Par **Camille LEBLANC**

Véhicule autonome aérien pour la localisation de sources sonores

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 4 décembre 2025
Spécialité : Acoustique et Vibrations – Automatique et Robotique : Laboratoire Roberval
D2886

UNIVERSITÉ de TECHNOLOGIE de COMPIÈGNE (UTC)

École Doctorale 71 : Sciences pour l'ingénieur

Thèse présentée pour l'obtention du grade de Docteur de l'UTC

Véhicule autonome aérien pour la localisation de sources sonores

Par **Camille LEBLANC**

Spécialité : Acoustique et Vibrations - Automatique et Robotique

Laboratoire Roberval
Unité de recherche en Mécanique, Énergie et Électricité

Thèse soutenue le 4 décembre 2025 devant le jury composé de :

M. Jérôme Antoni (Président)
Professeur des Universités, Laboratoire LVA, INSA Lyon, Villeurbanne
M. Jean-Daniel Chazot (Directeur de thèse)
Professeur des Universités, Laboratoire Roberval, Université de Technologie de Compiègne
M. Régis Marchiano (Directeur de thèse)
Professeur des Universités, Institut Jean Le Rond d'Alembert, Sorbonne Université
M. Sylvain Argentieri (Rapporteur)
Maître de Conférences, Institut des Systèmes Intelligents et de la Robotique (ISIR), Sorbonne Université
M. Vincent Valeau (Rapporteur)
Professeur des Universités, Institut Pprime, Université de Poitiers
Mme Reine Talj (Examinatrice)
Directrice de recherche CNRS, Laboratoire Heudiasyc, Université de Technologie de Compiègne

Invités **:**

M. Pedro Castillo
Directeur de recherche CNRS, Laboratoire Heudiasyc, Université de Technologie de Compiègne
M. Thibaut Gras
Ingénieur de recherche, Cente Technique des Industries Mécaniques (CETIM)

UNIVERSITÉ de TECHNOLOGIE de COMPIÈGNE (UTC)

École Doctorale 71 : Sciences pour l'ingénieur

Thèse présentée pour l'obtention du grade de Docteur de l'UTC

Autonomous unmanned aerial vehicle embedded sound source localization

By **Camille LEBLANC**

Spécialité : Acoustique et Vibrations - Automatique et Robotique

Laboratoire Roberval
Unité de recherche en Mécanique, Énergie et Électricité

Thesis defended on 4 December 2025

M. Jérôme Antoni (Président)
Professeur des Universités, Laboratoire LVA, INSA Lyon, Villeurbanne
M. Jean-Daniel Chazot (Directeur de thèse)
Professeur des Universités, Laboratoire Roberval, Université de Technologie de Compiègne
M. Regis Marchiano (Directeur de thèse)
Professeur des Universités, Institut Jean Le Rond d'Alembert, Sorbonne Université
M. Sylvain Argentieri (Rapporteur)
Maître de Conférences, Institut des Systèmes Intelligents et de la Robotique (ISIR), Sorbonne Université
M. Vincent Valeau (Rapporteur)
Professeur des Universités, Institut Pprime, Université de Poitiers
Mme Reine Talj (Examinatrice)
Directrice de recherche CNRS, Laboratoire Heudiasyc, Université de Technologie de Compiègne

Invités **:**

M. Pedro Castillo
Directeur de recherche CNRS, Laboratoire Heudiasyc, Université de Technologie de Compiègne
M. Thibaut Gras
Ingénieur de recherche, Centre Technique des Industries Mécaniques (CETIM)

Here, the straight path isn't the shortest.

Stalker (dir. Andrei Tarkovsky)

Remerciements

Ce mémoire est l'accomplissement d'un long travail, certes personnel, mais avant tout collectif. Que ce soit par vos retours constructifs et bienveillants, en me montrant le chemin, ou en partageant un kebab après une journée de dur labeur, vous avez tous joué un rôle dans la finalisation de ce projet, et je me dois de vous adresser mes remerciements sincères.

Mes premiers remerciements vont naturellement aux membres du jury qui ont accepté de rapporter ou d'examiner ce manuscrit. Merci à Sylvain Argentieri, Vincent Valeau, Jérôme Antoni et Reine Talj, pour le temps que vous avez consacré à l'évaluation de mon travail, et pour la richesse de vos retours.

Je tiens également à exprimer ma gratitude envers Jean-Daniel, Pedro et Régis, qui m'ont ouvert les portes du monde de la recherche et m'ont offert leur confiance tout au long de ces années.

Un grand merci à tous les doctorants et post-doctorants de l'équipe Acoustique du Laboratoire Roberval. Matthieu, Apolline, Thomas, Rita, les Alexandres, Najib, et Christophe : ça aurait été bien moins drôle sans vous ! Je suis en particulier reconnaissant à Simon, Grégoire et Aristide pour avoir toujours considéré mes problèmes comme les leurs, et pour le temps qu'ils ont passé à débloquer ce qui l'était.

Du côté du CETIM, je tiens à saluer Alexandre, Anaïs, Vincent, Bastien et Michel pour leur accueil chaleureux et leur aide précieuse.

Merci également aux doctorants du Laboratoire Heudiasyc, et tout particulièrement à Armando et Alejandro, sans qui les expériences présentées dans ce mémoire n'auraient pas vu le jour.

Pour ce qui est des résultats expérimentaux, leur crédit revient aussi en grande partie à l'expertise et au soutien de Guillaume, Thierry, Stéphane, Antoine et Benjamin. Merci pour les années de travail gagnées !

Enfin, un immense merci à Thibaut, pour son écoute permanente et pour m'avoir appris à concilier les différents aspects de la vie de thésard.

Merci à ma mère, à Virgile, à Amel, à Alma, d'avoir toujours cru en moi et de m'avoir poussé à saisir ma chance dans ce milieu.

Et enfin, merci à Siyue : j'espère être pour ta thèse un aussi bon partenaire que tu l'as été pour moi.

Abstract

The ability to locate sound sources is crucial for many applications, particularly industrial acoustic inspection. The objective is generally to cover a large and potentially dangerous area using microphone arrays and to locate emerging sound sources, whether they be gas leaks, partial electrical discharges, or machine malfunctions. In this context, mobile robots, and in particular unmanned aerial vehicles, offer unique advantages due to their mobility and access to hard-to-reach areas. This thesis focuses on the development and implementation of an experimental platform for autonomous sound source localization using a 32-microphone array initially mounted on a ground vehicle and then on an aerial vehicle. The systems operate in real time and are guided by the direction of arrival of sound waves from the source, estimated by acoustic imaging. In the case of the flying drone, the noise it generates is processed using a denoising algorithm based on a set of pre-recorded drone noise data. The system takes advantage of the circular motion obtained through the control algorithm to accurately locate the source using triangulation from estimates of the direction of arrival. The systems are tested in an indoor environment and locate stationary sound sources on the ground without prior knowledge of their position. The results obtained are satisfactory and rapid, and each vehicle corresponds to a different application scenario. The ground vehicle is suitable for cases where the sources are low-level and located close to the ground in a navigable environment. The drone-based solution, on the other hand, is more appropriate if the diagnosis needs to be as rapid as possible and the terrain is not accessible to a ground vehicle.

Résumé

La capacité à localiser les sources sonores est cruciale pour de nombreuses applications, notamment pour l'inspection acoustique industrielle. L'objectif est généralement de couvrir une zone étendue et potentiellement dangereuse à l'aide de réseaux de microphones, et de retrouver la position de sources sonores émergentes, qu'il s'agisse de fuites de gaz, de décharges électriques partielles, ou de défaut de fonctionnement sur des machines. Dans ce contexte, les robots mobiles, et en particulier les véhicules aériens sans pilote, offrent des avantages uniques en raison de leur mobilité et de leur accès à des zones difficiles à atteindre. Cette thèse se concentre sur le développement et la mise en place d'une plateforme expérimentale pour la localisation autonome de sources sonores en utilisant une antenne de 32 microphones embarquée dans un premier temps sur un véhicule terrestre, puis sur un véhicule aérien. Les systèmes fonctionnent en temps réel et sont guidés par la direction d'arrivée des ondes sonores de la source, estimée par imagerie acoustique. Dans le cas du drone volant, le bruit qu'il génère est traité à l'aide d'un algorithme de débruitage basé sur un ensemble de données préenregistrées du bruit du drone. Le système tire parti du mouvement circulaire obtenu grâce à l'algorithme de contrôle pour localiser la source avec précision en utilisant la triangulation à partir des estimations de la direction d'arrivée. Les systèmes sont testés dans un environnement intérieur et localisent des sources sonores stationnaires au sol sans connaissance préalable de leur position. Les résultats obtenus sont satisfaisants et rapides, et chaque véhicule correspond à un scénario d'application différent. Le véhicule terrestre s'avère approprié dans le cas où les sources sont de faible niveau sonore et qu'elles se situent près du sol, dans un environnement navigable. La solution embarquée sur drone, quant à elle, est plus appropriée si le diagnostic doit être le plus rapide possible et que le terrain n'est pas praticable pour un véhicule terrestre.

Abbreviations

A-SSL	Autonomous Sound Source Localization
ACF	Autocorrelation Function
API	Application Programming Interface
BPF	Blade Passing Frequency
C-CBF, C-MUSIC	Constrained CBF, Constrained MUSIC
CBF	Conventional BeamForming
CCW	CounterClockWise
CMF	Center of Mass
CSM	Cross-Spectral Matrix
CW	Clockwise
D&S	Delay-and-Sum
DA	Drone Audition
DOA	Direction of Arrival
FFT	Fast Fourier Transform
GEVD	Generalized Eigenvalue Decomposition
GPS	Global Positioning System
IMU	Inertial Measurement Unit
MEMS	Micro-Electro-Mechanical Systems
MUSIC	MUltiple SIgnal Classification
MWF	Multichannel Wiener Filter
PID	Proportional Integral Derivative
PSF	Point Spread Function
RMS	Root Mean Square
ROS	Robot Operating System
SIR	Signal-to-Interference Ratio
SNR	Signal-to-noise ratio
SOI	Source Of Interest
SPL	Sound Pressure Level
SSL	Sound Source Localization
TBL	Turbulent Boundary Layer
TCP	Transmission Control Protocol
UAL	Uniform Linear Array
UAV	Unmanned Aerial Vehicle
VRPN	Virtual-Reality Peripheral Network

Symbols

i	Imaginary unit
e	Euler's number
$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
$\|\|\|$	Euclidean norm
$\|\|\|_F$	Frobenius norm
^{T}	Transpose operator
^{H}	Hermitian transpose operator
$^\perp$	Orthogonal complement operator
tr	Trace operator
$\mathcal{I}_M$	Identity matrix of size $M \times M$
$\dot{}$	Time derivative
$\ddot{}$	Second time derivative
Δ	Laplacian operator
$\mathcal{L}$	Lagrangian operator
$\mathbb{E}$	Statistical expectation operator
μ	Mean of distribution
σ	Standard deviation of distribution
p	Acoustic pressure
t	Time
f	Frequency
c	Celerity of sound
k	Wave number
λ	Wavelength
$\vec{k}$	Wave vector
r	Radius coordinate
θ	Azimuth angle
ϕ	Coelevation angle
X	Position vector
$\mathbf{X}$	Positions matrix
X_s	Source position vector
$\hat{X}_s$	Source position estimate
$\hat{\theta}$	Source direction estimate
g	Green's function
V_s	Elementary sources' volume
s	Monopole source amplitude
M	Number of microphones
Q	Number of sound sources
R_F	Fraunhofer distance
D	Microphone array aperture
Γ	Microphone array directivity
τ_m	Time delay for microphone m
$Y_{\text{D\&S}}$	Delay-and-Sum beamformer output
$Y_{\text{FD\&S}}$	Frequency delay-and-Sum beamformer output
Y_{CBF}	Conventional beamformer output
Y_{MUSIC}	MUSIC beamformer output

Ω	Search domain
n_ϕ	Number of angular bins in the search space
δ_ϕ	Angular resolution
ϵ_φ	Phase error
θ_c	Interfering source direction
I	Number of interfering sources
ϵ_θ	Angular error
SPL_P	SPL measured at the robot position
SPL_{1m}	SPL at 1 m
S_i	Sound source indexed i
ϵ	Localization error
T_{20dB}	Time to reach a 20 dB decay
A_S	RMS amplitude of source signal
A_N	RMS amplitude of noise signal
$\mathbf{p}$	Microphone pressure vector
$\mathbf{s}$	Source amplitude vector
$\mathbf{n}$	Additive noise vector
$\mathbf{G}$	Propagation matrix
$\mathbf{G}_s$	Sources-to-microphones transfer function
$\mathbf{C}$	Theoretical cross-spectral matrix
$\mathbf{\Sigma}$	Sources cross-spectral matrix
$\mathbf{N}$	Noise cross-spectral matrix
n_m^2	Microphone m noise power
L	Number of snapshots for cross-spectral estimation
$\hat{\mathbf{C}}$	Estimated cross-spectral matrix
Δf	Frequency resolution
T_{acq}	Acquisition time
f_s	Sampling frequency
N_f	Number of samples
n_f	Number of frequency bins
ρ	Overlap factor
$\mathbf{U}_{\mathrm{signal}}$	Signal subspace matrix
$\mathbf{U}_{\mathrm{noise}}$	Noise subspace matrix
$\mathbf{\Lambda}$	Eigenvalues diagonal matrix
$\mathbf{p}_s$	Source pressure vector
$\mathbf{p}_c$	Correlated noise pressure vector
$\mathbf{p}_u$	Uncorrelated noise pressure vector
$\mathbf{C}_s$	Source cross-spectral matrix
$\mathbf{C}_c$	Correlated noise cross-spectral matrix
$\mathbf{C}_u$	Uncorrelated noise cross-spectral matrix
$\mathbf{C}_\times$	Theoretical ego-noise CSM
$\mathbf{C}_G$	GEVD-denoised CSM
$\tilde{\mathbf{p}}$	Denoised pressure vector
L_u	Number of snapshots for the uncorrelated noise estimation
T_c	Acquisition time for the stationary noise estimation
P	Robot position
l	Robot search area length
x_g, y_g	Position of the ground robot

γ_g	Orientation of the ground robot
d	Robot step size
V_g	Ground robot speed
l_f	Distance from the wheels to the center of mass of the ground vehicle
u_s	Steering control
$_d$	Desired value
T_m	Duration of one iteration for the robot
r_g	Turning radius of the ground robot
l_D	UAV arm length
$F_{i,z}$	Thrust force of the motor i
k_f	Thrust coefficient of the propeller
ω_i	Angular velocity of the motor i
u	Total lift force
m	Mass of the UAV
a_g	Gravitational acceleration
τ_i	Reactive torque at the motor i
τ_γ	Yaw torque
x, y, z	Cartesian coordinates
α, β, γ	Roll, pitch and yaw angles
T	Kinetic energy
$\mathcal{U}$	Potential energy
I_x, I_y, I_z	Moments of inertia along the principal axes
$\mathbf{Q}$	Generalized force vector
$\tau_\alpha, \tau_\beta, \tau_\gamma$	Roll, pitch and yaw torques
e_i	Tracking error on the variable i
$K_{P,i}, K_{I,i}, K_{D,i}$	Proportional, integral and derivative gains on the variable i
u_z	Desired altitude performance
R	Propeller radius
M_t	Tip Mach number

Contents

Abstract iii

Résumé iv

Abbreviations v

Symbols vi

Contents ix

Introduction 1

1 MEMS microphone arrays for real-time DOA estimation 9
- 1.1 Sound propagation fundamentals . . . 10
  - 1.1.1 The acoustic wave equation . . . 10
  - 1.1.2 Analytical solutions in free space . . . 10
  - 1.1.3 Green's function representation of acoustic fields . . . 11
  - 1.1.4 The forward acoustic problem . . . 11
  - 1.1.5 Discrete forward problem for microphone arrays . . . 12
  - 1.1.6 Modelling microphone array measurements . . . 12
- 1.2 The cross-spectral matrix . . . 12
  - 1.2.1 Theoretical formulation . . . 13
  - 1.2.2 Practical estimation . . . 13
  - 1.2.3 Estimation parameters . . . 14
  - 1.2.4 Parameter trade-offs . . . 15
- 1.3 Acoustic imaging . . . 15
  - 1.3.1 Beamforming . . . 16
  - 1.3.2 Subspace methods . . . 17
  - 1.3.3 Time-based approaches . . . 18
  - 1.3.4 Machine learning-based methods . . . 18
  - 1.3.5 Choosing a sound source localization algorithm . . . 18
  - 1.3.6 Scanning space . . . 18
- 1.4 Microphone arrays . . . 20
  - 1.4.1 Hemispherical microphone array . . . 20
  - 1.4.2 Type of microphones . . . 20
  - 1.4.3 Microphone array geometry . . . 21
  - 1.4.4 Calibration . . . 24
- 1.5 Implementation and algorithmic pipeline . . . 25
  - 1.5.1 Algorithm flow and steps . . . 25
  - 1.5.2 Computational considerations . . . 25
  - 1.5.3 Parameter selection for real-time performance . . . 26
- 1.6 Conclusion . . . 27

2 Trajectory planning for autonomous sound source localization 29
- 2.1 Sound source localization with microphone arrays . . . 30
  - 2.1.1 Acoustic imaging using a single microphone array . . . 30
  - 2.1.2 Fusion of multiple cross-spectral matrices from multiple arrays . . . 30

2.1.3 DOA fusion from synchronous arrays . . . 31
2.1.4 Triangulation from sequential DOA estimations . . . 32
2.2 Enabling triangulation for sound source localization . . . 32
2.2.1 Two-dimensional triangulation from asynchronous measurements . . . 32
2.2.2 Path-planning algorithm for effective triangulation . . . 34
2.2.3 Simulations results and performance analysis . . . 36
2.3 Handling mulitple sound sources via source masking . . . 39
2.3.1 Spatial filtering for source separation . . . 39
2.3.2 Integration into the SSL process and computational impact . . . 42
2.3.3 Experimental validation . . . 43
2.4 Path planning for autonomous multiple sound sources localization . . . 45
2.4.1 Simulation framework for multiple sound sources . . . 46
2.4.2 Conditional spatial filtering approach . . . 46
2.4.3 Validation of the conditional filtering approach . . . 49
2.4.4 Challenging scenarios . . . 50
2.4.5 Autonomous localization of multiple sources in simulations . . . 51
2.5 Discussions . . . 53
2.6 Conclusion . . . 54

3 Ground robot for sound source localization 57
3.1 Experimental platform . . . 57
3.1.1 Environment setup . . . 57
3.1.2 Ground robot description . . . 58
3.1.3 Communication architecture . . . 61
3.2 Experimental results . . . 62
3.2.1 Simulation and experiment comparison . . . 63
3.2.2 Real-time autonomous localization of multiple sources . . . 63
3.3 Achieving continuous real-time localization . . . 65
3.4 Discussions . . . 68
3.5 Conclusion . . . 70

4 Ego-noise generation and filtering for UAVs 73
4.1 UAV fundamentals . . . 74
4.1.1 Experimental platform . . . 74
4.1.2 Principles of UAV flight . . . 75
4.1.3 Dynamic model of a UAV . . . 76
4.1.4 UAV control strategy . . . 77
4.2 Noise generated by a quadcopter . . . 81
4.2.1 Propeller-induced acoustic sources . . . 81
4.3 Ego-noise mitigation techniques . . . 87
4.3.1 Passive noise reduction methods . . . 87
4.3.2 Numerical noise filtering techniques . . . 88
4.4 Practical implementation of denoising . . . 91
4.4.1 Estimation of the measurement CSM . . . 91
4.4.2 Estimation of the ego-noise CSM . . . 92
4.4.3 Temporal variations of the ego-noise CSM coefficients . . . 93
4.5 Experimental validation . . . 96
4.5.1 Case study: noise point source . . . 97

4.5.2 Case study: UAV interference during stationary hover . . . 99
4.6 Discussion . . . 99
4.7 Conclusion . . . 100

5 UAV-based sound source localization **103**
5.1 Platform and validation . . . 103
5.1.1 Experimental setup . . . 103
5.1.2 UAV control . . . 104
5.1.3 Path planning using true DOAs measurements . . . 106
5.2 UAV-based real-time sound source localization . . . 109
5.2.1 Experimental setup . . . 109
5.2.2 Results . . . 110
5.3 Discussions . . . 114
5.4 Conclusion . . . 115

General conclusions **117**

A Convergence of the path-planning algorithm **121**

B Cramér–Rao Bound for 2-DOAs Triangulation **123**

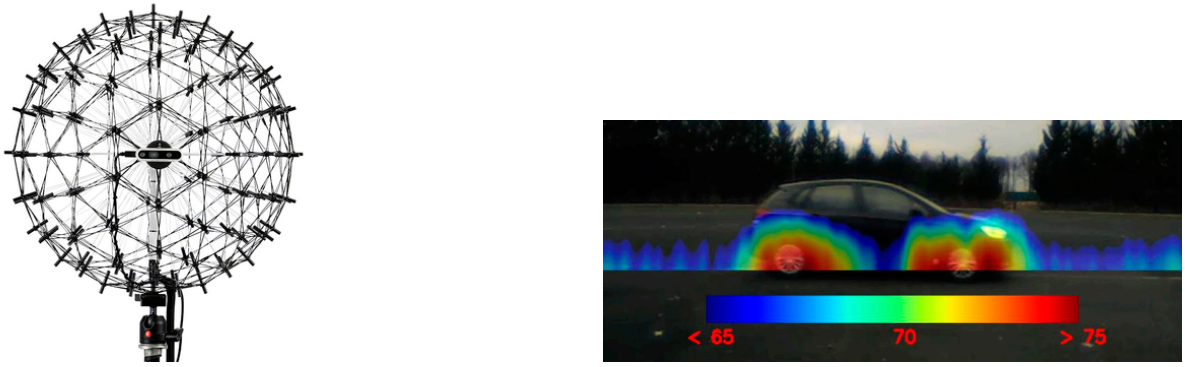
References **127**

Index **135**

Introduction

Sound is a powerful medium for conveying information. In both biological and artificial systems, it serves as a means of communication, detection, and situational awareness. Humans and animals rely on auditory cues to interact with their environment, detect threats, and coordinate actions [1]. For instance, bats and dolphins, amongst others, use echolocation to navigate and hunt in their surroundings, emitting sound waves and interpreting the echoes that return from objects in their path [2]. More generally, animals rely on the direction of sound to locate threats or prey. They are able to perceive it using their ears, which allow them to detect sound and interpret its direction based on the time difference of arrival and the sound level difference at each ear [3]. In an industrial setting, sound can provide information on the condition of machinery, allowing for predictive maintenance and early detection of faults. It is increasingly used for monitoring, diagnosis, and control in a wide range of applications.

Capturing and interpreting sound requires appropriate signal processing techniques. While a single microphone can provide basic amplitude and frequency information, it lacks the spatial information needed to determine the direction or distance of a source. This limitation is overcome using microphone arrays, which use the spatial distribution of microphones to perform Direction-Of-Arrival (DOA) estimation. Figure 1a illustrates a typical microphone array setup, where multiple microphones are arranged in a specific geometric configuration to capture sound from different directions. When associated with a camera, these microphone arrays can be referred to as acoustic cameras. They combine audio and visual data to create a comprehensive understanding of the sound field in a given environment. These tools enable precise localization of sound sources by producing acoustic images. Figure 1b illustrates the concept of acoustic imaging, where the sound field is represented as a two-dimensional map, allowing for the identification of sound sources based on their spatial characteristics. In this case, Leiba *et al.* [4] used a microphone array to visualize the sound sources on a driving car, demonstrating the strong pedagogical value of acoustic imaging in understanding sound fields.



a) b)

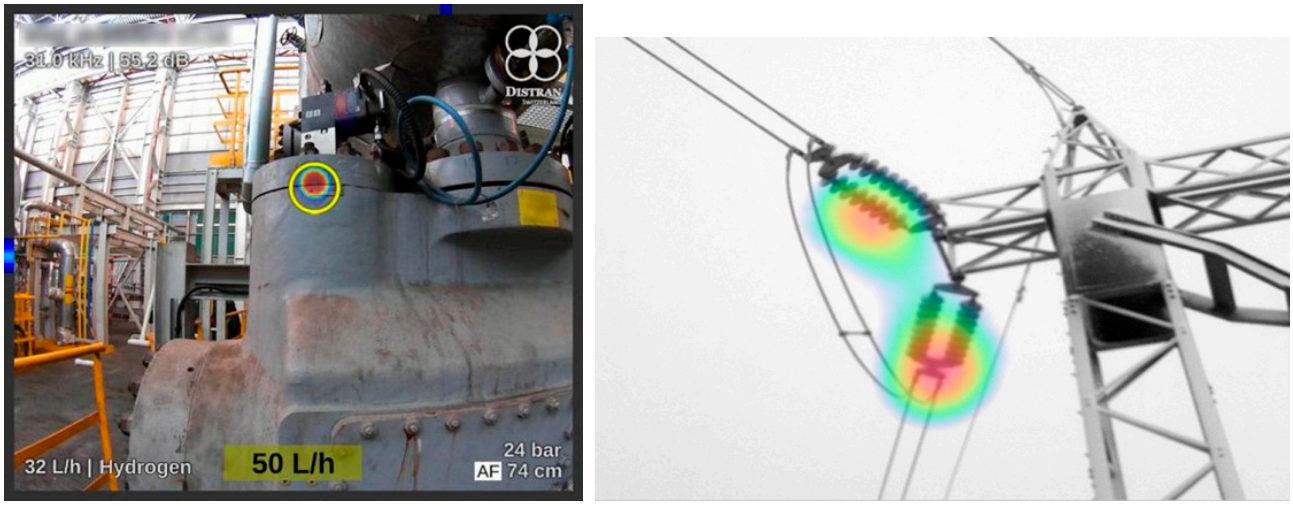
Figure 1. Microphone arrays and acoustic imaging. (a) illustrates a typical acoustic camera setup (credit gfai tech), while (b) shows the concept of acoustic imaging, where the sound field is represented as a two-dimensional map,from [4].

Sound source localization and acoustic diagnosis

Sound Source Localization (SSL) refers to the process of estimating the position of a sound source based on recorded audio signals. It plays a central role in numerous applications such as surveillance, industrial monitoring, and human-assistive technologies. In addition, acoustic diagnosis techniques allow engineers to identify faulty components or hazardous conditions by analyzing their acoustic signatures. These capabilities are further enhanced by source separation methods that isolate individual sources from complex auditory scenes.

Industrial applications

The use of acoustic cameras for acoustic diagnostics in industrial settings is particularly promising, as evidenced in the work from Fiebig *et al.* [5], in which an acoustic camera is used to identify the main sound sources in an industrial site as a first step towards lowering the total acoustic emissions, or DeCoensel *et al.* [6], who performed long term industrial noise surveillance using beamforming from a microphone array. It is also useful for diagnosing machinery, as demonstrated by Cardenas *et al.* in [7], or also to investigate fan noise, as shown by del Val *et al.* in [8]. Recently, industrial acoustic cameras have been used to detect gas leaks, as in the work of Liang *et al.* in [9], and electrical partial discharge, as illustrated in Fig. 2. These devices show great potential for improving safety and efficiency in industrial environments by providing real-time monitoring and analysis of sound emissions. However, they still require human intervention for data collection and analysis, which can be time-consuming. Moreover, the industrial settings often involve large-scale areas to be covered, or hazardous environments such as explosive atmosphere in the case of inflammable gas leaks or high voltage areas in the case of electrical partial discharge. Nevertheless, the advancements in the field of robotics and mobile sensing technologies open new avenues for automated acoustic diagnostics, enabling remote and flexible monitoring of large-scale industrial sites [10], despite hazardous conditions.



a) b)

Figure 2. Industrial applications of acoustic cameras for diagnostics. (a) Gas leak detection (credit Distran), and (b) electrical partial discharge detection (credit FLIR).

Mobile sensing and robot audition

Recent research increasingly focuses on embedding microphones on robotic platforms. This paradigm, known as robot audition, enables robot to react to their surroundings dynamically and provides them

with additional information about their environment. The concept of robot audition was proposed in the early 21th century by Nakadai et Okuto [11], and was initially targeting application in human-robot interaction, where robots are required to understand and respond to human speech. More recently, Zhang *et al.* [12] reviewed works including non-verbal sound in human-robot interaction. Their review highlights the importance of seemingly uninteresting sounds in the interaction process, supporting the idea that robots designed to interact with humans greatly benefit from sound source separation and sound source localization capabilities. In the review [13], Rascon *et al.* specifically focused on the use of SSL in robotics and reviewed a large body of work in this area. Their review demonstrates the surge of interest in this field over the last three decades, and summarizes the main challenges in mobile SSL, including the consideration of acoustic features in robot navigation. Such capabilities open new possibilities for robots to autonomously navigate and interact with their environment, which is crucial for applications such as surveillance, and industrial monitoring.

Autonomous robotic systems for acoustic measurements

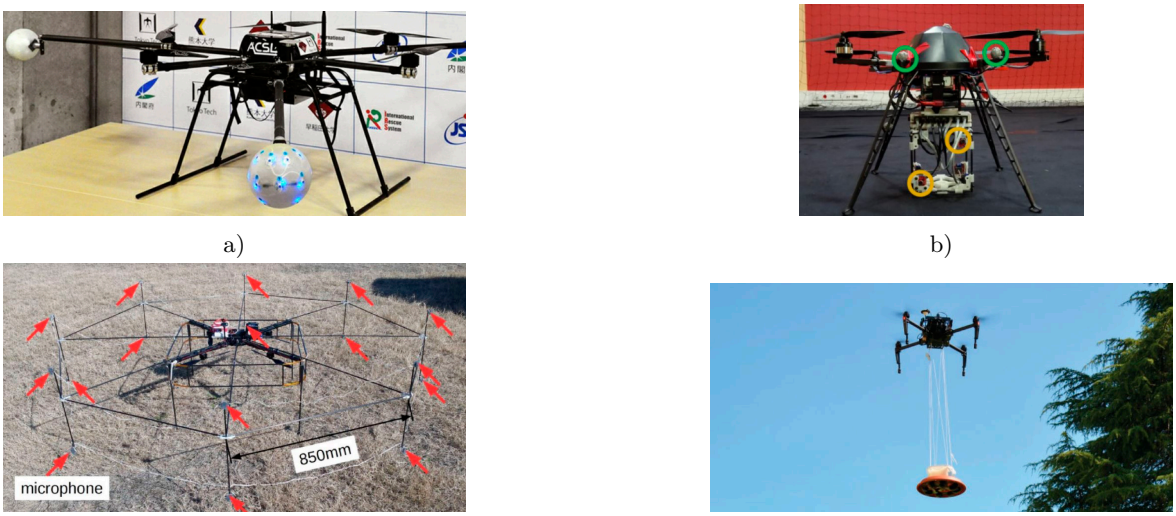
The development of autonomous robotic systems which actively move acoustic sensors to perform SSL has remained relatively limited, although it was one of the primary goals of robot audition, as stated by Nakadai et Okuto as early as 2000 in [11]. Later, in [14], Sasaki *et al.* demonstrated the possibility to exploit a 32-microphone-array-carrying robot motion to perform SSL of speech signal from loudspeakers in an indoor environment. The robot's motion is arbitrary, and the robot is able to localize multiple sound sources simultaneously by taking advantage of its own motion. Martinson *et al.* [15] later proposed an autonomous robot capable of localizing sound sources in a laboratory environment using an 8-microphone array and a patrol-loop algorithm. The robot explores an area following a pre-determined path, and extensively refines its spatial sampling in areas where a source is likely to be present. Although efficient, this method is time and measurement consuming. In [16], Evers *et al.* proposed an SSL method for arbitrarily moving robots in reverberant environments, and validated it on a two-microphone array simulated setup. The work did not focus on the robot's motion, but rather on the localization of the source and robustness to reverberations. It was extended in [17] to use a microphone array to perform acoustic Simultaneous Localization And Mapping (SLAM), which consist in jointly estimating the sources position and the robot's position. The method performed well in simulated scenarios with up to four sources, and was able to significantly improve the robot's self localization, which is a standard problem in mobile robotics. However, the algorithm is expected to be quite expensive computationally, making it challenging to implement on small moving robotic platforms. Although the method takes advantage of the robot's motion, its trajectory was set arbitrarily. In [18], Portello *et al.* used a binaural mobile microphone array to perform SSL of a source in an anechoic environment using a Bayesian filter and a source activation detection scheme. The approach was validated experimentally with arbitrary robot motion. It was proven to be robust to source intermittence and false measurements, but it is expected to struggle in presence of reverberations and was not designed to tackle multiple source scenario. This work was later extended by Magassouba *et al.* in [19], where the carrier robot movement is conditioned to perform SSL-related actions under different scenario. These include resolving the front-back ambiguity inherent to two-microphone setups, and approaching a sound source. This work seems to be the first to consider closed-loop control of the robot's motion based on the acoustic measurements. In [20], Nguyen *et al.* proposed a robot motion planning algorithm to find an optimal robot motion to minimize the uncertainty in the source localization based on a probabilistic occupancy grid. The robot's motion is driven towards maximum information gain, which is computed based on the estimated source position and extensive simulations of the room's acoustic properties. The results obtained through simulations confirm the performance of the approach with prior knowledge of the room's acoustic properties. More recently, in 2024, Yen *et*

al. [21] proposed a method to track moving sources using a swarm of UAVs equipped with microphone arrays. It uses a particle filter algorithm to fuse the estimated DOAs obtained from multiple agents at different positions. Simulation results demonstrated the feasibility of the approach, although the ego-noise of the drones heavily degrades the performance. This is the first attempt to use a multiagent mobile robotic system to localize sound sources. Note that most of the methods mentioned above use Bayes filters, which are typically sensitive to state modelling, parameters initialization and noise statistics prior [22]. The need for multiagent systems in SSL in the context of search and rescue was highlighted by Nakadai *et al.* in [23], where a proposition for an acoustic search and rescue multi-robot framework was proposed, aiming to be confronted to the feedback of rescue workers in the near future through a set of interviews. The advantage of multiagent systems over single robot is that they can perform the localization of non-stationary sources, which is out of the scope of this thesis. The main challenge currently facing autonomous SSL robotic systems is the ability to plan the robot's motion dynamically to localize multiple sound sources in an unknown environment. Although approaches have shown promising results in simulations, the experimental validation of these methods remains limited, and the performance of the algorithms in realistic scenarios is still an open question.

Drone audition

With the advent of lightweight computing and sensing technologies, autonomous robots are now capable of performing perception tasks without human intervention. Amongst these, Unmanned Aerial Vehicles (UAV) offer a unique platform for the monitoring of large and hazardous area. The integration of SSL capabilities on drones has been explored in several studies over the last decade, whether it is for search and rescue, wildlife monitoring, or industrial applications. For instance, in [24], Nakadai *et al.* proposed a drone equipped with a microphone array to localize sound sources in an outdoor environment, demonstrating the feasibility of drone-based SSL for a search-and-rescue application. This work was later extended by Wakayabashi *et al.* in [25] to include simultaneous multiple sources localization, however one major issue still remain before reaching real-world applications, which is the ability to mitigate the noise generated by the drone itself. Indeed, one of the most challenging part of robot audition is the ability to mitigate this self-generated noise, which is referred to as ego-noise. In the case of drone audition, the noise generated by the drone's propellers and the air flow over the microphone can significantly interfere with the measurements, greatly deteriorating the performance of SSL algorithms. In [26], Nakamura *et al.* proposed a method to numerically mitigate the ego-noise of a robot, and demonstrated its efficiency in an experimental setup. This method was later extended by Ohata *et al.* in [27] to UAV ego-noise to locate intermittent sound sources in an outdoor environment. A rise of interest in drone audition has been observed after these successful preliminary works, with several studies focusing on ego-noise mitigation and drone-based SSL. Public datasets such as the DREGON dataset [28] have been released to facilitate research in this area, and several algorithms have been proposed to mitigate the ego-noise of drones. These includes time-frequency filtering, proposed by Wang *et al.* [29], which selects the less noisy time-frequency bins to perform wideband acoustic images, and Manamperi *et al.* [30], who proposed a denoising algorithm based on prior knowledge of the UAV's ego-noise and the well known Wiener filter.

Regarding the location of the microphone array, various configurations illustrated in Fig. 3 have been proposed, including: Nakadai *et al.* [24] who placed the microphone array in the propellers plane, Go *et al.* [28], who placed the microphone array directly under the drone's body, Wakabayashi *et al.* [25] who placed the array around the drone in the propeller's plane, and finally Salvati *et al.* [31], who placed the array a few meters under the drone. The most popular configuration is to place the microphone array below the drone, although sparse studies suggests that UAV propeller noise should be minimized by placing the microphones in the propeller plane. No study has yet demonstrated



a) b)

c) d)

Figure 3. Different configurations of microphone arrays on drones. Microphone arrays located (a) in the propellers plane, from [24], (b) under the drone body, from [28], (c) around the drone in the propellers plane, and (d) a few meters under the drone, from [31].

the superiority of one configuration over another, and the choice of configuration is often driven by practical considerations such as weight distribution and stability of the drone. Another approach to tackle the ego-noise problem is to use a fixed-wing UAV, which emits less noise than a typical rotary-wing UAV. For instance, in [32], Basiri *et al.* proposed the fixed-wing UAV shown in Fig. 4, equipped with a microphone array to localize an emergency whistle sound in an outdoor wide-range environment. However, the use of fixed-wing UAVs limits the maneuverability and flexibility of the system, as they require a runway for takeoff and landing, and are less capable of hovering in place compared to rotary-wing UAVs.



Figure 4. The fixed-wing UAV setup used by Basiri *et al.* [32] for sound source localization.

