Acoustic Source Localisation in Constrained Environments

Elizabeth Vargas Vargas

a thesis submitted for the degree of
Doctor of Philosophy

Heriot-Watt University, Edinburgh
School of Engineering and Physical Sciences

· February 2020 ·

Abstract

Acoustic Source Localisation (ASL) is a problem with real-world applications across multiple domains, from smart assistants to acoustic detection and tracking. And yet, despite the level of attention in recent years, a technique for rapid and robust ASL remains elusive – not least in the constrained environments in which such techniques are most likely to be deployed.

In this work, we seek to address some of these current limitations by presenting improvements to the ASL method for three commonly encountered constraints: the number and configuration of sensors; the limited signal sampling potentially available; and the nature and volume of training data required to accurately estimate Direction of Arrival (DOA) when deploying a particular supervised machine learning technique.

In regard to the number and configuration of sensors, we find that accuracy can be maintained at state-of-the-art levels, Steered Response Power (SRP), while reducing computation sixfold, based on direct optimisation of well known ASL formulations. Moreover, we find that the circular microphone configuration is the least desirable as it yields the highest localisation error.

In regard to signal sampling, we demonstrate that the computer vision inspired algorithm presented in this work, which extracts selected keypoints from the signal spectrogram, and uses them to select signal samples, outperforms an audio fingerprinting baseline while maintaining a compression ratio of 40:1.

In regard to the training data employed in machine learning ASL techniques, we show that the use of music training data yields an improvement of 19% against a noise data baseline while maintaining accuracy using only 25% of the training data, while training with speech as opposed to noise improves DOA estimation by an average of 17%, outperforming the Generalised Cross-Correlation technique by 125% in scenarios in which the test and training acoustic environments are matched.

A mi familia ...

Acknowledgements

There are a lot of people who contributed to this thesis, directly or indirectly, and I want them to know how grateful I am for all their help and support.

First of all, I would like to express my gratitude to my supervisors, Dr. Keith Brown and Dr. Kartic Subr. I am very grateful to Keith for accepting to supervise me, even though it was not in his original plan; his guidance, encouragement and support — especially in the final stages — has been vital to the completion of this thesis. I would like also to thank Kartic for his faith in first entrusting me with this project, for all the lessons he has taught me along the way and for the white board discussions. My thanks go also to Dr. James Hopgood for all his contributions to this thesis. I appreciate the time he has dedicated to this research as well as his feedback, which was instrumental in finishing this thesis. I would also like to thank Dr. Neil Robertson for his trust and for recommending me for this position. Finally, I would like to thank Dr. Maria Trujillo, who has been my mentor from more than a decade now, and without whose encouragement and support I would not have made it this far. To all of you, I appreciate your infinite patience and support.

I gratefully acknowledge the funding for my research received from a James Watt Scholarship (JSW) in the School of Engineering & Physical Sciences. I am grateful to the heads of the Institute of Signals, Sensors and Systems (ISSS), Prof. Yvan Petillot and Prof. George Goussetis for their financial support that allowed me to attend conferences and disseminate my work within the signal processing community.

I also thank my colleagues for their company in this journey, the meetings of the KECTRA group and the nights out together. Special thanks go to Tatiana, friend, flatmate and like a little sister to me: I am glad we shared this adventure together during 3 years. I really appreciate all the fun times we had, but also all the support, love and attempts to make me laugh during the hard ones. I am also grateful to all others whom I have unjustly forgotten and who have been part of this journey.

A mi familia le agradezco por todo el apoyo y amor que me ha brindado desde que inicie mi educación cuando tenía 5 años. Desde las clases de ingles en el Colombo hasta las llamadas por Skype cada fin de semana, todo lo que he logrado hasta ahora es gracias a ustedes. Yo se que siempre se han esforzado por darme lo mejor y siempre voy a estar agradecida por todos sus sacrificios. Los amo.

Last but not least, I would like to thank Dan for all the support he has given me during these years. I know it has not been easy and there have been very stressful moments, but I consider myself very lucky to have such a wonderful human being by my side. Thank you very much for all your love and patience.

Research Thesis Submission

Name*:*	Elizabeth Vargas Vargas		
School:	School of Engineering and Physical Sciences		
Version: *(i.e. First, Resubmission, Final)*	Final	Degree Sought:	Doctor of Philosophy

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1. The thesis embodies the results of my own work and has been composed by myself
2. Where appropriate, I have made acknowledgement of the work of others
3. The thesis is the correct version for submission and is the same version as any electronic versions submitted*.
4. My thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5. I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.
6. I confirm that the thesis has been verified against plagiarism via an approved plagiarism detection application e.g. Turnitin.

ONLY for submissions including published works

7. Where the thesis contains published outputs under Regulation 6 (9.1.2) or Regulation 43 (9) these are accompanied by a critical review which accurately describes my contribution to the research and, for multi-author outputs, a signed declaration indicating the contribution of each author (complete)
8. Inclusion of published outputs under Regulation 6 (9.1.2) or Regulation 43 (9) shall not constitute plagiarism.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate*:*		Date:	

Submission

Submitted By *(name in capitals):*	ELIZABETH VARGAS VARGAS
Signature of Individual Submitting:	
Date Submitted:	

For Completion in the Student Service Centre (SSC)

Limited Access	Requested	Yes		No		Approved	Yes		No	
*E-thesis Submitted (**mandatory for final theses)***										
Received in the SSC by *(name in capitals):*						Date:				

Inclusion of Published Works

Declaration

This thesis contains one or more multi-author published works. In accordance with Regulation 6 (9.1.2) I hereby declare that the contributions of each author to these publications is as follows:

Citation details	**Elizabeth Vargas**, Keith Brown, Kartic Subr, "Impact of Microphone Array Configurations on Robust Indirect 3D Acoustic Source Localization", in *International Conference on Acoustics, Speech and Signal Processing* **(ICASSP)**, Calgary, Canada, April 2018. ***(Oral presentation)***
Elizabeth Vargas	Literature review, programming, experimental evaluation and manuscript writing.
Keith Brown, Kartic Subr	Revision and minor edits.
Signature:	
Date:	

Citation details	**Elizabeth Vargas**, James R. Hopgood, Keith Brown, Kartic Subr, "A Compressed Encoding Scheme for Approximate TDOA Estimation", in *European Signal Processing Conference* **(EUSIPCO)**, Rome, Italy, September 2018. ***(Oral presentation)***
Elizabeth Vargas	Literature review, programming, experimental evaluation and manuscript writing.
James R. Hopgood, Keith Brown, Kartic Subr	Revision and minor edits.
Signature:	
Date:	

Contents

List of Figures

List of Tables

List of acronyms

ASL	Acoustic Source Localisation
BLSTM	Bidirectional Long Short Term Memory
BPDN	Basis Pursuit Denoising
BSAR	Block Stationary Autoregressive
CBP	Continuous Basis Pursuit
CC	Cross-Correlation
CNN	Convolutional Neural Network
CRLB	Cramér-Rao Lower Bound
CRNN	Convolutional Recurrent Neural Network
CTLS	Constrained Total Least-Squares
CWLS	Constrained Weighted Least Squares
DNN	Deep Neural Network
DoG	Difference of Gaussians
DTFT	Discrete-Time Fourier Transform
DOA	Direction of Arrival
FFT	Fast Fourier Transform
GCC	Generalized Cross-Correlation
GAN	Generative Adversarial Network
GCC-PHAT	Generalized Cross-Correlation Phase Transform
GSG	Geometrically Sampled Grid

ICP	Iterative Closest Point
ISM	Image Source Method
LCLS	Linear-Correction Least-Squares
LS	Least-squares
LSTM	Long short-term memory
MAE	Mean Absolute Error
MCCC	Multichannel Cross-Correlation Coefficient
MFCC	Mel-frequency cepstral coefficients
MLW	Main Lobe Width
MSL	Maximum Sidelobe Level
MUSIC	MUltiple SIgnal Classification
NN	Neural Networks
SCOT	Smoothed Coherence Transform
RIR	Room Impulse Response
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SIFT	Scale-Invariant Feature Transform
SLAM	Simultaneous Localisation and Mapping
SLSQP	Sequential Least Squares Programming
SNR	Signal-to-Noise Ratio
SRP	Steered Response Power
SRP-PHAT	Steered Response Power Phase Transform
STFT	Short-Time Fourier Transform
TDOA	Time Difference of Arrival

TOA	Times of Arrival
UAV	Unmanned Aerial Vehicle
VAD	Voice Activity Detector
WCTLS	Weight Constrained Total Least-Square

Physical Constants

Speed of sound	$c = 343\,\mathrm{m\,s^{-1}}$

Nomenclature

α	Attenuation factor
Δx	Displacement
λ	Wavelength
$\mathbf{m_i}$	i-th microphone coordinate vector
$\mathbf{O}$	Microphone array centre
$\mathbf{R}_s(p)$	Source signal covariance matrix
$\mathbf{s}$	Source location
$\mathcal{G}$	Grid of candidate locations
ϕ	Elevation
ρ	Density
ρ_0	Background density
ρ_w	Perturbation density
$\sigma^2_{y_n}$	Expected value of $y_n^2(k)$
τ_{ij}	Time Difference of Arrival (TDOA) between microphones i and j
θ	Azimuth
$\tilde{\mathbf{s}}$	Estimated source location
$\tilde{\tau}_{ij}$	Estimated Time Difference of Arrival (TDOA) between microphones i and j
$\tilde{t}_i$	Estimated time for signal to travel from the source to the i-th microphone
A	Area
a	Acceleration

b_i	Binary vector at the i-th signal
c	Speed of sound in the air
D	Displacement
$E[\cdot]$	Mathematical expectation
F	Force
f	Wave frequency
f_i	Feature vector at the i-th spectrogram
K	Spectrogram frequencies of the STFT
M	Number of microphones
N	Number of extracted keypoints
n	Mass
P	Pressure
P_0	Background pressure
p_i	Spectrogram at the i-th signal
P_w	Perturbation in the pressure
r	Distance a light ray travels
T	Spectrogram times of the STFT
t	Time
t^*	Signal emission time
d_i	Distance between the source and the i-th microphone
t_i	Time for signal to travel from the source to the i-th microphone
Y_i^*	Complex conjugate of Y_i
Y_i	Signal at the i-th microphone in the frequency domain
y_i	Signal at the i-th microphone in the time domain

"*Sound is the most absorbent medium of all, soaking up histories and philosophical systems and physical surroundings and encoding them in something so slight as a single vocal quaver or icy harpsichord interjection.*"

Geoffrey O'Brien

Chapter 1

Introduction

Acoustic Source Localisation (ASL) refers to the ability to estimate the direction (or exact location) from which a sound is emitted, such as the speaker in a room, a barking dog and, when it is very precise, even a bee buzzing. Humans have the capability to perform ASL intuitively thanks to our ears, which receive acoustic signals, and relay them to the auditory apparatus, which is then able to estimate various audio cues, such as time and level differences between both ears, spectral information, timing analysis, correlation analysis, and pattern matching, and, ultimately, estimate the sound source location [6].

In signal processing, microphones act as replica ears, in the sense that they store the pressure of the emitted sound. This is what allows a group of two or more microphones organised in an array to perform ASL. The direct applications of this technology are many and varied, including smart assistants [7], acoustic target detection and tracking in poor light conditions [8], in addition to a variety of indirect applications, such as speech enhancement [9], acoustic Simultaneous Localisation and Mapping (SLAM) [10] and 3D reconstruction via SONAR [11].

1.1 Microphone Arrays

A microphone array is a group of microphones positioned in a way that captures spatial information [1]. Fig. 1.1 illustrates the problems that could potentially be solved using microphone arrays, including noise and echo reduction, dereverberation, localisation of one or multiple sources and the cocktail party problem. This thesis will focus on **localisation of a single source**, as highlighted in blue.

We will primarily consider omni-directional pressure sensors, and therefore some of the techniques we discuss will rely on the fact that there exists a Time Difference of Arrival (TDOA) between microphone pairs. The hardware used to record experiment data is a *gfai tech AC_Pro Acoustic Camera System* consisting of 72 microphones

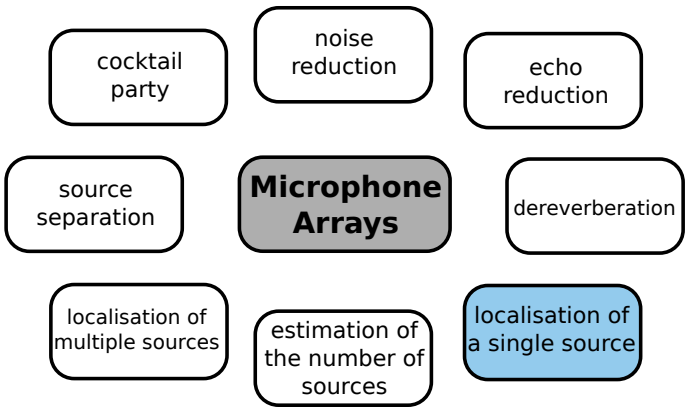


FIGURE 1.1: **Problems that can potentially be solved using microphone arrays.** This thesis focuses on localisation of a single source (blue).

sampled at 192kHz [12]. Although there exists a group of approaches focused on binaural localisation [13], that is techniques that use two ears (sensors), this is beyond the scope of this particular thesis.

1.2 Acoustic Source Localisation (ASL) in Constrained Environments

Given the vast amount of parameters at work in ASL, including sampling frequency, microphone array configuration and others, there are a large variety of scenarios in which a constraint on one of these parameters could be imposed. This thesis is focused on three types of constraints to ASL.

- **Number and configuration of microphones:** When there are only a limited number of microphones available or when only information for a specific number of microphones pairs can be accessed, this may lead to a variance in the accuracy of ASL. Moreover, the accuracy of the final source position estimation is potentially affected by the configuration in which these microphones are arranged.
- **Signal samples:** In certain scenarios, the use of the full length of an acoustic signal is either unavailable or computationally prohibitive. Therefore, it is necessary to investigate how to select signal samples in such a way as to preserve accuracy of source localisation.

- **Data available for training:** With the increase in machine learning and deep learning approaches, the study and use of training data has grown considerably in recent years. In some cases, there is either insufficient data available for training or the training data differs in terms of audio class with respect to the test data. It is important to study the effect that the amount or nature of training data has on the estimated source location.

1.3 Thesis Outline and Main Contributions

Chapter 2: The Physics of Sound

This chapter serves as an introduction to the basic concepts of sound from a physics perspective. The chapter begins with a brief summary of the wave equation and its core elements. Moreover, it explains how sound is measured and how frequency is represented. Next, we summarise the mechanics of reverberation, as well as the working principle of the Image Source Method (ISM), before concluding with a brief explanation of microphone arrays and the problems associated with them.

Chapter 3: Acoustic Source Localisation

This chapter summarises the state of the art in Acoustic Source Localisation (ASL), with a special focus on Time Difference of Arrival (TDOA)-based methods, which are used throughout this thesis. We also discuss the strengths and weaknesses of these approaches.

Chapter 4: Optimal Array Configuration and Number of Microphones

This chapter focuses on determining the optimal microphone array configuration, as well as the minimum number of microphones necessary to perform indirect source localisation. The main focus is the **number of microphones** constraint.

Part of the work presented in this chapter was published in (see Appendix C):

E. Vargas, K. Brown, K. Subr, "Impact of Microphone Array Configurations on Robust Indirect 3D Acoustic Source Localization", in *International Conference on Acoustics, Speech and Signal Processing* **(ICASSP)**, Calgary, Canada, April 2018. ***(Oral Presentation)***

Summary of contributions in Chapter 4

- We showed that direct optimisation of the well known formulation for ASL yields errors similar to the state of the art, Steered Response Power (SRP), with 6 times less computation.
- We showed using both simulated and real data that the method is robust to noise and reverberation.
- Our results have shown that circular arrays lead to higher localisation error than spiral and wheel configurations when considering large regions of space.

Chapter 5: A compressed encoding scheme for approximate TDOA estimation

This chapter focuses on TDOA estimation using signal samples derived from spectrogram features. In this work, we treated the spectrogram as an image and used Scale-Invariant Feature Transform (SIFT), a famous computer vision technique for keypoint detection, to detect the signal samples to be transmitted. The robustness of the approach was tested under different noise and reverberation conditions using various signals and source locations. The main focus is the **signal samples** constraint.

Part of the work presented in this chapter was published in (see Appendix C):

E. Vargas, J. R. Hopgood, K. Brown, K. Subr, "A Compressed Encoding Scheme for Approximate TDOA Estimation", in *European Signal Processing Conference*, **(EUSIPCO)**, Rome, Italy, September 2018. ***(Oral Presentation)***

Summary of contributions in Chapter 5

- We determined the signal keypoints to be transmitted in order to obtain an accurate TDOA estimation, at significantly lower data rates or improved accuracy compared with Generalized Cross-Correlation (GCC)-based solutions.
- We demonstrated the robustness of the proposed technique to different noise and reverberation conditions.

- We compared the proposed technique with another data-driven approach, that of audio fingerprinting, demonstrating that our algorithm is able to outperform an audio fingerprinting baseline while maintaining a compression ratio of 40:1.

Chapter 6: Impact of training data on Convolutional Neural Networks (CNNs) for Direction of Arrival (DOA) estimation

This chapter studies DOA using a CNN. The work is oriented to studying the impact of various data types for training purposes, including speech and music. We explore variations of these data, including the use of synthetic data, by means of a Generative Adversarial Network (GAN). The main focus is the **data available for training** constraint.

Summary of contributions in Chapter 6

- We showed that training with speech data, as opposed to noise, produces an average improvement of 3% on the accuracy of DOA estimates for test speech signals and 17% when the test signals belong to one of three other classes;
- We showed that training with music data from a dataset produces an average improvement of 19% in accuracy compared to training with noise;
- We proved that synthetic speech data generated using a state-of-the-art GAN [14], which can be generated automatically, is as effective in training as using real human speech;
- We concluded that music data performs better than speech data for training when obtained using real sound recordings: however, when they are synthetically generated using a GAN, speech data produces better results than music data;
- We compared with GCC, and showed that a Deep Neural Network (DNN) trained with speech is 125% more accurate when the test and training environments have similar reverberation, and comparable when the reverberation levels are different.

Chapter 7: Conclusions

This chapter summarises the conclusions derived from our work, as well as future directions to explore for further research.

1.4 Declaration of Authorship

The continuous use of "we" and "our" throughout the text is a matter of writing style. This thesis and the work presented in it are my own, and has been produced by me as the result of my own original research.

Part I

Background Literature Review

Chapter 2

The Physics of Sound

This chapter is intended as an introduction to the basic concepts of sound as a physical phenomenon. It begins with an explanation of the sound wave equation and its three main components: pressure, density and displacement. Next, it illustrates the way sound is measured and how frequency is interpreted. Third, it briefly summarises the main principle of reverberation. And finally, we explain the working principle of the Image Source Method (ISM) for simulated room acoustics. These concepts are fundamental to the understanding of the work presented in this thesis.

2.1 The Sound Wave

Sound is a wave, therefore it transfers energy from a source to a destination [15]. When someone is playing guitar, for instance, the guitar's strings vibrate, pushing air particles (molecules) back and forth. The particles are not in fact moving very far from their original position, but create a pattern that forms the wave. In this process, they transfer energy from the guitar to the ear. When the vibration reaches our ears, our brain interprets it as sound [16].

Sound requires a medium in order to be propagated. In a vacuum, for example, where there are no air molecules, there is no sound. In air on the other hand, the speed of sound is 343 m/s [15], concerned in the applications of this thesis.

Propagating a wave requires a restoring force, as well as some inertia in the medium [17]. In the guitar example, the vibration of the guitar strings moves the particles, causing the compression of the air molecules due to motion. The result of this is that more particles will be in the same area, causing an increase in density and pressure. The increase in pressure generates a restoring force, driving the motion of the particles (in both ways), and allowing the transmission of the wave [18]. Since the displacement of the medium is parallel to the propagation of the wave, we can refer

to sound as a *longitudinal wave*. In contrast, in transverse waves, the displacement of the medium is perpendicular to the direction of the wave [19].

In the case of sound, the pressure generates a force applied negatively to the pressure gradient that is, from the regions with high pressure to the ones with low pressure. When the air molecules are displaced away from the sound source, there is more compression, because they gather together with other particles. When they are displaced closer to the sound source, then the density of particles decreases, as well as the pressure [18].

Fig. 2.1 illustrates a longitudinal wave, where the pressure is represented as a function of distance. When the air molecule density is high, the pressure reaches its maximum point (compression). On the other hand, when the particles are scattered and the density is low, the pressure reaches its lowest point (rarefaction). The representation of sound by a sine wave is an attempt to illustrate the sinusoidal nature of the pressure-time fluctuations. It should not be concluded, however, that sound is a transverse wave.

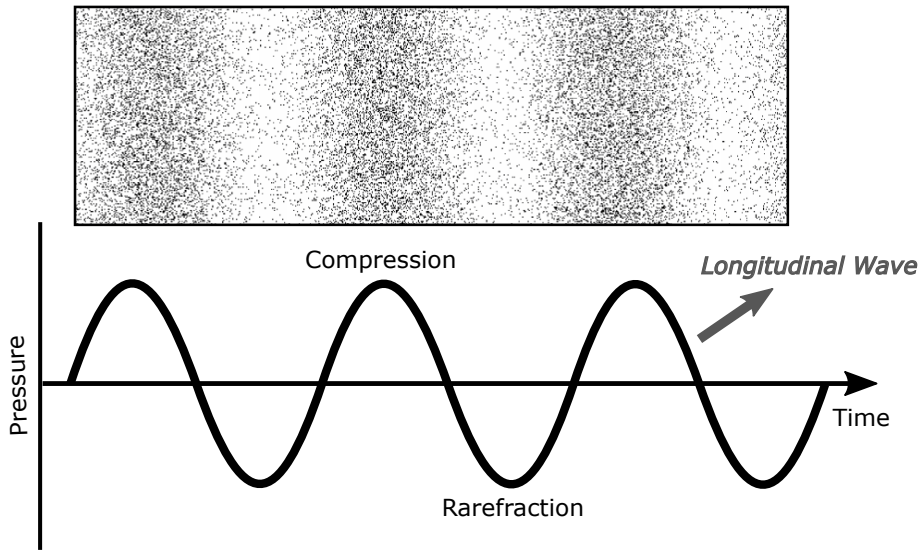


FIGURE 2.1: **Longitudinal wave with pressure as a function of time.** When the air molecule density is high, the pressure reaches its maximum point (compression) and when the particles are scattered and the density is low, the pressure reaches its lowest point (rarefaction).

2.1.1 Wave Equation

Three elements are present in the acoustic wave [19]:

- **Pressure (P):** related to the compression of the particles.
- **Density (ρ):** the number of particles per unit area.
- **Displacement (D):** how far the particles have moved away from their original position.

Fig. 2.2 illustrates the relationships between these elements. Our concern is to identify the equations underlying these relationships in order to find a mathematical representation of sound.

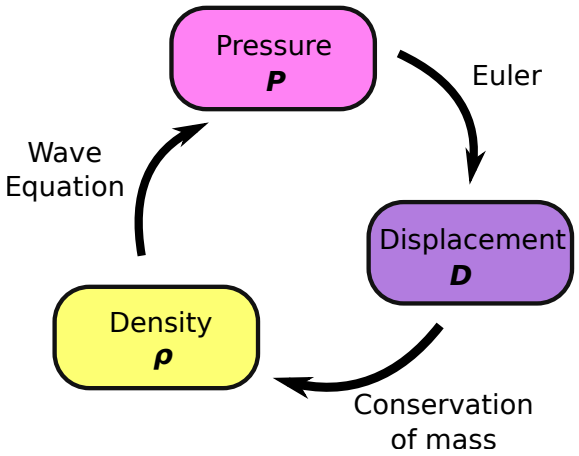


FIGURE 2.2: **Wave Equation.** Relationships between the wave equation elements: Pressure (**P**), Density (ρ) and Displacement (**D**)

Euler (P and D)

First of all, let us illustrate the relationship between P and D by a formula. We already know by Newton's Second Law that $F = ma$. Moreover, we know that volume is expressed by $(A\Delta x)$, where A represents the area in which the particles are located and Δx their displacement. Therefore, we can express the mass as the multiplication of volume by density, ρ, and the acceleration as the second derivative of displacement, D, as illustrated by Eq. 2.1.

$$F = (A\Delta x)\rho\frac{\delta^2 D}{\delta t^2} \tag{2.1}$$

We know by the definition of Pressure, P, in physics that $P = \frac{F}{A}$ [20], therefore we can express F in terms of the pressure, P, as stated in Eq. 2.2,2.3 and 2.4.

$$F = PA \tag{2.2}$$

$$F = A(P(x_1) - P(x_2)) \tag{2.3}$$

$$F = A\frac{\Delta x(P(x_1) - P(x_2))}{\Delta x} \tag{2.4}$$

Since both expressions of F are equivalent, we can derive Eq. 2.6, which is known as the **Euler Formula**. The Euler Formula demonstrates the relationship between Pressure, P and Distance, D in the sound wave.

$$\rho\frac{\delta^2 D}{\delta t^2} = \frac{(P(x_1) - P(x_2))}{\Delta x} \tag{2.5}$$

$$\rho\frac{\delta^2 D}{\delta t^2} = -\frac{\Delta P}{\Delta x} \tag{2.6}$$

Conservation of Mass (D and ρ)

The Displacement, D, and the density, ρ, are related through the pressure, P. Since $P = P_0 + P_w$ and $\rho = \rho_0 + \rho_w$, we can calculate $P(\rho)$ through a Taylor expansion, as presented in Eq. 2.7.

$$P(\rho) = P(\rho_0) + \frac{\delta P}{\delta \rho} + \cdots \tag{2.7}$$

Which is equivalent to $P = P_0 + P_w$. Combining the second term from Eq. 2.7 we obtain Eq. 2.8.

$$P_w = \frac{\delta P}{\delta \rho}\rho_w \tag{2.8}$$

We can determine $\frac{\delta P}{\delta \rho}$ experimentally by altering the density and measuring how much the pressure changes. This leads to $\frac{B}{\rho}$, which replaces $\frac{\delta P}{\delta \rho}$ in the previous equation, producing Eq. 2.10

$$P_w = \frac{B}{\rho}\rho_w \tag{2.9}$$

$$P_w = \frac{\rho_w}{\rho}B \tag{2.10}$$

Finally, to determine the relationship between ρ and D we need to analyse what happens to the displacement when we have large values for density. Fig. 2.3 illustrates what happens with the density with respect to the gradient of the displacement: when the gradient is negative, we have a higher density value, while when the gradient

is positive, we have a lower density value. This leads us to Eq. 2.11, known as the **Conservation of Mass**.

$$\rho_w = -\rho_0 \frac{\delta D}{\delta x} \tag{2.11}$$

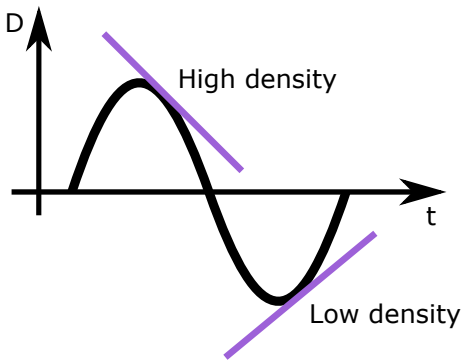


FIGURE 2.3: **Relationship between the density and the gradient of the displacement.** When the gradient (purple) is negative, we have a higher density value. In contrast, when the gradient is positive, we have a lower density value.

Wave Equation (ρ and D)

From the Euler Formula (Eq. 2.6) Eq. 2.10 we can derive Eq. 2.12.

$$\rho_0 \frac{\delta^2 D}{\delta t^2} = -\frac{B}{\rho_0} \frac{\delta \rho_w}{\delta x} \tag{2.12}$$

Moreover, from Mass Conservation (Eq. 2.11) we can derive Eq. 2.14.

$$-\frac{\rho_w}{\rho_0} = \frac{\delta D}{\delta x} \tag{2.13}$$

$$-\frac{1}{\rho_0} \frac{\delta \rho_w}{\delta x} = \frac{\delta^2 D}{\delta t^2} \tag{2.14}$$

Replacing Eq. 2.14 into Eq. 2.12 leads us to Eq. 2.16.

$$\rho_0 \frac{\delta^2 D}{\delta t^2} = B \frac{\delta^2 D}{\delta x^2} \tag{2.15}$$

$$\frac{\delta^2 D}{\delta t^2} - \frac{B}{\rho_0} \frac{\delta^2 D}{\delta x^2} = 0 \tag{2.16}$$

Setting c as the the speed of sound in the air, $c^2 = \frac{B}{\rho_0}$, Eq. 2.16 can be expressed as Eq. 2.17, which is known as the **Wave Equation**.

$$\frac{\delta^2 D}{\delta t^2} - c^2 \frac{\delta^2 D}{\delta x^2} = 0 \tag{2.17}$$

2.2 Sound Measurement

As explained in the previous section, sound is a wave that transfers air particles back and forth. When these particles reach a target, e.g. a microphone, the pressure that these particles apply can be measured in order to determine the loudness of the sound. In the International System of Units, the unit of sound pressure is the pascal (Pa), which is equivalent to one newton (N) of force applied over an area of one metre squared (m^2). However, since using such a large scale is hardly practical, a logarithmic scale in decibels (dB) was introduced [21]. Eq. 2.18 illustrates how the units in pascal relate to the decibel measurement. P is the sound pressure and P_0 is the reference sound pressure, equal to $20\mu Pa$, which corresponds to the lowest hearing threshold of a young and healthy ear [22].

$$L_P = 20 \log_{10}\left(\frac{P}{P_0}\right) \text{ dB}, \tag{2.18}$$

The scale of sound audible to the human ear ranges from 0 dB (hearing threshold) to 120-140 dB (pain threshold) [18].

2.3 Sound Frequency

Frequency is a property of sound related to pitch. In the International System of Units, frequency is measured in hertz (Hz), equivalent to 1 vibration per second [23].

Fig. 2.4 shows two sample Pressure vs Time plots. The first plot is an example of a sound wave with high frequency while the second one is an example of a sound wave with low frequency. The plots illustrate the time between successive high pressure points, known as the *period*. For a high frequency, the period is small, while for a low frequency, the period is high.

The *wavelength* (λ) is the spatial period of a periodic wave, that is, the distance (the period corresponds to the time) over which the wave repeats. The wavelength is traditionally calculated by measuring the distance between wave peaks; however, when we know the wave frequency (f) and the wave speed (c), we could use Eq. 2.19 to calculate the wavelength.

$$\lambda = \frac{c}{f} \tag{2.19}$$

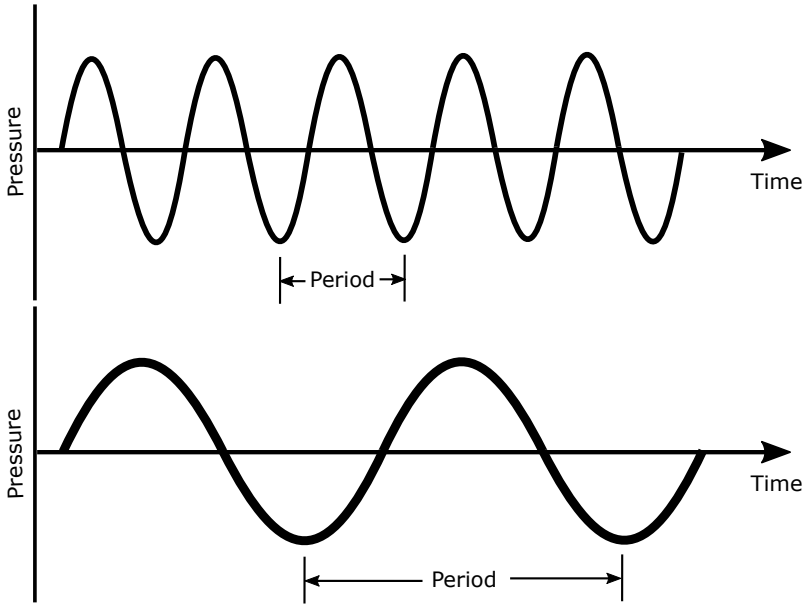


FIGURE 2.4: **Frequency.** Examples of Pressure vs Time plots illustrating a high and a low frequency respectively.

The scale of human hearing ranges from 20 Hz to 20 kHz. Speech is typically between 100 Hz to 1 kHz, while the peak sensitivity of human hearing is around 4 kHz. Sounds that have frequencies above the human hearing range are called ultrasound, while the ones below are called infrasound [24]. Some commonly known frequency values are the highest note of a soprano singer (2048 Hz), blue and fin whales (17 - 30 Hz), clapping (2.2 - 2.8 kHz), bat sonar clicks (25–80 kHz) and medical ultrasound (1–20 MHz) [23].

Sounds can be classified according to the frequency range they occupy. A sound is said to be *narrowband* when its energy is distributed over a relatively small section of the audible range. On the other hand, when a sound is distributed over a wide section of the audible range, it is classified as *wideband* or *broadband*.

2.4 Reverberation

The term reverberation refers to the persistence of sound after the sound itself is produced [25]. This phenomenon is caused by the interaction with the environment (walls, clothes, etc.) that occurs as a sound travels from the source to the target.

Examples of these interactions include transmission/refraction, reflection and/or diffraction [24]. Fig. 2.5 illustrates the path that the sound follows from the source (speaker) to the target (microphones) for a non-reverberant (black arrow) and reverberant (gray dotted arrow) situation. In the non-reverberant scenario, the sound follows a direct path straight from the speaker to the microphone, while in the reverberant scenario the sound reflects off the wall before reaching the microphone.

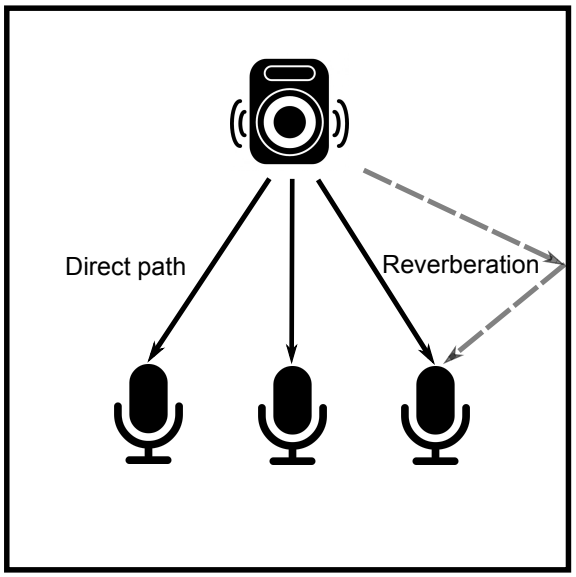


FIGURE 2.5: **Reverberation.** Path followed by a sound from the source (speaker) to the target (microphones) for a non-reverberant (black arrow) and reverberant (gray dotted arrow) situation.

Reverberations normally cause alterations in the sound amplitude, as well as in its spectrum [26]. The most common way to measure reverberation is in seconds, using a term known as T_{60}. This is defined as the time that it takes for the sound pressure level to reduce by 60 dB [25].

Echoes are normally associated with reverberation, however the main difference between the two is the time they take to reach the target. Reverberations reach their target within less than 50 ms, which results in the waves being perceived by the brain as a continuous sound. Echoes rearch their target between 50 ms and 100 ms [27], resulting in the sounds being perceived as separate events by the brain, as opposed to one extended event, as in the case of reverberations. In both cases, however, reverberations and echoes could last for several seconds.

Reverberations are normally linked to the room in which the sound source is present, and influenced by the size, shape and materials involved [26]. Examples of room reverberation values include kindergartens (0.4 s), offices (0.7 - 0.4 s),

classrooms (0.6 - 0.4 s), music rehearsal rooms (1.2 - 0.9 s), homes (0.9 s), bedrooms (0.5 s), and restaurants (0.8 - 0.7 s). This means that within a kindergarten, it would take 0.4 seconds for the sound pressure level to reduce by 60 dB, while inside a home the same reduction in the sound pressure would take 0.9 seconds.

2.5 Image Source Method (ISM)

The concepts presented in this chapter come together in the well-known ISM, originally proposed by Allen & Berkley in 1979 [28]. This method takes into account the physics of sound to generate a synthetic Room Impulse Response (RIR), meaning that it creates a transfer function between a sound and an acoustic sensor. What results are simulated signals for a microphone array inside a room. This is relevant in our work since we use a simulator based on this method for our experiments in future chapters.

The ISM assumes that the sound propagates along straight lines or rays. Based on these assumptions, the method mirrors the original sound source, such that it is located on a line perpendicular to the wall, at the same distance from the original source, with the result that there is a straight path (ray) between the mirrored source and the receiver. This mirrored source was called the "image source" [26]. Fig. 2.6 illustrates an example. Since the sound energy travels at a fixed speed along these rays, the energy in each ray decreases $1/r^2$, where r is the total distance travelled by the ray [29].

The original implementation is performed in the time-domain, leading to certain limitations. Lehmann & Johansson proposed an implementation using a frequency-domain simulation of the image sources, while implementing a phase inversion upon each sound reflection at the room boundaries [30]. We use this implementation to generate the simulations for our experiments.

2.6 Microphone Arrays

Microphone arrays are a fundamental tool in the experimental evaluation in future chapters. Since they take into account important physical properties of sound, this section will summarise those of them most relevant to our work.

One important aspect in the design of microphone arrays is the number of microphones they possess. It is well established in the theoretical and experimental literature on microphone arrays that the performance of a microphone array improves

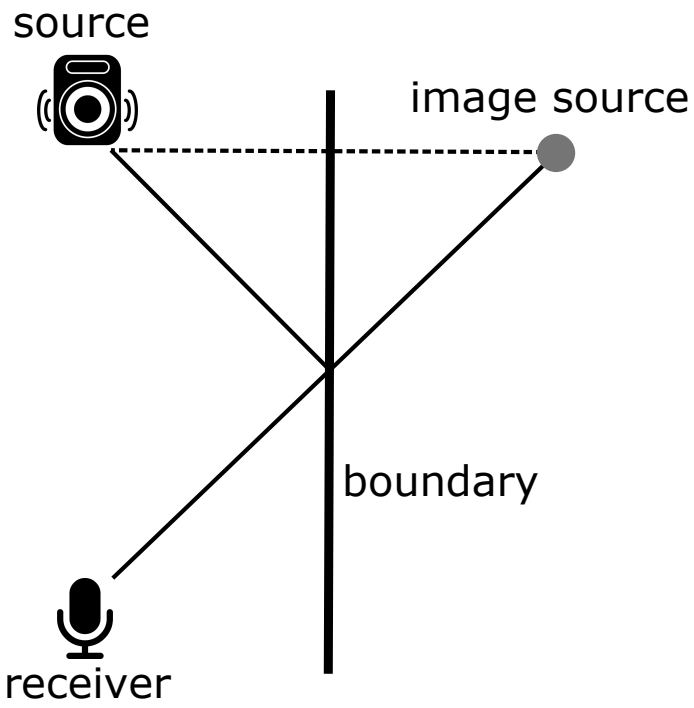


FIGURE 2.6: **Example of an "image source" representation.** An imaginary source is placed on a line perpendicular to the wall (gray dotted line), at the same distance from the original source, with the result that there is a straight path (black line) between the mirrored source and the receiver.

linearly as the size of the array grows [31]. This is, however, constrained by the spacing and availability of the microphones. Moreover, the increase in the number of microphone increases the computation needed to process the signals. Something similar occurs with the sampling frequency of each microphone. A high sampling frequency allows the calculation of more accurate Time Difference of Arrival (TDOA), as it will be presented in Chapter 3. However, as the length of the signal increases, the computation required also increases. Thus we can conclude that the number of microphones and the sampling frequency causes a trade-off between accuracy and efficiency.

Another factor considered in microphone array design is the distance between the microphones, which is closely related to the concept of aperture. The *aperture* refers to a spatial region that transmits or receives propagating waves. In acoustics, an aperture is an electroacoustic transducer that converts acoustic signals into electrical signals, e.g. microphone, or vice-versa, e.g. loudspeaker. The beam pattern/aperture directivity pattern is known as the aperture response, which is a function of the direction of arrival and frequency [32]. Therefore, the microphone array aperture tells us from which spatial region the array is receiving sound waves. The beamwidth is directly related to the spacing, d, between microphones. Increasing d leads to larger array aperture, which leads to more noise reduction. The issue arises when d is larger than half of the wavelength, causing *spatial aliasing* [33], which is an effect that causes different signals to become indistinguishable (or aliases of one another) when sampled.

2.7 Far-Field and Near-Field

One of the main applications of microphone arrays is the localisation of sound sources. This could be done by either estimating the Direction of Arrival (DOA) (angle) or the exact location in 3D space (x,y,z coordinates). The possibility to calculate one or the other depends on whether the source is located in the near field or in the far field.

Sound sources located at least 2 wavelengths away from the microphone array are said to be on the *far-field*. When the distance between the source and the array is smaller than 2 wavelengths, we say that the source is located on the *near-field* [34]. Since the wavelength is a function of frequency, concepts of far-field and near-field are also a function of frequency. Moreover, the distance between the source and the

array depends on the separation between the microphones, and therefore the far-field and near-field are also a function of the array size.

Far-field localisation assumes that the source is located far enough from the microphone array that the waves that reach the microphones are planar [35]. In this case, it is only possible to estimate the DOA.

On the contrary, when the source is assumed to be on the near-field, the wave fronts are spherical waves, and both the DOA and the range parameters (3D coordinates) should be determined to localise the source [36].

Chapter 3

Acoustic Source Localisation

Acoustic Source Localisation (ASL) refers to the ability to identify the spatial location of an object or person emitting a sound. This could be done either by determining the relative angle from which the sound emanates, also known as Direction of Arrival (DOA), (θ), by finding the exact location in 3D space (x, y, z) at which the object or person is positioned or by estimating the 2D angles, azimuth (θ) and elevation (ϕ), together with the range [6].

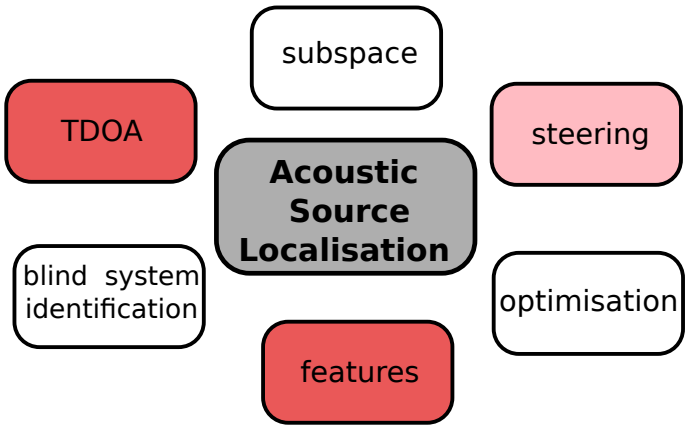


FIGURE 3.1: **Our Acoustic Source Localisation (ASL) literature classification.** It comprises six different type of methods. Our contributions (highlighted in bright pink) in this thesis are in Time Difference of Arrival (TDOA)-based approaches (Chapter 4 and Chapter 5) and feature-based approaches (Chapter 6). A steering-based approach (highlighted in light pink) will be used as the baseline (Chapter 4).

Given the application of ASL in multiple domains, the literature is rich in a variety of approaches to solve the problem. On the other hand, this has meant authors have different versions of how these approaches should be classified. An early classification was proposed in [37, 38], which included three types of approaches: subspace based techniques, microphone array beam scanning and TDOA-based approaches. In [39]

the authors add to this list adaptive multichannel time delay estimation using blind system identification based methods, probabilistic model based methods such as maximum likelihood method and methods based on histogram analysis of narrowband DOA estimates. Other authors in [40] propose a more general classification, based on criteria such as the source location, number of sources and type of features used.

Our classification of the literature is based on that proposed in [1] and our own criteria connected with the work presented in this thesis. Fig. 3.1 summarises our view of the ASL literature. This chapter will summarise each of these approaches, highlighting their strengths and weaknesses. Moreover, at the end of the chapter we will present a summary of the literature, stating the reason for choosing some of these approaches in the rest of the thesis. Finally, each of the following three chapters presents a more detailed review of the related work, in order to clarify our contributions to the ASL literature.

3.1 TDOA-Based Approaches

Time Difference of Arrival (TDOA)-based approaches are one of the most popular in Acoustic Source Localisation (ASL) estimation. Fig. 3.2 illustrates the TDOA-based ASL pipeline, which comprises three steps: the recording of the acoustic signal, considering the principles explained in Chapter 2, TDOA, explained in Section 3.1, and localisation, summarised in Section 3.1.2. Since various steps are involved, some authors refer to this group as *indirect methods*. Since the next couple of chapters involve the use of these sorts of methods, this section will present a summary of TDOA-based approaches in the literature.

FIGURE 3.2: **TDOA-based ASL pipeline.** It comprises three steps: the recording of the acoustic signal (Chapter 2), TDOA (Section 3.1) and localisation (Section 3.1.2).

3.1.1 Cross-Correlation Based Methods

The most important step in correctly estimating Direction of Arrival (DOA) or accurately localising the 3D position of a sound source is the estimation of TDOA, τ, among different microphone pairs. In this context, Cross-Correlation (CC) is

one of the oldest yet most common methods used to find the TDOA between microphones [41, 42].

