EVALUATION OF PAIR-WISE SIMILARITY SPOTFORMING ALGORITHM ON REAL OMNIDIRECTIONAL SIGNALS AND AMBISONIC SIGNALS WITH SEARCH FOR IMPROVEMENTS ON THE ALGORITHM

by

SOUCHAUD ANTOINE ROBERT

March 16 2024 - August 16 2024

Supervisor	Professor Ville Pulkki
Supervisor	Professor Emmanuel Saint-James
Sub-Supervisor	Doctoral Candidate Stefan Wirler

A THESIS

Presented to the Faculte des Sciences et Ingenierie
and the Division of Graduate Studies of the Sorbonne Université, Télécom Paris and IRCAM
in partial fulfillment of the requirements
for the degree of
Master in Informatique, "Parcours ATIAM"

September 2024
Thesis written during August 2024

THESIS ABSTRACT

Souchaud Antoine Robert

Master in Informatique, "Parcours ATIAM"

Faculte des Sciences et Ingenierie

September 2024

Title: Evaluation of pair-wise similarity spotforming algorithm on real omnidirectional signals and ambisonic signals with search for improvements on the algorithm

Abstract

This master's thesis explores, at first, the application of a spotforming algorithm developed by Stefan Wirler to real signals and real ambisonics signals, and secondly, attempts to better said algorithm through a heuristic pruning and fine-tuning of the algorithm and an analysis of harmonic structure in the context of harmonic signals within the Spot Of Interest (SOI). The spotforming algorithm is based on a post-filter created from the pair-wise coherence of distributed microphone arrays. The primary objective of this research is to confirm the results obtained through simulation of the algorithm and observe possible improvements when using ambisonic signals steered towards the SOI. A secondary objective of this research is to find improvements to the algorithm once the first objective is fulfilled. There is little research done on the topic of spotforming. However, interest in the topic has grown in recent years. This technology would prove useful in noisy environments where only an SOI is targeted, independent of any speaker or entity. This is indeed a different approach than classical source separation, which generally focuses on the characteristics of a certain speaker to extract the target speaker from the mixture. There are also spatial source separation techniques, the vast majority being beamforming algorithms, in which one separates sources by selecting a direction different from what we study during this thesis. A more in-depth explanation will be given in 1.2, 2.1 and 2.2

Objective evaluations have been conducted to assess the effectiveness of these approaches in terms of the source-to-interference ratio (SIR), source-to-noise ratio (SNR) and source-to-artefact ratio (SAR) for objective evaluations. The results indicate that using first-order ambisonic signals steered towards the SOI increases SIR

by an average of −3 dB, 13.3 dB and 1.5 dB for the best method (DSPF) in rooms with RT60 values of 0.33, 0.52 and 1.6 seconds respectively. The results for the other methods show a better trend using directional data. A more detailed view and explanation are given in 5 and 5.3. As for the research done on the criterion side, both the heuristic pruning and the harmonic structure-based criterion show an average improvement of 2.7 dB and 1.7 dB for the best method (DSPF) in the least reverberant case. More tests have been done in different situations, and both proposed methods show betterment on average as shown in 5 and 5.4. Furthermore, the proposed criterion also shows improvement on average compared to the criterion used in the original paper.

ACKNOWLEDGMENTS

I am deeply grateful to the individuals and organizations whose support and assistance were crucial in the completion of my master thesis. Your contributions were invaluable and I could not have done it without you. I am especially grateful to Stefan Wirler, for being an exceptional academic advisor that was always available for questions or when asking for input. I want to thank Professor Ville Pulkki for welcoming me in the Acousitcs Lab at Aalto University. I am thankful to Cyrielle Fiolet for assisting me with all aspects of administrative work and making this exchange possible. I am also very grateful to everyone who works in the Acoustics lab for a fantastic work environment that is fun and cooperative and pushes one to achieve more. I am thankful to Aalto University for giving me the opportunity to work at the Acoustics lab. I am thankful to Sorbonne Université and the ERASMUS+ program for helping fund this project. Finally, I would like to thank the institutions of IRCAM, Sorbonne Université, Télécom Paris, and all of the professors involved for offering such an interesting master program, the ATIAM master program, to students such as I who are passionate about music and audio and can study these subjects scientifically.

TABLE OF CONTENTS

Chapter Page

TABLE OF CONTENTS 6

LIST OF FIGURES 7

LIST OF TABLES 10

1. INTRODUCTION 11
   - 1.1. Background and context 11
   - 1.2. Research problems and objective 11
   - 1.3. Significance and motivation of the study 12
2. LITERATURE REVIEW 14
   - 2.1. Review of relevant literature 14
   - 2.2. Theoretical foundations and concepts 15
     - *Original method* 15
     - *Spherical harmonics and ambisonics* 16
3. PROPOSED METHOD 19
   - 3.1. Criteria for choosing pair-wise filters and microphone selection 19
   - 3.2. Additional processing for better extraction 23
4. METHODOLOGY 25
   - 4.1. Dataset acquisitions 25
   - 4.2. Directional signals 26
   - 4.3. Metrics for evaluation 28
5. RESULTS AND DISCUSSION 32
   - 5.1. Simulation results 32
   - 5.2. Omnidirectional signals results 34
   - 5.3. Directional signal results 36
   - 5.4. Criteria for choosing pair-wise filters and microphone selection results 39
   - 5.5. Addressing limitations 43
   - 5.6. Suggestions for further research 43
6. CONCLUSION 45

BIBLIOGRAPHY 46

LIST OF FIGURES

Figure Page

1.1. Illustration of the problems one can encounter with source separation when using a beamforming algorithm. The microphone array is represented in blue, the interfering source in red and the desired source in green. The beamformer has a spatial selectivity represented by the blue cone. Here an interfering source and the desired source are in the same direction as the microphone. This makes source separation difficult with beamforming algorithms. 12

1.2. Illustration of how spotforming algorithms could solve the problems encountered using a beamforming algorithm. The blue dots represent the microphone arrays, the red dot is the interfering source, the green dot is the desired source and the green area represents the SOI, the area where the algorithm can isolate sources from. 12

2.1. An illustration of the spherical harmonics, up to the 3rd degree. The lobes with a lighter shade have opposite polarity compared to those with a darker shade. Image taken from [ZF19]. 17

2.2. An illustration of the B-Format channels W, X and Y. The last channel, Z, would be perpendicular to the X and Y channels. 18

2.3. First and second order ambisonics in the 2D plane. 18

3.1. Illustration of the steered signals towards a SOI with first order ambisonics. The star represents the SOI, the blue dots represent the location of the microphones and the green square an interfering source. The polar patters represent the steered ambisonic signal. 19

3.2. Values of the variable G_{pq} for p = 7 and q = 8 at the time bin n = 44. This can be seen as the FFT of a signal that should have harmonic structure. 21

3.3. Values of the variable $\Phi_{pqpitch}$ for p = 7 and q = 8 at the time bin n = 44. . . . 21

3.4. Graph illustration the regions where different frequencies have a coherence higher than 0.7 given a specific microphone layout. The blue regions represent the 125 Hz frequency, the red regions represent the 250 Hz frequency, and the green region represents the 500 Hz frequency [WP24]. 23

3.5. Naive filter to boost harmonic signals found in the SOI. 24

4.1. Image of a Zylia ZM-1S microphone. Image taken from [Gui23] 26

4.2. Layout of the Zylia ZM-1S microphones and the Genelec 830A speakers in the Arni room. In red are the speaker locations, and in blue are the microphone locations. The arrows represent the look direction of the speakers and microphones. 26

5.1. Results of the simulation using the omnidirectional signal and the DaS signal when using the original criterion. The colored dots represent individual simulation results, whilst the white dot represents the mean for that method. 33

5.2. Results of the simulation using the omnidirectional signal and the DaS signal as a baseline to create the dipole and cardioid signals using only the harmonic criterion. The colored dots represent individual simulation results, whilst the white dot represents the mean for that method. 33

5.3. Results of the simulation using the omnidirectional signal and the DaS signal as a baseline to create the dipole and cardioid signals when using the conjunction of the original criterion and the harmonic criterion with the boosting filter. The colored dots represent individual simulation results, whilst the white dot represents the mean for that method. 34

5.4. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains a male speaker; there are 2 voices in total. 34

5.5. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains an English female speaker; there are 2 voices in total. 35

5.6. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains the male speaker, there are 3 voices total. 35

5.7. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains the English female speaker, there are 3 voices total. 36

5.8. Evolution of the objective metrics versus the RT60 values of the room for the 1st ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the male speaker, there are 2 voices total. 37

5.9. Evolution of the objective metrics versus the RT60 values of the room for the 1st ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the English female speaker, there are 2 voices total. 37
5.10. Evolution of the objective metrics versus the RT60 values of the room for the 1st ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the male speaker, there are 3 voices total. 38
5.11. Evolution of the objective metrics versus the RT60 values of the room for the 1st-order ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the English female speaker, there are 3 voices total. 38
5.12. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the male speaker, there are 2 voices total. 39
5.13. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the English female speaker, there are 2 voices total. 40
5.14. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the male speaker, there are 3 voices total. 40
5.15. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the English female speaker, there are 3 voices total. 41

LIST OF TABLES

Table Page

4.1. Hyperparameters used for the simulation 27

4.2. Position of the microphones in the simulated room 27

4.3. Hyperparameters used for the evaluation in real conditions. 31

5.1. Average difference in SIR between the original mixture and the DSPF method. 36

5.2. Average difference in SDR between the original mixture and the DSPF method. 36

5.3. Average difference in SIR between using omnidirectional signals and 1st order ambisonics. 39

5.4. Average difference in SDR between using omnidirectional signals and 1st order ambisonics. 39

5.5. Average difference in SIR between using the heuristic pruning criterion and the original criterion with omnidirectional signals. 41

5.6. Average difference in SDR between using the heuristic pruning criterion and the original criterion with omnidirectional signals. 41

5.7. Average difference in SIR between using the conjunction of the harmonic criterion and the energy criterion and the original criterion with the boosting filter and the omnidirectional signals. 42

5.8. Average difference in SDR between using the conjunction of the harmonic criterion and the energy criterion and the original criterion with the boosting filter and omnidirectional signals. 42

5.9. Average difference in SIR between using the conjunction of the harmonic criterion and the energy criterion and the original criterion with the boosting filter and omnidirectional signals without the outlier value. The outlier value is replaced by a *NaN* value. 43

5.10. Average difference in SDR between using the conjunction of the harmonic criterion and the energy criterion, and the original criterion with the boosting filter and omnidirectional signals without the outlier value. The outlier value is replaced by a *NaN* value. 43

CHAPTER 1

INTRODUCTION

In this section, we will go over the background of this thesis, the objectives set and the motivations for this study.