Embedded systems and real-time processing

A key requirement for autonomous and adaptive robot audition is the ability to process audio data in real-time. It heavily constrains the design of the algorithms, as they must be efficient enough to run on the limited computational resources available. In their review, Chevtchenko *et al.* [33] noted that multiple approaches have been proposed to address this challenge, including the use of remote servers and communication architectures to offload the processing from the drone to a more powerful computer. However, this approach introduces latency and requires a reliable communication link, which may not be feasible in all scenarios. Throughout their review study [33], authors identified the principal challenges in the field of drone audition: first, the choice of microphone array configuration, which is currently limited to a small number of microphones due to payload and computational constraints. Second, the intended applications of drone-based SSL, which are primarily focused on human speech

or whistle localization for search-and-rescue operations, and would benefit from extending this test cases to other practical applications, such as industrial inspection. Finally, most studies tackled single source localization, while multi-source localization remains relatively unexplored, especially in the context of drone audition.
The work proposed throughout this PhD thesis aims to address these challenges by proposing a method for autonomous sound source localization using a microphone array embedded on a quadcopter UAV. The proposed method is designed to be efficient enough to run on the drone's onboard computer, enabling real-time processing and autonomous operation.

Problem statement and scope of the thesis

The problem statement of this thesis might be summarized as follows: ***How to autonomously localize multiple sound sources using a microphone array embedded on a quadcopter UAV?*** The goal is thus to enable a drone to autonomously navigate an environment and estimate the location of one or multiple sound sources. From the literature review, it is clear that the main challenges in this field are: 1) the development of a control scheme that allows the drone to efficiently explore the environment and localize efficiently sound sources, 2) the implementation of a real-time processing architecture that enables the drone to process the audio data, plan the motion of the drone, and estimate the sound source locations in real-time, and 3) the mitigation of the ego-noise generated by the drone itself, which can significantly degrade the performance of the sound source localization algorithms.

Target scenario and hypotheses

This project was supported and funded by the Centre des Techniques des Industries Mécaniques (CETIM), a French engineering center mutualizing research and development efforts for the mechanical industry. CETIM's interest in this work stems from the potential applications of drone-based SSL in industrial settings, where it can be used for monitoring and inspection of large-scale industrial sites, as well as for predictive maintenance of machinery. The industrial problem can be express as the need to be able to localize sound sources in an unknown environment, without human intervention. This challenge is particularly aligned with the aforementioned current advances and challenges in drone audition and provides us with a target scenario for the system developed throughout this work. Considering the problem and the challenges identified in the literature, this thesis thus focuses on a specific scenario, under laboratory conditions. The UAV is tasked with localizing stationary, immobile, and omnidirectional sound sources, in an indoor environment. The operational setting is assumed to be free of obstacles and with minimal reverberation. The UAV flies above the scene and is guided solely by acoustic cues to approach and hover near the source.
The work presented serves as a proof of concept and is, to the best of our knowledge, the first attempt to validate experimentally a fully autonomous drone-based sound source localization system. The hypotheses are discussed in more details throughout the manuscript, and perspectives for future work are proposed to extend the system capabilities to more complex scenarios, such as moving or intermittent sources, outdoor environments, or even collaborative multiagent systems.

Plan of the manuscript

To meet the challenge raised by the target scenario, this manuscript relates the steps of designing a method and an experimental platform for multiple sound sources localization from a UAV. The manuscript is organized as follows.

- **Chapter 1**: The question of how to estimate the direction of a sound source using a MEMS microphone array is addressed. The fundamentals of acoustic imaging are described and the key concepts are illustrated through simulations. The choice of the acoustic imaging method is discussed, as well as the parameters influencing the performance of the method and their practical choices.

- **Chapter 2**: Building on the previous chapter, the means to estimate the position of a sound source from asynchronous measurements of its direction are explored. The automatic placement of the microphone array is solved conceptually by proposing an adaptive path-planning algorithm based on acoustic measurements. A triangulation scheme is introduced to localize the sound sources, and a spatial filter allows previously found sources to be filtered out. The performance of the method is demonstrated through a numerical study under disturbances.

- **Chapter 3**: The implementation and validation of the path-planning algorithm are carried out on an autonomous ground vehicle in an indoor environment. A communication architecture enabling real-time processing is proposed and validated in a test case involving two sound sources. The method is then extended to operate in closed loop without requiring the vehicle to stop to perform acoustic measurements.

- **Chapter 4**: After validating the approach on a silent vehicle, the case of a UAV carrying the microphone array is investigated. The challenges of reducing the vehicle ego-noise on the microphone measurements are summarized and the denoising algorithm used to mitigate this noise is introduced. The denoising is tested for different interfering sources: a loudspeaker and a stationary hovering UAV.

- **Chapter 5**: Finally, the experimental results in the target scenario are presented, validating the methodology on a real UAV. The robustness of the acoustic-based control is studied, as well as the localization performance of the system. The intricate relationship between UAV control and the performance of the localization algorithm is highlighted, and the benefits of the proposed control are evidenced.

This work resulted in the development and validation of an autonomous UAV-based experimental platform for sound source localization. Significant opportunities for future improvement remain, and several research directions are proposed.

Publications

The work presented in this manuscript has lead to scientific productions published in or submitted to peer-reviewed journals and conferences, which are listed below:

- C. Leblanc, J.-D. Chazot, R. Marchiano, T. Gras, P. Castillo, "Robot control architecture for autonomous sound source mapping", InterNoise 2024, conference proceedings.
- C. Leblanc, J.-D. Chazot, R. Marchiano, T. Gras, P. Castillo, "Real-time autonomous multiple sound sources mapping for unmanned aerial vehicle embedded applications", Quiet Drone 2024, conference proceedings.
- C. Leblanc, P. Castillo, J.-D. Chazot, R. Marchiano, T. Gras, "Adaptive Sound Sources Localization using a Mobile Robot in Real-time Experiments", under peer review.
- C. Leblanc, J.-D. Chazot, R. Marchiano, T. Gras, P. Castillo, A. Tevera-Ruiz, "Drone audition: Closed-loop acoustic-based control for autonomous UAV-embedded sound source localization", in preparation.

1. MEMS microphone arrays for real-time DOA estimation

Contents

1.1 Sound propagation fundamentals . . . 10
1.1.1 The acoustic wave equation . . . 10
1.1.2 Analytical solutions in free space . . . 10
1.1.3 Green's function representation of acoustic fields . . . 11
1.1.4 The forward acoustic problem . . . 11
1.1.5 Discrete forward problem for microphone arrays . . . 12
1.1.6 Modelling microphone array measurements . . . 12
1.2 The cross-spectral matrix . . . 12
1.2.1 Theoretical formulation . . . 13
1.2.2 Practical estimation . . . 13
1.2.3 Estimation parameters . . . 14
1.2.4 Parameter trade-offs . . . 15
1.3 Acoustic imaging . . . 15
1.3.1 Beamforming . . . 16
1.3.2 Subspace methods . . . 17
1.3.3 Time-based approaches . . . 18
1.3.4 Machine learning-based methods . . . 18
1.3.5 Choosing a sound source localization algorithm . . . 18
1.3.6 Scanning space . . . 18
1.4 Microphone arrays . . . 20
1.4.1 Hemispherical microphone array . . . 20
1.4.2 Type of microphones . . . 20
1.4.3 Microphone array geometry . . . 21
1.4.4 Calibration . . . 24
1.5 Implementation and algorithmic pipeline . . . 25
1.5.1 Algorithm flow and steps . . . 25
1.5.2 Computational considerations . . . 25
1.5.3 Parameter selection for real-time performance . . . 26
1.6 Conclusion . . . 27

This chapter presents the theoretical foundations, algorithmic approaches, hardware considerations, and practical challenges associated with Direction of Arrival (DOA) estimation using Micro-Electro-Mechanical Systems (MEMS) microphone arrays. The goal is to establish a complete understanding of

the problem, from the physical modeling of sound wave propagation to the algorithmic and hardware considerations necessary for real-world applications.
The chapter begins with a brief introduction to the physics of sound waves, presenting the sound wave equation and its classical solutions. The Green's function representation is discussed as a key tool for modeling wave propagation from point sources, leading to the formulation of the forward problem. A discrete version of the forward problem is then presented, forming the basis for understanding the measurement principles of microphone arrays. Following this, the Cross-Spectral Matrix (CSM) is introduced as a fundamental statistical tool for analyzing multichannel microphone signals. Both the theoretical formulation and the methods for practical estimation of the CSM are detailed, including the discussion of parameters such as averaging techniques and windowing functions.
The chapter then examines various Sound Source Localization (SSL) algorithms. Classical methods such as beamforming, subspace-based methods like MUSIC, time-based approaches, and emerging machine learning-based techniques are reviewed. For each category, the computational demands, robustness, and suitability for real-time, embedded implementation are considered.
The design of the microphone array is a critical component that directly impacts the quality of DOA estimation. The choice of the microphones is discussed, and a hemispherical array design is proposed to ensure wide spatial coverage.
Finally, technological and scientific choices made to answer our problem are discussed. Particular attention is given to achieving real-time operation.

1.1 | Sound propagation fundamentals

Sound waves are mechanical waves that propagate through a medium, typically air, as pressure variations. This section develops the mechanisms of acoustic wave propagation on which SSL methods are based upon.

1.1.1 | The acoustic wave equation

Assuming a compressible, inviscid, and isotropic fluid at rest, the sound wave equation can be derived from the conservation of mass, conservation of momentum, conservation of energy, and the equation of state [34]:

$$\Delta\bar{p}(X,t) + \frac{1}{c^2}\frac{\partial^2\bar{p}(X,t)}{\partial t^2} = 0 \tag{1.1}$$

where $\bar{p}(X,t)$ is the acoustic pressure at location $X \in \mathbb{R}^3$ and time t, c is the celerity of the sound in the fluid, and Δ is the spatial Laplacian operator. For harmonic sound sources, the propagation of the sound wave at frequency f can be conveniently described in the frequency domain by the Helmholtz equation:

$$(\Delta + k^2)p(X,f) = 0 \tag{1.2}$$

where $k = \frac{2\pi f}{c}$, with c the celerity of the sound in the medium, is the wave number, and $p(X,f)$ is the Fourier transform of the pressure $\bar{p}(X,t)$ at frequency f. As all signals considered in the following are harmonic, the time dependency $e^{i2\pi ft}$ is implicitly assumed in the notation.

1.1.2 | Analytical solutions in free space

Considering a sphere of radius $a \to 0$, pulsating at frequency f as the sound source, the solution to the Helmholtz equation in free field (homogeneous and unbounded medium) is given in spherical coordinates by

$$p(r) = A\frac{e^{-ikr}}{4\pi r} \tag{1.3}$$

with i the imaginary unit, A the complex amplitude of the source, and r is the radius coordinate. A point source radiating a pressure field of the form (1.3) is called a monopolar source.
Now consider a sphere of radius $a \to 0$, oscillating at frequency f along an axis z as the sound source. Let ϕ be the angle from the axis z. Then, the solution to the Helmholtz equation in free field is given in spherical coordinates by

$$p(r, \phi) = B(\frac{1}{r} + ik)\frac{e^{-ikr}}{4\pi r}\cos\phi \tag{1.4}$$

with B the complex amplitude of the source, and r is the radius coordinate. A point source radiating a pressure field of the form (1.4) is called a dipolar source.
Another fundamental solution to the Helmholtz equation is the plane wave, which models sound propagation in an unbounded medium without amplitude decay. A plane wave traveling in direction $N \in \mathbb{R}^3$ is expressed as:

$$p(X) = Ce^{-ikN\cdot X} \tag{1.5}$$

where C is a complex amplitude, N is a unit vector indicating the direction of propagation, and $N \cdot X$ is the projection of position X along the propagation direction.

1.1.3 | Green's function representation of acoustic fields

The Helmholtz equation in free-field admits a fundamental solution known as the Green's function, which describes the acoustic pressure field generated at an observation point X due to a unit-amplitude monochromatic point source located at $X_s \in \mathbb{R}^3$.
In free-field conditions, the Green's function for the Helmholtz equation is given by:

$$g(X, X_s) = \frac{e^{-ik\|X-X_s\|}}{4\pi\|X - X_s\|} \tag{1.6}$$

where $\|X - X_s\|$ is the Euclidean distance between the source and measurement points.
This function provides the complex acoustic pressure at X due to a monopole point source at X_s and translates both the amplitude decay and phase shift due to spherical wave propagation.

1.1.4 | The forward acoustic problem

The acoustic field solution to Eq. (1.2) in presence of volume velocity acoustic sources within a domain V_s can be expressed as a superposition of elementary fields generated by a continuous distribution of monopoles [34]. In free-field, the pressure radiated by a distribution of monopoles of complex amplitudes $s(X_s)$ can be expressed through the Kirchhoff Helmholtz integral formulation:

$$p(X) = \iiint_{V_s} g(X, X_s)s(X_s)\,dX_s = \iiint_{V_s} \frac{e^{-ik\|X-X_s\|}}{4\pi\|X - X_s\|}\,s(X_s)\,dX_s \tag{1.7}$$

where $g(X, X_s)$ is the free-field Green's function and V_s denotes the volume inside which elementary sources are distributed. The pressure at M locations $\mathbf{X} = \{X_m\}_{m=1}^{M}$ can also be expressed as:

$$\mathbf{p}(\mathbf{X}) = \iiint_{V_s} \mathbf{g}(\mathbf{X}, X_s)s(X_s)\,dX_s \tag{1.8}$$

where $\mathbf{p}(\mathbf{X}) = \begin{bmatrix} p(X_1) & \dots, p(X_M) \end{bmatrix}^T \in \mathbb{C}^{M\times 1}$ is the complex pressure vector at positions $\mathbf{X}$, and $\mathbf{g}(\mathbf{X}, X_s) = \begin{bmatrix} g(X_1, X_s) & \dots, g(X_M, X_s) \end{bmatrix}^T \in \mathbb{C}^{M\times 1}$ is a vector containing the Green function from the source position X_s to the locations $\mathbf{X}$.

1.1.5 | Discrete forward problem for microphone arrays

If the source distribution is discrete and finite, which generally holds as a first approximation, Eq. (1.8) can be rewritten as:

$$\mathbf{p}(\mathbf{X}) = \sum_{q=1}^{Q} \mathbf{g}(\mathbf{X}, X_q) s(X_q) \tag{1.9}$$

where Q is the number of point sound sources, and X_q is the position of source q. It can, in turn, be rewritten in matrix form as:

$$\mathbf{p}(\mathbf{X}) = \mathbf{G}(\mathbf{X}, \mathbf{X}_s)\mathbf{s}(f) \tag{1.10}$$

with $\mathbf{G}(\mathbf{X}, \mathbf{X}_s) = \begin{bmatrix}\mathbf{g}(\mathbf{X}, X_1) & \dots, \mathbf{g}(\mathbf{X}, X_Q)\end{bmatrix} \in \mathbb{C}^{M\times Q}$, the so-called propagation matrix from Q sources to M locations at frequency f, and $\mathbf{s}(f) = \begin{bmatrix}s(X_1) & \dots, s(X_Q)\end{bmatrix}^T \in \mathbb{R}^{Q\times 1}$, a vector of the sound source amplitudes. For the sake of clarity, the frequency dependency of each term will be implicit from now on.

Eq. (1.10) offers a direct model for the complex pressures obtained at given positions within an acoustic free field, assuming the source distribution is known and finite. However, the focus of this work is to solve the inverse problem, i.e., recover the location $\mathbf{X}_s = \{X_q\}_{q=1}^{Q}$ of the sound sources from a set of measurements $\mathbf{p}$. This is commonly known as the SSL problem.

1.1.6 | Modelling microphone array measurements

SSL is typically performed using a set of microphones arranged in an array. Let M microphones be deployed in an array at known positions $\mathbf{X} = \{X_m\}_{m=1}^{M}$, capturing a sound field generated by Q uncorrelated point sources located at $\mathbf{X}_s = \{X_q\}_{q=1}^{Q}$. Each source emits a signal with complex amplitude s_q at frequency f. The propagation between a source and each microphone is modeled by the propagation matrix $\mathbf{G}_s = \mathbf{G}(\mathbf{X}, \mathbf{X}_s) \in \mathbb{C}^{M\times Q}$, which contains the phase and amplitude modifications due to propagation, dependent on the Green's function and source-receiver geometry.

The resulting measured pressure vector $\mathbf{p}(\mathbf{X}) \in \mathbb{C}^{M\times 1}$ at the array under free-field and non-dissipative medium assumptions is then modeled as:

$$\mathbf{p}(\mathbf{X}) = \mathbf{G}_s\mathbf{s} + \mathbf{n} \tag{1.11}$$

where $\mathbf{n} \in \mathbb{C}^{M\times 1}$ models additive noise, which is unknown in practice. This model will serve as the basis for solving the inverse problem of source localization. This inverse problem is ill-posed due to its sensitivity to noise and the non-uniqueness of its solutions.

1.2 | The cross-spectral matrix

While the model in Eq. (1.11) provides a direct relationship between the sound source amplitudes and the measured pressures at the microphones, it is often insufficient in practice due to random noise, or phenomenon that are not accounted for in the model, such as reverberations or variability in the source signals. Moreover, measurements over a single time-frame may not offer enough robustness for reliable source localization, especially in the presence of strong additive noise. To address these challenges, statistical representations of the measured pressure are standard practice. In particular, the Cross-Spectral Matrix (CSM) provides a more reliable representation of the acoustic field. This matrix contains both amplitude and phase relations between microphone pairs and is central to most of the acoustic imaging algorithms introduced later in this chapter.

The CSM of the measured pressures $\mathbf{C} \in \mathbb{C}^{M \times M}$ is defined as:

$$\mathbf{C} = \mathbb{E}\{\mathbf{p}\mathbf{p}^H\} \tag{1.12}$$

where the superscript ^{H} denotes the conjugate transpose, and $\mathbb{E}\{\}$, the statistical expectation. The diagonal elements of $\mathbf{C}$ correspond to the microphone autospectra, representing the power of the pressure signals captured by each sensor over a given frequency band. Its off-diagonal elements, known as crossspectra, contain information about the phase differences between microphone pairs over a given frequency band, which are essential for acoustic imaging.

1.2.1 | Theoretical formulation

Under the assumptions of uncorrelated and stationary sound sources and white, uncorrelated additive noise $\mathbf{n}$, the theoretical CSM can be expressed as:

$$\mathbf{C} = \mathbf{G}_s \mathbf{\Sigma} \mathbf{G}_s^H + \mathbf{N} \tag{1.13}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$ is a diagonal matrix which elements represent the squared amplitudes (sound power) of the sources, s_q^2 for $q \in [1, Q]$, and $\mathbf{N}$ is a diagonal matrix which elements represent the additive noise power on each microphone, noted n_m^2 for $m \in [1, M]$. Note that the cross-terms $\mathbf{s}\mathbf{n}^H$ and $\mathbf{n}\mathbf{s}^H$ average out to zero due to the statistical uncorrelated assumption between sources and noise.

1.2.2 | Practical estimation

In practice, the theoretical CSM defined in Eq. (1.13) is not directly accessible, as it relies on unknown expectations over the source and noise distributions. Instead, it is approximated using a finite set of microphone array pressure measurements, typically collected over multiple time frames, called snapshots. For a single snapshot l, the instantaneous CSM estimation is given by:

$$\mathbf{C}_l = \mathbf{p}_l \mathbf{p}_l^H \tag{1.14}$$

where $\mathbf{p}_l$ is the complex pressure vector at snapshot l. However, this single-snapshot estimation is highly sensitive to both the randomness of the source signals and the additive noise.
To improve robustness, the assumptions that the sources are stationary, uncorrelated, and ergodic and that the noise is uncorrelated and white are exploited. These assumptions allow the CSM to be averaged over L snapshots to obtain a more stable estimation of the CSM:

$$\hat{\mathbf{C}} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{p}_l \mathbf{p}_l^H \tag{1.15}$$

Under these assumptions, this empirical average converges to the theoretical CSM as:

$$\hat{\mathbf{C}} \xrightarrow[L \to \infty]{} \mathbf{G}_s \mathbf{\Sigma} \mathbf{G}_s^H + \mathbf{N} \tag{1.16}$$

This averaging process, following Welch's [35] method for cross-spectral density estimation, significantly reduces the variance of the estimation and mitigates the impact of random noise. Indeed, by taking L large enough, the additive noise is whitened, and its contribution is localized only on the autospectra.

1.2.3 | Estimation parameters

In practice, the accuracy of the CSM estimated via Eq. (1.15) depends on the proper choice of several key parameters. These parameters influence the trade-off between spectral resolution and variance reduction in the estimator. In the following, we discuss the roles of the number of snapshots, frequency resolution, overlap between snapshots, acquisition time, and the apodisation window. For further considerations about these parameters and CSM estimation in general, the reader is referred to the work of Stoica and Moses [36].

Number of snapshots

A higher number of snapshots allows for more averaging, which reduces the variance of the estimated CSM. However, increasing the number of snapshots may require, if the measurement time is constrained, using shorter time segments, thus degrading the frequency resolution of the discrete Fourier spectrum. Additionally, the number of snapshots is linearly linked to the computational time of Eq. (1.15), and increasing it makes the estimation slower. Furthermore, the amplitude of the residual noise component on the cross-spectra follows a $\frac{1}{\sqrt{L}}$ law (assuming the noise follows a zero-mean Gaussian process and snapshots are independent). This means that the higher the number of snapshots, the lower the benefits in terms of noise reduction in the CSM compared to the increase in computational cost. In some cases where the lowest source signal amplitude $\min(s_q)$ is much greater than the highest noise amplitude $\max(\sigma_m)$, then a single snapshot might be sufficient to yield correct SSL. It is thus important to set the number of snapshots responsibly, with respect to the residual noise amplitude acceptable(i.e., $\frac{\max(\sigma_m^2)}{\sqrt{L}}$). Note that this method does not reduce the noise on the autospectra, which is not a problem in the case of SSL, as the source amplitude is not of interest. It is also common to artificially set the diagonal elements of the CSM to zero, to nullify the impact of noise on autospectra.

Frequency resolution

The frequency resolution Δf determines the length of each snapshot used to compute the Fast Fourier Transform (FFT). A higher frequency resolution (i.e., a smaller Δf) requires longer snapshots. While longer snapshots improve the resolution of spectral peaks, they may require a longer acquisition time to obtain a sufficient number of independent snapshots. In the case of wideband acoustic imaging, multiplying the number of frequency bins also increases the computational time. Indeed, most algorithm compute the wide-band image as the average of the narrow-band images. If the frequency resolution of each image is higher, the number of images to average increases, thus increasing the computational time.

Overlap

A critical assumption for Eq. (1.15) to converge to the theoretical CSM is the statistical independence of the snapshots. Overlapping snapshots can be used to increase the effective number of snapshots obtained from a fixed-duration signal. For example, an overlap of 50% means that two consecutive snapshots share half of their data. In practice, increasing the overlap does not have an adverse impact on DOA estimation accuracy [37]. However, excessive overlap may yield nearly redundant snapshots, so its selection should balance between obtaining more averaging and maintaining statistical independence.

Acquisition time

The acquisition time, T_{acq}, is the total duration over which pressure signals are collected. It is defined as:

$$T_{\text{acq}} = \frac{1}{\Delta f}[1 + (L-1)(1-\rho)] \tag{1.17}$$

where L is the number of snapshots, Δf is the spectral resolution, and ρ is the overlap factor. It must be chosen based on the desired number of snapshots and frequency resolution in order to keep an acceptable amount of overlap. Additionally, the acquisition time must be chosen so that the signal can be considered stationary over the measurement period. If the acquisition time is too short, the stationary hypothesis for the signals might not hold. A long acquisition time, together with a high enough number of snapshots, is expected to reduce the bias in the estimator of Eq. (1.15). In SSL, source signal stationarity is not particularly relevant, since finding the source amplitude is not of paramount importance. However, it becomes important when filtering an unwanted source, as is the case with a carrier robot noise, where an accurate estimation of the robot-noise (i.e., the unwanted source) CSM is required. Chapter 4 is dedicated to this issue; thus, in the general SSL scenario, when no particular filtering of the CSM is involved, the stationarity of the source signal is not assumed.

Apodisation window

The selection of the apodisation (or windowing) function is the final parameter influencing the CSM estimation. The window function (e.g., Hanning, Hamming, or Blackman) is applied to each data segment before computing their Fourier transform. It allows to reduce spectral leakage.

1.2.4 | Parameter trade-offs

Considering all these parameters, the estimator from Eq. (1.15) can be seen as a complex trade-off between reducing variance, maintaining high spectral resolution of the estimated CSM, and keeping the total computational time low. In our case, the choice of these parameters is constrained by the total computational time of the CSM estimation. Real-time processing requires the computing time to be lower than the acquisition time. The guidelines followed throughout this work are the following: 1) the number of snapshots is picked so that the residual noise is low enough to perform SSL (this will be developped in further sections), 2) the acquisition time is selected to ensure the reactiveness of the robot, 3) the frequency resolution is set to 10Hz and Hanning windowing is used to keep the spectral leakage acceptable, and finally, 4) the overlap is obtained through Eq. (1.17). If the overlap is larger than 90%, the acquisition time is increased accordingly.
Following these guidelines should allow a robust estimation of the CSM and poses strong foundations for the acoustic imaging algorithms described in the next section.

1.3 | Acoustic imaging

The estimation of the Cross-Spectral Matrix provides a robust statistical description of the sound field captured by the microphone array. However, the ultimate objective of SSL is to recover the source distribution generating this sound field. The following section introduces the fundamental concepts and methods behind acoustic imaging and discusses how the CSM enables the practical solution of the inverse problem in complex acoustic environments.
The literature offers a wide range of approaches to tackle SSL using microphone arrays, including time-based approaches, beamforming, subspace methods, acoustic holography, deconvolution methods, Bayesian inference, or deep learning based-methods. For the present application, the choice of method

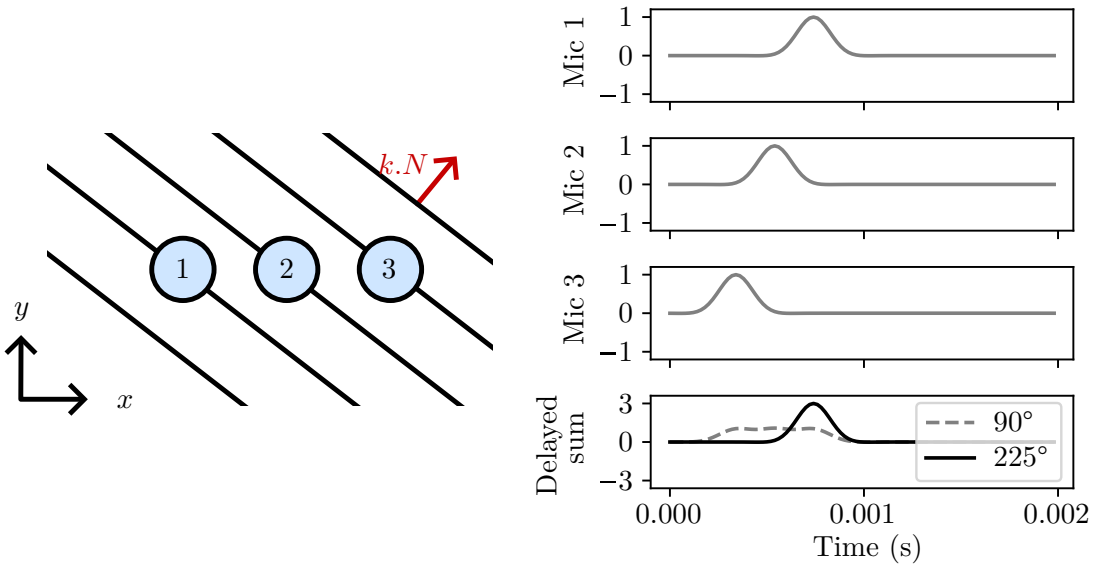


Figure 1.1. Principle of the delay-and-sum algorithm. A plane wave impinging on the 3-microphone linear array with a wave vector $k.N$, arriving at the array with an angle of 225°. The signals received by each microphone are delayed due to the propagation of the wave. The output of the beamformer is calculated for candidate source directions of 90° and 225°, which causes the signals to sum incoherently and coherently, respectively.

will primarily be determined by its computational efficiency. As comprehensive reviews of the most widely used SSL techniques are already available [13, 38, 39], this section will focus on the most viable, and thus popular, methods for our use case, with a focus on real-time processing capabilities.

1.3.1 | Beamforming

The most illustrative acoustic imaging algorithm is the Delay-and-Sum (D&S) beamformer [40]. Its principle is to apply time delays to signals measured by each microphone to compensate for the propagation effects. A time delay is derived analytically for each candidate location of a sound source. If the candidate position corresponds to that of a real source, the signals impinging from that source are summed coherently, yielding maximal output power. Conversely, if the source is located at a different position, the signals sum incoherently and average out. This principle is illustrated in Fig. 1.1. The D&S beamformer output can be expressed as:

$$Y_{\mathrm{D\&S}}(t, X, f) = \sum_{m=1}^{M} p_m(t + \tau_m(X)) \tag{1.18}$$

where $p_m(t)$ is the time signal from the m-th microphone, X is the candidate source location, and $\tau_m(X)$ is the time delay associated with microphone m for a source at X. The sum is the most coherent when $X = X_s$, and thus the source location, X_s, can be estimated by finding the position, X, that maximizes the power of $Y_{\mathrm{D\&S}}(t, X)$.

The operation of evaluating Eq. (1.18) for a set of candidate sources is referred to as beamforming, and the spatial representation of $Y_{\mathrm{D\&S}}(t, X)$ is often called the spatial spectrum, or beamforming map. In practice, the operation of delaying time signals is inconvenient, and the frequency equivalent of Eq. (1.18) is preferred. The frequency D&S beamformer [40] is given as:

$$Y_{\mathrm{FD\&S}}(X, f) = \sum_{m=1}^{M} p_m(f) e^{j2\pi f \tau_m(X)} \tag{1.19}$$

where $p_m(f)$ is the Fourier transform of the pressure measured by microphone m at frequency f, and $e^{j2\pi f\tau_m(X)}$ is actually equal to $\frac{g^H(X_m,X)}{\|g(X_m,X)\|}$. In the context of acoustic imaging, $g^H(X_m,X)$ is often referred to as the steering vector. In quadratic form, we get the Conventional Beamformer (CBF) [40], or Bartlett beamformer:

$$Y_{\text{CBF}}(X,f) = \frac{\mathbf{g}^H(X)\hat{\mathbf{C}}(f)\mathbf{g}(X)}{\|\mathbf{g}(X)\|^4}, \tag{1.20}$$

with $\hat{\mathbf{C}}$ the estimated CSM over the microphones and $\mathbf{g}(X)$ the Green function for a source located in X and radiating over our M microphone array. The source location within a search domain Ω can be found by maximizing the CBF output as:

$$\hat{\mathbf{X}}_{s\text{CBF}}(f) = \arg\max_{X\in\Omega} Y_{\text{CBF}}(X,f) \tag{1.21}$$

A wide variety of beamformers exist and fit different purposes; for instance, the Minimum Variance Distortionless Response Beamformer(MVDR-BF or Capon-BF) [40] is used as a spatial filter to reconstruct a source signal in the presence of strong interferences.

1.3.2 | Subspace methods

The most famous subspace-based SSL algorithm is the MUltiple SIgnals Classification (MUSIC) [41]. It is based on the central assumption that the signals from the sources and from the noise lie in orthogonal subspaces within the CSM. The eigenvalue decomposition of the CSM enables to separate the two subspaces such as:

$$\mathbf{C} = [\mathbf{U}_{\text{signal}}|\mathbf{U}_{\text{noise}}]\mathbf{\Lambda}[\mathbf{U}_{\text{signal}}|\mathbf{U}_{\text{noise}}]^{-1}, \tag{1.22}$$

where $\mathbf{\Lambda} \in \mathbb{C}^{M\times M}$ is a diagonal matrix of eigenvalues, $\mathbf{U}_{\text{signal}} \in \mathbb{C}^{M\times Q}$ denotes the so-called signal subspace, and $\mathbf{U}_{\text{noise}} \in \mathbb{C}^{M\times M-Q}$ the noise subspace. The eigenvalues are ordered in descending value, and the strongest values are expected to represent the main signals coming from the main sound sources. Each eigenvector is then associated with a “virtual" sound source, a linear combination of one or multiple “physical" sound sources [42]. While exploring possible source positions X in a search space Ω, it is possible to take advantage of the orthogonality between the source and noise subspaces. The most probable position of the sound source, $\hat{\mathbf{X}}_{\text{MUSIC}} \in \Omega$, is found by maximizing the MUSIC estimator

$$Y_{\text{MUSIC}}(X,f) = \frac{1}{\mathbf{g}^H(X)\mathbf{U}_{\text{noise}}\mathbf{U}_{\text{noise}}^H\mathbf{g}(X)}, \tag{1.23}$$

$$\hat{X}_{\text{MUSIC}}(f) = \arg\max_{X\in\Omega} Y_{\text{MUSIC}}(X,f)$$

This approach is known to achieve higher resolution than traditional beamformers but is more costly in terms of computational power because of the eigenvalue decomposition involved, although optimizations and adaptations exist, such as beamspace variations of MUSIC which can produce directly wide-bands images, as proposed by Argentieri *et* Danès in [43]. More recently, Huang *et al.* [44] proposed a greedier implementation of a wide-band MUSIC algorithm, optimized to run on a GPU in real time. MUSIC seems to be a good candidate for real-time SSL, although it remains heavier than CBF in terms of computational complexity.

1.3.3 | Time-based approaches

Time-based approaches to SSL rely on estimating the relative arrival times of an acoustic signal at different microphones within an array. The most prominent among these is the Time Difference of Arrival (TDOA) method, which forms the basis for algorithms such as the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) [45]. The GCC-based methods operate by computing the delay that maximizes the cross-correlation between signals received at microphone pairs, thereby allowing triangulation of the source direction or position. They have the advantage of performing broadband (in terms of frequency) estimations of the source location at a low computational cost. They yield an accurate estimation of the main sound source location, however, their performance degrades significantly in noisy environments, especially in the presence of a strong interfering source such as a carrier Unmanned Aerial Vehicle (UAV).
Time-based methods have also been used by Rakotoarisoa *et al.* [46] to characterize aeroacoustic sources, which tend to have a short duration, causing averaging-based method such as the one used throughout this work to fail.

1.3.4 | Machine learning-based methods

Machine learning approaches to SSL, and more particularly deep learning algorithms, have gained significant traction over the past decade. Grumiaux *et al.* [47] conducted a comprehensive survey and identified 156 papers published between 2011 and 2021 that proposed DL-based SSL algorithms. While many of these works demonstrated impressive localization performance under controlled conditions, the authors concluded to a few current limitations, and most notably that DL-based SSL methods often struggle with adaptability to dynamic acoustic environments. They typically require that the measurement setup (e.g., microphone array geometry and hardware) remain consistent between training and inference. These constraints are particularly limiting in our context, where both the operating environment and the UAV platform are expected to evolve during ongoing developments. Nevertheless, such ML-based methods remain promising candidates for future implementation, especially once a microphone array configuration on the carrier vehicle has been settled, and sufficient training data can be accumulated.

1.3.5 | Choosing a sound source localization algorithm

As highlighted in the above (non-exhaustive) review of the literature, a wide variety of SSL algorithms exist, each with their own strengths, assumptions, and limitations. Therefore, selecting the most suitable method for a given application can be a cumbersome task. However, the core contribution of this work is to explore the interaction between acoustic imaging algorithms and the control strategy of a robot equipped with a microphone array. Consequently, the SSL algorithm was selected based on general acceptance, robustness, and ease of integration. In this context, the CBF method was adopted due to its widespread use and strong consensus within the community. Its interpretability and straightforward implementation make it an ideal baseline. For comparison purposes, the MUSIC algorithm was also employed in the experimental phase. MUSIC remains a benchmark in the field due to its high spatial resolution capabilities and its denoising capabilities.

1.3.6 | Scanning space

A key parameter in beamforming algorithms is the choice of the scanning space Ω, which depends on the assumed sound source model. If sources are modeled as acoustic monopoles, the algorithm must search over a 3D spatial domain (x, y, z), requiring high computational effort but enabling accurate localization in near-field scenarios where the source-array distance is low. Note that more

complex steering vectors can be used for the sound wave propagation, for instance taking into account reflections on walls or the ground [48], non-static flow conditions [46], or even more complex sources such as dipoles [49], however, under the far-field and small array aperture assumption, sources can be sufficiently accurately modeled as plane waves with planar wavefronts over the array. The search then reduces to two angular parameters (θ, ϕ), significantly lowering computational costs while maintaining good localization performance for distant sources. This problem is known as the DOA estimation problem.

Beamforming in spherical coordinates

Using the plane wave model, the DOA estimation problem is conveniently formulated in spherical coordinates. The azimuth–coelevation framework employed throughout this work is introduced in Fig. 1.2. The direction of a point X of cartesian coordinates X_{cart}=(1,1,1) can be described in this framework as $\Theta(X) = (\theta, \phi) = (\frac{\pi}{4}, \frac{\pi}{4})$. This direction framework is sufficient when the source to be localized is located in the acoustic far field. The notion of far-field is detailed in the following paragraph.

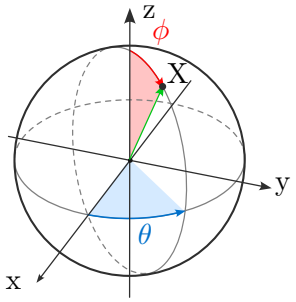


Figure 1.2. Azimuth(θ)-coelevation(ϕ) framework used in this work.

Acoustic far field for microphone arrays

Whether a sound source is located in the acoustic far-field or not determines the validity of the plane wave assumption. Thus, the far-field can be defined as the region in which the phase shifts on the microphone signals induced by the curvature of the wavefronts are negligible. The Fraunhofer distance [50] is defined as the minimal distance for which these phase errors on microphones are higher-bounded by ϵ_φ, and is given by:

$$R_F = \frac{\pi}{4}\frac{D^2}{\epsilon_\varphi \lambda} \tag{1.24}$$

with D the largest inter-microphone distance, also known as the array aperture, and λ the acoustic wavelength. For example, accepting a maximal error of $\frac{\pi}{8}$ rad, we obtain the well-known result $R_F = 2\frac{D^2}{\lambda}$. Conversely, assuming $D \leq \lambda_{max}$, keeping a distance of $5\lambda_{max}$ guarantees that the phase errors are kept lower than $\frac{\pi}{20} \approx 9$ deg. This means that if the array aperture is designed so that it is smaller than the largest wavelength of interest, keeping a distance of 5λ to the sources guarantees model approximation phase errors lower than 9 deg.