The underlying assumption is that the microphone array has only two microphones, and therefore this first set of methods will focus solely on the cross-correlation between two microphones.

The **CC** between two signals $y_1(k)$ and $y_2(k+p)$ is defined by Eq. 3.1 [43]. This equation represents a convolution between two signals in the time domain. $E[\cdot]$ denotes mathematical expectation.

$$r_{y_1y_2}^{CC}(p) = E[y_1(k)y_2(k+p)] \tag{3.1}$$

The value of p at which $r_{y_1y_2}^{CC}(p)$ reaches its maximum value is the point at which the signals are the most similar [44], and therefore represents the TDOA as illustrated by Eq. 3.2.

$$\tilde{\tau}^{CC} = \arg\max_p r_{y_1y_2}^{CC}(p) \tag{3.2}$$

The calculation of the cross-correlation is extended to the frequency domain in a method called **Generalized Cross-Correlation (GCC)** [43]. In this case, instead of a convolution between two signals, Eq. 3.3 is used.

$$r_{y_1y_2}^{GCC}(p) = \int_{-\infty}^{\infty} \vartheta(f)\phi_{y_1y_2}(f)e^{2\pi fp}df \tag{3.3}$$

where $\phi_{y_1y_2}(f)$ is expressed by Eq. 3.4, Y_1 represents the Discrete-Time Fourier Transform (DTFT) of the signal y_1 and Y_2^* is the complex conjugate of its DTFT. p is the time and f the frequency variable of the DTFT.

$$\phi_{y_1y_2}(f) = E[Y_1(f)Y_2^*(f)] \tag{3.4}$$

The value of τ is again calculated as the value that maximizes the GCC, as expressed by Eq. 3.5.

$$\tilde{\tau}^{GCC} = \arg\max_p r_{y_1y_2}^{GCC}(p) \tag{3.5}$$

Different GCC methods may be applied [1], according to the chosen value of $\vartheta(f)$ [1].

- **Classical Cross-Correlation:** Corresponds to the case where $\vartheta(f) = 1$ [45].

- **Smoothed Coherence Transform (SCOT):** To mitigate the impact of fluctuating levels in the speech source signal, it is common to pre-whiten the microphone outputs [46]. This is represented by Eq. 3.6.

$$\vartheta(f) = \frac{1}{\sqrt{E[|Y_1(f)|^2]E[|Y_2(f)|^2]}} \tag{3.6}$$

- **Generalized Cross-Correlation Phase Transform (GCC-PHAT):** Since the TDOA depends on the phase, rather than on the amplitude on the spectrum, the amplitude can be discarded, keeping only the phase [47]. This is represented by Eq. 3.7

$$\vartheta(f) = \frac{1}{|\phi_{y_1y_2}(f)|} \tag{3.7}$$

The methods we have discussed so far are applied when we have only two microphones. When we have a microphone array of more than two microphones, the redundant information, that is, the same signal in all of them, is taken into account to calculate the TDOA. The key is to use the spatial correlation matrix, and therefore this technique uses the **Multichannel Cross-Correlation Coefficient (MCCC)**, which measures the correlation among outputs of an array system [1]. Therefore, it could be seen as a generalisation of CC.

Eq. 3.9 shows the matrix [48].

$$R_a(p) = E[\mathbf{y}_a(k,p)\mathbf{y}_a^T(k,p)] \tag{3.8}$$

$$R_a(p) = \begin{bmatrix} \sigma_{y_1}^2 & r_{a,y_1y_2}(p) & \dots & r_{a,y_1y_N}(p) \\ r_{a,y_2y_1}(p) & \sigma_{y_2}^2 & \dots & r_{a,y_2y_N}(p) \\ \vdots & \vdots & \ddots & \vdots \\ r_{a,y_Ny_1}(p) & r_{a,y_Ny_2}(p) & \dots & \sigma_{y_N}^2 \end{bmatrix} \tag{3.9}$$

with the aligned (subscript a) signal vector, and

$$\sigma_{y_n}^2 = E[y_n^2(k)], n = 1, 2, ..., N \tag{3.10}$$

$$r_{a,y_iy_j}(p) = E\{y_i[k + \mathcal{F}_i(p)]y_j[k + \mathcal{F}_j(p)]\}, i, j = 1, 2, ..., N, \tag{3.11}$$

Eq. 3.12 represents a way in which the matrix can be factorised.

$$R_a(p) = \Sigma\tilde{R}_a(p)\Sigma \tag{3.12}$$

Eq. 3.13 shows the sigma matrix [49], which is a diagonal matrix.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{y_1} & 0 & \dots & 0 \\ 0 & \sigma_{y_2} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{y_N} \end{bmatrix} \tag{3.13}$$

Eq. 3.14 shows the $\tilde{R}_a$ matrix, where $\rho_{a,y_iy_j}(p)$ is the correlation coefficient between the *ith* and the *jth* aligned microphone signals as illustrated by Eq. 3.15.

$$\tilde{R}_a(p) = \begin{bmatrix} 1 & \rho_{a,y_1y_2}(p) & \dots & \rho_{a,y_1y_N}(p) \\ \rho_{a,y_2y_1}(p) & 1 & \dots & \rho_{a,y_2y_N}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{a,y_Ny_1}(p) & \rho_{a,y_Ny_2}(p) & \dots & 1 \end{bmatrix} \tag{3.14}$$

$$\rho_{a,y_iy_j}(p) = \frac{r_{a,y_iy_j}(p)}{\sigma_{y_i}\sigma_{y_j}} i,j = 1,2,...,N, \tag{3.15}$$

The MCCC is linked to the normalised spatial correlation matrix by Eq. 3.16 [33].

$$\rho^2_{a,y_1:y_N}(p) = 1 - \frac{\det[\mathbf{R_a}(p)]}{\prod_{n=1}^{N}\sigma^2_{y_n}} \tag{3.16}$$

The MCCC has the following properties [1]:

1. $0 \leq \rho^2_{a,y_1y_N}(p) \leq 1$
2. $\rho^2_{a,y_1:y_N}(p) = 1$ if two or more signals are perfectly correlated
3. $\rho^2_{a,y_1:y_N}(p) = 0$ if all signals are completely uncorrelated
4. If one signal is completely uncorrelated with the others, the MCCC will measure the correlation among the remaining signals.

Using these definitions, we can say that to find the delay we need to maximise the MCCC as expressed by Eq. 3.17.

$$\tilde{\tau}^{MCCC} = \arg\max_p \rho_{a,y_1:y_N}(p) \tag{3.17}$$

From Eq. 3.16, we can see that the problem is equivalent to minimising the determinant of $\mathbf{R_a}(p)$ [50], as expressed by Eq. 3.18.

$$\tilde{\tau}^{MCCC} = \arg\min_{p} \det[\mathbf{R_a}(p)] \tag{3.18}$$

The main advantage of the GCC methods is that they are computationally efficient. Moreover they have been well studied and perform well in moderately noisy and non-reverberant environments. Their main drawback is however their sensitivity to reverberation, given that they assume an ideal free field model of the room in which no reverberations are present.

3.1.2 Localisation

Once we have estimated the TDOA, we can proceed to localising the acoustic source, either by finding the direction (angle) at which the source is located, known as DOA, or by computing the exact 3D coordinate (x, y, z) at which the source is positioned [37]. The choice of one or the other depends on our assumption of where the source is located.

Sound sources located at least two wavelengths away from the microphone array are said to be on the *far-field*. When the distance between the source and the array is smaller than two wavelengths, we say that the source is located on the *near-field* [34]. Since the wavelength is a function of frequency, then concepts of far-field and near-field are also a function of frequency. When the source is assumed to be on the far-field, the wave fronts are plane waves, therefore we can only estimate the DOA. On the contrary, when the source is assumed to be on the near-field, the wave fronts are spherical waves, and both the DOA and the range parameters (3D coordinates) should be determined to localise the source [36].

This section summarises the methods used to localise acoustic sources both in the far-field (Section 3.1.2) and near-field (Section 3.1.2).

Far-field Localisation: Direction of Arrival (DOA)

Far-field localisation assumes that the source is located far enough from the microphone array that the waves that reach away from the microphones are planar [35]. Fig. 3.3 illustrates sound waves arriving at a linear microphone array consisting of two microphones.

This model assumes an array of N sensors, with outputs illustrated by Eq. 3.19. $n = 1, 2, ..., N$, α_n are the attenuation factors due to propagation effects, $s(k)$ is the

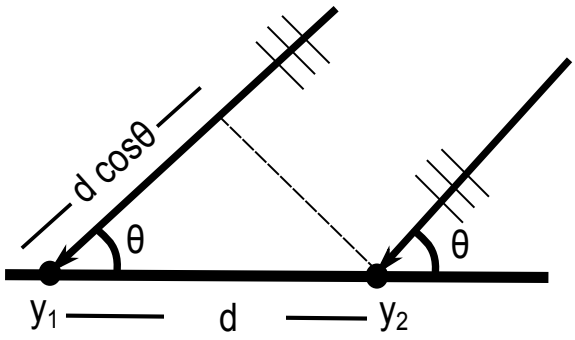


FIGURE 3.3: **Planar waves reaching the microphone array from a source located on the far-field.** θ is the incident angle of the planar wave, corresponding to the DOA.

source signal, t is the propagation time from source to sensor 1, v_n is an additive noise signal and τ is the TDOA [37].

$$y_n(k) = \alpha_n s[k - t - \mathcal{F}_n(\tau)] + v_n(k) \tag{3.19}$$

In the example of Fig. 3.3, with τ_{12} being the TDOA between microphones 1 and 2, Eq. 3.20 illustrates how this relates to the DOA, θ. d is the distance between the microphone pair and c the speed of sound in the air.

$$\tau_{12} = \frac{d\cos\theta}{c} \tag{3.20}$$

Since we know all the variables except θ, we could find its value using Eq. 3.21 [1].

$$\theta = \arccos\left(\frac{c\tau_{12}}{d}\right) \tag{3.21}$$

Generally, the TDOA in a microphone array with N equidistant microphones is represented by Eq. 3.22.

$$\mathcal{F}_n(\tau) = (n-1)\tau = \frac{(n-1)d\cos(\theta)}{c} \tag{3.22}$$

Far-field localisation will be studied in detail in Chapter 5, in which a TDOA-based algorithm will be proposed for DOA estimation.

Near-field Localisation: 3D Coordinates

When a sound source is located on the near-field, the wave fronts are spherical waves, and therefore both the DOA and the range parameters (3D coordinates) could be determined to localise the source. This section summarises the methods used

to localise the source in these scenarios and highlights the main advantages and drawbacks associated with them.

Let us start by assuming that the Time of Arrival (TOA), $\tilde{t_i}$, of the acoustic signal at the i-th microphone is given by Eq.3.23.

$$\tilde{t}(\theta_i) = t(\theta_i) + \eta \tag{3.23}$$

$t(\theta_i)$ is the time it takes the signal to travel from the source $\mathbf{s}$ to the i-th microphone, $\mathbf{m_i}$, as illustrated by the dotted lines in Fig. 3.4. t^* represents the delay in signal arrival common to all microphones, given by the time between the moment when microphones are turned on and the moment when the signal is transmitted. $\tilde{\mathbf{s}}$ is the estimated source location and η is the noise associated with the Times of Arrival (TOA) calculation.

The distance associated to $t(\theta_i)$, which corresponds to the euclidean distance between $\mathbf{s}$ and the microphone $\mathbf{m_i}$, is presented in Eq. 3.24.

$$\|\mathbf{m_i} - \mathbf{s}\| = \sqrt{(s_x - r\cos\theta_i)^2 + (s_y - r\sin\theta_i)^2 + {s_z}^2} \tag{3.24}$$

Therefore $t(\theta_i)$ can be expressed as Eq. 3.25.

$$t(\theta_i) = \frac{\|\mathbf{m_i} - \mathbf{s}\|}{c} + t^* \tag{3.25}$$

Lastly, the TOA can be expressed as Eq. 3.26.

$$\tilde{t}(\theta_i) = \frac{\|\mathbf{m_i} - \mathbf{s}\|}{c} + t^* + \eta \tag{3.26}$$

Near-field localisation relies on the estimation of TOA at the sensors (microphones) or TDOA between microphones pairs in order to leverage the most likely source location using multilateration and solve by means of Least-squares (LS) [51]. The main idea of LS methods is to minimise a function that is zero when there is no noise in the TOA or TDOA estimations. A variety of error functions have been defined, based on hypothesised parameters of the observed data, deriving various LS estimators. Traditionally, two approaches have become very popular: hyperbolic and spherical LS error functions.

In the case of the *hyperbolic LS error functions*, the idea is to minimise the total error between the measured TDOA and the TDOA predicted by the geometry, assuming a target position. This means that the LS optimisation function is using the TDOA among microphone pairs [52]. We assume that the TDOA, τ, between the i-th and j-th microphone are defined by Eq. 3.28.

$$\tau_{ij} = \tilde{t}(\theta_i) - \tilde{t}(\theta_j) \tag{3.27}$$

$$\begin{aligned} \tau_{ij} &= \tilde{t}(\theta_i) - \tilde{t}(\theta_j) \\ \tau_{ij} &= \frac{\|\mathbf{m_i} - \mathbf{s}\| - \|\mathbf{m_j} - \mathbf{s}\|}{c} \end{aligned} \tag{3.28}$$

$\tilde{\tau}_{ij}$ is the TDOA calculated using the microphone signals using methods such as GCC-PHAT. Eq. 3.29 is used as an objective function.

$$\arg\min_{\mathbf{s}} \sum_{i}^{M} \sum_{j}^{M} \left(\tilde{\tau}_{ij} - \tau_{ij}\right)^2 \tag{3.29}$$

Since this function is non-linear, minimising it leads to a computationally intensive solution as the number of microphones increases. Additionally, the hyperbolic function is very sensitive to noise, especially for far-field sources.

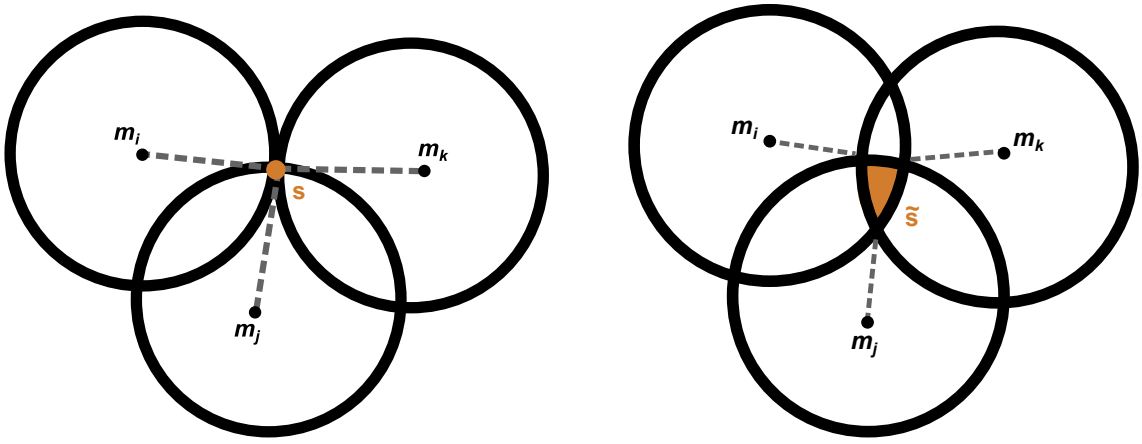
On the other hand, in the case of the *spherical LS error functions*, multilateration is defined as the intersection of spheres centred at the microphones [37], as illustrated by Fig. 3.4a. The spheres have been represented as circles for simpler visualisation. Fig 3.4a illustrates an ideal scenario, in which there is no noise in the calculations, and therefore the circles intersect exactly at a single point corresponding to the source location. Fig. 3.4 illustrates the case in which the TOA or TDOA measurements are noisy, requiring us to use LS optimisation in order to find the most likely source location [53].

The spherical error function is then defined by Eq. 3.30

$$\arg\min_{\mathbf{s}} \frac{1}{2} \sum_{i}^{M} \left(\|\mathbf{m_i} - \tilde{\mathbf{s}}\| - \|\mathbf{m_i} - \mathbf{s}\|\right)^2 \tag{3.30}$$

In general, TDOA-based localisation copes well with narrowband as well as broadband signals. Moreover, these methods can be used in different sampling rates and microphone array size. The main drawback is their dependency on the TDOA calculation, since high noise in the estimation could lead to very inaccurate localisation.

In Chapter 4 a method for near-field estimation will be evaluated using various microphone array configurations, while Chapter 5 will present an algorithm for TDOA estimation using few signal samples.



(A) Noise-free (B) Noisy

FIGURE 3.4: **Spherical Least-squares (LS) error function.** Represented by intersection of spheres (circles in 2D) centred at the microphones for the case in which (A) the calculations do not have noise (orange dot) and (B) when they do (orange area).

3.2 Subspace-Based Techniques

There is a set of methods that is based on high-resolution spectral analysis, which includes eigenanalysis-based techniques. The most popular of these methods is the well-known **MUltiple SIgnal Classification (MUSIC)** [54]. These methods perform statistical fit for Direction of Arrival (DOA) with respect to a spatio-spectral correlation matrix derived using the signals recorded at the sensors. The spatial correlation in this case is given by Eq. 3.31

$$\mathbf{R}_a(p) = \mathbf{R}_s(p) + \sigma_v^2 \mathbf{I} \tag{3.31}$$

where the source signal covariance matrix $\mathbf{R}_s(p)$ is represented by Eq. 3.32 [55], where τ is the Time Difference of Arrival (TDOA).

$$\mathbf{R}_s(p) = \begin{bmatrix} \sigma_s^2 & r_{ss,12}(p,\tau) & \dots & r_{ss,1N}(p,\tau) \\ r_{ss,21}(p,\tau) & \sigma_s^2 & \dots & r_{ss,2N}(p,\tau) \\ \vdots & \vdots & \ddots & \vdots \\ r_{ss,N1}(p,\tau) & r_{ss,N2}(p,\tau) & \dots & \sigma_s^2 \end{bmatrix} \tag{3.32}$$

and

$$r_{ss,ij}(p,\tau) = E\{s[k-t-\mathcal{F}_i(\tau)+\mathcal{F}_i(p)]s[k-t-\mathcal{F}_j(\tau)+\mathcal{F}_j(p)]\} \tag{3.33}$$

When $p = \tau$, the matrix has rank 1. Based on this property, the problem is formulated as a maximisation problem, as presented by Eq. 3.34, where $\mathbf{b}_n$ is an eigenvector of $\mathbf{R}_a(p)$ [56].

$$\tilde{\tau}^{BMUSIC} = \arg\max_{p} \frac{1}{\sum_{n=2}^{N} \mathbf{b}_n^T(p)\mathbf{R}_a(p)\mathbf{b}_n(p)} \tag{3.34}$$

The main limitation of these approaches is that they assume that the signal needs to be statistically stationary and originally narrowband (a broadband MUSIC has been developed [57]). Moreover the source needs to be located in the far-field of the sensor array and the multipath effect is not taken into account. These constraints mean that the algorithm is unsuitable for speech sources or highly reverberant environments.

3.3 Steering-Based Approaches

Steering-based approaches, also called *direct methods* by some authors, consist, as the names suggests, of steering a microphone array and scanning across a room for the highest energy output, which leads to an estimate of the direction of an active source [37]. Included in this category are methods such as Steered Response Power (SRP), in which a grid of candidate locations is used to find the sound source position, while others, such as beamforming, are focused on enhancing the signal coming from one direction.

In the case of SRP [58, 59], with $\mathcal{G}$ being the grid of candidate locations, Eq. 3.35 illustrates the SRP of a spatial point $\mathbf{s}$ for each microphone pair. The Generalized Cross-Correlation (GCC) is calculated by means of Eq. 3.3.

$$P(\mathbf{s}) = \sum_{m_i=1}^{M} \sum_{m_j=1}^{M} \tilde{\tau}_{m_i,m_j}^{GCC}(\mathbf{s}) \tag{3.35}$$

Eq.3.36 illustrates calculation for the source, $\tilde{\mathbf{s}}$, by selecting the position that maximises $P(\mathbf{s})$.

$$\tilde{\mathbf{s}} = \arg\max_{\mathbf{s}\in\mathcal{G}} P(\mathbf{s}). \tag{3.36}$$

Methods such as SRP generally present a trade-off between accuracy and efficiency. On one hand, when a very dense grid is used, the localisation is more accurate, but the algorithm computation time increases. The opposite happens when a grid is sparse. In Chapter 4, we present a more detailed examination of the literature including

these sorts of approaches. and SRP will be used as a baseline for comparison against a Time Difference of Arrival (TDOA)-based localisation method in the near-field.

The other set of steering-based approaches falls into the beamforming category. A beamformer is formulated as a spatial filter that operates on the outputs of a sensor array in order to form a desired beam (directivity) pattern, decomposed in two sub-processes: synchronization and weight-and-sum [1]. In simpler terms, the goal is to enhance the signal coming from one direction while suppressing noise and interference from other directions by steering the microphone array beamformer and scanning for the highest energy output in the room.

The problem is formulated by Eq. 3.37, where y_n are the N array outputs, given N microphones. The equation means that we can represent any signal by the delay of another plus some noise, where α is an attenuation factor due to the propagation effect.

$$\begin{aligned} y_n(k) &= \alpha_n s[k - t - \mathcal{F}(\tau)] + v_n(k) \\ &= x_n(k) + v_n(k), n = 1, 2, \cdots, N, \end{aligned} \tag{3.37}$$

Delay-and-sum is the most popular beamforming algorithm. As the name says, it has two parts: delay the signals to align them and then sum the aligned signals. Eq. 3.38 illustrates the delay part, in which a signal is shifted according to the imputed TDOA. In the Eq., $v_{a,s}(k) = v_n[k + \mathcal{F}(\tau)]$, where the subscript a implies that it is a copy of the signal.

$$\begin{aligned} y_{a,n}(k) &= y_n[k + \mathcal{F}(\tau)] \\ &= \alpha_n s(k - t) + v_{a,n}(k) \\ &= x_{a,n}(k) + v_{a,n}(k), n = 1, 2, ..., N, \end{aligned} \tag{3.38}$$

Eq. 3.39 illustrates the sum part, in which the aligned signals are added up. In this equation $\alpha_s = \frac{1}{N}\sum_{n=1}^{N}\alpha_n$ and $v_s(k) = \sum_{n=1}^{N} v_{a,n}(k)$.

$$\begin{aligned} z_{DS}(k) &= \frac{1}{N}\sum_{n=1}^{N} y_{a,n}(k) \\ &= \alpha_s s(k - t) + \frac{1}{N} v_s(k) \end{aligned} \tag{3.39}$$

Another way to visualise the delay-and-sum beamformer is by visualising the

corresponding beam pattern, defined as the magnitude of the the spatial filter's directional response [1]. Fig. 3.5. illustrates an example of a beam pattern.

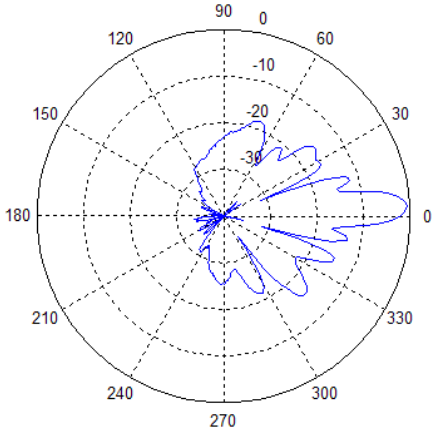


FIGURE 3.5: **Beam pattern (blue) [1].** The main lobe is pointed at a direction of 0° while the side lobes point in various orientations.

Using Eq. 3.40 the directional response can be calculated, where θ is the incident angle and ψ is a directional angle, such that $0 \leq \psi \leq \pi$. The beam pattern is written as Eq. 3.41.

$$\mathcal{S}_{DS}(\psi,\theta) = \frac{1}{N}\sum_{n=1}^{N} \exp^{-j2\pi(n-1)fd[cos(\psi)-cos(\theta)]/c} \tag{3.40}$$

$$\mathcal{A}_{DS}(\psi,\theta) = \left|\frac{sin[N\pi fd(cos\psi - cos\theta)/c]}{Nsin[\pi fd(cos\psi - cos\theta)/c]}\right| \tag{3.41}$$

In this context, there are two commonly used metrics to asses the performance of the beamformer. The first one is known as Maximum Sidelobe Level (MSL), which measures the difference between the main lobe and maximum power of the first side lobe in the beam pattern. In a beamformer with an ideal performance there would be no sidelobes, however in the best case scenarios a low MSL would be desired. Similarly, the Main Lobe Width (MLW) is also a commonly used metric. In general, small beam width is desired, in order to have a good separation between the lobes.

3.4 Blind System Identification

There are a set of approaches that adopt the real reverberant model instead of a free-field model, while also considering a single source and two microphones. Therefore, these methods first identify the two channel impulse responses from the source

to the two sensors and then measure the Time Difference of Arrival (TDOA) by detecting the two direct paths. Given that the source signal is unknown, the channel identification has to be a *blind method* [1].

One of these approaches is the **Adaptive Eigenvalue Decomposition Algorithm (AED)**. With $\mathbf{R}_{yy}$ as the covariance matrix of the two microphone signals and $\mathbf{w}$ is the two impulse responses, then $\mathbf{w}$ is found as the normalized eigenvector of $\mathbf{R}_{yy}$ corresponding to the smallest eigenvalue, as illustrated by Eq. 3.42, subject to $\|\mathbf{w}\| = 1$

$$\hat{\mathbf{w}} = \arg\min \mathbf{w}^T \mathbf{R}_{yy} \mathbf{w}, s.t. \|\mathbf{w}\| = 1 \tag{3.42}$$

This is solved in an adaptive manner using a constraint least mean squares (LMS) algorithm. After this algorithm convergence, the time difference between the direct paths of the identified channel impulse responses, $\hat{\mathbf{g}_1}$ and $\hat{\mathbf{g}_2}$, corresponds to the TDOA estimate as illustrated by Eq. 3.43.

$$\tilde{\tau} = \arg\max |\hat{\mathbf{g}_{1,l}}| - \arg\max |\hat{\mathbf{g}_{1,2}}| \tag{3.43}$$

Similarly, **Adaptive Blind Multichannel Identification Based Methods** calculate the TDOA by blindly identifying the impulse responses, but using more than two microphones instead [1]. The main idea is that, in a two-channel system, the zeros of the two microphones can be close, leading to an ill-conditioned system that is difficult to identify: and therefore, when more microphones are employed, this is less likely to happen.

3.5 Optimisation-Based Methods

There are another set of approaches again that frame the problem as an optimisation method, either for Time Difference of Arrival (TDOA) estimation or direct Direction of Arrival (DOA) or 3D localisation. Since some of these are based on very recent approaches, such as compressed sensing, we believe that they join to form a large set of approaches that can be condensed into a single category.

One large group of methods involves the use of compressive sensing. In [60] for example, the authors formulate the localisation problem as the sparse approximation of the measured signal in a specific dictionary of atoms. Compressive sensing and Fast Fourier Transform (FFT)-based feature extraction are used to create the atoms of the dictionary while sparsity is enforced via a circular grid. A similar approach is presented in [61], in which the authors presented an application in

distributed microphones arrays. They formulate the localisation problem as a sparse recovery problem based on the compressive sensing theory. Similarly to the previous approach, feature extraction is used, but using a discrete cosine transform (DCT). Moreover, they use a dictionary learning method, as well as an improved block-sparse reconstruction algorithm. Last but not least, another example of a compressed sensing-based approach is presented in [62]. The article introduces an approach for multiple source localisation in the near-field based on optimising the measurement matrix to enforce the restricted isometry property from compressive sensing and maximise the Signal-to-Noise Ratio (SNR).

Another example of this family of methods is presented in [63] using **Basis Pursuit Denoising (BPDN)** and **Continuous Basis Pursuit (CBP)**. These methods create a dictionary of shifted signals and search for the support within the dictionary and the corresponding coefficients that describe the received signal. BPDN uses the ℓ_1 penalty and has an advantage over greedy methods because it is guaranteed to converge to the global minimum solution. CBP surpasses BPDN by introducing a bilinear model that finds the atoms of the dictionary which best approximate the signal, and then improves that approximation by finding a coefficient for a corresponding dictionary which perturbs the approximation closer to the original signal. BPDN is not guaranteed to work since the signal may lie in an off-grid location, while CBP is limited to the assumption of sparsity in the scene.

A different use of optimisation is presented in [64], in which the authors frame the TDOA estimation as a constrained optimisation problem, using a known microphone array geometry in order to estimate the set of feasible TDOA. The method can be used in conjunction with an arbitrary number of non-coplanar microphones. The authors extend their approach in [65] and [66], including experiments with real data, the use of the branch and bound optimiser as well as arbitrarily-shaped microphone arrays. The main drawback of this method is that it is limited to a single source estimation and no reverberation is considered in the experimental setup.

3.6 Feature-Based Methods

Feature-based methods have appeared relatively recently with the rise of machine and deep learning techniques in the last decade. The main methodology of these approaches is either to extract certain features from the microphones' signals in order to estimate Time Difference of Arrival (TDOA) using a matching algorithm,

or to some features together with the acoustic properties of the room to estimate Direction of Arrival (DOA) directly.

An example of a recent work that extracts features is presented in [67], in which the authors present an approach based on audio fingerprint features. They subsequently perform self-localisation of an ad-hoc network of randomly distributed devices in open space with low reverberation but heavy noise. The device localisation framework calculates the distance using the difference between the maximum TDOA and the minimum TDOA. The main limitation of this approach is the need to locate sources at end-fire positions of the microphone array. This approach, as well as related ones, will be explained in further detail in Chapter 5, in which it is used as a baseline for comparison against our own feature-based method.

In the case of approaches that learn the features and the room acoustics in order to directly estimate the DOA, the literature is relatively recent, since it involves the use of Neural Networks (NN). However, given the popularity of NN, the number of approaches has grown very quickly. Examples of these include speaker localisation using a robot [68, 69], passive underwater sensing [70], antennas [71] and acoustic emission localisation on a pipeline [72]. Moreover, some approaches used for single localisation of a single speaker [4] have been extended to localise multiple sources [73]. A detailed review of these approaches will be presented in Chapter 6, in which the impact of various sound classes for training data in Convolutional Neural Network (CNN) is studied.

The main advantages of these approaches is the large amount of possibilities that could be explored, by considering new methodologies for feature extraction. Moreover, in the case of NN the work is relatively recent and new approaches to improve NN are developed everyday, some of which could be applied to Acoustic Source Localisation (ASL). The main drawbacks of these approaches are that they are very sensitive to training data and, in some cases, the NN have a tendency to overfitting.

3.7 Summary

The literature in Acoustic Source Localisation (ASL) could be summarised as follows:

- The most popular approaches for ASL are the Time Difference of Arrival (TDOA) based methods. They rely on the estimation of the TDOA among microphone pairs, that is, the difference in time at which the signal arrives at the different microphones. After this estimation, the algorithm proceeds to

estimate the location of either the Direction of Arrival (DOA) when the source is located in the far-field or the exact 3D source estimation when it is located in the near-field. On the one hand, the main drawback of these methods is that they are very dependent on a highly accurate TDOA estimation. On the other, their main advantage is that they cope well with both narrowband and broadband signals while being computationally efficient. Moreover, the localisation resolution can be flexibly adjusted by varying the sampling rate and the size of the microphone array. The research that we will present in Chapter 4 is concerned with these methods, because their flexibility with different microphone configurations and their efficiency will allow us to test them in a variety of signals and scenarios. We will show how the number of microphone pairs and their configuration could play an important role in the performance of TDOA-based methods, even when the TDOA estimation is very noisy.

- The most often used method to estimate TDOA is Cross-Correlation (CC). The main advantage of these methods is that they are computationally efficient while performing well in noisy conditions. Their main limitation is their low robustness to reverberation, given that they assume a free field room model. These methods are chosen in Chapter 4 given their robustness to noise, their efficient TDOA estimation, and their ability to work for various types of signals. Moreover, in Chapter 5 they are used in binary signals, given their flexibility to work for various types of input.

- After the estimation of TDOA among two or more microphone pairs, localisation by either DOA estimation or 3D calculation is the next step. In this context, the main limitations of TDOA-based approaches are their sensitivity to the TDOA estimation and, in the case of multilateration, the risk that the optimisation function falls into local minima. In Chapter 4 we will present our solutions, developed with the aim of overcoming these limitations.

- Subspace-based techniques, such as the popular MUltiple SIgnal Classification (MUSIC), perform a statistical fit for DOA with respect to the spatio-spectral correlation matrix that is derived from the recorded signals. The main advantage of these methods is their ability, as the name suggests, to estimate the DOA of multiple signals. However, their main limitation is the assumption that the signal needs to be statistically stationary and narrowband. Moreover,

the sources need to be located in the far-field and the algorithm performs poorly in highly reverberant environments.

- Steering-based approaches are one of the most accurate methods used in ASL. The main problem resides in the trade-off that appears when choosing the grid in which the source is going to be searched: a sparse grid will produce a very inaccurate localisation, however a very dense grid will carry a high computational cost. A instance of this family of methods, Steered Response Power (SRP), will be used in Chapter 4 as a baseline for comparison given the high accuracy obtained.

- Blind system identification was also a very popular technique for ASL, considering that it takes reverberation into account. The main drawback is that it relies on the blind estimation of the impulse responses to estimate TDOA. This can be complicated when the zeros in the two channels are close, leading to an ill-conditioned system that is either difficult to identify or unidentifiable.

- Optimisation-based approaches are a relatively new family of methods that, as the name indicates, formulate either the DOA or TDOA estimation using optimisation techniques. The main limitation of these methods is the assumptions made about the function, such as sparsity, in the case of compressive sensing, which limits its application in real scenarios.

- Finally, feature-based methods are the most recent set of approaches used in ASL. The main drawback is the difficulty in finding the right feature space or Neural Networks (NN) architecture for the problem. However, when these obstacles are overcome, these algorithms are known to be very accurate, given the advantage of learning techniques. Since they have been developed very recently, there are still a lot of paths to explore in their use. For this reason, in Chapter 5 we develop a feature-based algorithm oriented to compression, using computer vision-based techniques. Similarly, in Chapter 6 we study the use of NN for ASL, considering different types of training data.

Part II

Contributions

Chapter 4

Optimal Array Configuration and Microphone Pairs

4.1 Introduction

Localisation is the problem of estimating the position of objects in 3D space. Despite the advances in localisation using visual features, the use of audio sensing continues to offer important advantages, such as reliability under poor illumination, inexpensive sensing equipment and the use of signal processing (1D) tools. There have been attempts to use audio localisation both in robotics [74] and in scene understanding [75]. *Acoustic Source Localisation (ASL)* is typically achieved by leveraging known discrepancies in measurements of the emitted signal at multiple locations. ASL algorithms may exploit differences in time, amplitude, or both.

Some approaches to ASL, such as the Steered Response Power (SRP) [58, 59], solve directly for the most likely position of the sound source amongst a grid of candidate locations. In contrast, "indirect" methods first estimate the Times of Arrival (TOA) at the sensors (microphones) or the Time Difference of Arrival (TDOA) across pairs of microphones, and then use this information to infer the source position via multilateration [2, 51]. Although indirect methods are simpler to express as a least squares optimization [37], the resulting objective function is non-convex and often does not lend itself to an analytical solution. Various reformulations of these methods using weighted least squares, convex constrained least squares [76], total weighted least squares [77] and weight constrained total least squares [78] have been analysed in the literature. Direct methods are believed to be more robust to noise and reverberation [58]. Section 4.2 presents a more detailed review of these methods.

A uniform circular array of microphones [79, 80] or a ring configuration [81] are common choices for taking measurements since azimuthal angles to sources are considered more important than elevation. The advantage of *acoustic cameras* with

such arrays is that they can focus on specific targets [82,83], which is useful for speech processing. Recent studies have shown the resolution in elevation to be improved by using a 2.5D circular array [84]. While there have been a few studies examining the use of spherical arrays, multiple spheres [85], randomly placed microphones [86, 87] and spiral configurations [88], there is little analysis of the impact of an array's geometric structure on particular optimisation algorithms for ASL.

Given various applications in which the use of a variety of microphone array configurations has a positive impact on results, such as speech enhancement [82] or traffic noise analysis [89], our hypothesis is that using one particular microphone configuration over another could also bring more accuracy to the estimation of sound source localisation. Moreover, we believe that indirect methods could be fast and reliable when used in combination with the right amount of microphone pairs.

In this research, we adopt an optimisation (sequential least squares programming) approach to indirect ASL. We focus on localising a single source, though other work directed towards estimating TDOA for multiple sources is directly applicable. Although the objective function we choose is non-linear and non-convex, we show using *simulation and real data* that the method is robust to noise and reverberation. Our experiments verify that the technique is comparable to SRP for real data while being 6× more efficient to compute. Using this optimisation scheme, we study the localisation error resulting from different geometric structures in the microphone array. Our results show that circular arrays produce the highest errors (across space) and are therefore least desirable.

In summary, in this work:

- we showed that direct optimisation of the well known formulation for ASL yields errors similar to the state of the art (SRP) with 6 times less computation.
- we showed using both simulation and real data that the method is robust to noise and reverberation.
- our results have shown that circular arrays lead to higher localisation error than spiral and wheel configurations when considering large regions of space.

4.2 Related Work

This section summarises the literature related to our work in this chapter. It starts by presenting the literature in Acoustic Source Localisation (ASL) using Least-squares (LS) (Section 4.2.1) and Steered Response Power (SRP) (Section 4.2.1).

Next, we present a brief description of the methods that have used microphone array configurations as a means to improve results. Finally, we finish with a summary of the most closely related works, and the way they differ from our current work.

4.2.1 Acoustic Source Localisation (ASL)

As previously stated, ASL can be solved using either direct methods, such as SRP, or indirect ones such as multilateration. This section presents a succinct overview of the approaches that will be used throughout the rest of this chapter: LS estimators and SRP. A more general summary of ASL was presented in Chapter 3, and included a much broader classification.

Least-squares (LS)

Multilateration approaches estimate the Time Difference of Arrival (TDOA) across sensors and use this information, together with an optimisation function, in order to find the source location. LS approaches have been widely employed in such scenarios, with a primary focus on robustness to noise and reverberation by avoiding falls into local minima.

One of the early approaches to indirect ASL was presented in [51]. This proposes a Linear-Correction Least-Squares (LCLS) estimation method for acoustic source localisation. The criteria used for optimisation are the hyperbolic LS error function and the spherical LS error function. The proposed approach is compared against spherical interpolation, and quadratic-correction LS estimators with and without iterations in the second correction stage, showing that the proposed approach yields superior performance.

Various reformulations of these methods followed, with the aim of improving localisation, in [77], researchers propose an algorithm based on Constrained Total Least-Squares (CTLS). This is solved using a numerical method based on Newton's optimisation method, guaranteeing convergence to the global minimum. Results show that CTLS outperforms [51] in terms of location accuracy and computational complexity for experiments that use simulated data.

Building on the previous approach, [78] puts forward a method that involves Weight Constrained Total Least-Square (WCTLS). This algorithm not only exploits the structural information of the measurement matrix, but also uses the prior knowledge of TDOA measurements. The algorithm is evaluated using simulations and the results show that it outperforms [77].

In [90] the proposed approach is a Constrained Weighted Least Squares (CWLS) estimator for TDOA-based localisation. The problem this work addresses is the situation when a sound source is located in front of the centre ($x = 0$ and $y = 0$) of a circular array (located in the xy plane, with $z = 0$). The main approach is to separate the source coordinates and the additional variable to different sides of the linear equations. As a result, the matrix to be inverted has a smaller condition number than that of the conventional LS approach. The results show that the method works for their data: however, the experiments are limited to simulated data and assume sufficiently small noise conditions.

A similar approach is presented in [76], where the problem is reformulated into a convex optimization problem. The authors derive a primal-dual interior point algorithm to reach a global solution efficiently. Moreover, similarly to the previous approach, the algorithm is able to avoid the ill-conditioning problem, that is, when the array is circular and the source is located near the centre of the array. The method is tested in simulated mild conditions and has shown that it can theoretically achieve Cramér-Rao Lower Bound (CRLB) accuracy. In estimation theory and statistics, the CRLB establishes a lower bound on the error covariance matrix for any unbiased estimator [91].

Steered Response Power (SRP)

The original SRP technique, as first proposed in [92], is a very robust approach for ASL, especially in the presence of noise and reverberation. The algorithm consists of (i) an estimate of Generalized Cross-Correlation (GCC) and (ii) a search for the most likely source location over a grid of points. Therefore, there is a trade-off when choosing the grid resolution: a sparse grid leads to inaccurate results, while a dense grid makes it computationally expensive. Its main drawback, therefore, is the large computational cost it carries when a high accuracy is desired. For this reason, the SRP literature has mostly focused on decreasing the computational cost while maintaining the original algorithm robustness.

One of the earlier efforts to speed up computation in SRP was presented in [59], which proposes a stochastic region contraction (SRC) implementation of Steered Response Power Phase Transform (SRP-PHAT) in order to speed up calculations by 2-3 orders of magnitude. The basic idea is that, given an initial rectangular search volume, and using an iterative process, the original volume will be contracted until a small subvolume is reached, within which the global optimum is contained. The

authors test their approach with real human speech and show how their algorithm accurately localises the sound source.

In the same vein, in [58] the authors propose an efficient variation of SRP called refined volumetric SRP (RV-SRP). As opposed to the previous approach, in which the search space is reduced, this approach tries to modify the function that estimates acoustic activity, while still maintaining the entire search space. Therefore, the search space is divided into small volumes and the algorithm finds the one with the highest acoustic activity, using a proposed SRP functional. Afterwards, traditional SRP is applied to the small volume. Results are presented for real and simulated scenarios.

Similarly, [93] proposes a sensitivity-based region selection SRP algorithm (R-SRP). It first identifies whether the source is positioned in a high or low sensitivity region, using peak-to-peak ratio (PPR) and Geometrically Sampled Grid (GSG). Afterwards, it searches the acoustic source in the selected region using the sensitivity map to weight the power acoustic map. The experiments are performed using real and simulated data.

In [94], the authors present a similar approach applied to localisation of multiple sources. The main idea is still the refinement of a search grid, as in [59], however they use Steered Response Power Density (SRPD), a measure of the spatially averaged SRP and an associated, signal-adaptive search method called hierarchical grid refinement to reduce the number of steering directions needed to estimate Direction of Arrival (DOA). This method can then estimate the number of sources and localise them in a variety of simulated and real scenarios.

In [95] authors propose a mixed approach in which the spatial grid used by SRP-PHAT is designed using the discrete hyperboloids obtained from TDOA estimations. This is called the GSG algorithm. This is tested in simulated and real scenarios showing high localisation accuracy in areas of high sensitivity while in low sensitivity regions the performance is degraded.

4.2.2 Microphone Array Configurations

The configuration of the microphone array has been previously studied with application in a large variety of domains. Most of the cases study the optimal placement of the microphones rather than the impact of existing configurations. This section summarises these approaches and explains how they are related to each other.

In [89], authors present an approach that determines the microphone array geometry via optimisation. The array consists of 24 microphones located on a 2D plane. The iterative optimisation procedure is focused on minimising the Maximum

Sidelobe Level (MSL) using only a limited number of microphones, while the positions of other microphones are obtained by a unique mathematical relationship, which ensures an irregular structure. The experiments present the estimated configuration results, compared against four different microphone arrays presented in the literature. The results show that, when used on the designed frequency range, this approach exhibits smaller values of MSL.

Similarly, in [96] authors present a microphone array layout for two dimensional sound field recording and reproduction. As opposed to previous approaches, the array layout has a discrete rotationally symmetric geometry composed of several geometrically similar subarrays. With that in mind, the microphone positions are determined based on the following constraints: number of microphones, frequency range and microphone diameter. The array is compared against a circular array and presents better results in terms of sound field reconstruction.

In [97], authors present a microphone array that was designed using a bi-objective optimisation technique that allows a trade-off between array resolution (beamwidth) and maximum side-lobe levels. The authors found that 7 concentric circles with 9 microphones per circle yielded the best performance. The goal of the microphone array is to localise noise generation from a model wind turbine and the experiments were performed using a rotor rig.

In [98], the authors use optimisation to place the microphones, and HR CLEAN-SC to evaluate its impact in beamforming using simulations, applied to an array in an open-jet anechoic wind tunnel. MSL as well as Main Lobe Width (MLW) are used for the optimisation function. Their conclusion is that using only MSL is not sufficient, since the location of the sidelobes also matter. Moreover, to best exploit HR CLEAN-SC, the source marker constraint should be adjusted according to the MSL and the region of interest.

In [99–102], the authors propose a set of approaches based on iterative microphone array removal. The authors introduce their technique in [100, 101], where there are two array removal methods proposed: one-by-one microphone removal, which systematically removes one microphone at a time, such that this reduced array results in the smallest product of the frequency-averaged MSL and frequency-averaged MLW; and exponential decay profile removal, which uses a Exponential Decay Profile (EDP), allowing one or more microphones to be removed each time. Results show improvements in logarithmic spiral and randomised pattern arrays. In [99], the authors used their technique to design eight 48-channel arrays, and test them and compare them against beamforming maps. [102] presents a modification of the algorithm, which included changes in MSL and MLW with respect to the microphones

removed.

