marketing emphasis on VR (*"The new dimension of VR audio productions"* [12]), or the upcoming *Zylia ZM-1*, which is oriented towards instrument source separation[2]. Zoom announced compatibility with VR for the H2n[3]. The *Twirling720 Lite*, in pre-order at the moment of writting, will be the first Ambisonics microphone designed for mobile devices (as a USB-compliant microphone). The upcoming *8ball* microphone will feature 8 microphones in a circular array, providing first order horizontal recordings.
Table 2.2 lists several characteristics of some of the currently available Ambisonic microphones.

Ambisonics Microphone	**Number of Capsules**	**Ambisonics Order**	**Release Year**
SoundField	4	1	1978
Tetramic	4	1	2007
Eigenmike	32	4	2008
Ambeo	4	1	2016
Zoom H2n	3	1 horizontal	2016
Zylia ZM-1	19	3	(2017)
Twirling720 Lite	4	1	(2017)
8ball	8	1 horizontal	(2017)

TABLE 2.2: Comparison of Ambisonics microphones

2.2 Sound Source Localization from Microphone Arrays

2.2.1 SSL with Linear Microphone Arrays

Sound Source Localization, also referred as *Direction of Arrival (DOA) Estimation* in the acoustics field, has been an active research topic over last decades. SSL is a relevant field across diverse scientific disciplines, as for instance radar, seismology or telecommunications [14]. It often used, as well, as a preprocessing stage for other signal processing applications, specially sound source enhancement, identification and/or separation.

[1] The first SoundField microphone was manufactured in 1978. Starting in 1993, when the patent expired, other companies started to manufacture the model [11].

[2] In fact, the process of source separation has a preprocessing stage which requires every instrument to play alone and in a fixed position. This fact suggests that the source separation is given by static beamforming based on the preprocessed Direction of Arrival estimation

[3] The firmware version 2.00, released in 2016, allows to record in *Spatial Audio* mode, which is First Order Ambisonics in the horizontal plane [13]

The traditional approach for DOA estimation consists of calculating, for a single sound source, the *Time Difference of Arrival (TDoA)* of the sound wave between a pair of microphones. One of the most relevant algorithms is the *Generalized Cross-Correlation with Phase-Transform (GCC-PHAT)*, which exploits the coherence properties of the microphone signals [15]. *Multiple Signal Classification (MUSIC)* [16], and *Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT)* [17], on the other hand, are two of the most popular methods for the localization of several simultaneous sound emitters.

This topic, however, exceeds the main focus of the thesis. Therefore, the reader is encouraged to refer to [14] or [18] for further information.

2.2.2 SSL with Ambisonics Microphones

As introduced previously in Subsection 2.1.1, Ambisonics microphones are a special case of microphone arrays, in which the capsules are arranged around the surface of a sphere (hence the name *spherical arrays*). The signal might be then processed in the Ambisonic domain, exploiting the structural and geometric properties of this specific arrangement. Consequently, DOA estimation might benefit from this approach.

All methods presented in this section allow for multiple non-static source localization, and are relatively robust against room reverberation effects. Table 2.3 presents a comprehensive comparison of the described methods.

The main Ambisonics SSL algorithm was presented by Pulkki in [19], as a processing stage inside the so called *Directional Audio Coding (DirAC)*. Pulkki derived theoretically a method for computing DOA based on the energetic analysis of the incoming sound wave. More specifically, he defined the *Direction vector*, D, and the *Diffusseness*, Ψ, which can be computed directly from the zero-th and first Ambisonic order representation of the signal.
Assuming that humans are only able to instantaneously identify one sound source per critical band [20], Pulkki proposed to compute D and Ψ for each Time-Frequency bin after the STFT of the B-Format signal. This method is usually referred as the *Intensity Vector (IV)* method.

This proposal was refined by Thiergart and Schultz-Amling [21], modelling the DOAs with *Gaussian Mixture Models (GMM)*, and Tervo [22], who introduced the Von Mises circular distributions for azimuth estimation. Further contributions to the method were authored by Pavlidi, Pulkki *et. al.* [23], who proposed the *Single Source Zone* estimator for performance improvement, and again by Pulkki [24], partially extending the *IV*

concept to Higher Order Ambisonics. Recently, H. Chen and colleagues have proposed several improvements based on binary mask of DOA estimations, by means of *local DOA variance* analysis, accuracy estimation, beamforming or K-Means clustering [25, 26].

A closely related approach is the *Pseudo-Intensity Vector (PIV)*, which is computed from the function solutions (*eigenbeams*) of the spherical Fourier Transform of a HOA input [27]. This method has been applied together with K-Means algorithm for the clustering of the potential DOAs [28], and with the *Direct-Path Dominance (DPD) Test* as a way to reduce the solution space [29]. DPD Test was first presented in [30], as one of the steps of DOA estimation with spherical arrays based on *Planar-Wave Decomposition (PWD)* and spatial correlation in the Spherical Harmonics domain.

To conclude this section, we must mention some other methods based, as well, on the energetic analysis of the Ambisonics soundfield. On the one hand, the algorithm proposed by Berge and Barret, which attempts to decompose the soundfield into two plane waves [31, 32] - this method is commercially available under the name *HARPEX* [33]. On the other hand, the approaches by Dimoulas and colleagues, which described several methods for *Energy Based Localisation (EBL)*, considering as well arrays of SoundField microphones for full 3D source localisation. [34, 35].

Article	**Method**	**Ambisonics Order**	**Microphone**	**Number of Capsules**
Pulkki07 [19]	IV	1	-	-
Thiergart09 [21]	IV + GMM	1 horizontal	Custom circular	4
Tervo09 [22]	IV + vonMises MM	1 horizontal	Custom circular	4
Pavlidi15 [23]	IV + SSZ	1	Custom spherical	32
Pulkki13 [24]	Sectorial IV	HOA	-	-
He17 [25]	IV + local DOA + accuracy + FOSDA	1 horizontal	Custom circular	4
Ding17 [26]	IV + local DOA + accuracy + KMeans	1 horizontal	Custom circular	4
Jarret10 [27]	PIV	HOA	Eigenmike	32
Evers14 [28]	PIV + K-Means	HOA	Custom spherical	32
Moore15 [29]	PIV + DPD	HOA	Custom spherical	32
Nadiri14 [30]	PWD + SCM + DPD	HOA	Eigenmike	32
Berge10 [31]	Harpex	1	-	-
Thiergart12 [32]	Harpex	1	-	-
Dimoulas07 [34]	A-EBL	1	SoundField	4
Dimoulas09 [35]	(DWT/SWT)-JTF-A-EBL	1	SoundField	4

TABLE 2.3: Comparison of Ambisonics SSL methods

2.3 Blind Source Separation

2.3.1 BSS for Monophonic Sources

Blind Source Separation is a well-known problem, which appears in a variety of different scientific scopes as for example Biomedical Signals, Seismology or Radar. In the context of audio source separation, there are several consolidated methods that have been extensively used: *Independent Component Analysis (ICA)* [36], or the most recent *Degenerate Unmixing Estimation Technique* [37] are good examples. In the following sections, we will only focus on BSS approaches from multichannel recordings.

2.3.2 BSS for Multichannel Sources

Many researchers have been interested on the topic of Blind Source Separation for multichannel mixtures during last years. In words of Sawada, "As humans/animals have two ears, multichannel processing is a way of realizing a more general source separation capability because the spatial properties (directions or locations) of source signals can be exploited" [38]. Indeed, the possibility of taking advantage of the spatial information provided by the microphone array might be useful at one or more stages during the BSS processing - for instance, the characterization of the source positions contributes to overcome the well known *permutation problem* [39], by imposing spatial-temporal continuity on the sources [40].

We will distinguish in the following paragraphs two types of approaches. In the first group, we included methods which do not actively exploit DOA estimation (*raw* multichannel approaches). In the second group, the presented methods use deterministic SSL estimation (by means of some of the algorithms reviewed in Section 2.2) as a fundamental part of the separation process.