In this study, since the array is moved throughout space, its position can be controlled to maintain a sufficient distance from the sources, thereby ensuring that the far-field assumption holds, as introduced in Chapter 2. The relationship between phase error and DOA estimation error is complex and depends on the array geometry. However, increasing phase error inevitably leads to a larger DOA estimation error.

1.4 | Microphone arrays

The performance of SSL algorithms is inherently linked to the geometry and quality of the microphone array used. While a wide range of array configurations exists in the literature, their choice depends on practical considerations such as spatial constraints or desired localization coverage.
This section introduces the microphone array used in this work, along with its key performance metrics and the motivations behind the design choices.

1.4.1 | Hemispherical microphone array

In this work, we employ a custom-built hemispherical microphone array consisting of 32 MEMS microphones arranged over a dome with a diameter of 20 cm. This geometry was chosen to provide a compromise between spatial coverage and mechanical compatibility with the UAV platform.
The whole acquisition chain is illustrated in Fig. 1.3. During experiments, the microphones are covered with polyurethane foam to prevent turbulent noise generated by the motion. It measures 25 cm in diameter and 15 cm in height and, together with its acquisition card, weighs 450 g.

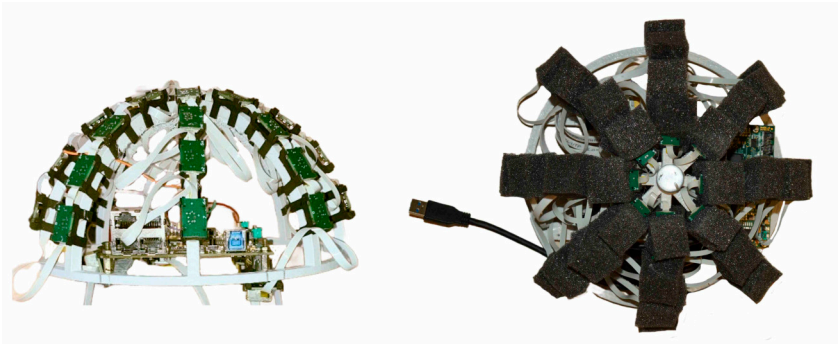


Figure 1.3. The Megamicros acquisition system and MEMS microphone array. (left) without the windshields, (right) with the polyurethane wind shields.

This array configuration was designed to support real-time acoustic imaging and SSL in 3D space, enabling the exploration of dynamic localization and control strategies for mobile robotic platforms.

1.4.2 | Type of microphones

The rapid development of MicroElectroMechanical Systems (MEMS) microphones over recent decades has enabled significant advances in the field of acoustic sensing, especially in embedded applications. While their individual performance may be inferior to that of traditional electret microphones, MEMS devices are widely adopted due to their low cost, compact size, and minimal weight, which makes them ideal for building dense and lightweight microphone arrays. Vanwynsberghe *et al.* [51] demonstrated that MEMS microphones have consistent sensitivity, near omnidirectional directivity, and flat frequency response between 80 Hz and 2 kHz, making them suitable for array signal processing and, more specifically, for acoustic imaging tasks.
In this study, the use of a MEMS-based array is essential for onboard integration with an UAV, where payload and space constraints are critical. This work builds upon the previous developments from the Megamicros project at the Institut ∂'Alembert, which produced a modular MEMS microphone acquisition system tailored for flexibility and portability.
The Megamicros system consists of 32 MEMS microphones that can be arranged freely to fit specific geometric requirements, along with a USB-compatible acquisition board. A dedicated Python library provides real-time signal processing capabilities in parallel to data acquisition, offering a convenient platform to develop embedded acoustic applications.

1.4.3 | Microphone array geometry

The spatial arrangement of microphones, commonly referred to as the array geometry, plays a central role in the performance of acoustic imaging algorithms. This section introduces key geometric parameters that influence spatial resolution and localization accuracy. To highlight the benefits of three-dimensional microphone placement, qualitative metrics of the array used in this work are compared with those of a uniform linear array with a 2 cm inter-microphone spacing. One key advantage of a 3D geometry is its ability to resolve the front-back ambiguity inherent in planar arrays, which occurs when all microphones lie on the same plane.

Point-spread function

The Point-Spread Function (PSF) is a fundamental concept in acoustic imaging, adapted from image processing. In this context, it characterizes the system's response to a unit-amplitude, far-field source located at a specific direction. Given a source at $(\theta_{\text{dir}}, \phi_{\text{dir}})$, the PSF describes how the array's beamforming output spreads across other directions in the scan space. Ideally, a perfect array would produce a Dirac delta centered at the true source direction. However, in practice, the response is spread over a main lobe and several side lobes. The width of the main lobe indicates the spatial resolution of the array, and narrower lobes mean better source separation. The level of the side lobes, on the other hand, reflects the robustness to interference and spatial noise. High side-lobe levels can result in false detections of sources and degraded localization accuracy.

The PSF for a plane wave coming from direction ($\theta_{\text{dir}} = 90$ °, and $\phi_{\text{dir}} = 90$ °) is compared for our geometry and the linear array at three different frequencies in Fig. 1.4. First, it is clear that the UAL suffers from spatial ambiguity due to its axial symmetry along the X-axis, making it impossible to recover the coelevation angle of the source. Since the UAL is also symmetric along the Z-axis, it also suffers from front-back ambiguity, which translates to a double main lobe in the PSF. The 3D geometry, however, does not suffer from this issue, and the main lobe resembles a Gaussian shape centered on the source. Since the number of microphones is high, the side-lobe level is relatively low compared to the main lobe; however, when the frequency is increased, side lobes appear, polluting the acoustic imaging. These two parameters, the main-lobe width and the side-lobe level, are linked to the geometry of the array. The main-lobe width is known to be related to the array aperture, and the side-lobe level to the number of microphones, i.e., the density of microphones. The latter is illustrated in Fig. 1.5, where the PSF at a frequency of 2000 Hz is computed for four sub-arrays sampled from the 3D geometry containing respectively, 4, 8, 16, and 32 microphones. The sub-arrays are taken so that the aperture is similar between each configuration. It can be seen that the main-lobe width is independent of the number of microphones, whereas the side-lobes level is considerably reduced when the number of sensors is increased.

Array directivity

Since array processing relies on the phase differences of an impinging sound wave across the microphones, the response of an acoustic imaging algorithm is inherently dependent on the geometry of the microphone array. This direction-dependent response is known as the directivity pattern of the microphone array.

To illustrate the impact of directivity on SSL, consider two sound sources of identical sound pressure levels. If the array exhibits strong directivity, it may favor one source over the other, potentially masking the second source entirely. This is problematic as it could lead to missing important sources while scanning the search space.

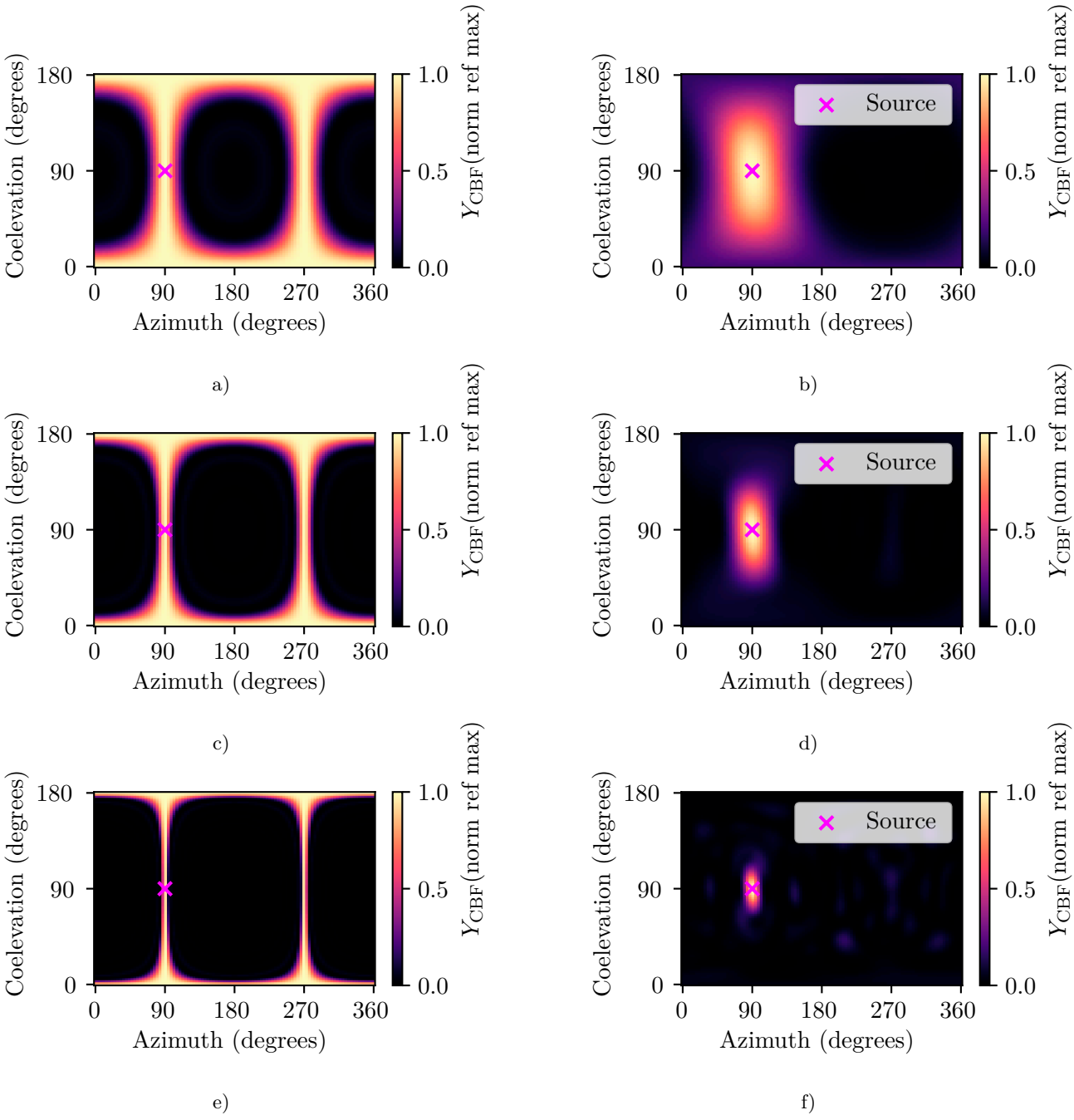


Figure 1.4. Point-spread functions for the uniform linear array at (a) 1000 Hz, (c) 2000 Hz, (e) 5000 Hz, and the 3D array at (b) 1000 Hz, (d) 2000 Hz, (f) 5000 Hz.

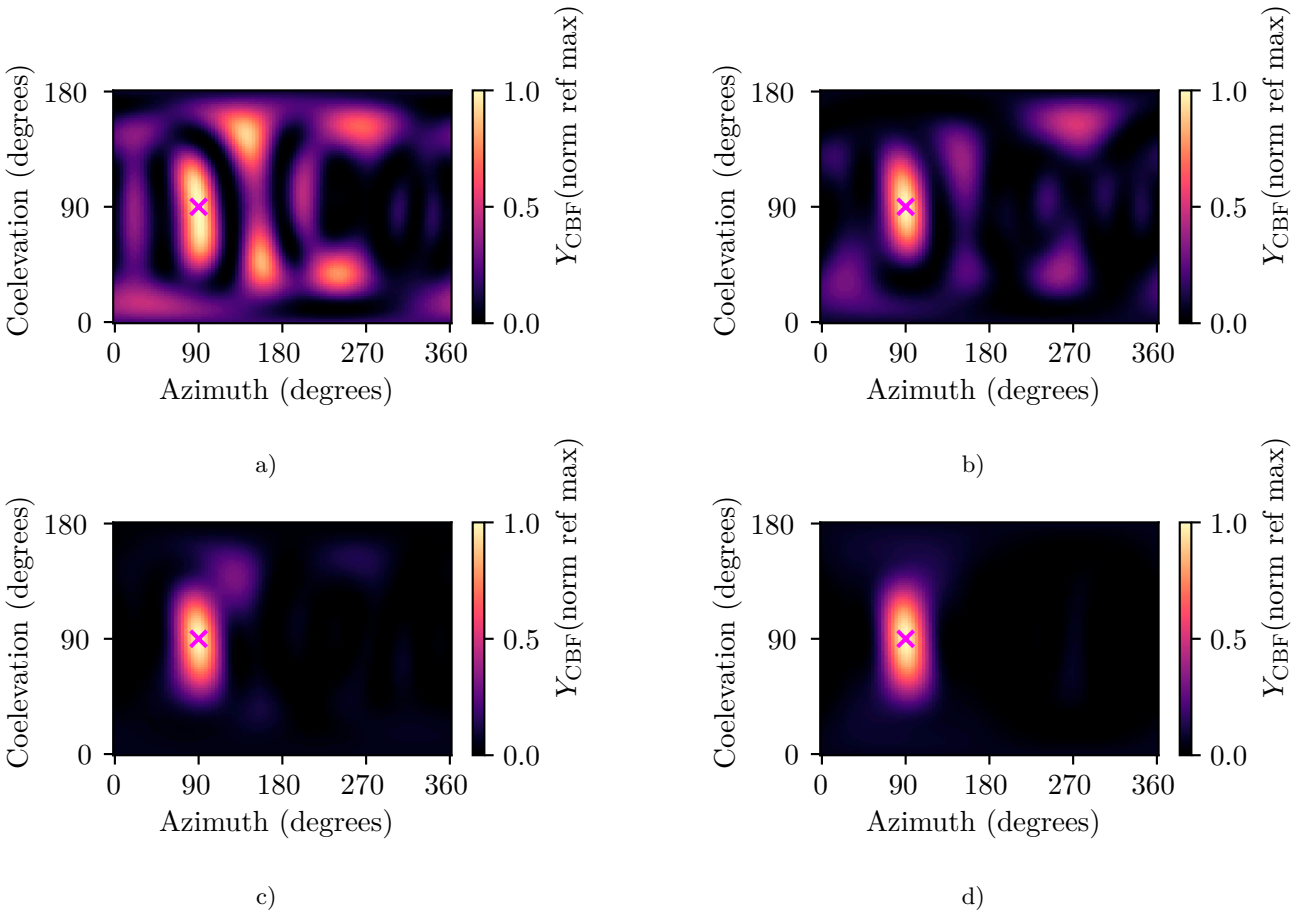


Figure 1.5. Point-spread functions at 2000 Hz for sub-arrays from the 3D array containing (a) 4, (b) 8, (c) 16, and (d) 32 microphones.

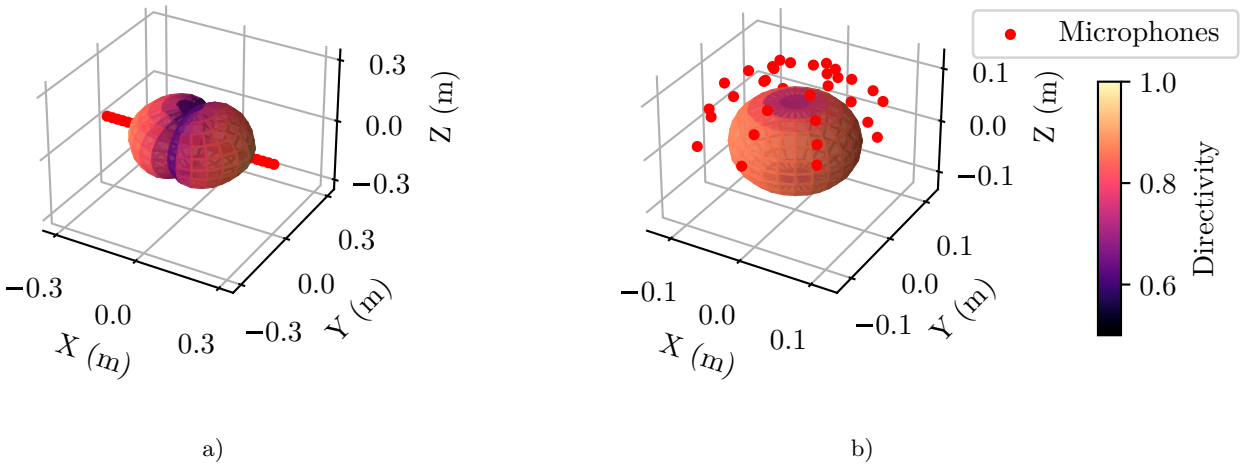


Figure 1.6. Directivity diagrams and array geometries for (a) a uniform linear array of spacing 2 cm, and (b) the array used throughout this work.

The directivity Γ of an array is defined here as one minus the average value of the PSF over all directions of arrival, following the concept introduced in [52]:

$$\Gamma(\theta_{\text{dir}}, \phi_{\text{dir}}) = 1 - \iint_{\Omega} \frac{1}{A_{\Omega}} \text{PSF}(\theta_{\text{dir}}, \phi_{\text{dir}}, \theta, \phi) d\theta d\phi \tag{1.25}$$

where A_{Ω} denotes the area of the domain Ω, the unit sphere in this case, $\text{PSF}(\theta_{\text{dir}}, \phi_{\text{dir}}, \theta, \phi)$ represents the point-spread function of the array, for a plane wave arriving from direction $\theta_{\text{dir}}, \phi_{\text{dir}}$.
If the PSF resembles a Dirac impulse centered on $\theta_{\text{dir}}, \phi_{\text{dir}}$, the integral in Eq. (1.25) approaches zero. In contrast, broader main lobes or significant side lobes cause the integral to increase, approaching one. As a result, directivity Γ tends to 1 when the response of the array is highly focused (ideal PSF) and to 0 when the response is more diffuse or contaminated by sidelobes. A nearly uniform directivity value indicates an omnidirectional array, while sharp peaks reveal directional sensitivity.
The directivity pattern of the microphone array used in this work is compared to that of a uniform linear array in Fig. 1.6. As expected, the ULA exhibits an axisymmetric directivity pattern and suffers from reduced directivity for sources arriving at $\theta = 90$ °, where inter-microphone distances are the smallest, thus more sensitive to errors. In contrast, the geometry of the 3D array demonstrates a more omnidirectional pattern, as intended by its design.

1.4.4 | Calibration

Various calibration techniques can be used to ensure the position of the microphone is known with high precision. Among them, the most popular are the of laser telemetry to pinpoint the microphone positions, or the use of a precisely printed 3D support. Vanwynsberghe *et al.* [53] proposed a time-difference of arrival calibration technique to estimate the position of each microphone in high number microphone arrays. In this work, since the microphone number is relatively low, and a single point-of-view is sufficient to locate all the microphones, we use an optical method based on camera imaging.

Procedure

The calibration of the microphone position in the array is crucial to obtaining satisfactory performance. The calibration procedure follows these steps: 1) the microphones are placed on the 3D-printed support structure, 2) a high-quality picture is taken, 3) geometric distortions and perspective effects within the picture are corrected using a planar homography image morphing algorithm [54], based on a prior calibration of the camera lens, 4) the microphone locations are automatically detected through image processing techniques, and finally 5) the detected 2D positions are then mapped onto the 3D model of the array, which was used to manufacture the support. This approach provides a fast and high-precision estimation of microphone positions using a minimal and simple setup. Moreover, using a high number of microphone is known to improve the acoustic imaging performances, thus the precision required for individual position calibration becomes less critical [55].

Calibration errors

The phase mismatch induced by a sensor positioning error is upper bounded, and we have the relation:

$$\epsilon_{\varphi} = \frac{2\pi\epsilon_X}{\lambda} \tag{1.26}$$

with ϵ_{φ} the standard deviation of phase errors, ϵ_X the standard deviation of position errors, and λ the wavelength. Thus, these phase errors follow a zero-mean Gaussian process of standard deviation ϵ_{φ}. Van Trees showed in reference [40] that the average squared value of the PSF for a source located

at a given position X_s can be linked to the number of sensors and the phase mismatches on the array, regardless of their origin (phase calibration or position calibration) as:

$$\mathbb{E}\{\mathrm{PSF}_{\mathrm{measured}}(X_s, X)^2\} \approx \mathrm{PSF}_{\mathrm{theo}}(\mathrm{X_s}, \mathrm{X})^2 \exp(-\epsilon_\varphi^2) + \frac{\epsilon_\varphi^2}{M^2} \tag{1.27}$$

with $\mathbb{E}\{\mathrm{PSF}_{\mathrm{measured}}(X_s, X)^2\}$, the statistical expectation of the squared actual PSF (after committing sensor location errors) at a given steering location X, $\mathrm{PSF}_{\mathrm{theo}}(X_s, X)$ the theoretical PSF, ϵ_φ the standard deviation of Gaussian phase errors, and M the number of sensors. The first term translates to an enlargement of the main lobe region, while the second term means an increase in the side-lobe level, which degrades the DOA estimation significantly. The latter can be compensated for by increasing the number of microphones, which confirms the intuition that an increased number of microphones decreases the effect of the positioning error. The error induced by microphone misplacement adds up to the modeling error from Section 1.3.6, and to each microphone imperfect phase calibration, hence the need to keep each of those to a minimum. Eq. (1.26) states that to obtain an error lower than $\frac{\pi}{20}$ at 2000 Hz, positioning errors should remain lower than 5 mm. The calibration procedure used in this work is assumed to yield sufficient precision.

1.5 | Implementation and algorithmic pipeline

One of the conclusions of the previous section is that increasing the number of microphones seems to be the solution to many problems, however, one aspect has not been tackled this far: the numerical complexity of the DOA estimation algorithm. Indeed, increasing the number of sensors adds more data to process, which becomes more and more challenging to operate in real-time. The current section begins by describing the algorithm flow from the microphone measurements to the estimated DOA, then lists all the computationally relevant parameters, and finally computes the algorithm's overall complexity concerning the different parameters.

1.5.1 | Algorithm flow and steps

The algorithm for real-time DOA estimation revolves around conventional beamforming techniques applied to a wideband microphone array. The total algorithm is organized as follows: the steering matrix $\mathbf{G}$ is computed for the given M-microphone array geometry, the evaluation grid, and n_f frequencies studied. It is stored to avoid unnecessary calculations at each iteration. The M microphone pressure signals are obtained through the Megamicros Python API. The Cross-Spectral Matrix (CSM) is computed using Eq. (1.15) over L snapshots, with frequency resolution Δf, over n_f frequency bins. The number of samples N_f in each signal is the relevant parameter to compute the complexity of the FFT spectrum, and is linked to the frequency resolution Δf and the sampling frequency f_s as $N_f = \frac{f_s}{\Delta f}$ The n_f beamforming maps are computed onto the evaluation grid following Eq. (1.19). The beamforming maps are averaged over the frequency axis to form the final wideband beamforming output. Finally, the DOA is estimated by locating the peak of the averaged map. This is summarized as a flowchart in Fig. 1.7.

1.5.2 | Computational considerations

The numerical complexity of each step is derived, and the relevant parameters are introduced. First, the computation of the steering matrix $\mathbf{G}$ requires $O(M \times n_\phi \times n_f)$ operations, where $n_\phi = \frac{180 \times 360}{\delta_\phi^2}$ is the number of points in the evaluation grid for a spherical grid of angular resolution δ_ϕ (in °). However, since $\mathbf{G}$ is precomputed and stored, this cost is negligible during real-time operation.

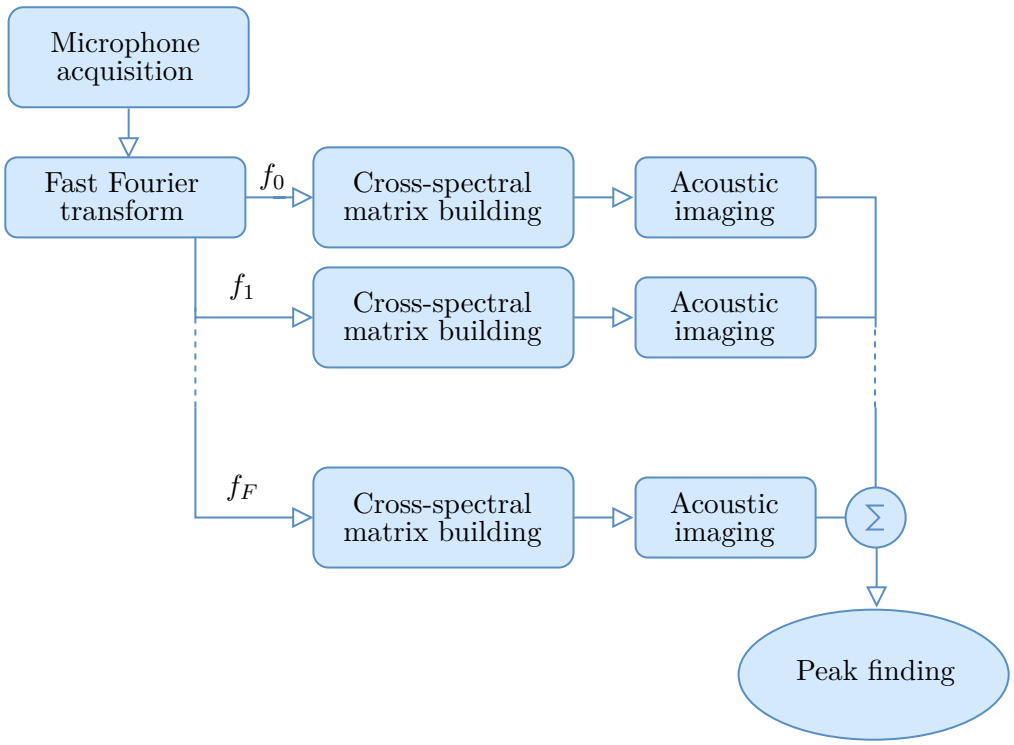


Figure 1.7. Flowchart of the DOA estimation algorithm.

The real-time computational load is dominated by: the segmentation and computation of a FFT spectrum of the pressure signals, totaling $O(M \times L \times N_f \log(N_f))$ operations; the estimation of the Cross-Spectral Matrix (CSM), which requires $O(M^2 \times n_f \times L)$ operations; the beamforming step, involving $O(M^2 \times n_\phi \times n_f)$ operations to compute the map at each frequency; the averaging across frequencies and the peak search on the map, which are of complexity $O(n_\phi \times n_f)$ and $O(n_\phi)$ respectively, and are considered negligible compared to other operations. The complexity of each step is summarized in Table 1.1. These values of complexity are especially informative when picking the values of each parameter. In particular, based on Table 1.1, the number of microphones plays a major role in the computational cost of each algorithm, and should thus remain limited to maintain real-time performances.

Step	**Complexity**
Steering matrix computation (offline)	$O(M \times n_\phi \times n_f)$
FFT of pressure signals	$O(M \times L \times N_f \log(N_f))$
CSM estimation	$O(M^2 \times n_f \times L)$
Beamforming map computation	$O(M^2 \times n_\phi \times n_f)$
Averaging over frequencies	$O(n_\phi \times n_f)$
Peak search	$O(n_\phi)$

Table 1.1. Summary of the computational complexity of each step.

1.5.3 | Parameter selection for real-time performance

The condition for real-time computing is here defined as "the processing time must be shorter than the acquisition time". It allows for continuous acquisition of acoustic data and ensures that no sample is missed. Additionally, there are benefits to keeping a short acquisition time from the robotic control point of view. More details on that point will be provided in Chapter 3. The acquisition time T_{acq} value is set to 125 ms to ensure that the system can operate in real-time, which corresponds to a

processing frequency of 8 Hz. As opposed to the numerical complexity, the computational time is dependent on the station running the program.
To evaluate the algorithm's real-time capability, we define the real-time factor as the ratio of computation time to acquisition time. A real-time factor below one ensures real-time operation. In this study, simulations were conducted to identify a suitable set of parameters meeting this condition and to determine the main computational bottlenecks. The chosen parameter set given in Table 1.2 yields a real-time factor of 0.3. Note that this process is largely hardware-dependent and further code optimization could reduce the value of the real-time factor. The implementation used for these simulations was developed in Python 3. No specific code optimizations were applied, and the execution was strictly linear, without parallel processing. All computations were performed on a ground station (Intel(R) Core(TM) i7-10750H CPU and 32 GB of RAM).

Parameter	Value	Definition
T_{acq}	125 ms	Acquisition time
M	32	Number of microphones
δ_ϕ	5°	Search grid resolution
n_f	1	Number of frequency bins
L	100	Number of snapshots
Δf	10 Hz	Frequency resolution
f_s	10 kHz	Sampling frequency

Table 1.2. Final set of parameters satisfying a real-time factor of 0.3.

Throughout this work, all acoustic imaging computations are performed on a single frequency band to demonstrate the feasibility of the approach, although it could be generalized to wide-band acoustic processing. Furthermore, the algorithms are suited for optimization through parallel computing. As illustrated in Fig. 1.7, the computations at each frequency band are independent and are well suited for parallel computing.

1.6 | Conclusion

This chapter presented the fundamental concepts and practical considerations involved in real-time DOA estimation using MEMS microphone arrays. Starting from the basic physics of sound waves, the inverse problem of acoustic imaging was introduced. The theoretical formulation of the cross-spectral matrix was developed, and practical methods for its estimation were described. Various sound source localization algorithms, ranging from classical beamforming to machine learning-based methods, were reviewed, highlighting their strengths and limitations for embedded applications. The choice of Conventional Beamforming as the DOA estimation algorithm was motivated by its simplicity and proven robustness. Additionally, its low computational cost offers real-time performance without needing advanced optimizations.
The design of the microphone array, including its geometry, the choice of MEMS microphones, and the calibration procedures, was discussed in detail to ensure robust and reliable measurements. Furthermore, the implementation strategy, including the algorithmic flow and numerical complexity analysis, was introduced, with a focus on achieving real-time performance under limited computational resources.

Summary table

Table 1.3 summarizes the choices made after carefully considering the trade-offs between performance and computational time.

Component	Choice	Motivations
Microphone type	MEMS Microphones	Low cost, small size, and light weight
Array geometry	Hemispherical Array	Omnidirectional and compact
CSM Estimation	Welch's algorithm	Reduces uncorrelated noise
SSL Algorithm	Conventional Beamforming	Robustness and real-time capability
Steering vector	Plane wave (angular formulation)	High fidelity in the Fraunhofer region and low computational

Table 1.3. Summary of choices for DOA estimation considering performance and computational constraints.

The concepts and tools presented here lay the foundations for developing autonomous sound source localization systems based on mobile robotic platforms, which will be the topic of Chapter 2.

2. Trajectory planning for autonomous sound source localization

Contents

2.1 Sound source localization with microphone arrays . . . 30
2.1.1 Acoustic imaging using a single microphone array . . . 30
2.1.2 Fusion of multiple cross-spectral matrices from multiple arrays . . . 30
2.1.3 DOA fusion from synchronous arrays . . . 31
2.1.4 Triangulation from sequential DOA estimations . . . 32
2.2 Enabling triangulation for sound source localization . . . 32
2.2.1 Two-dimensional triangulation from asynchronous measurements . . . 32
2.2.2 Path-planning algorithm for effective triangulation . . . 34
2.2.3 Simulations results and performance analysis . . . 36
2.3 Handling mulitple sound sources via source masking . . . 39
2.3.1 Spatial filtering for source separation . . . 39
2.3.2 Integration into the SSL process and computational impact . . . 42
2.3.3 Experimental validation . . . 43
2.4 Path planning for autonomous multiple sound sources localization . . . 45
2.4.1 Simulation framework for multiple sound sources . . . 46
2.4.2 Conditional spatial filtering approach . . . 46
2.4.3 Validation of the conditional filtering approach . . . 49
2.4.4 Challenging scenarios . . . 50
2.4.5 Autonomous localization of multiple sources in simulations . . . 51
2.5 Discussions . . . 53
2.6 Conclusion . . . 54

Chapter 1 explored the most popular acoustic imaging methods when it comes to estimating the direction of a sound source using microphone arrays. While acoustic imaging methods provide reliable estimations of a sound source's direction, practical acoustic diagnostics demand precise spatial Sound Source Localization (SSL). Here, a difference is made between Direction-of-Arrival (DOA) estimation, which estimates a given source's direction from the point of view of a microphone array, and SSL, which recovers the source position in space.

In our context, the ideal solution must be capable of locating the sound source's position within an industrial site, enabling an operator or an automated algorithm to identify and address issues efficiently. The challenge is thus to find a cost-effective and time-efficient method to locate as many sources as possible with high accuracy, in an unknown environment with an uncertain number of immobile sound sources.

The proposed approach aims to tackle this by integrating real-time cooperation between an acoustic imaging algorithm and an adaptive motion-planning strategy for a mobile robot. The acoustic imaging algorithm identifies the direction of the sound source, while the motion-planning algorithm efficiently determines the system's movements to locate the source's position.
The current chapter articulates around four main sections: in Section 2.1, a literature review of SSL methods is proposed, in Section 2.2, a path-planning algorithm is used to perform efficient autonomous SSL of a single source, Section 2.3 is focused on masking a sound source whose position is already known, and finally the multiple sound sources case is dealt with in Section 2.4.

2.1 | Sound source localization with microphone arrays

2.1.1 | Acoustic imaging using a single microphone array

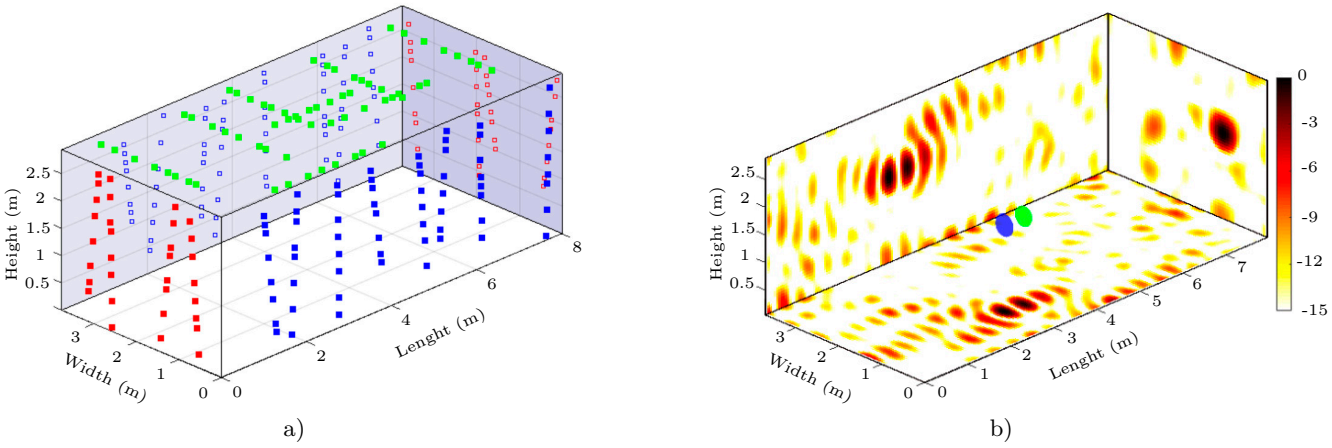


Figure 2.1. (a) The 256 MEMS microphones array geometry. (b) Reconstructed pressure field in the whole room using a room-reverberations-aware beamforming method in dB scale. The three orthogonal slices passing through the estimated location of the source are projected onto the room walls. The signed -3 dB isosurfaces are shown in blue and green. (figure from [56])

Acoustic imaging algorithms typically struggle to estimate the position of sound sources in 3D space using a single microphone array because they can only provide a single point of view of the acoustic scene [57]. However, the algorithms introduced in Chapter 1 can directly locate sound sources in 3D space by exploiting a microphone array deployed around the target region and searching for candidate source positions on a 3D grid. This approach has been used in studies such as [56, 58], where Demontis *et al.* covered the walls of a rectangular room with a 256 MEMS microphone array to create a 3D array wrapping the sound sources for characterization. Figure 2.1 illustrates the capability of 3D beamforming in localizing sound sources within the array's coverage. A similar method was also employed by Zhou *et al.* in [59] where a 3D array of microphones was mounted in the open section of an anechoic wind tunnel to obtain 3D acoustic images of aeroacoustic sources.
Notice that this method is efficient for characterizing sound sources in relatively small areas, but becomes impractical for large-scale regions due to the high number of sensors required.

2.1.2 | Fusion of multiple cross-spectral matrices from multiple arrays

To address the limitation of single arrays in covering large areas, an alternative approach is to fuse measurements from multiple arrays or from a single array at different positions. Multi-array methods are based on the concept of fusing Cross-Spectral Matrices (CSM) obtained from various asynchronous measurements into a single representation. This fusion can be achieved through different strategies.

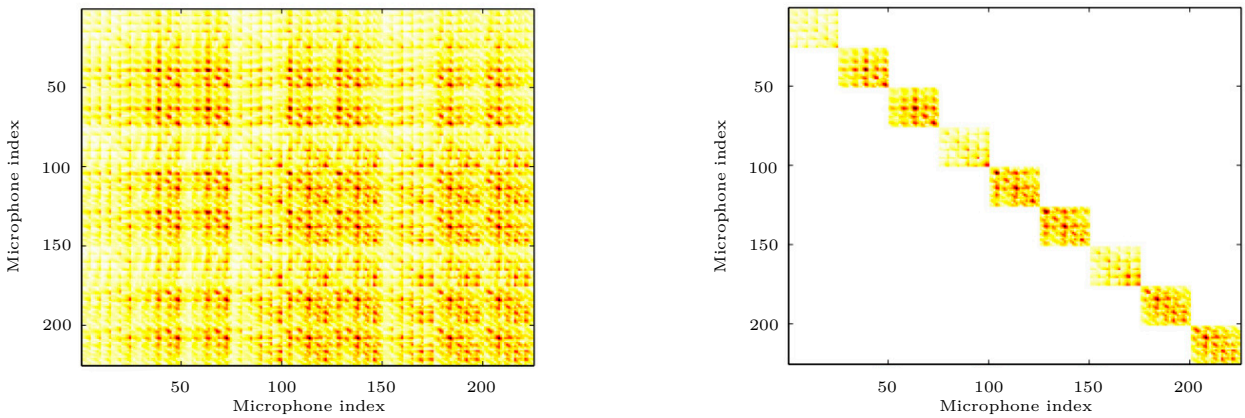


Figure 2.2. (a) The full CSM obtained from synchronous measurements. (b) The combined CSM is obtained by assembling asynchronous measurements. (figure from [60])

A common method is to assemble the estimated CSMs from multiple asynchronous measurements into a block-diagonal matrix, as shown in Fig. 2.2. This approach mimics a measurement with a synchronous array but suffers from the loss of phase information across the diagonal blocks. Although it is possible to apply classical imaging methods directly to the block-diagonal CSM using combined steering vectors [61, 62], advanced techniques aim to reconstruct the missing off-diagonal elements of the matrix to improve accuracy.