In [103], the authors present an approach that does not rely on numerical optimisation, but instead combines phylotaxis modelled by Vogel's spiral together with a weighting proposed by Hansen. Results are presented for one frequency and a 64-channel microphone array. In this case, the circle geometry leads to the most narrow beam width, but also has a low side lobe level.

In [82], the authors evaluate the performance of beamforming using six different array configurations with 63 microphones: Archimedean spiral, Dougherty log-spiral, Arcondoulis spiral, Multi-spiral, Underbrink array and Bruel & Kjer style array. The authors use various source locations, both in the near-field and far-field. The comparison metrics include the ability of the array to locate a source at a given frequency, known as resolution, as well as its ability to reject sources away from the main source, by measuring the MSL, determined by the next highest lobe in the array response. Results show that arrays based around multiple arms, with microphones evenly distributed around the array area, achieve the best resolution and adequate MSL. The opposite happens when the high density of microphones is located at the centre: high MSL is achieved with poor resolution.

4.2.3 Summary

The related work to this chapter could be summarised as follows:

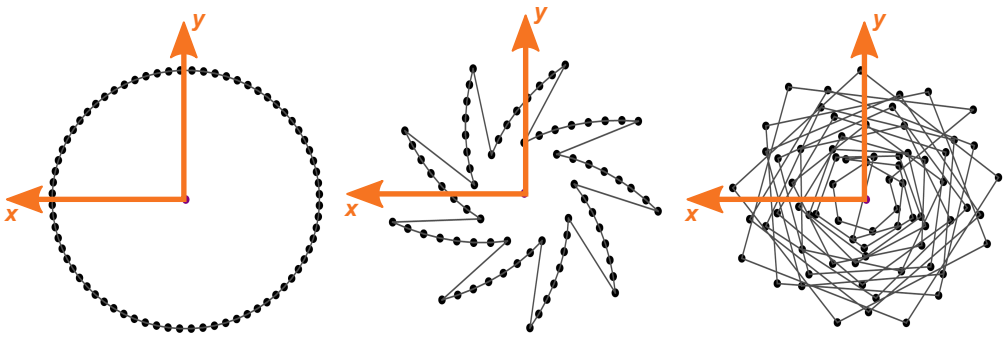
- There have been multiple efforts [51, 77, 78] to modify LS to achieve more accurate localisation, by evading local extrema, making the algorithm robust to noise and reverberation. Most of these efforts, however, have been evaluated only in simulated environments or with limited varieties of data.

- The approaches most closely related to our work are [90] and [76], given that their aim is to improve localisation when the array is circular and the source is located near its centre. As opposed to our approach, however, the authors do not analyse various source locations nor evaluate their approach for various audio classes. Moreover, their approaches are limited to mild acoustic conditions and experiments using simulated data.

- From the approaches proposed to improve SRP, [59] has not only been widely cited, but also its implementation has been released in MATLAB. Therefore, we decided to use this approach as a baseline for comparison for our work, as will be presented in Section 4.3.3.

- Most of the work related to microphone configurations is about optimal microphone placement in order to localise the source accurately and the evaluation metrics are those used in beamforming, such as MSL and MLW. The closest approach related to our work is presented in [82], since the authors compare a set of microphone configurations. Its main difference with ours is that the method does not take into account the source location, but rather tests the configuration for a couple of locations, with the metrics evaluated in terms of beamforming.

To the best of our knowledge, there is no evidence that there is an approach that compares configuration and its impact in localisation on 3D space in terms of performance. Moreover, no other author has used Sequential Least Squares Programming (SLSQP) for ASL or determined the optimal number of microphone pairs needed to accurately localise the sound source.



(a) ring (b) wheel (c) spiral

FIGURE 4.1: **Microphone configurations** Ring, wheel and spiral. The purple dot represents the centre of the array. The microphones are arranged in the xy plane with $z = 0$.

4.3 Methodology

This section introduces the formulation of Acoustic Source Localisation (ASL) with multilateration, using Times of Arrival (TOA) and Time Difference of Arrival (TDOA). Moreover, we present the baseline for comparison in our algorithms.

4.3.1 Multilateration

Consider a source at location $\mathbf{s}$ that emits an acoustic signal at some arbitrary time t^*. Let the measurements of the emitted sound be recorded by an array of

M microphones located at $\mathbf{m}_i,\ i = 1, 2, ..., M$ and the times taken by the signal to travel from $\mathbf{s}$ to $\mathbf{m}_i$ be t_i. If the distance between the source and the i^{th} microphone is $d_i \equiv \|\mathbf{m}_i - \mathbf{s}\|$, then $t_i = d_i/c + t^*$ where c is the speed of sound in air and t^* is not generally known.

Times of Arrival (TOA)

In the case that the times of arrival at the microphones are measured as $\tilde{t}_i$, we pose the ASL problem as one of jointly determining $\mathbf{s}$ and t^* as per the following:

$$O_1 : \quad \arg\min_{\mathbf{s},t^*} \sqrt{\sum_{i=1}^{M} (\tilde{t}_i - t_i)^2} \tag{4.1}$$

Time Difference of Arrival (TDOA)

Another possibility is to note the difference in measured times between a pair of microphones, $\tilde{\tau}_{ij} \equiv \tilde{t}_i - \tilde{t}_j$, or TDOA. The literature is rich in methods to estimate TDOA, as previously presented in Chapter 3. In this work, we choose the popular Generalized Cross-Correlation Phase Transform (GCC-PHAT) [104]. Then, we perform ASL by optimising [37]:

$$O_2 : \quad \arg\min_{\mathbf{s}} \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{M} \left(\tilde{\tau}_{ij} - \tau_{ij}\right)^2}, \tag{4.2}$$

where $\tau_{ij} = (t_i - t_j)$.

For both formulations O_1 and O_2, we know that the solution is constrained by the dimensions of the room, so we supply these constraints as linear inequalities.

We solve the constrained non-linear optimisation, illustrated using Sequential Least Squares Programming (SLSQP). Eq. 4.3 illustrates a formal definition of the general nonlinear programming problem.

$$\arg\min_{x \in R^n} f(x) \tag{4.3}$$

subject to

$$g_j(x) = 0, j = 1, ..., m_c \tag{4.4}$$

$$g_j(x) \geq 0, j = m_c + 1, ..., m \tag{4.5}$$

$$x_l \leq x \leq x_n, \tag{4.6}$$

where the problem functions $f : R^n \rightarrow R^l$ and $g : R^n \rightarrow R^m$ are assumed to be continuously differentiable and to have no specific structure.

SLSQP is an iterative procedure, and it is solved starting with a given vector of parameters x^0, the $(k+1)^{st}$ iterate x^{k+1} will be obtained from x^k by the step

$$x^{k+1} = x^k + \alpha^k d^k \tag{4.7}$$

where d^k is the search direction within the k^{th} step and α^k is the step length.

In each iteration, a constrained quadratic programming sub-problem is built so that the chain of solutions converges to a local minimum [105]. Each subproblem replaces the objective function with a local, quadratic approximation subject to local affine approximations of the constraints, as illustrated by Eq. 4.8. The optimiser used to solve each subproblem is a modified version of NNLS [106].

$$L(x, \lambda) = f(x) - \sum_{j=1}^{m} \lambda_j g_j(x) \tag{4.8}$$

We chose the step length using an L_1 test function, as illustrated by Eq.

$$\phi(x; \varrho) = f(x) + \sum_{j=1}^{m_e} \varrho_j |\varrho_j(x)| + \sum_{j=m_e+1}^{m} \varrho_j |\varrho_j(x)|_{\_} \tag{4.9}$$

with $|\varrho_j(x)|_{\_} = |min(0, g_j(x)|$, as a merit function $\varphi = R^1 \rightarrow R^1$

$$\varphi(\alpha) = \phi(x^k + \alpha^k d^k) \tag{4.10}$$

with x^k and d^k fixed, leads to a stepsize α guaranteeing global convergence for values of the penalty parameters ϱ_j greater than some lower bound.

We used a Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximation to update the Hessian matrix required for the local quadratic approximation.

We used the following parameters as inputs to the optimiser: iterations = 1500, accuracy = 1e-20, epsilon = 1.49e-08.

4.3.2 Bayesian Optimisation

Bayesian Optimisation is concerned with finding the minimum of a function $f(x)$ for some bounded set, χ, constructing a probabilistic model for $f(x)$ and then exploiting this model to make decisions about where in χ to evaluate the function, while integrating out uncertainty.

A Python implementation of Bayesian global optimisation with Gaussian processes was used. The algorithm was proposed in [107, 108] and the implementation is available from [109]. The main characteristics of the algorithm are as follows:

- It constructs a posterior distribution of functions (Gaussian process) that best describes the function the algorithm is trying to optimise.
- The posterior distribution improves as the number of observations grows. This implies that the algorithm becomes more certain of which regions in parameter space are worth exploring and which ones are not.
- The algorithm balances exploration and exploitation by taking into account what it knows about the target function. This means that, at each step, the algorithm is able to determine the next point that should be explored.
- It is adequate for situations in which sampling the function to be optimised is very expensive.

4.3.3 Baseline: Steered Response Power (SRP)

We used the implementation of SRP proposed in [59], whose implementation is available in MATLAB. As explained previously in Section 4.2, [59] proposes a stochastic region contraction (SRC) implementation of Steered Response Power Phase Transform (SRP-PHAT). The idea is that given an initial rectangular search volume, using an iterative process, the original volume will be contracted until a small subvolume is reached, so that the goal optimum is found inside it. The parameters used from this algorithm are: lower rectangular search boundary: (-2, -1, 0), upper rectangular search boundary: (2, 1, 4), number of random search points: 2500, best N points: 25.

4.4 Experimental Results

We performed our experiments using an *gfai tech AC_Pro Acoustic Camera System* consisting of 72 microphones sampled at 192 kHz. We used three different microphone configurations: ring, wheel, and spiral, located in the xy plane (with $z = 0$) and spanning the same area as illustrated by Fig. 4.1. For each configuration, we measured recorded sounds played by a *Bose Soundlink Bluetooth Mobile Speaker II, Model 404600* in five different calibrated positions within a room of size $12m \times 7m \times 3m$. The speaker was positioned, using a tripod, to be on the plane $y = -0.32$ for all

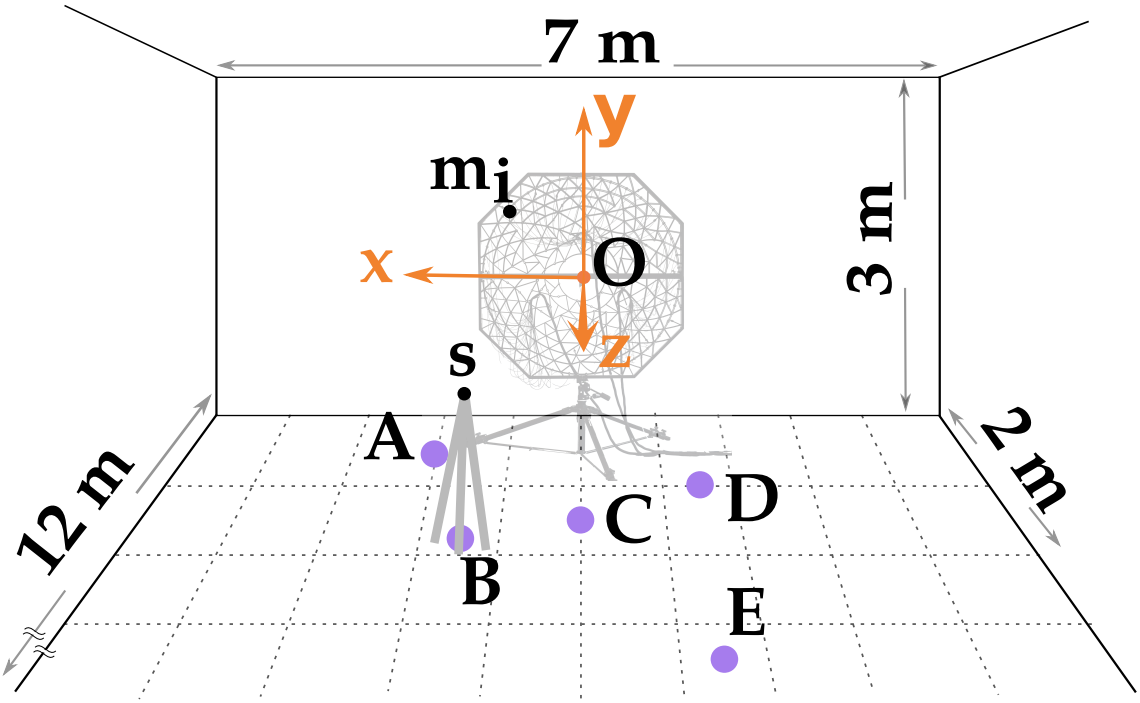


FIGURE 4.2: **Experimental setup and coordinate system (orange).** The room size is $12m \times 7m \times 3m$. The speaker was positioned, using a tripod, to be on the plane $y = -0.32$ for all five positions A, B, C, D and E (purple dots).

five positions A, B, C, D and E. For each position we acquired three recordings. Fig. 4.2 illustrates the setup. We repeated the experiments for 4 different audio signals [110]: chirp, gunshot, dogbark and speech.

We start by comparing various optimisation techniques to estimate the source location. Moreover, we evaluate the best technique in different microphone configurations in order to evaluate the impact of the configuration on localisation accuracy. Finally, we evaluate the optimisation technique with real data and we compare with Steered Response Power (SRP).

4.4.1 Simulation of Noisy TOA and TDOA

We tested the proposed optimisation by evaluating the relative error in localisation for different simulated degrees of noise σ in the estimated Times of Arrival (TOA) and Time Difference of Arrival (TDOA) values. To enable comparison across multiple source locations, we express σ for each source location as a percentage of the time taken for sound to travel from $\mathbf{s}$ to the center of the microphone array $\mathbf{O}$. We use a Gaussian model for the noise [111] in simulated TOA $\tilde{t}_i = t_i + \eta$ and for TDOA $\tilde{\tau}_{ij} = \tau + \eta$ where:

$$\eta \sim \mathcal{N}\left(0, \frac{\sigma}{100}\frac{\|\mathbf{s}-\mathbf{O}\|}{c}\right). \tag{4.11}$$

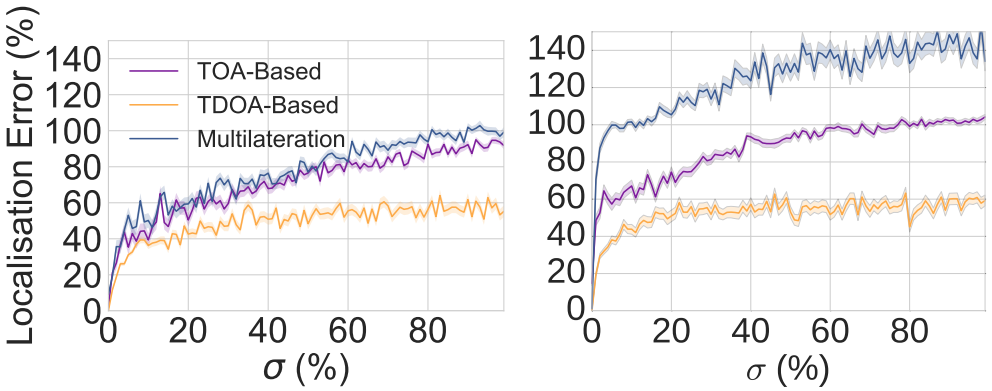
We measure relative error, expressed as a percentage of the distance from the source to the camera, as the evaluation metric for the accuracy of localisation:

$$\text{error}(\%) = \frac{\|\mathbf{s}-\tilde{\mathbf{s}}\|}{\|\mathbf{s}-\mathbf{O}\|} * 100, \tag{4.12}$$

where $\tilde{\mathbf{s}}$ is the source location estimated by the optimisation.

The relative error is used because we need to compare sources located in different positions across the space. Since the TDOA is affected by these positions, such that some TDOA are smaller than others, we decided to use relative error as a way to standardise the error and make it comparable across positions.

We started by comparing optimisations for TOA and TDOA with multilateration [2]. Our hypothesis is that when the microphones and the source are synchronised, that is, $t^* = 0$, the localisation error is going to be similar for most of the methods, but that when the microphones and the source are not synchronised, the error is going to be much higher for all the methods except the TDOA-based optimisation. Fig. 4.3 illustrates this comparison, by depicting plots of relative localisation error (Y-axis) as the noise in the simulation is increased (X-axis). We performed two



(a) synchronised ($t^* = 0$) (b) unknown t^* (10ms)

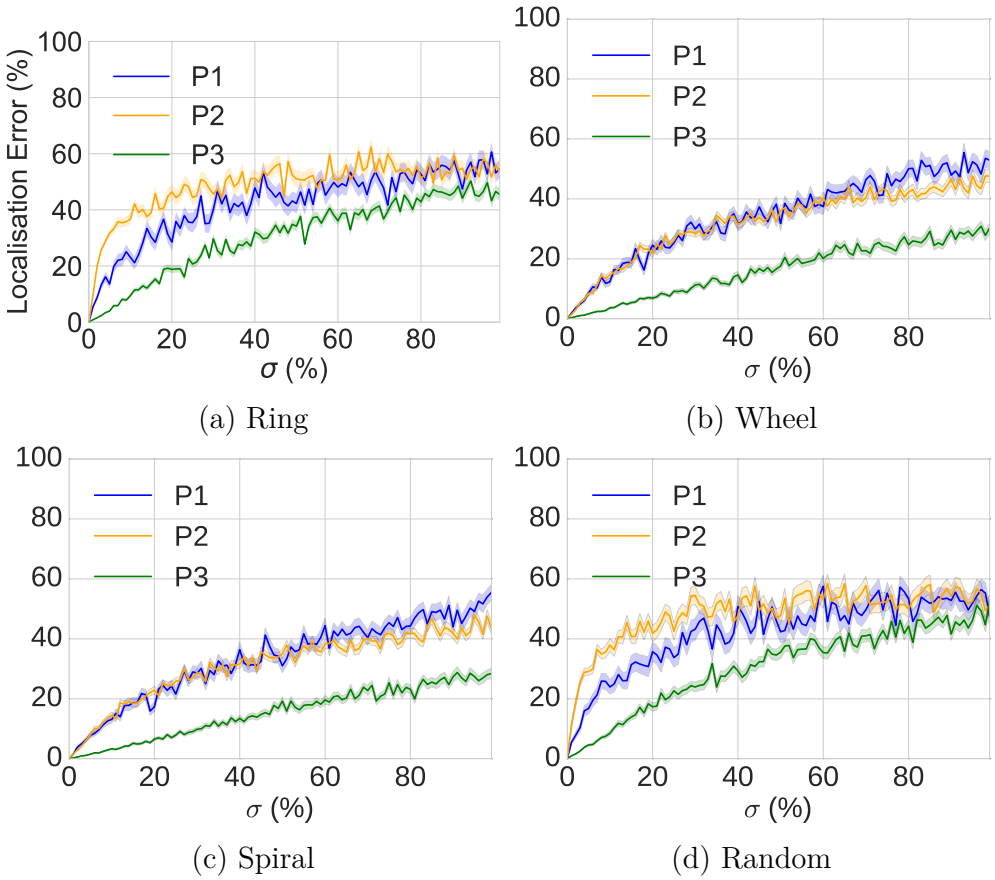
FIGURE 4.3: **Relative localisation errors.** Using O_1 (TOA), in purple, O_2 (TDOA), in orange, and multilateration [2], in blue, in the event that (a) speaker is synchronised with microphones and (b) time of emission is unknown.

versions of the experiment: one assuming that the microphones and the sound source are synchronised ($t^* = 0$ in Fig. 4.3a), and one without that assumption by setting $t^* = 0.01s$. This confirms our hypothesis that TDOA-based optimisation accurately locates the source when the microphones and the source are not synchronised.

4.4.2 Simulation of Microphone Configurations

We decided to estimate the localisation error at different points in space, obtained via simulation, for four different microphone configurations: ring, wheel, spiral and random. The microphones are arranged in the xy plane with $z = 0$. Our hypothesis was that the use of certain microphone configurations would yield better localisation accuracy than others for the exact same source location. We started with three positions $P1 \equiv (-2, -1, 4)$, $P2 \equiv (-1, 0.5, 3)$ and $P3 \equiv (0.4, 0.7, 1.05)$. We simulated some noise into the TDOA estimation, using the criteria previously presented in Section 4.4.1, and proceeded to estimate the localisation relative error (Eq. 4.12). Fig. 4.4 illustrates the obtained results. We plotted the localisation relative error as a function of noise for the four above-mentioned microphone configurations. This confirms our hypothesis that the ring configuration produces less accurate results than wheel and spiral, when evaluated for the same source location and with the same amount of noise in the TDOA estimation.

We decided to do further experiments to strengthen the confirmation of our hypothesis. To do this, we simulated a source located in each position of a grid of



(a) Ring (b) Wheel

(c) Spiral (d) Random

FIGURE 4.4: **Relative localisation error for increasing noise at three source locations.** P1: (-2,-1,4) in blue; P2: (-1,0.5,3) in yellow; P3: (0.4,0.7,1.05) in green.

points inside a room of dimension 2m $\times$ 2m. For each source position on the grid, we estimated the localisation relative error (Eq. 4.12) for three different microphone configurations: ring, wheel and spiral. The three configurations were identical to those used for real measurements with our acoustic camera, consisting of 72 microphones. Each configuration results in different TOA and TDOA values, due to the different microphone positions, and therefore we used the criteria previously presented in Section 4.4.1 to end up with a relative error that is comparable amongst source locations and microphone configurations. Our hypothesis for this experiment, judging by the results obtained previously in Fig. 4.4, is that the localisation error would be larger for the ring configuration for some region in the grid compared to spiral and wheel configurations. Fig. 4.5 visualises the resulting heatmaps for $\sigma = 0\%, 25\%, 50\%, 75\%, 100\%$ TDOA simulated error. This figure confirms our hypothesis since, when noise is added to these TOA and TDOA values, each

configuration reveals a characteristic heatmap for localisation relative error over space. The errors were averaged over 100 trials for each grid point. We chose a grid over $x = [-2, 2]$, $z = [0, 4]$ and $y = -0.32$, with a resolution of 10 cm, for a total of 1600 source locations, so that it matches our experiments with real data that will be presented in Section 4.4.3. Using the values obtained on the heatmaps, we represented the relative localisation error using histograms, as illustrated in Fig. 4.6, in order to observe the distribution of the error.

4.4.3 Real Data and Comparison with Steered Response Power (SRP)

We proceeded to perform experiments with real data in order to validate our hypothesis in real scenarios. Therefore, we used optimisation scheme O_2 to localise a speaker placed in five different positions $A \equiv (2.0, -0.32, 0.5)$, $B \equiv (1.5, -0.32, 2.0)$, $C \equiv (0.0, -0.32, 1.5)$, $D \equiv (-1.5, -0.32, 1.0)$ and $E \equiv (-1.5, -0.32, 3.5)$. These locations represent a variety of scenarios in which the impact of the three different microphone configurations is illustrated, according to the results obtained in Fig. 4.5. Our hypothesis, given the previous results, is that when the source is located in front of the microphone array, the localisation accuracy will decrease for any sound class when the ring configuration is used and it will remain the same for other configurations. Fig. 4.7 plots relative errors (Y-axes) for three different microphone configurations (X-axes) at the chosen five locations (columns). The three rows of plots correspond to results obtained using SLSQP, SRP [59] and Bayesian optimisation [107] respectively. Error bars (standard deviation) are shown with black lines on top of the bars. These plots confirm our hypothesis that when using ring configuration, the error increases for a source located in front of the microphone array, such as $C \equiv (0.0, -0.32, 1.5)$, while it is not affected in other configurations, such as wheel and spiral.

4.5 Discussion

4.5.1 Microphone Configurations

Our results suggest that circular (ring configuration) arrays perform worse than spiral or wheel configurations when considering relative localisation error over a wide range of positions. Our simulation results (Fig. 4.5 and Fig. 4.6) show regions (top view) that are prone to error when using circular arrays. In Fig. 4.5 for example,

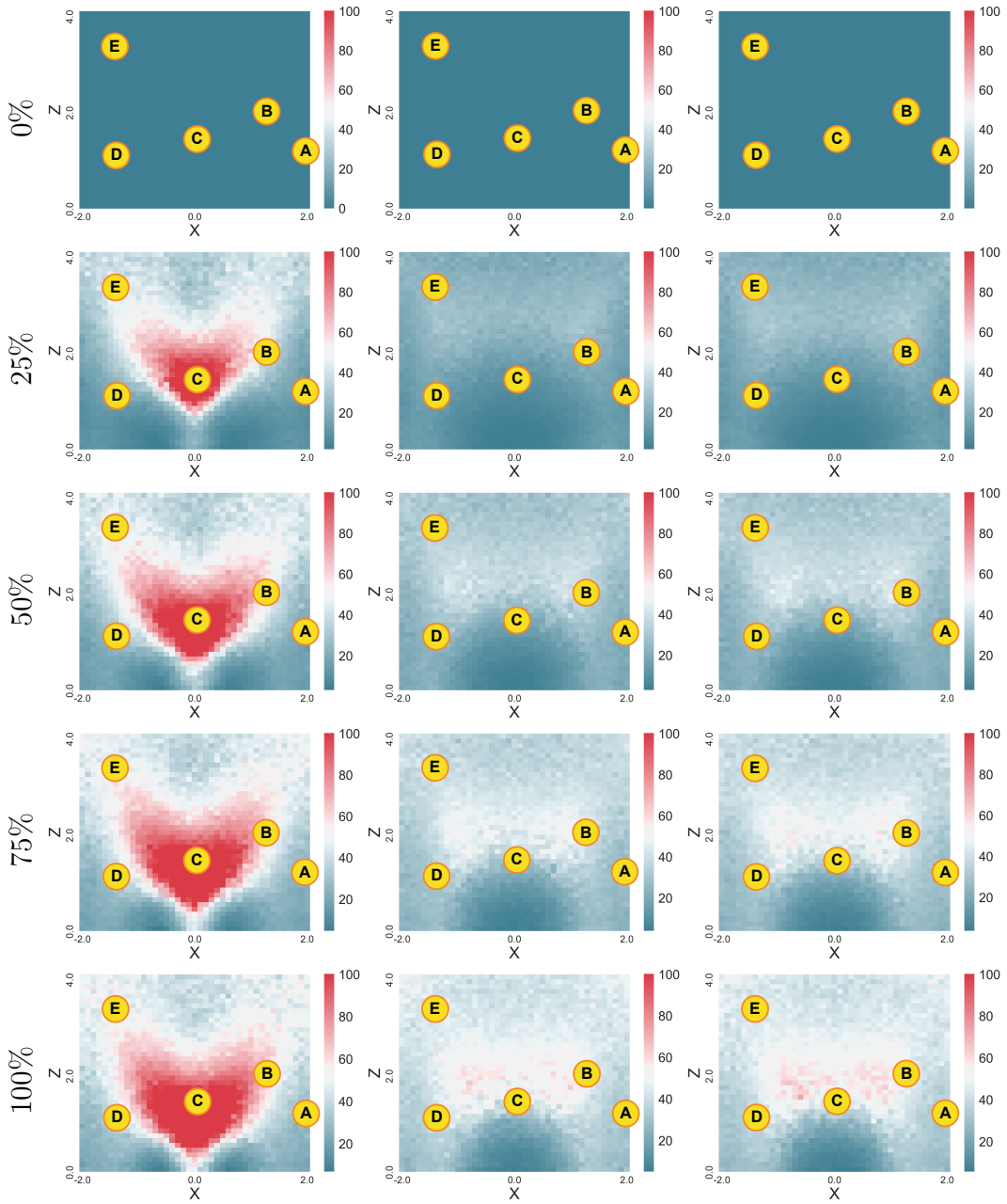


FIGURE 4.5: **Relative localisation error visualised as heatmaps for a 2m x 2m room.** Simulation of TDOA with various noise levels (0%, 25%, 50% 75% and 100%), expressed as percentages as explained in Section 4.4.1. Each location (x, z) in the heatmap represents a source location inside a 2m $\times$ 2m room. y is a fixed value for all the heatmaps, equal to -0.32. 100 estimates were averaged to determine the error estimate at each grid position. High levels of error are presented in red and low ones in blue.

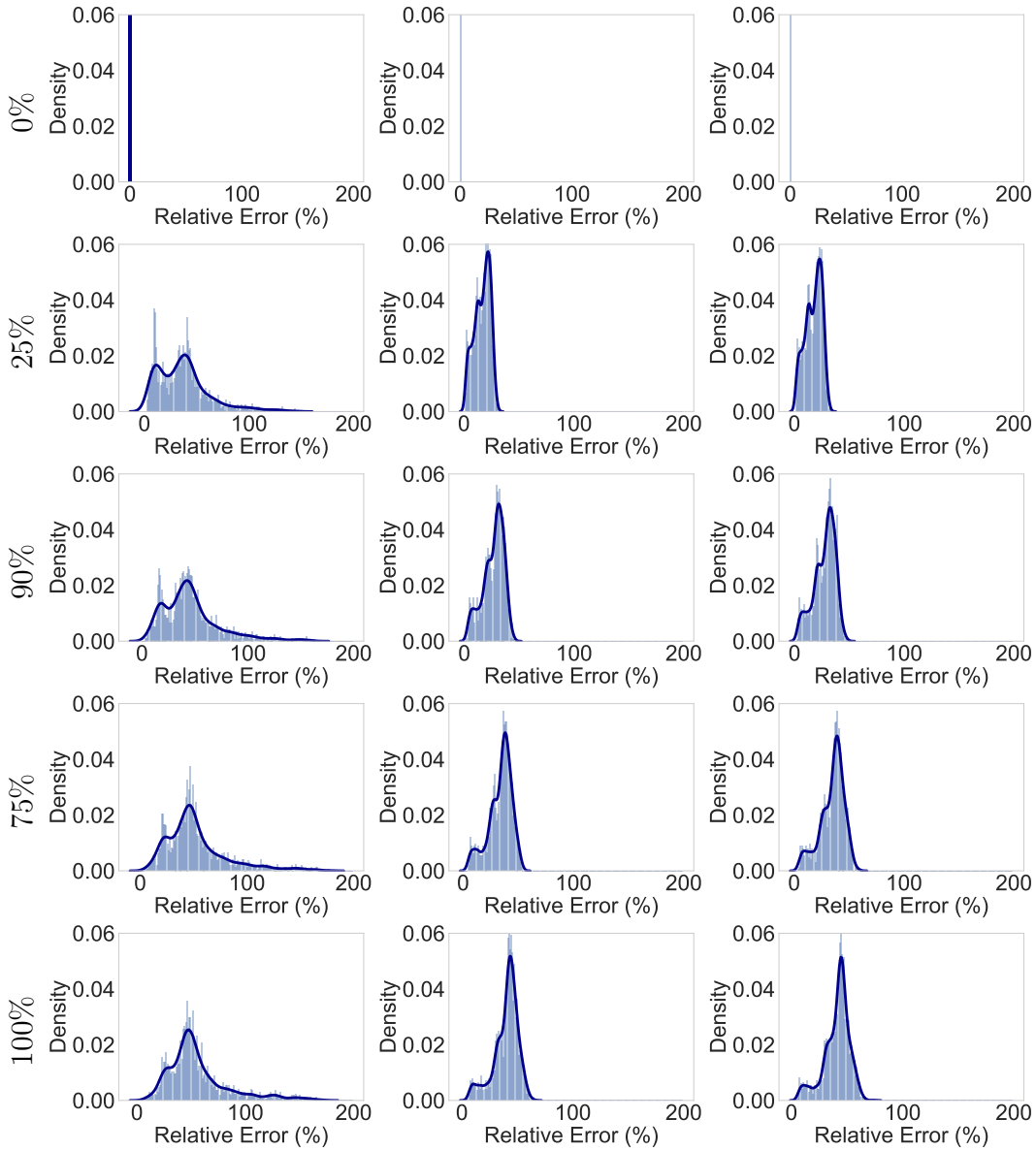


FIGURE 4.6: **Relative localisation error visualised as histograms (blue) for a 2m x 2m room.** Simulation of TDOA with various noise levels (0%, 25%, 50% 75% and 100%), expressed as percentages as explained in Section 4.4.1. The histograms depict the localisation error for 1600 source locations inside a 2m $\times$ 2m room. 100 estimates were averaged to determine the error estimate at each source position.

even for a small amount of noise (25% on the second row), it can be seen how the error in the ring configuration increases for a vast portion of the grid points, while for the wheel and spiral configurations the error remains below 50%. The histograms

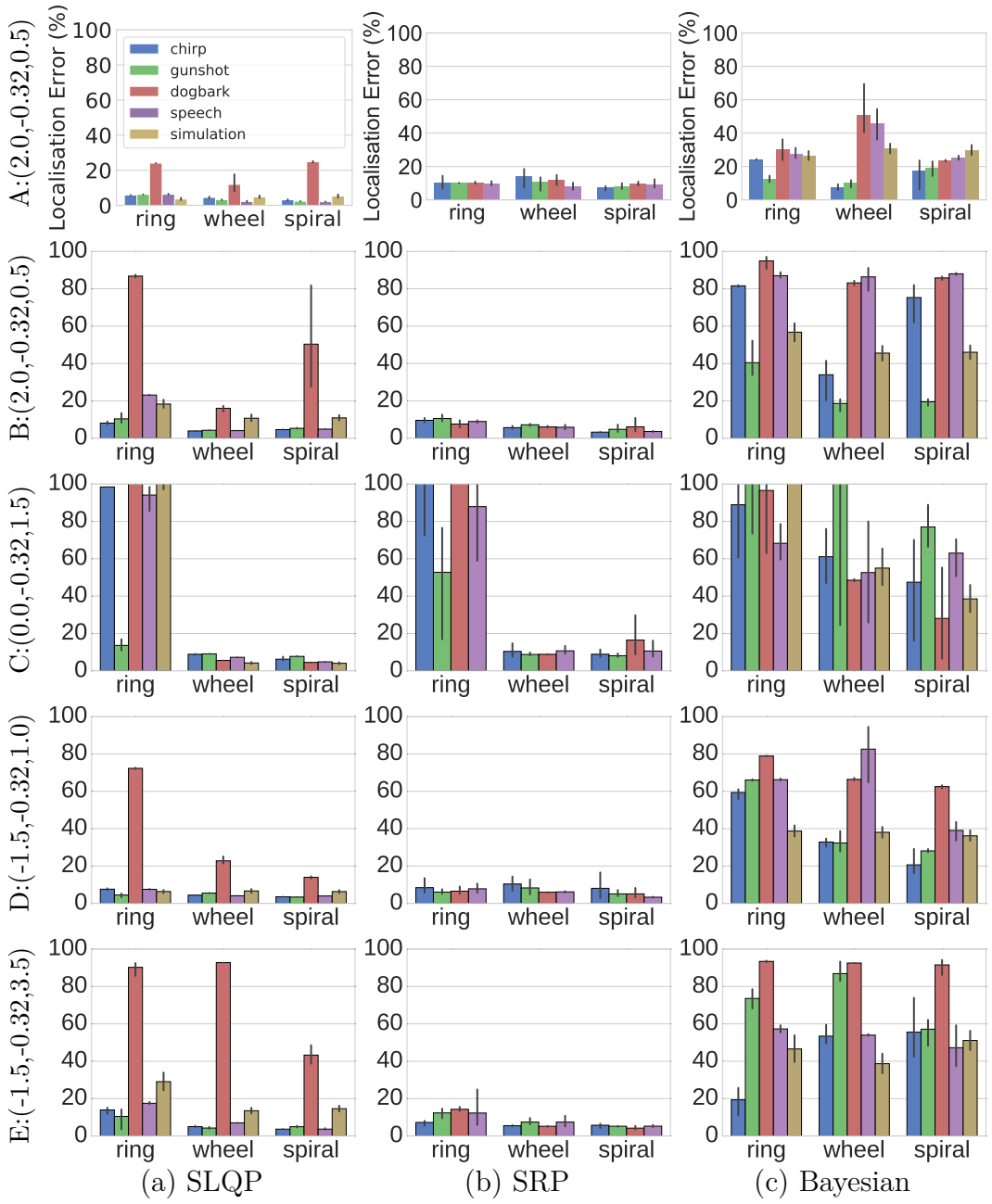


FIGURE 4.7: **Multilateration vs SRP vs Bayesian Optimisation.** Localisation Error using SQLP and simulation (left) SRP (middle) and Bayesian Optimisation (right), for source locations: A:(2.0,-0.32,0.5) 1st row; B: (1.5,-0.32,2.0) 2nd row; C: (0.0,-0.32,1.5) 3rd row; D: (-1.5,-0.32,1.0) 4th row; E: (-1.5,-0.32,3.5) 5th row in four different audio classes: chirp (blue), gunshot (green) dogbark (red), speech (purple), as well as in simulated error (yellow).

in Fig. 4.6 reassert this, since they show how the error distribution for the ring configuration has a tendency to be uniform between 0 and 100, while for wheel and

spiral configurations they present peaks around 50%.

This is also true of our real measurements (Fig. 4.7), where the results obtained for position C are worse for the ring configuration than for the wheel or spiral configurations using any of the three localisation techniques. The yellow bars in the first column show that the errors observed with real data correspond to errors obtained with about 10% noise in our simulation.

Additionally, Fig. 4.7 also illustrates how positions not directly facing the microphone array (e.g. A and D) yield better localisation accuracy even when the ring configuration is used. A clear exception occurs for the dogbark dataset, in which the error is high for all the three different configurations. This is explained by the histograms in the Appendix A, Fig. A.1, A.2 and A.3, which illustrate the Time Difference of Arrival (TDOA) relative error for each dataset. In each source location, it can be seen how the histogram shows a greater error for the dogbark dataset, arising from the use of the Generalized Cross-Correlation Phase Transform (GCC-PHAT) and the repetitive pattern of the signal. On the other hand, these histograms also show how the rest of the signals present a low TDOA relative error for the three different microphone configurations, albeit the ring configuration performs much worse than the rest of them.

TABLE 4.1: **Error using all pairs.** Table comparing errors and time for SRP as against TDOA optimisation using all pairs. Standard deviations are shown within parentheses.

signal	**SRP**		**TDOA** (all)	
	Rel. Err %	Time in min	Rel. Err %	Time in min
chirp	14.7 (25.9)	3 (0.2)	12.1 (23.2)	4.5 (0.03)
gunshot	11.0 (13.3)	2.58 (0.2)	6.4 (3.5)	2.4 (0.02)
dogbark	16.0 (28.5)	2.49 (0.1)	48.5 (44.6)	2.4 (0.02)
speech	13.2 (21.1)	2.63 (0.1)	12.9 (22.5)	2.5 (0.02)

4.5.2 Comparison with Multilateration

Our experiments showed that both optimisation strategies O_1 and O_2 result in lower relative errors than state of the art multilateration [2]. This is particularly true when the time of emission of the signal is unknown and when the emitter is not synchronised with the microphones ($t^* \neq 0$). When $t^* = 0$, our implementation of the multilateration algorithm has similar accuracy to optimising O_1 (Times of

Arrival (TOA)). Our proposed approach to optimising O_2 (TDOA) has the least degree of relative error and remains unaffected by t^*.

4.5.3 Comparison with Steered Response Power (SRP)

A common criticism of indirect methods is that the optimisation is not as robust as direct methods such as SRP. However, our results in Table 4.1 and Table 4.2 show that our localisation error is comparable to SRP while being more efficient for most of the signals we tested (with the exception of the dogbark). We used an efficient CPU implementation of SRP in MATLAB that leverages stochastic region contraction [59] and a naïve CPU implementation of our optimisation in Python. In both cases, the accuracy of the proposed optimisation may also be traded for performance. The experiments were carried out on a computer with 4th Generation Intel(R) Core(TM) i7-4790 processor and 24GB Dual Channel DDR3 1600MHz memory.

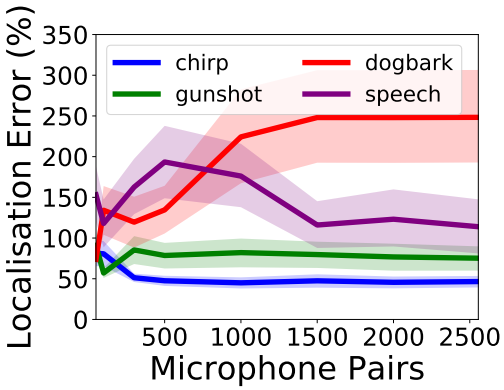


FIGURE 4.8: **Localisation accuracy vs microphone pairs.** Localisation accuracy for various numbers of microphone pairs. In cases of accurate TDOA estimation, such as chirp and gunshot, the curve stabilises when a relatively low number of microphone pairs (100) has been used. In the case of more challenging datasets, such as speech and dogbark, the use of fewer microphone pairs decreases the localisation error, since a large number of microphone pairs introduces more noise to the source estimation.

4.5.4 Accuracy vs Performance

One way to approximate the localisation is to modify the nested summation in O_2 to consider only some of the microphone pairs. We studied convergence plots of localisation error for different source positions, as the number of microphone pairs

is increased from just 1 pair to all pairs (C_2^{72}). Fig. 4.8 illustrates the localisation relative error for various numbers of microphone pairs. In cases of accurate TDOA estimation, such as chirp and gunshot, the curve stabilises when a relatively low number of microphone pairs (100) has been used. In the case of more challenging datasets, such as speech and dogbark, the use of fewer microphone pairs decreases the localisation error, since a large number of microphone pairs introduces more noise to the source estimation.

In Table 4.1 we run our algorithm using all possible microphone pairs and report our results in relative error accuracy and estimated time in minutes to execute the algorithm. In Table 4.2 we used the same metrics, but this time using only 100 microphone pairs. These microphone pairs were randomly chosen, without any prior consideration whether some microphones pairs might produce more accurate TDOA estimations or not. The results showed that in the latter scenario the error was comparable to that obtained when using all the microphones pairs, except with a sixfold increase in the efficiency of the computation time. The error generally drops below 20% for 100 mic pairs (see Table 4.2 for computation times), except for the dogbark signal, which exhibits high TDOA relative error calculation (see Appendix A, Fig. A.1, A.2 and A.3). Fig. 4.9a plots relative error averaged across spatial locations for all four test signals using only 100 microphone pairs.

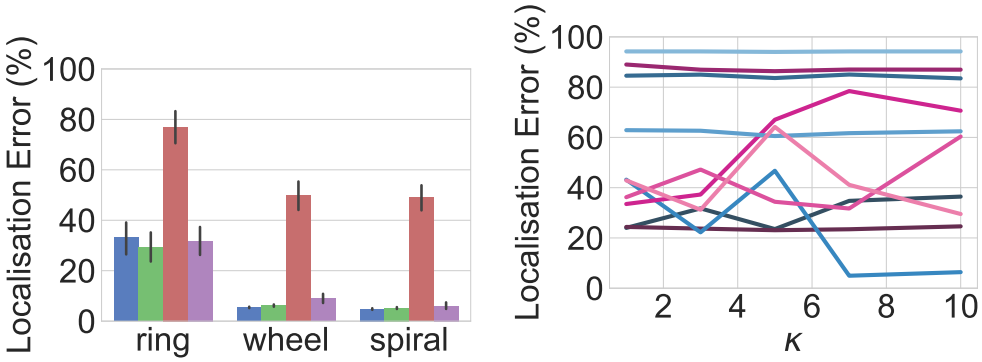
TABLE 4.2: **Error using 100 pairs.** Table comparing errors and time for SRP as against TDOA optimisation using 100 of the C_2^{72} microphone pairs. Standard deviations are shown within parentheses.

signal	**SRP**		**TDOA** (100)	
	Rel. Err %	Time in min	Rel. Err %	Time in min
chirp	14.7 (25.9)	3 (0.2)	14.2 (25.9)	0.5 (0.01)
gunshot	11.0 (13.3)	2.58 (0.2)	9.6 (12.8)	0.4 (0.02)
dogbark	16.0 (28.5)	2.49 (0.1)	58.9 (38.8)	0.4 (0.02)
speech	13.2 (21.1)	2.63 (0.1)	15.2 (23.5)	0.4 (0.02)

4.5.5 Bayesian Optimisation

We tested a Bayesian optimiser described in Section 4.3.2 with O_2 as its loss function ($\kappa = 1$). This took an order of magnitude longer than SQLSP and the resulting errors were larger. We tested with various degrees of the κ parameter to trade off exploitation versus exploration. The plot (Fig. 4.9b) shows that exploitation ($\kappa = 1$) performs better than exploration ($\kappa = 10$) in most cases. The number of iterations

and tolerance were set so that the optimiser converged to the reported solutions, suggesting that the problem is not due to multiple local minima.



(a) Using 100 mic pairs (b) Bayesian Optimisation

FIGURE 4.9: **Exploration vs Exploitation.** (a) Errors (real data) for four signals: chirp (blue), gunshot (green) dogbark (red), speech (purple), across spatial locations. (b) Exploitation ($\kappa = 1$) vs exploration ($\kappa = 10$) for dogbark (blue) and speech (purple) for spiral configuration.