Raw Multichannel BSS

Most of the works on this field consist of multichannel extensions of the established methods for monophonic BSS. The HOA extension for *Independent Component Analysis (ICA)*, first presented in [40], is a good exmaple. In that article, Epain and colleagues investigated the usage of 2nd order HOA sources through anechoic simulations. This line of research was continued by Baque *et. al.* [41], comparing several different ICA algorithms under reverberant simulations, and using DOA estimation as an evaluation metric.

It must be noticed that, when using ICA algorithms with HOA, the number of target sources must be smaller of equal to the number of spherical harmonics N for a given Ambisonics order L (Equation 2.1). This condition would be equivalent to the limit of the under-determined BSS in the Spherical Harmonics domain.

The method proposed by Ozerov and Fevotte [42] estimates a spectral model for each source by means of *Non-Negative Matrix Factorization (NMF)*. Despite the fact that their methodology used 2 microphones, complying with the SiSEC 2008 [43] specifications, it did not consider any spatial cue of the mix.

Duong and colleagues investigated on the usage of full-rank Gaussian *Spatial Covariance Models* for BSS [44], a method previously introduced for semi-blind source separation in [45]. Separation is provided by *Maximum Likelihood* estimation though EM and Wiener filters. It is interesting to mention that they considered DOA estimation in the algorithm, but just as a mean to help parameter initialization, and for solving the permutation problem.

By combining the spectral and spatial models from [42] and [44, 45], the algorithm by Arberet *et. at.* outperformed the previous results for music blind source separation with the audio mixture from two microphones [46].

Sawada and colleagues [38] proposed multichannel versions of several *cost functions*, such as Euclidean distance and Isakura-Saito divergence, and designed a complex-valued NMF procedure which involves a spatial matrix H.

To conclude this brief overview, we must mention the excellent and recent work by Gannot, Vincent, Markovich-Golan and Ozerov [47]. They featured the most extensive review, up to the present day, of the existing methods for speech enhancement with microphone arrays, considering only the situation with one static speaker.

Table 2.4 summarizes the most relevant information about the exposed Raw Multichannel BSS methods.

SSL-Based Multichannel BSS

As stated in the previous section, spatial information of the sound scene might be included in the blind source separation process. The following proposals exploit actively SSL techniques, usually as an initial step.
As in the case of Section 2.2.1, we will focus on the methods based on spherical microphone arrays. Therefore, we will not review the literature about Linear Micropone Arrays. The reader might refer, for example, to [48] for an overview on the topic.

Article	**Method**	**Microphone Array**	L	**Target Sound**	**Dataset**	**Evaluation Metrics**
Epain10 [40]	ICA	Custom spherical	2	Speech	custom	PESQ
Baque16 [41]	ICA (ERBM)	Custom spherical	2	Speech	custom	SDR, DOA
Ozerov09 [42]	NMF	Linear array	-	Music	SiSEC08	SDR, ISR, SIR, SAR
Duong11 [44]	Gaussian SCM + ML	Linear array	-	Speech	custom	SDR, ISR, SIR, SAR
Arberet11 [46]	Gaussian SCM + NMF	Linear array	-	Music	custom	SDR
Sawada13 [38]	Spatial CNMF	Linear array	-	Music	SiSEC11	SDR

TABLE 2.4: Comparison of Raw Multichannel BSS methods

Gunel and colleagues presented, in 2008, the first work featuring a preprocessing step of DOA estimation from Intensity Vectors on the TF domain computed from a SoundField microphone [49]. The histogram of the estimated DOAs is then used for obtaining the *probability density function* of the sources. Provided that the algorithm requires the source directions to be given, they are used to model the sources with von Mises distributions. Finally, a *directivity function* (TF softmask) is applied to the spectrogram.

This approach has been further refined by Riaz in his PhD Thesis [50]. He proposed a number of improvements, including microphone correction, adaptive filtering, location estimation for moving sources and an extensive experimental validation.

A similar approach was researched by Shujau *et. al.* [51]. In this case, however, an *Acoustic Vector Sensor (AVS)* was used - a device featuring pressure and velocity microphones, which is often found in the field of underwater acoustics [52]. The signal output from the AVS is equivalent to First Order Ambisonics, so the energetic analysis is performed in the same way. After the IV computation, the researchers proposed a binary mask (based on the *Voice Activity Detection (VAD)* algorithm) to select the candidate TF bins, and then the DOAs are computed and clustered for separation. Unlike Gunel [49], the information about number and position of the sources is not required, thus making this algorithm more flexible.

X. Chen *et. al.* [53] proposed some improvements over the works by Gunel [49] and Shujau [51]. Along with IV-based DOA estimation, they considered as well a *Mixing Vector (MV)* estimation (also referred as *Bin-wise classification*) [54]. Von Mises and Gaussian distributions are used to model both estimators, respectively, using an EM algorithm. A soft TF masking is performed as a last step for source separation. Evaluation is provided for both simulations and SoundField recordings, under a reverberant environment for the under-, even- and over-determined cases.

Table 2.5 synthesizes the main features of the aforementioned methods.

Article	Method	Ambisonics Microphone	L	Target Sound	Dataset	Evaluation Metrics
Gunel08 [49]	IV + vonMises MM + Softmask	SoundField	1 h	Speech	Music for Archimedes	SDR, SIR
Riaz15 [50]	Gunel + Mic Correction + Adaptive Filter + Location Estimation	SoundField	1 h	Speech, Music	Music for Archimedes	SDR, ISR, SIR, SAR
Shujau11 [51]	IV + VAD + DOA Clustering + Binary Mask	AVS	1 h	Speech	TIMIT	SDR, ISR, PESQ-MOS
Chen15 [53]	IV + MV + Softmask	SoundField	1 h	Speech	TIMIT	SDR, ISR, PESQ-MOS

TABLE 2.5: Comparison of Ambisonics SSL-Based Multichannel BSS methods

2.4 Multimodal Enhancement for BSS

2.4.1 Audiovisual SSL

Apart from the information that can be obtained from the audio signals, other information sources might be as well analysed in order to improve the performance of the algorithms. It is often the case, in the scope of cinematic VR, in which Ambisonics microphones are widely used, that some kind of stereoscopic or 360 camera system is simultaneously used.

Despite that image processing exceeds the scope of this proposal, information retrieval from video might be used to improve performance of SSL and BSS algorithms. In fact, as Gannot and colleagues point out in their considerations about future work on Blind Source Separation, "the area of audio-visual speech processing remains largely understudied despite its great promise" [47]. This approach was already explored by some of the works that we will briefly review in the following paragraphs, and summarize in Table 2.6.

In the context of audiovisual speech source localization, the simplest approach would be to cluster each sensor-data space separately, and then find an optimal common representation of each unimodal solution . The team of Khalidov proposed the *Conjugate Mixture Model* for joint GMM-clustering of sensor information - in that way, consistency across data spaces can be guaranteed [55]. This algorithm was shown to outperform the former approach of separate clustering, in both simulated and experimental environments with static and moving speakers.

Gebru *et al.* investigated another extension of GMMs applied to audiovisual source localization [56]. The *Weighted-Data GMM (WD-GMM)* proposed algorithm features a

weighting factor, which integrates the reliability of each unimodal data in the gaussian model. They evaluated the proposal through an experiment with multiple speakers, recorded with a binaural dummy head and a camera.

An application of multimodal speaker localization was considered by Khan and colleagues [57]. Their algorithm faces BSS problem through a SSL-based approach. Common monaural and binaural cues from binaural audio, as *MV*, *IDL* or *ILD*, are used for the separation stage. However, the localization stage is completely performed on the visual domain, by using information from two cameras though the *Markov chain Monte Carlo based particle filter (MCMC-PF)* method.

Article	**Localization Method**	**Separation Method**	**Target**	**Mic**	**Camera**
Khalidov11 [55]	Conjugate GMM	-	Speech	2 omni	2
Gebru14 [56]	Weighted-Data GMM	-	Speech	Binaural head	1
Khan13 [57]	MCMC-PF	GMM-EM	Speech	Binaural head	2

TABLE 2.6: Comparison of Multimodal SSL and BSS methods

2.5 DNN for BSS

2.5.1 DNN for Monophonic BSS

Despite its short lifetime, *Deep Neural Networks* have shown a great potential when applied to a variety of MIR problems, in many cases using the knowledge obtained by the Computer Vision research field. In this section, we will briefly overview some of the recent works on musical instrument source separation. Table 2.7 shows a comparative review of the selected works.