Most CSM reconstruction methods rely on the assumption that the measured CSM is low rank [60, 63, 64], reflecting the sparsity of sound sources generating the acoustic field. These methods use matrix nuclear norm minimization to recover the missing coefficients, but their computational cost makes them unsuitable for real-time applications.

Another popular approach involves the direct fusion of beamforming maps obtained independently from each measurement. These maps can be combined using methods such as arithmetic or geometric averaging [61], which offer comparable performance to block-diagonal CSM processing but at a significantly lower computational cost [65]. However, both CSM fusion and map fusion methods are post-processing techniques which require all measurements to be collected before processing, which becomes untractable in the large areas targeted within our work context.

2.1.3 | DOA fusion from synchronous arrays

An alternative to direct CSM or map fusion is to perform fusion at the level of DOA estimations. Instead of reconstructing or fusing acoustic images, this method uses DOA estimations from multiple measurements to recover the source position through geometric triangulation. This approach is computationally lighter and allows for real-time processing. For instance, in [66], Bavu *et al.* use multiple immobile microphone arrays and an optical device to estimate and track a UAV's position in real time. Each array provides a DOA estimation for the UAV, and the position is recovered using multimodal particle filtering.

The concept of SSL from multiple synchronous moving arrays is investigated by Yen *et al.* in [21]. In this work, multiple microphone arrays carried by UAVs are simulated to track the position of sound sources in real time using MUSIC and a particle filter. The UAVs adapt their positioning dynamically based on the source position estimations to minimize localization errors. The simulation results presented highlight the potential of this method for applications involving multiple mobile sound sources in complex environments. This work highlights the importance of the microphone array measurement positions, a central point in this chapter. However, some major known limitations of

the system, such as noise from a drone polluting the others, and practical synchronization issues, were not accounted for in the simulations.
The DOA-fusion-based SSL approaches have the advantage of being computationally efficient, as the data fusion assembles minimal data in the form of spatial coordinates and DOAs, while other methods fuse either pressure measurements or CSMs. They also enable flexibility in the acoustic imaging process, as they do not impose any requirement on the imaging method chosen. This makes them a robust and efficient solution, especially for applications valuing real-time adaptability.

2.1.4 | Triangulation from sequential DOA estimations

Although they are introduced for synchronous arrays in the previous examples, assuming quasi-static sources during the total measurement time, DOA fusion algorithms can be extended to asynchronous measurements from a single moving microphone array. In the context of this work, the sound sources are considered immobile and producing a stationary sound. The single microphone array SSL techniques are thus preferred here as they are less costly and sufficient to localize the targeted sound sources.
The main difference between multiple-array localization and single-array localization is that the latter requires advanced path-planning algorithms to avoid redundant measurements. In [20], Nguyen *et al.* use a ground vehicle carrying a microphone array to localize sound sources in a room. A motion planning method is proposed to find an optimal robot motion minimizing the SSL uncertainty. The robot motion is planned in real time in response to the acoustic measurements. A learning algorithm is trained using hundreds of simulations of the robot's motion in the presence of a sound source in the room. Information on the room characteristics is required to simulate the sound field accurately and to train the model properly. Once the algorithm has accumulated enough data, the robot can react to the current direction of the sound source and move to a measurement position that will minimize the localization error. In our context, the environment is unknown, making beforehand simulations of the experiment impossible.
Although robot-embedded SSL has grown in interest in the previous years, the autonomous localization of stationary and immobile sound sources remains an open problem. To fulfill the SSL in such an environment, we propose a method that plans the microphone array measurement position by reacting to each new measurement in a fully autonomous manner.

2.2 | Enabling triangulation for sound source localization

The concept of triangulation is central to the autonomous SSL problem. This section introduces the 2-DOAs triangulation scheme used throughout this work, as well as a path-planning algorithm leading to a measurement setup minimizing localization error.

2.2.1 | Two-dimensional triangulation from asynchronous measurements

When using a single array displaced to different positions, the problem of localizing a source can be solved using geometric triangulation to find the coordinate (x, y) of the source in a plane. This amounts to a two-dimensional problem illustrated in Fig. 2.3 where $S(x_s, y_s)$ stands for the source position, while $P_0(x_0, y_0)$ and $P_1(x_1, y_1)$ are two successive robot acoustic measurement positions. At each position, the acoustic imaging method yields an estimation $\hat{\theta}_i$ of the direction of the sound source. Using two estimated directions at two known vehicle positions, it is possible to recover the sound source coordinates $\hat{x}_s$ and $\hat{y}_s$ by solving a straightforward triangulation problem formulated as a system of two equations with the two unknowns:

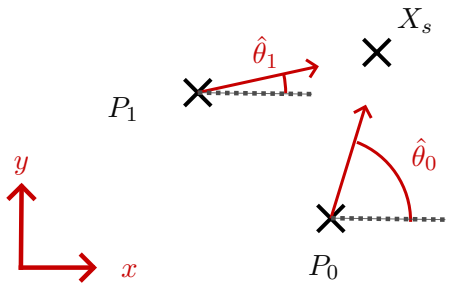


Figure 2.3. The 2D triangulation problem.

$$\mathbf{A_t}\ \hat{X}_s\ = \mathbf{b}_t \tag{2.1}$$

$$\text{with}\quad \mathbf{A_t} = \begin{bmatrix} \sin(\hat{\theta}_0) & -\cos(\hat{\theta}_0) \\ \sin(\hat{\theta}_1) & -\cos(\hat{\theta}_1) \end{bmatrix}, \quad \hat{X}_s = \begin{bmatrix} \hat{x}_s \\ \hat{y}_s \end{bmatrix}, \quad \mathbf{b}_t = \begin{bmatrix} x_0 \sin(\hat{\theta}_0) - y_0 \cos(\hat{\theta}_0) \\ x_1 \sin(\hat{\theta}_1) - y_1 \cos(\hat{\theta}_1) \end{bmatrix},$$

The estimated source coordinates are then directly obtained with a matrix inversion : $\hat{X}_s = \mathbf{A_t}^{-1}\mathbf{b}_t$ and it is possible to express the source estimated location $\hat{X}_s$ as

$$\hat{X}_s = \begin{bmatrix} x_0 + \frac{-(x_1-x_0)\sin\hat{\theta}_1 + (y_1-y_0)\cos\hat{\theta}_1}{\sin(\hat{\theta}_0-\hat{\theta}_1)} \cos\hat{\theta}_0 \\ y_0 + \frac{-(x_1-x_0)\sin\hat{\theta}_1 + (y_1-y_0)\cos\hat{\theta}_1}{\sin(\hat{\theta}_0-\hat{\theta}_1)} \sin\hat{\theta}_0 \end{bmatrix} \tag{2.2}$$

This 2-DOA geometric triangulation method's main advantage is its very low computational complexity compared to more advanced methods, such as Kalman filtering or particle filtering, which is a strong point in favor of its selection as the SSL strategy. Please note that, in the following, this triangulation scheme could be replaced by any other method without loss of generality.

Error propagation

Triangulation, regardless of the specific method used, is highly sensitive to the geometry of the triangle formed by the two measurement positions and the source. More precisely, its precision depends on the difference between the two estimated DOAs. Figure 2.4 illustrates the DOA estimation uncertainty as a cone around the true direction. Assuming similar uncertainties at both measurement positions, it is clear that the localization errors for the source position X_s are smaller in the second case, where the two DOAs are nearly orthogonal. This orthogonal configuration is the ideal case for triangulation, as it minimizes localization uncertainty. This property is proven in Appendix B. The worst-case scenario occurs when the two measurement positions are aligned with the source. In this case, the two DOAs are nearly identical, making triangulation impossible due to the lack of intersection between the direction estimations.

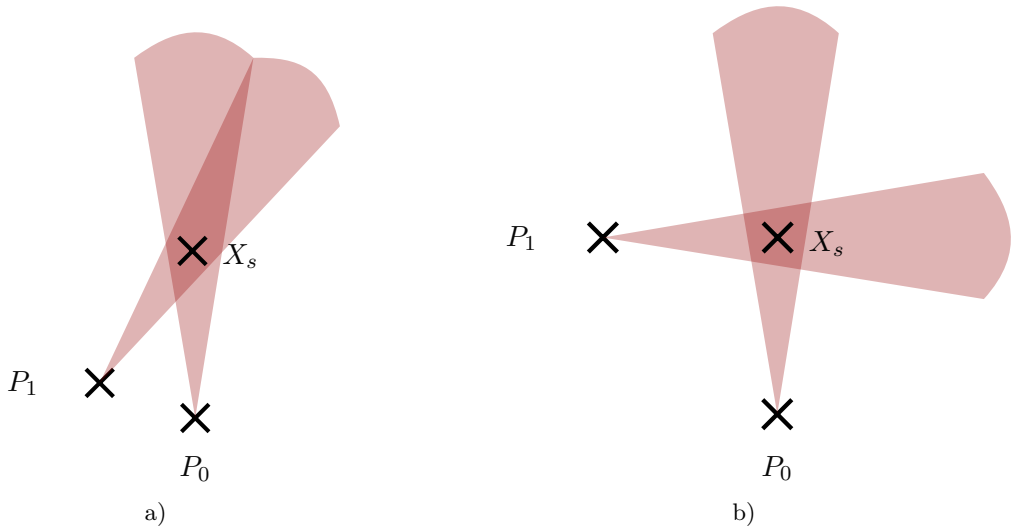


Figure 2.4. Visual interpretation of the link between DOA estimation uncertainty and localization uncertainty when DOAs are (a) closely separated, and (b) separated by 90°.

Constraints for precise triangulation

From this geometric approach of triangulation, the ideal scenario would be to place the microphone array in such positions that the estimated DOAs are orthogonal to each other.

From an acoustic consideration, DOA estimations are more reliable when the sound from the source emerges from background noise. On one hand, it is possible to increase the Signal-to-Noise Ratio by getting as close to the sound source as possible. On the other hand, as shown in Chapter 1, in order to benefit from the computational efficiency and generic formulation of a plane-wave steering vector, the measurement should take place in the source's acoustic far-field. The optimal measurement distance is thus the minimal distance satisfying the far-field assumption.

In summary, in order to minimize the triangulation error, two conditions must be met: (a) the two measurement positions must be close to the source to maximize the Signal-to-Noise Ratio (SNR) of the acoustic measurements while remaining in the acoustic far field, and (b) the two DOAs ($\hat{\theta}_0$ and $\hat{\theta}_1$) need to be orthogonal. The challenge is thus to fulfill these requirements without prior information on the sound source location.

2.2.2 | Path-planning algorithm for effective triangulation

We propose the following path-planning strategy for autonomous SSL to meet the requirements of the array's placement. At each iteration, acoustic measurements are performed, DOA is estimated, triangulation is performed, and the robot moves to the next location. The next robot's position is computed from its current position and the estimated DOA by moving one step of length d, at 45 ° to the left of the last estimated DOA. Turning always in the same direction prevents the robot from getting stuck and alternating between two positions. This path planning leads the robot closer to the source, following a swirl motion, allowing triangulation. Eventually, it gets close enough to the sound source and starts to circle it, while its successive positions follow a circle or radius $\frac{d}{\sqrt{2}}$ (a proof is given in Appendix A). This satisfies the two conditions to reduce the triangulation error to a minimum. Figure 2.5 illustrates the theoretical path followed by the robot starting at position P_0 in the presence of a sound source at position X_s. The robot's position at a given iteration j is noted P_j.

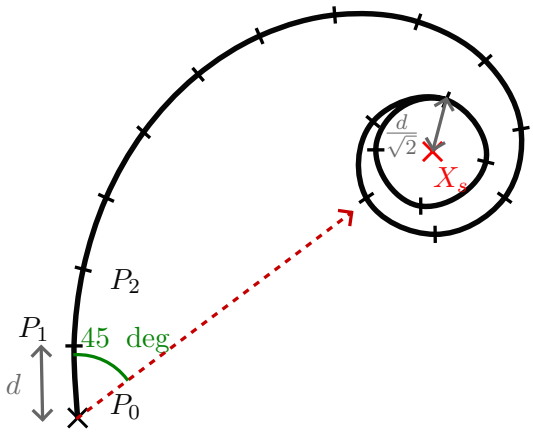


Figure 2.5. Theoretical trajectory with the proposed iterative method for a step d. At each iteration, the array is moved one step in the direction 45 ° to the left of the source DOA.

Algorithm 1 Adaptive trajectory planning for the single source case

while convergence is not achieved **do**
  Move one step at 45 ° to the left of the last DOA
  Estimate the new DOA
  Triangulate the source position
  Check for algorithm convergence
end while
return the averaged source position over the last 5 iterations

Convergence

The proposed convergence criterion is that the last five estimated positions lie within a circular neighborhood of radius equal to the wavelength, λ, around the most recent estimated position, as illustrated in Fig. 2.6.

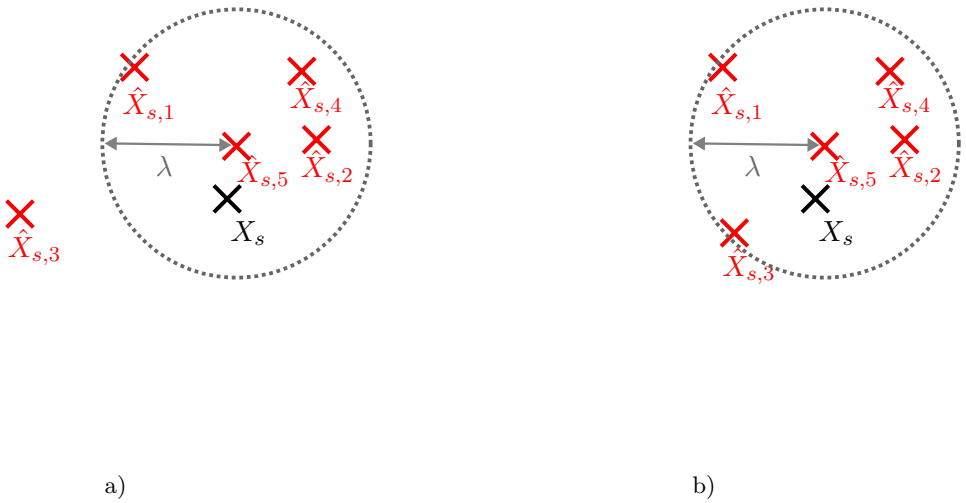


Figure 2.6. Two cases of triangulation. The position of source $\hat{X}_s$ is triangulated 5 times, yielding the estimations $\hat{X}_{s,1}, \hat{X}_{s,2}, \hat{X}_{s,3}, \hat{X}_{s,4}$ and $\hat{X}_{s,5}$ (in chronological order). In (a), the algorithm does not converge, whereas in (b) it converges.

This choice is motivated by the practical application of acoustic imaging in real-world scenarios. In

future developments, acoustic imaging will be paired with an optical device, such as a camera, to verify the defect's origin. Its primary function is to provide a sufficiently accurate estimation of the source's location for the image recognition algorithm to operate effectively. A precision of λ is considered adequate for this purpose. In addition, convergence is also triggered if the last ten triangulated positions are all contained within a circle of radius 5λ. This prevents the robot from being stuck indefinitely around a sound source, at the cost of reduced localization accuracy.
Once the convergence is reached, the algorithm returns the average of the last five triangulated positions and stops. The method for a single source is summarized in Algorithm 1.
The minimal distance to the source is directly related to the step length and equals $\frac{d}{\sqrt{2}}$. Here, the distance of 5λ is chosen for d, which guarantees that the robot circles the sound source at a minimal distance of $\frac{5}{\sqrt{2}}\lambda$. This is considered to be far enough to be situated in the acoustic far field of the source (which maintains model approximation phase errors lower than $9\sqrt{2}$ deg, as defined in Chapter 1), thus the plane wave model holds.

2.2.3 | Simulations results and performance analysis

In this section, several simulations are presented to validate the autonomous SSL strategy. The robustness of the motion planning strategy under uncertainties in DOA estimation is tested, followed by an investigation into how the number of snapshots affects convergence.

Motivations

The development of this simulator for this work was motivated by the complexity of leading experimental works for autonomous SSL. Indeed, localization success depends on numerous parameters, including environmental factors (e.g., source number and positions, background noise, and search area dimensions) and algorithmic settings (e.g., FFT resolution, snapshot count, and path planning criteria). Conducting extensive statistical studies to understand the influence of these parameters requires a large number of experiments, which is impractical in real-world scenarios. Simulations provide a feasible way to perform these studies, offering a controlled environment to explore parameter variations systematically. Additionally, the simulator serves as a platform for algorithm development and optimization.

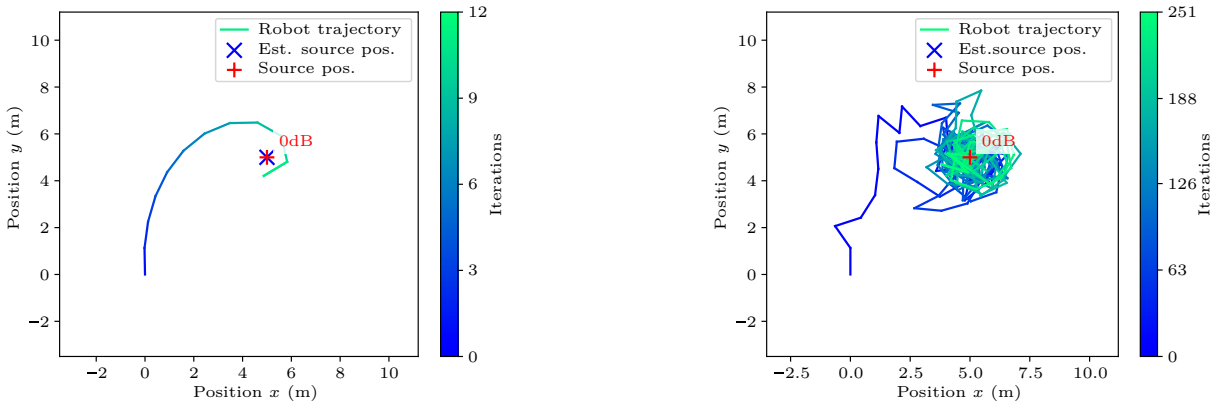
The simulator

The primary objective of the simulator is to assess the performance of the autonomous SSL algorithm under different environmental and signal processing configurations. The performance is evaluated through two scopes: first, the localization accuracy, which is given by the error between the position outputted by the algorithm and the real position of the source, and second, the number of iterations required to attain convergence. The number of iterations is equivalent to the number of acoustic measurements performed. It is thus directly linked to the duration of the autonomous SSL process. In other words, it gives an idea of the time a real-life robot would take to perform this task.
The simulator consists of two modules: an acoustic module, which models the acoustic wave propagation from the sources to the microphones, and a spatial module, which handles the movements of the robot and the triangulation of the source. The modules operate in a loop as outlined in Algorithm 1. To avoid computation of numerous Fast Fourier Transforms (FFT) and handling large amounts of time signals, the acoustic module does not simulate the acoustic time signals unless explicitly required (such as in the following snapshot number study), but directly models the frequency spectrum on each microphone. A spatial acoustic monopole transfer function calculates the spectrum based on the source's and microphones relative positions. Ambient noise is added, CSM is computed, and the DOA

estimation algorithm is applied. The spatial module then triangulates the sound source and updates the robot's position for the next measurement.
The simulator is implemented in Python 3, using libraries such as NumPy and SciPy for numerical computations, and Matplotlib for visualization.
The following sections investigate the impact of uncertainties in the DOA estimation on the convergence algorithm. We propose and validate tuning the number of snapshots in the CSM computation to improve convergence.

Robustness to direction estimation uncertainties and bias

In real cases, acoustic imaging algorithms might be subject to uncertainties in the DOA estimations. Uncertainties on the microphone array's position and orientation are also to be expected, yielding errors on the DOAs. The method's robustness to DOA estimation uncertainties is tested here through a first set of geometric simulations. To spare some computational time, the acoustic module is replaced by a custom direction of arrival estimation here. The real DOA is computed from the current relative positions of the robot and the source. Uncertainties, modeled by a Gaussian distribution of mean μ and standard deviation σ, are added to this real DOA at each iteration to simulate the estimated uncertain DOA. Parameter μ is related to a possible bias introduced by the method employed to move the robot, or by the method to estimate its location and orientation. Parameter σ is related to the uncertainty level. The same experiment was repeated 1000 times for each set of parameters μ and σ. We consider the experiment successful when the method can locate the sound source following the previously introduced convergence criterion, and the localization error is lower than the wavelength λ. Additionally, the simulation is stopped when the iteration counter exceeds 250, which is more than ten times the number of iterations necessary to converge in the noiseless case. Two sample trajectories are available in Fig. 2.7.



a) b)

Figure 2.7. Sample trajectories in (a) the noiseless case and (b) with a zero-mean Gaussian noise of standard deviation $\sigma = 45$ ° on the DOA estimation.

In the case without noise presented in Figure 2.7a, we can see that the robot can circle the sound source and locate it accurately. With DOAs uncertainties, the robot is not able to converge and estimate the position of the source using the proposed convergence criterion, as shown in Figure 2.7b. However, it gets closer to it, suggesting that the bottleneck for the localization of source is caused by the convergence criterion.

Figure 2.8a presents each configuration's success percentage. The proposed strategy proves robust under biases lower than 20 °, with a performance degradation when the DOA estimations become uncertain. These results confirm the importance of a good acoustic imaging method to reduce the

DOA estimation uncertainties, and the performance of the autonomous SSL strategy in an ideal scenario.

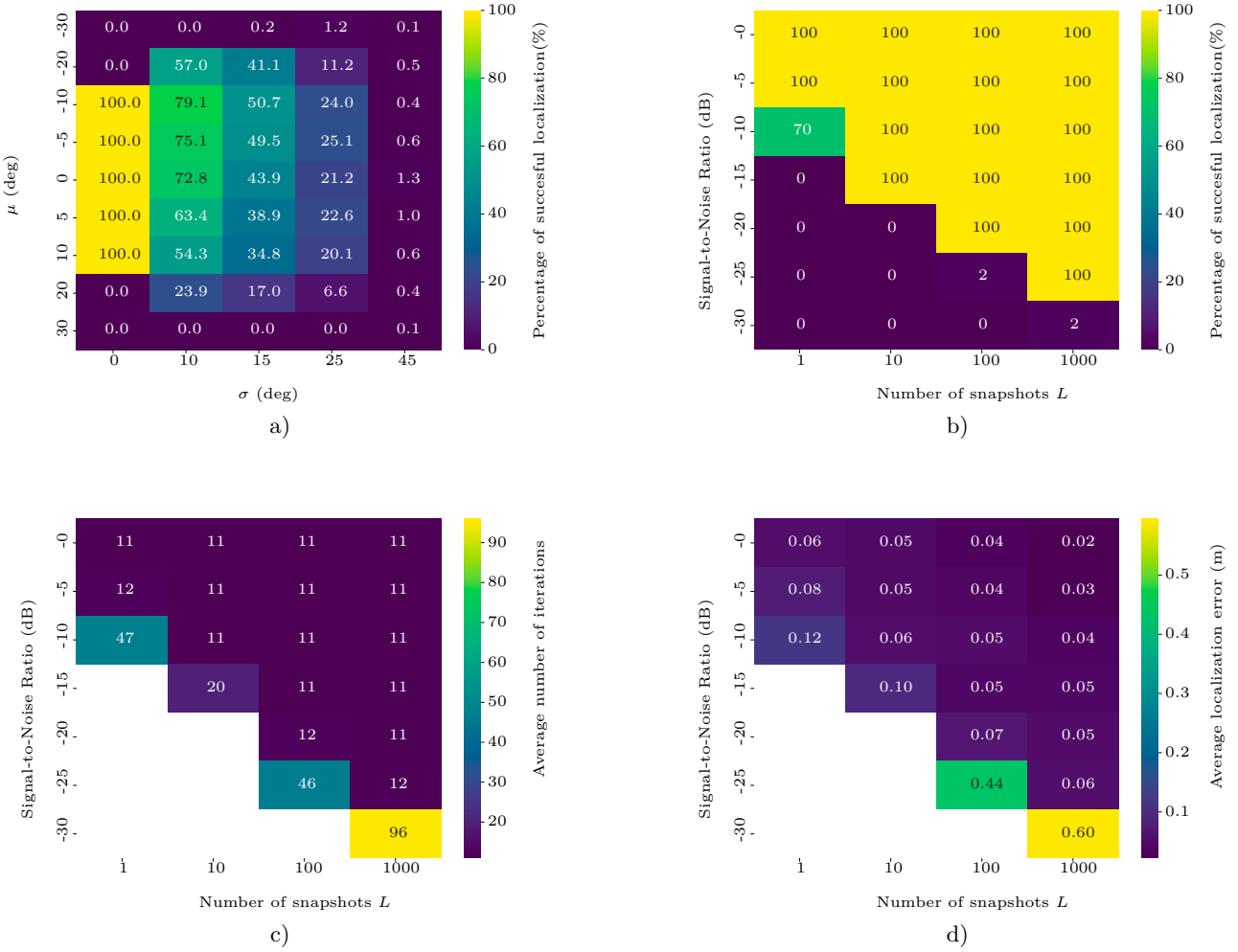


Figure 2.8. Simulation results : (a) Percentage of successful localization when introducing a Gaussian noise of mean μ and standard deviation σ on the DOA estimation, (b) Percentage of successful localization in the presence of ambient noise, (c) Average number of iterations required to localize the source in the presence of ambient noise, (d) Average localization error in the presence of ambient noise.

Influence of the number of snapshots

Background noise, i.e., uncorrelated noise over the microphones, is one of the main sources of uncertainties in real scenarios. A common denoising technique consists of increasing the number L of snapshots in the CSM estimation. Indeed, under the assumption that the background noise is uncorrelated over the microphones, the crossed-terms of the CSM are better estimated when averaged over a higher number of snapshots [67], thus improving SSL.

In the following simulations, an acoustic monopole is considered for the source, and a Gaussian uncorrelated noise is added to each microphone simulated signal. For these specific simulations, the whole time signals are simulated and split into snapshots. Each snapshot yields a CSM, and the CSMs are averaged to obtain the final CSM used in acoustic imaging for DOA estimation. The noise amplitude is defined by the SNR in dB with respect to the sound source level as

$$\mathrm{SNR} = 20\log_{10}\frac{A_S}{A_N} \tag{2.3}$$

with A_S and A_N representing the Root-Mean-Square (RMS) amplitude of signal and noise measured by the robot at the first iteration. After the robot moves towards the source, the SNR thus improves. The SNR varies from 0 dB (the noise received by the microphones is of the same magnitude as the signal received from the source) to -30 dB ($-30 \leq \text{SNR} \leq 0$ in dB). The number of snapshots varies from 1 to 1000. Larger snapshot numbers are not investigated because they would require extremely long acquisitions (or generate excessive overlap), and the CSM estimation would become too computationally demanding for our real-time scenario. The percentage of successful localizations, the number of iterations required to converge, and the localization error are recorded and compared for different scenarios, respectively in Figs. 2.8b–2.8d.

Larger numbers of snapshots are seen to drastically increase the percentage of successful localizations, as well as slightly reducing the localization error. The number of iterations to converge remains constant and around 11 for almost every successful case. These results suggest that increasing the number of snapshots can push the limit of the method in terms of SNR.

Additionally, increasing the number of snapshots also linearly increases the computational time of the CSM calculations, meaning there is a trade-off to achieve optimal real-time performance. In practical scenarios, additional knowledge of the SNR might be greatly beneficial to design the acquisition parameters adequately and to achieve this trade-off.

In the test cases presented further in this work, the number of snapshots is set as high as possible while maintaining a total computational time shorter than the acquisition time of the acoustic signals, while ensuring overlap remains acceptable.

2.3 | Handling mulitple sound sources via source masking

Algorithm 1 provides a method to identify the position of a single sound source. To enable autonomous SSL of multiple sources, it is now extended to include the capability of ignoring sources that have already been located. Our approach solves this issue by integrating a spatial filter. This section first presents the signal processing theory underlying the spatial filter, then examines its applicability to the proposed real-time approach, and finally validates its performance through experimental results.

2.3.1 | Spatial filtering for source separation

The constrained-MUSIC (C-MUSIC) algorithm [68] is used to filter the measured CSM. Based on directional information on the sound sources to filter, it removes their contributions in the measured CSM. To understand its principle, let $\theta_{\mathbf{c}} \in \mathbb{R}^I$ be the directions of the I sound sources to be filtered. The orthogonal complement $\mathbf{A}_c^{\perp} \in \mathbb{C}^{M \times M-I}$ of the matrix of steering vectors $\mathbf{a}(\theta_{\mathbf{c}}) \in \mathbb{C}^{I \times M}$ is built such as

$$\mathbf{a}(\theta_{\mathbf{c}})\mathbf{A}_c^{\perp} = 0 \in \mathbb{C}^{I \times M-I} \tag{2.4}$$

Thus, for effectively nullifying the contribution from sources in the direction $\theta_{\mathbf{c}}$, the CSM is projected on the orthogonal subspace, such as,

$$\mathbf{C}_c = \mathbf{A}_c^{\perp}\mathbf{A}_c^{\perp H}\mathbf{C}\mathbf{A}_c^{\perp}\mathbf{A}_c^{\perp H} \tag{2.5}$$

This allows to remove contribution from sources in the direction $\theta_{\mathbf{c}}$ from the measured CSM, effectively removing, both the main-lobe and the side-lobes of the associated source in the acoustic image. In the original method, the MUSIC algorithm is applied to the filtered CSM, $\mathbf{C}_c$. We propose to extend the method to Conventional BeamForming (CBF), and we refer to it as Constrained-CBF (C-CBF).

By applying the CBF equation to the filtered CSM, $\mathbf{C_c}$, the C-CBF response writes:

$$Y_{C-CBF}(\theta) = \frac{1}{\|\mathbf{a}(\theta)\|^4}\mathbf{a}(\theta)\mathbf{A}_c^{\perp}\mathbf{A}_c^{\perp H}\mathbf{C_c}\mathbf{A}_c^{\perp}\mathbf{A}_c^{\perp H}\mathbf{a}(\theta)^H \tag{2.6}$$

To interpret equation (2.6) in a more practical way, let us consider a C-CBF filtering a source in direction θ_c and observe the response of this beamformer to a source located in another direction. According to Eq. (2.4), if the source is located in direction θ_c, the output of the beamformer is null. In practice, the sound source to be filtered might not be exactly where it is expected to be, or another sound source might be close to it. The following question thus arises: what happens if the source is not in the direction θ_c? To answer this, we simulate a sound source located in direction θ_s, and observe the response in direction θ_s of the C-CBF filtering in direction θ_c. Since the formulation of C-CBF in Eq. (2.6) only depends on the microphone spatial distribution and the frequency of study, the output of C-CBF is observed as a function of these two parameters.

It consists of 32 microphones arranged on a hemisphere with a diameter of 20 cm. To examine the influence of geometry on the filtered beamforming map, four sub-geometries comprising 4, 8, 16, and 32 microphones were extracted from the original 32-element array presented in Chapter 1. The C-CBF outputs at 2 kHz for each geometry are given in Fig. 2.9. The CSM is computed on a single snapshot, 1 s long measurement without background noise for a 2 kHz monopolar source placed at a distance of 10 m from the array. Additionally, four frequencies were studied with the 32 microphones array: 1 kHz, 2 kHz, 5 kHz, and 10 kHz. The corresponding C-CBF outputs are presented in Fig. 2.10.

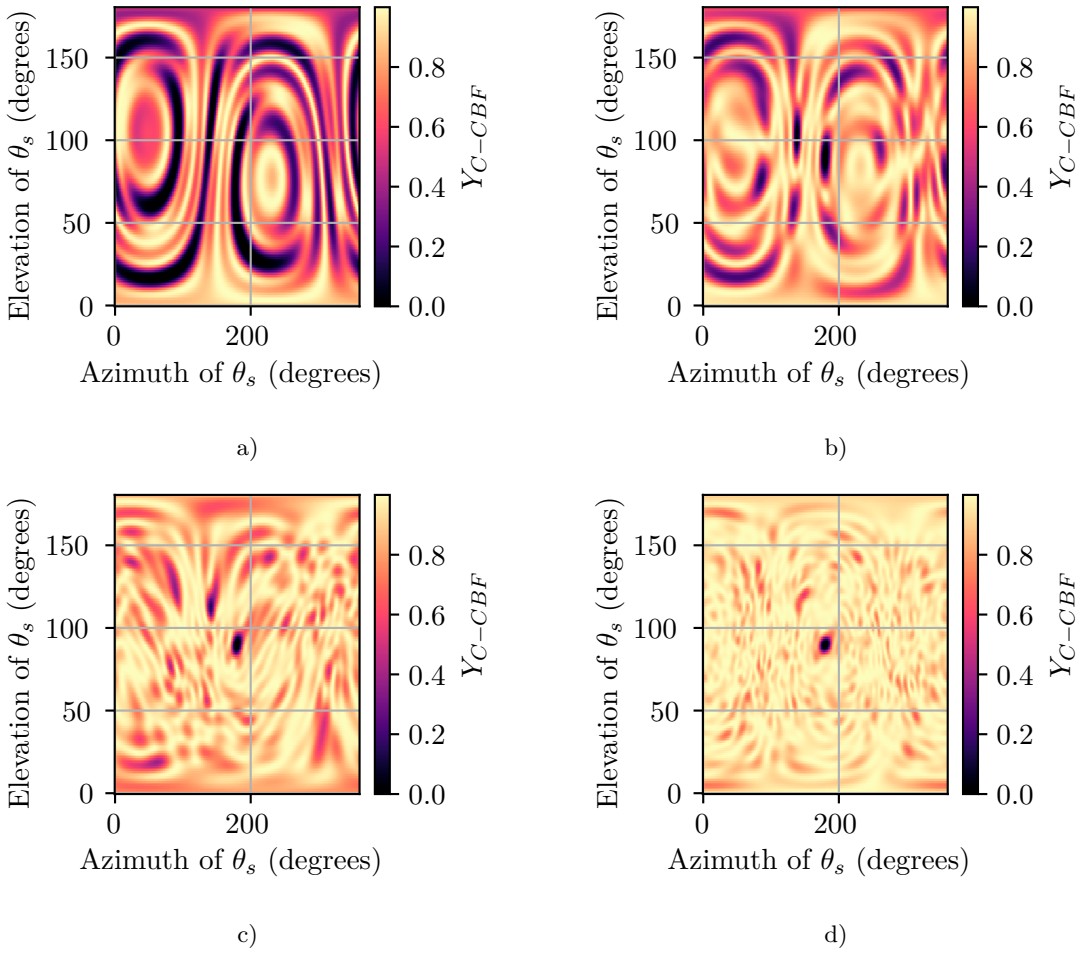


Figure 2.9. Effect of the filter, steered at an azimuth of 180 ° and a coelevation of 90 °, on a sound source located at θ_s for different microphone numbers: (a) 4, (b) 8, (c) 16, and (d) 32.

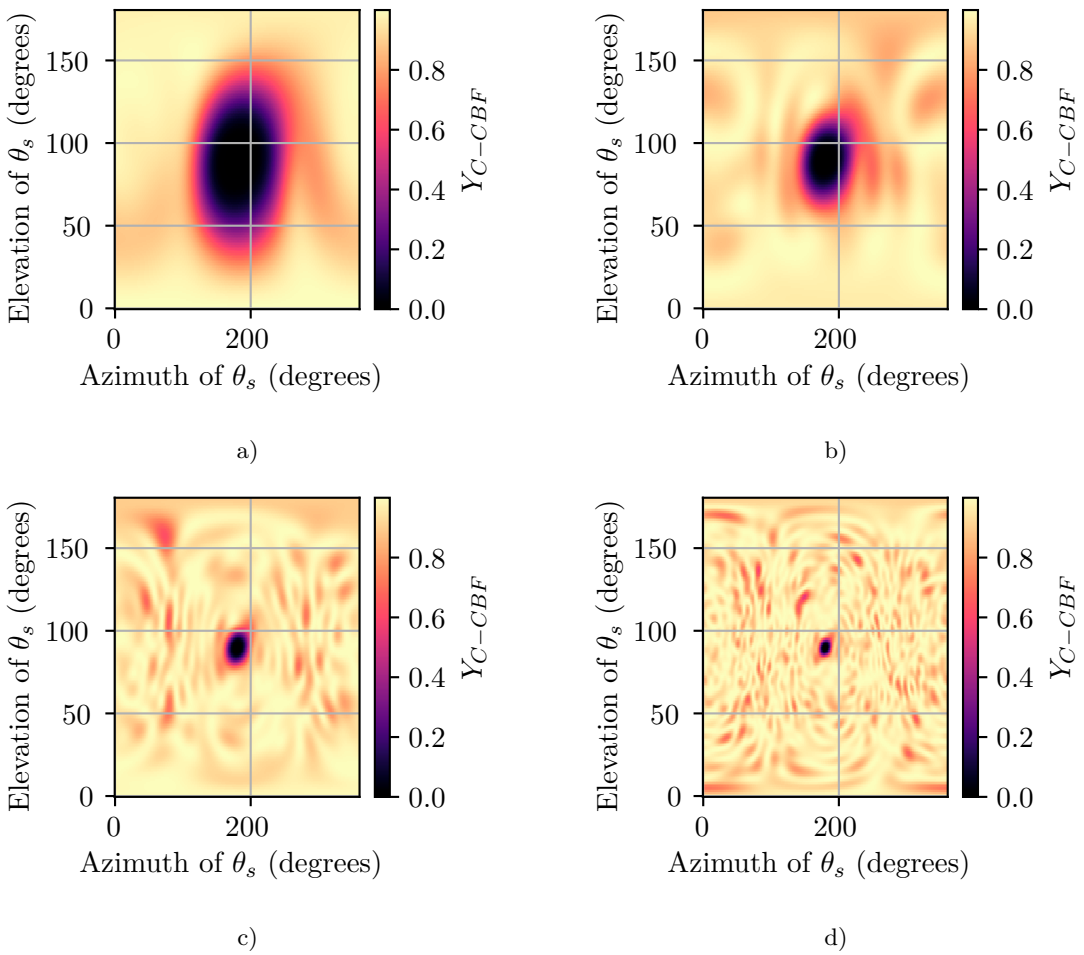


Figure 2.10. Response of C-CBF filtering in direction θ_c (180 ° azimuth and 90 ° coelevation), on a sound source located at θ_s at different frequencies: (a) 1 kHz, (b) 2 kHz, (c) 5 kHz, and (d) 10 kHz.