4.5.6 Limitation

One drawback of indirect localisation achieved by minimising O_2 is its dependency on the estimated TDOA values. Although our results show that GCC-PHAT is accurate enough to yield localisation errors comparable to SRP, the former performs worse when dealing with signals with repeating patterns such as the barking of a dog (red bar in Fig. 4.7). Our localisation was more robust to reverberation (when the source was placed at room boundaries) than to repetitive macro-structures. Perhaps using full signal correlation matrices, as adopted by spectral estimation techniques, would resolve this problem.

4.6 Conclusions

In this chapter, we presented novel findings regarding direct optimisation of Acoustic Source Localisation (ASL) and the impact of microphone array configuration on localisation accuracy.

First of all, we used a direct optimisation of ASL to find the source location. Using various numbers of microphone pairs, we found that, when using 100 randomly

chosen of these pairs, the localisation accuracy is the same as for Steered Response Power (SRP), but 6 times faster in three out of four different datasets.

Parallel to this, we observed that, for some random source locations, there was an increase in the relative error in localisation when the ring configuration was used, compared with other configurations, such as wheel and spiral. Moreover, when these results were extended in order to visualise the error in a simulated room, it was found that the pattern persists, particularly when the sources are located in front of the microphone array (in the xy plane, with $z > 0$). Finally, experiments with real data showed that, for various audio classes, the localisation error is much higher when a source is placed in front of the microphone array and the ring configuration is used.

We tested both parts of our system (microphone configuration and localisation) using a microphone array to record sounds from a static source. Moreover, the sounds recorded were from a variety of audio classes from real life scenarios. Therefore, we believe that our findings could be used and applied to any microphone array system.

Our contribution, then, could be summarised as follows:

- we showed that using a limited number of microphone pairs and Sequential Least Squares Programming (SLSQP) yields accurate source localisation, but using 6 times less computation than an optimised version of SRP.

- we demonstrated that circular microphone arrays are the least desirable configuration for ASL, since noise in the Time Difference of Arrival (TDOA) estimation leads to higher localisation error than in other configurations, such as wheel and spiral.

In conclusion, we have shown that direct optimisation of the well known formulation for ASL yields errors similar to the state of the art (SRP) with 6 times less computation. Moreover, we demonstrated, using both simulated and real data, that the method is robust to noise and reverberation. Our results have shown that circular arrays are the least desirable configuration.

In the future, we plan to perform further experiments in a wide range of scenarios, including different microphone array sizes, to generalise the performance limitations of the ring array. Moreover, the approach could also be extended to estimate angle errors, in 3D, in order to complement the relative localisation error already presented.

Chapter 5

Signal Samples Selection for TDOA Estimation

5.1 Introduction

The ability to ascertain the position of objects by relying only on the sounds emitted by these objects has many applications, e.g. smart assistants (Amazon Echo-7, Google Home-2, Apple Airpods-2) [7], 3D reconstruction via SONAR, multilateration for detecting hostile targets, and many others [112, 113]. *Acoustic localisation* is possible because sounds emitted by an object (source) reach multiple sensors (microphones) at different times. When the relative positions of the sensors is fixed (or known), the goal is to identify the relative delay, or Time Difference of Arrival (TDOA), in events recorded by different pairs of microphones [114] and to leverage this information to obtain the source location. This is particularly challenging in the presence of noise [115] and/or reverberations [116]. An implicit, and common, assumption is that there is access to all the signals recorded by all the microphones. However, it is wasteful (both in terms of bandwidth and energy) and often impossible for all sensors to transmit all recorded data. We therefore address the problem of accurate acoustic localisation under constraints on the amount of data that is broadcast by the sensors.

Traditionally, the methods used to calculate TDOA rely on Generalized Cross-Correlation (GCC), using the entire signal to perform the estimation. These techniques yield good results and are used in wide-ranging scenarios for various types of microphone arrays. Section 5.2 presents a more in-depth review of these methods. In current audio systems, however, the amount of data available has increased considerably, as a result of the boost in sampling frequency [117] and the amount of sensors available. In scenarios in which the data transmission is constrained or the bandwith is low, this presents a potential problem. Examples of such scenarios

include the use of microphone arrays in robots [118–120], underwater sensors [121] and Unmanned Aerial Vehicles (UAVs) [122, 123]. Although there are approaches that include data compression (as presented in Section 5.2), in most of the cases they have not been tested in a large variety of scenarios, as we will further explain.

Our hypothesis is that it is not necessary to use the whole signal to accurately estimate TDOAs. The idea behind this is that it would require only a small portion of the signal (e.g. a word) detected at each sensor to calculate the TDOA.

Our main contributions in this chapter are:

- Determining the signal keypoints to be transmitted in order to obtain an accurate TDOA estimation, at significantly lower data rates or improved accuracy compared with GCC based solutions.
- Demonstrating the robustness of the proposed technique to different noise and reverberation conditions.
- Comparing the proposed technique with another data-driven approach, that of audio fingerprinting.

5.2 Related Work

In this section, we summarise the literature related to Time Difference of Arrival (TDOA) estimation, as well as the use of Scale-Invariant Feature Transform (SIFT) features on a signal spectrogram.

5.2.1 Time Difference of Arrival (TDOA) Estimation

The literature in source localisation is rich in approaches dedicated to estimating the TDOA. The first and probably most well-known family of methods are those based on Cross-Correlation (CC), which uses microphone pairs to estimate TDOA. It can be applied in the time domain, in which case it is known as Generalized Cross-Correlation (GCC) [44], or it can be applied in the frequency domain, in which case a spectral normalisation is used, referred to as the Generalized Cross-Correlation Phase Transform (GCC-PHAT) [47]. An extension of these approaches uses the redundant information across multiple sensors in order to estimate TDOA.

There are a group of methods that rely on geometrical information to calculate the TDOA. In [64–66], the authors present a method that uses the positions of a non-coplanar microphone array to formulate a constrained optimisation problem

that simultaneously estimates the TDOA and the source position. Similarly, in [124], the authors create a TDOA mapping from a 2D scenario to a range of estimations (Times of Arrival (TOA)). They extend this approach in [125] to remove outliers.

Although in most of the cases these algorithms estimate the TDOA with a high degree of accuracy, their main limitation is that they rely on the use of the entire signal. This is a problem in scenarios in which transmission is constrained or the bandwidth is low.

5.2.2 Time Difference of Arrival (TDOA) Estimation Using Feature-Based Approaches

There are a group of methods that calculate the TDOA by locating features in each audio signal [114]. These features are then further matched according to their similarity, measured by a variety of metrics. Using the time at which the feature is present in the signal, the TDOA is calculated on the basis of the difference between those times.

In [67], the authors propose a method for the self-localisation of an ad-hoc network of randomly distributed devices in an open space with low reverberation but significant noise. The device localisation framework calculates the distance using the difference between the maximum TDOA and the minimum TDOA. First of all, the Short-Time Fourier Transform (STFT) of each signal is calculated, then a group of spectral peaks are selected and finally pairs of these peaks form an audio landmark [126]. The matches of landmarks across multiple microphone pairs is used to estimate the TDOA. The framework is set to work only when the sources are in end-fire locations, the number of sources is sufficient and the reverberation is low. A fairly similar approached is presented by Wang et al. in [127, 128], in which a database is constructed based on the fingerprint positions. The localisation consists on estimating the fingerprinting of the current audio signal and finds the closest match in the database. This approach is however very limited, since it is highly dependent on the construction of the initial database.

In [129], a different application of audio fingerprinting landmarks is introduced: alignment of unsynchronised meeting recordings. The authors combine audio fingerprinting with spectrotemporal eigenfilters to create an unsupervised learning algorithm. The method begins by performing a rough alignment of the signal, before improving the alignment using Hamming distance.

In [130], the authors propose a source localisation and counting approach that combines time-frequency (TF) clustering on the signal spectrogram with GCC-PHAT

to estimate TDOA. Their results are mainly focused on the correct estimation of the number of sources, using the F-score to calculate the number of successes.

A review of binaural source localisation methods is introduced in [131], where the Interaural Level Difference (ILD), Interaural Time Difference (ITD) and Interaural Phase Difference (IPD) are presented as main features for performing horizontal localisation. In [132], this approach is extended by obtaining a composite feature vector derived from analysing the mutual information between different spatial cues and estimating the optimum feature combination that minimises the angular localisation error in three-dimensional space. This allows these features to work at different noise levels. The experiments do not consider reverberation, however, and are performed only on steady sources.

In [133], a Time Delay Estimation (TDE) algorithm is proposed, based on the deduction of a Multichannel Frequency-domain Adaptive Filter, representing the impulsive characteristics of the speech spectrum. After the signals are filtered, the TDOA is calculated by comparing the time differences of the direct-path components between different channels. The experiments are conducted using a single sound source, with simulated added White Gaussian Noise.

5.2.3 Fingerprinting Variations

Our study focuses on using audio fingerprinting as a baseline for comparison. Therefore, we have investigated the methods that present variations of it to gauge how much the method could be improved. Since it is an audio retrieval algorithm, the changes are mostly to improve the matching of songs under different constraints, but no significant change could improve the TDOA calculation.

The main advantage with fingerprinting is that it remains robust in the face of differing variations in the song (e.g. acoustic versions, background noise, etc.). In [134] for example, the authors are focused on improving audio fingerprinting for pitch shifting by means of cosine filters. Similarly, in [135] the focus is on handling time-scale and pitch modifications by drawing on three components: local maxima in a time-frequency representation, triplets of events and Constant-Q transform. Lastly, Sonnleitner et al. [136] aim for robustness in the face of large time and frequency scale distortions, as well as efficient operation for large reference audio collections, by choosing groups of four peaks (instead of the traditional peak pairs). Since in our problem there are no considerable modifications to the pitch, these improvements to the fingerprinting method are not relevant for TDOA estimation.

Another group of approaches are focused on content-based copy detection (CBCD). One example of this is [137], in which the authors generate different copies of the spectrogram (for various noise values) and evaluate whether or not it is a copy by using the preserved fingerprintings. In [138], they use a different representation of signals called time-chroma representation, in which severe pitch and tempo change is calculated using fingerprinting. Finally, in [139] salient regions of binary images derived from the spectrogram matrix are combined with fingerprinting to identify copies.

In [140], fingerprinting is based on wavelet creation. Different wavelets need to be created, therefore a lot of computation is needed and the top-t wavelets are selected, making this algorithm unsuitable for compression.

5.2.4 Scale-Invariant Feature Transform (SIFT) Using the Spectrogram of the Signal

The idea of using SIFT on the spectrogram signal has been explored in the past [141–143]. Authors have used it for different applications, which are briefly summarized below.

Early works include the use of the spectrogram as a 2D image and Viola-Jones is used to find a descriptor to perform music identification [141]. This was the beginning of applications relating to music, as in [142], in which SIFT features are extracted for various music applications including genre classification, music mood classification, and cover song identification. Similarly, in [143] the authors present a new algorithm for audio fingerprinting using SIFT features in the signal spectogram, overcoming two of three challenges in music identification: time stretching and pitch shifting. Additionally, Nguyen et al. [144] use SIFT on the signal spectrogram for speech classification (words classification) using the local naïve Bayes classifier.

5.2.5 Time Difference of Arrival (TDOA) Estimation with Compression

There have been some previous attempts to estimate TDOA using compressed approaches, not only for TDOA estimation, but also in wireless communications.

Early works include [145], in which the authors are focused on transmitting data from one sensor to another, using operational rate-distortion viewpoint with a distortion measure based on Fisher information of the estimation problem. They extend their method in [146–148]. Later, in [149], the authors employ an event

detection algorithm that estimates the point in time at which an event occurs (TOA). Taking this information from multiple sensors, the algorithm uses a consistency function to calculate which source location matches the estimated times, based on the amount of sensors with similar TOA. The main limitation of this approach is that the consistency function must be evaluated for every point in space at which the source might be present, making it computationally expensive. The authors transmit 1.1% of the raw signal, but they limited their experiments to a single scenario under specific noise and reverberation conditions. Similarly, Fuyong et al. [150] present a compression algorithm tested using compression ratios between $4:1$ and $8:1$. Additionally, there are authors who focus on sensor networks for low-bandwidth localization in [151, 152]; however, these approaches differ from ours in the sense that they are active sensing methods, therefore sensors may emit calibration signals. Lastly, in [153] vector quantisation is used to process the signal.

5.2.6 Summary

In general, it can be seen that the literature encompasses the use of the following approaches related to our work:

- TDOA estimation using standard techniques that rely on the use of the entire signal [44,47,64]. This does not consider scenarios in which there are constraints either due to transmission, or because the bandwidth is low.
- estimation of TDOA using feature-based approaches based on fingerprinting [67], time-frequency (TF) clustering [130], binaural features [131], mutual information [132] and filtering [133].
- SIFT features extracted from the signal spectrogram to perform music identification [142], music genre classification [143] and speech classification [144].
- TDOA estimation with compression using a consistency function [149] and Discrete-Time Fourier Transform (DTFT) [150].

However, there is a significant gap in the literature, since the use of SIFT features in the spectrogram for TDOA estimation with compression has not yet been explored. The rest of this chapter is dedicated to leveraging this using simulated and real speech signals.

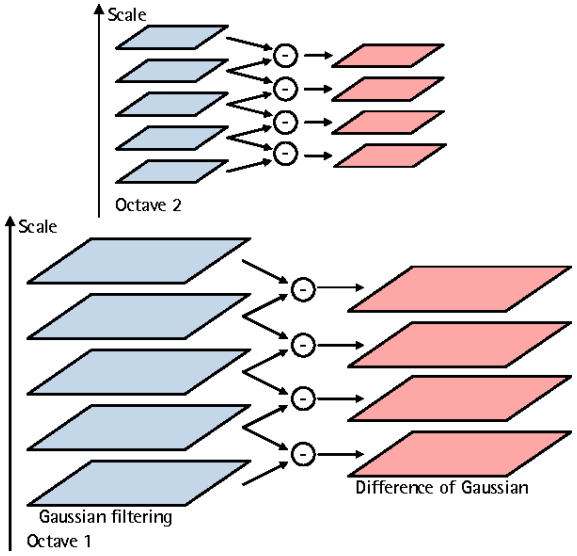


FIGURE 5.1: **Difference of Gaussians (DoG).** A new image (pink) is generated by subtracting two consecutive blurred images (blue) for each octave [3].

5.3 Methodology

This section summarises the developed algorithm, starting with a brief introduction to Scale-Invariant Feature Transform (SIFT), one of the core elements of the proposed approach. Afterwards, the algorithm is explained in detail, together with a summary of unsuccessful approaches. Following on from this, we explain the error metric considered to evaluate the algorithm. Finally, the spectrogram parameter selection is summarised.

5.3.1 Scale-Invariant Feature Transform (SIFT)

SIFT is a well-known feature detection algorithm in computer vision, used to detect and describe local features in images. It was originally proposed by David Lowe in [154] and it has been widely used since then.

SIFT is composed of a detector and a descriptor. The detector is in charge of selecting keypoints in the image while the descriptor assigns a feature vector to each of these points. Since only the detector will be used in our algorithm, a brief explanation of its working principle is presented below.

1. **Scale-space.** SIFT starts by taking the original image and adding some blur. Moreover, the original image is resized to half its size and the blurring is applied again to the down-sampled image. Each resized image contains various blur

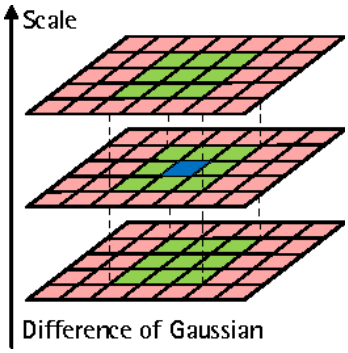


FIGURE 5.2: **Local maxima/minima in DoG images.** Checking the pixels on a 3 by 3 window (green) and comparing it with the neighbours above and below. The point is marked as a keypoint (blue) if it is the greatest amongst all 26 neighbours [3].

levels, which together form an octave. The author proposed originally 4 octaves and 5 blur levels.

2. **Difference of Gaussians (DoG).** Fig. 5.1 illustrates the calculation of the DoG, where a new image (pink) is generated by subtracting two consecutive blurred images (blue) for each octave.

3. **Keypoint localisation.** Keypoints are located at local extrema (maxima or minima) in DoG images. Fig. 5.2 illustrates how these points are detected: checking the pixels on a 3 by 3 window (green) and comparing it with the neighbours above and below. The point is marked as a keypoint (blue) if it is the greatest amongst all 26 neighbours. The subpixel maxima/minima is calculated using a Taylor expansion, that is, an interpolation of the function inside regions "in-between" pixels is estimated and the maxima/minima is calculated within this region, obtaining a subpixel as a result.

4. **Discarding low-contrast keypoints.** Since the amount of keypoints generated by the previous step is large, it is necessary to prune some of these points. First of all, the low contrast features are removed by defining a threshold and removing the pixel intensity of the original image below this threshold. Next two perpendicular gradients are calculated in order to remove edges. If both gradients are big, it means it is a corner, therefore the point is accepted as a keypoint, otherwise it is rejected.

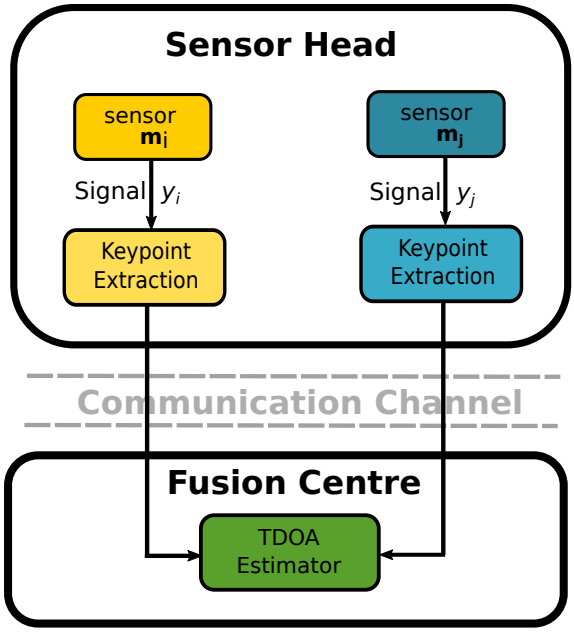


FIGURE 5.3: **Overview of the system architecture.** Keypoint extraction (yellow and blue) occurs at the Sensor-Head (SH). These keypoints are then communicated to a Fusion Centre (FC), which may be either a centralised node, or simply another sensor node, where the Time Difference of Arrival (TDOA) is calculated (green).

5.3.2 Algorithm Overview

The proposed approach is based on Fig. 5.3, in which keypoint extraction occurs at the Sensor-Head (SH). These keypoints are then communicated to a Fusion Centre (FC), which may be either a centralised node, or simply another sensor node. The communications channel is assumed to be low-bandwidth, such that minimal communication is desirable to ensure low-latency in the full localisation system. The sensors considered in this chapter are microphones, but could naturally be any passive transducer, such as hydrophones, or RF.

Algorithm 5.3.1 Calculate TDOA

function CALCULATE_TDOA(y_i, y_j)
  $p_i, p_j \leftarrow$ SPECTROGRAM(y_i, y_j) ▷ SH
  $f_i, f_j \leftarrow$ SIFT(p_i, p_j) ▷ SH
  $b_i, b_j \leftarrow$ BINARY_VECTOR(p_i, p_j, f_i, f_j) ▷ SH
  TRANSMIT(b_i, b_j)
  $\tau \leftarrow$ CROSS_CORRELATION(b_i, b_j) ▷ FC
end function

The sensors (microphones) $\mathbf{m_i}$ and $\mathbf{m_j}$ measure signals, y_i and y_j. The proposed

algorithm for estimating TDOA, for that pair of microphones, is summarised by Algorithm 5.3.1. It consists of the following key steps:

1. **At the Sensor-Head (SH):** Calculate the spectrograms, p_i and p_j at each microphone, from the recorded signals y_i and y_j. The dimension of each spectrogram is K by T, where K is the number of rows corresponding to frequencies and T is the number of columns corresponding to time. We determined, as explained in Section 5.3.5, that the optimum parameters for calculating the spectrogram were window size = 256, overlap = 204 and the final number of sampling points in the discrete Fourier transform = 1024, as will be explained in detail in Section 5.3.5, optimised for this particular scenario;

Algorithm 5.3.2 Row Selection

function SELECT_ROWS(p,X,f)
  $sum \leftarrow \sum_K p$
  $sorted \leftarrow \text{SORT}(p, sum)$
  $n \leftarrow X \times K$
  $rows2use \leftarrow \text{P\_TOP\_ROWS}(sorted, n)$
  $feat2use \leftarrow f \cap rows2use$
  return $feat2use$
end function

2. **At the Sensor-Head (SH):** Calculation of the Scale-Invariant Feature Transform (SIFT) [154] on the normalised spectrogram magnitude, in order to detect N keypoints from each spectrogram. We create a vector of keypoints, $\mathbf{q_i}$ and $\mathbf{t_i}$ for the *i-th* microphone. The n^{th} keypoint has coordinates (q_n, t_n), which corresponds to the time-frequency location at which the keypoints are detected. The values that will be transmitted are integers (encoded in 32 bits in order to keep a high level of precision) and we only transmit the $t-$coordinates. It was found that adding in the frequency information did not improve the TDOA relative error, as will be explained in detail in Section 5.3.3. Therefore, the total number of data bits that need to be transmitted to the fusion centre is $N \times 32$. Additional compression could be used, but it is not considered here, since we opted to use integers to represent time coordinates, encoded in 32-bits ($2^{32} - 1$ values). For short signals, however, the encoding could be done in 8 (255 values) or 16-bits (65, 536 values).

   We experimented with the number of keypoints that need to be transmitted in order to obtain an acceptable margin of error in the Time Difference of Arrival (TDOA). In light of this, we selected keypoints with the highest energy

frequency coefficients, i.e. points that belong to rows of the spectrogram in which the sum of coefficients at key points is large. We selected X-rows each time, where X varies between 0.1 and 1. Algorithm 5.3.2 presents the steps followed in this procedure, while Fig. 5.5 illustrates the keypoints on the spectrogram for various compression ratios;

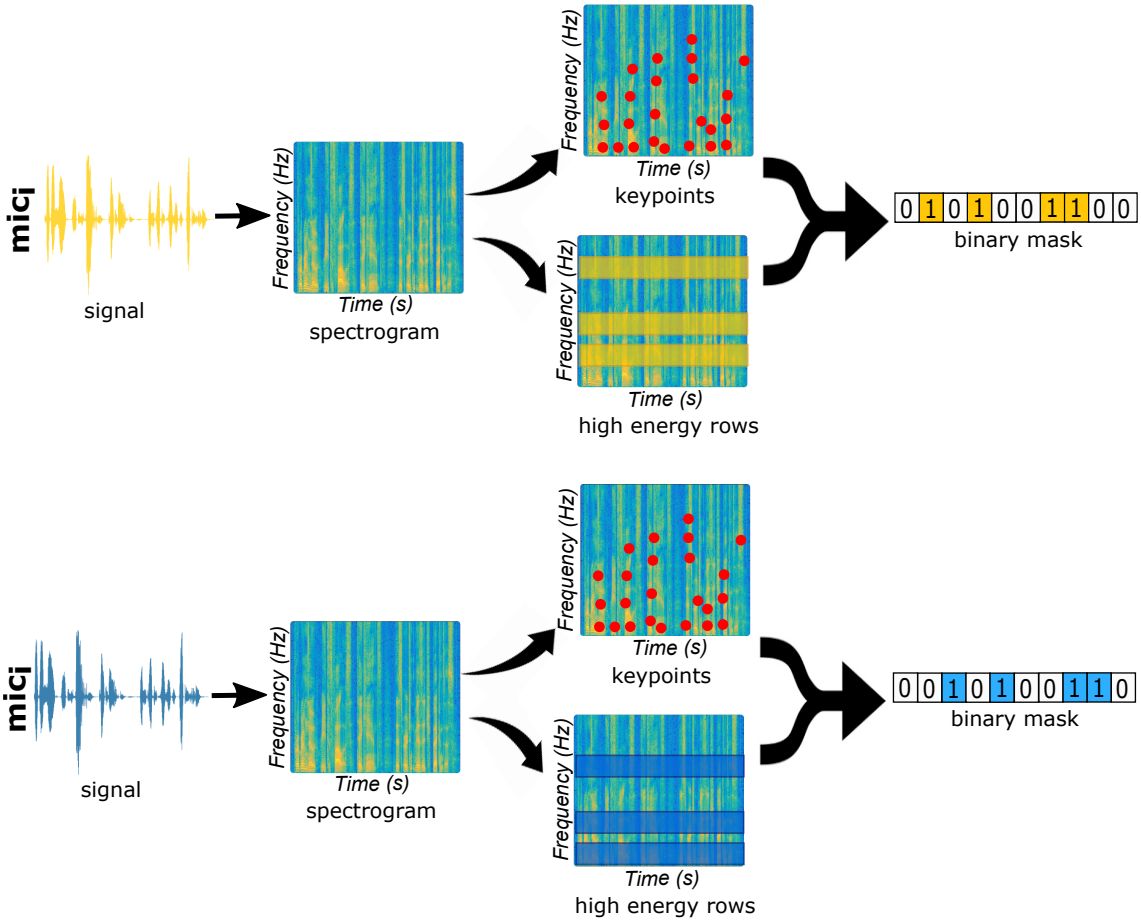


FIGURE 5.4: **Pipeline before data transmission to the Fusion Center (FC).** The signal spectrogram is computed and the SIFT keypoints are extracted. Using the steps illustrated by Algorithm 5.3.2, a binary mask is created using the extracted keypoints (yellow and blue) and the top x high energy rows. An index is filled with 1 if there is a keypoint in that position and that point lies in one of the high energy rows.

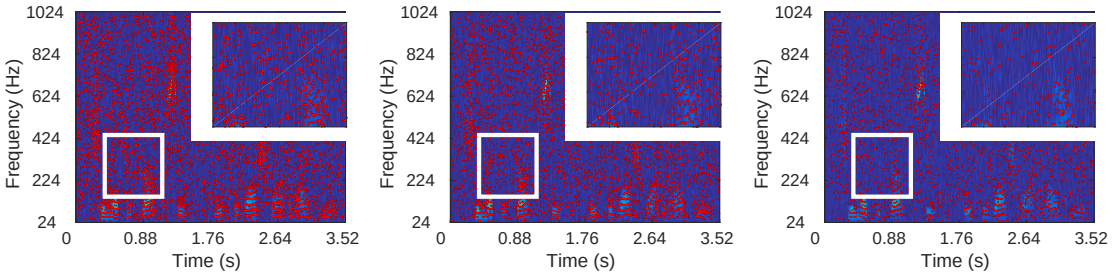
At this point, the processing in the Sensor-Head (SH) is complete. Fig. 5.4 presents a visual illustration of the whole pipeline before transmitting the data to the fusion centre.

3. **At the Fusion Centre (FC):** After the data is transmitted, two new vectors, b_i and b_j, of the same size as y_i and y_j are created at the fusion centre. We are assuming that all the sensors are synchronised and therefore started recording at the same instant. We can map keypoint locations to vectors by pre-calculating the times that correspond to the t-coordinates. The vector is filled with 1's in indices where a SIFT keypoint was detected and with 0's otherwise;

$$b_i(l) = \begin{cases} 1 & \text{if } l \in t_i \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

4. **At the Fusion Centre (FC):** Calculation of Generalized Cross-Correlation (GCC) (defined by the $\star$ operator) between both vectors in the time domain. Since the cross-correlation is now on a binary vector, there is no need for the spectral normalisation as in PHAT.

$$\tau_{\text{delay}} = \arg\max_t((b_i \star b_j)(t)) \tag{5.2}$$



(a) 40 : 1 compression (b) 55 : 1 compression (c) 90:1 compression

FIGURE 5.5: **SIFT keypoints (indicated in red) in the signal spectrogram, for different compression ratios.** For each spectrogram, a patch (white rectangle) is selected and magnified at the upper right corner to provide a clearer visualisation of the SIFT keypoints (red). This illustrates how the selected SIFT features are not necessarily spectrogram peaks and how our features differ from the peak picker approaches.

5.3.3 Unsuccessful Approaches

During the design of the algorithm, there were various approaches pursued in order to improve the obtained results. Since some of them were not successful, this section presents a summary of these approaches.

- One of the early approaches included the use of Mel-frequency cepstral coefficients (MFCC), commonly used for speech recognition, to select keypoints from the sound signals and use a nearest neighbour approach to match the chosen points from each signal. Early experiments showed that it was not possible to calculate TDOA using this technique. Moreover, given the size of the MFCC feature vector, it would not have been suitable for compression.

- Local extrema, that is, local maxima and minima, were used to detect keypoints before using Iterative Closest Point (ICP) [155] to calculate the TDOA. This approach also proved to be insufficient to estimate TDOA.

- The use of the SIFT descriptor was also considered for TDOA calculation: however, similarly to what happened with MFCCs, the size of the descriptor would not have made this approach suitable for compression.

- An initial version of the algorithm used the original values of the signal instead of 1's in the mask in order to perform GCC. This also proved to generate inaccurate TDOA estimations, while the use of a binary mask provided better results.

- An approach consisting of using various binary masks, one for each frequency level, was also explored. However, this resulted in inaccurate TDOA estimations.

5.3.4 Error Metric

Since TDOA is in the order of milliseconds for some source locations and centiseconds for others, it is necessary to standardise the error in order to make a fair comparison among source positions, similarly as explained previously in Chapter 4. Using the Ground Truth (GT), the relative error is computed using the TDOA estimation error in Equation 5.3. Similarly, we use the same principle to estimate the Direction of Arrival (DOA) relative error in Eq. 5.4.

$$\text{tdoa error}(\%) = \frac{\|\text{tdoa} - \text{gt}\|}{\|\text{gt}\|} * 100 \tag{5.3}$$

5.3.5 Spectrogram Parameter Estimation

Finally, we estimated the parameter configuration that produces the most accurate TDOA estimation. We considered the following parameters:

TABLE 5.1: **Spectrogram parameters.** Set of values for the chosen experiments

parameter	values
window	1024, 512, 256
overlap	80%, 90%, 95%
nfft	1024, 512
complex magnitude	1, 0
normalise	1, 0

- **window**: number of samples per window in the spectrogram.
- **overlap**: number of overlapped samples in the spectrogram windows.
- **nfft**: number of sampling points in the discrete Fourier transform.
- **complex magnitude**: use only the spectrogram magnitude.
- **normalise**: convert values of the spectrogram to the scale 0 to 1.

Table 5.1 presents the list of parameters and the values considered for each of them.

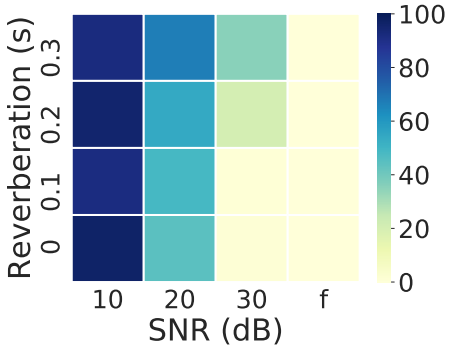
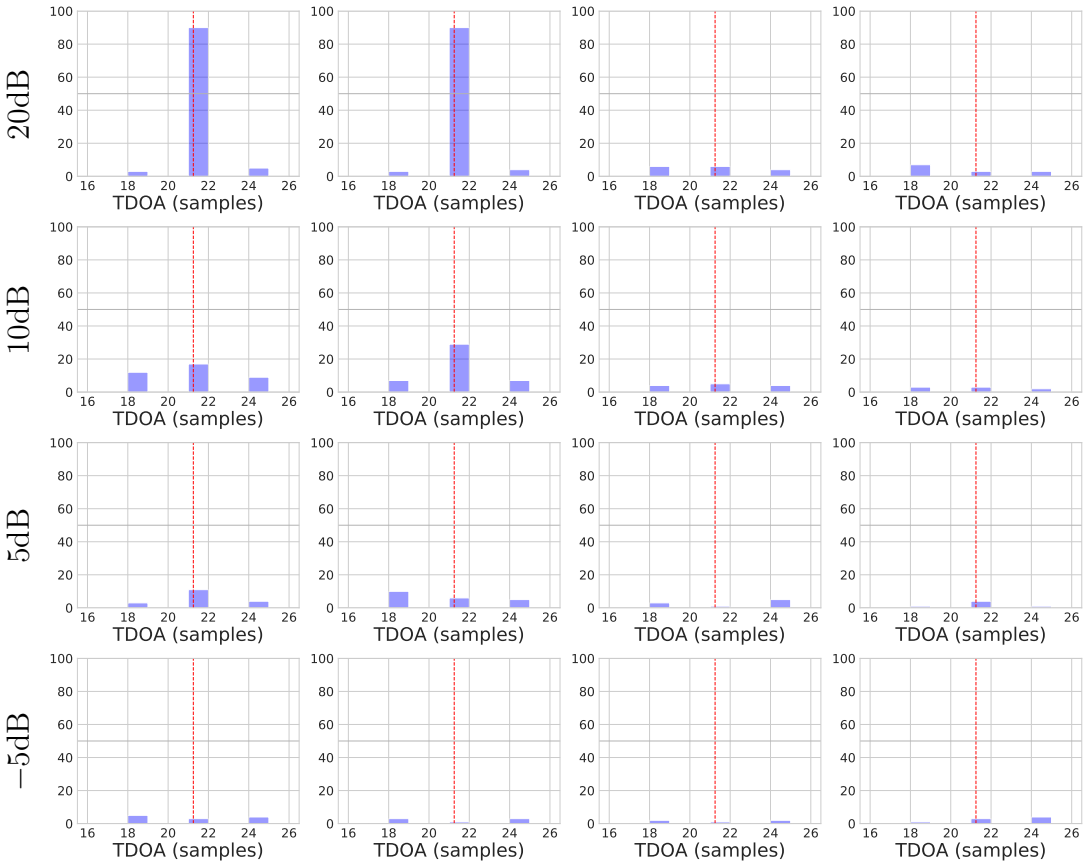


FIGURE 5.6: **Matrix of relative error for the parameters that produce the lowest TDOA estimation error for a signal sampled at 44 kHz.** window = 256, overlap = 204, nfft = 1024, complex magnitude = 1, normalise = 0. The heatmap represents the TDOA relative error from low (yellow) to high (blue).

A sound source located at 30° was simulated under different noise (noise free, 30 dB, 20 dB and 10 dB) and reverberation (0s, 0.1s, 0.2s, 0.3s) conditions. Using the horizontal microphone pair of our circular microphone array, we generated heatmaps of the TDOA relative error for each scenario by using 100 monte carlo simulations. Fig. 5.6 illustrates the heatmap obtained with the parameters that produce the

lowest error when a signal of 44 kHz is used. The resulting parameters, obtained for this particular scenario, were window = 256, overlap = 204 (80% of the window samples), nfft = 1024, complex magnitude = 1 and normalise = 0.

5.4 Experimental Results



(a) $T_{60} = 0$ s (b) $T_{60} = 0.1$ s (c) $T_{60} = 0.3$ s (d) $T_{60} = 0.5$ s

FIGURE 5.7: **Algorithm initial validation.** The histogram (blue) illustrates the estimated Time Difference of Arrival (TDOA) using our algorithm for 100 monte carlo simulations for a fixed microphone pair and source location $(x = 2, y = 1, z = 5)$, with various Signal-to-Noise Ratio (SNR) (rows) $(20dB, 10dB, 5dB, -5dB)$ and reverberations (columns) $(0s, 0.1s, 0.3s, 0.5s)$. The TDOA ground truth estimated in samples is 21.25 (red line).

The experiments detailed in this section were performed using speech signals from the TIMIT database [156] and simulated environments by means of the Image Source Method (ISM) [30]. We simulated two microphones in a linear array, separated by a

distance of 4 metres and sampled at 16kHz. The simulated room has a size of 25m $\times$ 3m $\times$ 12m.

The ISM was previously introduced in Chapter 2. It simulates Room Impulse Response (RIR) in a room, for a given reverberation, expressed as T_{20} or T_{60}, and noise, represented as the desired SNR level of additive Gaussian noise, computed as a time average across all sensors.

5.4.1 Algorithm Validation

We first validated our algorithm by running it on 100 monte carlo simulations for a fixed microphone pair and source location $(x = 2, y = 1, z = 5)$, with various SNR $(20dB, 10dB, 5dB, -5dB)$ and reverberation $(0s, 0.1s, 0.3s, 0.5s)$. The TDOA ground truth estimated in samples is 21.25 and minimum compression $(40 : 1)$ is used, that is, all Scale-Invariant Feature Transform (SIFT) keypoints are selected. Our hypothesis is that the distribution of the estimated TDOA will have a peak around the ground truth. Fig. 5.7 illustrates the obtained results for our experiment. This figure validates our hypothesis for low noise and reverberation levels, in which it can be clearly seen that our algorithm correctly estimates TDOA, since the distribution of the estimation has a peak that coincides with the ground truth (red line). It is also important to note that when the noise and reverberation increases, the peak height decreases (in the case of $10dB$) and it disappears when noise, reverberation or both are high.

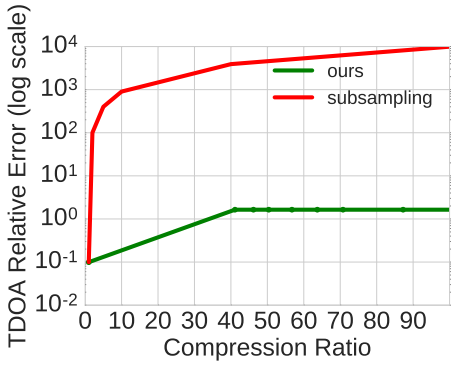


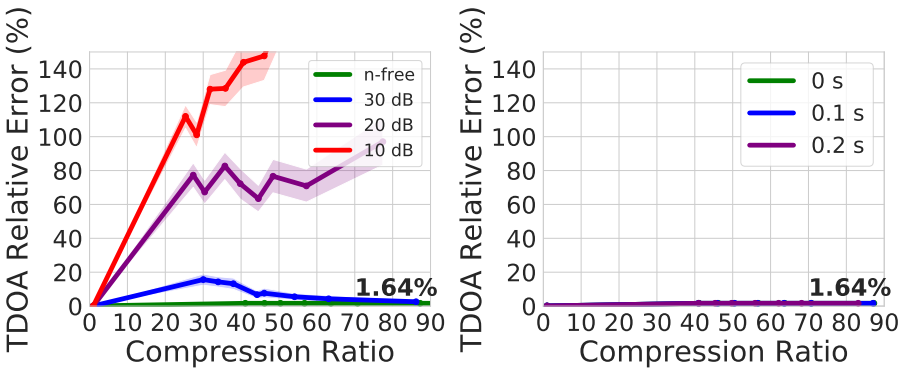
FIGURE 5.8: **Subsampling vs our algorithm in a noise-free environment.** TDOA Relative Error achieved for different compression ratios for a source located at Direction of Arrival (DOA) 45°. The figure shows the TDOA relative error for our algorithm (green) compared with a baseline (red) in which the signal is compressed by subsampling. We used the logarithmic scale on the Y-axis given that the error for the subsampling approach is much higher than our error.

5.4.2 Accuracy vs Compression

As previously mentioned, the compression ratio was varied in order to determine how much compression we could achieve while obtaining a reasonable TDOA relative error. We used the subsampling strategy presented in Sec. 5.3, where we selected keypoints with the highest energy frequency coefficients, i.e. points that belong to rows of the spectrogram in which the sum of coefficients at key points is large. Fig. 5.5 shows the spectrogram SIFT keypoints for different compression ratios.

Our hypothesis is that the error will decrease when the compression is increased and that it will be much lower than using a naive approach that compressed the signal by subsampling it. Fig. 5.8 illustrates the TDOA relative error with respect to compression ratio. In this experiment, the source was located at a DOA of 45°. It shows the error for an environment free of noise and reverberation using the proposed method and compares it with an approach in which compression is achieved by subsampling the signal. Since subsampling the signal increases the error dramatically even for low compression ratios, we decided to use a logarithmic scale on the Y-axis. Fig. 5.9 shows the relative error for various noise (left) and reverberation (right) conditions. While our hypothesis about outperforming subsampling is validated by Fig. 5.8, our hypothesis about the accuracy decreasing appeared to be true only for changes in the noise conditions, since in the absence of noise and restricted to reverberation only (Fig. 5.9) the accuracy remains the same for high compression ratios. This may be caused because the spectrogram is not drastically changed when there is a change in reverberation, as opposed to a change in noise, therefore the SIFT features extracted are different in the latter case.

Fig. 5.10 shows how noise and reverberation separately affect the maximum compression ratio. From previous experiments, our hypothesis is that our algorithm is more sensitive to noise than to reverberation. We calculated the maximum value of compression that produced a TDOA relative error smaller than $5\%, 10\%, 50\%$ for the given noise and reverberation conditions. In this scenario, the source is located at DOA 45°. In Fig. 5.10(a), a white Gaussian noise of -10 dB, 0 dB, 10 dB, 20 dB, 30 dB and 40 dB signal-to-noise ratio per sample was added to the original signal. Note how the compression improves as the signal-to-noise ratio (SNR) gets higher. In the case, of 5% and 10%, the compression ratios are identical, therefore we can only visualise one line. We used T_{60} as a measurement of reverberation, interpreted as the time it takes a signal to drop by 60dB. In Fig. 5.10(b), reverberation values of $T_{60} = \{0.1k,\, k \in \{1, \ldots, 10\}\}$ seconds are simulated. In this case we can see that there is no compression value for which the error is smaller than 5%, however for



(a) noise (b) reverberation

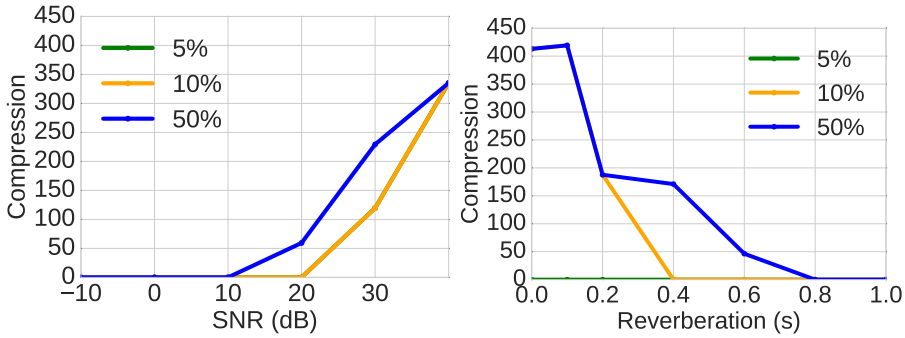
FIGURE 5.9: **Accuracy vs compression for various noise and reverberation conditions.** The left-hand side of the figure shows the TDOA Relative Error for a noise-free signal and for signals with various SNR values: noise-free (green), $30dB$ (blue), $20dB$ (purple) and $10dB$ (red) for a source located at DOA 0° (challenging DOA estimation). The right-hand side, in contrast, shows the relative error for various reverberation levels: $0s$ (green), $0.1s$ (blue) and $0.2s$ (purple). To estimate the relative error for each compression ratio, we used 100 simulations. The challenging location of the source in this scenario means that the error does not get below 5% in (b). This differs from the result presented in Fig. 5.8 (in which the error remains at 1.64%), given that, in that case, the source is located in a less challenging location (end-fire).

10% and 50% we achieved high compression ratios for low reverberation values (up to 0.6), after which the compression decreases to zero. These experiments confirm our hypothesis that our algorithm is more robust to noise than it is to reverberation.

5.4.3 Accuracy vs Source Location

After determining the noise and reverberation conditions under which our algorithm performs accurately, we decided to evaluate it in terms of the location of the sound source. Our hypothesis is that for some DOA the algorithm will perform better since the number of TDOA samples to estimate is higher. Therefore, we started by evaluating the TDOA estimation and then proceeded to calculate how it affects the DOA.

Fig. 5.11 and 5.12 illustrate the TDOA relative error and the DOA relative error for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds and 3 noise conditions: noise-free, $30dB$ and $20dB$ SNR. We randomly selected 10 different sounds from the TIMIT dataset, which included speech signals from 5 men and 5



(a) noise (b) reverberation s

FIGURE 5.10: **Maximum compression for various noise and reverberation levels.** Maximum compression when the TDOA relative error is $\leq 5\%$ (green), 10% (yellow), 50% (blue) for a source located at DOA 45° for different values of noise and reverberation. In (a), white Gaussian noise of -10 dB, 0 dB, 10 dB, 20 dB, 30 dB and 40 dB signal-to-noise ratio per sample was added to the original signal. For 5% and 10%, the compression ratios are identical, therefore we can only visualise a single line. In (b), we simulated reverberation values of $T_{60} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ seconds.

women (labeled A to J in Fig. 5.12). We simulated 19 different source locations (DOA), from 0° to 180°, with a step size of 5°. We ran 5 different simulations for each of these sources and reverberation values.