Huang and Kim [58] proposed several DNN architectures (RNN, DRNN, stacked RNN) for the problem of separating singing-voice from background. They included an extra output layer which performs a TF softmask, instead of applying it in as a separated stage.

Uhlich and colleagues investigated the usage of DNN with ReLU layers, applied to instrument separation [59]. The instrument types must be given as a problem parameter. In a further work [60], Uhlich's team proposed two DNN architectures capable of separating four instruments independently (vocals, drums, bass and *other*).

Sebastian and Murthy studied the influence of using *Modified Group-Delay (MOD-GD)* instead of the usual magnitude spectrum [61].

Recently, Chandna and colleagues [62] proposed a low-latency DNN structure with orthogonal convolutional layers, each one for modelling a different axis in the magnitude spectrogram.

Article	**DNN Architecture**	**Target**	**Dataset**	**Evaluation Metrics**
Huang14 [58]	DNN, DRNN, sRNN	Singing Voice	MIR-1k	SDR, SIR, SAR
Uhlich15 [59]	ReLU DNNU	Predefined Instrument	TRIOS	SDR, SIR, SAR
Uhlich17 [60]	Feed-Forward, Bi-LSTM	Vocal, Bass, Drum, other	DSD100	SDR,R
Sebastian16 [61]	MOD-GD DRNN	Singing Voice, Vocal-Violin	MIR-1k	SDR, SIR, SAR
Chandna17 [62]	DNN, 2 convolutional layers	Vocal, Bass, Drum, other	DSD100, MSD100	SDR, SIR, SAR, ISR

TABLE 2.7: Comparison of DNN methods for Mono BSS

2.5.2 DNN for Multichannel BSS

In a similar fashion to Section 2.3.2, DNN algorithms applied to BSS might take profit of the largest amount of information provided by a microphone array from a given sound scene. Again, we will make a distinction between the proposals which provide an explicit DOA estimation step prior to source separation, and the proposals who just exploit the extra amount of data. We will start reviewing the latter option.

DNN for Raw Multichannel BSS

Nugraha, Liutkus and Vincent proposed two variants of a method considering 2 and 6 audio channel inputs, respectively [63, 64]. The first one addresses the problem of singing voice separation, while the second one treats the problem of speech enhancement. The methodology is similar to Duong11 [44] and Arberet11 [46], since the problem is reduced to model PSDs and SCMs for each source, from the magnitude spectrogram. However, in this case, DNNs are introduced for two different steps: a DNN for spectrogram initialization, and a DNN for PSD-spectrogram fitting for each target.

On the other hand, Wisdom *et. al* [65] modelled the two speakers separation problem by means of multichannel GMM, and applied deep unfolding in order to derive a specific DNN architecture, based on *Markov Random Fields*. It is noticeable that the multichannel GMM processes complex TF values, instead of magnitude spectrogram.

We must mention the work of Erruz [66], still not publicly available at the moment of writing. He proposed a CNN architecture for music source separation, based on stereo

and synthetic binaural mixtures. Despite the fact that SSL is not performed explicitly, the network takes advantage of the *ILD spectrogram* to integrate spatial information in the learning stage.

Table 2.8 summarizes the reviewed methods.

Article	**Method**	**Microphone Array**	**#Mics**	**Target Sound**	**Dataset**	**Evaluation Metrics**
Nugraha16 [63]	DNN-PSD + SCM	Custom linear	2	Vocals	DSD100	SDR, SIR, ISR, SAR
Nugraha16(2) [64]	DNN-PSD + SCM	Custom linear	6	Speech	CHiME-3	SDR, SIR, ISR, SAR
Wisdom16 [65]	DMCGMM	Custom circular	8	Speech	WSJCAM0 REVERB	SDR
Erruz17 [66]	ILD-CNN	-	2	Music	DSD100	SDR, SIR, ISR, SAR

TABLE 2.8: Comparison of DNN Raw Multichannel BSS methods

DNN for SSL

Before we move forward with the DNN for BSS review, we want to mention a couple of recent works that have been proposed on the scope of Sound Source Localization using DNN. These works are relevant to the present review, since they show how DNNs might be exploited across different aspects of the BSS problem. An overview of the most relevant features of the methods is shown in Table 2.9.

Xiao and colleagues proposed a *Multi-Layer Perceptron (MLP)* architecture to model the non-linear relationship between the TDoAs and the real DOA, by using an 8-capsule circular microphone array, under noise and reverberant conditions [67]. They used the known *GCC-PHAT* method, and modelled the problem as a 360-class pattern classification task (using angular resolution of 1 degree in the horizontal plane). Experimental results showed a great performance increase compared to the traditional *LS* method, specially under very noisy environments.

Chakrabarty and Habets extended Xiao's method in several aspects [68]. The most relevant improvement consisted of using exclusively phase information (without magnitude spectrum) for DOA estimation. In that way, training of the network might be performed by synthesized noise signals, thus removing the necessity of a large database.

DNN for SSL-Based Multichannel BSS

Araki and colleagues [69] approached the speech enhancement problem wit a *Denoising Auto-Encoder (DAE)* architecture, which is a specific neural network used to enhance

Article	Method	Microphone Array	#Mics	Target Sound	Dataset	Evaluation Metrics
Xiao15 [67]	MLP GCC-PHAT	Custom circular	8	Speech	WSJCAM0	DOA RMSE, MAE
Chakrabarty17 [68]	phase spectrogram CNN	Custom linear	4	Speech	Synthesized Noise	SDR

TABLE 2.9: Comparison of DNN SSL methods

speech and minimize noise effect. It is interesting to highlight the usage of analytically computed binaural cues, *ITD* and *ILD*, as a part of input feature vector of the DAE.

Another binaural approach to multichannel BSS was presented by Jiang *et. al.* [70]. ITD and ILD are computed, for each TF bin, from the *normalized Cross-Correlation Function*. They used as well monaural features, as the *Gammatone Frequency Cepstral Coefficients (GFCC)*, computed from the left signal, and used a DNN for processing the speech segregation.

Based on their previous work, Xiao and colleagues proposed a methodology for speech recognition with microphone arrays, featuring a DNN in each of its two stages [71]. The first step is a straightforward implementation of their SSL method [67], with a Feed-Forward NN which estimates DOA from the *GCC-PHAT* - the estimated DOA then is used to perform a beamforming softmask to the complex spectrum. As a second step, monaural features as *log Mel filterbanks* are computed, and given to a speech Acoustic Model DNN, implemented as a LSTM network.

Article	Method	Microphone Array	#Mics	Target Sound	Dataset	Evaluation Metrics
Araki15 [69]	ITD, ILD DAE	Binaural	2	Speech	PASCAL CHiME	SSNR, CD
Jiang14 [70]	ITD, ILD, GFCC DNN	Binaural	2	Speech	Custom speech, ROOM-SIM	HIT, FA, HIT-FA, SNR
Xiao16 [71]	FF GCC-PHAT, LSTM AM	Custom circular	8	Speech	WSJCAM0, REVERB	WER

TABLE 2.10: Comparison of DNN SSL-Based Multichannel BSS methods

2.6 Summary

We have reviewed in this section the basic concepts and the most relevant works on the edge between Ambisonics and Blind Source Separation. In order to provide a global, top-down perspective, we have designed an *Euler Diagram*[4] which shows the most relevant relationships between the methods and concepts for our proposal.
Figure 2.3 shows the labelled fields and the overlapping between scopes. Fields that go beyond this proposal, as for example *Multichannel SSL* (TDoA-like methods), or *Mono BSS with NN* (as for example applied to speech) are not represented.
Figure 2.4 adds some information to the previous diagram. More precisely, we have highlighted the fields for which we reviewed the *State-of-the-Art* in the present section. There are two number codes associated with each mark. The top one represents the Section number of the topic, while the above one shows the Table number with the corresponding article review.