The source direction θ_s is described by its azimuth and coelevation coordinates. The results demonstrate that at the position of the constraint, the source is correctly filtered, as $Y_{C-CBF}(\theta_\mathbf{c}) = 0$. This is observed across all configurations of the study frequency and the number of microphones. Sound sources close to θ_c are attenuated, with the degree of attenuation depending on both the frequency and the number of microphones. Much like the point-spread function of a microphone array, the C-CBF output exhibits a main lobe and multiple side-lobes, whose width and amplitude vary based on the microphone array geometry and the frequency. The filter's quality improves at higher frequencies and with an increased number of microphones. This observation suggests that optimizing the array geometry could further enhance the filter's performance. It is interesting to note that the main lobe having a non-zero width means that this method is tolerant to some uncertainty regarding the filtered source direction. Overall, the results confirm that the spatial filter is efficient across a broad frequency range for the 32-microphone array developed in this study.

An interesting limitation of this method is the number of constraints applicable to the CSM. Indeed, considering the case where the number of constraints I is equal to the number of microphones in the array M, the rank of the matrix of steering vectors to be filtered $\mathbf{a}(\theta_\mathbf{c}) \in \mathbb{C}^{I\times M}$ is M. Thus, its orthogonal complement is the null matrix, which makes the C-CBF response null in every direction. Further sections will propose a strategy to only use this filter if necessary, thus reducing the number of constraints applied simultaneously.

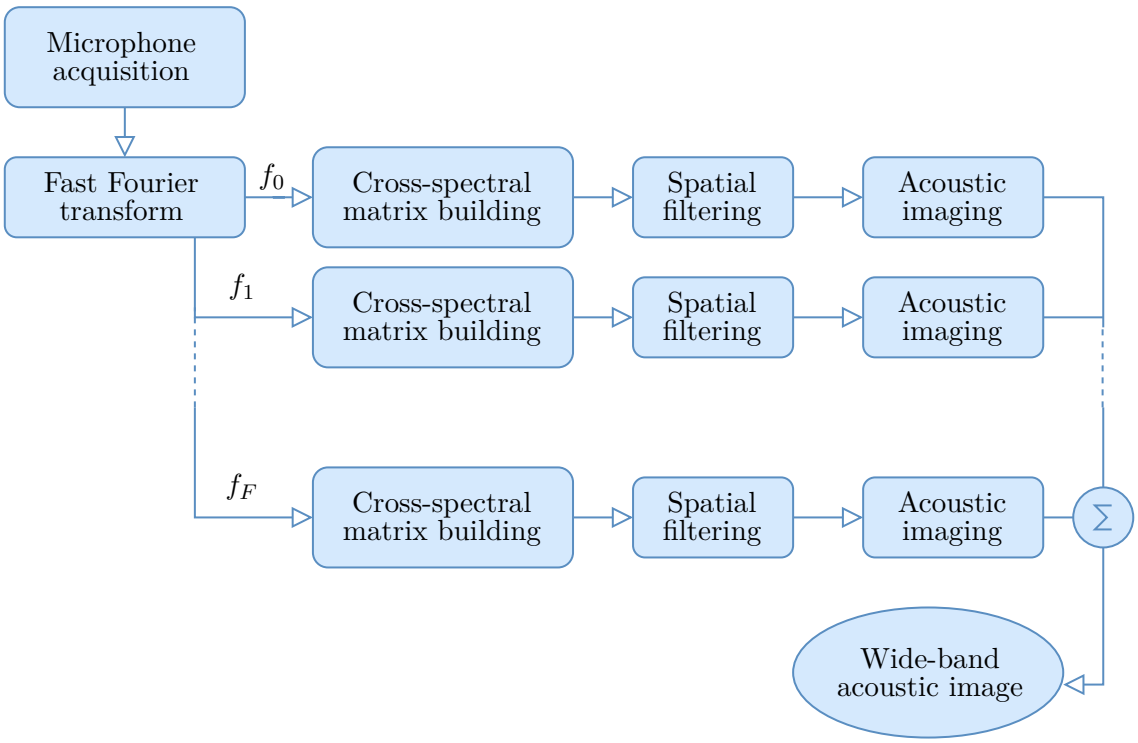


Figure 2.11. Flowchart of the acoustic imaging algorithm, incorporating the spatial filtering step. The process is repeated over n_f frequency bands, and the resulting acoustic maps are summed to produce a wide-band image.

2.3.2 | Integration into the SSL process and computational impact

The spatial filter must be applied to each CSM, which is subsequently transformed into an acoustic map using either CBF or MUSIC, following the computational flowchart illustrated in Fig. 2.11. When computing a wide-band acoustic image across n_f frequency bands, the filter is applied n_f times, once for each CSM. The computational complexity of the algorithms is a critical consideration, as each additional processing step increases the computational load, making real-time processing harder to achieve.

The computation of a vector's orthogonal complement involves a singular value decomposition (SVD), with a complexity that depends on the matrix dimensions, $O(MI^2)$ in this case. The matrix multiplication step has a complexity of $O(M^3)$. The overall computational complexity of the filtering step is $O(M^3 + MI^2)$, depending only on the number of microphones, M, and the number of directions to be filtered, I. To assess the computational load, the filtering step's processing time was compared to the times required for CSM computation and the generation of the corresponding acoustic image under realistic parameters. These parameters included a sampling frequency of 10 kHz, an acquisition duration of 1 s, a frequency resolution of 10 Hz for the FFT, 100 snapshots with 90% overlap for CSM computation, $M = 32$ microphones, $I = 1$ direction to be filtered, and a spatial angular resolution of 5 ° for the angular beamforming map. The evaluation was performed for a single frequency bin.

These parameters resulted in real-time factors of 0.30 and 0.68, for the CBF and MUSIC algorithms, respectively, meaning that the computation time is 30% and 68% of the acquisition time, respectively. Since these values are lower than 100%, the algorithms can run in real-time without causing delays in acquisition. Averaged results over 1,000 runs are presented in Table 2.1. The spatial filter was implemented in Python 3, and the computations were performed on a laptop with an Intel(R) Core(TM) i7-10750H CPU and 32 GB of RAM.

The proportion of time dedicated to the spatial filtering step appears negligible to the overall SSL process. This validates the suitability of incorporating the filter for real-time applications.

Algorithm	Computation time (ms)	% of total time (using CBF)	% of total time (using MUSIC)
CSM	250.1	82.8%	36.9%
Spatial filter	1.1	0.4%	0.2%
CBF imaging	50.7	16.8%	
MUSIC imaging	426.0		62.9%

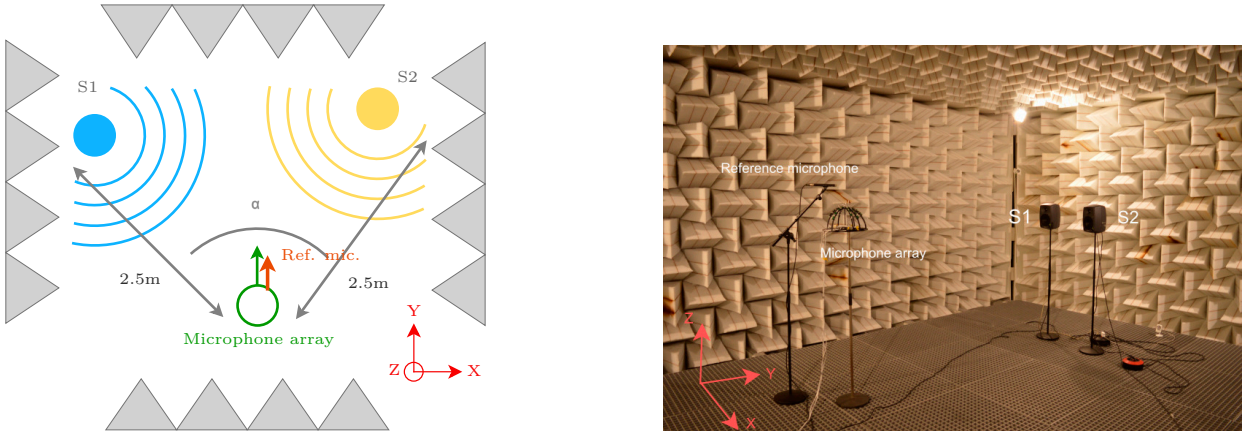
Table 2.1. Computational times of the different algorithmic steps (averaged over 1,000 runs).

2.3.3 | Experimental validation

To validate the spatial filter, an experiment was conducted in the anechoic room at the Institut ∂'Alembert. This experiment aimed to create an experimental dataset to evaluate the algorithm's performance in finding the direction of a sound Source Of Interest (SOI) while filtering an interfering sound source.

Experimental setup

Two identical sound sources are positioned at 2.5 m from the microphone array in the free-field environment of the anechoic room. A reference condenser microphone (DPA4006A) is placed next to the microphone array. The experimental configuration is illustrated in Fig. 2.12.



a) b)

Figure 2.12. (a) Diagram of the experimental setup. (b) Picture of the experimental setup.

Parameters studied

The two following parameters are studied to explore the spatial filter's limitations:
Source-to-Interference Ratio (SIR): The ratio of Root-Mean-Square (RMS) Sound Pressure Level (SPL) between the SOI and the interfering source:

$$A_{2/1dB} = 20\log_{10}(\frac{A_2}{A_1}) \tag{2.7}$$

with A_1, the SPL of the interfering sound source in Pa, measured at the array position, and A_2, the SPL of the SOI in Pa, also measured at the array position. A wide range of SIR is tested, from +5 dB to -30 dB.

Angular distance between sources : the angle between the two sound sources. It varies from 0 to 90 °, with a step of 10 °.

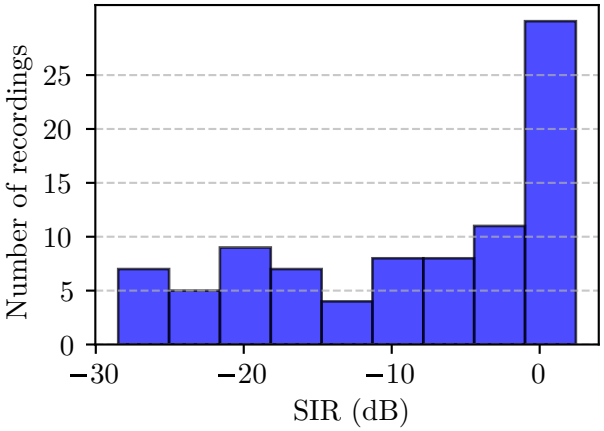


Figure 2.13. Distributions of SIR values.

The 32 MEMS microphone Megamicros device records 10 s sound pressure measurements at 50 kHz across all microphones for each test case. To complement these test cases, individual recordings of each sound source are taken for all sound levels and angular positions. The reference microphone signal is also recorded for each Megamicros dataset at a sampling frequency of 44.1 kHz. Sources emit 2 kHz sine signals of fixed amplitude throughout the recording. The background noise was measured and found to be below 25 dB. The lowest level emitted by a sound source in this experiment was 65 dB, which is large enough to consider the background noise negligible.

Processing

Four algorithms are tested: CBF and MUSIC, serving as reference methods, along with their spatial filter-enhanced counterparts, C-CBF and C-MUSIC. The spatial filter's purpose is to recover the position of source S2 given the known position of source S1. To evaluate the filter's performance, the localization error $|\hat{\theta}_2 - \theta_2|$, with $\hat{\theta}_2$ the estimated direction of the SOI, and θ_2 its real direction, is normalized by the angular distance between the sources. This normalized localization error is defined as $\epsilon_\theta = \frac{|\hat{\theta}_2 - \theta_2|}{|\theta_1 - \theta_2|}$, with θ_1, the direction of the interfering source. This approach allows for meaningful comparisons across different angular distance scenarios. For instance, a 10 ° localization error is acceptable for a 90 ° angular separation but unacceptable for a 20 ° separation. In the experiment, the sound sources are manually positioned. To measure the sound source positions, CBF is applied to the single-source recordings, and the direction is estimated from the beamformer output. The SIR values are calculated from single-source reference recordings. A total of 88 recordings are analyzed, and the resulting SIR distributions are shown in Fig. 2.13.

Although the distributions are not uniform, they span values from -28 to +4 dB in sufficient numbers to produce statistically meaningful results in each of the 5 dB SIR bands. After examining the first results, the recordings with an angular distance lower than or equal to 20 ° are discarded because none of the algorithms could achieve good results in those cases. For each microphone array recording, all the data processing is performed on a 1 s signal extracted from the measured data. The parameters for the CSM computation are the following: a sampling frequency of 50 kHz, a frequency resolution of 10 Hz, and 100 snapshots with 90% overlap. Acoustic images are produced over a 2D angular grid of fineness 5 °.

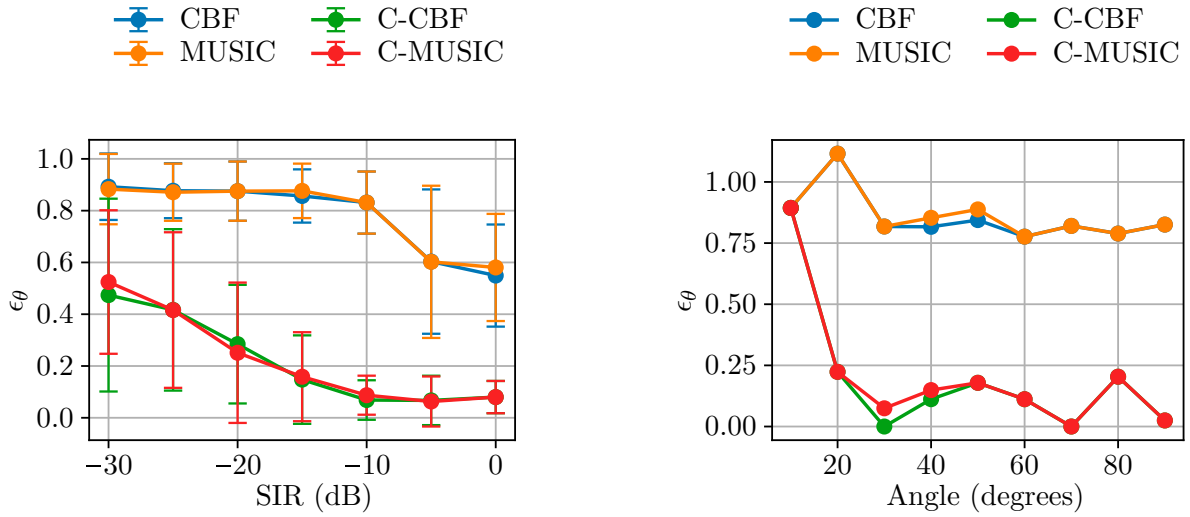


Figure 2.14. (a) Algorithms' performances averaged over angular distances larger than or equal to 20 °. The standard deviations of the normalized localization errors are plotted as bar plots. (b) Algorithms' performances averaged over SIR comprised between -15 and -5 dB.

Results

First, the limitations of the filter in terms of SNR are investigated. Figure 2.14a shows the normalized localization error ϵ_θ for all measurements of angular distance greater than 20 °, averaged over 5 dB SIR bands. A clear trend can be observed: performances and uncertainties degrade when SIR values drop below -15 dB. This suggests that the filter is efficient and reliable for negative SIR, down to -10 dB. It is also clear that C-CBF and C-MUSIC share very similar performances in this experimental setting.

The spatial filter's limitations on angular distance between sources are observable in Fig. 2.14b. Angular errors of less than 20% are considered acceptable. The filter satisfies this condition for sources spaced more than 20 °. The question of closely spaced sound sources will be investigated later using the simulator introduced in Section 2.2.3. No explanation was found regarding the high error observed for an angular distance of 80 °. The post-process is rigorously identical for each measurement, suggesting this outlier is due to a problem in the data acquisition or the experimental setup.

In the following, CBF is preferred to MUSIC for our real-time application because of its lower computational cost. However, in outdoor scenarios, MUSIC is expected to outperform CBF in terms of robustness to noise. Additionally, with optimized implementation, MUSIC has been proven suitable for real-time use [44].

2.4 | Path planning for autonomous multiple sound sources localization

Being able to ignore sound sources whose position is already known is not sufficient to perform multiple sources autonomous SSL. This section tackles some issues arising from the presence of multiple sound sources and the use of the previously introduced spatial filter. The single-source case path planning algorithm is adapted and validated through simulations.

2.4.1 | Simulation framework for multiple sound sources

The simulator introduced in Section 2.2.3 is adapted for multiple sources scenarios. An illustration of typical outputs of the simulator for the multi-source case is given in Fig. 2.15. These simulations include two sound sources of random one-meter SPL, located randomly within a 10 m by 10 m square area. The one-meter SPL of sources is given in dB with a reference pressure of 2×10^{-5} Pa. They provide insights into the robot's behavior. Under the hypothesis of free-field and omnidirectional sources, the SPL of a source at position X_s measured by the robot at position P, SPL_{P,X_s} can be expressed as

$$\mathrm{SPL}_{P,X_s} = \mathrm{SPL}_{1m,X_s} - \log_{10} \|X_s - P\| \tag{2.8}$$

with SPL_{1m,X_s} the one-meter SPL of the source, and $\|X_s - P\|$ the distance between the robot and the source positions P and X_s. At a given iteration, the robot will orient itself toward the source located at position X_s that yields the highest SPL. Since the robot position changes from one iteration to another, SPL_{P,X_s} also changes dynamically, meaning the robot can switch focus when passing close to a source. To make the interpretation clearer, the value of SPL_{1m,X_s} is normalized by the largest one-meter SPL of the sources, meaning the most powerful source has a one-meter SPL of 0 dB.
The frequency of study is 2 kHz with an FFT resolution of 10 Hz, the number of snapshots to compute the CSM is 100, the background noise is null, the robot's step is 5λ, and the microphone array geometry is the one introduced in Chapter 1, which is composed of 32 microphones placed on a hemisphere.
The simulated robot's behavior aligns well with our expectations. It approaches a sound source in a spiral pattern, circling it until the estimated position stabilizes. The switch in focus due to the dynamic evolution of the source SPL measured on the array, SPL_{P,X_s}, is illustrated in Fig. 2.15c. The robot initially moves towards the 0 dB sound source. However, as it progresses to the right, the -8 dB source's SPL_{P,X_s} becomes larger, eventually drawing the robot to converge around it. This behavior is a key feature of the method, enabling the robot to locate lower-level sound sources that might otherwise be overlooked. It effectively serves as a way to increase the exploratory capabilities of the robot. Since the trajectory is updated solely based on the previous estimation, the robot remains "adaptable," allowing it to be influenced by its surroundings and explore multiple sound sources effectively.

2.4.2 | Conditional spatial filtering approach

To localize multiple sound sources ($\mathcal{V}$) autonomously, we propose extending the single-source autonomous SSL algorithm with a path-planning strategy that systematically identifies the most dominant sound source until it is localized. Once localized, the source is masked using the spatial filter introduced in Section 2.3.1, allowing the algorithm to focus on the next dominant source. This process is repeated until all source positions are recovered.

However, directional filtering can inadvertently mask other sources of interest, as shown in Fig. 2.16, where the robot fails to localize the SOI when the two sources align with the robot. This issue also arises in simulation, as depicted in Fig. 2.17, where the robot misses the middle source due to unintended filtering. This behavior reduces convergence speed by introducing errors in the estimated source direction.
To address this, we propose to apply the spatial filter only when necessary. When the robot approaches a new source, that source typically becomes dominant due to its proximity, making additional filters unnecessary. At each step, the sound levels of known and focused sources are compared. If the sound level of a known source is lower than the focused source, its filter is disabled. The SPL of the focused source is estimated using a CBF response, as given by:

$$Y_{CBF}(\theta_{\mathbf{p}}) = \frac{\mathbf{a}(\theta_{\mathbf{p}})\mathbf{C}\mathbf{a}(\theta_p)^H}{\|\mathbf{a}(\theta_{\mathbf{p}})\|^4}$$

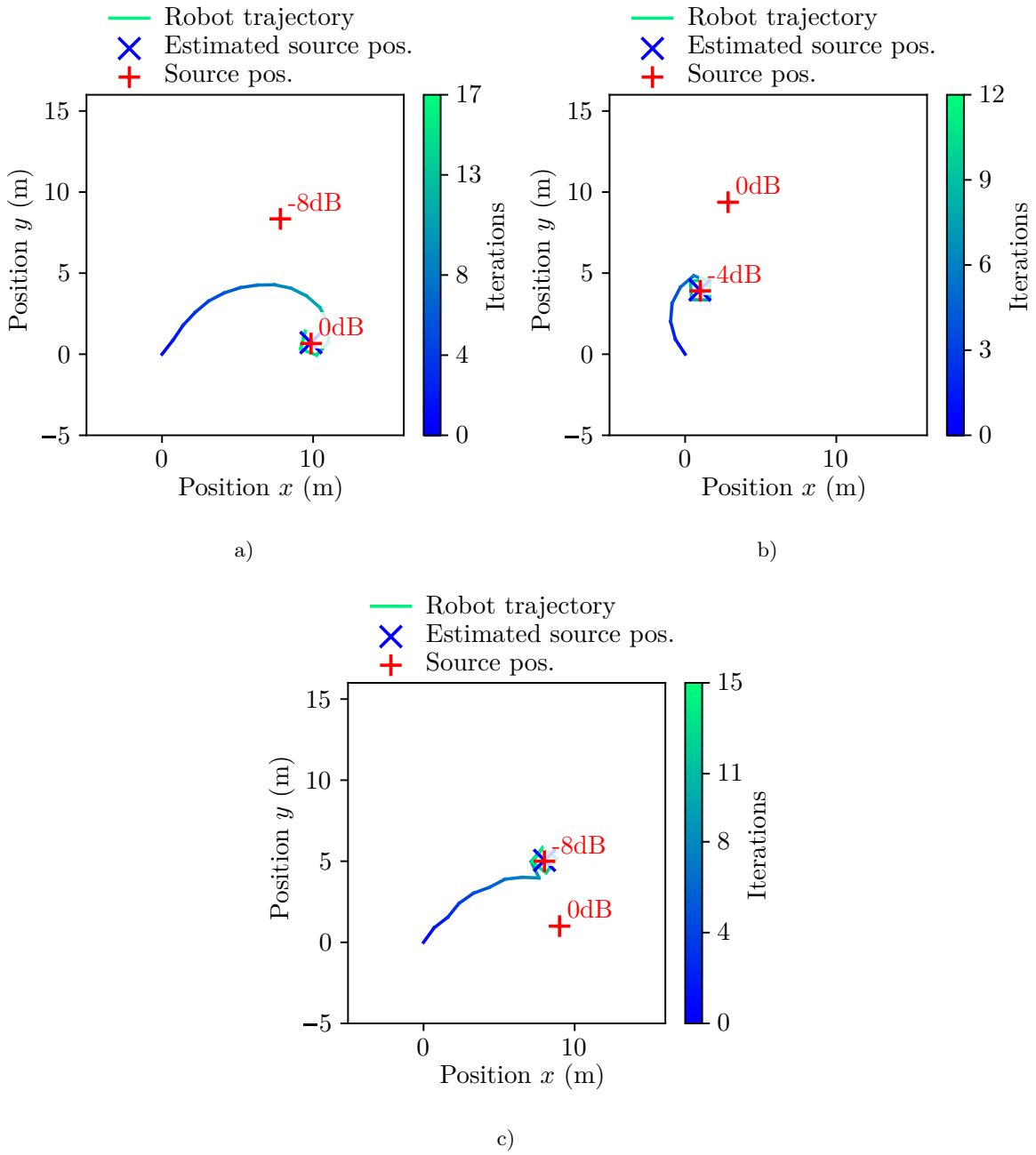


Figure 2.15. Simulated trajectories of the robot without background noise and in the presence of two sound sources of respective sound pressure levels at 1 m (a) -8 dB and 0 dB, (b) 0 dB and -4 dB, and (c) 0 dB and -8 dB.

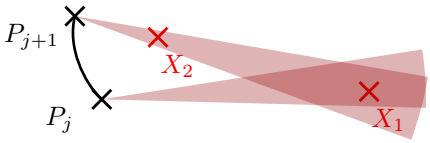


Figure 2.16. The spatial filtering problem in the multi-source case. The masked areas at each iteration are represented by the red triangles. At iteration $j+1$, the spatial filter masks both the already located source position, X_1, and the targeted source position, X_2.

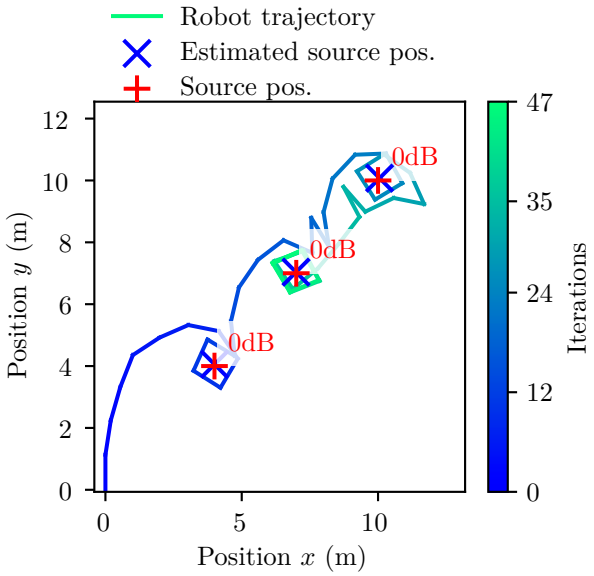


Figure 2.17. Simulated trajectory of the robot in a case where the spatial filter unintentionally masks an unknown source. No background noise was added.

where θ_p is the candidate source's direction, $\mathbf{C}$ is the measured CSM, and $\mathbf{a}(\theta_p)$ is the steering vector. Since the position of the known source is available, its SPL can be estimated using its one-meter SPL, which is obtained once the source is located. This strategy is summarized in Algorithm 2.

Algorithm 2 Command strategy for the multiple sources case

```
while V sources are not located do
    while the positions of the focused source have not converged do
        Move one step 45 ° to the left of the last DOA
        Estimate the sound levels of known and focused sources
        if the level of a source is greater than that of the focused source then
            Apply the spatial filter for that source
        else
            Disable the spatial filter for that source
        end if
        Triangulate the focused source's position
        Check for convergence
    end while
    Store the final position of the source as the average of the last 5 estimated positions
    Increase the source counter
end while
return the final positions of all sources
```

2.4.3 | Validation of the conditional filtering approach

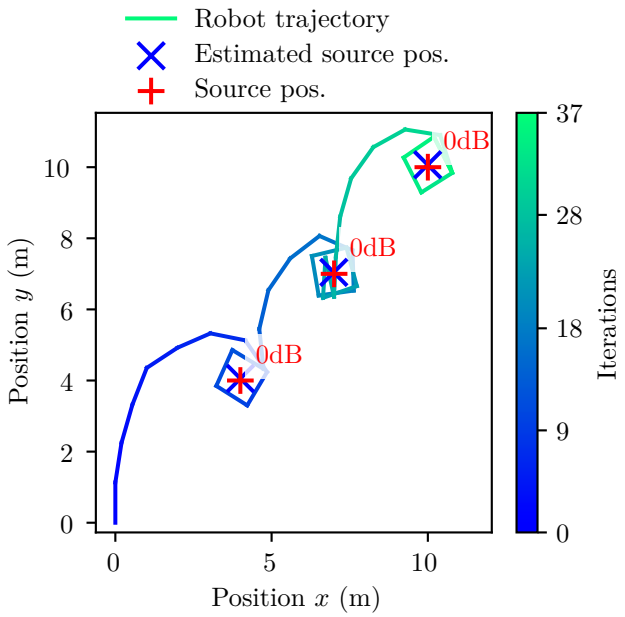


Figure 2.18. Simulated trajectory of the robot in a tricky case after implementation of the conditional activation of the spatial filter. The robot localizes efficiently the 3 sources. No background noise was added.

The previous problematic case from Fig. 2.17 is now tested after implementing Algorithm 2. The trajectory of the robot is given in Fig. 2.18. In this case, the robot efficiently localizes all three sources in 37 iterations compared to 47 iterations without conditional filtering. Even in this simple case, this represents a significant improvement. It is important to underline that the addition of the conditional

activation of the spatial filter does not add significant computational cost to the Algorithm 1, since it can be estimated as the 1m-SPL divided by the distance to the source from Eq. (2.9). They are both trivial operations, and their cost is negligible compared to CSM building or acoustic map computation.

To validate the proposed methodology, several simulations were carried out, firstly for the two sound sources case, then for a larger number of sources and finally, to assess the strategy's performance emulating real scenarios.

2.4.4 | Challenging scenarios

Closely spaced sources

Localizing two closely spaced sound sources is known to be a challenging task in acoustic imaging. This section investigates the performance of the proposed autonomous SSL strategy in the case of closely spaced sources. To do so, we used the simulator from Section 2.2.3 under the proposed following test scenario. The idea is to place two sources close to each other, separated by a distance Δ_S, and a third one far enough so that the robot can locate it without being disturbed by the two other sources, but only after locating and filtering the first two. Three sound sources are simulated in a square area of width $l = 10d$, with d the robot movement step. The first source position is generated randomly in an area close to the robot starting position. The second source is placed at a distance Δ_S from the first source in a random direction. Finally, the third source is placed so that the first source is exactly in the middle between the robot's starting position and the third source. Since all sources have equal one-meter SPL, and due to the geometric configuration of the sources, the robot is forced to locate the closely spaced sources first before moving on to the third one. The configuration of sources is illustrated in Fig. 2.19.

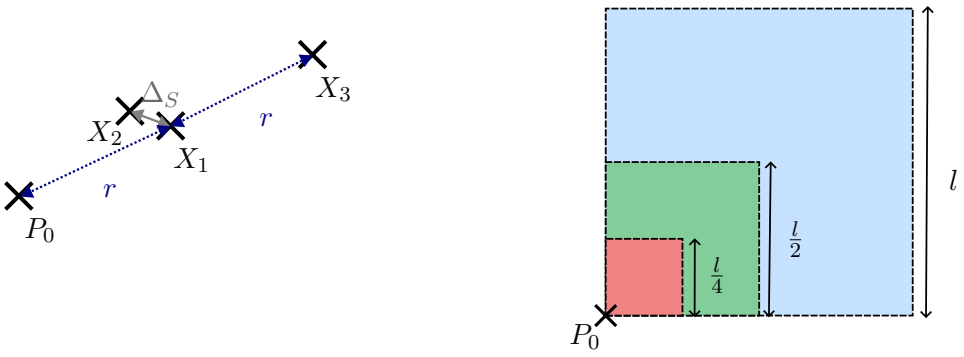


Figure 2.19. Configuration of the simulated experiment. (a) Two sources at positions X_1 and X_2 are placed close, separated by a distance Δ_S. A third source at position X_3 is placed further away. The goal is to verify that the robot, starting from position P_0 will not be stuck around X_1 and X_2. (b) Generation of sound source positions. X_1 is generated within the green area, X_2 is placed at a distance of Δ_S around X_1, and X_3 is generated at twice the distance between the robot initial position, P_0, and X_1, ending up in the blue area.

For each run, the different outcomes listed in Table 2.2 are possible. Several values of Δ_S were tested, and the experiment was repeated 100 times for each. The step of the robot in this experiment

Case	Success/Fail	Interpretation
1) S_1, S_2, and S_3 are found	Success	The sources are far enough for the algorithm to locate each of them individually
2) S_1 or S_2, and S_3 are found	Success	S_1 and S_2 appear as one single source for the robot
3) S_1 and/or S_2 are found, but not S_3	Fail	The robot struggled to filter S_1 and S_2 and could not escape their vicinity after locating them
4) Only S_3 is found	Fail	Based on the experiment design, this case should not happen
5) None of the sources are found	Fail	The robot stuck around S_1 and S_2 and could not locate either of them, indicating the convergence criterion is too strict

Table 2.2. Possible outcomes of the closely-spaced sound sources experiment, and their interpretation.

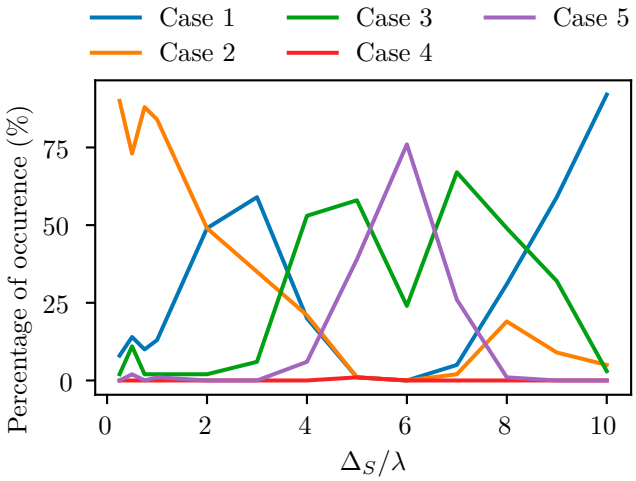


Figure 2.20. Occurrences of the different outcome scenarios when sources are placed close to each other.

was $d = 5\lambda$. Results are presented in Fig. 2.20. A clear trend appears: when $\Delta_S \leq \lambda$, case 2 dominates, indicating that the algorithm can consider two sources as a unique one. As $\Delta_S \gg \lambda$, case 1 becomes predominant, meaning the robot does not struggle to locate each source individually. When $\Delta_S \in [3\lambda, 9\lambda]$ (which can be interpreted as $\Delta_S \approx d$), problematic cases occur in most of the simulations, showing a limitation of the current approach.

These issues could likely be overcome by implementing an adaptive step strategy, for example, by increasing the step if the robot is seen to spend too many iterations in a specific location of the search area. These considerations are left for future works, and the distance between sources will be controlled in the rest of this study to avoid facing these known limitations.

2.4.5 | Autonomous localization of multiple sources in simulations

We have extended our results to the multi-source case, intending to validate the method's capabilities to robustly localize. After reviewing preliminary simulation results, it was observed that, in some cases, the robot localizes the same sound source multiple times. As a result, it misses other sound sources by stopping the simulation prematurely. It can happen if the source is located with high uncertainty, which in turn deteriorates the efficiency of the spatial filter. To solve this issue, an additional condition

was added to consider a sound source is found: its position must be distant from at least 2λ from a previously located source. As seen in the previous section, the algorithm does not separate sound sources spaced closer than 2λ, hence this choice. In practice, the robot might localize a source multiple times, adding filters steered toward each of the positions estimated, but not incrementing the number of found sources. After adding several filters, the robot is able to locate another source and escape the vicinity of the previous one. This is coherent with the understanding of the filter's behavior, as developed in Section 2.3.2. Although a few more iterations are spent around the source, this solution solves the problem efficiently. This behavior is illustrated in Fig. 2.21, where the robot localized the first source 3 times before reaching the second one.

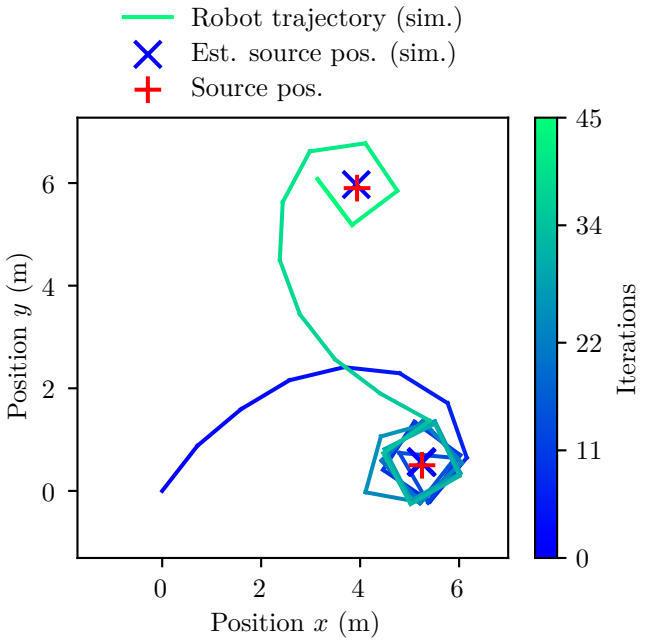


Figure 2.21. Simulated trajectory of the robot. In this case, the robot found the first source but struggled to filter it efficiently. Consequently, it kept circling it until finding it two more times, adding a second and a third filter. After these filters are added, the robot escapes the source's vicinity and focuses on another source. Both sources have the same one-meter SPL, and no background noise was introduced.

In order to validate the proposed methodology on multiple sound source cases, we considered one hundred simulations with different SNR values, following the definition given in Eq. (2.3). The simulated scenario is the following: five sources are randomly positioned in a 10 m square search area. An ambient uncorrelated noise of amplitude N is added to the simulation. The first source has a one-meter SPL set to 0 dB. Each other source has a random one-meter SPL, ranging from 0 to -10 dB. The sources are distant from at least two steps, d, from each other, otherwise, they would be detected as only one source by our algorithm. The simulation stops after $K = 750$ iterations, as it is assumed that increasing the number of iterations will not yield more results. All the simulations use $L = 100$ snapshots for the CSM computation, a sampling frequency of 10 Hz, a frequency of study of $f = 2000$ Hz, and a step $d = \frac{5}{\lambda}$. Table 2.3 shows the results, averaged for the 100 tries. In some cases, the algorithm finds some sound sources where there is nothing, so we consider a sound source is found when the localization error ϵ is $\epsilon < 2\lambda$. Overall, the localization is accurate for ambient noise lower or equivalent to 5 dB. As expected, increasing ambient noise increases the number of iterations necessary to converge. In each tested case, $\epsilon < \lambda$. The behavior of the robot is depicted in a sample case where localization is successful in Fig. 2.22.