Fig. 5.11 shows the TDOA relative error for each DOA. The compression ratio is 40 : 1 for each signal. It can be seen from the plots that for environments with low reverberation, $T_{60} = 0.1, 0.2$ seconds, the TDOA relative error is smaller than 20% for most DOA, except for 80°and 100°, in which case the error rises above 40%. The reason for this behaviour is the small magnitude of TDOA values at such locations, which makes its calculation very challenging. This will be analysed in further detail in Section 5.4.4. Similar results are obtained for $30dB$ SNR, where most relative errors are below 40% for low reverberations. In the case of $20dB$ SNR however, the relative error increases to 60% for the most accurate source locations. Appendix B includes a version of this plot with unlimited y-axis, illustrating the error TDOAs of small magnitude.

Fig. 5.12 shows the DOA localisation error. The x-axis presents 10 different datasets (labeled A to J). Three different compression ratios are used: 40 : 1, 45:1 and 50:1. For noise-free and low reverberation, $T_{60} = 0.1, 0.2$ seconds, the DOA relative error remains less than 20% for different compression ratios and sources. When reverberation $T_{60} = 0.3$ seconds, the TDOA relative error increases dramatically for

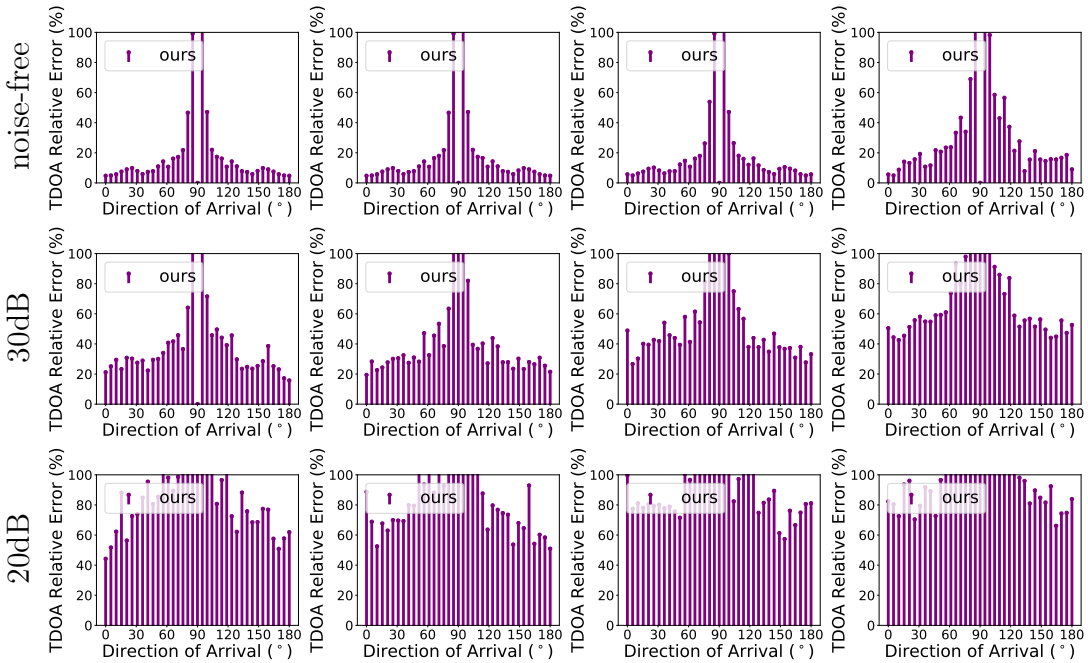


FIGURE 5.11: **TDOA relative error vs DOA.** TDOA relative error (purple) for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds and 3 noise conditions: noise-free, $30dB$ and $20dB$ SNR. The results are from 10 speech signals, at 19 different locations (DOA), from 0° to 180°, with a step size of 5°. We ran 5 different simulations for each of these sources and reverberation values. The compression ratio is 40 : 1 for each signal. A version of these plots without the 100 limit is presented in Fig. B.1. The high errors for sources located in front of the microphone array is because they are below the resolution I am able to calculate, as explained in Section 5.4.4.

most DOA, especially for 80°and 100°, in which case it is close to 80%. This large TDOA error has little impact on the DOA estimation, however. Even though the DOA relative error is above 20% in this case, the error in general remains less than 40%. When the SNR is 30 or $20dB$, the DOA average error increases for all the datasets. It is important to keep in mind that the error is averaged amongst 19 DOA and, as seen previously in Fig. 5.11, the error increases dramatically for a source located at 90°, affecting the average error per dataset.

$$\text{doa error}(\%) = \frac{\|\text{doa} - \text{gt}\|}{\|\text{gt}\|} * 100 \tag{5.4}$$

These experiments validated our hypothesis that for some sound source locations, especially those in front of the microphone array, the TDOA error is higher, given

that the ground truth corresponds to very small values. We could also validate that the algorithm performance remains similar for various speech signals.

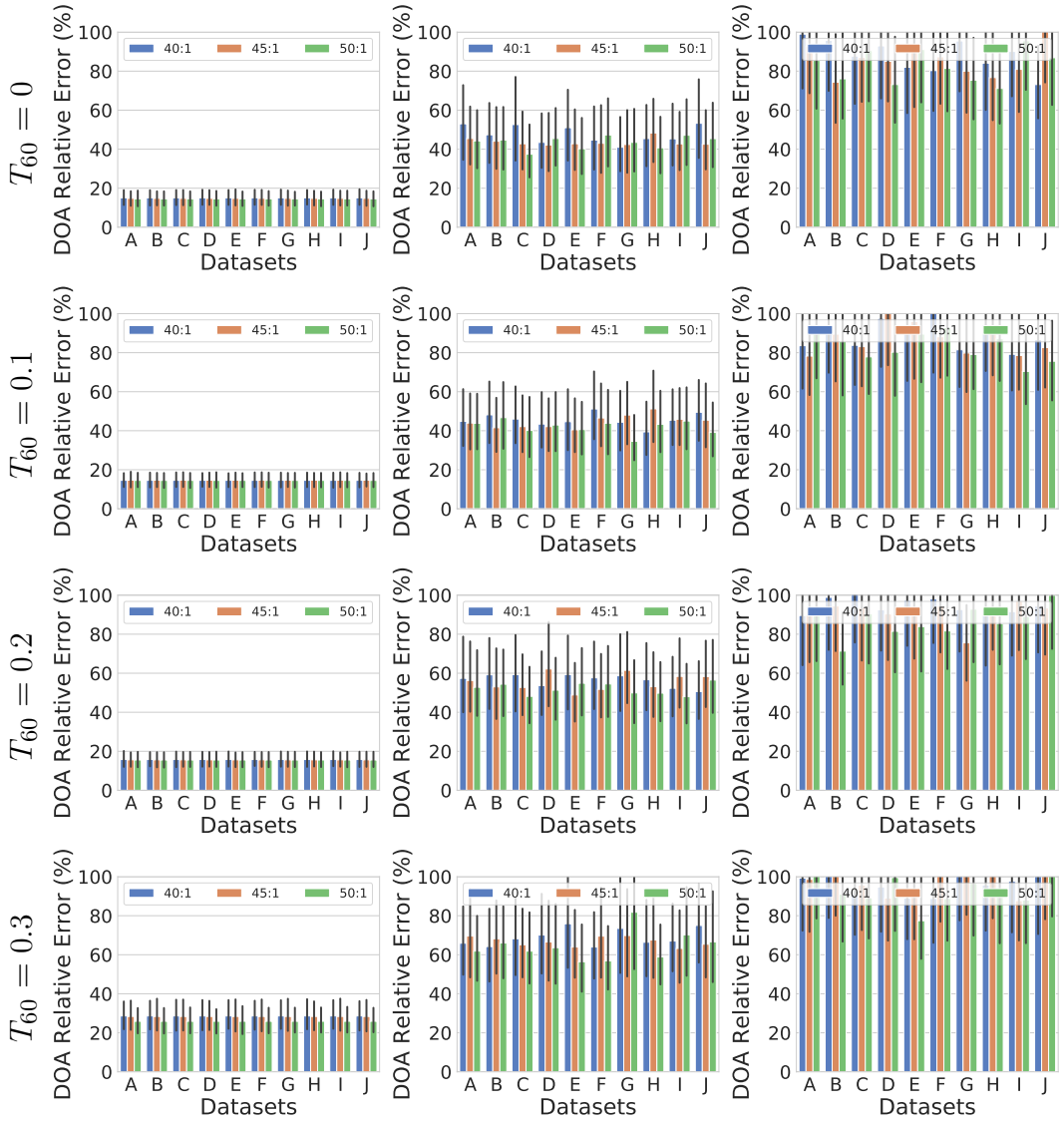
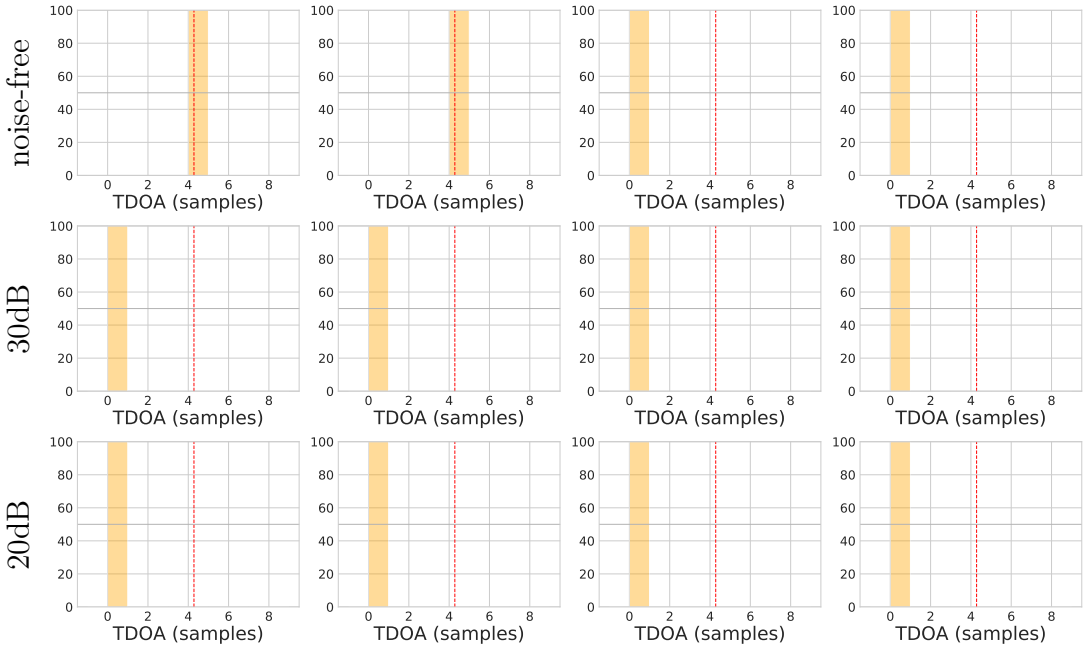


FIGURE 5.12: **DOA localisation error per dataset for three different compression ratios:** $40:1$ **(blue),** $45:1$ **(orange) and** $50:1$ **(green).** TDOA relative error and the DOA relative error for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds and 3 noise conditions: noise-free, $30dB$ and $20dB$ SNR. The results are from 10 speech signals (labelled A to J), at 19 different locations (DOA), from 0°to 180°, with a step size of 10°. We ran 5 different simulations for each of these sources and reverberation values.



(a) $T_{60} = 0$ s (b) $T_{60} = 0.1$ s (c) $T_{60} = 0.3$ s (d) $T_{60} = 0.5$ s

FIGURE 5.13: **Histogram of estimated TDOA for small microphone separation.** Simulated source located at $(x = 2, y = 1, z = 5)$ in various noise and reverberation conditions. The ground truth TDOA in samples is 4 (red line). The histogram (yellow) illustrates the estimated TDOA using our algorithm for 100 monte carlo simulations. The algorithm fails for some noise and reverberation conditions by inaccurately estimating 0 as the TDOA.

5.4.4 TDOA of Small Magnitude Estimation

We decided to explore the behaviour studied in the previous section, relating to the estimation of TDOA values of small magnitude. Our hypothesis is that, when the TDOA magnitude is very small (in the order of 1×10^{-4}), our algorithm will estimate the TDOA as zero. Fig. 5.13 confirms our hypothesis. In this case, the TDOA of a source located at $(x = 2, y = 1, z = 5)$ was estimated using our approach under various noise (free, $30dB$ and $20dB$) and reverberation ($T_{60} = 0s, 0.1s, 0.3s, 0.5s$) conditions. The ground truth TDOA corresponds to 4 samples. It can be seen that even when the noise is low ($30dB$) most of the samples are estimated as zero.

5.4.5 Baseline: Fingerprinting

Last but not least, we compared our approach against fingerprinting, an approach that was previously mentioned in Section 5.2 and has been used in the past for calibration, locating sources in end-fire locations. We simulated a sound source located

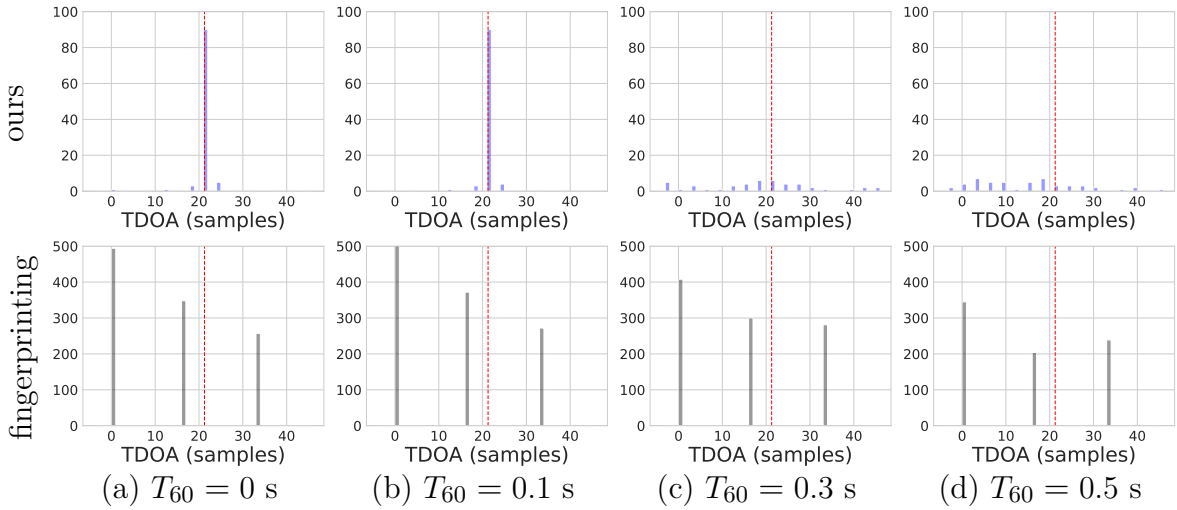


FIGURE 5.14: **Our algorithm vs fingerprinting**. Histogram of estimated TDOA values for a source located at $(x = 2, y = 1, z = 5)$ and $20dB$ SNR. The ground truth of TDOA is 21.25. While our approach (blue) presents a clear peak in the TDOA distribution, the fingerprinting approach (gray) presents the highest peak at zero.

at 45° DOA. Our hypothesis is that for non-end-fire locations our algorithm will outperform fingerprinting. Fig 5.14 validates this hypothesis, by showing that, while our algorithm presents clear peaks in the estimated TDOA samples, fingerprinting does not have a clear peak and the highest one corresponds to zero.

5.5 Discussion

5.5.1 Keypoints for Compression

We found that, by applying the computer vision technique of Scale-Invariant Feature Transform (SIFT) to the spectrogram of a speech signal, it is possible to detect keypoints that contain relevant information about the signal. We were able to use these keypoints to select the signal samples used to estimate Time Difference of Arrival (TDOA) within a reasonable margin of relative error.

Our mechanism for improving the compression rate is to use subsampling of the SIFT keypoints in the spectrogram constructed at each sensor (microphone). Our strategy was to select the highest energy frequency coefficients, i.e rows of the spectogram in which the sum of coefficients at key points is large. This proved to be effective in scenarios in which there is little noise, as illustrated by Fig. 5.9b and Fig. 5.10a.

We ran our algorithm for various source locations and speech signals. We determined that the highest error in estimating the TDOA was caused in positions where the source was located in front of the microphone array, either at 80°or 100°. This happens because the TDOA is very small for these positions, which complicates the estimation. For 90°, where the TDOA is zero, and for 0°and 180°, where the separation is maximum, the relative error is closer to zero. On the other hand, given a similar position and the same noise and reverberation conditions, our algorithm performs very similarly across the test speech signals we used.

Considering that the experiments were conducted in simulated scenarios, but using real speech data in a large variety of acoustic conditions, it would be possible to extend this approach to real-life scenarios. This should be in the context of low noise and reverberation environments, in which the microphone separation is considerably large in order to obtain TDOA of large magnitude.

5.5.2 SIFT vs Baseline

We used an audio fingerprinting approach to estimate TDOA [67] as a baseline for comparison. The authors used this method for calibration, therefore, their approach is limited to sources located at end-fire positions and controlled noise and reverberation environments. We used the implementation of audio fingerprinting presented in [30], in which the input signal is subsampled to 8kHz to calculate the spectrogram. The number of sections is 64ms and the overlap is 32ms. We selected 50 landmarks per signal to perform our comparison. We simulated a source located at DOA 45°. We demonstrated how our algorithm shows a clear peak in the sample distribution while fingerprinting's highest value was located at zero.

5.5.3 Limitations and Future Work

The algorithm's main drawback is its sensitivity to noise, as is evidenced in Fig. 5.10. This may be attributable to SIFT keypoints chosen from noise rather than from the original signal, leading to choosing different points in each sensor. One strategy to overcome this problem might be to estimate the probability of the keypoints being noise based on the amplitude of neighbouring keypoints.

Another limitation studied is the low accuracy when the magnitude of the TDOA value to be estimated is very small. In this case, we are limited to the sampling frequency, which makes it impossible to calculate some TDOA values. In the cases in which it is possible, but our algorithm is assuming it to be zero, one strategy could

be to increase the overlap in the spectrogram calculation together with a denoising algorithm, which might improve SIFT keypoint detection.

5.6 Conclusions

In this chapter, we presented novel findings regarding Time Difference of Arrival (TDOA) estimation using only a few signal samples.

We started by testing our algorithm in a microphone pair using 100 Monte Carlo simulations, and we realised that for low noise and reverberation conditions, the mode of the estimated TDOA corresponded to the ground truth. We then proceeded to compare our algorithm against a compression approach using subsampling and we demonstrated how our algorithm estimated the TDOA accurately, even when the compression ratio increased a lot. Therefore, we continued our tests and showed how our algorithm estimated the TDOA for high compression ratios in various noise and reverberation conditions.

We continued our test using various signals from the TIMIT dataset located at different Direction of Arrival (DOA). We showed how our algorithm accurately estimated the TDOA and DOA with high compression ratios, in scenarios in which the noise and reverberation were low.

Finally, we compared our algorithm against an approach that uses audio fingerprinting for TDOA estimation and we showed how our algorithm outperformed the baseline.

While the scenarios in which the algorithm was tested were generated by means of a simulator, the data used for testing was real speech from a well-known dataset. Therefore, we believe that our system capabilities could easily be extended to real life scenarios.

In the current version of this work, we are using only a single microphone pair: however, the algorithm could also be extended so that it works for multiple pairs of microphones. This could potentially benefit the performance of the algorithm, since the redundancy in the information could be used to remove outliers from the selected keypoints, making the algorithm more robust to noise and reverberation. This is left as future work.

Our contribution, then, could be summarised as follows:

- We proposed an algorithm that determined the signal keypoints to be transmitted in order to accurately estimate TDOA obtaining a signal compression ratio of $40:1$

- We compared our technique against a baseline that uses audio fingerprinting and showed that our approach presents superior results.

In conclusion, in this work, we showed that, by applying a computer vision approach to the spectrogram of a speech signal, it was possible to identify samples of the signal allowing for an estimation of TDOA within a reasonable margin of relative error. We tested the robustness of the proposed technique under different noise and reverberation conditions using different speech signals and source locations. We showed that our algorithm can estimate TDOA and the source location within an acceptable error range when the compression ratio of the signal is $40:1$.

In the future, we plan to modify our algorithm by improving on its robustness to noise and reverberation. We intend to do this by estimating the probability of keypoints representing reverberation or noise.

Chapter 6

Training Data on CNNs for DOA Estimation

6.1 Introduction

Estimation of the spatial direction from which a sound is emitted, commonly known as Direction of Arrival (DOA), is an important and well-studied problem in Acoustic Source Localisation (ASL) with applications in numerous domains [68,70]. The advent of smart assistants (e.g. Amazon Echo, Google Home, Apple HomePod) [7], equipped with arrays of microphones, has facilitated the generation of large datasets and has motivated research into the use of data-driven methods for DOA. In particular, learning via a Deep Neural Networks (DNNs) architecture – deployed effectively for computer vision applications [157] and audio processing [158] – is emerging as an effective tool for ASL estimation [73].

Traditional methods to perform ASL have been widely studied in the literature [131], the most common of which are: (i) Time Difference of Arrival (TDOA)-based approaches, which normally employ Generalized Cross-Correlation (GCC) [159, 160], (ii) beamforming-based approaches, including the well-known Steered Response Power (SRP) [58], which solve directly for the most likely source position among a grid of candidate locations; and (iii) MUltiple SIgnal Classification (MUSIC), which uses the signals subspaces to estimate multipe DOA. A summary of the literature review in TDOA estimation is found in Chapter 3. Neural networks have been applied for various problems related to ASL including speaker localisation using a robot [68, 69], passive underwater sensing [70], antennas [71] and acoustic emission localisation on a pipeline [72]. Chakrabarty et al [4] perform single source localisation by treating ASL as a classification problem, where the discretised DOA corresponds to a class, which they solve using a Convolutional Neural Network (CNN). This method has been extended to multiple sources [73] using synthetic noise data to train

the network. CNNs combined with Long short-term memory (LSTM) [161] have been shown to be useful for estimating DOA by using Generalized Cross-Correlation Phase Transform (GCC-PHAT) as input data. Some approaches use neural networks to perform pre-processing such as time-frequency (TF) masking [162–164] or denoising and dereverberation [165].

For multiple source localisation, the use of planar arrays include an extension of [4] for multiple sources [39, 166] using synthetic data and an improvement to their architecture in [73] by incorporating systematic dilation of the convolution filter in each layer of the CNN. There are also some papers that use ambisonics [167], which is a representation of the sound field as a decomposition into spherical harmonics. Among these, [168] employs Layerwise Relevance Propagation (LRP) extended in [169] to testing in unseen rooms. In [170] the authors use Recurrent Neural Networks (RNN) tested in different sound classes and extend their approach by performing joint localisation and detection [171]. Finally, [172] localises and detects sound events using quaternion-valued data processing. Section 6.2 summarises the literature review for single and multiple DOA estimation using Neural Networks (NN).

Despite the widespread use of CNNs in applications related to ASL, numerous questions regarding the quality and quantity of the training data remain unanswered. In [170, 171], data from different sound classes is randomly used for training and testing, while in [173] the authors propose a method of data augmentation for the task of room classification from reverberant speech using a Generative Adversarial Network (GAN). In [174] deep CNN and data augmentation are used for environmental sound classification. On the other hand, Pons et al [175] use few training samples (from 1 to 100) per class to train an event and acoustic scene classifier. In this chapter, we test the impact of various sound classes for training on the accuracy of DOA estimation. Our hypothesis is that using speech and music data for training will provide more accurate DOA estimation than using noise, as in the current literature. Our reasoning is that speech and music data contains frequency information that helps the CNN learn the room acoustics much better than white noise. Our conclusion is that using real speech data augmented with synthetic speech data (using GAN-based methods) performs best for a wide range of test audio classes and different incident directions.

Our main findings in this work are that:

- training with speech data, rather than noise, produces an average improvement of 3% on the accuracy of DOA estimates for test speech signals and 17% when the test signals belong to one of three other classes;

- training with music data from a dataset produces an average improvement of 19% in accuracy compared to training with noise;
- synthetic speech data generated using a state-of-the-art GAN [14], which can be generated automatically, is as effective in training as using real human speech;
- music data performs better than speech data for training when obtained using real sound recordings: however, when they are synthetically generated using a GAN, speech data produces better results than music data;
- compared with GCC, a DNN trained with speech is 125% more accurate when the test and training environments have similar reverberation, and comparable when the reverberation levels are different.

6.2 Related Work

The literature is divided between methods that calculate Direction of Arrival (DOA) and those that estimate the 3D source location.

6.2.1 Direction of Arrival (DOA)

DOA methods are subdivided depending on whether they estimate the DOA for a single source, or multiple sources.

Single Source

The use of planar arrays is very common in single-source DOA estimation. In [68], for instance, the authors train a Deep Neural Network (DNN) to localise sources using a microphone array embedded on a NAO robot. Localisation is presented as a binary classification problem, in which the algorithm returns either 1 or 0, depending on the existence (or not) of a source at a given direction. The main contributions arising out of this work are the use of a directional activator, similar to MUltiple SIgnal Classification (MUSIC), and the use of this activator to treat complex numbers (from the spectrogram) at each sub-band. The evaluation was performed using real data from a Japanese dataset as training and testing sets (with different data used for each set), and accuracy computed for 72 different DOA and blocks of 200ms. The main limitation of this work is that the DNN is unable to localise sources located in positions that not appear on the trainning set. The authors propose a new approach

to overcome these limitations in [69], using unsupervised learning together with a parameter adaption layer and early cessation of the parameter updates. These changes result in improvements for some of the DOA angles, but with a deterioration for others. A similar approach is presented in [4], in which the authors use the phase information of the Short-Time Fourier Transform (STFT) coefficients together with a single-class classifier to train a Convolutional Neural Network (CNN) that outputs the DOA of a group of signals from a microphone array. The DOA is modelled as a single-class classification problem, in which the classes are 37 different angles (DOA). The network is trained with synthetic data and tested with speech signals from the TIMIT dataset. The results are presented as accuracy level per frames, that is to say, the number of frames that correctly classify the DOA, similarly to [68]. Since this article is the base for our work, Section 6.3.1 will go into this in further detail. Lastly, in [161] the authors use a CNN combined with a Long short-term memory (LSTM) to estimate DOA. The main contribution of this method is its adaptability to a new microphone array and the use of a very small amount of data, since the network uses Generalized Cross-Correlation Phase Transform (GCC-PHAT) as input, rather than the spectrogram as in previous cases.

There are a set of approaches that use Neural Networks (NN) as a pre-processing step, including [163], in which the authors use a Bidirectional Long Short Term Memory (BLSTM) for time-frequency (TF) masking to arrive at a clean phase Time Difference of Arrival (TDOA) estimation. They use this to improve conventional Cross-Correlation (CC), beamforming and subspace-based algorithms for Acoustic Source Localisation (ASL). They perform experiments with a binaural setup, judging the estimation as accurate when the error is within 5 degrees. This approach is extended in [164] where DOA is calculated directly using monaural spectral information for mask estimation during training, and therefore this approach could be extended to different microphone configurations. Similarly to [162], the authors use a CNN to predict a time-frequency (TF) mask for emphasising the direct path speech signal in time-varying interference. This approach is applied in combination with Steered Response Power (SRP) to estimate the DOA. The main limitation is that it only works on the same type of training data while the main assumption is that there is only one main interference with the target of interest. The experiments were conducted using speech (English for training and Japanese for testing) mixed with everyday sounds (office printer background or household noise) to train and test the NN for both static and moving speech sources. Finally, Wang et al [165] propose the use of an Acoustic Vector Sensor (AVS) to estimate DOA, in conjunction with a network for denoising and dereverberation. The authors' hypothesis is that clean

features are better classified than unclean ones, therefore they used a DNN for Signal Denoising and Reverberation (DNN-SDD), which maps noise and reverberant speech features to their clean versions and uses them as input for a DNN that calculates DOA. The method is evaluated in small-sized microphone arrays, with the Mean Absolute Error (MAE) as Root Mean Square Error (RMSE) used as evaluation metrics.

Finally, there are some works that describe ASL using NN in planar arrays for very specific applications. In [70], the authors present an application of CNN for DOA to passive underwater sensors, a technique that uses cepstrograms and generalized cross-correlogram as input to estimate range and bearing. The network is trained using real, multi-channel acoustic recordings of a surface in a shallow water environment. Another application is presented in [71], in which DOA estimation using DNN is used in antennas. The main contributions of the work are a proposed end to end DNN for general (not only acoustic) DOA estimation, the use of an autoencoder for pre-processing and training with various outputs of a certain array, so the network is robust to imperfections. The authors train and test their approach based on simulated data and use MUSIC as a baseline for comparison. Finally, in [72] we are presented with an application of acoustic emission localization on a pipeline, generated when energy is released within a material. The experiments showed an accuracy of 97% and execution time of 0.963 milliseconds.

Multiple Sources

There are various approaches focused on estimating DOA when there are multiple sources present. In [39, 166] for instance, an extension of their work in [4], the authors use a CNN for multi-class multi-label classification, with the last layer using Sigmoid activation. The main assumption is W-disjoint orthogonality, which means that two speakers cannot be active at a given time-frequency point. One of the main novelties in comparison with previous work is the generation of synthetitic data for training by creating separate sources and then concatenating and randomising their spectrograms. The method was tested with simulated data from the TIMIT and LIBRI datasets. The experiments considered the generalisation to unseen acoustic conditions and unseen noise type, as well as the influence of source-array distance and number of convolutional layers. Steered Response Power Phase Transform (SRP-PHAT) and MUSIC were used as a baseline. The authors extended this approach in [73] by incorporating systematic dilation of the convolution filter in each layer of the CNN, which expands the receptive field of the filter and reduces computational cost while

keeping the memory the same. The results, however, demonstrate that, in so doing, the accuracy decreases and the best they are able to achieve is the same accuracy as the original CNN, but reducing the computational cost by 40%.

Alternatively, [176] proposes a joint localisation and classification of acoustic sources. The input is the raw spectrogram of the acoustic signals and the output is the likelihood of DOA, as well as a Speech/No-Speech Classification per frame. The metrics are calculated in terms of precision vs recall curves. The network is tested with real data and the results are compared against SRP-PHAT, highlighting the advantages of using a joint approach.

In [177], researchers investigate two domain adaptation methods using NN for multiple sound source localisation: weak supervision and domain adversarial training. The authors used a pre-trained network, adapted with both source and target domain, using recordings from the Pepper robot and loudspeaker data for adaption. The evaluation used both loudspeaker and human speech and presented precision vs recall curves. The experiments showed an improvement in models adapted with weak supervision: however, the combination of domain-adversarial training does not further improve the performance according to the results presented.

Finally, there is a set of methods that uses ambisonics, a sound format that covers sources above and below the listener, commonly obtained using spherical microphone arrays. In [168], researchers use a Convolutional Recurrent Neural Network (CRNN) as a DOA estimation system for multi-source localisation. The authors use layerwise relevance propagation (LRP) as input, features derived from the acoustic intensity vector. The training was done using simulated signals and the test using recordings from Eigenmike. The same authors introduce a similar approach in [169], in which they train the NN on a large variety of simulated rooms and test it on unseen rooms. They evaluate their algorithm on DOAs that lie anywhere on the sphere and not only on the same discrete grid used for training. [170] uses Recurrent Neural Network (RNN) to estimate both azimuth and elevation by sampling the unit sphere uniformly and predicting the probability of sound source at each direction. The main advantage of this approach is that it does not algorithmically limit the number of directions to be estimated. The authors use synthetic data for training and testing, while they evaluate their method using real spherical harmonic input signals in different sound classes. Their approach is extended in [171] by jointly performing sound event localization and detection. The RNN input is the magnitude and phase of the spectrogram, while the output is DoA and Event Detection. This method is outperformed in [172], by using quaternion-valued data processing.

6.2.2 3D Localisation

While not as wide-ranging as that looking at DOA, there has been some previous work that investigates 3D localisation of acoustic sources using NN. In [178] the authors propose an approach for single sound source azimuth and distance estimation using a binaural setup. The network learns azimuth and distance using CC of the two channels as input. The training and test data are obtained using a loudspeaker to play audios from the TIMIT dataset. Similarly, in [179] the authors propose an end-to-end method to perform 3D localisation of a sound source, using a CNN. The input of the CNN are the raw signals and the output is the 3D position. The authors fine-tune the network using a small amount of real data, dataset AV 16.3, the same one as used for testing.

6.2.3 Applications

It is important to mention some applications that, even though not related to ASL directly, are very related to this work. Inspired by few-shot learning to learn from few training data, in [175] the authors use few training samples (from 1 to 100) per class to train an audio classifier for event and acoustic scene classification. In their experiments they consider regularisation, prototypical networks, transfer learning and a combination of them for the classification task. They conclude that transfer learning is a powerful tool, but that prototypical networks show promising results in the absence of external or validation data. Another interesting application is presented in [173], in which the authors proposed a method for data augmentation for the task of room classification from reverberant speech. Generative Adversarial Networks (GANs) are trained to generate artificial data as if they were measured in real rooms. The representation is based on a sparse model for the early reflections, a stochastic model for the reverberant tail and a mixing mechanism between the two. In the experiments shown, the proposed data augmentation method increases the test accuracy of a CNN-RNN room classifier from 89.4% to 95.5%.

6.2.4 Summary

In general, we could summarise that the literature in deep neural networks as applied to ASL is focused on creating neural network architectures and methodologies that generalise the following:

- **Room Acoustic Conditions:** The network goal is to be robust to new acoustic conditions, such as noise and reverberation, different from those used

during training. One of the clearest examples is [4], in which the network is trained and tested with different room sizes and reverberations. Moreover, [164] test their pre-processing T-F mask in various noise and reverberant environments. Perotin et al, [169], train their NN on a large variety of simulated rooms and test it on unseen rooms.

- **Source Locations:** The objective is to be able to estimate source locations different from those present in the training set. In [4] the authors considered in their experiments the influence of source-array distance. Similarly, [169] evaluated their algorithm on DOAs that lie anywhere on the sphere rather than on the same discrete grid used for training.

- **Microphone Configuration:** The NN should be able to be tested on any microphone configuration, independent of the one(s) it was trained with. This is partially achieved in [161], in which the authors use GCC-PHAT as input to the NN, therefore the microphone configurations of training and testing could be different, provided that the microphones are located at the same distance. A better generalisation is presented in [164], in which the NN uses monoaural information: however, this is only for T-F mask estimation as a pre-processing step, rather than DOA estimation directly.

Even though the literature covers a lot of work in generalising the learning process, there is a gap in the efforts to generalise the **nature of training data**. The closest effort has been presented in [170], in which the authors use various data classes for training and testing the network: however, they limit their work to using the same audio class for training and testing. Accordingly, we have focused this work on studying the impact of the quality and quantity of training data when it comes to DOA estimation.

6.3 Methodology

6.3.1 Baseline: Direction of Arrival (DOA) estimation using Convolutional Neural Networks (CNNs)

The focus of this work is on analyzing the impact of training data, therefore we use an existing architecture [4] and follow the methodology presented in this section for training and testing.

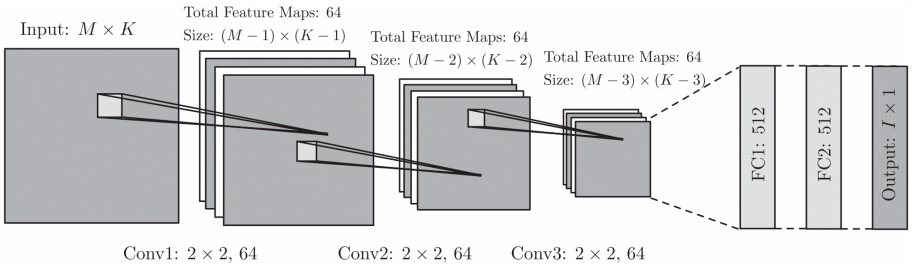


FIGURE 6.1: **Convolutional Neural Network (CNN).** CNN architecture used in [4]. The input of the diagram is a matrix M by K estimated per frame, where M is the number of microphones and K is the number of frequencies on the Short-Time Fourier Transform (STFT). The first four layers are convolutional layers while the last two are fully connected layers (FC). The output is a vector of size I, with zeros in all entries and one in the frame class. I represents the number of DOA classes. The total number of parameters is $426,946$.

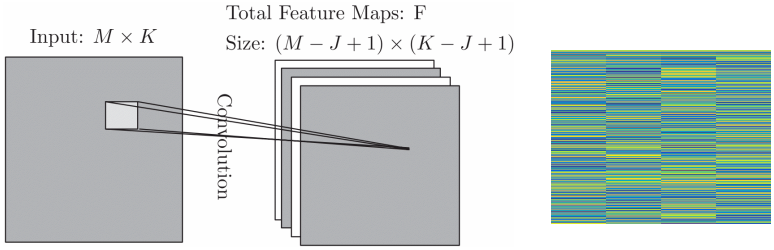
The CNN, initially proposed in [4] and used in [39,73,166] , is based on a standard CNN [180] architecture. These networks typically consist of a set of "convolution layer", which act as filters on the input, resulting in the set of features that the network learns. The convolution is followed by an activation layer, operating point-wise over each element of the feature map. Later on, a pooling operation is applied to reduce the feature map. In the final step, the fully connected layers aggregate information from all different positions to perform classification.

In this particular application, the authors use the CNN architecture presented in Fig. 6.1, which has the following characteristics:

- The CNN treats the phase of the STFT as an image and the input is a matrix of size M by K, as illustrated by Fig. 6.2 (a), where M is the number of microphones and K the resolution of the STFT in the frequency domain. Fig. 6.2 (b) shows the input feature. It is important to note that the input is a time frame of the total signal.

- The number of parameters for each convolutional layer is given by `((shape of width of filter*shape of height filter+1)*number of filters)` $= ((2 * 2) + 1) * 64 = 320$. Similarly, the number of a parameters for each fully connected layer is given by `((current layer n*previous layer n)+1)` $= 512 + 320 + 1$. Adding the parameters estimated for each layer, we obtain $320 + 320 + 320 + 163,814 + 262,145 = 426,946$ [180].

- The authors use the rectified linear units (ReLU) as activation function.



(a) Input (b) Features

FIGURE 6.2: **Convolution operation and features visualisation.** CNN input matrix [4] and visualisation of this input with our data. (a) Illustrates the convolution operation when F different locals filters are used, each of size J by J. (b) Visualisation of the Input matrix M by K for one of the audio signals used for training.

- The CNN does not have any pooling layer, since it decreases the performance of the network.
- The last layer uses softmax activation function to perform classification.
- The network was trained using the Adam optimiser [181], with a learning rate of 0.001, for 5 epochs, and uses categorical cross-entropy as loss function.
- The output of the CNN is the posterior probabilities of the input belonging to one of 37 DOA classes (discrete values from 0 to 180, with a gap of 5 degrees).

We tested the performance of this network to have a baseline for comparison. Fig. 6.4 illustrates this. It also presents the results of the sample experiments available in [5].

6.3.2 Acoustic Conditions

The training and testing conditions are summarised in Table 6.1. These conditions are the same as those described in [4], to aid comparison. Although the inter-microphone distance is the same for both training and test, the arrays are positioned in different locations within the rooms. The training data is composed of 5.6 million frames, while the test data is composed of 100 audio files per audio class, which are used to generate Room Impulse Response (RIR) for 9 different DOA: 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°and 150°. The RIR simulation is performed using the Image Source Method (ISM) [28].

TABLE 6.1: **Training and Testing Conditions**. Inter-microphone distance, source-array distance and reverberation conditions for training and testing simulations.

Parameter	Train	Test
Inter-mic distance	8 cm	8 cm
Source-array distance	1 m and 2 m	2 m
T_{60}	0.3 s, 0.2 s	0.1 s

6.3.3 Training Audio Classes

We used two different audio classes to train the CNN: speech and music. The reason behind this choice is the availability of speech and music data in datasets, as well as the frequency information they provide. For each of these classes we used different variations to produce this data, either by using datasets or methods to synthesise these sounds.

Speech

Six different types of speech training data are used, in order to improve the DOA estimation accuracy in different audio classes. The methods used for generating the training data are the following:

1. **Speech (TIMIT)** Data from the TIMIT dataset [182], containing data of 630 speakers from 8 major dialects of American English, who are reading phonetically rich sentences. The dataset was originally designed as a database of speech data for acoustic-phonetic studies, as well as the development and evaluation of automatic speech recognition systems. This data set includes silent frames, usually when the speaker pauses inbetween words, where there is little signal energy.

2. **Speech and Voice Activity Detector (VAD) (TIMIT+VAD)** The TIMIT speech data is pre-processed using a VAD [183], a technique in speech processing, used to detect absence of human speech. In this case, silent frames were detected using a VAD and later removed from the signal before training the Neural Networks (NN).

   In general, a VAD algorithm consists of three steps: first, there is a noise reduction stage; then, some features are extracted from a section of the signal (which is what is described here as a frame); and, finally, a classification technique is applied in order to evaluate whether the frame contains speech

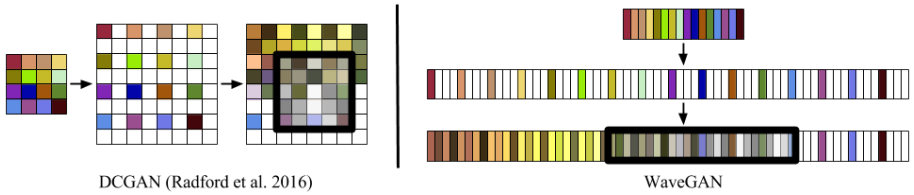


FIGURE 6.3: **Flattening process of DCGAN into WaveGAN.** Illustration of the transposed convolution operation for the first layers of the DCGAN. DCGAN uses 5 by 5 two-dimensional filters, while WaveGAN uses length-25 one-dimensional filters. The colours represent the position of the NN elements for a 2D input vs 1D input and how they are equivalent.

or not. In this step, the algorithm proposed in [184] is employed, using an implementation available in [183]. The authors use end-point detection to determine where speech begins and ends, and also to determine a speech threshold for initial estimation of silent frames. Moreover, they compute the zero crossing rate in the vicinity of endpoints, that is, the number of successive signal samples that have different algebraic signs. If frames above the initial threshold have considerable changes in zero-crossing rate, the endpoints are re-designed to the points at which the changes take place.

3. **Synthetic Speech (BSAR)** Synthetic speech signal, modelled by using a Block Stationary Autoregressive (BSAR) process [185]. Eq. 6.1 illustrates how the signal, $s(t)$, is modelled. $s(t)$ is partitioned into $\mathcal{M}$ continguos blocks, with block i begining at sample t_i and $e(t)$ the excitation process with variance σ.

$$s(t) = -\sum_{q=1}^{Q_i} b_i(q)s(t-q) + e(t), e(t) \sim \mathcal{N}(\mu, \sigma^2) \tag{6.1}$$

4. **Generative Adversarial Network (GAN) Speech (GAN-TIMIT)** Synthetic speech signal generated using an implementation of GAN, known as WaveGAN [14], trained with TIMIT speech data. WaveGAN is a machine learning algorithm based on GANs, which uses real audio samples to learn to synthesise raw waveform audio. The implementation provided by the authors is capable of learning up to 4 seconds of audio at 16kHz.

   GANs, originally proposed in [186], are composed of two NNs: a discriminator, D, and a generator, G. Being, P_X the distribution over data, x, and P_Z a prior on input noise variables, z, Eq. 6.2 illustrates the value function that G

is trained to minimise and D is trained to maximise. This means that D is trained to determine if an example is real or not using training data, while G is trained to try to fool the discriminator into thinking its output is real. The generator commonly uses randomized input as initial seed.

$$V(D,G) = \mathbb{E}_{x\sim P_X}[\log D(x)] + \mathbb{E}_{z\sim P_Z}[\log(1 - D(G(z)))] \tag{6.2}$$

The approach proposed in [14] is based on a two-dimensional deep convolutional GAN (DCGAN) proposed in [187], used for image synthesis. The authors bootstrap DCGAN to work on spectrograms, proposing an approach called SpecGAN. Moreover, they use a waveform approach called WaveGAN, which flattens the DCGAN architecture to work on one dimension. Fig. 6.3 illustrates the flattening process, in which DCGAN uses 5 by 5 two-dimensional filters, while WaveGAN uses length-25 one-dimensional filters. Moreover, they increased the stride factor for all convolutions, removed batch normalisation from generator and discriminator and finally trained using the WGAN-GP [188] strategy.

5. **GAN Speech (GAN-SC09)** Synthetic speech signal generated using WaveGAN [14], trained with Speech Commands Zero through Nine (SC09) data.

6. **GAN for Speech Data Augmentation (TIMIT+GAN-TIMIT)** Half of the data is from Speech (TIMIT) while the other half is synthetically generated using a waveGAN and no VAD is used.