[4] "An Euler diagram is a diagrammatic means of representing sets and their relationships. [...] Unlike Venn diagrams, which show all possible relations between different sets, the Euler diagram shows only relevant relationships" (from [72]).

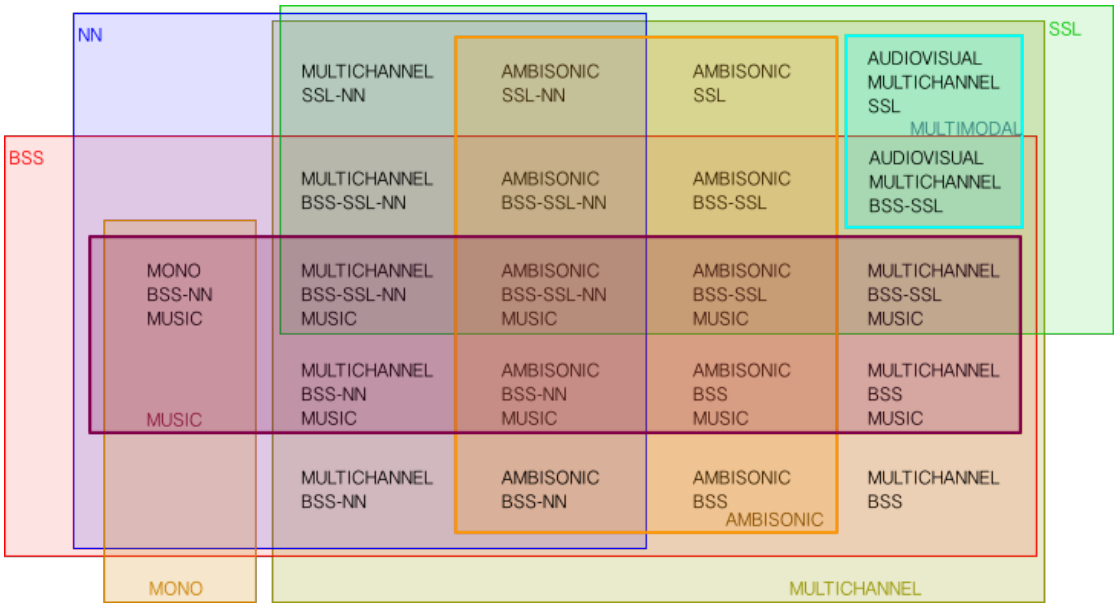

FIGURE 2.3: Euler Diagram showing the proposal's related topics

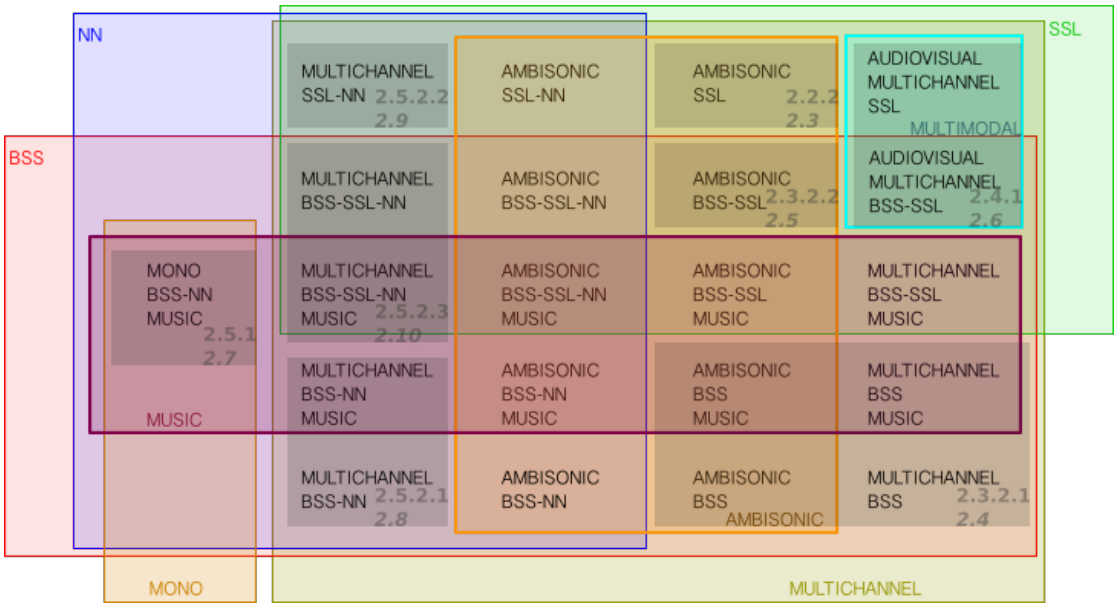

FIGURE 2.4: Euler Diagram highlighting the reviewed topics

Research Proposal

3.1 Goals and Contributions

In Chapter 2 we have broadly reviewed the most relevant works on the intersection of the related domains: Ambisonics, Sound Source Localization, Blind Source Separation, Multimodality and Deep Neural Networks. The Euler's Diagram from Figures 2.3 and 2.4 help to conceptually organize the ideas, and to highlight the conclusions, which will lead to our research proposals.

In short, the idea behind the proposal is based on the following assumption. Specific knowledge about the sound scene, as the microphone characteristics or the nature of the target sound, might be exploited in order to improve the quality of the Blind Source Separation result. This consideration is aligned with the *guidelines* that Gannot *et al.* propose for upcoming BSS methods [47].

- **Conclusions**
  Ambisonics analytical approaches to SSL estimation are consolidated (Table 2.3), and its application to BSS is as well mature (Table 2.5). However, most of the approaches only make use of First Order Ambisonics in the horizontal plane, and target speech sources in most of the cases (Table 2.4)
  On the other hand, recent Deep Learning based methods have shown promising results, both on the SSL domain 2.9) and in the BSS domain (Tables 2.7 and 2.8). Again, the main target situations were pointing speech recognition/enhancement and microphone arrays.
  Lastly, the usage of multimodal data for SSL and BSS enhancement, mainly from the image scope, represents a promising new research area 2.6.

- **Research Goal**
  The main research line is focused on the investigation, adaptation and improvement of existing state-of-the-art algorithms from all the domains involved, in order

to design new methods for Blind Source Separation applied to Higher Order Ambisonics audio, and specially focusing on the musical application domain.

- **Collateral Contributions**

  I A new HOA SSL method based on DNNs, for noisy and reverberant environments.

  II The application of *Contribution I* to the BSS problem, potentially focusing on musical sources.

  III The investigation on DNN-based, raw multichannel BSS approaches to HOA musical signals, including the definition of Music Descriptors in the Ambisonics Domain.

  IV A new audiovisual multimodal approach to SSL and BSS focused on musical instruments for source localization and spectral modelling, eventually considering 360 video images.

In order to visualize the proposed goals and contributions from a top-down perspective, we have mapped them into the Euler Diagram, resulting in Figure 3.1. The red boxes represent the Research Goal and the Collateral Contributions, each one with its roman number label as previously presented.

Furthermore, this information has been also mixed with the scientific background representation in Figure 3.2. The arrows represent the temporal and hierarchical relationships between research areas.