1.1 Background and context

The research focus of this master's internship emerged from the desire to evaluate the spotforming algorithm proposed by Wirler et al. on real signals. While the algorithm was initially assessed using simulated data in the submitted paper [WP24], it was necessary to validate its performance with real-world signals. Additionally, a potential enhancement involved utilizing ambisonic signals and directing them toward the signal of interest (SOI), although this modification has yet to be explored in the submitted paper. The method was intended to be evaluated in various environments, including rooms with different reverberation times (RT60) and configurations with varying numbers of speakers placed at different positions. Throughout the evaluation process, any potential improvements to the algorithm were thoroughly tested, forming a crucial component of the master's thesis.

1.2 Research problems and objective

The typical approach to enhancing or separating a source using microphone arrays is known as beamforming. This technique involves applying specific processing to a compact array of antennas to emit or receive signals from a designated direction. Initially introduced in the field of radio communication, beamforming has since been extensively adopted in acoustics. Various beamforming algorithms exist, including those described in [Vor13], [EA00], and [Van02]. These algorithms are classified as beamforming methods because they can direct the array towards a specific direction. The critical distinction between beamforming and spotforming algorithms lies in their focus on direction versus location. For instance, in the scenario illustrated in Fig. 1.1, beamforming algorithms may struggle to separate two sources, while a spotforming algorithm, provided the correct location of the desired source, could effectively isolate it from other present sources, as depicted in Fig. 1.2. The overarching goal of this research is to develop a beamforming algorithm that also incorporates distance as a critical factor.

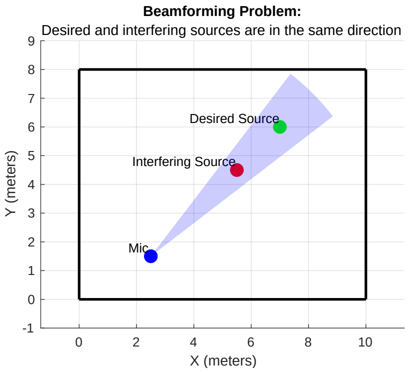


Figure 1.1. Illustration of the problems one can encounter with source separation when using a beamforming algorithm. The microphone array is represented in blue, the interfering source in red and the desired source in green. The beamformer has a spatial selectivity represented by the blue cone. Here an interfering source and the desired source are in the same direction as the microphone. This makes source separation difficult with beamforming algorithms.

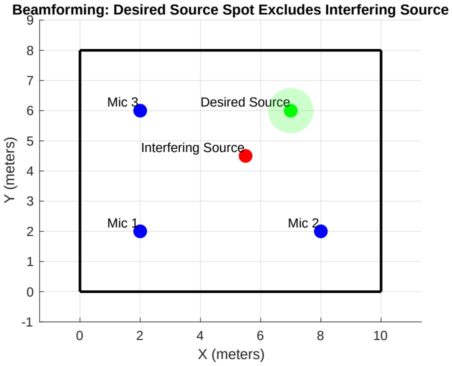


Figure 1.2. Illustration of how spotforming algorithms could solve the problems encountered using a beamforming algorithm. The blue dots represent the microphone arrays, the red dot is the interfering source, the green dot is the desired source and the green area represents the SOI, the area where the algorithm can isolate sources from.

1.3 Significance and motivation of the study

In the previously discussed scenario, repositioning the microphone may resolve the issue for a beamforming algorithm. However, one can imagine a ring of undesired noise surrounds the desired target. Such scenarios are often referred to as the "cocktail party" problem, which exemplifies the challenges of speech recognition in crowded environments. In this context, if a microphone array is positioned around the sources in a cocktail party setting, and the desired source is located within the array, a traditional beamforming solution would likely be inadequate. Conversely, a spotforming approach could potentially address this issue by isolating the desired source from the surrounding noise.

CHAPTER 2

LITERATURE REVIEW

This section will first present the state of the art in spot forming. Afterwards, we will describe the method proposed in the paper this master's thesis is based on. Finally, we will briefly describe the concepts of spherical harmonics and ambisonics.

2.1 Review of relevant literature

Spotforming is a relatively unexplored field ; consequently, limited literature is available on the subject. One notable approach to achieving spotforming was introduced by Kagimoto et al. [Kag+22] through the use of Non-Negative Matrix Factorization (NMF). The core concept involves first generating beamformed signals from multiple angles, ensuring that the beams overlap at the signal of interest (SOI) so that all beams capture the SOI's information. Subsequently, NMF is applied to the concatenated beamformed signals to identify a common basis matrix. The activation matrices are then re-learned to detect the information activated simultaneously across all microphones. This method assumes that the information activated simultaneously originates from the SOI. A binary mask is then created and applied to the activation matrix to extract the information corresponding to the SOI.

Another approach to spotforming is presented by Taseska and Habets [TH16], who propose a method based on Minimum Variance Distortionless Response (MVDR) spotforming. This technique utilizes power spectral density matrices to distinguish between desired and undesired speakers and estimate the probability of the desired speaker being located within the signal of interest (SOI).

Additionally, the method proposed by Suzuki and Honjo [SH15] closely aligns with our proposed approach. Suzuki and Honjo introduce a spotforming technique that relies on overlapping signals captured by two shotgun microphones. A shotgun microphone is characterized by its directional selectivity, akin to a beamformed or ambisonic signal. Their method involves delaying and summing the two signals to focus on a specific SOI. This approach essentially represents the initial stage of the algorithm proposed by Wirler, as described in 2.2.

Original method

The following section provides a brief overview of the method introduced in the original work. This explanation assumes an anechoic signal.

In a system with M microphones, the signal received by the m-th microphone is written as,:

$$x_m(t) = s(t - \tau_m). \tag{2.1}$$

Given a pair of microphones p and q. The inter-microphone time-delay function (IMTDF) for the signal to reach both microphones can be written as :

$$\tau_{pq}(r_s, r_p, r_q) = \frac{\|r_s - r_p\| - \|r_s - r_q\|}{c}, \tag{2.2}$$

with c the speed of sound, $\mathrm{r_s}$, $\mathrm{r_p}$ and $\mathrm{r_q}$ the positions of the source, the p-th microphone and the q-th microphone respectively.

For the remainder of this document, we will describe the signals in the time-frequency domain.

The equation given in Eq. 2.1 can be rewritten as :

$$X_m(k, n) = S(k, n)e^{-j\omega(k)\tau_m}, \tag{2.3}$$

where ω(k) denotes the angular frequency of the k-th frequency bin.

We can then derive this equation :

$$e^{-j\omega(k)\tau} = \frac{X_p X_q^*}{|X_p X_q^*|}, \tag{2.4}$$

where τ is the time IMTDF for the p-th microphone and the q-th microphone

Given this, we can aim in a specific direction for every pair of microphones. Overlapping these directions makes us aim for an SOI because the microphones are distributed throughout the room.

In the original paper, a combination of the real part of the phase-corrected normalized cross-power spectrum for each microphone pair is used to create the post-filter, given that a pair passes the criterion. This is written as :

$$\Phi_{pq}(k,n) = \frac{X_p X_q^*}{|X_p X_q^*|} e^{j\omega(k)\tau_{pq}}, \tag{2.5a}$$

$$G_{pq}(k,n) = \max(\lambda_{pq}, \mathrm{Re}\{\Phi_{pq}(k,n)\}). \tag{2.5b}$$

For each frequency bin k and time bin n, the value of $\mathrm{G_{pq}(k,n)}$ is between λ_{pq} and 1. Values closer to 1 correspond to information gathered from the SOI. There is then a combination of the $\mathrm{G_{pq}(k,n)}$ that have passed the criterion. The original criterion given in the paper is written as :

$$\frac{\sum_{k=1}^{K} |X_p(k,n)X_q^*(k,n)|}{\sum_{k=1}^{K} \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{l=i+1}^{M} |X_l(k,n)X_i^*(k,n)|} > 1. \tag{2.6}$$

After the combination scheme, more processing is done, such as temporal smoothing or time-frequency flooring. The final post filter is multiplied by the time-frequency signal of one of the microphones or the Delay-and-Sum (DaS) signal created from the combination of all the microphone signals or the output signal of a beamformer, as :

$$Y(k,n) = G(k,n)S(k,n). \tag{2.7}$$

In this equation $Y(k,n)$ corresponds to the STFT of the output of the algorithm, $G(k,n)$ is the estimated post-filter, and $S(k,n)$ is the STFT of the input signal, whether the omnidirectional signal or the DaS signal.

We then apply an ISFTF on Y to get the time-domain signal.

Spherical harmonics and ambisonics

The following section briefly introduces the concept of spherical harmonics and ambisonics. The original algorithm is extended by incorporating directional signals derived from the ambisonic signals. With the advancements of 3D audio, a growing interest to have better spatialisation than the classic stereo setup came about, especially with the development of virtual reality (VR), which needs 3-dimensional 360° audio. The concept of ambisonics was to consider a 360° sphere where audio is coming from with the centre of the sphere, the listening point. This is the connection between spherical harmonics and ambisonics. Spherical harmonics are functions defined on the surface of a sphere. Hence, ambisonics corresponds to a truncated version of spherical harmonics used in audio.

Here, we will not show how to derive these functions. However, a representation of the spherical harmonics can be seen in Fig. 2.1. The most common use of ambisonics is the so-called B-format. This is the sound field decomposition up to the first order of spherical harmonics. This allows one to place or listen to audio with four channels : the omnidirectional channel (W), the front-back channel (X), the left-right channel (Y) and the up-down channel (Z). An illustration can be seen in Fig. 2.2. An application of this would be, for example, when recording a sound field with a microphone capable of producing B-Format ambisonics, if a source is 90° on the xy plane to the microphone, then the X and Z channels should not pick up anything from that source. Ambisonics also work sound field reproduction. Given a setup with four speakers regularly spaced around a listening point, ambisonics can recreate the sound field with spatial localisation of the source in 2D. If we have an audio source S and we want to make it seem as if it is coming from an azimuth angle θ and an elevation angle $\phi = 0$, the corresponding W, X, Y and Z channels are created following :

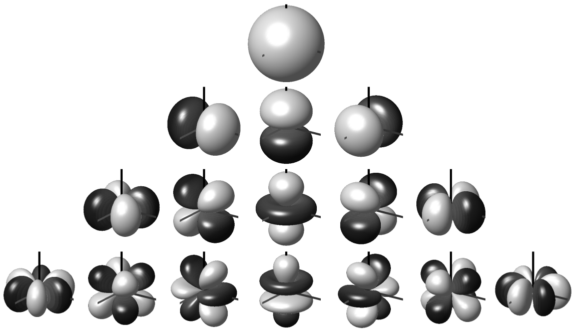


Figure 2.1. An illustration of the spherical harmonics, up to the 3rd degree. The lobes with a lighter shade have opposite polarity compared to those with a darker shade. Image taken from [ZF19].

$$W = \frac{1}{\sqrt{2}} \cdot S, \tag{2.8a}$$

$$X = S \cdot cos(\theta) \cdot cos(\phi), \tag{2.8b}$$

$$Y = S \cdot sin(\theta) \cdot cos(\phi), \tag{2.8c}$$

$$Z = S \cdot sin(\phi). \tag{2.8d}$$

Ambisonics is not limited to B-Format. Given microphones with more capsules or speaker setups with more speakers, we can encode (record) or decode (play) in higher

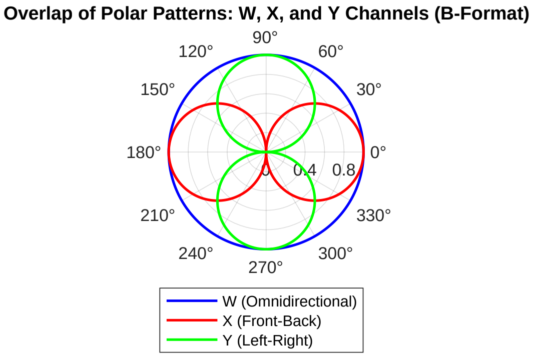


Figure 2.2. An illustration of the B-Format channels W, X and Y. The last channel, Z, would be perpendicular to the X and Y channels.

orders of spherical harmonics. These higher orders allow us to have greater spatial selectivity, as shown in Fig. 2.3 represented by the thickness of the lobe.