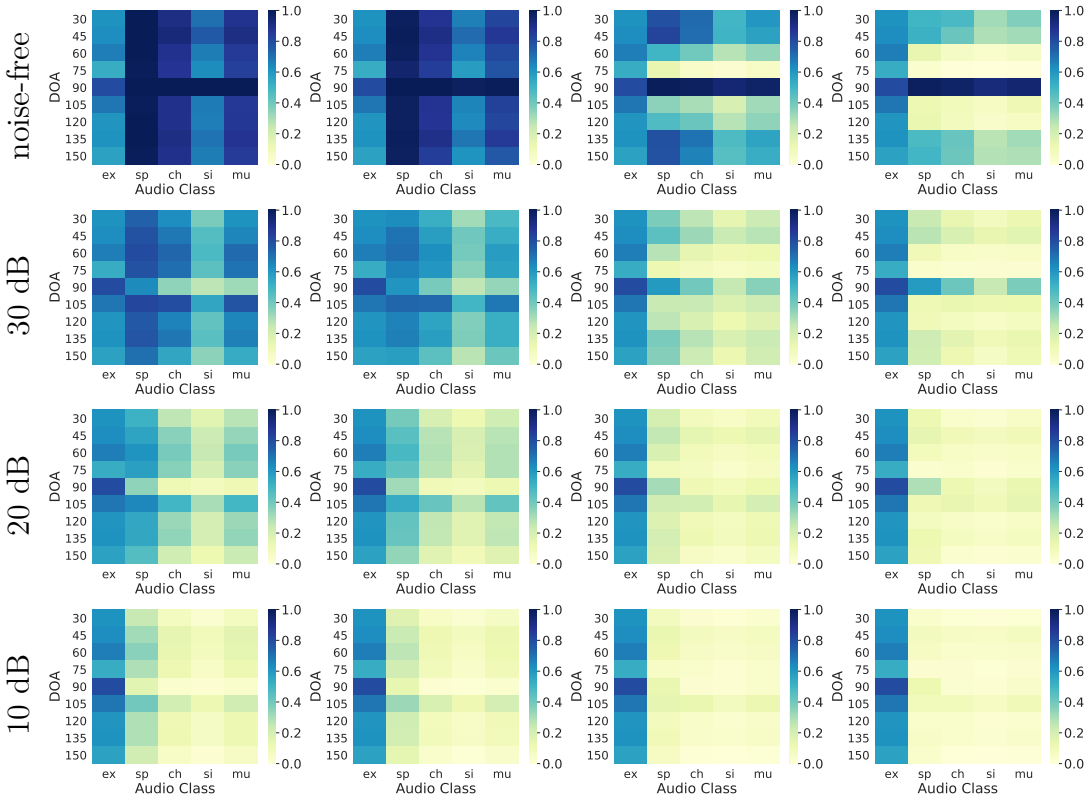
Music

1. **Street Music (StMu):** Data from the *UrbanSounds8k* dataset [189], which contains 27 hours of audio across 10 sound classes. The authors downloaded all sounds returned by Freesound search engine when using the class (e.g. "street music") as query. They then manually checked the recordings, kept the field recordings and label the start and end times of every ocurrence using Audacity. Signals from the class "street music" were selected to train the CNN.

2. **Street Music and VAD (StMu+VAD):** The Street Music data is preprocessed using a VAD [183] in order to remove silent frames.

3. **GAN Piano (GAN-Piano):** Synthetic speech signal generated using WaveGAN [14], trained with Piano data.

4. **GAN Drums (GAN-Drums):** Synthetic speech signal generated using WaveGAN [14], trained with Drums data.



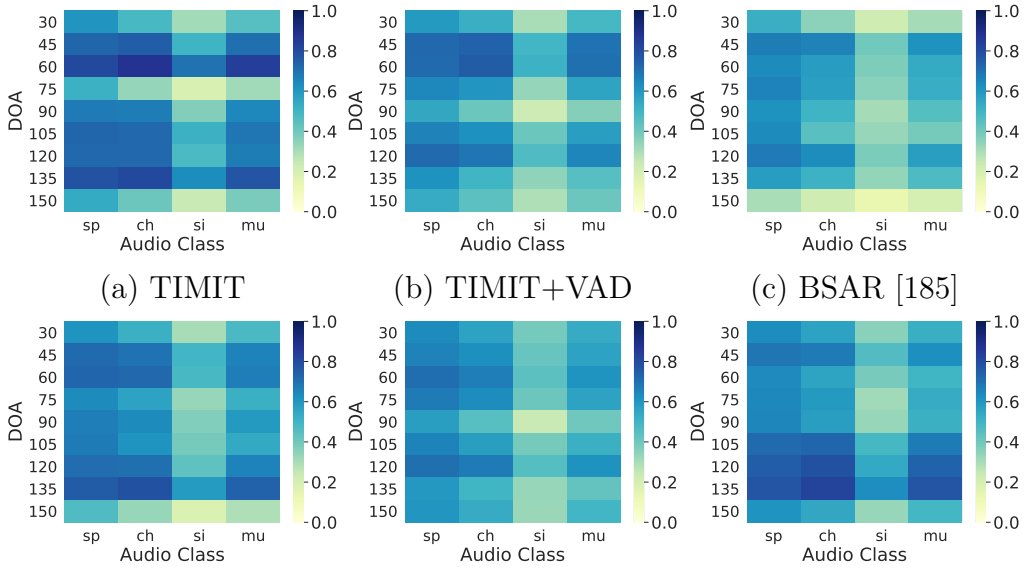
(a) Reverb. 0 s (b) Reverb. 0.1 s (c) Reverb. 0.2 s (d) Reverb. 0.3 s

FIGURE 6.4: **Accuracy of testing the pre-trained network.** Four different noise (noise free, 30dB, 20dB and 10dB Signal-to-Noise Ratio (SNR) from top to bottom) and reverberation (0s, 0.1s, 0.2s and 0.3s from left to right) conditions. The heat-maps illustrate the accuracy from 0 (yellow) to 1 (blue). The pre-trained network performed accurately for the speech class: however, the performance decreased when it was presented with new audio classes for testing, particularly in noisy and reverberant scenarios.

6.3.4 Testing Audio Classes

We tested the implementation in the following audio classes:

- **Example (ex):** Sample test speech data provided in [5], created when convolving a 13 sec long speech signal with Measured RIRs from the Bar-Ilan Multi-Channel Impulse Response Database [190].



(a) TIMIT (b) TIMIT+VAD (c) BSAR [185]

(d) GAN-TIMIT [14] (e) GAN-SC09 [14] (f) TIMIT+GAN-TIMIT

FIGURE 6.5: **A comparison of Direction of Arrival (DOA) estimation accuracy by training with different sources of speech data.** The heat-maps illustrate the accuracy from 0 (yellow) to 1 (blue). Using speech from the TIMIT dataset (a) or waveGAN (d) yields the best performance. However, training with any speech achieves higher accuracy than the baseline (second row of fig. 6.4) across audio classes. The test data is the same as that used for the baseline, with 30 dB SNR and 0.1 s reverberation.

- **Speech (sp):** The TIMIT dataset [182], as described above.
- **Urban Sounds:** Data from the *UrbanSounds8k* dataset [189], which contains 27 hours of audio across 10 sound classes. The classes used were: **Children playing (ch)**, **Siren (si)** and **Street music (mu)**.

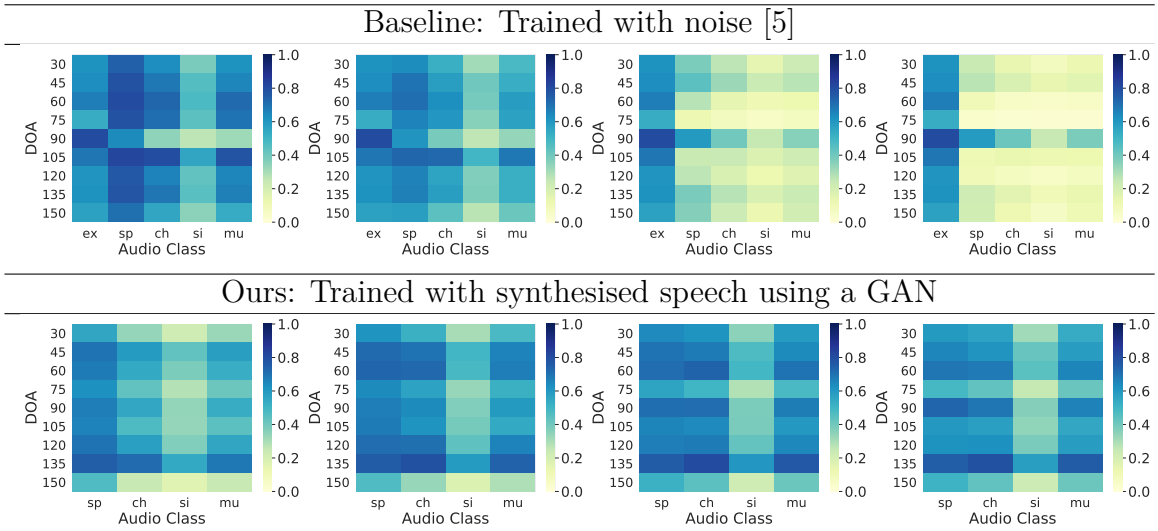
6.3.5 Evaluation Metric

In order to evaluate the trained network, *accuracy* is used as a performance metric, similarly to [5] and [68]. Accuracy is calculated as N_c/N_t, where N_c is the number of correctly classified frames and N_t is the total number of frames.

6.4 Experimental Results

For all experiments in this chapter, we use simulation [28] to mimic transport of the source signals to the microphone. The simulation introduces the appropriate delay

and adds noise and reverberation.

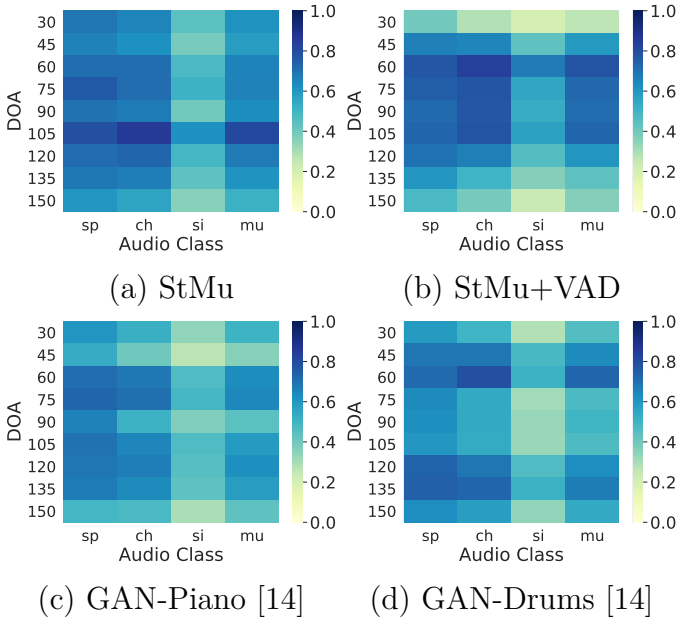


(a) Reverb. 0 s (b) Reverb. 0.1 s (c) Reverb. 0.2 s (d) Reverb. 0.3 s

FIGURE 6.6: **A comparison of the DOA accuracy (colorbar) for different audio classes (X-axes) and multiple incident directions (Y-axes).** The heat-maps illustrate the accuracy from 0 (yellow) to 1 (blue). The baseline (top row) performs well for speech signals (particularly at 90°) or when reverberation levels are low. Training with speech (bottom row) is more robust to incident directions as well as audio classes. The test data consists of simulated Room Impulse Responses using the Image Source Method, for 30 dB SNR. Legend: example [5] test data (ex), speech (sp), children playing (ch), siren (si) and street music (mu).

6.4.1 Baseline

In order to establish a baseline for comparison, we tested the performance of a pre-trained network available in [5], on different test audio classes. Fig. 6.4 illustrates the accuracy of testing the pre-trained network (trained using white noise) for four different Gaussian noise (noise free, $30dB$, $20dB$ and $10dB$ Signal-to-Noise Ratio (SNR)) and reverberation ($0s$, $0.1s$, $0.2s$ and $0.3s$) conditions. The test data is described in Section 6.3 and the room conditions are summarised by Table 6.1. Our hypothesis was that the pre-trained network would perform accurately for the speech class, but that, the performance would decrease when presented with new audio classes for testing. The results shown in the top row of Fig. 6.4, are good for speech data under low reverberation. For other audio classes, the accuracy drops by about 60% for higher reverberation simulations, confirming our hypothesis.



(a) StMu (b) StMu+VAD

(c) GAN-Piano [14] (d) GAN-Drums [14]

FIGURE 6.7: **A comparison of DOA estimation accuracy by training with different sources of music data.** The heat-maps illustrate the accuracy from 0 (yellow) to 1 (blue). Using speech from the Street Music class from Urban Sounds 8K (a) or WaveGAN trained with Drums (d) yields the best performance. However, training with any variation of music achieves higher accuracy than the baseline (second row of fig. 6.4) across audio classes. The test data is the same as that used for the baseline, with 30 dB SNR and 0.1 s reverberation.

6.4.2 Training with Speech

We trained the Convolutional Neural Network (CNN) using the six types of speech training data described in Section 6.3, and tested them on the same data as the baseline (see Table 6.1 for details). Our main hypothesis is that using speech for training the CNN will provide accurate results and will outperform the ones obtained with the baseline.

Fig. 6.5(a) illustrates the results obtained when the TIMIT database is used for training. It presents high accuracy for most angles (except 30°, 75°and 150°) and most audio classes (except the siren, which is the most challenging). Fig. 6.5(b) presents the results obtained when training with signals from the TIMIT dataset, pre-processed using a Voice Activity Detector (VAD). In comparison to Fig. 6.5(a), the accuracy decreased in general for most audio classes and angles, except for 45°, 60°, and 120°, where it is still above 60%. Fig. 6.5(c) shows the results obtained when the network is trained using synthetic speech from a Block Stationary Autoregressive (BSAR). This does not perform very well. Fig. 6.5(d) and (e) show the results using data

generated using WaveGAN, using TIMIT and SC09 respectively. Even though both generate accurate results, using the WaveGAN trained with TIMIT provides more accurate results than using the WaveGAN trained with SC09, particularly for 135° when it is very accurate. These results are comparable to the results using TIMIT. Finally, Fig. 6.5(f) illustrates the results obtained when the data from TIMIT is augmented using WaveGAN with TIMIT input. This latest approach is the one that presents the best results amongst speech, surpassing even the ones obtained with TIMIT for this particular CNN architecture. These experiments confirm our hypothesis that using speech for training the CNN provide accurate results for DOA estimation.

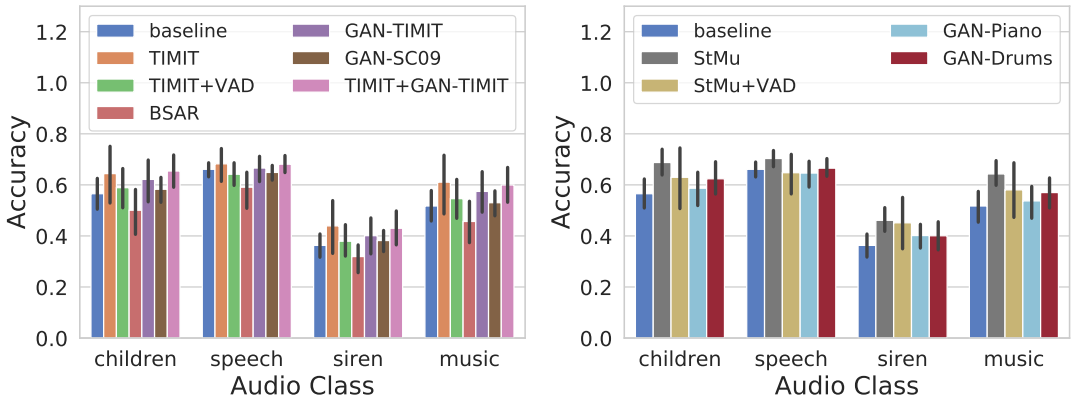
Fig. 6.6 presents the results obtained when using the pre-trained network from the baseline compared with the results obtained when we use synthesised speech from WaveGAN with TIMIT as input. The results show that our results are superior to the ones obtained by the baseline, particularly when the reverberation levels are high. This confirms our hypothesis that training the CNN using speech data outperforms the results obtained when the CNN is trained with noise.

6.4.3 Training with Music

We trained the CNN using the four types of music training data described in Section 6.3, and tested them on the same data as the baseline (see Table 6.1 for details). Our hypothesis in this case is that using music for training will provide accurate results, outperforming those of the baseline, though not as robust as those obtained with speech, since speech data uses specially recorded speech, while street music is recorded in urban scenarios, as explained in Section 6.3.

Fig. 6.7(a) illustrates the results obtained when training with Street Music signals, as recorded in the Urban Sounds 8K dataset. It shows that the accuracy is very high for all the tested angles and audio classes, except for siren, where the accuracy is around 40%. When using a VAD to remove silent frames, the accuracy obtained is decreased, as presented in Fig. 6.7(b). On the other hand, the use of WaveGAN to generate synthetic music data generates accurate results in both scenarios, but it shows better performance when the Generative Adversarial Network (GAN) is trained with Drums, Fig. 6.7(d), in comparison to when it is trained with Piano, Fig. 6.7(c). These results support our hypothesis that using music for training generates accurate results, outperforming those obtained using the baseline.

6.4.4 Speech vs Music



(a) Trained with Speech (b) Trained with Music

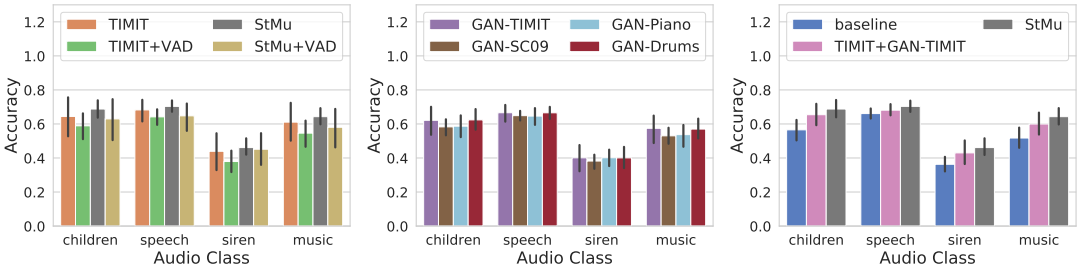
FIGURE 6.8: **Using synthesized speech (GAN) is marginally worse than using real speech data (TIMIT).** However, augmenting real speech with synthetic (TIMIT+GAN) performs similarly to TIMIT and with a lower standard deviation. Each bar depicts the accuracy averaged over 9 different DOA angles and 4 different audio classes, in a simulated scenario with 30 dB SNR and 0.1 sec reverberation.

Fig. 6.8(a) compares the average accuracy for all DOA on the test set for the different test audio classes, obtained when using a CNN trained with variations of speech data. In general, training the neural network using data from the TIMIT dataset presents the most accurate DOA estimation, not only for the test that uses speech, but also for the rest of the audio classes. Similar results are obtained when using data generated from WaveGAN for training. In both cases, the accuracy outperforms that obtained using the pre-trained network (baseline). In contrast, training using a VAD to pre-process the signals or using synthetic speech from a BSAR process decreases the accuracy of the DOA estimation.

Similarly, Fig. 6.8(b) compares the average DOA accuracy, when the network was trained with variations of music. In this case, the best results are obtained when training directly with Street Music (StMu), even when a VAD is used. The use of synthetic data from a GAN is not as accurate as in the case of speech: however, they outperform the results obtained using the baseline for children, siren and music audio classes.

In Fig. 6.9 we compare the various variations we used for training among themselves in order to determine the best training strategy depending on the test scenario. Fig. 6.9(a) illustrates the case in which the datasets and VAD are used for training.

In this case, Street Music generates the best results for all the test audio classes, even when a VAD is used. In contrast, Fig. 6.9(b) illustrates the comparison when data from WaveGAN is used. In this scenario, the best results are obtained when TIMIT speech data is used as input for the GAN. Finally, Fig. 6.9(c) compares the best results for each type of training data against the baseline. This confirms that training with either speech or music produces more accurate results than using the baseline and the best results are obtained when training with Street Music data. This also confirms that our hypothesis that training with speech is better than training with music is not completely accurate, since the best results are obtained using Street Music. However, it is important to remember that when using data from WaveGAN, it is better to use speech rather than music. This behaviour could be caused because the ability of the CNN to learn the room acoustics is much better when trained with music data, given the frequency information that it provides. However, when the data is synthesised using a GAN, this frequency information might not be represented as accurately as in the dataset.



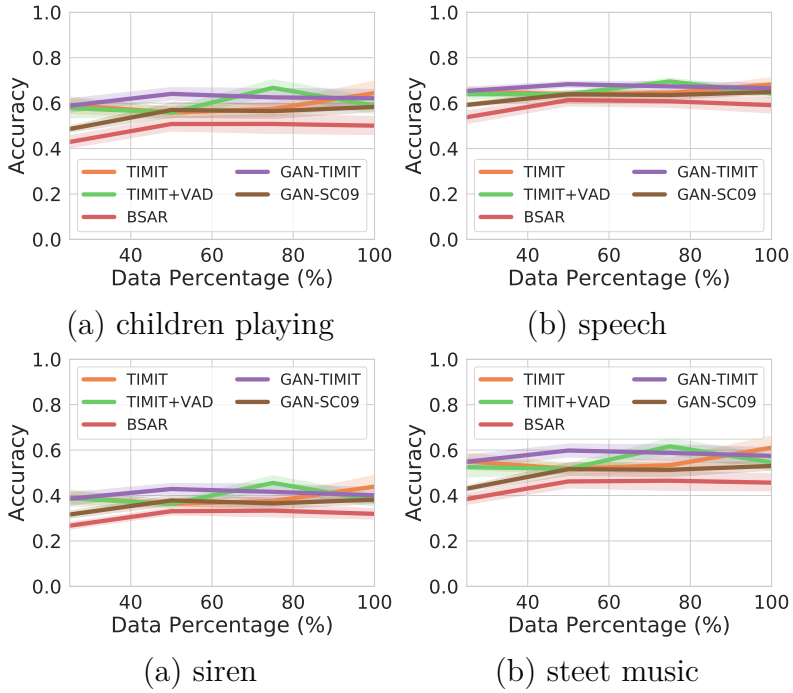
(a) Speech vs music (b) GANs (c) Best vs baseline

FIGURE 6.9: **Comparison of training strategies.** Datasets and synthetic data from speech and music.

6.4.5 Amount of Data

We investigated the impact of decreasing the amount of training data on the accuracy of DOA estimation. Our hypothesis is that the data from datasets will be more affected by the change in the amount of data, rather than the data from the GAN, since the first one has more variation between samples, while the latter one is more homogeneous.

Fig. 6.10 presents the results of this study for a network trained with speech. We used different percentages of the original training data, 25%, 50% and 75%. In general, the five proposed training methods do not present a high variation in



(a) children playing (b) speech

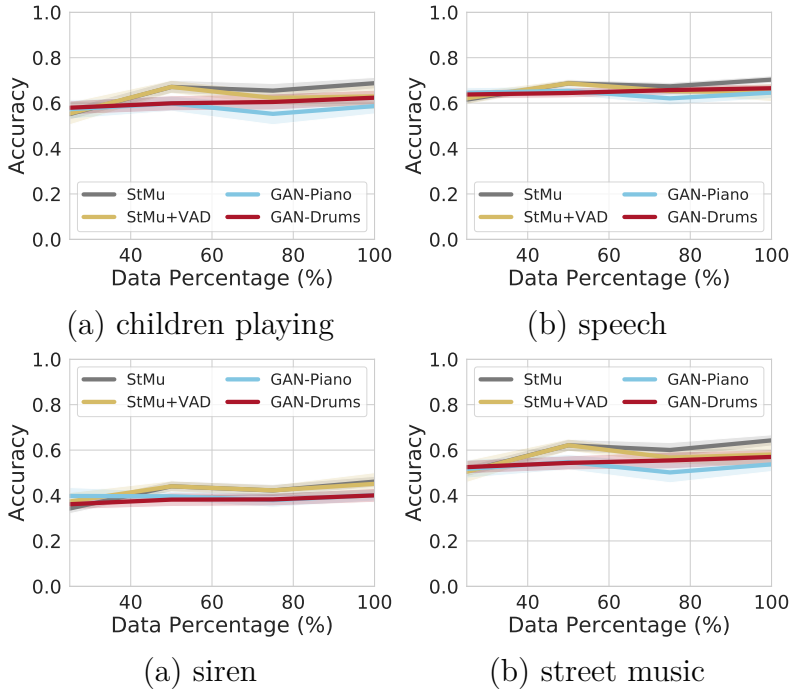
(a) siren (b) steet music

FIGURE 6.10: **Impact of the volume of training data (X-axes) on accuracy (Y-axes) for five different speech training datasets.** Training with synthesized speech, BSAR and GAN, exhibits the lowest variation across different training volumes with the latter performing better. 100% corresponds to the full training data used in other experiments.

accuracy; however, training with WaveGAN yields the least change in accuracy, even when the amount of data used is 25% of the original set.

Fig. 6.11 presents the same results, but for a network trained with music. Similarly to the speech case, there is a large variation in the accuracy; however, using data generated with WaveGAN produces a smaller change in accuracy than it does to use data from the dataset directly or even using a VAD, which produces the highest variation.

These experiments slightly confirmed our hypothesis that data generated from GAN produces the smallest variation in the output when the amount of training data is considerably decreased. However, overall, the change in the accuracy is so small for all the training methodologies that it does not yield a meaningful conclusion. Moreover, since 25% is a considerable reduction in the amount of data, we decided not to pursue any further decrease in our experiments.



(a) children playing (b) speech

(a) siren (b) street music

FIGURE 6.11: **Impact of the volume of training data (X-axes) on accuracy (Y-axes) for four different music training datasets.** Training with synthetic data from a GAN exhibits the lowest variation across different training volumes. 100% corresponds to the full training data used in other experiments.

6.4.6 Learning vs Cross-Correlation

Finally, we compare our method against a traditional approach that uses Generalized Cross-Correlation (GCC), to understand the relative merits of machine learning. Fig. 6.12 illustrates the DOA estimation accuracy under two different reverberation conditions, one that was used during training (0.3 s) and one that was not (0.1 s). For 0.1 s, it can be seen that both GCC and GAN perform very similarly across the four audio classes. For 0.3 s, however, both GAN clearly outperform GCC, especially for DOA 30°, 45°, 135°and 150°, where the accuracy improves 16× on average. This suggests that the CNN is potentially learning information about the room acoustics, whereas GCC assumes a free-field environment.

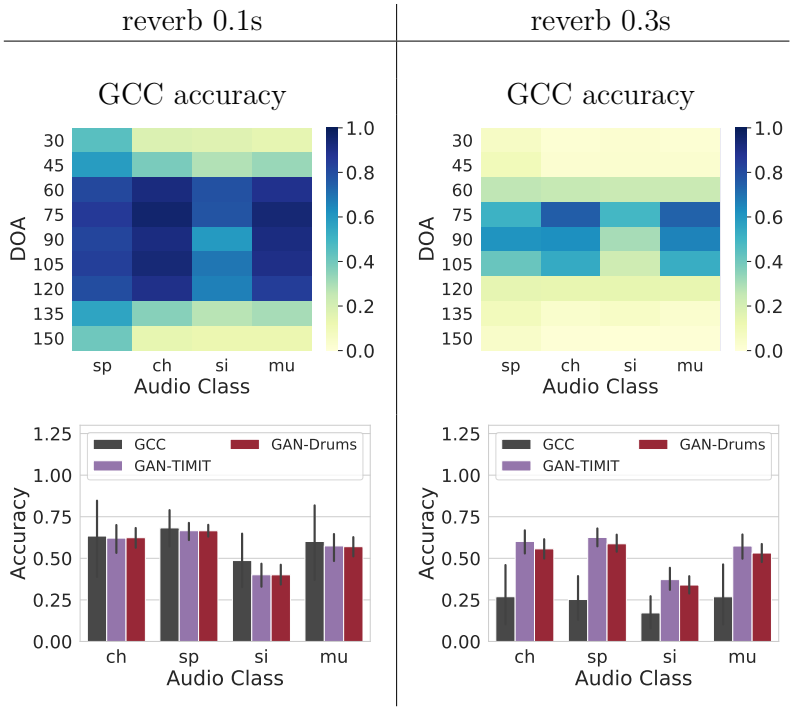


FIGURE 6.12: **Comparison of waveGAN-trained network with GCC under different reverberation conditions.** The heatmaps illustrate the accuracy from 0 (yellow) to 1 (blue). The network used was trained with reverb. of 0.3s. When the test environment is different (left), the network performs similarly on average (but with lower variance). When the test condition matches training (right), the network outperforms GCC. As expected, GCC's performance suffers when the reverberation is increased. An advantage of using supervised learning is that the method can be trained to handle such difficulties.

6.5 Discussion

6.5.1 Nature and Volume of Training Data

The main finding in this chapter is that using real music data for training Convolutional Neural Networks (CNNs) yields the most accurate Direction of Arrival (DOA) estimation. This is followed by training using augmented real speech data with synthetic speech. Curiously, we observed that generating synthetic speech with WaveGAN yields about 15% improvement in accuracy over methods such as synthesis using a Block Stationary Autoregressive (BSAR) model. On the other hand, using a Voice Activity Detector (VAD) decreases the accuracy around 8% when it is used in speech data, but only in 3% when used on music data. The use of WaveGAN to generate training data provides higher accuracy for speech than for music, but only by 2% on

average. In practical terms, when training with audios from datasets, it is better to use music rather than speech, however when using data from WaveGAN, it is better to use speech rather than music.

We also observed that using only 25% of the training data (as reported in other experiments in this chapter) was sufficient to obtain similar accuracy. Furthermore, for a given method (and training data), we find that the accuracy is not very dependent on the amount of training data. This appears particularly true when synthetic data is used in training. Further investigation is required to understand why this might be true.

6.5.2 Advantage of Learning

Training can be viewed as advantageous when certain aspects of the test conditions might be known *a priori*. For example, training data may be generated specific to the acoustic behavior of a particular auditorium if the goal is to track only speakers in that auditorium. Although traditional methods such as Generalized Cross-Correlation (GCC) do not require training, this can be seen as a shortcoming since such specific information cannot be encoded. For example, if reverberation within the auditorium is known to be high, it is not trivial to develop a method that augments GCC with that information.

6.5.3 Limitations and Future Work

The main limitation of supervised learning is its difficulty in generalisation. For example, training a CNN to suit a variety of acoustic environments incurs a penalty (of lower accuracy). Further investigation is required to ascertain the details of this trade-off between accuracy and generalisation.

6.6 Conclusions

In this chapter, we presented novel findings regarding the training data used to train a Convolutional Neural Network (CNN) for Direction of Arrival (DOA) estimation.

First of all, we observed that training using noise was not very robust to test signals that involved various audio classes different from speech, therefore we decided to use variations of speech and music data, which come from either data sets or synthetic approaches. We discovered that training with music data performs better

than training with speech data and both of them performed better than training with noise.

Next, we compared variants of speech and music data. The speech data included a speech dataset (TIMIT), pre-processed speech data using a Voice Activity Detector (VAD), synthetic data using a Block Stationary Autoregressive (BSAR) process and synthetic data using a Generative Adversarial Network (GAN). Our results indicate that using a combination of real and synthetic (using WaveGAN) data performs best. The music data, on the other hand, included a street music dataset (StMu), pre-processed data using a VAD, synthetic data using a GAN from two different instruments, piano and drums. Our experiments showed that using the data from the dataset (StMu) performed best. Moreover, when comparing the results obtained when training with speech and music, we concluded that, when using data from recorded datasets, the best results are obtained when using music; however, when using synthetic data from GAN, the best results are obtained using speech.

We also investigated the impact of the amount of data used for training the CNN. It is encouraging to note that using just 25% of the training data does not notably reduce estimation accuracy, either with speech or music. Synthetic data generated with GAN is slightly less prone to changes in the accuracy than real data from datasets.

Finally, we showed how the use of a learning-based approach overcomes the limitations of the Generalized Cross-Correlation (GCC) approach in scenarios in which there is some a priori knowledge of the test environment.

We trained and tested our algorithm in simulated environments, using the same microphone configuration in both scenarios. A reproduction of this set-up in real scenarios could lead to results similar to the ones obtained in our experiments.

The decision to use speech and music is related to the variation in amplitude in the frequency range of these sounds. This is opposed to the baseline that used white noise, which has the same amplitude in the whole range of frequencies. This variation in the frequency range allows the CNN to learn a better representation of the room acoustics, increasing the DOA estimation accuracy.

Although the results presented in this chapter have only been tested using a single CNN architecture, it should be possible to extend them to further learning-based methods. In principle, the characteristics of speech and music sounds, mentioned in the previous paragraph, could favour the learning process of any Neural Networks (NN): however, due to lack of implementations available online, the test for different CNN architectures is left as future work.

Our contribution, then, could be summarised as follows:

- we demonstrated the positive impact of the use of variations of speech and music data in the training of a CNN for DOA estimation, when the test data involves a variety of audio classes. Our results suggest an improvement of 19% compared to a baseline that trains with noise.

- we showed that using synthetic signals generated using a GAN for training produces results as accurate as those obtained using signals from datasets.

- we compared the CNN against GCC and demonstrated that the learning algorithm performs better in conditions in which the train and testing environments are very similar, while performing comparatively when they are not.

In conclusion, we have shown that a variation/variations of speech and music data could be used to train a CNN for DOA estimation. These variations included the use of a GAN, which performs similarly to data from datasets. Moreover, we showed that using only 25% of the training data generates the same performance as if the total amount were used. Finally, we showed the advantages of using a CNN when the training and testing data present similar acoustic conditions.

Future work includes the use of simulated data for training and real data for testing, using transfer learning. Moreover, the use of different NN architectures will allow us to reach broader conclusions. Finally, including additional audio classes for testing would be another interesting future direction.

Part III

Conclusions

Chapter 7

Conclusions

This thesis presented work on Acoustic Source Localisation (ASL) in constrained environments. The three constraints studied were the number and configuration of sensors; the signal samples; and training data, with the main findings summarised as follows:

- In regard to the number and configuration of sensors, accuracy can be maintained at state-of-the-art levels (SRP) while reducing computation sixfold.
- In regard to signal sampling, the algorithm presented in this work outperforms an audio fingerprinting baseline while maintaining a compression ratio of 40:1.
- In regard to training data, music training data is used to record an improvement of 19% against a noise data baseline using only 25% of the training data.

7.1 Summary

This thesis presented work examining Acoustic Source Localisation (ASL) in constrained environments. Experiments were conducted using simulated and real data.

Chapter 1 presented an introduction to the topic of acoustic source localisation. It started with an explanation of the various applications that ASL offers and, later on, it explained how microphone arrays are used as a tool to solve the problem. Moreover, this chapter introduced the three types of constraints in ASL studied in this thesis: the number of microphones, the amount of signal samples needed, and the data available for training. The chapter finished with an outline of the thesis and highlighted the main contributions that were going to be presented in each chapter.

Chapter 2 summarised the basic concepts relating to sound from a physics perspective. It started with the derivation of the wave equation, followed by a brief

explanation of how sound is measured. There were also explanations of the concepts of frequency and reverberation. The chapter ended with an introduction to the Image Source Method (ISM), a description of the problems related to microphone arrays, and the definition of far-field and near-field in ASL.

Chapter 3 presented an overview literature review of the work undertaken on ASL. It attempted to classify the ASL literature, as well as demonstrating the advantages and disadvantages of various approaches. Examples of these include subspace-based techniques, steering-based approaches, blind system identification, optimisation-based methods and feature-based methods. The chapter was particularly focused on Time Difference of Arrival (TDOA)-based methods, since these are widely used throughout this thesis.

Chapter 4 was focused on the **number and configuration of sensors** constraint and presented novel findings concerning direct optimisation of ASL and the impact of the sensor array configuration on localisation accuracy. We started by comparing various indirect ASL approaches, based on Times of Arrival (TOA) and TDOA, establishing that, when the source is not synchronised with the microphones – that is, the emission time is unknown – indirect approaches based on TDOA are more robust than TOA-based ones as noise increases. This prompted us to apply a TDOA-based approach using real data to estimate the location of a variety of sound sources from different audio classes. We found that, when using 100 randomly chosen microphones from these pairs, the localisation accuracy is the same as that obtained using a state-of-the-art technique, Steered Response Power (SRP), but with **6 times less computation** in three out of four different datasets. In parallel, we also applied this TDOA-based optimisation to three different microphone configurations: ring, wheel and spiral. We observed that, for some random source locations, there was an increase in the relative error in localisation when the ring configuration was used, compared with the other two configurations. Further experiments in a simulated room showed that this pattern persists, particularly when the sources are located in front of the microphone array. Experiments with real data confirmed that the ring configuration produces the highest localisation error when the source is located in front of the microphone array. The experiments in this chapter were conducted using simulated data by means of the ISM and validated using data recorded using a static source and a microphone array in an uncontrolled environment. Moreover, the sounds recorded were from a variety of audio classes from real-life scenarios. Therefore, we believe that our findings could be used and applied to any microphone array system.

Chapter 5 concentrated on the **signal samples** constraint and presented novel

findings regarding accurate Direction of Arrival (DOA) estimation using only a compressed version of the input signal. Our main outcome was the design of an algorithm that selects certain signal samples to accurately estimate TDOA. We were inspired by the use of a famous computer vision algorithm, Scale-Invariant Feature Transform (SIFT), to detect keypoints in the signal spectrogram. We used the detected keypoints together with the cross-correlation algorithm to estimate the TDOA. We started by testing our algorithm in a microphone pair using 100 montecarlo simulations, and we realised that, for low noise and reverberation conditions, the mode of the estimated TDOA corresponded to the ground truth. We then proceeded to compare our algorithm against a compression approach using subsampling, and were able to demonstrate how our algorithm estimated the TDOA accurately, even when the compression ratio was significantly increased. Therefore, we continued our tests and demonstrated how our algorithm estimated the TDOA for high compression ratios in various noise and reverberation conditions. We continued our test using various signals from the TIMIT dataset located at different DOA. We showed how our algorithm accurately estimated the TDOA and DOA with high compression ratios, in scenarios in which the noise and reverberation were low. We showed that our algorithm can achieve a **40:1 compression ratio**. Finally, we compared our algorithm against an approach based on audio fingerprinting and demonstrated that we were able to outperform the baseline. While the scenarios in which the algorithm was tested were generated by means of a simulator, the data used for testing was real speech from a well-known dataset. Therefore, we believe that our system capabilities could easily be extended to real-life scenarios.

Chapter 6 was concerned with the **training data** constraint and we presented novel findings regarding the training data used to train a Convolutional Neural Network (CNN) for DOA estimation. We used a CNN that performs DOA estimation and was trained with noise, in order to predict the DOA in datasets from a variety of audio classes. First of all, we observed that, while the network worked well with speech, it failed to correctly estimate DOA for other audio classes. As a result, we decided to use variations of speech and music data as input to train the CNN. Our main finding was that using music data for training produces more accurate results than using speech data, and both of them perform better than using noise for training. Next, we compared variants of speech and music data. The speech data included a speech dataset (TIMIT), pre-processed speech data using a Voice Activity Detector (VAD), synthetic data using a Block Stationary Autoregressive (BSAR) process and synthetic data using a Generative Adversarial Network (GAN). Our results indicate that using a combination of real and synthetic (using WaveGAN)

data performs best, with an **improvement of 17%** with respect to the baseline. The music data, on the other hand, included a street music dataset (StMu), pre-processed data using a VAD, and synthetic data using a GAN from two different instruments, piano and drums. Our experiments showed that using the data from the dataset (StMu) performed best, with an **improvement of 19%** with respect to the baseline. Moreover, when comparing the results obtained when training with speech and music, we concluded that, when using data from recorded datasets, the best results are obtained when using music; however, when using synthetic data from GAN, the best results are obtained using speech. We also investigated the impact of the amount of data used for training the CNN. It is encouraging to note that using just 25% of the training data does not notably reduce estimation accuracy, either with speech or music. Synthetic data generated with GAN is slightly less prone to changes in the accuracy than real data from datasets. Finally, we showed how the use of a learning-based approach overcomes the limitations of the Generalized Cross-Correlation (GCC) approach in scenarios in which there is some a priori knowledge of the test environment. We trained and tested our algorithm in simulated environments, using the same microphone configuration in both scenarios. A reproduction of this set-up in real scenarios could lead to results similar to the ones obtained in our experiments.

7.2 Future Work

Array Configuration and Microphone Pairs

The research presented in this thesis is limited to experiments with a fixed-size microphone array, where only the microphone configuration could be modified. One possible extension to this work could be to include the use of a larger range of scenarios, involving different microphone array sizes and more configurations beyond the three ones studied. Moreover, the error quantisation could be extended to the use of angular errors in 3D, which could provide a new understanding of the impact of configuration.

Signal Samples

The work presented in regard to the signal samples constraint is not robust to high levels of noise and reverberation. Future work in this field might include the use of neighbouring pixel information to identify reverberations and noisy values and remove them from the final binary mask. Moreover, this work dealt solely with

speech signals, which are rich in features. Nevertheless, the use of simpler signals might present a challenge for the current version of the algorithm, and therefore studying a different type of audio class might provide interesting results that could be applied to a large variety of situations. Last but not least, there is also scope to consider the estimation of Time Difference of Arrival (TDOA) values when they are small.

Training Data

The work presented was focused solely on an existing neural network architecture, therefore one possibility to extend the research in this area could be the comparison with further network architectures. Moreover, the metric used to evaluate the algorithm was the classification error, which could potentially be extended to the relative error or the fine error (in degrees). Another option would be to include new audio classes for testing. Finally, future work could also include the use of simulated data for training the neural network and real data for testing, using transfer learning.

7.3 Conclusion

This work presented research into Acoustic Source Localisation (ASL) for three types of constraints: number and configuration of sensors (Chapter 4), signal samples (Chapter 5) and training data (Chapter 6).

In our work on the first constraint, the number and configuration of sensors, we were able to take a formulation of ASL and apply it to estimate the location of a sound source with errors similar to the state-of-the-art Steered Response Power (SRP), but with **6 times less computation**, using a limited number of microphone pairs. Moreover, we were able to determine that the use of circular arrays yields higher localisation errors than spiral and wheel configurations for large regions of space.

Turning our attention to the signal samples constraint, we proposed a novel encoding scheme for estimating Time Difference of Arrival (TDOA). This was achieved by applying a well-known computer vision technique to select signal samples from the spectrogram of a speech signal. This subsequently allowed us to estimate to use these samples of the signal to estimate TDOA within a reasonable margin of relative error. We tested the robustness of the proposed technique under different noise and reverberation conditions using different speech signals and source locations. The results demonstrate that our algorithm is able to estimate TDOA and the source location within an acceptable error range while **maintaining a signal compression**

ratio of 40:1. We compared our technique against a baseline that uses audio fingerprinting and showed that our approach presents superior results.

Lastly, our work on the third and final constraint examined the training data required to estimate Direction of Arrival (DOA) using a deep learning-based approach. Our findings indicate that using variations of music and speech data for training produces more accurate results for various audio classes than those obtained with noise using the state of the art approach. These variations included the use of a Generative Adversarial Network (GAN), which performs similarly to data from datasets. Our results suggest an **improvement of 19%** compared to a baseline that trains with noise. Moreover, we showed that using only 25% of the training data generates the same performance as if the total amount were used. Finally, we compared the Convolutional Neural Network (CNN) against Generalized Cross-Correlation (GCC) and demonstrated that performance is comparable when the reveberation levels are different in training and testing, and significantly improved in conditions in which the training and testing environments are very similar, being 125% more accurate than GCC.

In closing, then, we were able to demonstrate material improvements to ASL in regard to the three constraints we set out to consider in this work: number and configuration of sensors; signal sampling; and training data, and hope that this will serve as a basis for further study and investigation into these aspects of the field.