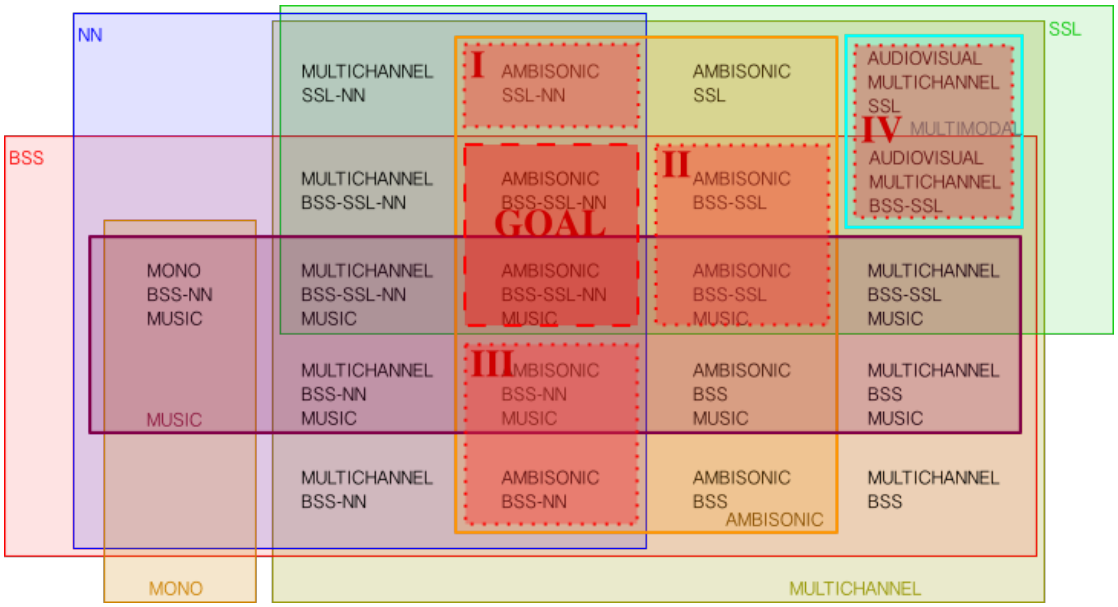

FIGURE 3.1: Euler Diagram with the proposal contributions

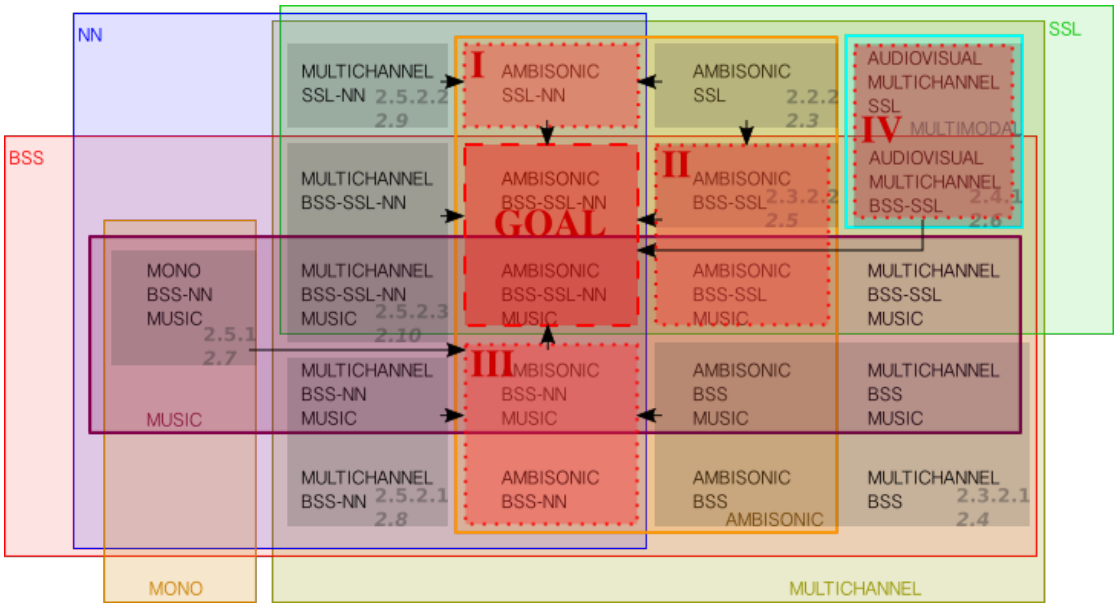

FIGURE 3.2: Euler Diagram with reviewed topics and proposal contributions

Before concluding this section, we must briefly clarify the research methodology related with the multimodal approach. As already stated, the research field of image segmentation is out of the proposal's scope. However, preliminary meetings and proposals have been addressed in oder to collaborate with researchers on the field, as O. Slizovskaia (*Music Technology Group, UPF*), and R. Redondo and C. Bosch (*Multimedia Technologies Group, Eurecat*).

3.2 Research Methodology

3.2.1 Data Generation and Acquisition

The nature of the BSS problem and its evaluation requires two different kinds of data. On the one hand, solo anechoic recordings of speech and music are needed for groundtruth validation and procedural mixture generation. For that purpose, several high-quality standard datasets might be used, as the *MSD100* [73], *DSD100* [74] or *MIR 1k* [75] for music recordings, and the *TIMIT* [76] or *WSJCAM0* [77] for speech.

On the other hand, the BSS problem is specially challenging on the convolutive mixture case, *i.e.*, when reverberation is present on the recording due to the room acoustics - this is, however, the most realistic scenario. Therefore, the most flexible option is to gather Ambisonics IR data, either recorded or simulated. Synthetical sound mixtures can then be generated, taking anechoic recordings and convolving them with the desired IR. Such an approach would allow for arbitrary combinations of sound sources and IRs.

Ambisonics IR data is, though, more difficult to access. To the best of the author's knowledge, there is only one publicly accessible, high-quality, research-oriented database of IRs: the *OpenAir* library, supported by the University of York [78]. At the moment of writing, 26 sets of recorded IRs were available under the tag "B-Format (4 Channel Files)", but any under the category "Higher Order Ambisonics".

Apart from real recordings, IRs might be obtained through simulations by using some of the methods from the computational acoustics field. In the specific case of Ambisonics IRs, the tool *SMIR Generator* [79, 80], created by Jarret and his team at the *International Audio Laboratories Erlangen*, extends the original image method from Allen and Berkley [81] to provide a solution around a sphere, in the spherical harmonics domain. Despite there is commercial software for acoustic simulations capable of providing Ambisonics IRs, we will not consider it on the present proposal, since it exceeds the academic scope.

Therefore, the need for a tool which procedurally creates sound scenes arise. To the best of the author's knowledge, there is any software with such specifications. Furthermore, it would be also desirable to use a systematic scene description methodology, in order to automatize localization evaluation. One existing proposal in that direction is *Spatial Sound Description Interchange Format (SpatDIF)* [82], which comes from the computer music scope. Other options will be further investigated.

- **Contribution V**
  A tool for custom Ambisonic sound scene creation and description, given a database of individual mono recordings and Ambisonics IRs.

A different approach would consist of using real Ambisonics recordings. However, in this case, there are some major drawbacks. The first one is the commercial perspective of most of the Ambisonics libraries available on the internet (*Spheric Collection*[1], *Ambyss*[2], *ProSound Effects*[3] or *A Sound Miner*[4], to cite some of the most relevant). At the moment of writing, *FreeSound*[5] featured 48 sounds under the tag "ambisonics" and 10 under "b-format", but only 7 under "ambisonics" and "4 channel". No HOA recording was available.

Another drawback consists of the ambience-centered Ambisonics recordings. It is certainly true that soundscapes and ambiences might be easily created or recorded through Ambisonics, and therefore it might be argued that most of the recordings available online follow this tendency.

The last point, related with the former one, is the lack of scene description content. Neither explicit geometrical descriptions, nor associated data (for example images) which might be susceptible of being manually annotated is usually provided[6]. Therefore, a proper evaluation of the sound scene in terms of localization might be impossible.

However, it would be very desirable to evaluate the proposed algorithm's performance in a real experimental scenario, without being limited to the simulation scope. The solution would imply to perform real recordings under an acoustically controlled scenario. In order to preserve quality and integrity of the original sources, sound might be played through speakers, either directly or using one of the existing sound spatialization techniques (which would be more suitable for non-static sources). In that way, the sound scene can be again procedurally created and evaluated.

Finally, we must consider the image data necessary for the multimodal processing. On the one hand, regarding the learning stage, several standard datasets might be used, such as the *Youtube-8M*[7] for video, or the *ImageNet*[8] for images. On the other hand, the gathering of 360 images and videos suffers from the same problems as in the case of

[1] http://spheric-collection.com
[2] http://www.ambyss.com
[3] https://shop.prosoundeffects.com/collections/ambisonic
[4] https://www.asoundeffect.com/sound-category/misc-sounds/ambisonics/
[5] http://www.freesound.org
[6] The only exception can be found on the *Ambyss* database, which provides a 360 image preview for each sound scene
[7] https://research.google.com/youtube8m/
[8] www.image-net.org

audio recordings: lack of availability for research and lack of annotations or descriptive data. Again, the solution for experimental setups might lie in the *ad hoc* recording of scenes. As already mentioned, we will not enter into further details of image segregation methods.