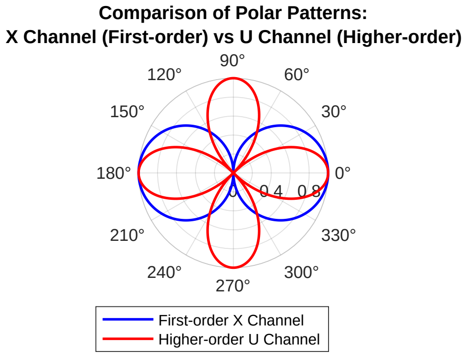


Figure 2.3. First and second order ambisonics in the 2D plane.

Ambisonics can decompose an omnidirectional signal in its spatial components, allowing us to create different directional patterns. For example, we can create a cardioid pattern with a linear combination of the W channel and the X channel, as following :

$$Cardioid = \frac{W+X}{2}. \tag{2.9}$$

CHAPTER 3

PROPOSED METHOD

This chapter describes the modifications used to improve the algorithm's performance.

A way of improving the proposed method is to use directional signals steered towards the SOI instead of omnidirectional signals to estimate the post-filter. The hypothesis is that there should be fewer unwanted signals, such as less influence of the interfering sources and less diffuse reverberation and early reflections captured by the directional signals outside the SOI. In this study, only signals of the first order were used to measure the influence of directional signals on the algorithm's performance. An illustration of this can be seen in Fig. 3.1. We can see the polar patterns are all steered towards the SOI. This idea aims to eliminate interfering sources from other locations. In this example, we can see that microphones 1 and 2 might be unusable due to the location of the interfering source ; however, the other microphones should suppress the interfering source as it is not in the path of the other microphones.

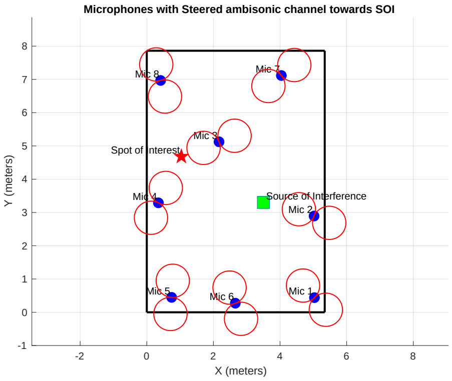


Figure 3.1. Illustration of the steered signals towards a SOI with first order ambisonics. The star represents the SOI, the blue dots represent the location of the microphones and the green square an interfering source. The polar patters represent the steered ambisonic signal.

3.1 Criteria for choosing pair-wise filters and microphone selection

During this study, the importance of which microphone the algorithm has access to overall and which microphone pair it uses at each time step to estimate the individual post-filters was noticed. This relates to the criterion we use as it discriminates whether or not to use a particular microphone pair at each time step. In the Wirler et al. paper, the criterion given is Eq. 2.6. The 1 in the Eq. 2.6, discriminated against microphone pairs with less energy than the average of all pairs. This value of 1 can be refined to achieve greater separation. We can also prohibit the algorithm from using some of the microphones. For example, during a heuristic search for the best microphones to use, we noticed that excluding microphone 5 in most cases seemed to make the separation better.

However, in both cases, which microphone to use or the value for the criterion, the optimal choice for these depends on the situation, whether we are targeting an SOI with a male or female voice or the value of the room's RT60. This means that to get the optimal values in a specific situation, a new heuristic search has to be applied.

Performing a heuristic search every time is time-consuming and not optimal. Elaborating a more refined criterion to correctly find the optimal microphone pair at each time bin autonomously would be the best way to improve this algorithm. The criterion we propose to overcome this issue is what we call the "harmonic structure" criterion. This criterion is intended for signals with harmonic structure within the SOI, such as speech or an instrument.

The central assumption of the method proposed in the original manuscript is that all frequencies coming from the SOI should have a coherence close to 1 in an ideal scenario. Therefore, if a signal with a harmonic structure is present in the SOI, the fundamental frequency of all harmonics should have coherence close to 1. Furthermore, different frequencies can have high coherence at different positions in the room. However, only the SOI should have all of the frequencies with high coherence, as seen in Fig. 3.4.

The proposed method takes advantage of this by applying a pitch-detection algorithm that works on the harmonic structure of the signal. The pitch-detection algorithm used in our method is the method proposed by A.P.Klapuri et al. [Kla03].

The first implementation of the proposed method estimated fundamental frequencies in all of the $\Phi_{pq}(k,n)$. Then, a k-means algorithm would be performed on the estimated fundamental frequencies with k being the number of speakers (with harmonic structure) in the room. The assumption here is that the most commonly found fundamental frequency should be that of the speaker in the SOI. We would then select the $\Phi_{pq}(k,n)$ for the combination scheme that would belong to the most significant clusters. This processing is done for each time bin n.

This first approach failed to detect the fundamental frequency because the frequencies that were not present in either speech would have high coherence and make

$\Phi_{pq}(k, n)$ too noisy to correctly detect the actual fundamental frequency, as seen in Fig. 3.2.

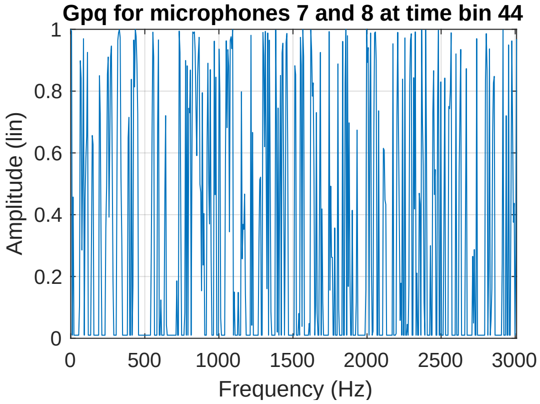


Figure 3.2. Values of the variable G_{pq} for p = 7 and q = 8 at the time bin n = 44. This can be seen as the FFT of a signal that should have harmonic structure.

To resolve this issue, a threshold is applied to the real part of the non-normalized phase-shifted cross-spectrum. We call this quantity $\Phi_{pqpitch}$. This is written as :

$$\Phi_{pqpitch} = \mathrm{Re}\{X_p X_q^* e^{j\omega(k)\tau_{pq}}\}. \tag{3.1}$$

As we can see in Fig. 3.3, the application of this technique leads to less noisy estimations and therefore, the pitch should be easier to correctly estimate for the algorithm.

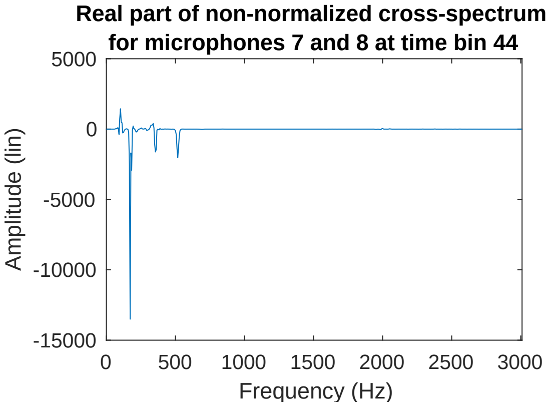


Figure 3.3. Values of the variable $\Phi_{pqpitch}$ for p = 7 and q = 8 at the time bin n = 44.

However, a few problems are revealed in this figure. The pitch detection algorithm is intended to work on FFTs. Therefore, we want something comparable to an FFT. As shown in Fig. 3.3, we have negative values in our coherence, which are impossible in an FFT ; we need positive values. In the original paper, Wirler et al. took the maximum between the real part of the Φ_{pq} and a hyperparameter λ_{pq}, as seen in Eq. 2.5. This is for a limited half-wave rectification. So in this case, we could also do the same here. However, if we look closer at our coherence values, we notice that the negative minima correspond to the harmonics of the desired source. We have some hypotheses as to why we have negative coherence values for frequencies from the SOI. The first one is that the phase rectification might be too imprecise. Another possibility for these negative values is the simplicity of the signal model. We can see in Eq. 2.1 that the signal model we use only considers direct sound. Reverberation is not taken into account here. We could use more elaborate models that consider the room's effect, affecting the phase shift we apply. We would then achieve better results, especially when testing in higher reverberation settings. In any case, the solution to this problem we used was to apply the absolute function to the estimated coherence as :

$$\Phi_{pqpitch} = \text{abs}\{\mathfrak{Re}\{X_p X_q^* e^{j\omega(k)\tau_{pq}}\}\}. \tag{3.2}$$

We apply some additional processing, such as more flooring, to make the estimation more robust against noise, and normalizing is done to make the estimation independent of the gain of the microphone. We therefore add these 2 steps, Eq. 3.3 and Eq. 3.4, before estimating pitch :

$$\Phi_{pqpitch} = \begin{cases} \Phi_{pqpitch} & \text{if } \Phi_{pqpitch} > 50 \\ 0 & \text{otherwise} \end{cases}, \tag{3.3}$$

$$\Phi_{pqpitch} = \frac{\Phi_{pqpitch}}{\max(\text{abs}\{\Phi_{pqpitch}\})}. \tag{3.4}$$

The pitch estimation is then applied on $\Phi_{pqpitch}$. This process gives us an estimated pitch for every combination of p and q. We then apply a k-means clustering with k equal to the number of sources with harmonic structure in the room.

Once we have our clusters, we select the Φ_{pq} for the combination scheme that has a corresponding $\Phi_{pqpitch}$ with an estimated pitch inside of the most significant cluster.