Part IV

Appendices

Appendix A

TDOA Errors

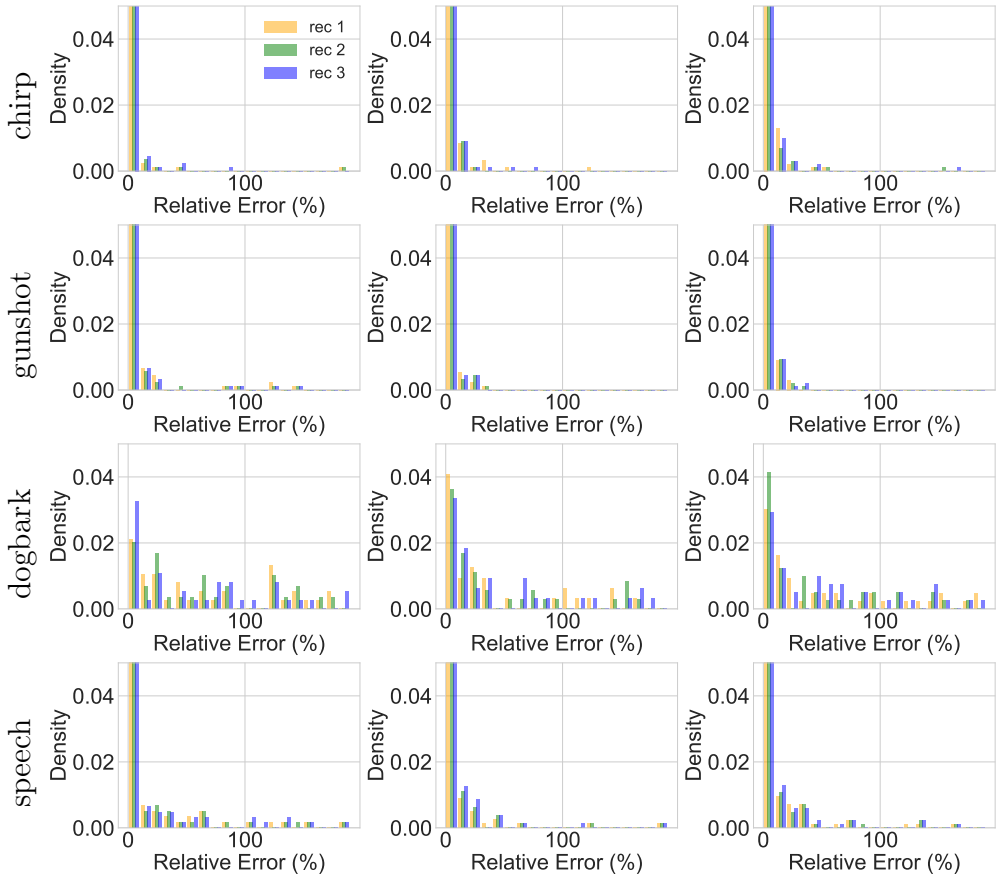
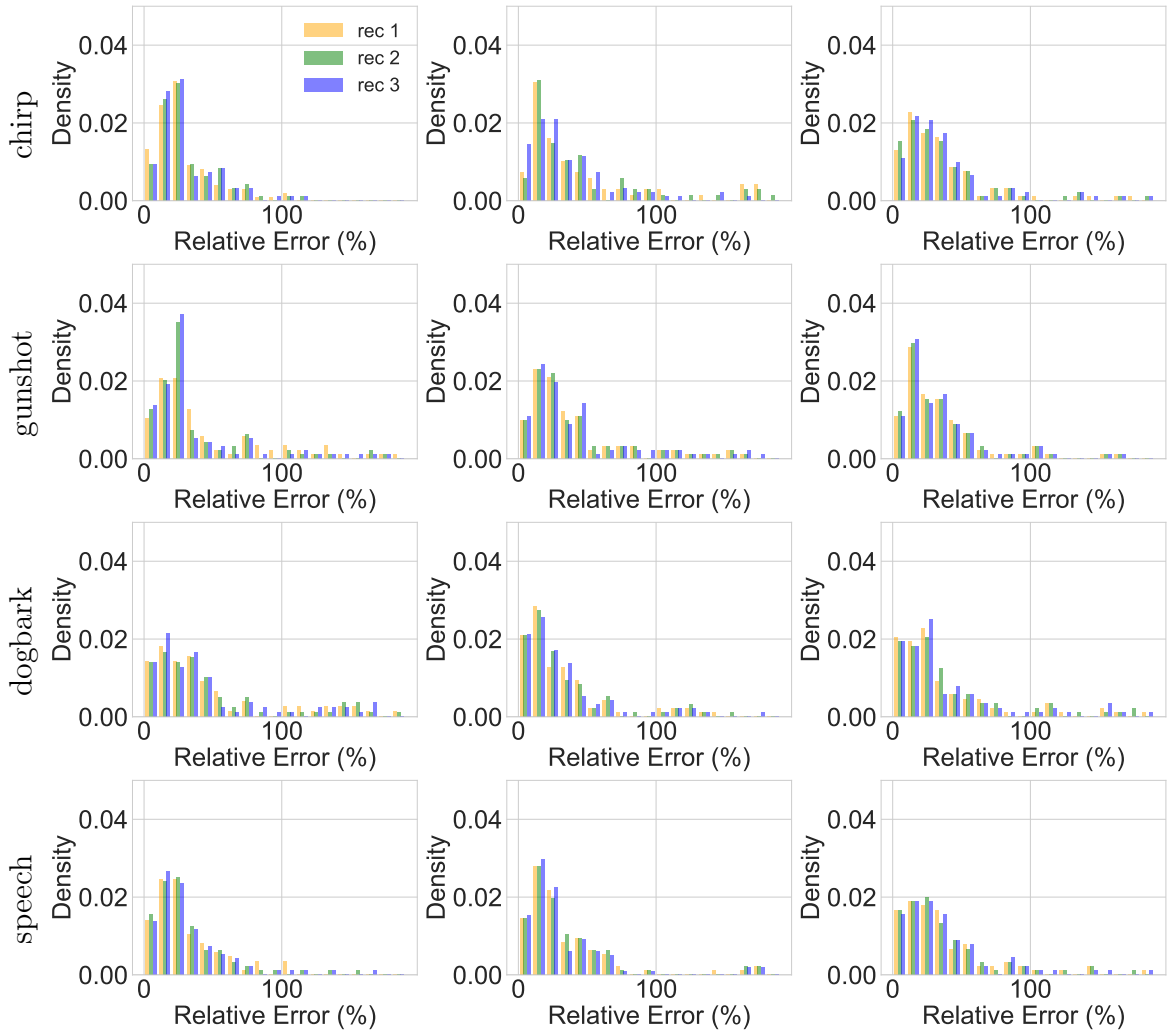
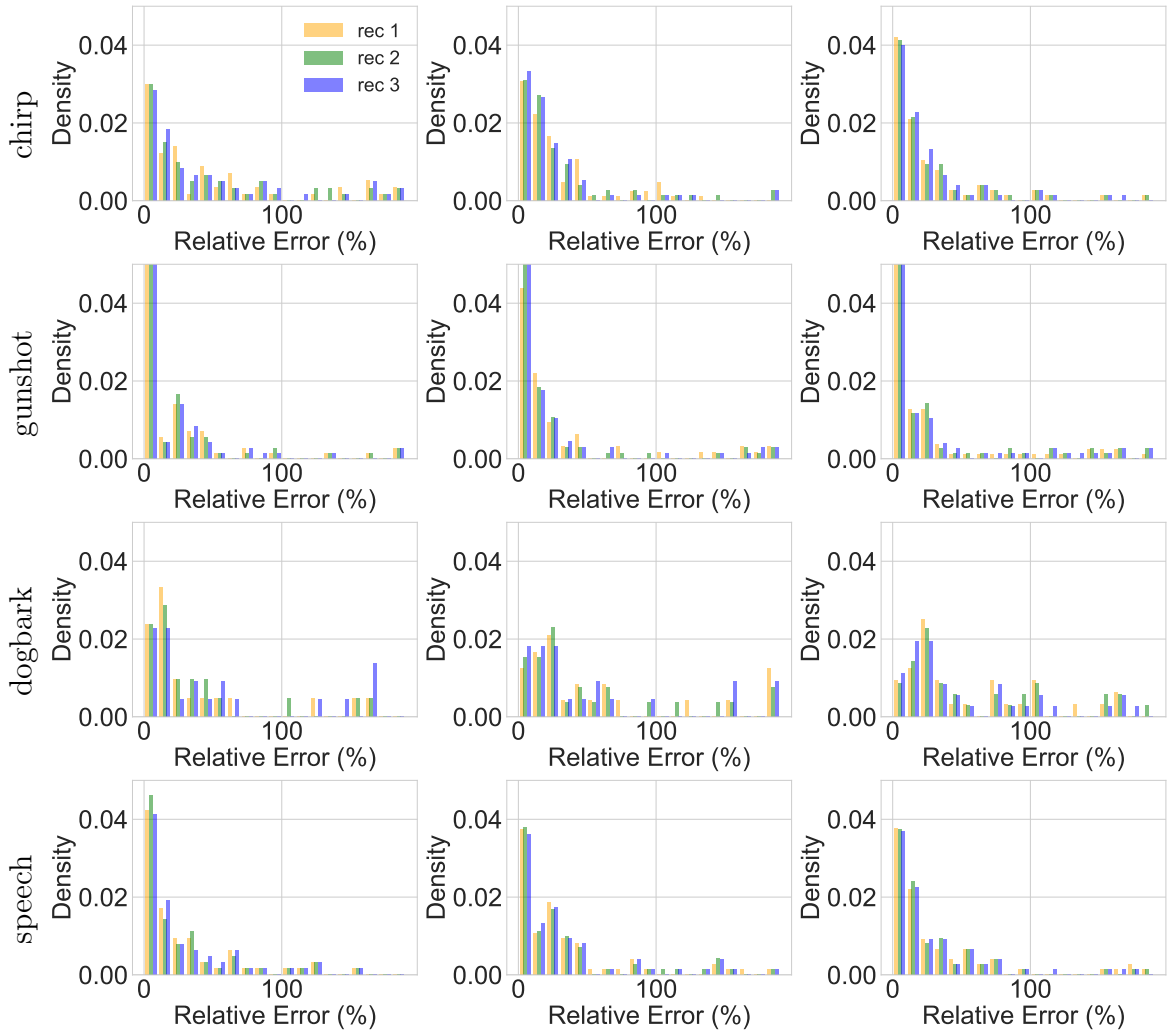


FIGURE A.1: **TDOA errors for source located at A:(2.0,-0.32,0.5).** Each row represents the results of a dataset: chirp, gunshot, dogbark and speech respectively. For each of them, the results for three different recordings are illustrated by the color bars. The histogram shows a larger error for the *dogbark* dataset, arising from the use of the Generalized Cross-Correlation Phase Transform (GCC-PHAT) and the repetitive pattern of the signal.



(a) Ring (b) Spiral (c) Wheel

FIGURE A.2: **TDOA errors for source located at C: (0.0,-0.32,1.5).** Each row represents the results of a dataset: chirp, gunshot, dogbark and speech respectively. For each of them, the results for three different recordings are illustrated by the colour bars. The histogram shows a greater error for the *dogbark* dataset, arising from the use of the GCC-PHAT and the repetitive pattern of the signal. The rest of the signals present a low TDOA relative error for the three different microphone configurations

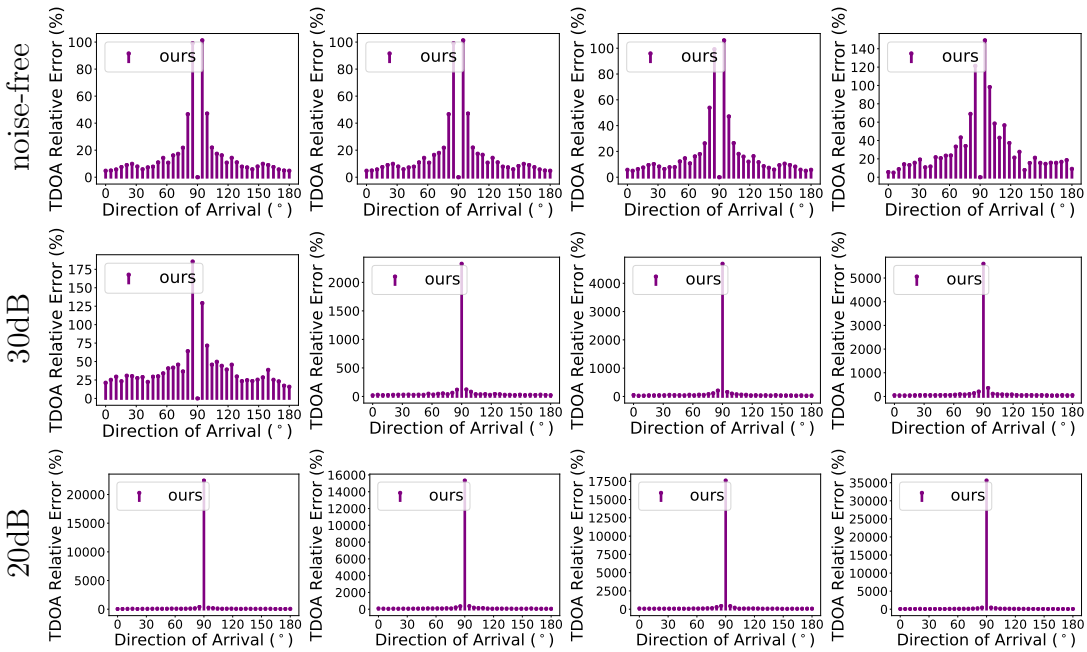


(a) Ring (b) Spiral (c) Wheel

FIGURE A.3: **TDOA errors for source located at E: (-1.5,-0.32,3.5).** Each row represents the results of a dataset: chirp, gunshot, dogbark and speech respectively. For each of them, the results for three different recordings are illustrated by the colour bars. The histogram shows a greater error for the *dogbark* dataset, arising from the use of the GCC-PHAT and the repetitive pattern of the signal. The rest of the signals present a low TDOA relative error for the three different microphone configurations

Appendix B

TDOA vs DOA



(a) $T_{60} = 0$ s (b) $T_{60} = 0.1$ s (c) $T_{60} = 0.2$ s (d) $T_{60} = 0.3$ s

FIGURE B.1: **TDOA relative error vs DOA.** TDOA relative error (purple) for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds and 3 noise conditions: noise-free, $30dB$ and $20dB$ Signal-to-Noise Ratio (SNR). The results are from 10 speech signals, at 19 different locations (DOA), from 0°to 180°, with a step size of 10°. We ran 5 different simulations for each of these sources and reverberation values. The compression ratio is $40:1$ for each signal.

Appendix C

Publications

IMPACT OF MICROPHONE ARRAY CONFIGURATIONS ON ROBUST INDIRECT 3D ACOUSTIC SOURCE LOCALIZATION

Elizabeth Vargas, Keith Brown

Heriot-Watt University
Edinburgh, United Kingdom

Kartic Subr

University of Edinburgh
Edinburgh, United Kingdom

ABSTRACT

Acoustic source localization (ASL) is an important problem. Despite much attention over the past few decades, rapid and robust ASL still remains elusive. A popular approach is to use a circular array of microphones to record the acoustic signal followed by some form of optimization to deduce the most likely location of the source. In this paper, we study the impact of the configuration of microphones on the accuracy of localization. We perform experiments using simulation as well as real measurements using a 72-microphone acoustic camera which confirm that circular configurations lead to higher localization error than spiral and wheel configurations when considering large regions of space. Moreover, the configuration of choice is intricately tied to the optimization scheme. We show that direct optimization of well known formulations for ASL yield errors similar to the state of the art (steered response power) with $6\times$ less computation.

***Index Terms*—** 3D acoustic source localization, microphone array configuration

1. INTRODUCTION

The problem of estimating the 3D position of objects is called *localization*. Despite the advancement in localisation using visual features, the use of audio sensing has important advantages such as reliability under poor illumination, inexpensive sensing equipment and the use of signal processing (1D) tools. There have been attempts to use audio localization in robotics [1] and in scene understanding [2]. *Acoustic source localization* (ASL) is typically achieved by leveraging known discrepancies in measurements of the emitted signal at multiple locations. ASL algorithms may exploit differences in time, amplitude or both.

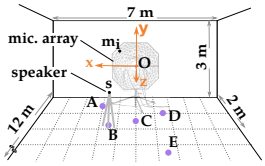
Some approaches to ASL, such as Steered Response Power (SRP) [3, 4], directly solve for the most likely position of the source amongst a grid of candidate locations. ‘‘Indirect" methods first estimate the times of arrival (TOA) at the sensors (microphones) or time differences of arrival (TDOA) across pairs of microphones and then use this information to infer the source position via multilateration [5, 6]. Although indirect methods are simpler to express as a least squares optimization [7], the resulting objective function is non-convex and often does not lend itself to an analytical solution. Various reformulations of these methods using weighted least squares, convex constrained least squares [8], total weighted least squares [9] and weight constrained total least squares [10] have been analyzed. Direct methods are believed to be more robust to noise and reverberation [3].

A uniform circular array of microphones[11, 12] along with a ring configuration [13] is a common choice for taking measurements since azimuthal angles to sources are considered more important than elevation. The advantage of *acoustic cameras* with such arrays is that they can focus on specific targets [14, 15], which is useful for speech processing. The resolution in elevation has recently been shown to be improved by using a 2.5D circular array [16]. While there have been a few results examining the use of spherical arrays, multiple spheres [17], randomly placed microphones [18, 19] and spiral configurations [20], there is little analysis of the impact of the geometric structure of the array on particular optimization algorithms for ASL.

We adopt an optimization (sequential least squares programming) approach for indirect ASL. We focus on localizing a single source, but other work towards estimating TDOA for multiple sources is directly applicable. Although the objective function we choose is non-linear and non-convex, we show using *simulation and real data* that the method is robust to noise and reverberation. Our experiments verify that it is comparable to SRP for real data while being $6\times$ more efficient to compute. Using this optimization scheme, we study the localization error resulting from different geometric structure for the microphone array. Our results show that circular arrays produce the highest errors (across space) and are therefore least desirable.

2. OBJECTIVE FUNCTION AND OPTIMIZATION

Consider a source at location $\mathbf{s}$ that emits an acoustic signal at some arbitrary time t^*. Let the measurements of the emitted sound be recorded by an array of M microphones located at $\mathbf{m}_i,\ i = 1, 2, ..., M$ and the times taken by the signal to travel from $\mathbf{s}$ to $\mathbf{m}_i$ be t_i. If the distance between the source and the



(a) Setup

signal	SRP		TDOA (100)		TDOA (all)	
	Rel. Err %	Time in min	Rel. Err %	Time in min	Rel. Err %	Time in min
chirp	14.7 (25.9)	3 (0.2)	14.2 (25.9)	0.5 (0.01)	12.1 (23.2)	4.5 (0.03)
gunshot	11.0 (13.3)	2.58 (0.2)	9.6 (12.8)	0.4 (0.02)	6.4 (3.5)	2.4 (0.02)
dogbark	16.0 (28.5)	2.49 (0.1)	58.9 (38.8)	0.4 (0.02)	48.5 (44.6)	2.4 (0.02)
speech	13.2 (21.1)	2.63 (0.1)	15.2 (23.5)	0.4 (0.02)	12.9 (22.5)	2.5 (0.02)

(b) Errors and computation time comparison across all microphone configurations

Fig. 1. (a) Our setup and coordinate system. (b) Table comparing errors and time for SRP with TDOA optimization using 100 of the C_2^{72} mic pairs (middle) and using all pairs. Standard deviations are shown within parantheses.

i^{th} microphone is $d_i \equiv \|\mathbf{m}_i - \mathbf{s}\|$, then $t_i = d_i/c + t^*$ where c is the speed of sound in air and t^* is not generally known.

Time of arrival In the case that the times of arrival at the microphones are measured as $\tilde{t}_i$, we pose the ASL problem as one of jointly determining $\mathbf{s}$ and t^* as

$$O_1: \quad \arg\min_{\mathbf{s},t^*} \sqrt{\sum_{i=1}^{M} (\tilde{t}_i - t_i)^2} \tag{1}$$

Time Difference of Arrival (TDOA) Another possibility is to note the difference in measured times between a pair of microphones, $\tilde{\tau}_{ij} \equiv \tilde{t}_i - \tilde{t}_j$, or TDOA. The literature is rich in methods to estimate TDOA. We choose the popular Generalized Cross-Correlation Phase Transform (GCC-PHAT) [21]. Then, we perform ASL by optimizing [7]:

$$O_2: \quad \arg\min_{\mathbf{s}} \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{M} \left(\tilde{\tau}_{ij} - \tau_{ij}\right)^2}, \tag{2}$$

where $\tau_{ij} = (t_i - t_j)$.

For both formulations O_1 and O_2, we know that the solution is constrained by the room dimensions, so we supply these constraints as linear inequalities. We solve the constrained non-linear optimization using Sequential Least Squares Programming (SLSQP) which is an interative procedure. In each iteration, a constrained quadratic programming sub-problem is built so that the chain of solutions converges to a local mininum [22]. Each subproblem replaces the objective function with a local, quadratic approximation subject to local affine approximations of the constraints. We used a Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximation to update the Hessian matrix required for the local quadratic approximation and chose the step length using an L_1 test function. The optimizer used to solve each subproblem is a modified version of NNLS [23]. We used the following parameters as inputs to the optimizer: iterations = 1500, accuracy = 1e-20, epsilon = 1.49e-08.

2.1. Experiments

We performed experiments using an *gfai tech AC_Pro Acoustic Camera system* consisting of 72 microphones sampled at 192kHz. We used three different microphone configurations: ring, wheel and spiral, spanning the same area. Using each configuration, we measured recorded sounds played by a *Bose Soundlink Bluetooth Mobile Speaker II, Model 404600* in five different calibrated positions within a room of size $12m \times 7m \times 3m$. The speaker was positioned, using a tripod, to be on the plane $y = -0.32$ for all five positions A, B, C, D and E. For each position we acquired three recordings. Fig. 1 illustrates the setup. We repeated the experiments for 4 different audio signals [24]: chirp, gunshot, dogbark and speech.

Simulation: noisy TOA and TDOA We tested the proposed optimization by evaluating the relative error in localization for different simulated degrees of noise σ in the estimated TOA and TDOA values. To enable comparison across multiple sources locations, we express σ for each source location as a percentage of the time taken for sound to travel from $\mathbf{s}$ to the center of the microphone array $\mathbf{O}$. We use a Gaussian model for the noise in simulated TOA $\tilde{t}_i = t_i + \eta$ and for TDOA $\tilde{\tau}_{ij} = \tau + \eta$ where

$$\eta \sim \mathcal{N}\left(0, \frac{\sigma}{100}\frac{\|\mathbf{s} - \mathbf{O}\|}{c}\right). \tag{3}$$

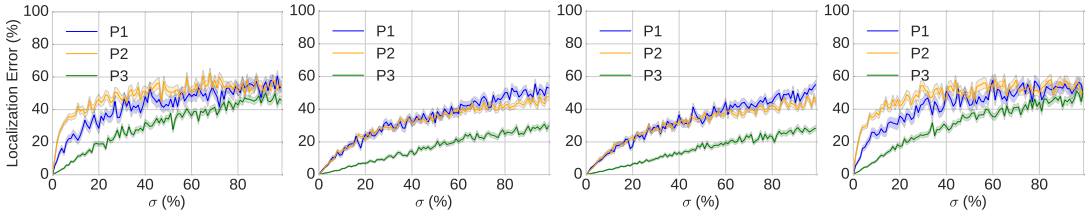
We measure relative error, expressed as a percentage of the distance from the source to the camera, as the evaluation metric for the accuracy of localization:

$$\text{error}(\%) = \frac{\|\mathbf{s} - \tilde{\mathbf{s}}\|}{\|\mathbf{s} - \mathbf{O}\|} * 100, \tag{4}$$

where $\tilde{\mathbf{s}}$ is the source location estimated by the optimization.

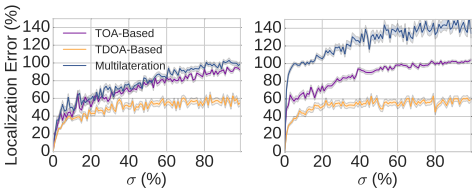
We compared optimizations for TOA and TDOA with multilateration [6]. Fig. 3 depicts plots of relative localization error (Y-axis) as the noise in the simulation is increased (X-axis). We performed two versions of the experiment: one assuming that the microphones and the sound source are synchronised ($t^* = 0$ in Fig. 3a), and one without that assumption by setting $t^* = 0.01s$.

Simulation: microphone configuration We estimated the localization error at different points in space, obtained via simulation. For each source position on a grid, we estimated the localization errors for three microphone configurations. The three configurations were identical to those used for real measurements with our acoustic camera, using 72 microphones. Each configuration results in different TOA and TDOA values, due to the different microphone positions. When noise is added to these TOA and TDOA values, each configuration reveals a characteristic heat-map for localization error over space. Fig. 4 visualizes these heatmaps for $\sigma = 100\%$ simulated error, along with the



(b) Ring (c) Wheel (b) Spiral (c) Random

Fig. 2. Relative localization error for increasing noise at three source locations: P1: (-2,-1,4), P2: (-1,0.5,3), P3: (0.4,0.7,1.05).



(a) synchronized ($t^* = 0$) (b) unknown t^* (10ms)

Fig. 3. Relative localization errors using O_1 (TOA), O_2 (TDOA) and multilateration [6] (a) speaker is synchronized with microphones and (b) time of emission is unknown.

corresponding error histograms. The errors were averaged over 100 trials for each grid point. We chose a grid over $x = [-2, 2]$, $z = [0, 4]$ and $y = -0.32$, with a resolution of 10 cm, so that it matches our experiments with real data. For three positions $P1 \equiv (-2, -1, 4)$, $P2 \equiv (-1, 0.5, 3)$ and $P3 \equiv (0.4, 0.7, 1.05)$, we plotted error as a function of noise for four different microphone configurations (Fig. 2).

Real Data: Comparison with SRP [4] We used optimization scheme O_2 to localize a speaker placed in five positions $A \equiv (2.0, -0.32, 0.5)$, $B \equiv (1.5, -0.32, 2.0)$, $C \equiv (0.0, -0.32, 1.5)$, $D \equiv (-1.5, -0.32, 1.0)$ and $E \equiv (-1.5, -0.32, 3.5)$. Fig. 5 plots relative errors (Y-axes) for three different microphone configurations (X-axes) at the chosen five locations (columns). The three rows of plots correspond to results obtained using SLSQP, SRP and Bayesian optimization [25] respectively. Errorbars (standard deviation) are shown with black lines on top of the bars.

2.2. Results and discussion

Microphone configurations Our results suggests that circular (ring configuration) arrays perform worse than spiral or wheel configurations when considering relative localization error over a wide range of positions. Our simulation results (Fig. 4) show regions (top view) that are error prone when using circular arrays. This is also true for our real measurements (Fig. 5), where the results obtained for position C are worse for ring than for wheel or spiral using any of the three localization techniques. The yellow bars in the first row show that the errors observed with real data correspond to errors obtained with about 10% noise in our simulation.

Comparison with multilateration Our experiments showed that both optimization strategies O_1 and O_2 result in lower relative errors than state of the art multilateration [6]. This is particularly true when the time of emission of the signal is unknown and when the emitter is not synchronized with the microphones ($t^* \neq 0$). When $t^* = 0$, our implementation of the multilateration algorithm has similar accuracy to optimizing O_1 (TOA). Our proposed approach to optimizing O_2 (TDOA) has the least relative errors and remains unaffected by t^*.

Comparison with SRP A common criticism of indirect methods is that the optimization is not as robust as direct methods such as SRP. However, our results (Table 1) show that our localization error is comparable to SRP but is more efficient. We used an efficient implementation of SRP that leverages stochastic region contraction [4] and a naïve implementation of our optimization in python. In both cases, the accuracy of the proposed optimization may also be traded for performance.

Accuracy vs performance One way to approximate the localization is to modify the nested summation in O_2 to consider only some of the microphone pairs. We studied convergence plots of localization error for different source positions, as the number of microphone pairs is increased from just 1 pair to all pairs (C_2^{72}). The error generally drops below 10% for 100 mic pairs (see Table 1 for computation times), except for the dogbark signal. Figure 6a plots relative error averaged across spatial locations for all four test signals using only 100 microphone pairs.

Bayesian optimization We tested a Bayesian optimizer with O_2 as its loss function ($\kappa = 1$). This took an order of magnitude longer than SQLSP and the resulting errors were larger. We tested with various degrees of the κ parameter to trade-off exploitation versus exploration. The plot (Fig. 6b) shows that exploitation ($\kappa = 1$) performs better than exploration ($\kappa = 10$) in most cases. The number of iterations and tolerance were set so that the optimizer converged to the reported solutions, suggesting that the problem is not due to multiple local minima.

Limitation One drawback of indirect localization achieved by minimizing O_2 is its dependency on the estimated TDOA values. Although our results show that GCC-PHAT is accu-

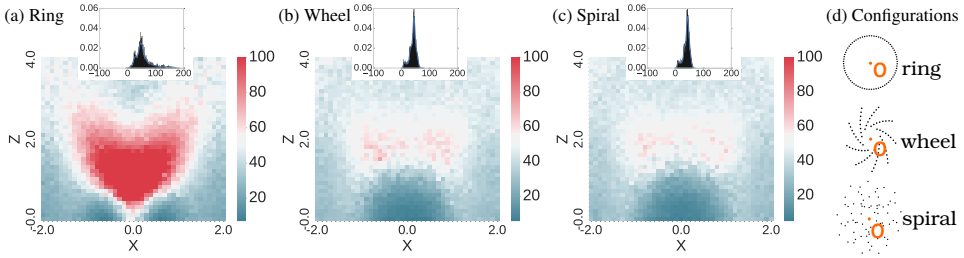
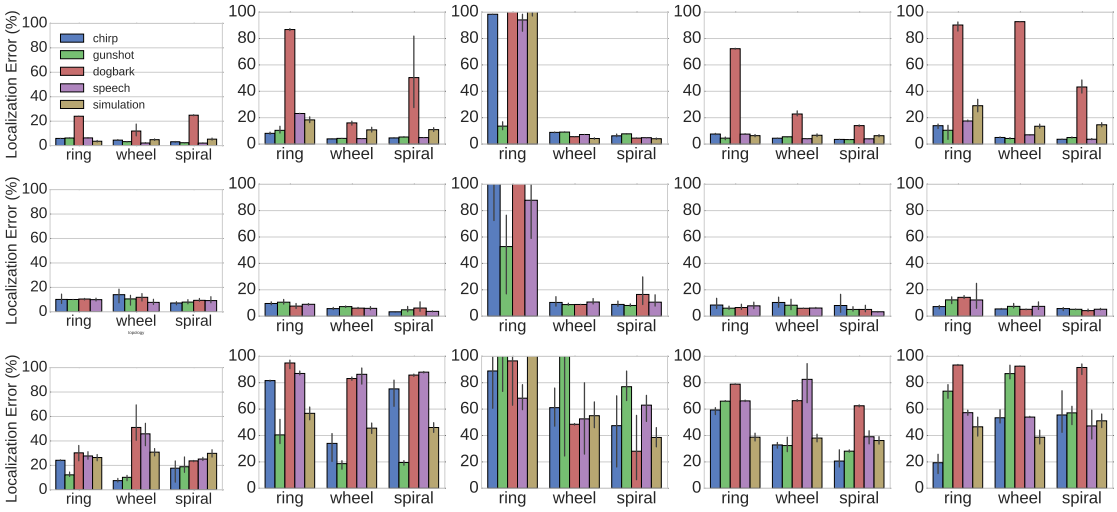


Fig. 4. Relative error percentages visualized as heatmaps obtained using simulations, at 100% noise, for a $2\mathrm{m} \times 2\mathrm{m}$ room. 100 estimates were averaged for the error estimate at each grid position. The insets show the distributions of errors as histograms.



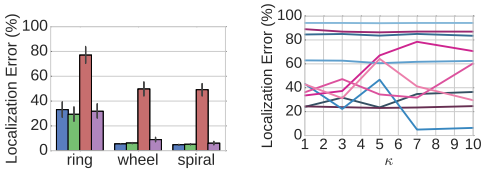
(a) A: (2.0,-0.32,0.5) (b) B: (1.5,-0.32,2.0) (c) C: (0.0,-0.32,1.5) (d) D: (-1.5,-0.32,1.0) (e) E: (-1.5,-0.32,3.5)

Fig. 5. Localization Error using SQLP and simulation (top row) SRP (2nd row) and Bayesian Optimization (3rd row)

rate enough to yield localization errors comparable to SRP, the former performs worse when dealing with signals with repeating patterns such as the barking of a dog (red bar in Fig. 5). Our localization was more robust to reverberation (when the source was placed at room boundaries) than to repetitive macro-structures. Perhaps using full signal correlation matrices, as adopted by spectral estimation techniques, would resolve this problem.

3. CONCLUSIONS

We have shown that direct optimization of the well known formulation for ASL yields error similar to the state of the art (SRP) with 6 times less computation. Moreover, we showed using both simulation and real data that the method is robust to noise and reverberation. Our results showed that circular arrays are least desirable configuration. In the future we plan to perform further experiments in a wide range of scenarios to generalize the ring arrays' performance limitations.



(a) Using 100 mic pairs (b) Bayesian Optimization

Fig. 6. (a) Errors (real data) for four signals across spatial locations. (b) Exploitation ($\kappa = 1$) vs exploration ($\kappa = 10$) for dogbark (blue) and speech (purple) for spiral configuration.

4. REFERENCES

[1] Ivan Marković and Ivan Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robotics and Autonomous Systems*, vol. 58, no. 11, pp. 1185–1196, 2010.

[2] Ryosuke Kojima, Osamu Sugiyama, and Kazuhiro Nakadai, "Scene understanding based on sound and text information for a cooking support robot," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2015, pp. 665–674.

[3] Markus VS Lima, Wallace A Martins, Leonardo O Nunes, Luiz WP Biscainho, Tadeu N Ferreira, Maurício VM Costa, and Bowon Lee, "A volumetric srp with refinement step for sound source localization," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1098–1102, 2015.

[4] Hoang Do, Harvey F Silverman, and Ying Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 1, pp. I–121.

[5] Yiteng Huang, Jacob Benesty, Gary W Elko, and Russell M Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[6] Orhan Oçal, Ivan Dokmanic, and Martin Vetterli, "Source localization and tracking in non-convex rooms," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Ieee, 2014, pp. 1429–1433.

[7] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, *Springer handbook of speech processing*, Springer Science & Business Media, 2007.

[8] Xiaomei Qu and Lihua Xie, "An efficient convex constrained weighted least squares source localization algorithm based on tdoa measurements," *Signal Process.*, vol. 119, no. C, pp. 142–152, Feb. 2016.

[9] K. Yang, J. An, X. Bu, and G. Sun, "Constrained total least-squares location algorithm using time-difference-of-arrival measurements," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 3, pp. 1558–1562, March 2010.

[10] Cao Jing-min, Wei He-wen, and Yu Jian, *Weighted Constrained Total Least-Square Algorithm for Source Localization Using TDOA Measurements*, p. 739746, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[11] Despoina Pavlidi, Matthieu Puigt, Anthony Griffin, and Athanasios Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2625–2628.

[12] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[13] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking.," in *MLMI*. Springer, 2004, pp. 182–195.

[14] Zebb Prime and Con Doolan, "A comparison of popular beamforming arrays," *Australian Acoustical Society AAS2013 Victor Harbor*, vol. 1, pp. 5, 2013.

[15] David Ayllón, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera, "An evolutionary algorithm to optimize the microphone array configuration for speech acquisition in vehicles," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 37–44, 2014.

[16] Mingsian R Bai, Chang-Sheng Lai, and Po-Chen Wu, "Localization and separation of acoustic sources by using a 2.5-dimensional circular microphone array," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 286–297, 2017.

[17] X. Pan, H. Wang, F. Wang, and C. Song, "Multiple spherical arrays design for acoustic source localization," in *2016 Sensor Signal Processing for Defence (SSPD)*, Sept 2016, pp. 1–5.

[18] Mohammad J Taghizadeh, Saeid Haghighatshoar, Afsaneh Asaei, Philip N Garner, and Hervé Bourlard, "Robust microphone placement for source localization from noisy distance measurements," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2579–2583.

[19] Roberto Macho-Pedroso, Francisco Domingo-Perez, Jose Velasco, Cristina Losada-Gutierrez, and Javier Macias-Guarasa, "Optimal microphone placement for indoor acoustic localization using evolutionary optimization," in *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*. IEEE, 2016, pp. 1–8.

[20] Chiong Lai, Sven Nordholm, and Yee-Hong Leung, "Design of robust steerable broadband beamformers with spiral arrays and the farrow filter structure," in *Proceedings of IWAENC 2010*. Ortra, 2010, vol. 90, pp. 653–669.

[21] Pasi Pertila, Matti S Hamalainen, and Mikael Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2393–2402, 2013.

[22] Philip E Gill and Elizabeth Wong, "Sequential quadratic programming methods," in *Mixed integer nonlinear programming*, pp. 147–224. Springer, 2012.

[23] Dieter Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[24] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[25] Jasper Snoek, Hugo Larochelle, and Ryan P Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

2018 26th European Signal Processing Conference (EUSIPCO)

A compressed encoding scheme for approximate TDOA estimation

Elizabeth Vargas*, James R. Hopgood†, Keith Brown* and Kartic Subr‡
*Institute of Sensors, Signals and Systems, Heriot-Watt University
†Institute for Digital Communications, University of Edinburgh
‡Institute of Perception, Action and Behaviour, University of Edinburgh
Edinburgh, United Kingdom
Email: ev42@hw.ac.uk, james.hopgood@ed.ac.uk, k.e.brown@hw.ac.uk, k.subr@ed.ac.uk

***Abstract*—Accurate estimation of Time-Difference of Arrivals (TDOAs) is necessary to perform accurate sound source localization. The problem has traditionally been solved by using methods such as Generalized Cross-Correlation, which uses the entire signal to accurately estimate TDOAs. However, this could pose a problem in distributed sensor networks in which the amount of data that can be transmitted from each sensor to a fusion center is limited, such as in underwater scenarios or other challenging environments. Inspired by approaches from computer vision, in this paper we identify Scale-Invariant Feature Transform (SIFT) keypoints in the signal spectrogram. We perform cross-correlation on the signal using only the information available at those extracted keypoints. We test our algorithm in scenarios featuring different noise and reverberation conditions, and using different speech signals and source locations. We show that our algorithm can estimate Time-Difference of Arrivals (TDOAs) and the source location within an acceptable error range at a compression ratio of $40:1$.**

***Index Terms*—microphone arrays, time difference estimation, signal compressed encoding**

I. Introduction

The literature on estimation of Time-Difference of Arrivals (TDOAs) is rich with a variety of approaches. One of the most common methods is Generalised Cross-Correlation (GCC), which is used to find the TDOA in a microphone array [1]. Methods based on cross-correlation are classified into two groups: ones that use a pair of microphones, and ones that draw on the redundancy among the microphones in the array. The first group includes the Smoothed Coherence Transform (SCOT) [2] and Generalized Cross-Correlation Phase-Transform (GCC-PHAT) [3] techniques, which are an extension of the cross-correlation into the frequency domain using a spectral normalization parameter. The second group of methods uses a spatial correlation matrix (MCCC) to determine the TDOA values that minimize the cross-correlation between each pair of signals. The most common of these methods is MUltiple SIgnal Classification (MUSIC) [4], which uses eigenvectors to estimate the TDOA.

Estimating TDOAs across a distributed sensor network is of increasing relevance as decentralised ad-hoc devices become more and more widespread. In such situations, the sensors need to exchange information to estimate the TDOAs. For example, to estimate TDOA using GCC would require transmission of the entire signal, or at the least a down-sampled version (which will lead to temporal quantisation). In scenarios in which the communications bandwidth is limited, or in which there are constraints on the amount of data that can be transmitted, approaches based on the full signal information are not very useful. Typical scenarios include underwater sensors [5], inexpensive ad-hoc mobile networks with energy constraints [6], and cases in which a high-speed communications network is either denied or unavailable (for example, disaster zones). Simon et al. [7] have developed an algorithm that relies on event detection of the signals in order to decide which parts of the signals to transmit. The authors transmit 1.1% of the raw signal, but they limited their experiments to a single scenario under specific noise and reverberation conditions. Similarly, Fuyong et al. [8] present a compression algorithm tested using compression ratios between $4:1$ and $8:1$. Additionally, there are authors who focus on sensor networks on low-bandwidth localization in [9], [10]; however, these are active sensing methods, in that sensors may emit calibration signals.

Previous studies have used different methods that involve feature extraction from the audio signals, including music identification [11], [12] and alignment of unsynchronized meeting recordings [13]. The most popular of these is known as audio fingerprinting [14], commonly used for music identification. It uses the signal spectrogram to select spectral peaks, provided that their power spectral amplitude is above a given threshold. These peaks are grouped into pairs to form a landmark, which is indexed using a hashing function. A set of these landmarks combines to characterize a song. Audio fingerprinting is used to perform *self-localization* in an ad-hoc microphone array in [15]. The problem in this instance is to localize sensors rather than sound sources, so the sources are placed in end-fire locations (i.e. points that lie on a straight line between two microphones, excluding the points that lie between the microphones) to guarantee a maximum TDOA.

In contrast to existing work that performs peak detection based on thresholding, we propose to detect audio landmarks using the Scale-Invariant Feature Transform (SIFT), a common approach in computer vision. Although there is evidence in the literature that authors have previously used SIFT on spectrograms [16]–[18], this is the first time to the best of our

ISBN 978-90-827970-1-5

knowledge that such an approach has been applied with a focus on data compression. In this paper, we present an approach based on estimating certain specific samples of the signal to be transmitted so as to estimate the delay using GCC. We use the SIFT algorithm to extract keypoints in the spectrogram, which is treated as an image.

Our main contributions in this paper are:

- Determining the signal keypoints to be transmitted to obtain an accurate TDOA estimation, at lower data rates or improved accuracy as against GCC solutions.
- Demonstrating the robustness of the proposed technique to different noise and reverberation conditions.
- Comparison of the proposed technique with another data-driven approach, namely audio fingerprinting.

II. Methodology

The proposed approach is based on Fig. 1, in which keypoint extraction occurs at the sensor-head. These keypoints are then communicated to a fusion center, which may either be a centralised node, or simply another sensor node. The communications channel is assumed to be low-bandwidth, such that minimal communication is desirable to ensure low-latency in the full localisation system. The sensors considered in this paper are microphones, but could naturally be any passive transducer, such as hydrophones, or RF.

The sensors $\mathbf{s}_i$ and $\mathbf{s}_j$ measure signals, m_i and m_j. The proposed algorithm for estimating TDOA, for that pair of microphones, consists of the following key steps:

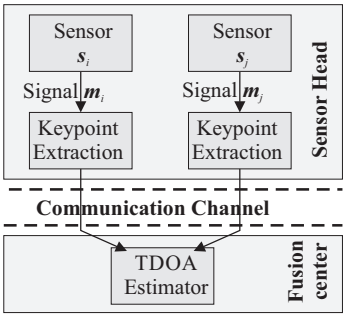


Fig. 1. Overview of the system architecture.

1) **At the Sensor-Head:** Calculate the spectrograms, $\widetilde{m}_i$ and $\widetilde{m}_j$ at each microphone, from the recorded signals m_i and m_j. The dimension of each spectrogram is F by T, where F is the number of rows corresponding to frequencies and T is the number of columns corresponding to time. We determined the optimum parameters for calculating the spectrogram were window size = 256, overlap = 204 and the final number of sampling points in the discrete Fourier transform = 1024;
2) **At the Sensor-Head:** Calculation of the Scale-Invariant Feature Transform (SIFT) [19] on the normalized spectrogram magnitude, in order to detect n keypoints from each spectrogram. We create a vector of keypoints, $\mathbf{f_i}$ and $\mathbf{t_i}$ for the *i-th* microphone. The k^{th} keypoint has coordinates (f_k, t_k), which corresponds to the time-frequency location at which the keypoints are detected. The values that will be transmitted are integers (encoded in 32 bits in order to keep high precision) and we only need to transmit the $t-$coordinates. It was found that adding in the frequency information did not improve the Time-Difference of Arrival (TDOA) relative error. Therefore, the total number of data samples that need to be transmitted to the fusion center is $n \times 32$. We experimented with the number of keypoints that need to be transmitted in order to obtain an acceptable margin of error in the Time-Difference of Arrival (TDOA). In light of this, we selected keypoints with the highest energy frequency coefficients, i.e. points that belong to rows of the spectrogram in which the sum of coefficients at key points is large. We selected k-rows each time, where k varies between 0.1 and 1;
3) **At the Fusion Center:** After the data is transmitted, two new vectors, $\widehat{\mathbf{m_i}}$ and $\widehat{\mathbf{m_j}}$, of the same size as m_i and m_j are created at the fusion center. We are assuming that all the sensors are synchronised and therefore started recording at the same instant. We can map keypoint locations to vectors by pre-calculating the times that correspond to the t-coordinates. The vector is filled with 1's in indices where a SIFT keypoint was detected and with 0's otherwise;

$$\widehat{\mathbf{m_i}}(l) = \begin{cases} 1 & \text{if } l \in \mathbf{t_i} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

4) **At the Fusion Center:** Calculation of Generalised Cross-Correlation (GCC) (defined by the $\star$ operator) between both vectors in the time domain. Since the cross-correlation is now on a binary vector, there is no need for the spectral normalisation as in PHAT.

$$\tau_{\text{delay}} = \arg\max_t((\widehat{\mathbf{m_i}} \star \widehat{\mathbf{m_j}})(t)) \tag{2}$$

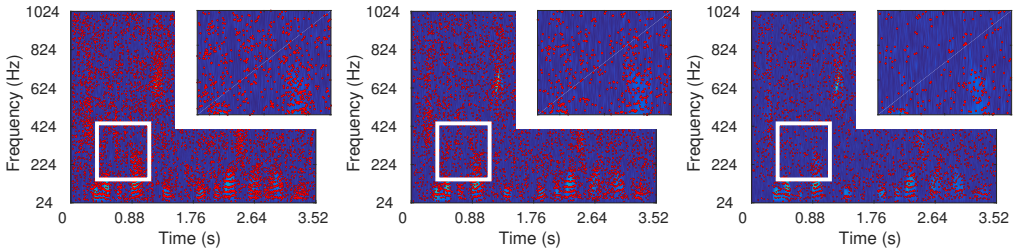
III. Experimental Results

We performed experiments using speech signals from the TIMIT database [20] and simulated environments by means of the image-source method [21]. We simulated two microphones in a linear array, separated by a distance of 4 metres and sampled at 16kHz. The simulated room has a size of 25m $\times$ 3m $\times$ 12m.