3.2.2 Evaluation

In order to assess the validity of the proposals, evaluation method and metrics must be defined for each of the different sub-tasks. Regarding SSL problem, the most common approach is to use objective geometrical evaluation metrics, such as Euclidean Distance or Angular Difference to the groundtruth - combined with measures as the *Root Mean Square Error* or the *Mean of Absolute Error*.

In the context of evaluating the performance quality of a separation algorithms, several objective signal measurements have been broadly used: for instance, *Signal to Distorsion Ratio (SDR)*, *Signal to Interference Ratio (SIR)* or *Signal to Artifacts Ratio (SAR)*. In the case of speech segregation/enhancement, other specific metrics, such as the *Perceptual Speech Quality (PESQ)*, might be considered. The option of subjective evaluations has been also widely explored. For a comprehensive review of the topic, the reader is encouraged to refer to the extensive work of Emiya and colleagues about the topic [83].

3.3 Schedule and Dissemination

The proposed work schedule, in the form of a Gannt Diagram, is showed in Figure 3.3. It is divided in two main sections. The top part, under the label *Research*, shows the different work units, divided by contributions, as stated in Section 3.1. The last month in every work unit, represented with a darker colour, indicates as well that a dissemination action might be performed upon the state of the research at that moment, for the given work unit. Apart from that, the fields *Literature Review* and *Thesis Review and Writing* are also scheduled. The darker colour represents an active, main dedication to the task, while the lighter colour means a background, secondary activity dedication.

The second part collects a list of appropriate conferences where potential results might be disseminated. The selected conventions are the following:

- *LVA ICA*: International Conference on Latent Variable Analysis and Signal Separation

- *ICASSP*: IEEE International Conference on Acoustics, Speech and Signal Processing
- *AES*: Audio Engineering Society Convention
- *SMC*: Sound and Music Computing Conference
- *EUSIPCO*: European Signal Processing Conference
- *ICSA*: International Conference on Spatial Audio
- *DAFx*: International Conference on Digital Audio Effects
- *INTERSPEECH*
- *MLSP*: IEEE International Workshop on Machine Learning for Signal Processing
- *ISMIR*: International Society for Music Information Retrieval Conference

For each event we have marked the date in which it will take place, or in which it is expected to take place.

The diagram also refers to some *Evaluation Challenges* that are related with the present proposal. Date (or expeted date) is again pointed out, as well as the related conference in which it takes place. The challenges are:

- *SiSEC* (LVA ICA)
- *CHiME* (INTERSPEECH)
- *MIREX* (ISMIR)

Finally, at the bottom of Figure 3.3, we have included a list of Journals where we could potentially publish our results:

- *Journal of the Acoustical Society of America*
- *IEEE Transactions on Audio, Speech and Language Processing*
- *IEEE Transactions on Multimedia*
- *Journal of Electrical and Computer Engineering*
- *Journal of New Music Research*

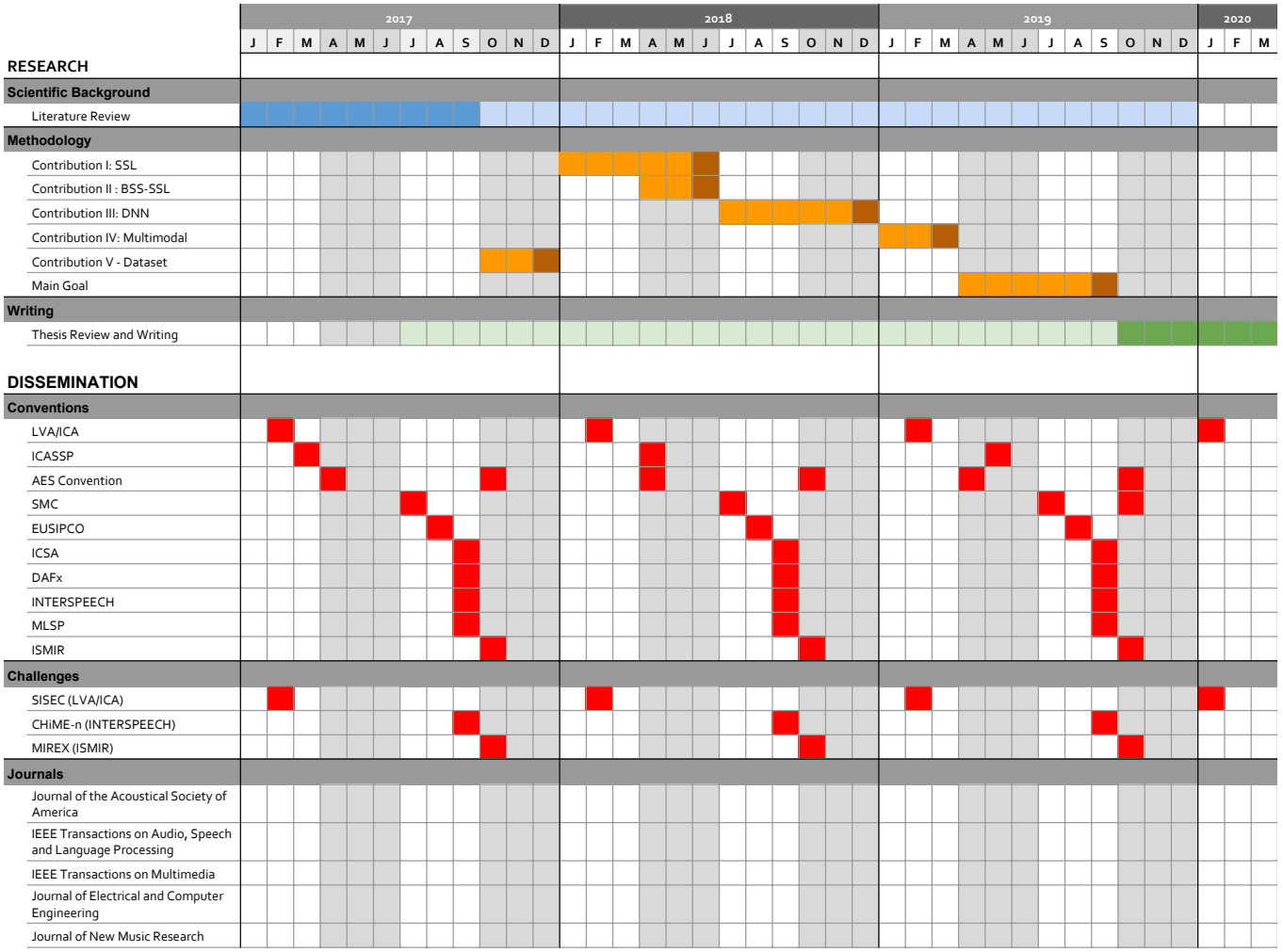

FIGURE 3.3: Gannt Diagram with the proposed Schedule and Dissemination

Bibliography

[1] Sarxos. Harmoniki sferyczne dla parametrw l=0..3 oraz m=-l..l. URL https://commons.wikimedia.org/wiki/File:Harmoniki.png. Accesed 2017-08-24.

[2] Fourier transform. URL http://i.stack.imgur.com/Y5EAf.png. Accesed 2017-08-24.

[3] Michael A. Gerzon. Periphony: With-Height Sound Reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, 1973. ISSN 00047554.

[4] Jerôme Daniel. Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. *Noûs*, (April):319, 2001. URL http://pcfarina.eng.unipr.it/Public/phd-thesis/jd-these-original-version.pdf.

[5] Nils Peters, Deep Sen, Moo-Young Kim, Oliver Wuebbolt, and S Merrill Weiss. Scene-based audio implemented with higher order ambisonics (hoa). In *Annual Technical Conference and Exhibition, SMPTE 2015*, pages 1–13. SMPTE, 2015.

[6] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Höldrich. A 3D ambisonics based binaural sound reproduction system. *Proceedings of the Audio Engineering Society 24th international conference*, (March):237–241, 2003.