N (dB)	Averaged K	Averaged ϵ (m)	Sources found (%)
- ∞	115.3	0.06	96.6
0	169.6	0.06	95.7
5	197.7	0.05	96.2
10	264.52	0.05	79.8

Table 2.3. Statistical performances of the algorithm for 5 sources at 2 kHz, averaged over 100 tries for each noise amplitude value.

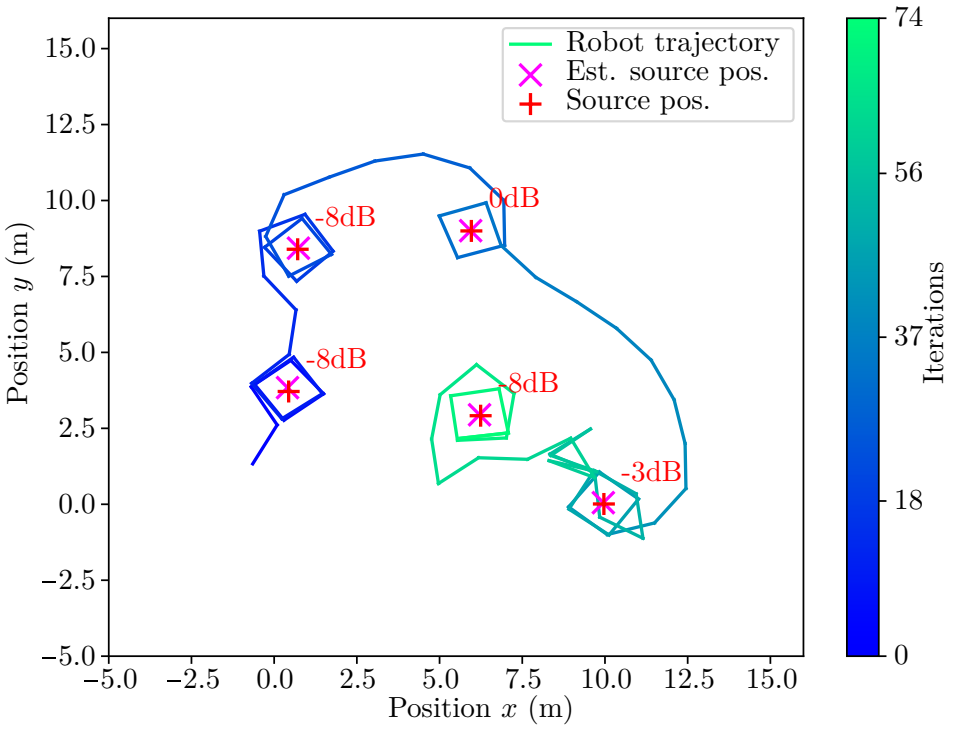


Figure 2.22. Sample trajectory of the robot in the noiseless case, for a random distribution of source positions and sound levels.

This simulation study demonstrates that autonomous SSL in the multiple sources scenario is possible using the proposed strategy. Moreover, it is robust to noisy DOA estimations, as well as random distributions of positions and sound levels of sources.

2.5 | Discussions

This chapter introduced a novel autonomous SSL methodology designed to locate multiple omnidirectional, stationary, and immobile sound sources of varying amplitudes and spatial distributions within an acoustic free-field environment. In this section, we discuss the assumptions underlying the method, its current limitations, and potential directions for future work.

A key assumption of the proposed method for stopping the algorithm, is that the number of sound sources to be localized, denoted as $\mathcal{V}$, is known in advance. In this way, the algorithm continuously searches for the remaining source positions until all have been successfully located, as outlined in Algorithm 2. Nevertheless, this stopping criterion can be replaced with an alternative criterion based

on the sound pressure level (SPL) measured by the array. To ensure robust DOA estimation, the SPL of each source must exceed a critical threshold. By knowing the lowest detectable SPL, the minimum one-meter SPL of the sources of interest, and the critical detectability distance, the robot can adjust its path based on the current acoustic image, ensuring that it covers regions where sources of sufficient SPL are expected to be present. This path could be structured as a grid, with its resolution determined by the critical detectability distance.

The method also assumes that the sound sources are stationary and immobile. While this assumption may hold in certain industrial applications, such as detecting gas leaks or machinery faults, it would be beneficial to extend the approach to handle moving and/or intermittent sound sources. Intermittent sources could be addressed using specialized triangulation methods that are more resilient to outlier DOA estimations. Previous work by Nguyen *et al.* [20], demonstrated the effectiveness of Kalman filtering for tracking intermittent sources in a single-source scenario, but the multi-source case remains unexplored. For moving sources, a promising solution could involve the use of particle filtering, as described by Yen [21], where multiple mobile arrays are employed to triangulate source positions in real-time. These extensions offer opportunities for enhancing the robustness of the proposed framework.

Another assumption made is the use of the plane wave model for the sound waves measured by the microphone array. While this model proved effective in the experiments conducted, it may not be suitable for large-scale real-world sources, particularly when their size exceeds the wavelength λ. One potential solution is to leverage the robot's mobility. By positioning the robot at a sufficient distance, the large source could be approximated as a monopole, which would allow the system to bypass this limitation. This adaptation could be incorporated into the path-planning algorithm, making it more flexible for real-world deployment.

2.6 | Conclusion

This chapter addressed the key challenges involved in developing an autonomous SSL method. A path-planning algorithm was proposed to facilitate triangulation based on successive DOA estimations, enabling the iterative localization of multiple sound sources. The performance of this algorithm was evaluated using a spatial filter, and several simulation studies were conducted to validate the method in controlled environments. These studies also helped identify and resolve various conceptual issues, anticipating the implementation on a mobile robot in real-world scenarios.

Summary tables

Table 2.4 provides a detailed overview of the main accomplishments discussed in this chapter. Additionally, Table 2.5 provides an overview of the perspectives and future directions for enhancing the framework. This includes potential improvements to handle moving and intermittent sources, addressing model limitations, and exploring real-world applications for large-scale deployments.

Aspect	Achievement
Path-planning algorithm	Developed an iterative path-planning algorithm for triangulation based on DOA estimations.
Sound source localization	Introduced a novel approach for locating multiple stationary, omnidirectional sound sources in an acoustic free-field.
Spatial filtering	Implemented and validated a spatial filter to tackle iterative multiple sources detection and localization.
Simulation validation	Validated the methodology through simulation studies.

Table 2.4. Summary of Key Achievements in Autonomous SSL Methodology

Aspect	Perspective
Intermittent sources	Extend the approach to handle intermittent sound sources using methods like Kalman or particle filtering for multi-source tracking.
Moving sources	Investigate synchronous triangulation techniques using multiple autonomous robots.
Plane wave model limitations	Adapt the path-planning algorithm to accommodate large sources.
Real-world deployments	Investigate the performance of the system in large-scale or outdoor environments, where real-time sensor feedback and robust filtering are required.

Table 2.5. Perspectives and Future Directions for Autonomous SSL

Building on this foundation, the following chapter focuses on the real-world implementation of the proposed autonomous SSL framework, emphasizing experimental validation, continuous real-time operation, and the challenges of autonomous localization in unknown environments.

3. Ground robot for sound source localization

Contents

3.1 Experimental platform 57
3.1.1 Environment setup 57
3.1.2 Ground robot description 58
3.1.3 Communication architecture 61
3.2 Experimental results 62
3.2.1 Simulation and experiment comparison 63
3.2.2 Real-time autonomous localization of multiple sources 63
3.3 Achieving continuous real-time localization 65
3.4 Discussions 68
3.5 Conclusion 70

The methodology introduced in Chapter 2 is tested using a dedicated experimental setup, demonstrating its real-time efficiency in a series of practical scenarios.
This chapter contains three main sections: Section 3.1 introduces the experimental platform developed during this work, Section 3.2 presents the tests performed to validate the autonomous Sound Source Localization (SSL) methodology, and Section 3.3 pushes the method to its limits to reach continuous and real-time SSL.

3.1 | Experimental platform

This section presents the platform developed and tested in the Heudiasyc laboratory and the computation architecture proposed to achieve real-time performances.
The goals of this platform are: (a) to program the proposed algorithm into the microphone array to estimate in real-time the acoustic source Direction Of Arrival (DOA) direction, (b) to use the estimated information as commands in the control algorithm of the ground vehicle to move it autonomously using the strategy described in Algorithm 2, and (c) to validate, in real time, the performance of the autonomous SSL strategy using the mobile robot.

3.1.1 | Environment setup

The experimental validation was conducted indoors, within the test area of the Heudiasyc laboratory. The flight area offers a convenient playground to test or validate novel robotic algorithms on both ground and aerial vehicles. Figure 3.1 gives a picture of the test area.



Figure 3.1. The robotic test area of the Heudiasyc laboratory.

Motion capture

This room is equipped with an array of 48 infrared cameras. Using a motion capture software (Motive, OptiTrack[1]), it is possible to track optical markers in the room with submillimetric precision at a frequency of 100 Hz. By attaching multiple of these markers to an object, its position and orientation can be monitored. This information is broadcast in real-time via the Robot Operating System (ROS) software via Wi-Fi, making it accessible to any computer connected to the Wi-Fi network. This provides precise and robust information streaming, essential to the control algorithms introduced further.

Acoustic considerations

The background noise sound level was measured using a reference electret microphone at 56 dB (wideband) and 21 dB (on the 2 kHz third-octave band). The reverberation time T_{20dB} was measured at 0.85 s (both wideband and on the 2 kHz third-octave band). The temperature of the room is maintained at 21 °C. The sound sources used are Jabra wireless loudspeakers.

3.1.2 | Ground robot description

Microphone array

The acoustic array consists of 32 MEMS microphones and a 3D-printed microphone support. The acquisition card and the electronics of the array have been designed beforehand as part of the Megamicros project at the Institut ∂'Alembert - Sorbonne Université (France) [69]. The whole acquisition chain is illustrated in Fig. 3.2.
The Megamicros Python API allows continuous acquisition of signals up to 50 kHz and importing and manipulating the live acoustic data directly in Python. The distances between microphones range from 2 cm to 20 cm, offering a wide frequency range of exploitation.

Ground vehicle

The ground robot used in this study is a Nvidia JetRacer[2], equipped with a Jetson Nano for control and light onboard computations, Fig. 3.3a.
The vehicle is capable of moving forward and backward at various speeds, as well as turning left or right and offers several advantages for our application. First, it is relatively quiet in the studied

[1] https://www.optitrack.com/
[2] https://developer.nvidia.com/embedded/community/jetson-projects/jetracer

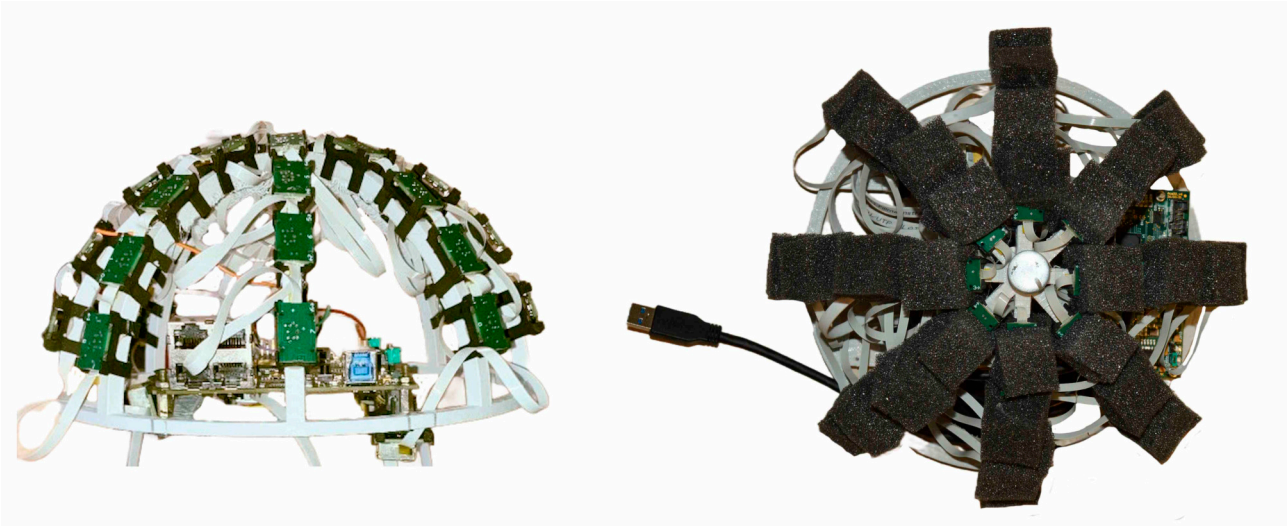
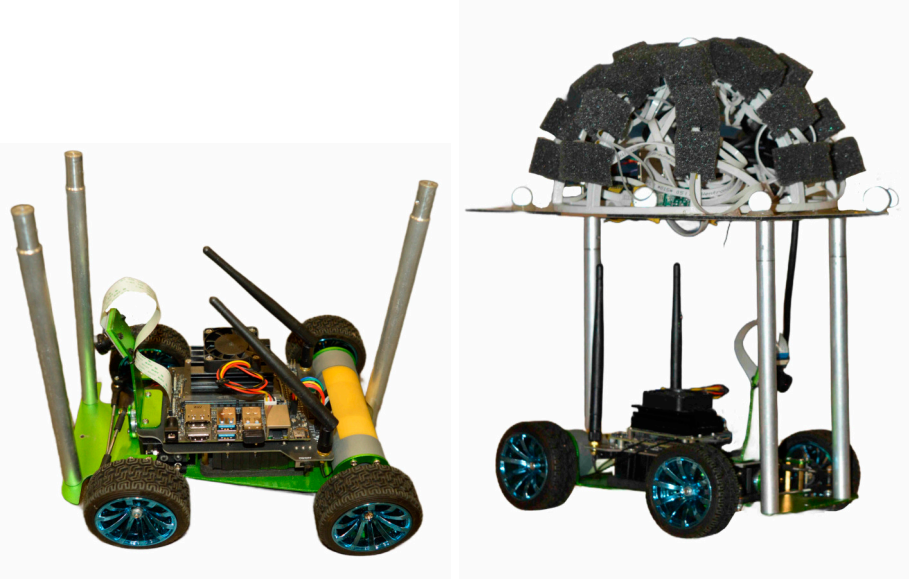


Figure 3.2. The Megamicros acquisition system and MEMS microphone array. (left) without the windshields, (right) with the polyurethane windshields.



a) b)

Figure 3.3. (a) The ground vehicle (Nvidia JetRacer). (b) The assembled ground robot.

frequency range (around 2 kHz), minimizing interference with acoustic measurements. A spectrogram of the car during motion is given in Fig. 3.4. The car starts moving at 8 s, moves in a straight motion, then stops at 10 s. It starts again at 18 s and moves while steering its wheels to the right until 21 s. Each time, shortly after stopping, the car wheels go back to their null-steering positions. This can be seen by an impulsive, high-frequency noise. The sound level by frequency bands remains very low, even during motion. Second, it supports autopiloting, which is crucial for A-SSL tasks. Third, it integrates an onboard computer hosting an Ubuntu 18 operating system capable of interfacing with the Megamicros system, ensuring smooth data processing and communication. Finally, it operates at a relatively high speed—typically around 0.8 m/s in our tests. This is a key advantage, as simulations suggest a high number of movements are required for accurate localization. In a practical scenario, the robot must complete localization within a reasonable time frame because of its battery autonomy, thus, it is interesting to use a fast vehicle to challenge the capabilities of the acoustic algorithm in this context.

The JetRacer operates within the ROS (Robot Operating System) environment, with an existing ROS

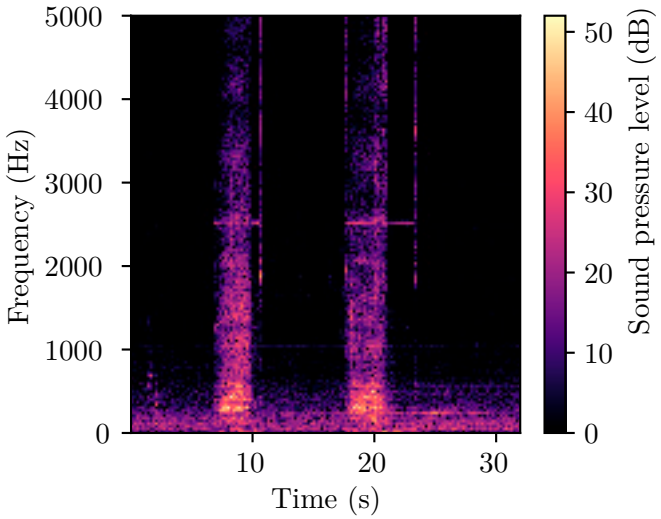


Figure 3.4. Spectrogram of the car during a start and stop motion.

package allowing for manual control. To facilitate autonomous operation, a dedicated ROS node was developed prior to this work to control the motor throttle and steering angle, see Figure 3.5.

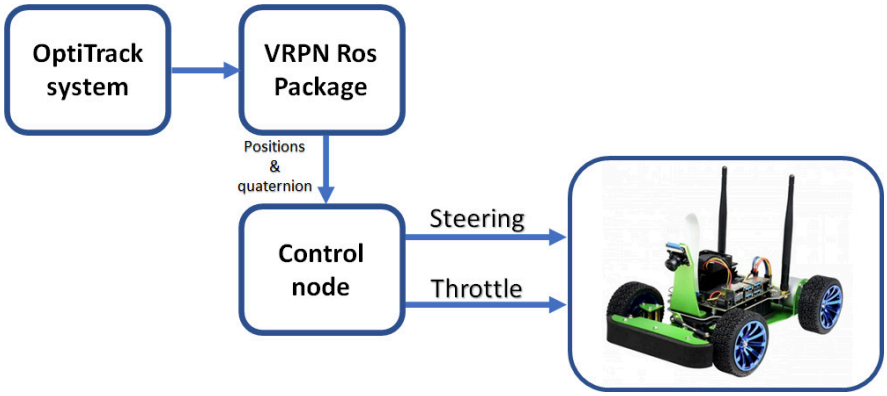


Figure 3.5. Ground vehicle architecture.

The OptiTrack system is used as the ground truth. A Virtual-Reality Peripheral Network (VRPN)[3] node was used to obtain the position and orientation of the ground vehicle using the OptiTrack system. In addition, a control node to guide the ground vehicle and control its speed was also developed. This node contains three stages: estimation of the ground vehicle's velocities, the definition of the desired positions based on the acoustic data, and the control design.

Assembled robot

The microphone array and the Megamicros acquisition system are mounted on a platform carried by the ground vehicle. To avoid damaging the material during experiments, the platform is equipped with a protective net, whose structure is acoustically transparent at the frequency range of study. Optical markers are placed over the robot's surface to track its position and orientation in the experiment room. A picture of the assembled robot is given in Fig. 3.3b.

[3] https://vrpn.github.io/

Model of the ground robot & control

The vehicle's motion can be described by the following mathematical model:

$$\begin{aligned} \dot{x}_g &= V_g \cos\gamma_g \\ \dot{y}_g &= V_g \sin\gamma_g \\ \dot{\gamma}_g &= \frac{V_g}{l_f}\tan u_s \end{aligned}$$

where (x_g, y_g) represent the positions of the ground vehicle, while γ_g is the yaw angle. V_g denotes the vehicle's speed, and l_f is the distance from the wheels to the vehicle's center of mass. The steering control is represented by u_s, and V_g can also be interpreted as the throttle control. Since the vehicle will be moving relatively slowly to ensure the optimal performance of the acoustic data processing, we can assume, without loss of generality, that the angles involved are small. Under this assumption, the previous equations simplify to:

$$\dot{x}_g = V_g \qquad \dot{y}_g = V_g\gamma_g \qquad \dot{\gamma}_g = \frac{V_g}{l_f}u_s$$

To control the system, we take V_g as the control input and propose the relationship $V_g = -k_1(x_g - x_d)$, where $k_1 > 0$ is a positive gain. This results in a closed-loop system described by $\dot{x}_g = -k_1(x_g - x_d)$, with x_d being the desired value. Consequently, the system tends towards $x_g \to x_d$. Furthermore, $\dot{y}_g$ and $\dot{\gamma}_g$ can be represented as:

$$\dot{\mathbf{y}} = A\mathbf{y} + Bu_s \tag{3.1}$$

where $\mathbf{y} = [y_g \;\; \gamma_g]^T$ and:

$$A = \begin{bmatrix} 0 & V_g \\ 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ \frac{V_g}{l_f} \end{bmatrix} \tag{3.2}$$

We propose a controller of the form $u_s = -K\mathbf{y}$, with $K = [k_2 \;\; k_3] > 0$. This leads to the equation: $\dot{\mathbf{y}} = \bar{A}\mathbf{y}$, $\bar{A}$ being a Hurwitz matrix (its eigenvalues are both negative), the system is stable [70], and $\mathbf{y} \to 0$. Without loss of generality, $\mathbf{y}$ can be rewritten as: $\mathbf{y} = [(y_g - y_d) \;\; (\gamma_g - \gamma_d)]^T$. This implies that $y_g \to y_d$ and $\gamma_g \to \gamma_d$.

Finally, it is important to note that the values of x_g, y_g, and γ_g are measured using the OptiTrack system, while x_d and y_d are dictated by the acoustic information obtained by the microphone array.

3.1.3 | Communication architecture

This section sums up the different elements that communicate together to attain real-time operation of the robot.

Before diving into the communication architecture of the system, it is important to understand the basic principles of the Robot Operating System (ROS), which is a flexible framework for building robot applications. ROS provides tools, libraries, and conventions that simplify the design and implementation of complex robotic systems. It is particularly well-suited for distributed systems where multiple processes, possibly running on different computers or devices, need to exchange data in real-time. At the core of ROS is its publish-subscribe model: the **nodes** are individual programs that communicate with each other via the ROS network. Each node can perform specific task and consists either of a C++ or a Python script. Nodes can publish data to a **topic**, or subscribe to receive data from a topic. A topic is a channel over which data can be sent. This enables nodes to exchange information asynchronously. Data sent over topics is structured in the form of **messages**. Messages can be very basic, such as numbers or character chains, or more complex, like matrices, or any mix of various data

types. The **ROS Master** coordinates the network of nodes, enabling them to find each other and communicate. In a distributed system, different nodes might run on different machines, and the ROS Master acts as an orchestra conductor and helps manage these communications.
In the proposed platform, three machines exchange information and data via a Wi-Fi ROS network. A dedicated computer hosts the VRPN ROS node and publishes the position and angle information measured via the OptiTrack on the VRPN topic. The Jetson Nano embedded in the robot supports three ROS nodes. The first one is the ROS node interfacing with Megamicros. It is in charge of acquiring the acoustic signals and publishing them on the microphone data ROS topic. The data is put in the shape of a matrix containing all the raw samples. The second one is a low-level control node in charge of interfacing the steering and throttle commands directly with the motors. It thus subscribes to the steering and throttle ROS topics. The third one is the ROS Master node. Finally, a ground station (HP Zbook i9 vPRO, 32Go RAM) hosts the path-planning node. This node has multiple tasks. First, it subscribes to the microphone data topic and estimates the DOA of the sound signals measured. It also subscribes to the VRPN topic and blends both DOA and position/orientation information to perform the triangulation. It checks the convergence of the triangulated positions and adapts the acoustic processing according to Algorithm 2. Last but not least, it is in charge of computing the next desired position of the car, tracking its position, and computing the appropriate values to publish on the steering and throttle topics. To do so, it uses the proportional controller introduced earlier.
A diagram of the communication setup is available in Fig. 3.6.

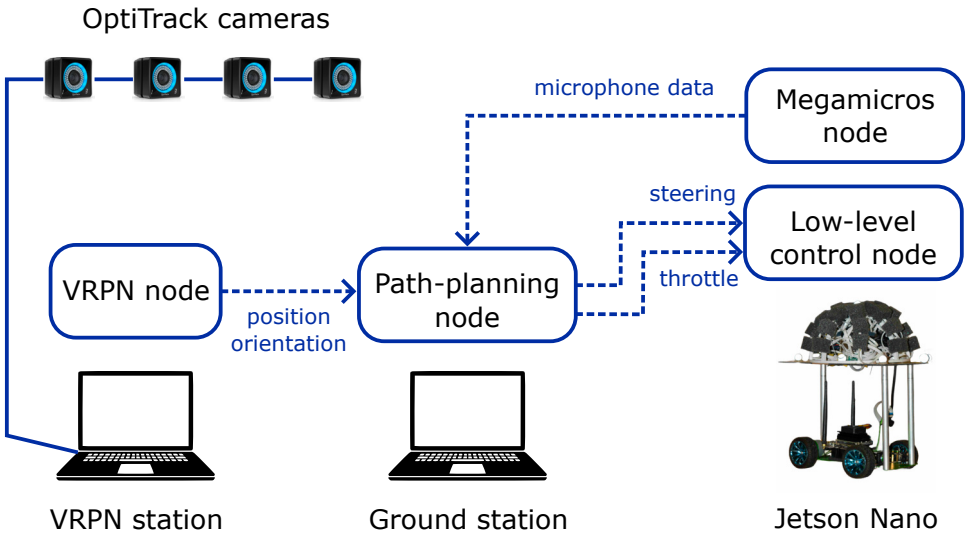


Figure 3.6. Diagram of the experimental setup and the communication architecture. Dashed arrows depict ROS topics broadcasting data via Wi-Fi. The arrow points from the publisher to the subscriber. The nature of the data transiting through the topic appears next to it in light blue. The ROS Master node is not represented for the sake of clarity.

The architecture allows parallelizing computations on different devices and thus achieves the very low latency required for real-time A-SSL.
Note that acoustic imaging can be efficiently implemented by using parallel computing [69], which would allow embedded real-time computing. Nevertheless, it is out of the scope of this study.

3.2 | Experimental results

The performance of the A-SSL method is now tested in laboratory conditions within the experimental setup introduced earlier. Two strategies are presented, the first is similar to the one depicted in the simulations in Section 2.2.3. The second performs continuous acoustic acquisition while maintaining real-time computing. For each of those strategies, the experimental setup is first tested with a single

sound source. Then, a test case with 2 sound sources is used to demonstrate the localization and filtering capabilities of the robot.

Experimental parameters

The sound sources are Jabra wireless loudspeakers emitting 2 kHz sine waves loud enough to consider the background noise negligible. Their positions are measured with the OptiTrack system. Sound pressures are measured on the microphone array with a 10 kHz sampling frequency, 10 Hz spectral resolution of 0.125 s, with 100 snapshots. Only the frequency bin corresponding to the 2 kHz source signal is used in the computations. Since Signal-to-Noise Ratio (SNR) does not represent a problem in this experiment, the robot step d is set to 1.8 m (10λ) to ensure measurements take place in the acoustic far field of the source.

Coarseness of the beamforming grid

The acoustic images are performed on a 2D angular grid, covering the whole space around the robot. The grid points are regularly spaced in azimuth and coelevation with a spacing of 5 °. The error induced by the coarseness of this grid are assumed to follow a uniform law over the interval $[-2.5, 2.5]$ °. Under the assumption of small errors and that the convergence of the SSL algorithm is mainly driven by the variance of the DOA estimation, we approximate the uniform distribution of these errors by a Gaussian distribution with similar mean ($\mu = 0$) and variance ($\sigma^2 = \frac{5^2}{12} \approx 2$). According to the simulation results presented in Chapter 2, see Fig. 2.8a, these errors are not expected to influence the convergence of the SSL algorithm. The method is thus robust to relatively coarse angular grids. The plane wave model introduced in Chapter 1 is used.

3.2.1 | Simulation and experiment comparison

Figure 3.7 compares simulated and experimental results for the single source using the autonomous silent vehicle. The robot closely follows the expected trajectory and circles the acoustic target. The final localization error in this experiment was $\frac{\lambda}{2}$, which is comparable to the loudspeaker dimensions. This first experiment supports the simulation results presented in Section 2.2.3.

3.2.2 | Real-time autonomous localization of multiple sources

After validating the A-SSL in the single source scenario, the masking capabilities of the spatial filter (2.6) are tested using the proposed experimental framework. Two sound sources are placed in the room. The robot autopilot, following Algorithm 2, searches for a first sound source, and once it has located it, applies the spatial filter and looks for a second sound source. It stops once it has found two sound sources. The robot trajectory is illustrated in Fig. 3.8. From our experience, this case is well representative of the two-source scenario.

The localization required 23 iterations to determine the source positions. The errors between the estimated positions and the ground truth are 3 cm for S_1 and 27 cm for S_2. The simulation in the multiple sound sources case predicts an average error below the wavelength, which is close to what is obtained here. These errors are expected to be statistically slightly higher than in the simulations, because of the real environment challenges, with echoes of the sources on the walls disturbing acoustic measurements and introducing DOA outliers and biases. If this is the case, the precision could most likely be improved by using a stricter convergence criterion or using more averages of the triangulated positions to estimate the final position of the source.
It can also be observed that the robot follows a slightly more erratic trajectory than the simulation. This is due to both uncertainties in the DOA estimation and the mismatch between the targeted

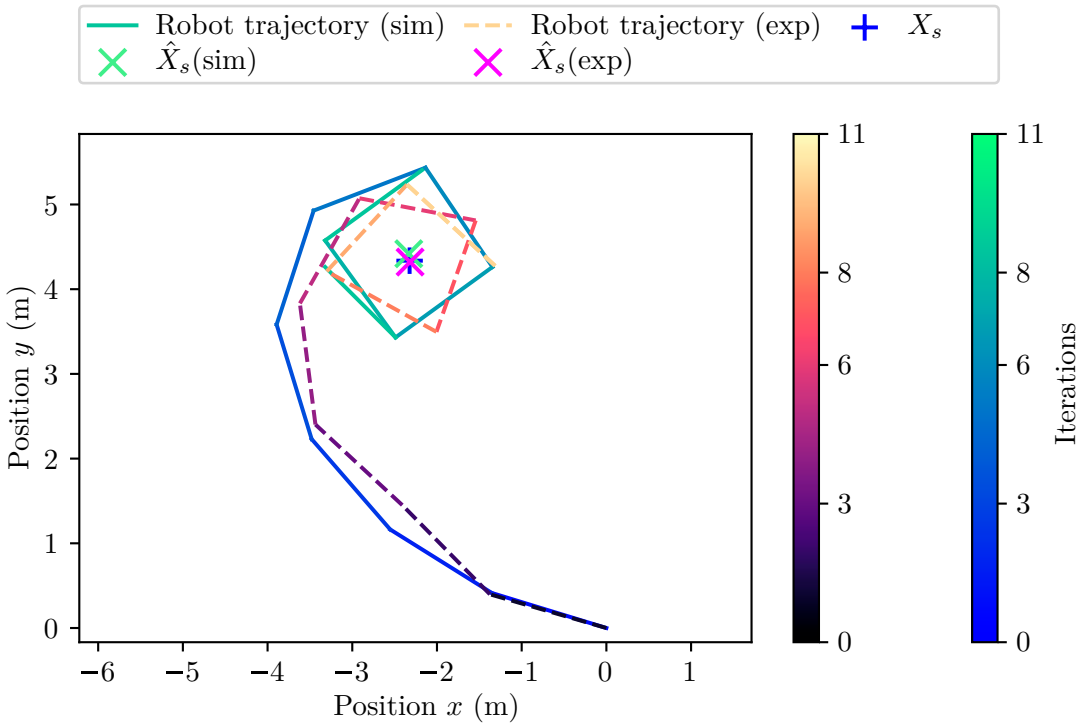


Figure 3.7. Comparison between simulated and experimental trajectories for the silent vehicle and single source case. Localization error is $\frac{\lambda}{2}$

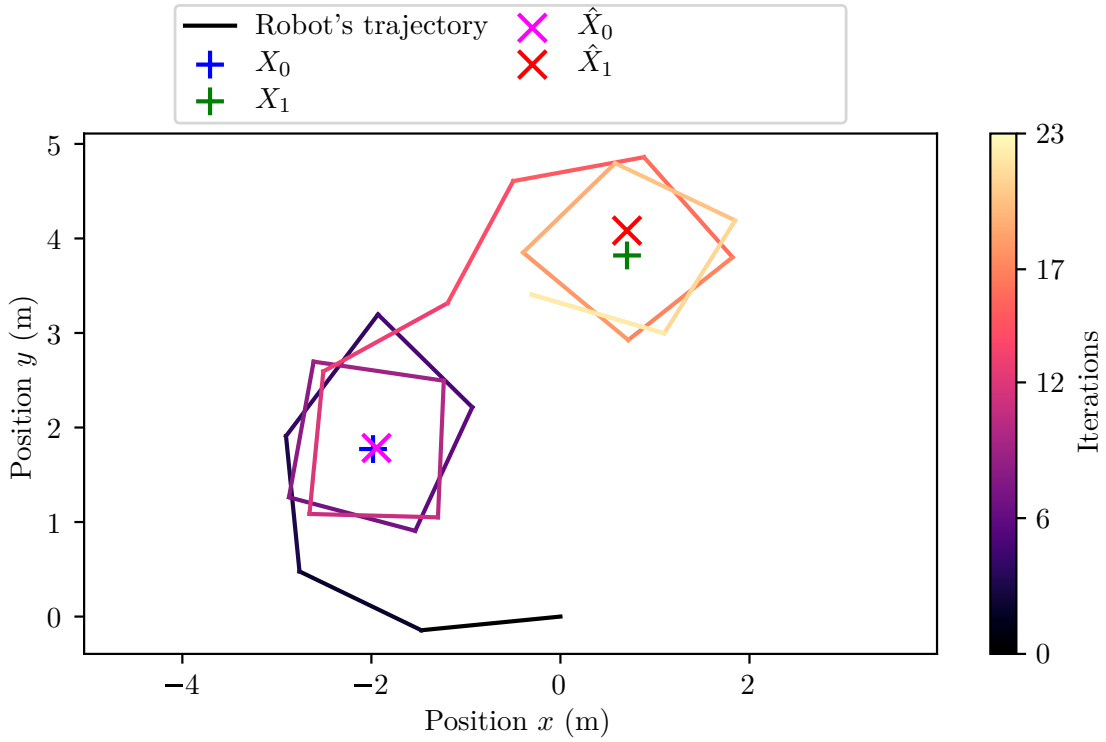


Figure 3.8. Robot's measured trajectory in the presence of two sources. Both sound sources have equal sound levels. The final localization error is of the order of magnitude of λ.

position and the attained position of the vehicle. However, this does not influence triangulation since the real position of the robot is used, rather than the desired position. The experiment was repeated several times, and similar results were observed. This demonstrates the proposed method's efficiency in an indoor environment for multiple sound sources.

3.3 | Achieving continuous real-time localization

Continuous acoustic sampling

So far, the experimental results fairly replicate the simulation. In the experiments presented previously, the robot requires time to physically reach its next position, and that travel time does not contribute to data collection. The duration of one iteration T_m can be derived from the average speed of the car $V_g = 0.8$ m/s, the step $d = 1.8$ m, and the acquisition time $T_{\text{acq}} = 0.125$ s. It can be decomposed into two phases: first, the robot moves to the measurement position, traveling during a time $\frac{d}{V_g}$, and second, it performs the acquisition during a time T_{acq}. Even neglecting the computational time to obtain the DOA, the robot spends at least a time $T_m = \frac{d}{V_g} + T_{\text{acq}} = 2.375$ s to perform one iteration. To address this inefficiency, we utilized the ability of the Megamicros device to sample sound pressures continuously, even while the robot is moving. By setting the acquisition time shorter than the robot's travel time, it becomes possible to update the robot's next position dynamically and eliminate unnecessary stops. This reduces the iteration time to the smallest value possible, which is $T_m = T_{\text{acq}}$. For the example values given earlier, this amounts to reducing T_m by 95%.
This strategy will be referred to as continuous sampling, as opposed to the point-by-point sampling presented earlier. This approach also has a key advantage over point-by-point sampling: it ensures no information is missed during the robot's movement. Instead of risking gaps in data collection, the robot continuously gathers sound pressure measurements, which can be processed in real-time.

Vehicle control

Note that, in the continuous sampling case, since the desired position of the vehicle is updated after each measurement, the step d becomes a function of the car's speed and the acquisition time, and becomes effectively smaller than the wavelength (typically 10 cm based on the speed and acquisition time introduced in the previous section). Chapter 2 showed that the vehicle converges to a circle of diameter $d\sqrt{2}(\approx 14$ cm in our case) around the sound source, thus making the acoustic far field assumption required by the DOA estimation algorithm unrealistic.
However, this reasoning ignored the kinematic model of the vehicle so far. In the case of the ground vehicle, the steering radius of the vehicle does not allow it to reach any desired position through a straight line. Thus, it can only circle the sound source at a distance equal to its steering radius, 0.8 m in our case. This situation benefits the triangulation algorithm as it allows it to automatically keep a sufficient distance from the sound source. Controlling the steering radius of the car thus conveniently allows controlling the robot-to-source distance. It is set to 0.8 m for the rest of this study.

Triangulation in the continuous sampling case

Following the continuous sampling strategy, consecutive DOA estimations are no longer orthogonal in the vicinity of the source. However, assuming the robot is already circling the source, it is possible to pick the most recent measurement that gave a DOA orthogonal to the last estimated DOA. This amounts to finding the number of iterations between two measurements of respective indices i and j so that their associated DOAs are orthogonal to each other. This number $|i - j|$ can be expressed as

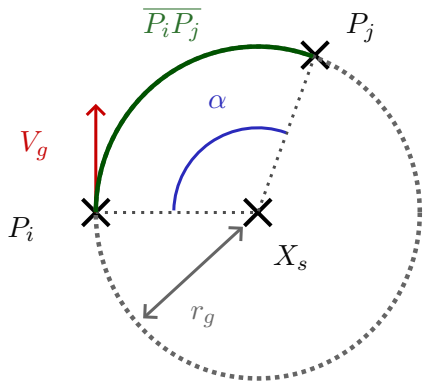


Figure 3.9. Diagram of two measurements at positions P_i and P_j.

a function of the desired angle α between two measured DOAs, the robot speed, V_g, the acquisition time, T_{acq}, and the turning radius of the car, r_g. The problem is illustrated in Fig. 3.9.