Since Time-Difference of Arrival (TDOA) is in the order of milliseconds for some source locations and centiseconds for others, it is necessary to standardize the error in order to make a fair comparison among source positions. Using the Ground Truth (GT), the relative error is computed using the TDOA estimation error in Equation 3. Similarly, we use the same principle to estimate the Direction of Arrival (DOA) relative error in Eq. 4.

$$\text{tdoa error}(\%) = \frac{\|\text{tdoa} - \text{gt}\|}{\|\text{gt}\|} * 100 \tag{3}$$

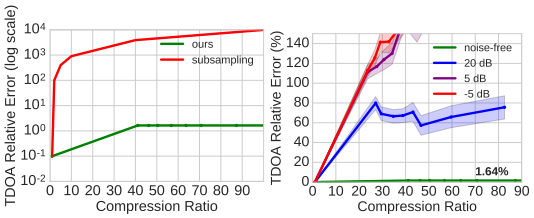
ISBN 978-90-827970-1-5



(a) $40:1$ compression (b) $55:1$ compression (c) $90{:}1$ compression

Fig. 2. SIFT keypoints (indicated in red) in the signal spectrogram, for different compression ratios. For each spectrogram, a patch (white rectangle) is selected and magnified at the upper right corner to provide a clearer visualization of the SIFT keypoints. This illustrates how the selected SIFT features are not necessarily spectrogram peaks and how our features differ from the peak picker approaches.

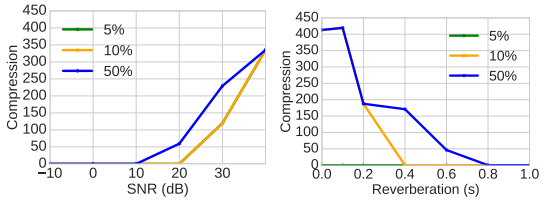
As previously mentioned, the compression ratio was varied in order to determine how much compression we can achieve while obtaining a reasonable TDOA relative error. We used the subsampling strategy presented in Sec. II, where we selected keypoints with the highest energy frequency coefficients, i.e points that belong to rows of the spectrogram in which the sum of coefficients at key points is large. Fig. 2 shows the spectrogram SIFT keypoints for different compression ratios. Fig. 3 illustrates the TDOA relative error with respect to compression ratio. In this experiment, the source was located at a DOA of 45°. Fig. 3a shows the error for an environment free of noise and reverberation using the proposed method and compares it with an approach in which compression is achieved by subsampling the signal. Since subsampling the signal increases the error dramatically even for low compression ratios, we decided to use a logarithmic scale on the Y-axis. Fig. 3b shows the relative error for a non-reverberant environment for different levels of noise. For a signal with SNR 20dB the TDOA error remains below 100%.



(a) Subsampling vs Ours (b) Various Noise Levels

Fig. 3. TDOA Relative Error achieved for different compression ratios for a source located at DOA 45°. The figure of the left shows the TDOA relative error for our algorithm compared with a baseline in which the signal is compressed by subsampling. We used the logarithmic scale on the Y-axis given that the error for the subsampling approach is much higher than our error. The right-hand side of the figure shows the TDOA Relative Error for a noise-free signal and for signals with various SNR values. To estimate the relative error for each compression ratio, we used 100 simulations.

Fig. 4 shows how noise and reverberation separately affect the compression ratio. We calculated the minimum value of compression that produced a TDOA relative error smaller



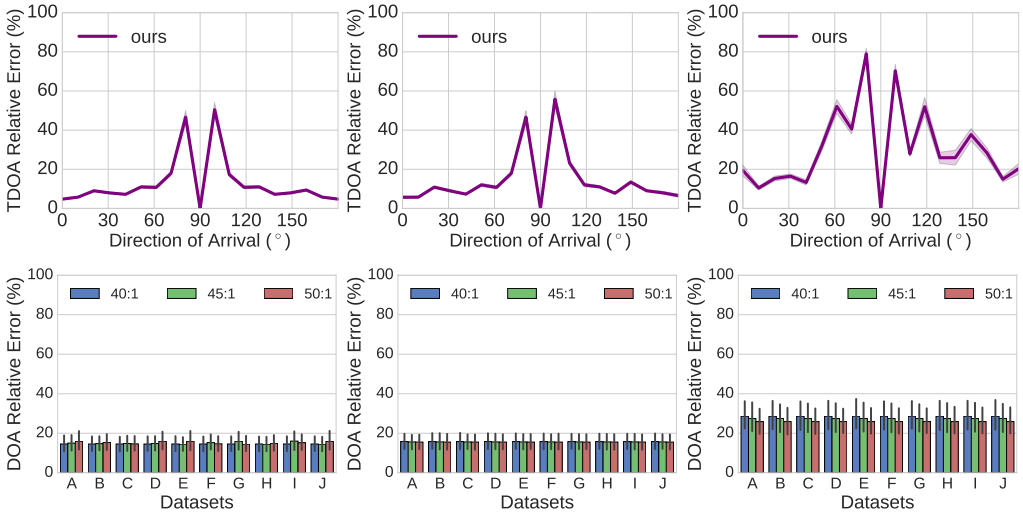
(a) Noise (b) Reverberation

Fig. 4. Maximum compression when the TDOA relative error $\leq 5\%, 10\%, 50\%$ for a source located at DOA $45°$ for different values of noise and reverberation. In 4(a), white Gaussian noise of -10 dB, 0 dB, 10 dB, 20 dB, 30 dB and 40 dB signal-to-noise ratio per sample was added to the original signal. For 5% and 10%, the compression ratios are identical, therefore we can only visualize a single line. In 4(b), we simulated reverberation values of $T_{60} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ seconds.

than $5\%, 10\%, 50\%$ for the given noise and reverberation conditions. In this scenario, the source is located at DOA $45°$. In Fig. 4(a), a white Gaussian noise of -10 dB, 0 dB, 10 dB, 20 dB, 30 dB and 40 dB signal-to-noise ratio per sample was added to the original signal. Note how the compression improves as the signal-to-noise ratio (SNR) gets higher. In the case, of 5% and 10%, the compression ratios are identical, therefore we can only visualize one line. We used T_{60} as a measurement of reverberation, interpreted as the time it takes a signal to drop by 60dB. In Fig. 4(b), reverberation values of $T_{60} = \{0.1k,\ k \in \{1, \ldots, 10\}\}$ seconds are simulated. In this case we can see that there is no compression value for which the error is smaller than 5%, however for 10% and 50% we achieved high compression ratios for low reverberation values (up to 0.6), after which the compression decreases to zero.

Fig. 5 illustrates the TDOA relative error and the DOA relative error for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds. We randomly selected 10 different sounds from the TIMIT dataset, which included speech signals from 5 men and 5 women (labeled A to J). We simulated 19 different source locations (DOA), from 0° to 180°, with a step size of 10°. We ran 5 different simulations for each of these sources and reverberation values. The first row of

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) reverberation $T_{60} = 0.1$ (b) reverberation $T_{60} = 0.2$ (c) reverberation $T_{60} = 0.3$

Fig. 5. TDOA relative error and the DOA relative error for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds. The results are from 10 speech signals (labelled A to J), at 19 different locations (DOA), from 0°to 180°, with a step size of 10°. We ran 5 different simulations for each of these sources and reverberation values. The first row shows the TDOA relative error for each DOA. The compression ratio is $40:1$ for each signal. The second row shows the DOA localization error per dataset for three different compression ratios: $40:1$, $45:1$ and $50:1$.

Fig. 5, shows the TDOA relative error for each DOA. The compression ratio is $40:1$ for each signal. It can be seen from the plots that for environments with low reverberation, $T_{60} = 0.1, 0.2$ seconds, the TDOA relative error is smaller than 20% for most DOA, except for 80°and 100°, in which case the error rises above 40%. The reason for this behavior is the small values of TDOA at such locations, which makes its calculation very challenging. The second row of Fig. 5 shows the DOA localization error. The x-axis presents 10 different datasets (labeled A to J). Three different compression ratios are used: $40:1$, 45:1 and 50:1. For low reverberation, $T_{60} = 0.1, 0.2$ seconds, the DOA relative error remains less than 20% for different compression ratios and sources. When reverberation $T_{60} = 0.3$ seconds, the TDOA relative error increases dramatically for most DOA, especially for 80°and 100°, in which case it is close to 80%. This large TDOA error has little impact on the DOA estimation, however. Even though the DOA relative error is above 20% in this case, the error in general remains less than 40%.

$$\text{doa error}(\%) = \frac{\|\text{doa} - \text{gt}\|}{\|\text{gt}\|} * 100 \tag{4}$$

IV. Discussion

We found that, by applying computer vision techniques, Scale-Invariant Feature Transform (SIFT), to the spectrogram of a speech signal, it is possible to detect keypoints that contain relevant information about the signal. We were able to use these keypoints to select the signal samples used to estimate Time-Difference of Arrival (TDOA) within a reasonable margin of relative error.

Our mechanism for improving the compression rate is to use subsampling of the SIFT keypoints in the spectrogram constructed at each sensor (microphone). Our strategy was to select the highest energy frequency coefficients, i.e rows of the spectogram in which the sum of coefficients at key points is large. This proved to be effective in scenarios in which there is little noise, as illustrated by Fig. 3b and Fig. 4a.

We ran our algorithm for various source locations and speech signals. We determined that the highest error in estimating the TDOA was caused in positions where the source was located in front of the microphone array, either at 80°or 100°. This happens because the TDOA is very small for these positions, which complicates the estimation. For 90°, where the TDOA is zero, and for 0°and 180°, where the separation is maximum, the relative error is closer to zero. On the other hand, given a similar position and the same noise and reverberation conditions, our algorithm performs very similarly across the test speech signals we used.

The algorithm's main drawback is its sensitivity to reverberation, as is evidenced in Fig. 4 and Fig. 5. This may be attributable to SIFT keypoints chosen from reverberations rather than from the original signal. One strategy to overcome this problem might be to estimate the probability of the keypoints being reverberations based on the amplitude of neighboring keypoints.

Table I shows a comparison between our algorithm and audio fingerprinting [15] for TDOA estimation. Both algo-

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I
FINGERPRINT VS OURS: TDOA RELATIVE ERROR

	Fingerprint Noise (SNR)			Ours Noise (SNR)		
Reverb	**noise-free**	**-5 dB**	**-10 dB**	**noise-free**	**-5 dB**	**-10 dB**
0	75.93 (43.80)	140.86 (71.35)	115.38 (55.48)	0.94 (0.68)	114.94 (38.62)	102.77 (12.05)
0.1	77.61 (50.29)	136.34 (69.54)	108.78 (59.71)	0.94 (0.68)	94.92 (29.44)	97.84 (20.87)
0.2	80.10 (40.72)	143.88 (69.69)	115.22 (60.60)	1.33 (0.69)	97.84 (14.91)	99.08 (8.68)
0.4	86.90 (43.88)	147.81 (80.15)	121.74 (63.04)	18.81 (35.14)	107.08 (13.68)	94.61 (12.13)
0.6	89.82 (92.69)	149.26 (76.36)	129.49 (65.35)	36.13 (26.12)	87.06 (31.63)	99.38 (20.24)

rithms were run on 10 different speech signals from the TIMIT database [20]. A value of noise (from 0 to -10 dB) and reverberation (from 0 to 0.6) was added to the signal. For each of these noise and reverberation values, the algorithm was executed 50 times. The table presents the mean TDOA relative error with the standard deviation (in brackets). In this scenario, the source was located at DOA 45°. We used the implementation of audio fingerprinting presented in [21], in which the input signal is subsampled to 8kHz to calculate the spectrogram. The number of sections is 64ms and the overlap is 32ms. We selected 50 landmarks per signal to perform our comparison. Table I shows how audio fingerprinting error is larger than ours for this particular source location and these particular speech signals.

V. CONCLUSIONS AND FUTURE WORK

In this work, we showed that, by applying a computer vision approach to the spectrogram of a speech signal, it was possible to identify samples of the signal allowing for an estimation of Time-Difference of Arrival (TDOA) within a reasonable margin of relative error. We tested the robustness of the proposed technique under different noise and reverberation conditions using different speech signals and source locations. We showed that our algorithm can estimate TDOA and the source location within an acceptable error range when the compression ratio of the signal is $40:1$.

In the future, we plan to modify our algorithm by improving on its robustness to noise and reverberation. We intend to do this by estimating the probability of keypoints representing reverberation or not depending on the amplitude of its neighbors. Moreover, we would like to perform experiments in open spaces in order to evaluate how the high reverberation values affect our algorithm.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[2] H.-s. Wang, J. Li, Z.-q. Sun, M.-h. Cao, and H.-w. Xie, "Accurate delay extraction for indoor pulse sound source location," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 298–301.

[3] P. Pertila, M. S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2393–2402, 2013.

[4] H.-K. Hao, H.-M. Liang, and Y.-W. Liu, "Particle methods for real-time sound source localization based on the multiple signal classification algorithm," in *Intelligent Green Building and Smart Grid (IGBSG), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[5] Y. G. Kim, K. M. Jeon, Y. Kim, C.-H. Choi, H. K. Kim, and L. Nex, "Underwater acoustic sensor array signal lossless compression based on valid channel decision approach," *Int J Image Signal Syst Eng*, vol. 1, no. 1, pp. 21–28, 2017.

[6] S. Zhou and L. Ying, "On delay constrained multicast capacity of large-scale mobile ad hoc networks," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5643–5655, 2015.

[7] G. Simon and L. Sujbert, "Acoustic source localization in sensor networks with low communication bandwidth," in *Intelligent Solutions in Embedded Systems, 2006 International Workshop on*. IEEE, 2006, pp. 1–9.

[8] Q. Fuyong, G. Fucheng, J. Wenli, and M. Xiangwei, "Data compression based on DFT for passive location in sensor networks," *Procedia Engineering*, vol. 29, pp. 3091–3095, 2012.

[9] D. O. Zion and H. Messer, "Envelope only tdoa estimation for sensor network self calibration," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 IEEE 8th*. IEEE, 2014, pp. 229–232.

[10] N. El Gemayel, H. Jakel, and F. K. Jondral, "Error analysis of a low cost tdoa sensor network," in *Position, Location and Navigation Symposium-PLANS 2014, 2014 IEEE/ION*. IEEE, 2014, pp. 1040–1045.

[11] R. Sonnleitner and G. Widmer, "Robust quad-based audio fingerprinting," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 409–421, 2016.

[12] S. Baluja and M. Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern recognition*, vol. 41, no. 11, pp. 3467–3480, 2008.

[13] T. Tsai and A. Stolcke, "Robust and efficient multiple alignment of unsynchronized meeting recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 833–845, 2016.

[14] A. Wang *et al.*, "An industrial strength audio search algorithm." in *Ismir*, vol. 2003. Washington, DC, 2003, pp. 7–13.

[15] T.-K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1623–1636, 2015.

[16] M. Zanoni, S. Lusardi, P. Bestagini, A. Canclini, A. Sarti, and S. Tubaro, "Efficient music identification approach based on local spectrogram image descriptors," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[17] X. Zhang, B. Zhu, L. Li, W. Li, X. Li, W. Wang, P. Lu, and W. Zhang, "Sift-based local spectrogram image descriptor: a novel feature for robust music identification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 6, 2015.

[18] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 597–604.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus, 1993," *Linguistic Data Consortium, Philadelphia*.

[21] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

ISBN 978-90-827970-1-5

Bibliography

[1] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[2] Orhan Oçal, Ivan Dokmanic, and Martin Vetterli, "Source localization and tracking in non-convex rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, IEEE, pp. 1429–1433.

[3] Nico Mentzer, Guillermo Payá-Vayá, Holger Blume, Nora von Egloffstein, and Werner Ritter, "Instruction-set extension for an asip-based sift feature extraction," in *International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)*, Samos, Greece, July 2014, IEEE, pp. 335–342.

[4] Soumitro Chakrabarty and Emanuël AP Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk Mountain House New Paltz, NY, United States, October 2017, IEEE, pp. 136–140.

[5] Soumitro Chakrabarty, "Single speaker localization," 2017.

[6] Arthur N Popper, Richard R Fay, and Arthur N Popper, *Sound source localization*, Springer, 2005.

[7] Eric Bezzam, Robin Scheibler, Juan Azcarreta, Hanjie Pan, Matthieu Simeoni, Rene Beuchat, Paul Hurley, Basile Bruneau, Corentin Ferry, and Sepand Kashani, "Hardware and software for reproducible research in audio array signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, United States, March 2017, IEEE, pp. 6591–6592.

[8] Russell Braunling, Randy M Jensen, and Michael A Gallo, "Acoustic target detection, tracking, classification, and location in a multiple-target environment," in *Peace and Wartime Applications and Technical Issues for Unattended Ground Sensors*. International Society for Optics and Photonics, 1997, vol. 3081, pp. 57–66.

[9] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[10] Christine Evers and Patrick A Naylor, "Acoustic slam," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[11] Enrique Coiras, Yvan Petillot, and David M Lane, "Multiresolution 3-d reconstruction from side-scan sonar images," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 382–390, 2007.

[12] gfai tech GmbH, "The Acoustic Camera," `https://www.acoustic-camera.com/en/support/frequently-asked-questions/knowledge-base/the-acoustic-camera-system.html`, 2019, [Online; accessed 13-July-2019].

[13] Yan-Chen Lu and Martin Cooke, "Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners," *Speech Communication*, vol. 53, no. 5, pp. 622–642, 2011.

[14] Chris Donahue, Julian McAuley, and Miller Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations (ICLR)*, New Orleans, LA, United States, May 2019.

[15] Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders, "Fundamentals of acoustics," *Fundamentals of Acoustics, 4th Edition, by Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, James V. Sanders, pp. 560. ISBN 0-471-84789-5. Wiley-VCH, December 1999.*, p. 560, 1999.

[16] Richard E Berg, David G Stork, and Brian Holmes, *The physics of sound*, Pearson, 3 edition, August 2004.

[17] Ivan Dokmanić, *Listening to Distances and Hearing Shapes: Inverse Problems in Room Acoustics and Beyond*, PhD Dissertation. Ecole Polytechnique Federale de Lausanne (EPFL), 1 edition, 2015.

[18] Allan D Pierce and Robert T Beyer, *Acoustics: An Introduction to Its Physical Principles and Applications.*, American Institute of Physics, December 1990.

[19] Y.H. Kim, *Sound Propagation: An Impedance Based Approach*, Wiley, 2010.

[20] David Halliday, Robert Resnick, and Jearl Walker, *Fundamentals of physics*, John Wiley & Sons, 2013.

[21] David A Bies, Colin Hansen, and Carl Howard, *Engineering noise control*, CRC press, 2017.

[22] David M Howard and Jamie Angus, *Acoustics and psychoacoustics*, Focal press, 2017.

[23] F Alton Everest and Ken Pohlmann, *Master handbook of acoustics*, McGraw-Hill Education, 5 edition, July 2009.

[24] Frank J Fahy, *Foundations of engineering acoustics*, Elsevier, 2000.

[25] Michael Valente, Holly Hosford-Dunn, and Ross J Roeser, *Audiology: treatment*, Thieme New York, NY, 2000.

[26] Heinrich Kuttruff, *Room acoustics*, Crc Press, 2014.

[27] Heinrich Kuttruff, *Acoustics: an introduction*, CRC Press, 2006.

[28] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[29] Michael Vorländer, *Auralization: fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*, Springer Science & Business Media, 2007.

[30] Eric A Lehmann and Anders M Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[31] Eugene Weinstein, Kenneth Steele, Anant Agarwal, and James Glass, "Loud: A 1020 node microphone array and acoustic beamformer," Tech. Rep., Courant Institute of Mathematical Sciences New York United States, 2007.

[32] Alexander M Haimovich, Rick S Blum, and Leonard J Cimini, "Mimo radar with widely separated antennas," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 116–129, 2008.

[33] Joseph H DiBiase, Harvey F Silverman, and Michael S Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, pp. 157–180. Springer, 2001.

[34] Futoshi Asano, Hideki Asoh, and Toshihiro Matsui, "Sound source localization and separation in near field," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 83, no. 11, pp. 2286–2294, 2000.

[35] Anders Johansson, Nedelko Grbic, and Sven Nordholm, "Speaker localisation using the far-field srp-phat in conference telephony," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Honolulu, HI, United States, November 2002.

[36] Anastasios S Lyrintzis, "Surface integral methods in computational aeroacoustics—from the (cfd) near-field to the (acoustic) far-field," *International Journal of Aeroacoustics*, vol. 2, no. 2, pp. 95–128, 2003.

[37] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, *Springer handbook of speech processing*, springer, 2007.

[38] Afsaneh Asaei, Hervé Bourlard, Mohammad J Taghizadeh, and Volkan Cevher, "Model-based sparse component analysis for reverberant speech localization," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, IEEE, pp. 1439–1443.

[39] Soumitro Chakrabarty and Emanuël AP Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[40] Caleb Rascon and Ivan Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, 2017.

[41] John C Murray, Harry Erwin, and Stefan Wermter, "Robotics sound-source localization and tracking using interaural time difference and cross-correlation," in *Proceedings of NeuroBotics Workshop*, Ulm, Germany, September 2004, pp. 89–97.

[42] Jean-Marc Valin, François Michaud, Jean Rouat, and Dominic Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, United States, October 2003, IEEE, vol. 2, pp. 1228–1233.

[43] Mordechai Azaria and David Hertz, "Time delay estimation by generalized cross correlation methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 280–285, 1984.

[44] Jacob Benesty, Jingdong Chen, and Yiteng Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 12, no. 5, pp. 509–519, 2004.

[45] Jan Scheuing and Bin Yang, "Disambiguation of tdoa estimates in multi-path multi-source environments (datemm).," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. 837–840.

[46] J Kuhn, "Detection performance of the smooth coherence transform (scot)," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tulsa, OK, United States, April 1978, IEEE, vol. 3, pp. 678–683.

[47] Byoungho Kwon, Youngjin Park, and Youn-sik Park, "Analysis of the gcc-phat technique for multiple sources," in *International Conference on Control Automation and Systems (ICCAS)*, Gyeonggi-do, Korea, October 2010, IEEE, pp. 2070–2073.

[48] JC VanDecar and RS Crosson, "Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares," *Bulletin of the Seismological Society of America*, vol. 80, no. 1, pp. 150–169, 1990.

[49] Kenichi Kumatani, John McDonough, Jill Fain Lehman, and Bhiksha Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, Scotland, May 2011, IEEE, pp. 1–6.

[50] Jingdong Chen, Yiteng Huang, and Jacob Benesty, "Time delay estimation via multichannel cross-correlation [audio signal processing applications]," in

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, United States, March 2005, IEEE, vol. 3, pp. iii–49.

[51] Yiteng Huang, Jacob Benesty, Gary W Elko, and Russell M Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[52] Amir Beck, Petre Stoica, and Jian Li, "Exact and approximate solutions of source localization problems," *IEEE Transactions on signal processing*, vol. 56, no. 5, pp. 1770–1778, 2008.

[53] Ka Wai Cheung, Hing-Cheung So, W-K Ma, and Yiu-Tong Chan, "Least squares algorithms for time-of-arrival-based mobile location," *IEEE Transactions on Signal Processing*, vol. 52, no. 4, pp. 1121–1130, 2004.

[54] Ralph Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[55] Fred K Gruber, Edwin A Marengo, and Anthony J Devaney, "Time-reversal imaging with multiple signal classification considering multiple scattering between the targets," *The Journal of the Acoustical Society of America*, vol. 115, no. 6, pp. 3042–3047, 2004.

[56] Marian-Daniel Iordache, José M Bioucas-Dias, Antonio Plaza, and Ben Somers, "Music-csr: Hyperspectral unmixing via multiple signal classification and collaborative sparse regression," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 4364–4382, 2014.

[57] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Mohonk Mountain House New Paltz, NY, United States, October 2007, IEEE, pp. 18–21.

[58] Markus VS Lima, Wallace A Martins, Leonardo O Nunes, Luiz WP Biscainho, Tadeu N Ferreira, Maurício VM Costa, and Bowon Lee, "A volumetric srp with refinement step for sound source localization," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1098–1102, 2015.

[59] Hoang Do, Harvey F Silverman, and Ying Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, HI, United States, April 2007, IEEE, vol. 1, pp. I–121.

[60] Mehdi Banitalebi Dehkordi, Hamid Reza Abutalebi, and Mohammad Reza Taban, "Sound source localization using compressive sensing-based feature extraction and spatial sparsity," *Digital Signal Processing*, vol. 23, no. 4, pp. 1239–1246, 2013.

[61] Wei Ke, Xiunan Zhang, Yanan Yuan, and Jianhua Shao, "Compressing sensing based source localization for controlled acoustic signals using distributed microphone arrays," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[62] Kun Yan, Hsiao-Chun Wu, Hailin Xiao, and Xiangli Zhang, "Novel measurement matrix optimization for source localization based on compressive sensing," in *2014 IEEE Global Communications Conference*. IEEE, 2014, pp. 341–345.

[63] Hisham Bedri, Micha Feigin, Petros Boufounos, and Ramesh Raskar, "Exploring the resolution limit for in-air synthetic-aperture audio imaging," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Santiago, Chile, December 2015, pp. 26–31.

[64] Xavier Alameda-Pineda and Radu Horaud, "Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August 2012, IEEE, pp. 1309–1313.

[65] Xavier Alameda-Pineda, Radu Horaud, and Bernard Mourrain, "The geometry of sound-source localization using non-coplanar microphone arrays," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House New Paltz, NY, United States, October 2013, IEEE, pp. 1–4.

[66] Xavier Alameda-Pineda and Radu Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1082–1095, 2014.

[67] Tsz-Kin Hon, Lin Wang, Joshua D Reiss, and Andrea Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1623–1636, 2015.

[68] Ryu Takeda and Kazunori Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016, IEEE, pp. 405–409.

[69] Ryu Takeda and Kazunori Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, United States, March 2017, IEEE, pp. 2217–2221.

[70] Eric L Ferguson, Stefan B Williams, and Craig T Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, April 2018, IEEE, pp. 2386–2390.

[71] Zhang-Meng Liu, Chenwei Zhang, and S Yu Philip, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7315–7327, 2018.

[72] Hoo Yu Heng, Jeeva Sathya Theesar Shanmugam, Madhavan al Balan Nair, and Ezra Morris Abraham Gnanamuthu, "Acoustic emission source localization on a pipeline using convolutional neural network," in *IEEE Conference on Big Data and Analytics (ICBDA)*, Langkawi Island, Malaysia, November 2018, IEEE, pp. 93–98.

[73] Soumitro Chakrabarty and Habets Emanuël AP, "Multi-scale aggregation of phase information for complexity reduction of cnn based doa estimation," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, September 2019, IEEE, pp. 1–5.

[74] Ivan Marković and Ivan Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robotics and Autonomous Systems*, vol. 58, no. 11, pp. 1185–1196, 2010.

[75] Ryosuke Kojima, Osamu Sugiyama, and Kazuhiro Nakadai, "Scene understanding based on sound and text information for a cooking support robot," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2015, pp. 665–674.

[76] Xiaomei Qu and Lihua Xie, "An efficient convex constrained weighted least squares source localization algorithm based on tdoa measurements," *Signal Process.*, vol. 119, no. C, pp. 142–152, Feb. 2016.

[77] K. Yang, J. An, X. Bu, and G. Sun, "Constrained total least-squares location algorithm using time-difference-of-arrival measurements," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 3, pp. 1558–1562, March 2010.

[78] Cao Jing-min, Wei He-wen, and Yu Jian, *Weighted Constrained Total Least-Square Algorithm for Source Localization Using TDOA Measurements*, p. 739–746, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[79] Despoina Pavlidi, Matthieu Puigt, Anthony Griffin, and Athanasios Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, IEEE, pp. 2625–2628.

[80] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[81] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez, "Av16. 3: an audio-visual corpus for speaker localization and tracking," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.

[82] Zebb Prime and Con Doolan, "A comparison of popular beamforming arrays," *Australian Acoustical Society (AAS)*, vol. 1, pp. 5, 2013.

[83] David Ayllón, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera, "An evolutionary algorithm to optimize the microphone array configuration for speech acquisition in vehicles," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 37–44, 2014.

[84] Mingsian R Bai, Chang-Sheng Lai, and Po-Chen Wu, "Localization and separation of acoustic sources by using a 2.5-dimensional circular microphone array," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 286–297, 2017.

[85] X. Pan, H. Wang, F. Wang, and C. Song, "Multiple spherical arrays design for acoustic source localization," in *2016 Sensor Signal Processing for Defence (SSPD)*, Edinburgh, United Kingdom, September 2016, pp. 1–5.

[86] Mohammad J Taghizadeh, Saeid Haghighatshoar, Afsaneh Asaei, Philip N Garner, and Hervé Bourlard, "Robust microphone placement for source localization from noisy distance measurements," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, IEEE, pp. 2579–2583.

[87] Roberto Macho-Pedroso, Francisco Domingo-Perez, Jose Velasco, Cristina Losada-Gutierrez, and Javier Macias-Guarasa, "Optimal microphone placement for indoor acoustic localization using evolutionary optimization," in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2016, pp. 1–8.

[88] Chiong Lai, Sven Nordholm, and Yee-Hong Leung, "Design of robust steerable broadband beamformers with spiral arrays and the farrow filter structure," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tel Aviv, Israel, August 2010, Ortra, vol. 90, pp. 653–669.

[89] Miloš Bjelić, Miodrag Stanojević, Dragana Šumarac Pavlović, and Miomir Mijić, "Microphone array geometry optimization for traffic noise analysis," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3101–3104, 2017.

[90] Lanxin Lin, Hing-Cheung So, Frankie KW Chan, Yiu-Tong Chan, and KC Ho, "A new constrained weighted least squares algorithm for tdoa-based localization," *Signal Processing*, vol. 93, no. 11, pp. 2872–2878, 2013.

[91] Steven M Kay, *Fundamentals of statistical signal processing*, Prentice Hall PTR, 1993.

[92] Maurizio Omologo and Piergiorgio Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), Adelaide, Australia, April 1994, IEEE, vol. 2, pp. II–273.

[93] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti, "Sensitivity-based region selection in the steered response power algorithm," *Signal Processing*, vol. 153, pp. 1–10, 2018.

[94] Mert Burkay Coteli, Orhun Olgun, and Huseyin Hacihabiboglu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2215–2229, 2018.

[95] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *The Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 586–601, 2017.

[96] Wei-Hsiang Liao, Yuki Mitsufuji, Keiichi Osako, and Kazunobu Ohkuri, "Microphone array geometry for two dimensional broadband sound field recording," in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[97] Zebb Prime, Con Doolan, and Branko Zajamsek, "Beamforming array optimisation and phase averaged sound source mapping on a model wind turbine," in *Inter-Noise and Noise-Con Congress and Conference Proceedings*, Melbourne, Australia, November 2014, Institute of Noise Control Engineering, vol. 249, pp. 1078–1086.

[98] Salil Luesutthiviboon, Anwar Malgoezar, Mirjam Snellen, Pieter Sijtsma, and Dick Simons, "Improving source discrimination performance by using an optimized acoustic array and adaptive high-resolution clean-sc beamforming," in *7th Berlin beamforming conference*, Berlin, Germany, March 2018, pp. 1–27.

[99] Elias Arcondoulis, Pengwei Xu, and Yu Liu, "An experimental verification of iterative microphone removal beamforming arrays," in *Proceedings of ACOUSTICS*, 2018, vol. 7.

[100] Elias Arcondoulis and Yu Liu, "Acoustic beamforming array design using an iterative microphone removal method," in *2018 AIAA/CEAS Aeroacoustics Conference*, Atlanta, GA, United States, June 2018, p. 2807.

[101] Elias Arcondoulis and Yu Liu, "An iterative microphone removal method for acoustic beamforming array design," *Journal of Sound and Vibration*, vol. 442, pp. 552–571, 2019.

[102] Elias JG Arcondoulis and Yu Liu, "Adaptive array reduction method for acoustic beamforming array designs," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. EL156–EL160, 2019.

[103] Ennes Sarradj, "Optimal planar microphone array arrangements," *Fortschritte der Akustik, DAGA 2015, Nürnberg, 41. Jahrestagung für Akustik*, 2015.

[104] Pasi Pertila, Matti S Hamalainen, and Mikael Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2393–2402, 2013.

[105] Philip E Gill and Elizabeth Wong, "Sequential quadratic programming methods," in *Mixed integer nonlinear programming*, pp. 147–224. Springer, 2012.

[106] Dieter Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[107] Jasper Snoek, Hugo Larochelle, and Ryan P Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, United States, December 2012, pp. 2951–2959.

[108] Eric Brochu, Vlad M Cora, and Nando De Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *arXiv preprint arXiv:1012.2599*, 2010.

[109] Fernando Nogueira, "Bayesianoptimization," 2019.

[110] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, United States, March 2017.

[111] Daobilige Su, Teresa Vidal-Calleja, and Jaime Valls Miro, "Simultaneous asynchronous microphone array calibration and sound source localisation," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany, September 2015, IEEE, pp. 5561–5567.

[112] Quan V Nguyen, Francis Colas, Emmanuel Vincent, and François Charpillet, "Long-term robot motion planning for active sound source localization with monte carlo tree search," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, United States, March 2017, IEEE, pp. 61–65.

[113] Xiaodong Du, Fengdan Lao, and Guanghui Teng, "A sound source localisation analytical method for monitoring the abnormal night vocalisations of poultry," *Sensors*, vol. 18, no. 9, pp. 2906, 2018.

[114] Ui-Hyun Kim, Jinsung Kim, Doik Kim, Hyogon Kim, and Bum-Jae You, "Speaker localization using the tdoa-based feature matrix for a humanoid robot," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Munich, Germany, August 2008, IEEE, pp. 610–615.

[115] François Grondin and François Michaud, "Noise mask for tdoa sound source localization of speech on mobile robots in noisy environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, May 2016, IEEE, pp. 4530–4535.

[116] Charles Blandin, Alexey Ozerov, and Emmanuel Vincent, "Multi-source tdoa estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[117] Gunnar Heilmann, Andy Meyer, and Dirk Döbler, "Time-domain beamforming using 3d-microphone arrays," *Berlin Beamforming Conference (BeBeC)*, March 2008.

[118] Hiroshi G Okuno and Kazuhiro Nakadai, "Robot audition: Its rise and perspectives," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, IEEE, pp. 5610–5614.

[119] Robin R Murphy, Satoshi Tadokoro, and Alexander Kleiner, "Disaster robotics," in *Springer Handbook of Robotics*, pp. 1577–1604. Springer, 2016.

[120] Tahmid Latif, Eric Whitmire, Tristan Novak, and Alper Bozkurt, "Sound localization sensors for search and rescue biobots," *IEEE Sensors Journal*, vol. 16, no. 10, pp. 3444–3453, 2016.

[121] Yong Guk Kim, Kwang Myung Jeon, Y Kim, Chang-Ho Choi, Hong Kook Kim, and LIG Nex, "Underwater acoustic sensor array signal lossless compression based on valid channel decision approach," *International Journal on Image Signal Systems Engineering*, vol. 1, no. 1, pp. 21–28, 2017.

[122] Lin Wang and Andrea Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, 2017.

[123] S Lana, KNKNH Takahashi, and T Kinoshita, "Consensus-based sound source localization using a swarm of micro-quadrocopters," *Robotics Society Japan*, pp. 1–4, 2015.

[124] Marco Compagnoni, Roberto Notari, Fabio Antonacci, and Augusto Sarti, "A comprehensive analysis of the geometry of tdoa maps in localization problems," *Inverse Problems*, vol. 30, no. 3, pp. 035004, 2014.

[125] Marco Compagnoni, Alessia Pini, Antonio Canclini, Paolo Bestagini, Fabio Antonacci, Stefano Tubaro, and Augusto Sarti, "A geometrical–statistical approach to outlier removal for tdoa measurements," *IEEE Transactions on Signal Processing*, vol. 65, no. 15, pp. 3960–3975, 2017.

[126] Avery Wang et al., "An industrial strength audio search algorithm.," in *International Society for Music Information Retrieval (ISMIR)*. Washington, DC, 2003, vol. 2003, pp. 7–13.

[127] SP Wang, H Sun, and P Yang, "Indoor sound-position fingerprint method based on scenario analysis," *Journal Beijing Univerty of Technology*, vol. 2, pp. 224–229, 2017.

[128] Shuopeng Wang, Peng Yang, and Hao Sun, "Fingerprinting acoustic localization indoor based on cluster analysis and iterative interpolation," *Applied Sciences*, vol. 8, no. 10, pp. 1862, 2018.

[129] TJ Tsai and Andreas Stolcke, "Robust and efficient multiple alignment of unsynchronized meeting recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 5, pp. 833–845, 2016.

[130] Lin Wang, Tsz-Kin Hon, Joshua D Reiss, and Andrea Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 6, pp. 1079–1093, 2016.

[131] Sylvain Argentieri, Patrick Danès, and Philippe Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[132] Xiang Wu, Dumidu S Talagala, Wen Zhang, and Thushara D Abhayapala, "Spatial feature learning for robust binaural sound source localization using a composite feature vector," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, March 2016, IEEE, pp. 6320–6324.

[133] Hongsen He, Jingdong Chen, Jacob Benesty, Yingyue Zhou, and Tao Yang, "Robust multichannel tdoa estimation for speaker localization using the impulsive characteristics of speech spectrum," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, United States, March 2017, IEEE, pp. 6130–6134.

[134] Mathieu Ramona and Geoffroy Peeters, "Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, IEEE, pp. 818–822.

[135] Joren Six and Marc Leman, "Panako: a scalable acoustic fingerprinting system handling time-scale and pitch modification," in *15th International Society for Music Information Retrieval Conference (ISMIR-2014)*, Taipei, Taiwan, October 2014.

[136] Reinhard Sonnleitner and Gerhard Widmer, "Robust quad-based audio fingerprinting," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 409–421, 2016.

[137] Chahid Ouali, Pierre Dumouchel, and Vishwa Gupta, "A robust audio fingerprinting method for content-based copy detection," in *International Workshop on Content-Based Multimedia Indexing (CBMI)*, Bordeaux, France, June 2014, IEEE, pp. 1–6.

[138] Mani Malekesmaeili and Rabab K Ward, "A local fingerprinting approach for audio copy detection," *Signal Processing*, vol. 98, pp. 308–321, 2014.

[139] Chahid Ouali, Pierre Dumouchel, and Vishwa Gupta, "Fast audio fingerprinting system using gpu and a clustering-based technique," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 6, pp. 1106–1118, 2016.

[140] Shumeet Baluja and Michele Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern recognition*, vol. 41, no. 11, pp. 3467–3480, 2008.

[141] Yan Ke, Derek Hoiem, and Rahul Sukthankar, "Computer vision for music identification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, United States, June 2005, IEEE, vol. 1, pp. 597–604.

[142] Tomoko Matsui, Masataka Goto, Jean-Philippe Vert, and Yuji Uchiyama, "Gradient-based musical feature extraction based on scale-invariant feature transform," in *European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, September 2011, IEEE, pp. 724–728.

[143] Xiu Zhang, Bilei Zhu, Linwei Li, Wei Li, Xiaoqiang Li, Wei Wang, Peizhong Lu, and Wenqiang Zhang, "Sift-based local spectrogram image descriptor: a novel feature for robust music identification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 6, 2015.

[144] Quang Trung Nguyen et al., "Speech classification using sift features on spectrogram images," *Vietnam Journal of Computer Science*, vol. 3, no. 4, pp. 247–257, 2016.

[145] Mark L Fowler and Mo Chen, "Fisher-information-based data compression for estimation using two sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 3, pp. 1131–1137, 2005.

[146] Fowler Mark L Pourhomayoun, Mohammed, "Data compression for complex ambiguity function for emitter location," International Society for Optics and Photonics, 2010, vol. 7799, p. 77990E.

[147] Mohammed Pourhomayoun and Mark L Fowler, "Exploiting cross ambiguity function properties for data compression in emitter location systems," Baltimore, MD, United States, June 2011, IEEE, pp. 1–5.

[148] Mohammed Pourhomayoun and Mark Fowler, "An svd approach for data compression in emitter location systems," Pacific Grove, CA, United States, November 2011, IEEE, pp. 257–261.

[149] Gyula Simon and László Sujbert, "Acoustic source localization in sensor networks with low communication bandwidth," in *International Workshop on Intelligent Solutions in Embedded Systems*, Vienna, Austria, June 2006, IEEE, pp. 1–9.

[150] Qu Fuyong, Guo Fucheng, Jiang Wenli, and Meng Xiangwei, "Data compression based on dft for passive location in sensor networks," *Procedia Engineering*, vol. 29, pp. 3091–3095, 2012.

[151] Dan Ohev Zion and Hagit Messer, "Envelope only tdoa estimation for sensor network self calibration," in *Sensor Array and Multichannel Signal Processing Workshop (SAM)*, A Coruna, Spain, September 2014, IEEE, pp. 229–232.

[152] Noha El Gemayel, Holger Jakel, and Friedrich K Jondral, "Error analysis of a low cost tdoa sensor network," in *IEEE/ION Position, Location and Navigation Symposium*, Monterey, CA, United States, May 2014, IEEE, pp. 1040–1045.

[153] Zhiwei Zhuang, Yi Zhan, and Jian Qian, "Data compression technology of distributed cooperative passive location," *Hans Journal of Wireless Communications*, 2014.

[154] David G Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision (ICCV)*, Corfu, Greece, September 1999, IEEE, vol. 2, pp. 1150–1157.

[155] Paul J Besl and Neil D McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*. International Society for Optics and Photonics, 1992, vol. 1611, pp. 586–606.

[156] John S Garofolo, "Ld consortium et al," *TIMIT: acoustic-phonetic continuous speech corpus. Linguistic Data Consortium*, 1993.

[157] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, United States, December 2012, pp. 1097–1105.

[158] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational IntelligenCe Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[159] Thomas Padois, Olivier Doutres, Franck Sgard, and Alain Berry, "Time domain source localization technique based on generalized cross correlation and generalized mean," *Canadian Acoustics*, vol. 44, no. 3, 2016.

[160] Elizabeth Vargas, James R Hopgood, Keith Brown, and Kartic Subr, "A compressed encoding scheme for approximate tdoa estimation," in *European Signal Processing Conference (EUSIPCO)*, Rome, Italy, September 2018, IEEE, pp. 346–350.

[161] Qinglong Li, Xueliang Zhang, and Hao Li, "Online direction of arrival estimation based on deep learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, April 2018, IEEE, pp. 2616–2620.

[162] Pasi Pertilä and Emre Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, United States, March 2017, IEEE, pp. 6125–6129.

[163] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, "Robust tdoa estimation based on time-frequency masking and deep neural networks," in *Annual Conference of the International Speech Communication Association (Interspeech)*, Hyderabad, India, September 2018, pp. 322–326.

[164] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 1, pp. 178–188, 2019.

[165] Disong Wang and Yuexian Zou, "Joint noise and reverberation adaptive learning for robust speaker DOA estimation with an acoustic vector sensor," in *Annual Conference of the International Speech Communication Association (Interspeech)*, Hyderabad, India, September 2018, pp. 821–825.

[166] Soumitro Chakrabarty and Emanuël AP Habets, "Multi-speaker localization using convolutional neural network trained with noise," *Conference on Neural Information Processing Systems (NIPS) Workshops*, December 2017.

[167] Matthias Frank, Franz Zotter, and Alois Sontacchi, "Producing 3d audio in ambisonics," in *Audio Engineering Society Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*. Audio Engineering Society, 2015.

[168] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin, "Crnn-based multiple doa estimation using ambisonics acoustic intensity features," 2018.

[169] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin, "Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, September 2018, IEEE, pp. 241–245.

[170] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *European Signal Processing Conference (EUSIPCO)*, Rome, Italy, September 2018, IEEE, pp. 1462–1466.

[171] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[172] Danilo Comminiello, Marco Lella, Simone Scardapane, and Aurelio Uncini, "Quaternion convolutional neural networks for detection and localization of 3d sound events," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019, IEEE, pp. 8533–8537.

[173] Constantinos Papayiannis, Christine Evers, and Patrick A Naylor, "Data augmentation of room classifiers using generative adversarial networks," *arXiv preprint arXiv:1901.03257*, 2019.

[174] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[175] Jordi Pons, Joan Serrà, and Xavier Serra, "Training neural audio classifiers with few data," in *IEEE International Conference on Acoustics, Speech and*

Signal Processing (ICASSP), Brighton, United Kingdom, May 2019, IEEE, pp. 16–20.

[176] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," *Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 312–316, September 2018.

[177] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019.

[178] Mariam Yiwere and Eun Joo Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12384–12389, 2017.

[179] Juan Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, pp. 3418, 2018.

[180] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.

[181] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[182] John S Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[183] Urmila Shrawankar and Vilas Thakare, "Noise estimation and noise removal techniques for speech recognition in adverse environment," in *International Conference on Intelligent Information Processing*, Manchester, United Kingdom, October 2010, Springer, pp. 336–342.

[184] RG Bachu, S Kopparthi, B Adapa, and Buket D Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*, pp. 279–282. Springer, 2010.

[185] Christine Evers and James R Hopgood, "Parametric modelling for single-channel blind dereverberation of speech from a moving speaker," *IET Signal Processing*, vol. 2, no. 2, pp. 59–74, 2008.

[186] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, December 2014, pp. 2672–2680.

[187] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[188] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, United States, December 2017, pp. 5767–5777.

[189] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia*, Glasgow, United Kingdom, April 2014, ACM, pp. 1041–1044.

[190] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot, "Multichannel audio database in various acoustic environments," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, September 2014, IEEE, pp. 313–317.