Another possible use of this idea is to use one of the values calculated during the pitch estimation algorithm. To estimate pitch, this algorithm adds or multiplies the values of the harmonics together and chooses the fundamental frequency with the corresponding highest value. This value corresponds to the amalgamation of the energy in the harmonics. The original criterion was based on the energy in the cross product of

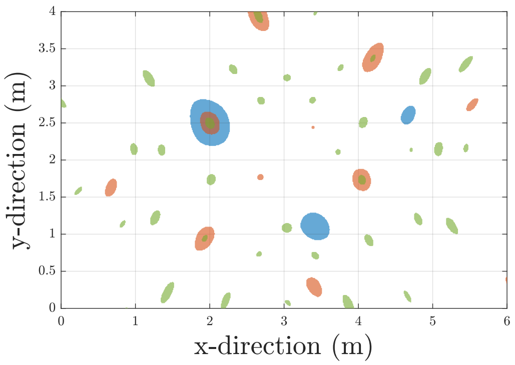


Figure 3.4. Graph illustration the regions where different frequencies have a coherence higher than 0.7 given a specific microphone layout. The blue regions represent the 125 Hz frequency, the red regions represent the 250 Hz frequency, and the green region represents the 500 Hz frequency [WP24].

$X_p X_q^*$. Another idea explored during this investigation was to use a criterion based on the energy in the harmonics instead of the whole signal. However, this thesis does not present those results as they were not as fruitful as the final proposed method.

Lastly, the idea of taking the overlap of both criteria as a new criterion was explored. The final criterion is the Φ_{pq} with an associated estimated pitch in the largest cluster that also satisfies an energy requirement.

3.2 Additional processing for better extraction

Another improvement proposed, in the case of harmonic signals inside of the SOI, was to create a filter that boosts the fundamental frequencies and harmonics that were estimated. The filter implemented was a naive implementation, as can be seen in Fig. 3.5. This addition to the algorithm is not specific to the rest of the algorithm, so any harmonic signal enhancement algorithm that works on FFTs should work here. This filter is synthesized by creating a Gaussian function centred around the fundamental frequency and its harmonics. This is then floored at an arbitrary value ; in this case, 0.2 is used. A slightly more intelligent filter design would consider the source's inharmonicity. The pitch estimation used here considers this and finds the harmonics with the inharmonicity.

Furthermore, it was noticed that, in general, the more restrictive the criterion, the better the results. This means fewer Φ_{pq} in the combination scheme. It could be so re-

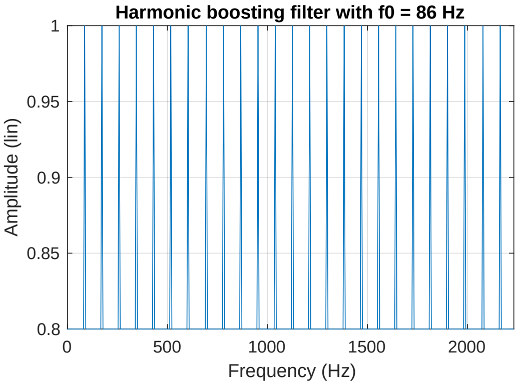


Figure 3.5. Naive filter to boost harmonic signals found in the SOI.

strictive at some time bins that no Φ_{pq} passed the criterion. In this case, we would set G_{pq} to a floor value, as :

$$G_{pq}(k,n) = \lambda. \tag{3.5}$$

The separated SOI signal is then calculated as described in Eq. 2.7.

CHAPTER 4

METHODOLOGY

In the following section, we will first present the data acquisition proposal. We will then explain the simulation setting and parameters. Afterwards, we will explain how the directional signal is steered towards the SOI. Finally, we will present the different metrics and signals we use for our evaluation.

4.1 Dataset acquisitions

The first step needed to evaluate the method in real-world scenarios is to gather data under different conditions. To do so, the Acoustics lab at Aalto University has a room with variable acoustics. This is the Arni room, as shown in [uni23]. This room has 55 individually controllable panels on the walls that can be opened (absorptive) or closed (reflective) to change the amount of absorption of the sound waves by the walls. The data was acquired for three different room acoustic configurations. The RT60 values are 1.6 s, 0.52 s, and 0.33 s from most reverberant to least reverberant, respectively. These were calculated by following the ISO 3382 guidelines [Int09], which say to "take averages over the six one-third-octave bands from 400 Hz to 1 250 Hz".

To record the signals, the Digital Audio Workstation (DAW) used was Reaper [Rea23]. This DAW allows us to record many channels at a time. Indeed, we needed to record 152 channels, as there were 8 Zylias during recording, with 19 capsules each.

To get these directional signals in natural conditions, we must use microphones capable of recording 3D audio. We used spherical microphone arrays, the Zylia ZM-1S microphones, as seen in Fig. 4.1 [Zyl23]. The 19 capsules of the Zylia allow an ambisonic recording up to the third order. To get the ambisonic signals from the 19 channels, Zylia provides an application that converts the 19 channels to the 3rd-order ambisonic signals. Ambisonics has multiple conventions for channel order and normalization. Here, we used the ACN channel order convention and the N3D normalization convention. Using this convention is the standard when using ambisonics.

The layout of the recording can be seen in Fig. 4.2. This layout was chosen to have the microphones well distributed in the room and the speakers to have different facing directions.

The gathered impulse responses are convolved to create the acoustic situations with an anechoic recording of the different speakers. This would be equivalent to play-

Figure 4.1. Image of a Zylia ZM-1S microphone. Image taken from [Gui23]

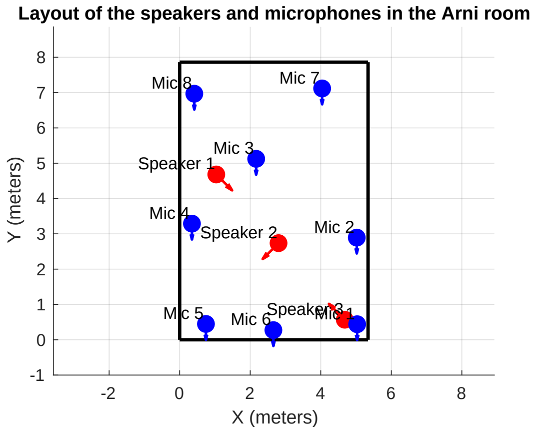


Figure 4.2. Layout of the Zylia ZM-1S microphones and the Genelec 830A speakers in the Arni room. In red are the speaker locations, and in blue are the microphone locations. The arrows represent the look direction of the speakers and microphones.

ing those speeches from the speakers directly. However, the impulse response method makes it easier to make any speaker seem to come from any loudspeaker. To get the impulse response, we applied the exponential sine-sweep method [Far00].

By the end of the recording, we had the impulse response for every microphone-speaker pair for three different amounts of reverberation.

More recordings were made in different situations, such as a moving source alone and with background noise, recorded with Opti-track, but this data was not used in this master's thesis and will be used for ongoing research.

4.2 Directional signals

Simulations were conducted to test the algorithm with the different signals and criteria to gather a first indication of the performance. The signals tested were a baseline signal, the first-order ambisonic signal and a cardioid signal created from that baseline signal. The baseline signal would be either the omnidirectional signal or the DaS signal. From this, we can evaluate the method's effectiveness for different positions of an interfering source. Unlike the evaluation with the real signal in which we apply the post filter to the microphone 3, here we apply the post filter to the closest microphone to the source. The interfering source moves through a regularly spaced grid of 18 positions distant 1 m from each other. The following tables show the parameters used for this simulation : Tab. 4.1 and 4.2.

Parameter	**Value**
Sampling Frequency	44100 Hz
Window length	4096 Samples
Window overlap	50 %
Window	square-root Hann
Reverberation time RT_{60} by octave band	[0.4, 0.3, 0.3, 0.25, 0.25, 0.2, 0.2, 0.2] s
Octave band	[63, 125, 250, 500, 1000, 2000, 4000, 8000] Hz
Room Size (x y z)	[6, 4, 2.5]
Source position (x y z)	[2, 2.5, 1.5]
Temporal smoothing factor	0.3
$(lambda_{pq}$/lambda	0.05
F_0 search range	[50 500] Hz
Boosting filter floor	0.2

Table 4.1. Hyperparameters used for the simulation

x	**y**	**z**
0.2500	0.2500	1.5000
5.7500	0.2500	1.5000
5.7500	3.7500	1.5000
0.2500	3.7500	1.5000
3.0000	0.2500	1.5000
2.7500	3.7500	1.5000

Table 4.2. Position of the microphones in the simulated room

As for the real signals, once we have obtained the directional signals, we need to steer them towards the direction of the desired speaker. Ambisonic signals can be rotated virtually to aim towards a specific location. To do so, we must first calculate the yaw and pitch angles between the X channel of the ambisonic signals and the speaker in the SOI. The yaw-pitch-roll convention refers to the intrinsic rotations as opposed to the Euler angles, which correspond to an extrinsic rotation. During the layout of the microphones in the Arni room, the microphones were all placed so that the X channel would face towards the negative Y-axis, which would be towards the bottom of the image in Fig. 4.2, which corresponds to the vector,

$$\mathrm{x}_{look} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}.$$

To calculate the yaw and pitch angle between the look direction of the microphone and the SOI with the targeted speaker, we first calculate the vector defined by the difference between the position of the SOI and the current microphone, as shown in Eq. 4.1, with

$$\mathrm{x}_{\mathrm{SOI}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

the posisiton of the SOI and

$$\mathrm{x}_{spk} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

the position of the speaker.

$$\mathrm{x}_{\mathrm{d}} = \mathrm{x}_{\mathrm{SOI}} - \mathrm{x}_{\mathrm{spk}} \tag{4.1}$$

We then project the vector onto the XY plane to calculate the pitch angle and the YZ plane to calculate the yaw angle. We calculate the angles using the following :

$$\theta = \arctan\left(\frac{||\mathrm{x}_{\mathrm{d}} \times \mathrm{x}_{\mathrm{look}}||}{\mathrm{x}_{\mathrm{d}} \cdot \mathrm{x}_{\mathrm{look}}}\right).$$

Once these angles are found, we use the Matlab functions developed by Politis et al. [Pol16], which rotate ambisonic signals with the N3D normalization. Once the rotations are performed, the X channel of the current microphone will be steered towards the SOI as in Fig. 3.1.

4.3 Metrics for evaluation

For the objective evaluation, three different metrics are used : SIR, SDR and SAR. During this master, greater attention was given to bettering the SIR, which roughly corresponds to the quality of the separation between the competing sources, whilst the

SDR and SAR indicate the audio quality of the output. To calculate these values, we assume that the algorithm's output is the sum of the target, interferer, noise, and artefact signals as shown in Eq. 4.2.

$$s(t) = s_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t) \tag{4.2}$$

These metrics are calculated after the following equation 4.3

$$\text{SDR} := 10\log_{10}\left(\frac{||s_{\text{dist}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2}\right) \tag{4.3a}$$

$$\text{SIR} := 10\log_{10}\left(\frac{||s_{\text{dist}}||^2}{||e_{\text{interf}}||^2}\right) \tag{4.3b}$$

$$\text{SAR} := 10\log_{10}\left(\frac{||s_{\text{dist}} + e_{\text{interf}} + e_{\text{noise}}||^2}{||e_{\text{artif}}||^2}\right) \tag{4.3c}$$

BSS eval Matlab Toolbox [FGV05] is a Matlab Toolbox that allows us to calculate these values easily. We also used the **sig_align** function from the same toolbox. This function aligns a target signal with a reference signal. In our case, we align the algorithm's output with the reference signal we want to isolate.