[7] Matthias Kronlachner and Franz Zotter. Spatial transformations for the enhancement of Ambisonic recordings. *2nd International Conference on Spatial Audio*, (2): 1–5, 2014.

[8] Stephan Werner, Georg Gtz, and Florian Klein. Influence of head tracking on the externalization of auditory events at divergence between synthesized and listening room using a binaural headphone system. In *Audio Engineering Society Convention 142*, May 2017. URL http://www.aes.org/e-lib/browse.cfm?elib=18568.

[9] Use spatial audio in 360-degree and vr videos. URL https://support.google.com/youtube/answer/6395969. Accesed 2017-08-24.

[10] Introducing spatial audio for 360 videos on facebook. URL https://media.fb.com/2016/10/07/introducing-spatial-audio-for-360-videos-on-facebook/. Accesed 2017-08-24.

[11] Wikipedia. Soundfield microphone, . URL https://en.wikipedia.org/wiki/Soundfield_microphone. Accesed 2017-08-24.

[12] Sennheiser. Ambeo vr mic. URL https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic. Accesed 2017-08-24.

[13] Zoom. New h2n firmware with spatial audio for vr. URL https://www.zoom.co.jp/H2n-update-v2. Accesed 2017-08-24.

[14] Jingdong Chen, Jacob Benesty, and Yiteng Huang. Time delay estimation in room acoustic environments: An overview. *Eurasip Journal on Applied Signal Processing*, 2006(i):1–19, 2006. ISSN 11108657. doi: 10.1155/ASP/2006/26503.

[15] Charles H. Knapp and G. Clifford Carter. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976. ISSN 00963518. doi: 10.1109/TASSP.1976.1162830.

[16] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986. ISSN 0096-1973. doi: 10.1109/TAP.1986.1143830. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1143830.

[17] Richard Roy and Thomas Kailath. ESPRIT - Estimation of Signal Parameters Via Rotational Invariance Techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989. ISSN 00963518. doi: 10.1109/29.32276.

[18] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. Comparison Between Different Sound Source Localization Techniques Based on a Real Data Collection. *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, pages 69–72, 2008. doi: 10.1109/HSCMA.2008.4538690.

[19] Ville Pulkki. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc*, 55(6):503–516, 2007. ISSN 0004-7554. URL http://www.aes.org/e-lib/browse.cfm?elib=14170.

[20] D. R. Perrott. Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays. *Acoustical Society of America Journal*, 76:1704–1712, December 1984. doi: 10.1121/1.391617.

[21] Oliver Thiergart and R Schultz-Amling. Localization of sound sources in reverberant environments based on directional audio coding parameters. *Audio Engineering ...*, pages 1–14, 2009. URL http://www.aes.org/e-lib/browse.cfm?conv=127{&}papernum=7853.

[22] Sakari Tervo. Direction estimation based on sound intensity vectors. *European Signal Processing Conference*, (Eusipco):700–704, 2009. ISSN 22195491.

[23] Despoina Pavlidi, Symeon Delikaris-manias, Ville Pulkki, and Athanasios Mouchtaris. 3D Localization of Multiple Sound Sources With Intensity Vector Estimates in Single Source Zones. pages 1581–1585, 2015.

[24] Ville Pulkki, Archontis Politis, Giovanni Del Galdo, and Achim Kuntz. Parametric spatial audio reproduction with higher-order B-format microphone input. *AES Convention 134*, 2013.

[25] Saijuan He and Huawei Chen. Closed-Form DOA Estimation Using First-Order Differential Microphone Arrays via Joint Temporal-Spectral-Spatial Processing. *IEEE Sensors Journal*, 17(4):1046–1060, 2017. ISSN 1530437X. doi: 10.1109/JSEN.2016.2641449.

[26] Shaowei Ding and Huawei Chen. DOA estimation of multiple speech sources by selecting reliable local sound intensity estimates. *Applied Acoustics*, 127:336–345, 2017. ISSN 0003682X. doi: 10.1016/j.apacoust.2017.07.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0003682X16305357.

[27] Daniel P Jarrett, A P Habets, and Patrick A Naylor. 3D Source Localization in the Spherical Harmonic Domain Using a Pseudointensity Vector. (5):442–446, 2010.

[28] Christine Evers, Alastair H Moore, and Patrick A Naylor. Multiple Source Localization in the Spherical Harmonic Domain. *14th International Workshop on Acoustic Signal Enhancement (IWAENC) MULTIPLE*, (September):8–11, 2014.

[29] Alastair H Moore, Christine Evers, Patrick A Naylor, David L Alon, and Boaz Rafaely. Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test. pages 2341–2345, 2015.

[30] Or Nadiri and Boaz Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE Transactions on Audio, Speech and Language Processing*, 22(10):1494–1505, 2014. ISSN 15587916. doi: 10.1109/TASLP.2014.2337846.

[31] Svein Berge and Natasha Barrett. A new method for B-format to binaural transcoding. *40th AES International conference.*, pages 8–10, 2010. URL http://www.aes.org/e-lib/browse.cfm?elib=15527.

[32] Oliver Thiergart. Robust direction-of-arrival estimation of two simultaneous plane waves from a B-format signal. (November 2012), 2012. doi: 10.1109/EEEI.2012.6376899.

[33] Harpex. URL http://www.harpex.net/. Accesed 2017-08-25.

[34] CA Dimoulas, KA Avdelidis, GM Kalliris, and GV Papanikolaou. Sound source localization and B-format enhancement using soundfield microphone sets. *Audio Engineering Society Convention*, pages 1–8, 2007.

[35] CA Dimoulas, GM Kalliris, KA Avdelidis, and GV Papanikolaou. Improved localization of sound sources using multi-band processing of ambisonic components. *Audio Engineering Society Convention*, 2009.

[36] P Common. Independent component analysis, A new concept? *Signal Processing*, 36:287–314, 1994.

[37] Scott Rickard. The DUET Blind Source Separation Algorithm. *Blind Speech Separation*, pages 217–241, 2007. doi: 10.1007/978-1-4020-6479-1_8.

[38] Hiroshi Sawada, Hirokazu Kameoka, Shoko Araki, and Naonori Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech and Language Processing*, 21(5):971–982, 2013. ISSN 15587916. doi: 10.1109/TASL.2013.2239990.

[39] Hiroshi Sawada, Ryo Mukai, Shoko Araki, and Shoji Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12(5):530–538, 2004. ISSN 10636676. doi: 10.1109/TSA.2004.832994.

[40] Nicolas Epain, Craig Jin, and André Van Schaik. Blind source separation using independent component analysis in the spherical harmonic domain. *International Symposium on Ambisonics and Spherical Acoustics*, 71:S54–S59, 2010. ISSN 01681699. doi: 10.1016/j.compag.2009.07.014. URL http://ambisonics10.ircam.fr/drupal/?q=proceedings/o10.

[41] Mathieu Baque, Alexandre Guerin, and Manuel Melon. Separation of Direct Sounds from Early Reflections Using the Entropy Rate Bound Minimization Algorithm. *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*, (February):1–8, 2016. URL http://www.aes.org/e-lib/browse.cfm?elib=18072.

[42] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures. With application to blind audio source separation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 3137–3140, 2009. ISSN 15206149. doi: 10.1109/ICASSP.2009.4960289.

[43] Sisec 2008. URL http://sisec2008.wiki.irisa.fr/tiki-index.html. Accesed 2017-08-26.

[44] Ngoc Duong, Emmanuel Vincent, Ngoc Duong, and Emmanuel Vincent. Under-determined convolutive blind source separation using spatial covariance models. 2011.

[45] Ngoc Duong, Emmanuel Vincent, and Remi Gribonval. Spatial covariance models for under-determined reverberant audio source separation. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 129–132, 2009. ISSN 1931-1168. doi: 10.1109/ASPAA.2009.5346503.

[46] Simon Arberet, Alexey Ozerov, Ngoc Duong, Emmanuel Vincent, Pierre Vandergheynst, and Irisa Umr. Nonnegative Matrix Factorization and Spatial Convariance Model For Under-Determined Reverberant Audio Source Separation. *Information Systems*, (M):1–4, 2011.