The distance traveled by the ground vehicle on its circular trajectory between the two measurement points, $\overline{P_iP_j}$, is given by $\overline{P_iP_j} = r_g\alpha$, but can also be expressed as $\overline{P_iP_j} = V_g|i-j|T_{\text{acq}}$. Thus it follows that

$$|i-j| = \frac{r_g\alpha}{V_gT_{\text{acq}}} \tag{3.3}$$

This allows us to associate the latest measurement with the last measurement that yields an orthogonal DOA, $\alpha = 90$ deg, and to apply them two-DOAs triangulation scheme (2.2). This is expected to make the localization more robust to noise in the DOA estimations, following the principle introduced in Section 2.1.4. This method replaces the previous consecutive DOAs triangulation and the convergence criterion remains unchanged from the previous study.

Acoustic imaging in a circular motion

Moving the microphone array during acquisition typically blurs acoustic images and compromises DOA estimation, unless the motion is negligible over the acquisition window. However, the circular nature of the robot's motion mitigates this issue. Once the robot approaches the source and begins its circular trajectory centered around it, the source appears nearly stationary in the array's coordinate system, as illustrated in Fig. 3.10.

The direction of the source within the array coordinates corresponds to the angular distance between the array's orientation (blue) and the source direction (red). Under the planar wave propagation model $\mathbf{a}(\theta_q) = \mathrm{e}^{\mathrm{j}\mathbf{k}(\theta)\mathbf{r}}$, where $\mathbf{k}$ is the wave vector, θ the source direction, and $\mathbf{r}$ the microphone positions relative to the source, the received signal depends solely on $\mathbf{k}$ and $\mathbf{r}$. In our setup, both remain constant throughout the measurement path. As a result, the steering vector aligned with the source direction consistently matches the Green's function over the trajectory. This coherence minimizes the influence of the robot's motion on the A-SSL process, enhancing robustness at higher speeds. Furthermore, under the acoustic far-field assumption, the robot's motion is tangential to the circular wavefronts, effectively nullifying Doppler effects.

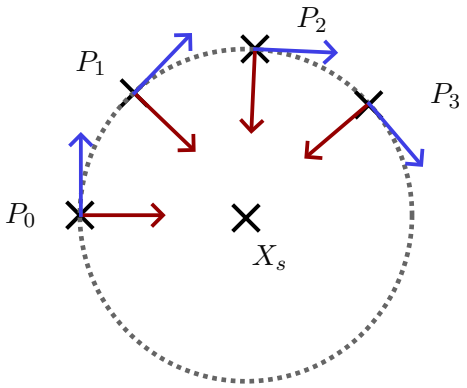


Figure 3.10. The array motion around the source. The orientation of the array is depicted in blue, and the direction of the source in red.

Processing parameters and latency

Another key assumption concerns processing time: it must remain shorter than the acquisition time to minimize delays between data capture and processing. This constraint was validated experimentally, and a suitable set of parameters was identified that met this condition without observable latency. Specifically, an acquisition time of 0.125 s, a sampling frequency of 10 kHz, a spectral resolution of 10 Hz, and 100 snapshots for CSM computation were used. A regularly spaced 2D angular grid with 5 ° steps in both azimuth and coelevation was also employed. These parameters were retained for the subsequent experiments presented in this section.

Simulation/experiment comparison

Fig. 3.11 compares simulated and experimental results for the single source using the autonomous ground vehicle and the continuous sampling strategy. Similar to Fig. 3.7, the robot closely follows the simulated trajectory and circles the source. The final localization error in this experiment was $\frac{\lambda}{2}$, which is similar to the point-by-point sampling strategy. Note that in the case of point-by-point sampling, the experiment took about 30 s to be performed, whereas the continuous sampling strategy achieved similar results in about 8.25 s. These results suggest that continuous sampling might achieve similar performances as the point-by-point sampling strategy, in a much shorter time of operation.

Two sources experiments

The continuous sampling strategy is now tested in the multiple source case, similarly to Section 3.2.2. Two sources of equal sound level are placed in the robotic test area, and the robot is placed at different starting points. Its position is recorded by the OptiTrack at each measurement, and the robot is stopped after locating the two sources. Two representative cases are presented in Fig. 3.12. The robot required 188 and 148 measurements, respectively (resp. 23.5 and 18.5 s) to find the two sources. Comparatively, the similar test case presented in Fig. 3.8 running the point-by-point sampling took about 60 s to be completed. Errors between the estimated positions and the ground truth are on average lower than λ, yielding a similar precision as the point-by-point strategy.
Figure 3.13 presents an interesting case where the robot first finds S_0, but seems to struggle to mask it accurately. This can be explained by a mismatch between the estimated and real position of the

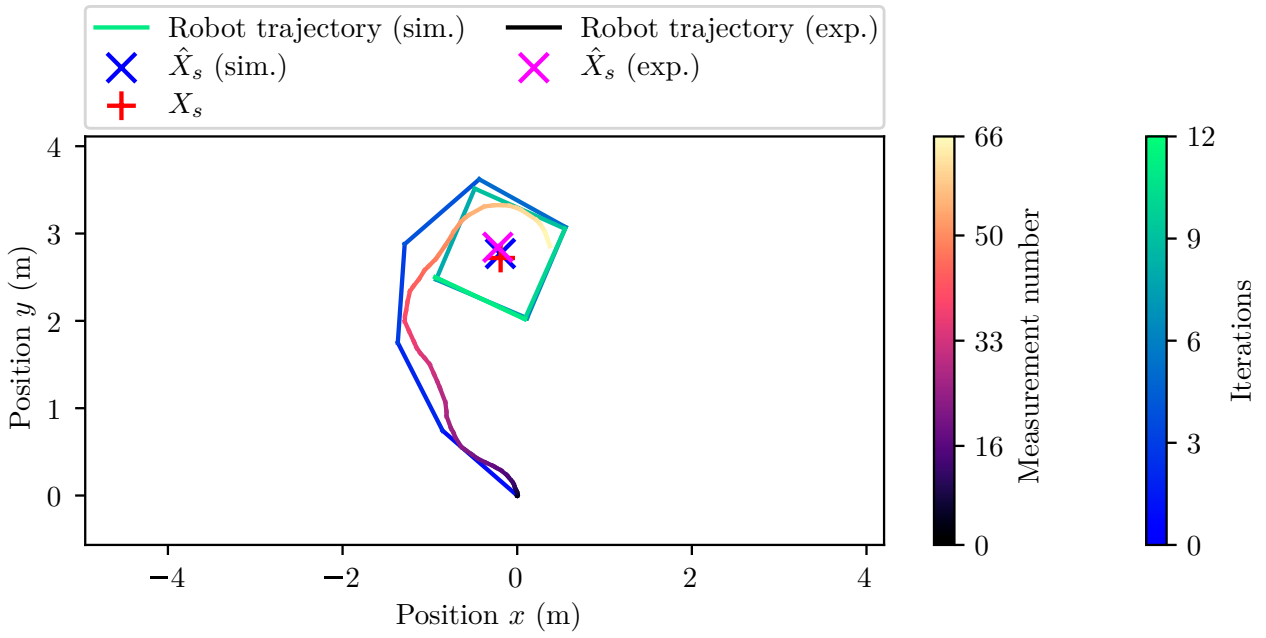


Figure 3.11. Comparison between simulated and experimental trajectories for the ground vehicle and single source case using the continuous sampling strategy with an acquisition time of 0.125 s. Localization error is $\frac{\lambda}{2}$.

source. The robot keeps circling S_0 before locating it one more time, adding another filter. After the second filter is added, the robot can break free and find S_1. In total, the robot took 237 measurements (i.e. 29.6 s) to locate both sources for localization errors under the wavelength λ. This illustrates the methodology introduced in Section 2.4.5 which ensures the robot considers that the two closely spaced estimations are linked to a single real sound source, and thus that only two sound sources are present in this experiment. This behavior was also observed in the simulations presented in Section 2.4.5 and demonstrates the ability of the robot to solve problems of accuracy of the position given to the spatial filter.

The continuous sampling strategy is thus validated as a more time-efficient alternative to the point-by-point sampling.

3.4 | Discussions

This experimental study demonstrated the feasibility of the proposed A-SSL methodology for localizing a known number of omnidirectional, stationary, and immobile sound sources in an unknown indoor environment.

The development of a dedicated experimental platform for implementing and validating autonomous SSL approaches in real time paves the way for various future studies and applications. Notably, the system's capabilities could be extended to track non-stationary or moving sources through the integration of advanced triangulation techniques such as Kalman or particle filtering, thereby enhancing localization robustness.

The vehicle's control currently relies on accurate positional feedback from the OptiTrack system, which, while extremely precise, is impractical for large-scale or outdoor deployments. In real-world scenarios, alternative sensors such as GPS, cameras, or Lidars would be required. The simulation study on convergence under DOA estimation uncertainty in Chapter 2 provides insights into the method's robustness to such inaccuracies. Nonetheless, a more comprehensive sensitivity analysis would be

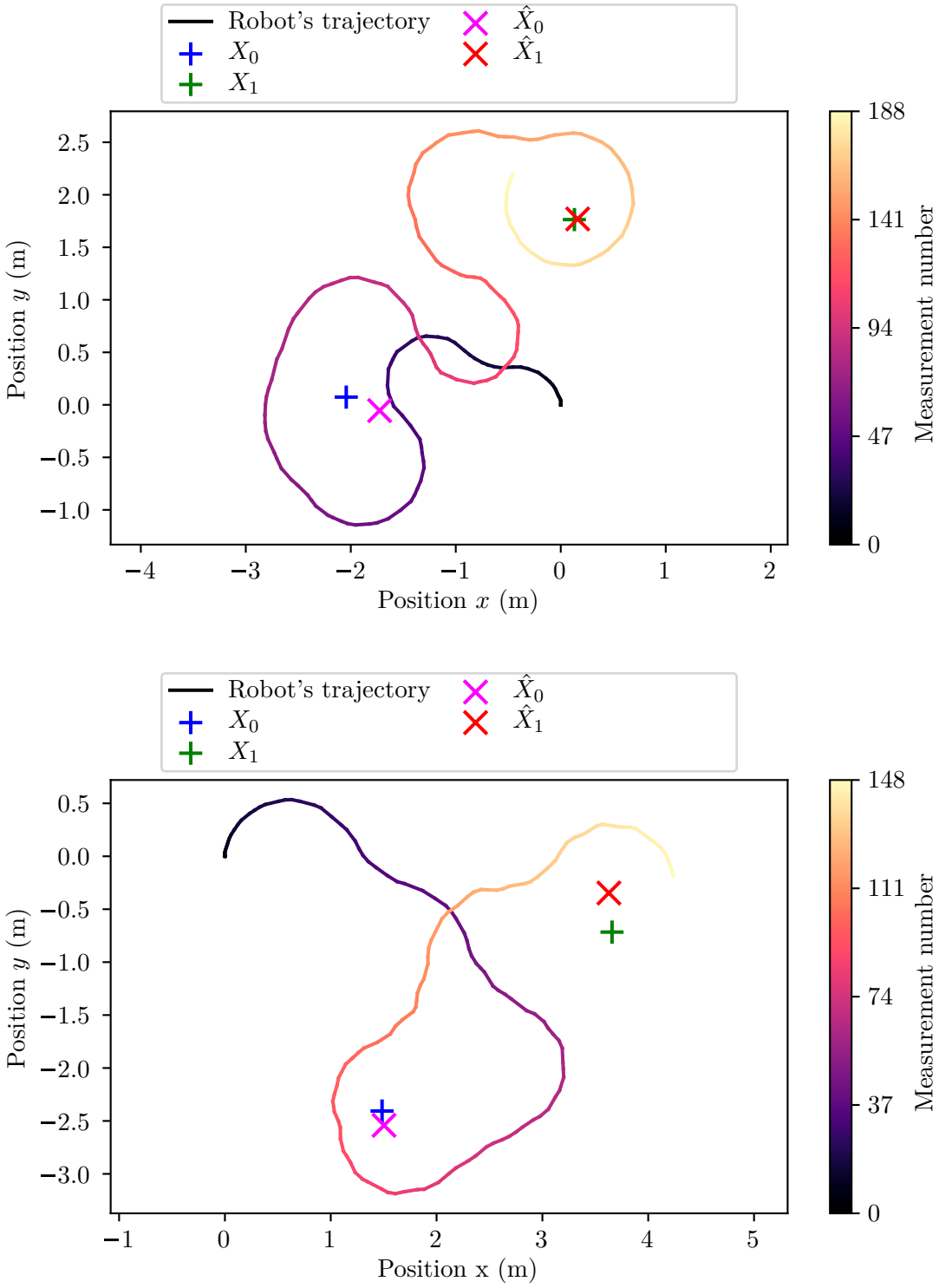


Figure 3.12. Robot trajectories in the presence of two sources using the continuous sampling strategy with an acquisition time of 0.125 s. Both sound sources have equal sound levels. The final localization error is of the order of magnitude of λ. The robot's starting point influences the trajectories. The localization errors remain lower than 2λ.

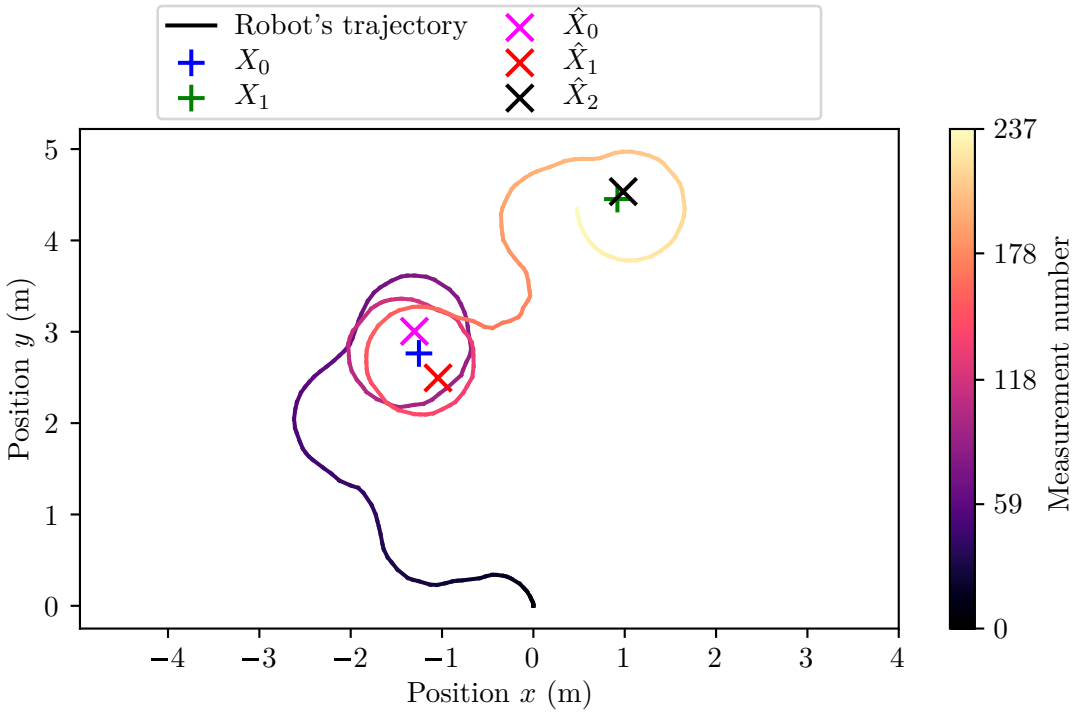


Figure 3.13. Robot trajectory in the presence of two sources using the continuous sampling strategy with an acquisition time of 0.125 s. Both sound sources have equal sound levels. The final localization error is of the order of magnitude of λ. In this case, the robot found the first source twice before finding the second source. The localization errors remain lower than 2λ.

beneficial before considering the replacement of OptiTrack. Moreover, optimizing the implementation of both the acoustic imaging and path-planning algorithms could enable fully embedded processing onboard the robot, thus eliminating the need for an external ground station.

The system presented in this Chapter is meant for silent vehicles, however, noisy vehicle show some interesting assets, particularly unmanned aerial vehicles (UAVs) or drones. UAVs offer a more flexible and maneuverable platform for deploying the A-SSL system than the radio-controlled car, which opens up broader use scenarios. However, a challenge arises from the vehicle's own noise, which can interfere with the acoustic measurements. The techniques and approaches employed to reduce the influence of vehicle noise on localization accuracy are discussed in detail in Chapter 4, where the results and theoretical foundations of the filtering strategies will be presented.

3.5 | Conclusion

This chapter explored a new experimental framework for autonomous SSL in controlled but unknown environments. A complete acquisition and processing platform was developed and validated, demonstrating the feasibility of localizing a known number of omnidirectional and stationary sources using a continuous, motion-based sampling strategy.

The experimental results confirmed that the circular motion of the robot around the source mitigates the impact of movement on localization accuracy. Once the trajectory stabilizes around the source, the relative geometry becomes nearly constant, preserving the integrity of the acoustic measurements and ensuring compatibility with the underlying localization model.

The proposed system achieved sub-wavelength localization precision within a significantly shorter time compared to conventional point-by-point approaches, while operating under real-time constraints. The parameters selected for the processing — including snapshot duration, resolution, and angular grid —

were shown to provide a good compromise between computational cost and estimation reliability. The capabilities of the proposed approach were demonstrated in the two sound sources case during real-time experiments using the iterative localization/filtering approach introduced in Chapter 2.

Summary tables

The key achievements introduced in this chapter are summarized in Table 3.1, and the main perspectives discussed are listed in Table 3.2.

Aspect	Achievement
System feasibility	Developped and validated an experimental platform for real-time A-SSL in unknown indoor environments.
Source localization	Achieved sub-wavelength localization accuracy for multiple stationary sound sources.
Path-planning strategy	Demonstrated that circular motion enhances localization precision and enables continuous, real-time operation.
Continuous operation	Continuous real-time sampling significantly reduces total acquisition time compared to point-by-point strategies.

Table 3.1. Summary of the main achievements of the proposed A-SSL framework

Aspect	Perspective
Source dynamics	Extend the system to localize moving or non-stationary sources using advanced triangulation methods (e.g., Kalman or particle filtering).
Sensor independence	Replace the OptiTrack system with onboard sensors (GPS, camera, LiDAR) for deployment in larger-scale or outdoor environments.
Embedded processing	Optimize algorithmic implementation to enable fully embedded, onboard processing and remove dependence on a ground station.
Non-silent vehicles adaptations	Investigate the adaptations required to use Unmanned Aerial Vehicles as a carrier for the microphone array.

Table 3.2. Future perspectives for the proposed A-SSL framework

This work serves as a foundation for further developments in A-SSL, both in terms of algorithmic robustness and system autonomy. The next chapter investigates the adaptations necessary to use an Unmanned Aerial Vehicle (UAV) as the carrier robot, with a focus on the noise generated by the flying vehicle polluting microphone measurements.

4. Ego-noise generation and filtering for UAVs

Contents

4.1 UAV fundamentals . . . 74
4.1.1 Experimental platform . . . 74
4.1.2 Principles of UAV flight . . . 75
4.1.3 Dynamic model of a UAV . . . 76
4.1.4 UAV control strategy . . . 77
4.2 Noise generated by a quadcopter . . . 81
4.2.1 Propeller-induced acoustic sources . . . 81
4.3 Ego-noise mitigation techniques . . . 87
4.3.1 Passive noise reduction methods . . . 87
4.3.2 Numerical noise filtering techniques . . . 88
4.4 Practical implementation of denoising . . . 91
4.4.1 Estimation of the measurement CSM . . . 91
4.4.2 Estimation of the ego-noise CSM . . . 92
4.4.3 Temporal variations of the ego-noise CSM coefficients . . . 93
4.5 Experimental validation . . . 96
4.5.1 Case study: noise point source . . . 97
4.5.2 Case study: UAV interference during stationary hover . . . 99
4.6 Discussion . . . 99
4.7 Conclusion . . . 100

Unmanned Aerial Vehicles (UAVs) have become valuable tools in a wide range of acoustic sensing applications, including environmental monitoring, search and rescue, and industrial diagnostics. Their mobility and flexibility enable the deployment of microphones in otherwise inaccessible or hazardous locations. However, the acoustic utility of UAVs is fundamentally limited by the strong ego-noise they generate, primarily from their propulsion systems. This interference can severely degrade the quality of recorded signals and the detection and localization of external sound sources.

The aim of this chapter is to investigate the impact of UAV-generated noise on acoustic measurements and to evaluate practical denoising strategies available. To that end, Section 4.1 first introduces the key characteristics of UAV platforms, with a particular focus on multirotor vehicles and their control systems, which play a central role in noise generation. Section 4.2 then reviews the main ego-noise sources arising from the propellers' operation. It highlights the complex nature of UAV noise, and the challenges it poses for acoustic signal processing. Section 4.3 is dedicated to denoising techniques, which are divided into passive and numerical approaches. Passive methods aim to reduce noise transmission through physical design changes, while numerical methods attempt to separate ego-noise from external signals through statistical processing. The challenge of denoising in realistic



Figure 4.1. The Intel Aero quadrotor.

conditions, involving short acquisition time, and limited compuational power, is investigated in Section 4.4. Finally, the proposed framework is validated in Section 4.5 through a series of experiments, analyzing the performance of a CSM whitening algorithm in real UAV recordings.

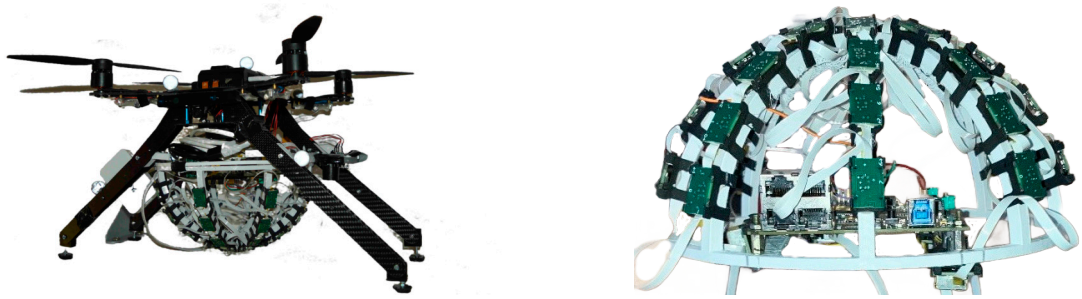
4.1 | UAV fundamentals

UAVs vary in design and so do their acoustic emissions. They are often separated into two classes: fixed-wing UAVs (FW-UAV) and rotary-wing UAVs (RW-UAV). FW-UAVs are more efficient in terms of energy consumption and can also perform gliding flight [71], which is much less noisy, making them best suited for SSL in theory. They struggle to perform hover maneuvers restraining their use to wide open area which limit their use in practice. RW-UAVs are noisier, with sound power levels varying drastically from one model to another, but allow for vertical landing and hovering in a given position. They also allow greater control of their position and orientation. Although the noise emitted by RW-UAVs can pose a serious challenge for acoustic imaging, they were preferred in this work to their fixed-wing counterparts because of their maneuverability. RW-UAVs models can have various numbers of propellers, the most common configuration being the four propeller one, called a quadcopter. This study thus focuses specifically on quadcopter UAVs noise, although the general results could be expanded to other designs.

4.1.1 | Experimental platform

In order to illustrate the fundamental concepts of UAV noise reduction, the quadcopter used in this work is introduced now. The configuration chosen is an Intel Aero quadcopter, see Fig. 4.1. It can carry a payload of up to one kilogram and offers a convenient platform to develop our solution. It comes fully assembled and equips an Intel Aero Compute Board running under a Linux environment, a Flight Controller unit, a GPS, a compass, and an Inertial Measurement Unit (IMU). It is operated within the Heudiasyc robotic area (described in Chapter 3) using the OptiTrack camera array to monitor its position and orientation. It is powered by a 14.7 V LiPo battery.

This aerial vehicle was modified to carry the microphone array, to avoid placing microphones directly under the propeller's flow, and to host the Megamicros acquisition card, as shown in Fig. 4.2. The legs were extended using carbon plates, and an optical sensor was added to the side of the array to increase the robustness of the altitude estimation.



a) b)

Figure 4.2. (a) The Intel Aero quadrotor equipped with the microphone array, (b) the embedded microphone array and the Megamicros acquisition card.

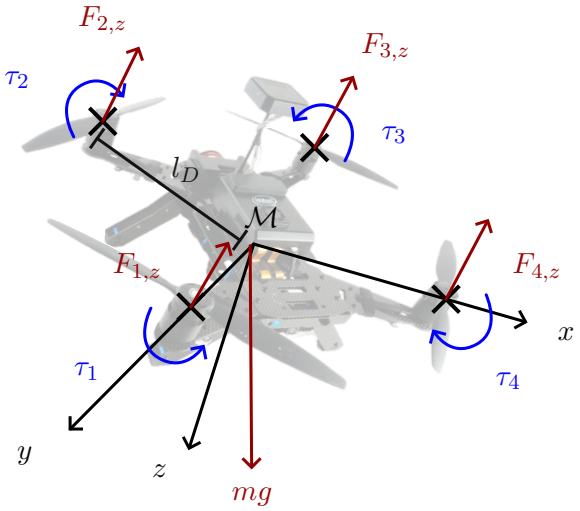


Figure 4.3. Forces and torques on the aerial quadcopter vehicle.

4.1.2 | Principles of UAV flight

The noise generated by these aerial vehicle is inherently linked to their flight control mechanisms. To understand the strong link between the UAV control and the noise it generates, it is important to introduce the basic principles of UAV control.

Consider an aerial quadcopter vehicle equipped with four propellers arranged along two orthogonal axes, as shown in Fig. 4.3. Each propeller spins at a defined angular velocity ω_i (with $i \in \{1, 2, 3, 4\}$), generating both lift and thrust forces. In this work, we consider that each propeller produces a vertical lift force $F_{i,z}$ given by:

$$F_{i,z} \approx k_f \omega_i^2$$

where k_f is the thrust coefficient related to the propeller's aerodynamic properties. The total thrust force, which acts along the drone's body-fixed vertical axis and is responsible for lifting the UAV, is the sum of the four individual rotor forces. We denote this total force as u, and define it as:

$$u = \sum_{i=1}^{4} k_f \omega_i^2 \tag{4.1}$$

In steady flight, this thrust force u balances the gravitational force ma_g, maintaining the UAV's altitude, where m is the mass of the UAV and a_g is the gravitaional acceleration. The rotation of the

propeller also generates a reactive torque τ_i, given by:

$$\tau_i = k_\tau \omega_i^2$$

where k_τ is the drag-related torque coefficient. As the propellers are posed such that two rotate clockwise and the others counter-clockwise, the sum of the reactive torques produced gives a net yaw moment acting on the UAV, given as:

$$\tau_\gamma = k_\tau(\omega_1^2 + \omega_3^2 - \omega_2^2 - \omega_4^2) \tag{4.2}$$

which can be controlled by varying the relative speeds of the rotors. To prevent unwanted rotational motion due to these reactive torques, the following spin configuration is used: rotors 1 and 3 rotate clockwise (CW) and rotors 2 and 4 rotate counterclockwise (CCW). Since each rotor produces a torque in the opposite direction of its spin, the torques induced by opposing rotors cancel out when all four rotors operate at the same speed, ensuring no net yaw torque at the UAV's center of mass in steady flight. This balance prevents unwanted rotation and allows for precise yaw control by introducing speed differentials between the clockwise and counterclockwise rotors.

Therefore, appropriately selecting the angular velocities ω_i allows to control the UAV roll, pitch, yaw, and vertical acceleration. The corresponding moments about the principal axes (assuming a symmetric frame with arm length l_D) are:
roll, α (rotation about x-axis):

$$\tau_\alpha = l_D k_f(\omega_1^2 - \omega_3^2) \tag{4.3}$$

pitch, β (rotation about y-axis):

$$\tau_\beta = l_D k_f(\omega_2^2 - \omega_4^2) \tag{4.4}$$

yaw, γ (rotation about z-axis), was described in Eq. (4.2)

These four control inputs exploit the horizontal components of thrust (u), and allow control of the full quadcopter attitude motion via its torques ($\tau = [\tau_\alpha, \tau_\beta, \tau_\gamma]^T$). Note that this is an underactuated system, as it has more degrees of freedom (6) than control inputs (4). Therefore, forward, backward, and lateral translations are achieved indirectly by tilting in the longitudinal and lateral axes the UAV.

4.1.3 | Dynamic model of a UAV

To formalize the link between the control inputs and the drone's resulting motion, a dynamic model can be derived using the Euler–Lagrange method. For the quadcopter, the six generalized coordinates of interest correspond to its positions (x, y, z) and orientation (roll α, pitch β, yaw γ). The time derivatives of these coordinate variables are denoted by the superscript $\dot{}$, and their second order time derivatives by the superscript $\ddot{}$. The kinetic energy T of the system can be decomposed into a translational component T_{trans} and a rotational component T_{rot} which write:

$$T_{\text{trans}} = \frac{1}{2} m \left(\dot{x}^2 + \dot{y}^2 + \dot{z}^2\right) \tag{4.5}$$

$$T_{\text{rot}} = \frac{1}{2} \left(I_x \dot{\alpha}^2 + I_y \dot{\beta}^2 + I_z \dot{\gamma}^2\right) \tag{4.6}$$

where m is the mass of the UAV, and I_x, I_y, I_z are the moments of inertia about the body axes. The potential energy $\mathcal{U}$ is given by:

$$\mathcal{U} = m a_g z \tag{4.7}$$

The Lagrangian L is then defined as the difference between the kinetic and potential energies:

$$L = T - \mathcal{U} \tag{4.8}$$

The equations of motion can be derived from the Lagrangian using the Euler-Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = \mathbf{Q} \tag{4.9}$$

where q represents the generalized coordinates $(x, y, z, \alpha, \beta, \gamma)$, and $\mathbf{Q} = [\mathcal{R}u, \tau]^T$ is the generalized force vector, and $\mathcal{R}(\alpha, \beta, \gamma)$ is the rotation matrix. $\mathbf{Q}$ is defined as

$$\mathbf{Q} = \begin{bmatrix} -u\sin(\beta) \\ u\cos(\beta)\sin(\alpha) \\ u\cos(\beta)\cos(\alpha) \\ \tau_\alpha \\ \tau_\beta \\ \tau_\gamma \end{bmatrix} \tag{4.10}$$

Computing (4.9), and considering small angles (no Coriolis forces) and no external aerodynamic disturbances, the resulting equations of motion can be expressed as (see [72]):

$$\begin{cases} m\ddot{x} = -u\sin(\beta) \\ m\ddot{y} = u\cos(\beta)\sin(\alpha) \\ m\ddot{z} = u\cos(\beta)\cos(\alpha) + ma_g \\ I_x\ddot{\alpha} = \tau_\alpha \\ I_y\ddot{\beta} = \tau_\beta \\ I_z\ddot{\gamma} = \tau_\gamma \end{cases} \tag{4.11}$$

4.1.4 | UAV control strategy

Remark that the goal of this work is not to propose a new control algorithm for this aerial vehicle. To regulate its motion, we use a cascaded PID control architecture, commonly found in literature for multirotor systems. This architecture is separated into two layers: an outer-loop position controller, which regulates the UAV's position in 3D space, and an inner-loop attitude controller, which stabilizes the UAV's orientation. Notice that we are not considering aggressive maneuvers, thus small angles are assumed for the design of the controller.

Outer loop: translational control

This loop computes the desired attitude angles α_d and β_d, and vertical thrust u, based on position tracking errors e_x, e_y, e_z. Thus, denoting the error for the x-axis as:

$$e_x(t) = x(t) - x_d(t)$$

where x is the state of the vehicle and x_d is its desired value, the PID controller is proposed as:

$$\beta_d(t) = K_{P,x}e_x(t) + K_{I,x}\int_0^t e_x(s)\,ds + K_{D,x}\frac{de_x(t)}{dt} \tag{4.12}$$

where $\beta_d(t)$ is the desired pitch angle, $K_{P,x}, K_{I,x}, K_{D,x}$ are positive gains defining the proportional, integral, and derivative terms of the controller, respectively.
Similarly, for the y-axis, the error is defined as:

$$e_y(t) = y(t) - y_d(t)$$

where y is the state of the vehicle and y_d is its desired value. The PID controller is proposed as:

$$a_g \alpha_d(t) = -K_{P,y} e_y(t) - K_{I,y} \int_0^t e_y(s)\, ds - K_{D,y} \frac{de_y(t)}{dt} \tag{4.13}$$

where $\alpha_d(t)$ is the desired roll angle, $K_{P,y}, K_{I,y}, K_{D,y}$ are positive gains defining the proportional, integral, and derivative terms of the controller, respectively.
Finally, for the z-axis, the error is defined as:

$$e_z(t) = z(t) - z_d(t)$$

where z is the state of the vehicle and z_d is its desired value. The PID controller is proposed as:

$$u_z(t) = -K_{P,z} e_z(t) - K_{I,z} \int_0^t e_z(s)\, ds - K_{D,z} \frac{de_z(t)}{dt} \tag{4.14}$$

where $u_z(t)$ is the desired performance for the altitude, and $K_{P,z}, K_{I,z}, K_{D,z}$ are positive gains defining the proportional, integral, and derivative terms of the controller, respectively.

Inner loop: attitude control

This loop stabilizes the UAV's angular orientation based on attitude tracking errors. Defining the attitude errors as:

$$\begin{aligned} e_\alpha(t) &= \alpha(t) - \alpha_d(t) \\ e_\beta(t) &= \beta(t) - \beta_d(t) \\ e_\gamma(t) &= \gamma(t) - \gamma_d(t) \end{aligned}$$

where α, β, γ are the current roll, pitch, and yaw angles of the UAV, and $\alpha_d, \beta_d, \gamma_d$ are their desired values. The PID controller for the roll angle is proposed as:

$$I_x \tau_\alpha(t) = -K_{P,\alpha} e_\alpha(t) - K_{I,\alpha} \int_0^t e_\alpha(s)\, ds - K_{D,\alpha} \frac{de_\alpha(t)}{dt} \tag{4.15}$$

where $K_{P,\alpha}, K_{I,\alpha}, K_{D,\alpha}$ are positive gains defining the proportional, integral, and derivative terms of the controller, respectively. Similarly, the PID controller for the pitch angle is proposed as:

$$I_y \tau_\beta(t) = -K_{P,\beta} e_\beta(t) - K_{I,\beta} \int_0^t e_\beta(s)\, ds - K_{D,\beta} \frac{de_\beta(t)}{dt} \tag{4.16}$$

where $K_{P,\beta}, K_{I,\beta}, K_{D,\beta}$ are positive gains defining the proportional, integral, and derivative terms of the controller, respectively. Finally, the PID controller for the yaw angle is proposed as:

$$I_z \tau_\gamma(t) = -K_{P,\gamma} e_\gamma(t) - K_{I,\gamma} \int_0^t e_\gamma(s)\, ds - K_{D,\gamma} \frac{de_\gamma(t)}{dt} \tag{4.17}$$

where $K_{P,\gamma}, K_{I,\gamma}, K_{D,\gamma}$ are positive gains defining the proportional, integral, and derivative terms of the controller, respectively.

The outer-loop controller produces the desired attitudes for pitch and roll that are passed to the inner-loop attitude controllers, which in turn generate angular rate commands to be tracked by the motor control layer. Remark that the desired yaw angle, γ_d, is directly obtained from the acoustic DOA estimation algorithm.

Stability analysis

Rewriting the altitude dynamic of the aerial vehicle, thus from (4.11)

$$m\ddot{z} = u\cos(\beta)\cos(\alpha) + ma_g \tag{4.18}$$

Notice that (4.14) represents the desired behavior for the altitude of the aerial vehicle. Thus, to control it, we can propose

$$u = \frac{u_z/m - ma_g}{\cos(\beta)\cos(\alpha)} \tag{4.19}$$

Therefore the closed-loop system becomes

$$\ddot{z} = u_z \tag{4.20}$$

Observe that the previous equation can be rewritten as

$$\dot{\mathbf{z}} = A\mathbf{z} + Bu_z \tag{4.21}$$

with

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \dot{\mathbf{z}} = \begin{bmatrix} e_z \\ \dot{e}_z \\ \ddot{z} \end{bmatrix} \tag{4.22}$$

Notice also that (4.14) can be rewritten as $u_z = -\mathbf{K}\mathbf{z}$, with $\mathbf{K}_z = [K_{I,z}\ K_{P,z}\ K_{D,z}]$. Therefore, the closed-loop system becomes

$$\dot{\mathbf{z}} = (A - B\mathbf{K})\mathbf{z} \tag{4.23}$$

Thus, if $(A - B\mathbf{K})$ is Hurwitz, then the system is stable [70], and $e_z \to 0$ and $z \to z_d$.