These evaluations were evaluated on different mixtures. The different cases are as follows :

- 1 Danish male voice, 1 English female voice with Danish male voice in SOI
- 1 Danish male voice, 1 English female voice with English female voice in SOI
- 1 Danish male voice, 1 English female, 1 Danish female voice with Danish male voice in SOI
- 1 Danish male voice, 1 English female, 1 Danish female voice with English female voice in SOI

The voices were kept at the same position for all tests, as in that each individual voice was convoluted with the same speaker for each situation. The male Danish voice was "placed" at the location of the first speaker, the female English voice was "placed" at the location of the second speaker and the female Danish voice was "placed" at the third speaker's location. First, the second and third speaker refers to the corresponding speaker numbers as seen in Fig. 4.2.

The original algorithm was evaluated using real signals, including omnidirectional and directional signals derived from first-order ambisonics recordings. This evaluation employed four distinct voice configurations across various combination schemes and signal types, as outlined in the original study. To maintain consistency and ensure the audience's familiarity with the material, the notation utilized in the original paper will be consistently adopted in this paper to denote the different types of signals. An explanation follows :

- O : This refers to the original signal from the microphone without any processing done to it. Improvements in other signals should be compared to this baseline.
- DS : This refers to the beamformed signal created with a typical delay-and-sum beamformer performed on the distributed microphones. This is used as a naive algorithm we can compare our contributions to.
- OPF : This refers to the signal created by applying the synthesised post-filter using the multiplicative combination scheme on the O signal.
- OPF2 : This refers to the signal created by applying the synthesised post-filter using the additive combination scheme on the O signal.
- DSPF : This refers to the signal created by applying the synthesised post-filter using the multiplicative combination scheme on the DS signal.
- DSPF2 : This refers to the signal created by applying the synthesised post-filter using the additive combination scheme on the DS signal.

For the other part of the paper (harmonic structure criterion), we evaluated the proposed criterion and other additions using only the best-performing DSPF signal.

As for any algorithm, hyperparameters must be set. The chosen parameters for the evaluation are shown in the following Tab. 4.3

Parameter	**Value**
Sampling Frequency	44100 Hz
Window length	4096 Samples
Window overlap	50 %
Window	square-root Hann
Reverberation time RT_{60}	[0.33, 0.52, 1.6] s
λ_{pq}	0.01
λ	0.001
Temporal smoothing factor	0.4
Number of estimated harmonics	3
F_0 search range	[50 500] Hz
Boosting filter floor	0.2

Table 4.3. Hyperparameters used for the evaluation in real conditions.

CHAPTER 5

RESULTS AND DISCUSSION

In this section, we will first present the simulation results, and then the results for evaluating the proposed method on omnidirectional signals. Afterwards, we will present the results using directional signals, and finally, the results obtained through our work on the criterion will be presented.

In the following sections, "speaker 1" refers to the position of speaker 1 and "speaker 2" refers to the position of speaker 2 in Fig. 4.2. The Danish male voice is emitted from speaker 1, and the English female voice from speaker 2. We must remember that the term "aiming for speaker x" does not mean we are targeting a specific signal but the location in which that speaker is located.

The tables presented in sections 5.2 to 5.4 are calculated by averaging over different quantities. The left sub-table is calculated by averaging the difference in SIR or SDR over the different speakers, whilst the right sub-table is calculated by averaging the difference in SIR or SDR over the different reverberation times. The spk_conf number refers to the number seen in List 4.3.

In this section, we call the original criterion the "energy" criterion, which discriminates against microphone pairs with less cross-spectrum energy than the average of all microphone pairs.

Only the SDR and SIR values are shown here because the different methods and configurations all have very similar SAR values.

5.1 Simulation results

To give ourselves an indication of the effectiveness of the proposed methods before being evaluated on real scenarios, we conducted simulations as explained in Section. 4.2.

As shown in the violin plots figs. 5.1 to 5.3, when using only the proposed harmonic criterion, the metrics, whether in SIR or SDR, do not seem to change much compared to the original criterion. However, the conjunction of the proposed criterion and the original criterion with the boosting filter does perform better than the original criterion on average. This reinforces the observation that the more constrictive the criterion, the better the results in terms of SIR. The cardioid and dipole also perform better than the baseline signal, whether when using the omnidirectional or DaS signals. As shown

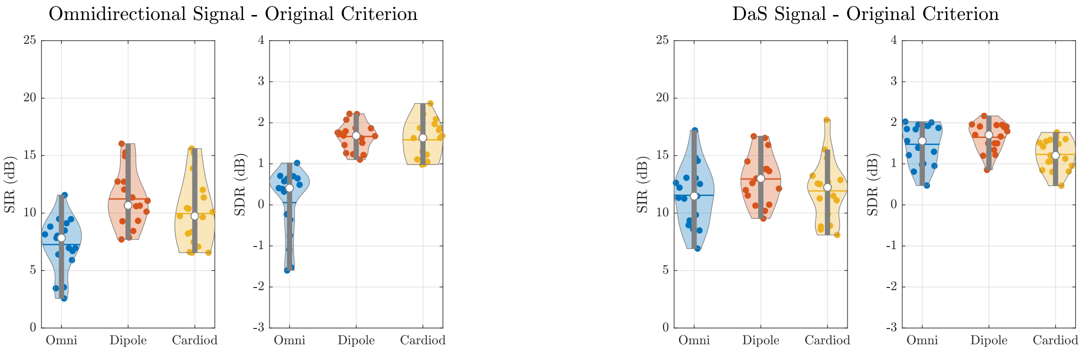


Figure 5.1. Results of the simulation using the omnidirectional signal and the DaS signal when using the original criterion. The colored dots represent individual simulation results, whilst the white dot represents the mean for that method.

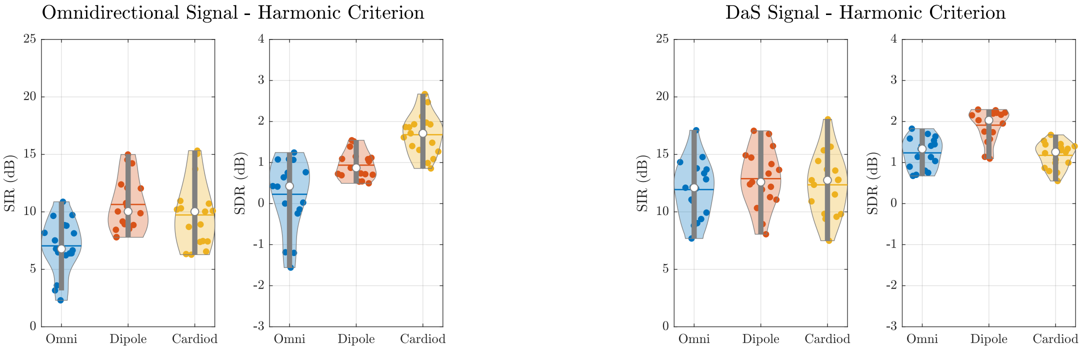


Figure 5.2. Results of the simulation using the omnidirectional signal and the DaS signal as a baseline to create the dipole and cardioid signals using only the harmonic criterion. The colored dots represent individual simulation results, whilst the white dot represents the mean for that method.

in all three figs. 5.1 to 5.3, the dipole performs slightly better than the cardioid in terms of SIR. The SIR is the primary metric we are focusing on during this thesis, so we chose only to use the dipole during the tests with real scenarios. We can also notice a slight decrease in SDR when using the conjunction criterion with the boosting filter. This is thought to be caused by the harshness of the boosting filter.

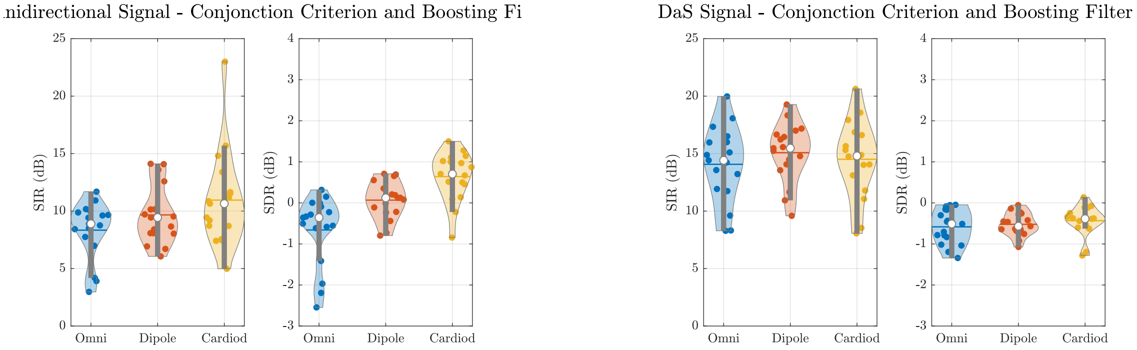
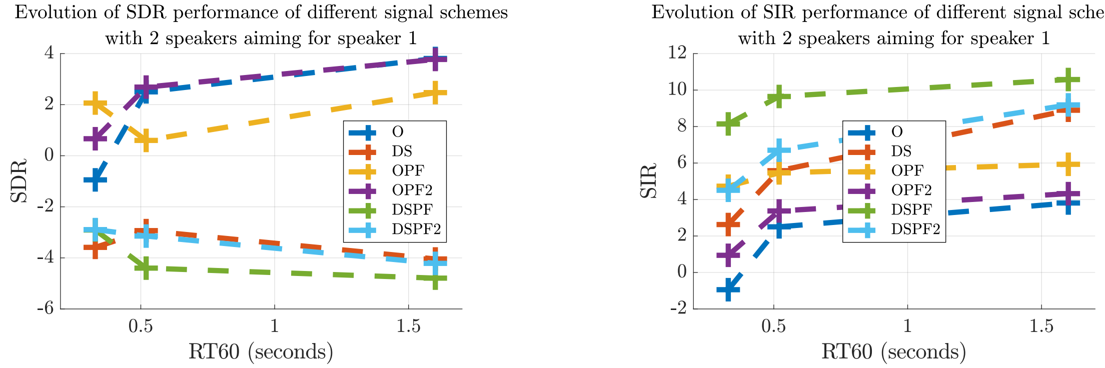


Figure 5.3. Results of the simulation using the omnidirectional signal and the DaS signal as a baseline to create the dipole and cardioid signals when using the conjunction of the original criterion and the harmonic criterion with the boosting filter. The colored dots represent individual simulation results, whilst the white dot represents the mean for that method.