[47] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-golan, Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-golan, Alexey Ozerov A, Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-golan, and Alexey Ozerov. A consolidated perspective on multi-microphone speech enhancement and source separation. 2017.

[48] Hiroshi Saruwatari, Toshiya Kawamura, Tsuyoki Nishikawa, Akinobu Lee, and Kiyohiro Shikano. Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2):666–678, 2006. ISSN 15587916. doi: 10.1109/TSA.2005.855832.

[49] Banu Günel, Hüseyin Hachabibolu, and Ahmet M. Kondoz. Acoustic source separation of convolutive mixtures based on intensity vector statistics. *IEEE Transactions on Audio, Speech and Language Processing*, 16(4):748–756, 2008. ISSN 15587916. doi: 10.1109/TASL.2008.918967.

[50] A Riaz. Adaptive Blind Source Separation Based on Intensity Vector Statistics. (December), 2015.

[51] M. Shujau, C. H. Ritz, and I. S. Burnett. Separation of speech sources using an acoustic vector sensor. *MMSP 2011 - IEEE International Workshop on Multimedia Signal Processing*, 2011. doi: 10.1109/MMSP.2011.6093797.

[52] M. Shujau, C. H. Ritz, and I. S. Burnett. Designing acoustic vector sensors for localisation of sound sources in air. *European Signal Processing Conference*, (August 2017):849–853, 2009. ISSN 22195491.

[53] Xiaoyi Chen, Wenwu Wang, Yingmin Wang, Xionghu Zhong, and Atiyeh Alinaghi. Reverberant speech separation with probabilistic time frequency masking for B-format recordings. *SPEECH COMMUNICATION*, 68:41–54, 2015. ISSN 0167-6393. doi: 10.1016/j.specom.2015.01.002. URL http://dx.doi.org/10.1016/j.specom.2015.01.002.

[54] Hiroshi Sawada, Shoko Araki, and Shoji Makino. A Two-Stage Frequency-Domain Blind Source Separation Method for. *2007 IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics*, (3):139–142, 2007.

[55] Vasil Khalidov, Florence Forbes, and Radu Horaud. Conjugate mixture models for clustering multimodal data. *Neural computation*, 23(2):517–57, 2011. ISSN 1530-888X. doi: 10.1162/NECO_a_00074. URL http://www.ncbi.nlm.nih.gov/pubmed/21105829.

[56] Israel D. Gebru, Xavier Alameda-Pineda, Radu Horaud, and Florence Forbes. Audio-visual speaker localization via weighted clustering. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2014. ISSN 21610371. doi: 10.1109/MLSP.2014.6958874.

[57] Muhammad Salman Khan, Syed Mohsen Naqvi, Ata ur Rehman, Wenwu Wang, and Jonathon Chambers. Video-Aided Model-Based Source Separation in Real Reverberant Rooms. *IEEE Transactions on Audio, Speech and Language Processin*, 21(0):1900–1912, 2013.

[58] Po Sen Huang and Minje Kim. Singing-Voice Separation From Monaural Recordings Using Deep Recurrent Neural Networks. *International Society for Music Information Retrieval*, pages 477–482, 2014. ISSN 15206149. doi: 10.1109/ICASSP.2014.6853860. URL http://cal.cs.illinois.edu/papers/huang-ismir2014.pdf.

[59] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2015-Augus, pages 2135–2139, 2015. ISBN 9781467369978. doi: 10.1109/ICASSP.2015.7178348.

[60] Stefan Uhlich, Marcello Porcu, Franck Gimd, Michael Enenkl, and Thomas Kempl. IMPROVINGMUSIC SOURCE SEPARATION BASED ON DEEP NEURAL NETWORKS THROUGH DATA AUGMENTATION AND NETWORK BLENDING. *ICASSP 2017*, (March):261–265, 2017. doi: 10.1109/ICASSP.2017.7952158.

[61] J. Sebastian and H.A. Murthy. Group delay based music source separation using deep recurrent neural networks. *2016 International Conference on Signal Processing and Communications, SPCOM 2016*, (August):1–5, 2016. doi: 10.1109/SPCOM.2016.7746672. URL http://ieeexplore.ieee.org/document/7746672/.

[62] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.

[63] Aditya Arie Nugraha, Antoine Liutkus, Emmanuel Vincent, Aditya Arie Nugraha, Antoine Liutkus, Emmanuel Vincent, Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel music separation with deep neural networks. 2016.

[64] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio*, 24(10):1652–1664, 2016. doi: 10.1109/TASLP.2016.2580946⟩. URL https://hal.inria.fr/hal-01163369v5.

[65] Scott Wisdom, John Hershey, Jonathan Le Roux, and Shinji Watanabe. Deep unfolding for multichannel source separation. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016-May:121–125, 2016. ISSN 15206149. doi: 10.1109/ICASSP.2016.7471649.

[66] Gerard Erruz. *Binaural Source Separation with Convolutional Neural Networks*. PhD thesis, Universitat Pompeu Fabra, 2017.

[67] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L. Jones, Eng Siong Chng, and Haizhou Li. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015-Augus:2814–2818, 2015. ISSN 15206149. doi: 10.1109/ICASSP.2015.7178484.

[68] Soumitro Chakrabarty and Emanuël. A. P. Habets. Broadband DOA estimation using Convolutional neural networks trained with noise signals. pages 0–4, 2017. URL http://arxiv.org/abs/1705.00919.

[69] Shoko Araki, Tomoki Hayashi, Marc Delcroix, Masakiyo Fujimoto, Kazuya Takeda, and Tomohiro Nakatani. Exploring multi-channel features for denoising-autoencoder-based speech enhancement. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015-Augus(ii):116–120, 2015. ISSN 15206149. doi: 10.1109/ICASSP.2015.7177943.

[70] Yi Jiang, De Liang Wang, Run Sheng Liu, and Zhen Ming Feng. Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Transactions on Speech and Language Processing*, 22(12):2112–2121, 2014. ISSN 23299290. doi: 10.1109/TASLP.2014.2361023.

[71] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu. Deep Beamforming Networks for Multi-Channel Speech Recognition. 2016. URL http://www.merl.com.

[72] Wikipedia. Euler diagram, . URL https://en.wikipedia.org/wiki/Euler_diagram. Accesed 2017-08-24.

[73] The mixing secret dataset 100 (msd100). URL https://sisec.inria.fr/sisec-2015/2015-professionally-produced-music-recordings/. Accesed 2017-08-31.

[74] Demixing secret dataset 100 (dsd100). URL https://sisec.inria.fr/home/2016-professionally-produced-music-recordings/. Accesed 2017-08-31.

[75] Mir-1k dataset. URL https://sites.google.com/site/unvoicedsoundseparation/mir-1k. Accesed 2017-08-31.

[76] John Garofolo et al. Timit acoustic-phonetic continuous speech corpus ldc93s1. URL https://catalog.ldc.upenn.edu/ldc93s1. Accesed 2017-08-31.

[77] Wsjcam0 corpus and recording description. URL https://catalog.ldc.upenn.edu/docs/LDC95S24/wsjcam0.html. Accesed 2017-08-31.

[78] Openair - impulse response database. URL http://www.openairlib.net/. Accesed 2017-08-31.

[79] D P Jarrett, E a P Habets, M R P Thomas, and P a Naylor. Rigid sphere room impulse response simulation: Algorithm and applications. *Journal of the Acoustical Society of America*, 132(3):1462–1472, 2012. ISSN 00014966. doi: 10.1121/1.4740497.

[80] Smir generator. URL https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator. Accesed 2017-08-31.

[81] Jont B. Allen. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943, 1979. ISSN 00014966. doi: 10.1121/1.382599.

[82] Niels Peters, Trond Lossius, and Jan C. Schacher. The Spatial Sound Description Interchange Format: Principles, Specification, and Examples. *Computer Music Journal*, 37(1):11–22, 2013. ISSN 1531-5169. doi: 10.1162/COMJ.

[83] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and Objective Quality Assessment of Audio Source Separation. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2046 –2057, 2011. ISSN 1558-7916. doi: 10.1109/TASL.2011.2109381.