5.2 Omnidirectional signals results

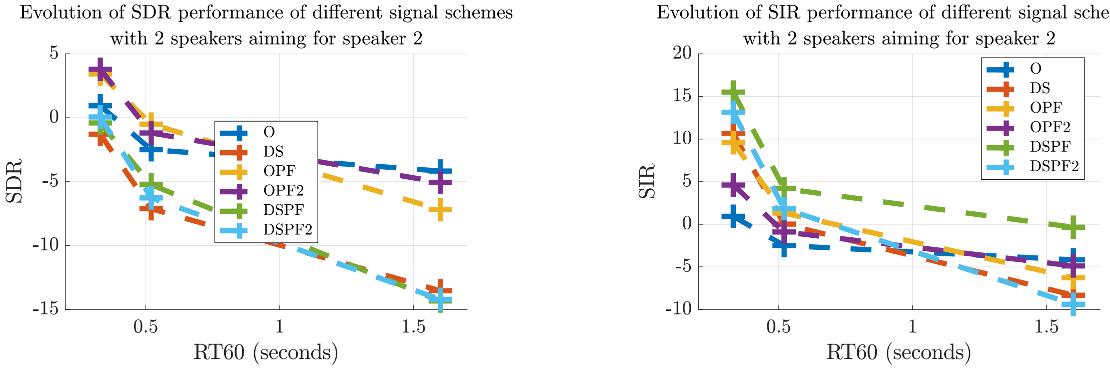


(a) SDR (dB)

(b) SIR (dB)

Figure 5.4. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains a male speaker; there are 2 voices in total.

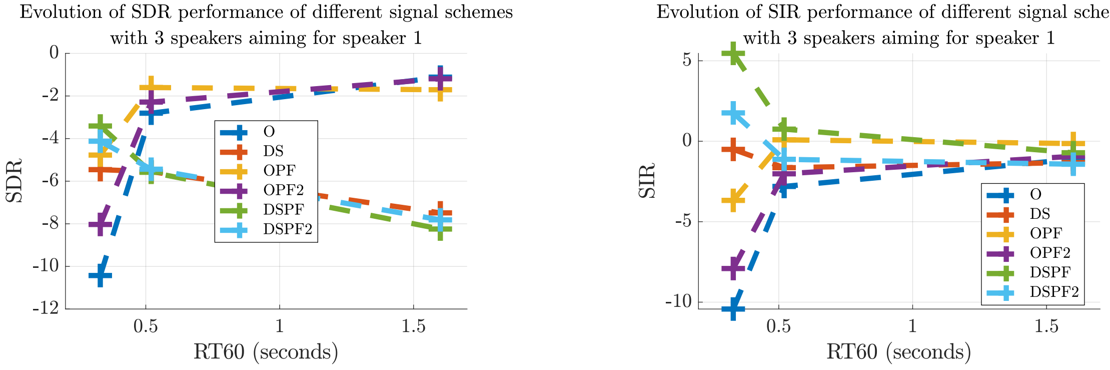
With the results obtained when evaluating the original method on real omnidirectional signals, we can confirm that, in terms of SIR, the DSPF method is the best-performing method, as shown in figs. 5.4b, 5.5b, 5.6b and 5.7b, which is also what is observed in the original paper. However, in terms of SDR, the OPF and OPF2 methods outperform the other methods, which was not observed in the source paper as shown in figs. 5.4a, 5.5a, 5.6a and 5.7a.



(a) SDR (dB)

(b) SIR (dB)

Figure 5.5. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains an English female speaker; there are 2 voices in total.



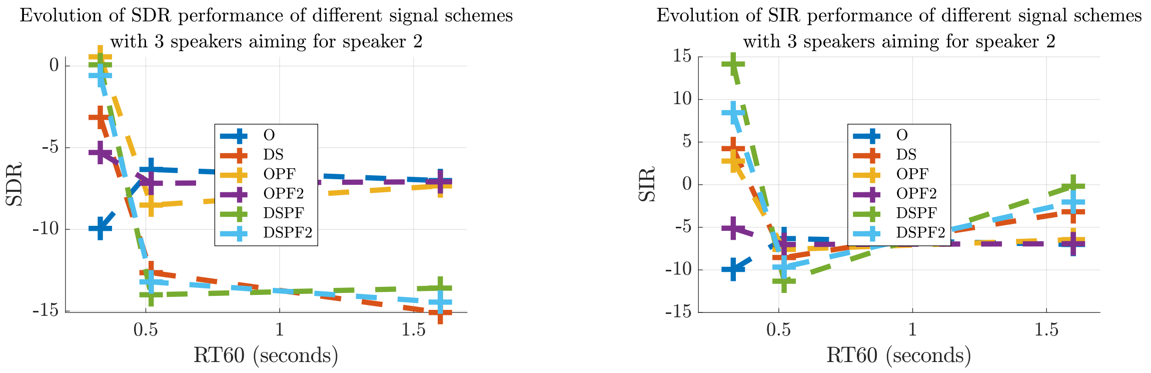
(a) SDR (dB)

(b) SIR (dB)

Figure 5.6. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains the male speaker, there are 3 voices total.

We can see that in almost all configurations, whether method, number of speakers, speaker in the SOI or reverberation time, the proposed methods have a better SIR than the original mixture, as we can see in the Tab. 5.1. However, we can find 1 configuration in which this is not the case. The configuration with 3 speakers aiming for speaker 2 with an RT60 of 0.52 s. These results are very good, as even with higher reverberation, we can achieve close to 4 dB of gain in SIR, as shown in Tab. 5.3a. Even when testing with more speakers, we can achieve high gain in SIR, as shown in Tab. 5.3b. Averaging over all configurations and reverb time, we observe an average difference gain of 7.7 dB.

The low results in terms of SDR, as shown in Tab. 5.2, can be explained by the low



(a) SDR (dB)

(b) SIR (dB)

Figure 5.7. Evolution of the objective metrics versus the RT60 values of the room for the omnidirectional signals. The evolution is plotted for all of the tested signals. The SOI contains the English female speaker, there are 3 voices total.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	13.5 dB	6.0 dB	3.7 dB

(a) Averaged over the different speaker configurations

spk_conf	1	2	3	4
Avg diff	7.7 dB	8.3 dB	6.6 dB	8.3 dB

(b) Averaged over different reverberation times.

Table 5.1. Average difference in SIR between the original mixture and the DSPF method.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	0.6 dB	-3.8 dB	-9.0 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	-5.8 dB	-4.7 dB	-1.0 dB	-4.7 dB

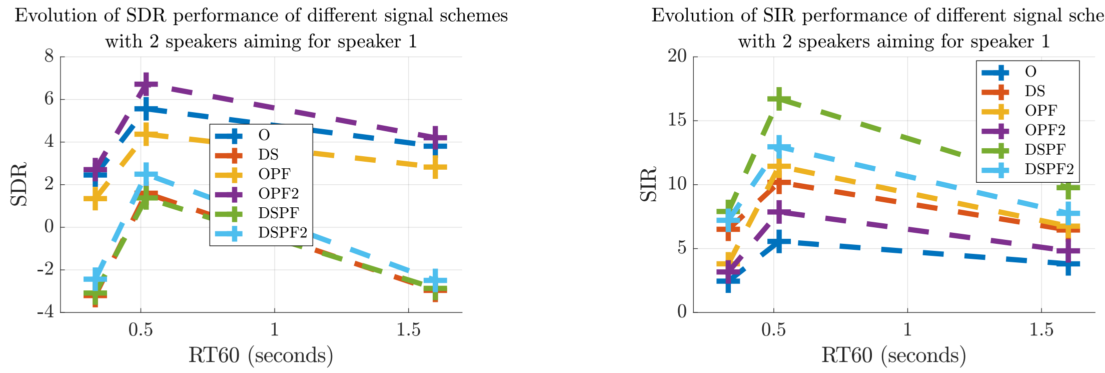
(b) Averaged over different reverberation times.

Table 5.2. Average difference in SDR between the original mixture and the DSPF method.

floor values used for the masking. With higher spectral floor values, the SDR values are expected to improve as the harshness of the filtering is diminished, letting more of the unwanted signal pass through the filter; however, the SIR values should also decrease at the same time.

5.3 Directional signal results

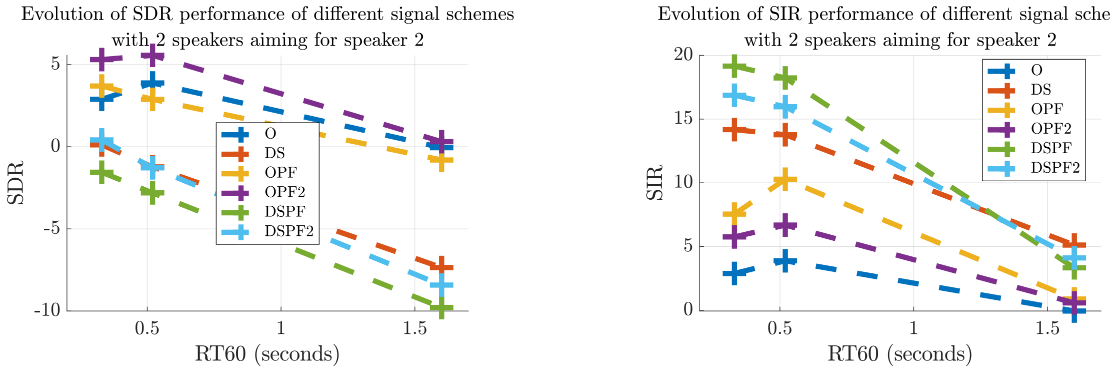
As we can see, in general, the first-order ambisonic signals show better results than we using the omnidirectional signals, especially for acoustic scenes with moder-



(a) SDR (dB)

(b) SIR (dB)

Figure 5.8. Evolution of the objective metrics versus the RT60 values of the room for the 1st ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the male speaker, there are 2 voices total.

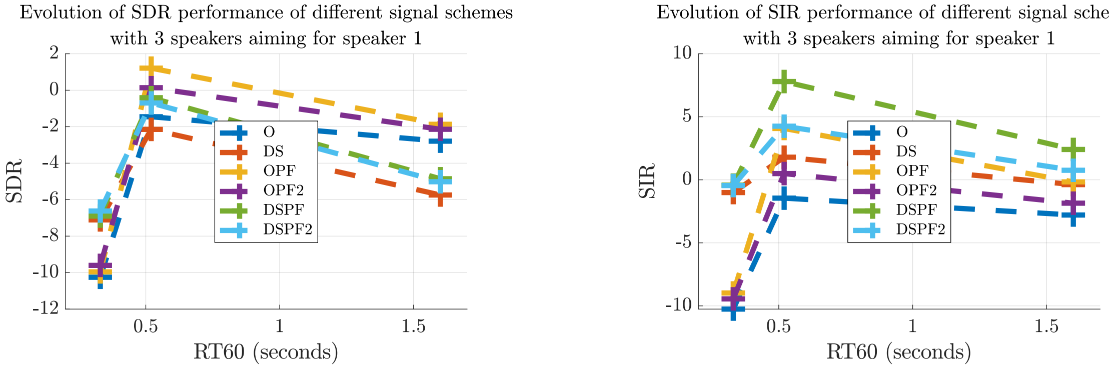


(a) SDR (dB)

(b) SIR (dB)

Figure 5.9. Evolution of the objective metrics versus the RT60 values of the room for the 1st ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the English female speaker, there are 2 voices total.

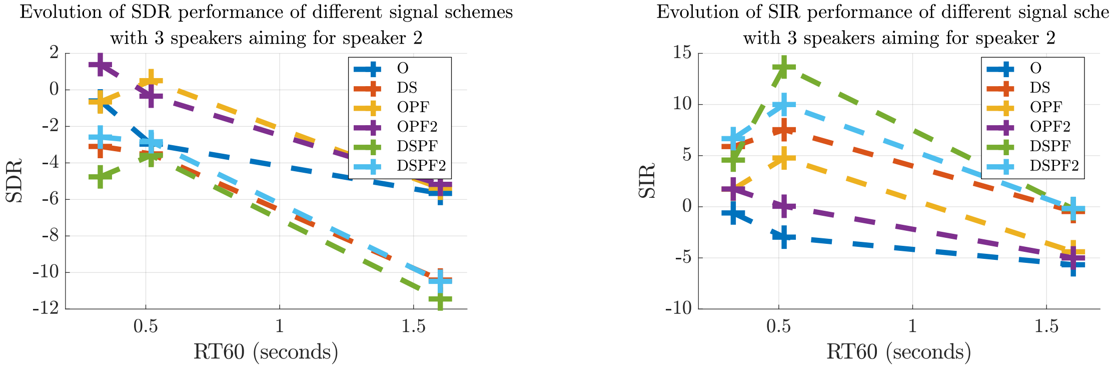
ate amounts of reverberation, as seen in the figs. 5.8 to 5.11. As shown in Tab. 5.3, the SIR for the directional signals improve compared to the omnidirectional signals. The -3 dB in Tab. 5.3a shows worse performance for the directional signals than the omnidirectional signals in acoustically dry environments. This can possibly be explained by poor steering towards the SOI, which would be less critical in situations with higher reverberation. An amplification of the interfering source could also explain it. Indeed, for the microphones aligned with the interfering source and the SOI, the amplification of the interfering source could worsen the overall quality of the separation. However, we can say that overall, using 1st-order ambisonic signals steered towards the SOI improves the



(a) SDR (dB)

(b) SIR (dB)

Figure 5.10. Evolution of the objective metrics versus the RT60 values of the room for the 1st ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the male speaker, there are 3 voices total.



(a) SDR (dB)

(b) SIR (dB)

Figure 5.11. Evolution of the objective metrics versus the RT60 values of the room for the 1st-order ambisonic signals steered towards the SOI. The evolution is plotted for all of the tested signals. The SOI contains the English female speaker, there are 3 voices total.

quality of the separation done by the algorithm, as shown in tables 5.3 and 5.4. Calculating the average difference overall in all configurations and RT60 values, we can observe an increase of 3.9 dB in SIR. Doing the same analysis for the SDR, we can see the values in this Tab. 5.4. These show similar gains as for the SIR. The overall average gain in SDR is 2.2 dB. These results reflect the differences observed during the simulation between the directional and omnidirectional signals, even showing better improvements than in the simulation. With these results, we can conclude that using these directional signals improves the algorithm's performance, thus confirming the original hypothesis for this internship to be true.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	-3 dB	13.3 dB	1.5 dB

(a) Averaged over the different speaker configurations

spk_conf	1	2	3	4
Avg diff	2 dB	7.1 dB	1.4 dB	5.1 dB

(b) Averaged over different reverberation time

Table 5.3. Average difference in SIR between using omnidirectional signals and 1st order ambisonics.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	-2.4 dB	5.9 dB	3.0 dB

(a) Averaged over the different speaker configurations.

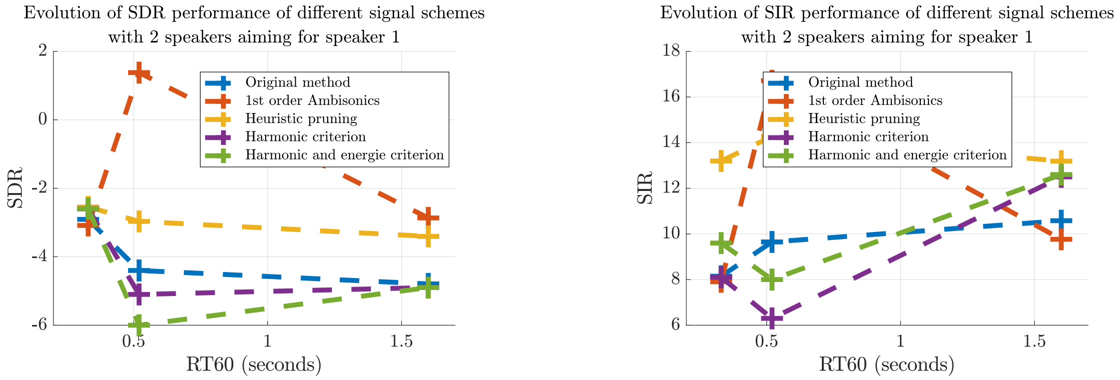
spk_conf	1	2	3	4
Avg diff	2.5 dB	2.9 dB	1.7 dB	2.8 dB

(b) Averaged over different reverberation time.

Table 5.4. Average difference in SDR between using omnidirectional signals and 1st order ambisonics.

5.4 Criteria for choosing pair-wise filters and microphone selection results

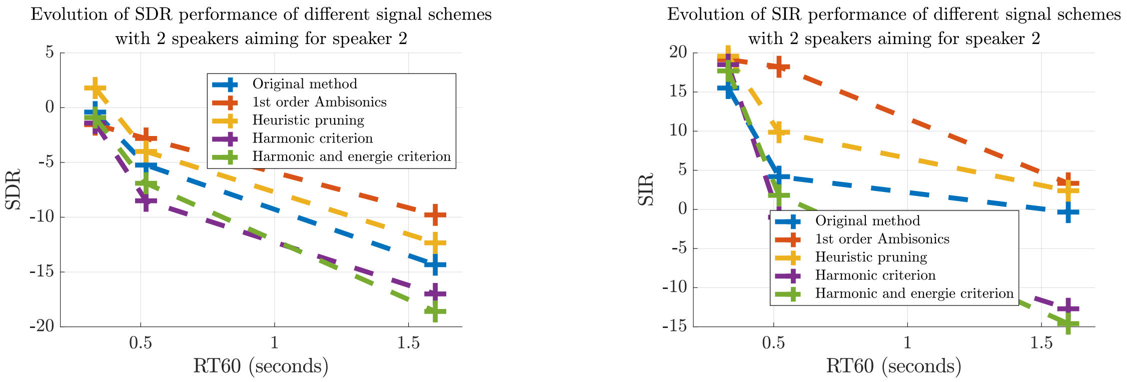
For this section, the heuristic pruning and the harmonic structure criterion were evaluated only using the DSPF signal, as it demonstrated the best results in terms of SIR in the previous sections.



(a) SDR (dB)

(b) SIR (dB)

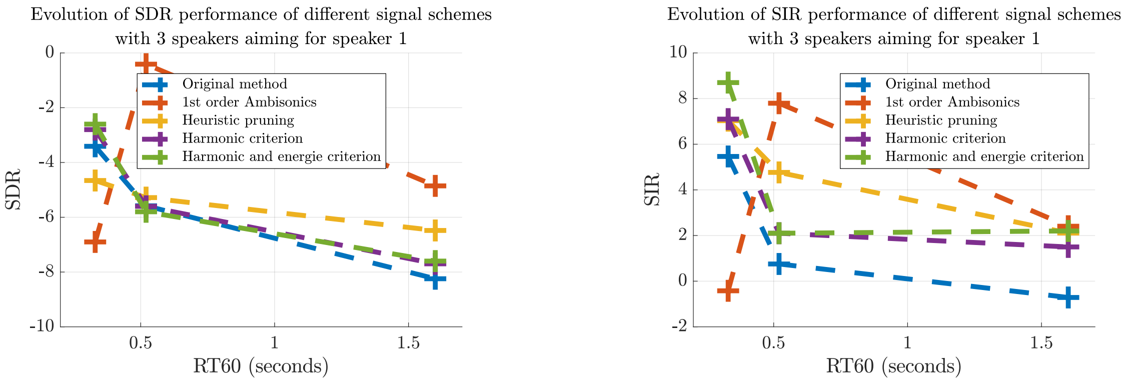
Figure 5.12. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the male speaker, there are 2 voices total.



(a) SDR (dB)

(b) SIR (dB)

Figure 5.13. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the English female speaker, there are 2 voices total.

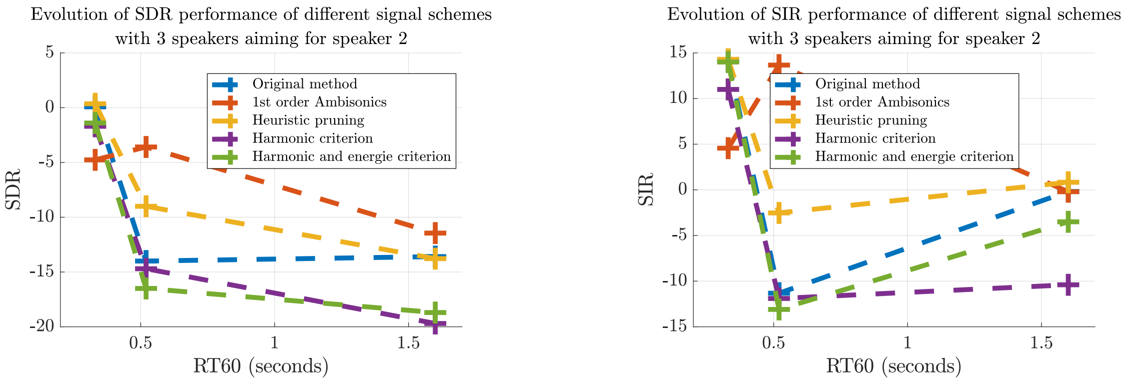


(a) SDR (dB)

(b) SIR (dB)

Figure 5.14. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the male speaker, there are 3 voices total.

Out of the two different criteria tested, heuristic and harmonic, the heuristic criterion performs the best. However, in drier settings, the harmonic criterion can be comparable with the heuristic pruning criterion, mainly if we use the conjunction of both criteria with the boosting filter, as we can see in figs. 5.12 to 5.15. Doing the same analysis as the previous parts, we can see the gains brought by changing the criterion in tables 5.5 and 5.6. The average difference in SIR using the heuristic pruning criterion is 3.6 dB, and 1.2 dB for the SDR. Comparing the results obtained with the directional signals and the heuristic pruning, it seems that, while the heuristic pruning might not



(a) SDR (dB)

(b) SIR (dB)

Figure 5.15. Evolution of the objective metrics versus the RT60 values of the room. Here, we compare the effectiveness of the criterion and the directional signals to the original algorithm. The SOI contains the English female speaker, there are 3 voices total.

show as good results as the directional signals in some instances, the heuristic pruning does seem much more stable over the different configurations.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	2.7 dB	5.8 dB	2.3 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	4.1 dB	4.2 dB	2.8 dB	3.3 dB

(b) Averaged over different reverberation times.

Table 5.5. Average difference in SIR between using the heuristic pruning criterion and the original criterion with omnidirectional signals.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	0.4 dB	2.0 dB	1.2 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	1.1 dB	1.8 dB	0.3 dB	1.7 dB

(b) Averaged over different reverberation times.

Table 5.6. Average difference in SDR between using the heuristic pruning criterion and the original criterion with omnidirectional signals.

As for the harmonic criterion and the conjunction of both criteria with the boosting filter, given the same analysis as the previous cases, the results initially appear less favourable than anticipated. As shown in tables 5.7 and 5.8, in most cases, the proposed criterion performs worse than the original method. However, if we take a closer look at the results, one case seems to be an outlier value, making the average much lower. If we take the same averages without taking into account that case, we can see that the proposed criterion does perform better than the original criterion in most cases, as shown in these tables 5.9 and 5.10. It must also be clear that the harmonic criterion is stochastic as the k-means centroids are initialised randomly, as opposed to the original criterion and the heuristic pruning criterion, which are deterministic. To obtain these results, we did an average of 10 trials. Regarding SIR, we see modest improvements compared to the original criterion. We get an overall average improvement of 0.3 dB when taking out the outlier value. As for the SDR, we get worse results than the original criterion, though this is thought to be caused by the boosting filter applied afterwards. We can see an overall average change of -1.0 dB. These results do not reflect the improvements found in the simulations as shown in Fig. 5.3, where we achieve close to 2 dB in gains in SIR. These results can be explained by the same reasons mentioned before, such as the simplicity of the model used, the incoherence of the signals, or the possible poor steering of the directional patterns. Anecdotally, the perception of the quality of the separation and the SIR values did not match, though an actual perceptual study should be done to confirm these impressions.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	1.7 dB	-1.1 dB	-3.1 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	0.6 dB	-4.8 dB	2.5 dB	-1.7 dB

(b) Averaged over different reverberation.

Table 5.7. Average difference in SIR between using the conjunction of the harmonic criterion and the energy criterion and the original criterion with the boosting filter and the omnidirectional signals.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	-0.2 dB	-1.5 dB	-2.2 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	-0.5 dB	-2.1 dB	0.4 dB	-3.0 dB

(b) Averaged over different reverberation time.

Table 5.8. Average difference in SDR between using the conjunction of the harmonic criterion and the energy criterion and the original criterion with the boosting filter and omnidirectional signals.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	1.7 dB	-1.1 dB	0.5 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	0.6 dB	-0.1 dB	2.5 dB	-1.7 dB

(b) Averaged over different reverberation time.

Table 5.9. Average difference in SIR between using the conjunction of the harmonic criterion and the energy criterion and the original criterion with the boosting filter and omnidirectional signals without the outlier value. The outlier value is replaced by a NaN value.

RT60	0.33 s	0.52 s	1.6 s
Avg diff	-0.2 dB	-1.5 dB	-1.5 dB

(a) Averaged over the different speaker configurations.

spk_conf	1	2	3	4
Avg diff	-0.5 dB	-1.1 dB	0.4 dB	-3.0 dB

(b) Averaged over different reverberation times.

Table 5.10. Average difference in SDR between using the conjunction of the harmonic criterion and the energy criterion, and the original criterion with the boosting filter and omnidirectional signals without the outlier value. The outlier value is replaced by a NaN value.

5.5 Addressing limitations

This research was thought to bring solutions to problems such as the cocktail party ; however, during testing, at most, there were only three concurrent speakers, which is still somewhat different from what the cocktail problem is thought to be. Adding babble background noise would make this study reflect more realistic settings. Furthermore, the automatic criterion proposed only works with harmonic signals. This is an improvement ; however, if we use harmonic signals of a specific target, more conventional algorithms still perform better. The idea here is if a harmonic signal moves into the SOI. Finding a criterion which works for all types of signals would make this algorithm much more robust. Also, as we have addressed in the "Methodology" chapter 4, the signal model used is very simplistic, as shown in Eq. 2.1, therefore the phase-shift is undoubtedly wrong, especially in acoustic settings with high RT60 values.

5.6 Suggestions for further research

In the context of this master thesis, different ideas for improvements were proposed. The first is to use different directional beam patterns. The beginning of this internship was based a lot on the CroPAC algorithm [DP13]. This paper proposes more selective beam patterns with fewer "side lobes". In this master thesis, we evaluated the possible improvements of first-order ambisonics and cardioid beam patterns, which are large and not very selective. An evaluation should be done on whether more selective patterns such as the CroPAC, higher ambisonics, or adaptive beam patterns such as MVDR beamforming improve the algorithm. Another idea that should be investigated is to make more hyperparameters that depend on RT60 values and frequency. For example, making the spectral flooring frequency and RT60 dependent. Something that should have been done during this internship was to sample the speech signals with a much lower sampling frequency. Especially when treating speech signals, sampling at 44.1 kHz is unnecessary. Indeed, the original paper's sampling frequency was 16 kHz. This would particularly help with the harmonic criterion. We would get much greater frequency precision with the same amount of samples for our FFT, allowing us to better discriminate between speakers. Another idea was to analyse the quality of the signals using Direct-to-Reverberant (DRR) metrics as a selection criterion. This is likely to improve the quality of the algorithm. These are ideas that are easy to implement. There are, however, more complicated aspects to research, such as the problem with the size of the SOI. As seen in Fig. 3.4, the size of the SOI is frequency-dependent and can become very small for high frequencies. Researching a way to make the size of the SOI frequency independent and controllable with a hyperparameter would be optimal. Lastly, as shown in this manuscript, finding the correct pairs during processing is critical to the algorithm's functioning. The proposed method does function better than the original criterion, but it should only work on harmonic signals, and improvements are limited compared to the heuristic pruning. Therefore, finding better pruning schemes should be the priority when conducting research on this algorithm.

CHAPTER 6

CONCLUSION

During this master thesis, we have gathered data in which different speakers are present in different acoustic environments with different levels of background noise. The data contains recordings of these acoustic scenes from 8 different microphones, capable of omnidirectional and directional recordings up to the third order. We have evaluated a proposed algorithm for spotforming on real and simulated signals and have shown the effectiveness of the proposed algorithm in achieving spotforming. We have evaluated different alterations of the base algorithm with different combination schemes or base signals. Most of the alterations can separate the audio coming from the SOI from the original mixture in most configurations. The DSPF method shows the best separation in most cases, whereas the OPF method shows the best SDR values. Our evaluation of the proposed method using ambisonic signals steered towards the SOI revealed a crucial factor in the process: the selection of the microphone pairs or the criterion used. We demonstrated this through two proposed criterion that, when applied, filter out some microphone pairs. Both the heuristic criterion and the criterion that identifies coherent harmonic signals outperformed the original DSPF method. However, the heuristic criterion showed significant improvements, underscoring the importance of the choice of the criterion in the evaluation process. Spotforming is a subject with much work to be done, whether to be done on this algorithm or in the broader field.

[DP13] S. Delikaris-Manias and V. Pulkki, *Cross Pattern Coherence Algorithm for Spatial Filtering Applications Utilizing Microphone Arrays*, IEEE Transactions on Audio, Speech, and Language Processing **21** (2013), no. 11, 2356–2367, DOI: 10.1109/TASL.2013.2277928.

[EA00] H. Elkamchouchi and M. Adam, *A new constrained fast null steering algorithm*, IEEE Antennas and Propagation Society International Symposium. Transmitting Waves of Progress to the Next Millennium. 2000 Digest. Held in conjunction with: USNC/URSI National Radio Science Meeting (C, Vol. 2, 2000, 926–929 vol.2, DOI: 10.1109/APS.2000.875371.

[Far00] A. Farina, *Simultaneous measurement of impulse response and distortion with a swept-sine technique*, Audio engineering society convention 108, Audio Engineering Society, 2000.

[FGV05] C. Févotte, R. Gribonval, and E. Vincent, *BSS_EVAL Toolbox User Guide – Revision 2.0* (2005).

[Gui23] GuitareInteractiveMagazine, *GuitareInteractiveMagazine*, URL: https://guitarinteractivemagazine.com/review/zylia-zm-1/, 2023.

[Int09] International Organization for Standardization, *Acoustics - Measurement of room acoustic parameters - Part 1: Performance spaces*, 2009.

[Kag+22] Y. Kagimoto et al., *Spotforming by NMF Using Multiple Microphone Arrays*, 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Kyoto, Japan, 2022, 9253–9258, DOI: 10.1109/IROS47612.2022.9981808.

[Kla03] A. Klapuri, *Multiple fundamental frequency estimation based on harmonicity and spectral smoothness*, IEEE Transactions on Speech and Audio Processing **11** (2003), no. 6, 804–816, DOI: 10.1109/TSA.2003.815516.

[Pol16] A. Politis, *Microphone array processing for parametric spatial audio techniques*, PhD thesis, Aalto Univesity, 2016.

[Rea23] Reaper, *Reaper website*, URL: https://www.reaper.fm, 2023.

[SH15] M. Suzuki and T. Honjo, *Spot-forming method by using two shotgun microphones*, 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE, Hong Kong, 2015, 188–191, DOI: 10.1109/APSIPA.2015.7415500.

[TH16] M. Taseska and E. A. Habets, *Spotforming: Spatial Filtering With Distributed Arrays for Position-Selective Sound Acquisition*, IEEE/ACM Transactions on Audio, Speech, and Language Processing **24** (2016), no. 7, 1291–1304, DOI: 10.1109/TASLP.2016.2540815.

[uni23] A. university, *Aalto Acoustic lab facilities*, URL: https://www.aalto.fi/en/aalto-acoustics-lab/aalto-acoustics-lab-facilities, 2023.

[Van02] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*, John Wiley & Sons, 2002.

[Vor13] S. A. Vorobyov, *Principles of minimum variance robust adaptive beamforming design*, Signal Processing **93** (2013), no. 12, 3264–3277, DOI: 10.1016/j.sigpro.2012.10.021.

[WP24] S. Wirler and V. Pulkki, *Spatially selective sound capture based on aggregated pair-wise similarity measures*, 2024, Unpublished.

[ZF19] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, *Vol. 19*, Springer Topics in Signal Processing, Springer International Publishing, Cham, 2019, DOI: 10.1007/978-3-030-17207-7.

[Zyl23] Zylia, *Zylia website*, URL: https://www.zylia.co, 2023.