Similarly for the γ dynamics, it follows that

$$I_z\ddot{\gamma} = \tau_\gamma \tag{4.24}$$

Introducing (4.17) into the previous, it follows

$$\ddot{\gamma} = -K_{P,\gamma}e_\gamma(t) - K_{I,\gamma}\int_0^t e_\gamma(s)\,ds - K_{D,\gamma}\frac{de_\gamma(t)}{dt} \tag{4.25}$$

Making a similar analysis that for the $\bar{\gamma}$ dynamics, it yields

$$\dot{\bar{\gamma}} = A_\gamma\bar{\gamma} + B_\gamma\tau_\gamma \tag{4.26}$$

with

$$A_\gamma = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad B_\gamma = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad \dot{\bar{\gamma}} = \begin{bmatrix} e_\gamma \\ \dot{e}_\gamma \\ \ddot{\gamma} \end{bmatrix} \tag{4.27}$$

Notice also that (4.17) can be rewritten as $\tau_\gamma = -\mathbf{K}_\gamma\bar{\gamma}$, with $\mathbf{K}_\gamma = [K_{I,\gamma}\ K_{P,\gamma}\ K_{D,\gamma}]$. Therefore, the closed-loop system becomes

$$\dot{\gamma} = (A_\gamma - B_\gamma\mathbf{K}_\gamma)\bar{\gamma} \tag{4.28}$$

Thus, if $(A_\gamma - B_\gamma \mathbf{K}_\gamma)$ is Hurwitz, then the system is stable, and $e_\gamma \to 0$ and $\gamma \to \gamma_d$

Introducing (4.19) into the y dynamics in (4.11), it follows that

$$m\ddot{y} = ma_g \tan(\alpha) \tag{4.29}$$
$$I_x\ddot{\alpha} = \tau_\alpha \tag{4.30}$$

As explained previously, considering small angles, it yields

$$\ddot{y} \approx a_g\alpha \tag{4.31}$$
$$I_x\ddot{\alpha} = \tau_\alpha \tag{4.32}$$

Introducing (4.15) into (4.32) and following an analysis similarly for z and γ dynamics, it follows that $e_\alpha \to 0$ implying that $\alpha \to \alpha_d$.

Introducing then (4.13) into (4.31), and making the same stability analysis, it yields that $e_y \to 0$ implying that $y \to y_d$.
Analyzing the x and β dynamics, it follows

$$m\ddot{x} = -\frac{ma_g \tan(\beta)}{\cos\alpha} \tag{4.33}$$
$$I_y\ddot{\beta} = \tau_\beta \tag{4.34}$$

Notice that after a time T_1, $\alpha << 1$, making that $\cos\alpha \approx 1$, and considering small angles, it yields

$$\ddot{x} \approx -a_g\beta \tag{4.35}$$
$$I_y\ddot{\beta} = \tau_\beta \tag{4.36}$$

Introducing (4.15) into (4.36) and with the same stability analysis, it follows that $e_\beta \to 0$ implying that $\beta \to \beta_d$.

Introducing then (4.12) into (4.35), and making the same stability analysis, it yields that $e_x \to 0$ implying that $x \to x_d$.

Motor control layer

Notice that the physical UAV is controlled through four independent motor thrusts (angular velocities - ω_i), which collectively generate total lift and torques allowing it to control the six degrees of freedom: its attitude (roll (α), pitch (β), yaw (γ)) and its translations (x, y, z).

The motor control layer receives the values of torques and thrust computed by the different PIDs. These commands are mapped to motor rotational speeds $[\omega_1, \omega_2, \omega_3, \omega_4]^T$ through an allocation matrix A, such as

$$\begin{bmatrix} \tau_\alpha \\ \tau_\beta \\ \tau_\gamma \\ u \end{bmatrix} = A \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \end{bmatrix} \tag{4.37}$$

This allocation matrix depends on the configuration and physical parameters of the aerial vehicle (numbers of rotors, position, etc). In addition, this approach simplifies controller tuning and enhances

robustness by separating the slower attitude dynamics from the fast motor response. The last step, in the actuators control part, is to convert the obtained motor rotational speeds into voltage Power Width Modulation (PWM) signals that will be sent to the motor using a classical electronic brushless motor controller.

UAV control architecture

The UAV control architecture is performed via the Flair software [73] developed at the Heudiasyc laboratory, which provides a high-level interface to the UAV's flight controller. It allows for easy configuration of the PID parameters and provides a real-time feedback loop for monitoring the UAV's state.

4.2 | Noise generated by a quadcopter

While the previous section detailed the control architecture that governs the UAV's motion and stability, this section highlights the direct link between the UAV's control actions and the noise it generates. Indeed, the operation of motors and propellers, driven by continuous control adjustments, is a primary source of noise. Understanding the characteristics and sources of this noise is essential to design efficient signal processing strategies in the context of Sound Source Localization (SSL) using a microphone array directly mounted on the UAV body. Due to the proximity of the microphones to the noise sources, the Signal-to-Noise Ratio (SNR) is typically very low. The ego-noise produced by a quadcopter is generated by several physical mechanisms, mostly aeroacoustic in nature. These include tonal emissions linked to propeller rotation, broadband noise from turbulent flow around the UAV structure, and flow-induced noise on the microphones themselves. Additionally, external environmental conditions (e.g., turbulence, recirculation) and flight control strategies introduce variability and unsteadiness in the noise signal.

4.2.1 | Propeller-induced acoustic sources

A quadcopter is equipped with four propellers, each of which generates noise through its rotation and interaction with the surrounding air. Each propeller also generates an airflow, which in turn interacts with the UAV's structure. In aeroacoustics, the speed of the air flow around the propeller determines which of these noise sources dominates. A flow is typically described by its Mach number, defined as the ratio of the flow speed to the speed of sound in air. In the case of a rotating propeller, the highest Mach number occurs at the tip of the propeller blades, and is called the tip Mach number, defined as:

$$M_t = \frac{R\omega}{c} \tag{4.38}$$

where R is the propeller radius, ω is the angular velocity of the propeller, and c is the speed of sound in air. The propellers used in this work have a radius is approximately 0.1 m, and the speed of sound in air is around 340 m/s. The typical angular velocity of the propellers in our setup is around 600 rad/s, leading to a tip Mach number of approximately 0.18, which is relatively low and enables us to assess the relative importance of different noise mechanisms.

The aeroacoustic noise-generating mechanisms of a propeller have been widely studied in the aeroacoustics literature, and can be classified into four main sources: thickness noise, loading noise, blade–vortex interaction noise, and turbulence noise [74]. The first noise source is the thickness noise. It is a tonal noise at the Blade-Passing Frequency (BPF), generated by air being pushed outside the propeller disc as the blades rotate. For dual-blade propellers, the BPF corresponds to the frequency at which the

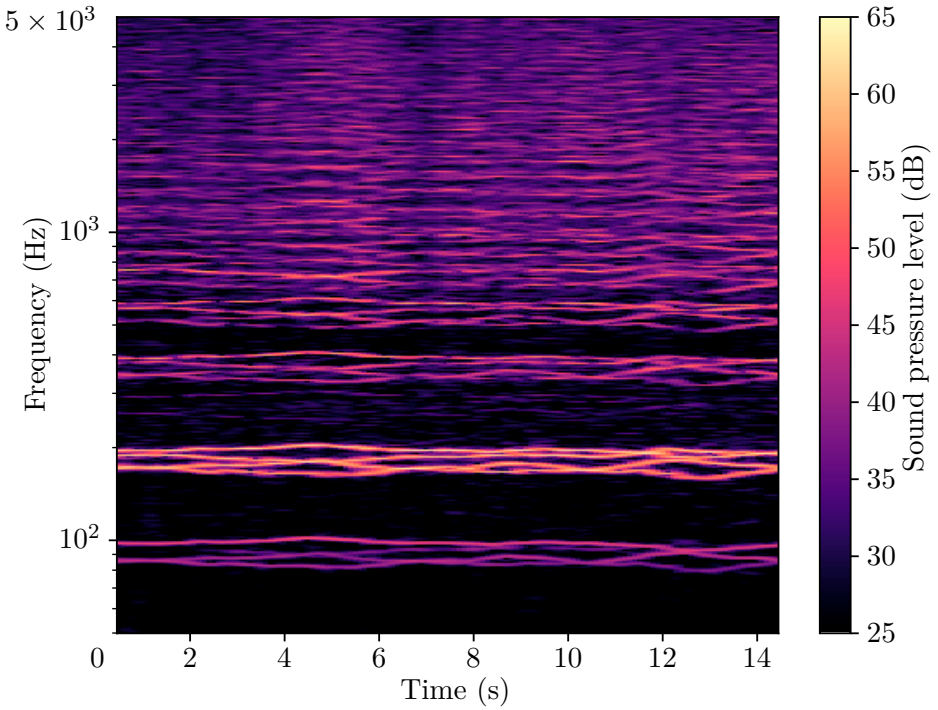
blades pass a fixed point in space. Note that each rotor spins at its own speed, and the BPF for each rotor is given by:

$$f_{\mathrm{BPF},i} = \frac{2}{2\pi}\omega_i = \frac{1}{\pi}\omega_i \tag{4.39}$$

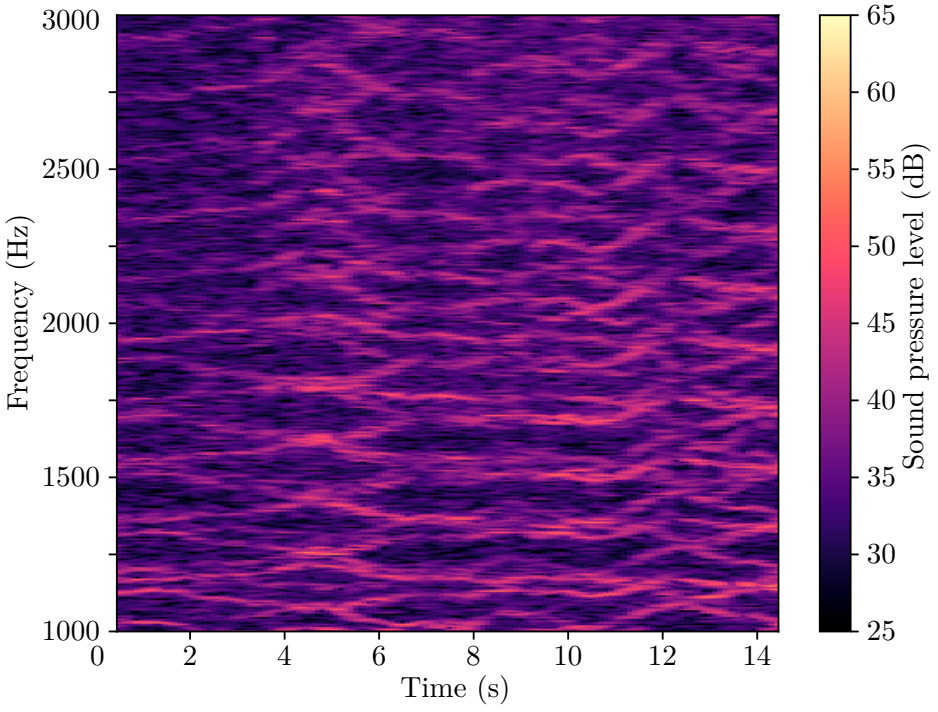
where ω_i is the angular velocity of rotor i. The thickness noise is typically weak at low Mach numbers such as in our case. The second noise source is the loading noise, which is both tonal and broadband, and is caused by the unsteady aerodynamic forces acting on the blades as they generate lift and thrust. This noise dominates at low Mach numbers, and is distributed along the whole blade. Remark that the tonal noise sources result from periodic rotor motion and structured aerodynamic loads. They create tonal features at harmonics of the rotor speed, which are modulated by control inputs described in Section 4.1.4. Zhong *et al.* [75, 76] demonstrated that even minor fluctuations ($\approx 1\%$) in rotor speed generate sideband harmonics comparable in amplitude to the main tone. These fluctuations arise naturally from flight control adjustments and result in frequency modulation of the tonal components. In addition, the rotor itself produces noise around its rotational speed frequency, ω_i, whose amplitude is typically lower than the BPF noise and its main harmonics. The third noise source is the blade–vortex interaction noise, which is a broadband noise generated by the interaction between the blades and vortices shed by preceding blades or wakes. This noise is typically stronger at higher Mach numbers, but can still be significant at low Mach numbers, especially when blade-wake interactions are strong, during forward movement for instance. It is located near the blade tip and along the trailing edge. Finally, the fourth noise source is the turbulence noise, which is a broadband noise generated by the turbulent flow around the propeller and the UAV structure. In free space, turbulence noise is generally negligible compared to the other sources, but it can become significant when turbulences interact with objects, such as the UAV body or the microphones themselves. A specific case is turbulence within a Turbulent Boundary Layer (TBL), where microphones lying on the same surface are simultaneously exposed to correlated pressure fluctuations. Note that these pressure fluctuations are not acoustic waves, but hydrodynamic pressure variations. Since turbulence propagates along the flow, microphones located within the same turbulent structure may record correlated noise. Several models have been proposed to predict this correlation, the most widely used being the Corcos model [77], which links the coherence of the pressure fluctuations between microphones to their spacing and the convective properties of the flow. In the present work, the array used (see Fig. 4.2) is open, meaning that the microphones do not lie on the same surface and therefore do not share the same TBL. As a result, TBL-induced noise is assumed largely uncorrelated, although it can not be neglected. If another configuration of microphone array were to be used, where the correlation of TBL noise could be non-negligible, strategies based on the Corcos model, inspired by the work of Dinsenmeyer *et al.* [78], could be implemented to specifically mitigate its impact. When scattered over the surface of the UAV body itself, the turbulences can also generate wideband noise, whose characteristics strongly depend on local geometry and flow conditions [79, 80]. This can become significant since the microphones are located close to the UAV body, where these phenomena occur. In addition, environmental influences, such as inflow turbulence ingestion, flow recirculation, and external perturbations (wind) further increase the variability. Kingan *et al.* [80] emphasized that this environment-dependent component of ego-noise makes generalized modeling difficult, especially in unknown or dynamic outdoor conditions.

To illustrate these noise sources, Fig. 4.4 shows spectrograms of UAV noise measured during hover in a hemi-anechoic chamber. The rotor-induced tonal noise (80–100 Hz) and the BPF tones (150–200 Hz) are visible. Despite the UAV holding position, frequency modulation is evident over time, and harmonic overlap becomes significant above 1000 Hz, gradually transitioning into broadband-like behavior, as seen in Fig. 4.4b.

To examine how this radiated noise scales with rotor speed, stationary flights with varying payloads were conducted and sound power level measurements were performed following the ISO3744 norm



a)



b)

Figure 4.4. Sound pressure level spectrograms of the hovering UAV noise in a hemi-anechoic room (a) from 50 Hz to 5 kHz, (b) zoomed in on the 1000 Hz to 3000 Hz frequency band. The signal was collected with a microphone situated 2 m away from the UAV.

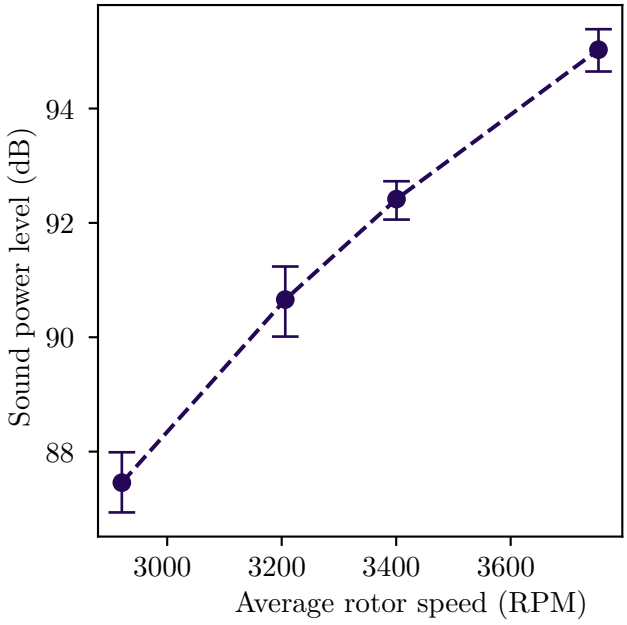


Figure 4.5. The sound power levels measured on the ground (outside the air flow) as a function of the average rotor speed.

[81]. As seen in Fig. 4.5, the emitted sound power level scales approximately linearly with rotor speed, validating a first-order proportionality between acoustic power and control input. Note that the acoustic power is measured sufficiently far from the UAV to neglect near-field effects or noise generated by the flow directly on the microphone surface.

The noise generated by a quadcopter is thus a complex mixture of various acoustic sources, which are spatially spread over the UAV body and the propellers, both tonal and wideband, and both correlated and uncorrelated over the microphones. This combination of characteristics make the noise generated by UAVs particularly challenging to filter out. A summary of the main operating UAV noise sources and their nature is given in Table 4.1.

Source	**Spectral nature**	**Dominance**	**Location**	**Spatial coherence**
Thickness noise	Tonal (BPF)	Low	Blade area	High
Loading noise	Tonal and broadband	High	Along the blade span	High
Blade-vortex interaction	Broadband	Medium	Blade tip, trailing edge	Low to moderate
Open air turbulence	Broadband	Low	Flow below the propellers	Low
Turbulence scattered	Broadband	High	Over the UAV's body	Low
Non-acoustic TBL excitation	Broadband	High	Not an acoustic source	Low

Table 4.1. Summary of the main UAV noise components at low Mach number, with their spectral nature, dominance, location, and whether their acoustic field is correlated over the microphones.

In situ analysis of UAV ego-noise

To analyze the unsteadiness and statistical nature of the ego-noise, motor controller outputs were recorded during stationary hover. These control signals, ranging from 0 (off) to 1 (full power), are directly linked to rotor speed, and hence to acoustic emission strength. Figure 4.6 displays time series and distributions of motor outputs. The distributions follow Gaussian shapes, with small variations around different mean values for each motor. These differences in mean values are attributed to asymmetries in mass distribution, e.g., due to battery placement, and to motor aging effects. Given the linear relationship between rotor speed and acoustic power established from Fig. 4.5, and between control signal and rotor speed, the acoustic output during hover can be modeled as a Gaussian process at first order. However, while total power appears stable, narrowband representations such as the CSM are more sensitive to unsteady features. Figure 4.4 shows that even under steady hover, spectral fluctuations are visible in single frequency bands.

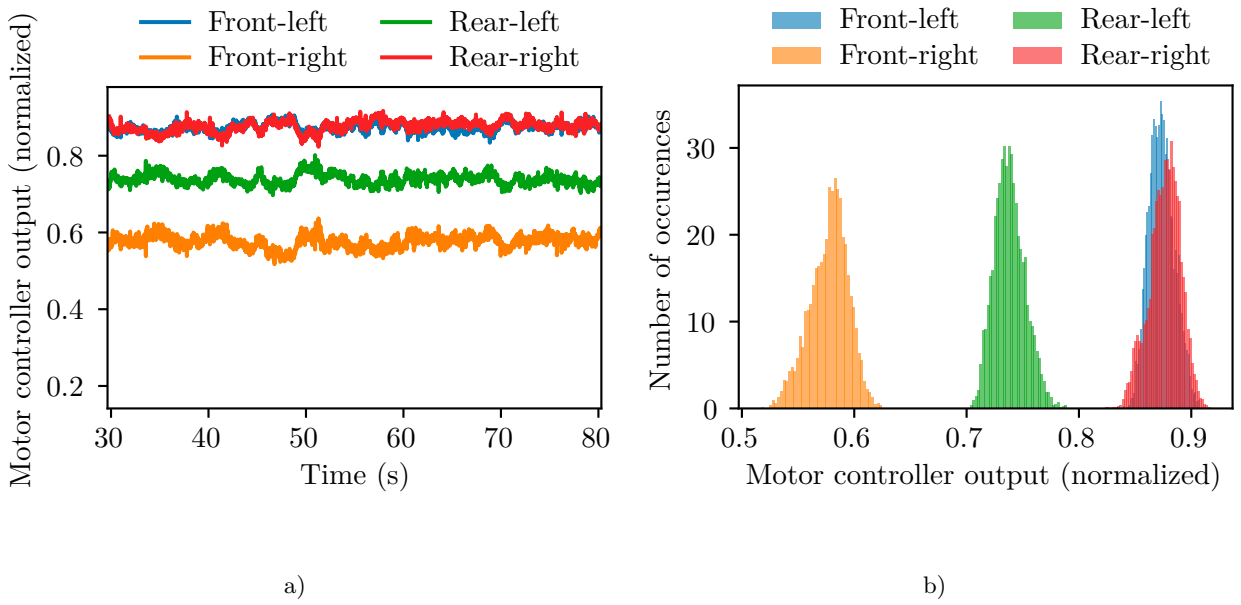
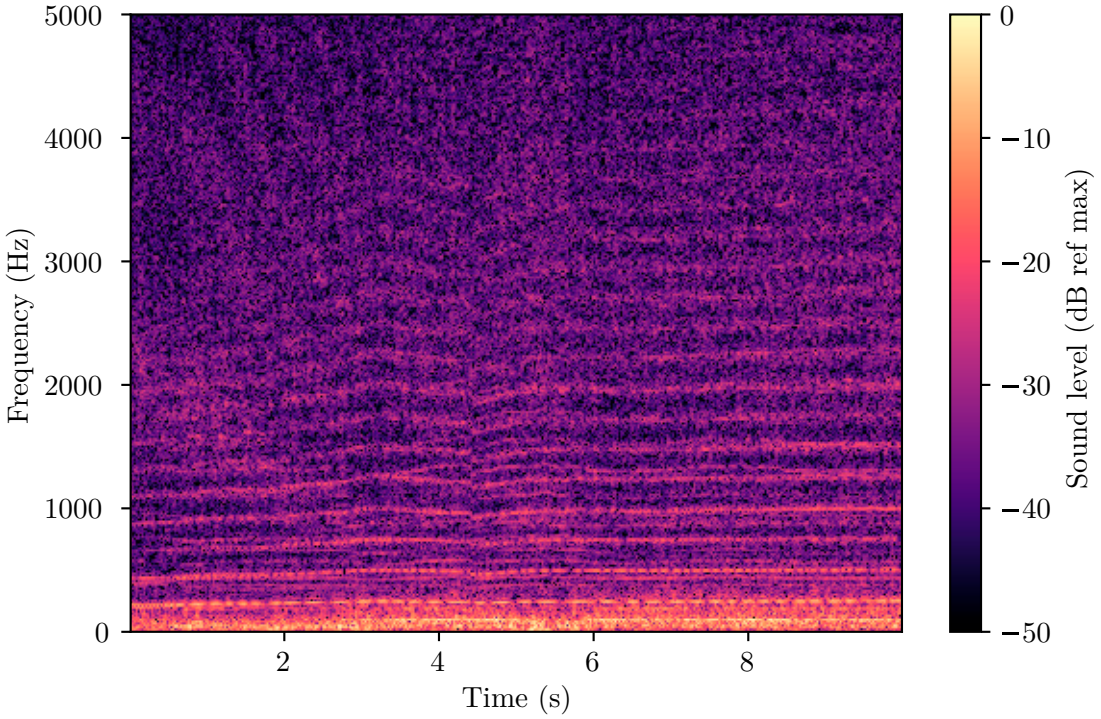
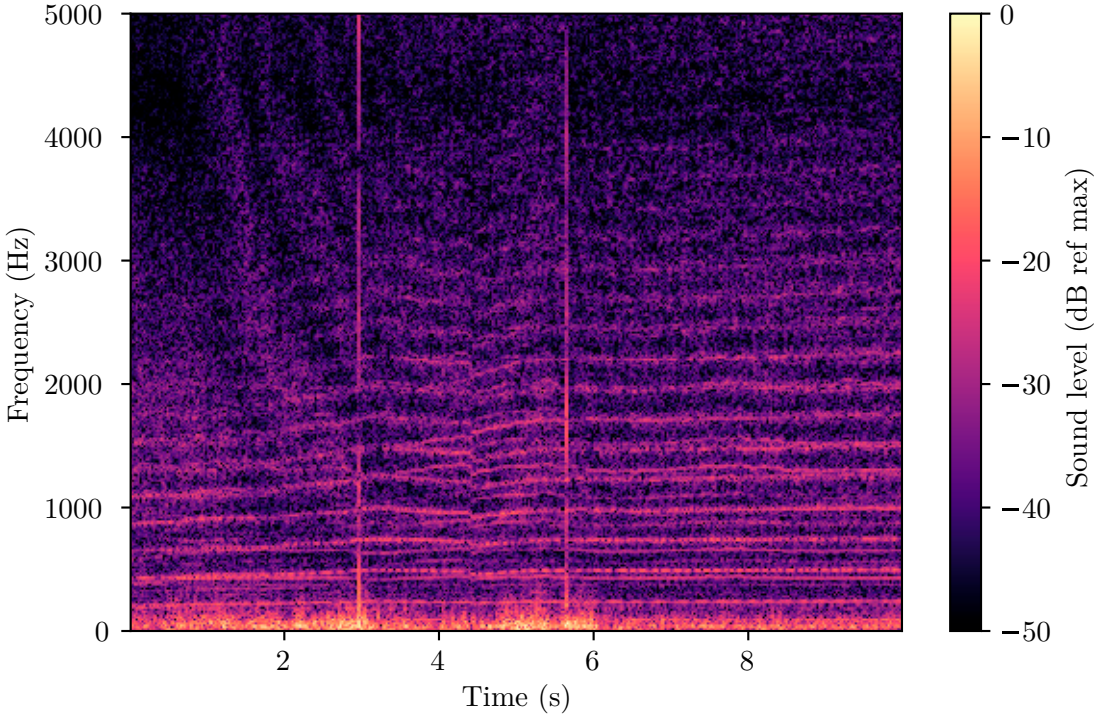


Figure 4.6. Data obtained from a stationary hover flight. (a) Normalized motor controller output for each motor as a function of time, and (b) distribution of the motor controller output values for each motor.

Therefore, although the ego-noise may be approximated as stationary in the time domain, care must be taken when interpreting its spectral statistics, particularly when estimating or interpreting the CSM used in acoustic imaging techniques. Figure 4.7 presents sound pressure level spectrogram captured by two microphones from the embedded array. It illustrates the different noise components at stake during a stationary hover maneuver. Tonal noise at the BPF and its harmonics emerges from the wide-band noise and appears to be of similar magnitude for both microphones. The wide-band noise is more important in the vicinity of the propeller's wake, as seen on microphone number 24, which is closer to the propeller than microphone number 28. The environmental and control-induced effects can be seen on both microphones around 4.5 s, where an abrupt shift in frequency occurs. This is attributed to noise on the position sensor, indeed, packet loss or misdetection of the UAV can cause a sudden change in the control scheme, leading to quick adjustment of the rotor speeds to realign the UAV with its desired position. Additionally, impact noise can be seen on microphone number 28, around 3 s and 5.5 s. This was also observed on nearby microphones, suggesting that an element on the UAV body was not properly attached and might have hit the microphone support structure near microphone number 28. Environmental effects can also be observed by comparing spectrograms from Fig. 4.4 and Fig. 4.7. The difference in frequency modulations is stronger in the first case. This could



a)



b)

Figure 4.7. Sound pressure level spectrograms of the hovering UAV noise recorded from the embedded array (a) from microphone number 24, (b) from microphone number 28. Both microphones are from the same branch, but microphone number 24 is closer to the propeller.



Figure 4.8. The Intel Aero quadcopter equipped with the 32 windshielded MEMS microphone array.

be explained by the stronger recirculation effects inside the hemi-anechoic room, which has a much smaller volume than the flight area where the second tests were performed. Moreover, in the first tests, the mobile OptiTrack setup was used, and only eight cameras were tracking the UAV, while 48 cameras were used in the flight area, making position estimation less noisy, hence a more stable motor control, leading to a more stable spectrogram.
These observations suggest that the broadband noise is dominated by the turbulence-induced noise on the microphone surface, which is lower in magnitude than the tonal noise generated by the propeller motion for frequencies between 100 Hz and 5 kHz. These order of magnitude considerations will be important later when designing the denoising strategy.

The great complexity of the mechanisms generating the UAV noise makes the task of modeling the pressures measured by an embedded microphone array challenging. Nevertheless, the noise generated by an operating UAV can be distinguished into two components, one correlated and one uncorrelated, both time-varying and potentially masking totally sound sources located on the ground. The following section presents the available solutions to mitigate the noise generated by the UAV.

4.3 | Ego-noise mitigation techniques

This section presents the method adopted to tackle the denoising of the microphone signals recorded by the embedded array. Passive denoising techniques are first presented before the numerical filtering method is introduced.

4.3.1 | Passive noise reduction methods

The uncorrelated noise generated in the TBL leads to high instantaneous pressure levels on the microphones. If these pressure fluctuations exceed the maximum input level of the microphone's analog-to-digital converter, the signal clips, which significantly degrade the quality of the cross spectra. To reduce this noise, windshields are added over the microphone surfaces, see Fig. 4.8. They help to reduce the flow speed near the microphone while letting the acoustic wave propagate through it.
Approaches using microphone arrays located relatively far from the propellers (either on the side [82] or beneath it [31, 83]) have shown success through experiments. In general, adding weight far from the Center of Mass (CM) of the drone reduces the maneuverability and makes the UAV harder to control. The most stable way to add the weight is to put it underneath the drone body. Using our hemispheric microphone array, we considered three options. The first one is to put the array directly in contact with the drone body, which is approximately 20 cm from its CM. Second, place it 40 cm below the CM, and third, 60 cm below the CM. These three configurations are tested within

the anechoic chamber at the Jussieu campus of the Institut ∂'Alembert. The portable OptiTrack setup is used to control the UAV's stationary hover over the fixed microphone array. The hover heights are set to 'low' (20 cm), 'medium' (40 cm), and 'high' (60 cm), and the drone maintains these positions for 15 s each. Microphone spectrograms are recorded and studied qualitatively to assess the pertinence of the array placement. The results are presented in Fig. 4.9. As expected, placing the microphone array further from the array slightly denoises its spectrogram, mostly its wide-band component. However, as introduced in Section 4.2, this wide-band component is most likely due to the TBL over the microphone surface, which can be conveniently mitigated using windshields and by using independent snapshots averaging when estimating the cross-spectral matrix. More specifically, on the 2 kHz band, the average power spectral density is reduced by half between the 'low' scenario and the 'high' scenario, corresponding to a 3 dB decrease. The benefits of placing the array further from the drone body, which is done at the cost of UAV maneuverability, were judged too dim to be implemented, thus the configuration where the array is directly attached to the UAV body was chosen. Other passive denoising techniques focus on parts design, and most notably on optimizing the propellers' design, or propellers placement [84]; however, since these methods are complementary to others, they were not studied in this work.

4.3.2 | Numerical noise filtering techniques

Before diving into the filtering techniques, it is worth mentioning the control-based noise reduction approaches. For instance, synchro-phasing techniques have shown promising results to mitigate far-field tonal noise emissions on a quadcopter UAVs by constraining the phase alignment of propeller blades during motion to minimize the radiated noise [85]. Additionally, active noise control (ANC) techniques have been proposed to reduce the noise radiated by UAVs by using additional microphones and speakers to generate destructive interference with the noise field [86]. Since these approaches are based on generating destructive interference, they typically work for far-field noise reduction, but are not fit for noise reduction on multiple microphones in the near field.

Numerical denoising of UAV-embedded microphone recordings has been drawing significant attention over the past decade. Often referred to as Drone Audition (DA), this field aims to provide drones with auditory capabilities, serving multiple purposes, either for wildlife monitoring [83], for search-and-rescue scenarios [24, 87], or for speech reconstruction [88].

While uncorrelated noise can be dealt with efficiently by increasing the number of snapshots, as explained in Chapter 1, correlated noise is typically difficult to tackle. The spatial filter introduced in Chapter 2 targets correlated noise coming from a point sound source. Having additional knowledge of the propagation operator between the source and the microphones, it could be adapted to tackle this problem. However, the previous section provided us with insights into how complex, heavily time-varying, and environment-dependent these sources are. Modeling propagation from these spatially spread sources to the microphones seems unrealistic, making the spatial filter unadapted. The following paragraph will be focused on the existing methods to denoise the measurement CSM from a correlated noise over the microphones.

The problem of measurement CSM denoising can be formulated as follows. From the analysis of the ego-noise, it follows that the Fourier transform $\mathbf{p} \in \mathbb{C}^{M\times 1}$ of the pressure measured by a microphone array embedded on a UAV, and in the presence of other Sources Of Interest (SOI), can be expressed as:

$$\mathbf{p} = \mathbf{p}_s + \mathbf{p}_c + \mathbf{p}_u \tag{4.40}$$

with $\mathbf{p}_s$ the pressure signal generated by the SOI, $\mathbf{p}_c$ the correlated component of the ego-noise, and $\mathbf{p}_u$ its uncorrelated component. Both $\mathbf{p}_s$ and $\mathbf{p}_c$ are assumed spatially correlated and temporally

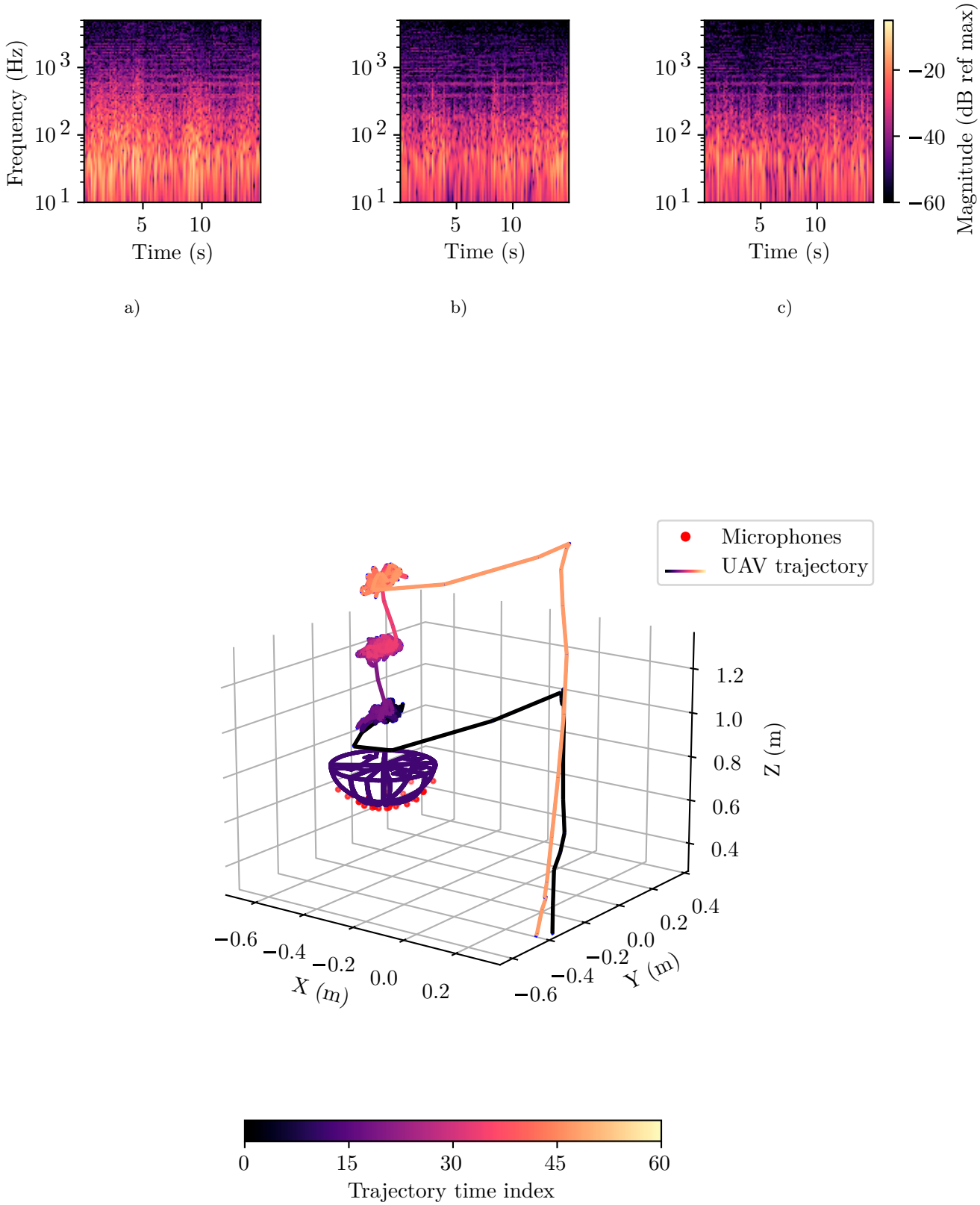


Figure 4.9. Spectrograms measured by a randomly picked microphone during stationary hover over the array at different altitudes: (a) low (20 cm), (b) medium (40 cm), and (c) high (60 cm). (d) The 3D rendering of the measurement setup and the drone trajectory over time.

coherent stochastic processes, whereas $\mathbf{p}_u$ is assumed to be a temporally and spatially white zero-mean stationary stochastic process. Under these assumptions, we can write

$$\mathbf{C} = \mathbb{E}\{\mathbf{p}\mathbf{p}^H\} = \mathbb{E}\{\mathbf{p}_s\mathbf{p}_s^H\} + \mathbb{E}\{\mathbf{p}_c\mathbf{p}_c^H\} + \mathbb{E}\{\mathbf{p}_u\mathbf{p}_u^H\} = \mathbf{C}_s + \mathbf{C}_c + \mathbf{C}_u \tag{4.41}$$

where $\mathbb{E}\{.\}$ is the expectation operator, $\mathbf{C}_s$, the SOI CSM, $\mathbf{C}_c$, the correlated ego-noise CSM, and $\mathbf{C}_u$, the uncorrelated ego-noise CSM which is a diagonal matrix of diagonal elements the power of the noise on each microphone. The theoretical ego-noise CSM is denoted $\mathbf{C}_\times = \mathbf{C}_c + \mathbf{C}_u$. The main strategy adopted by numerical denoising methods for microphone arrays consists of estimating the ego-noise CSM and using this information in a filtering scheme. To date, no consensus has been reached regarding a standard algorithm, although the two following methods appear to be predominantly favored in the existing literature addressing this issue.

Multichannel Wiener filtering

The first approach is to use the well-known Multichannel Wiener Filter (MWF) to reduce the UAV ego-noise. In [30], Manamperi *et al.* introduced an MWF coupled with a Gaussian Mixture Model-based Wiener Filter (GMM-WF) to enhance speech components in the time signals. The GMM-WF post-filter is based on a parametric Wiener Filter (WF) whose principle is to enhance speech by weighting the frequency spectra of the ego-noise and speech mixture using information on the ego-noise spectra. The GMM part is used to estimate the noise and speech power spectral density. In other words, it acts as a spectral smoothing algorithm that attenuates noisy frequency bins. For our purpose, since the information sought is the location of the sound source rather than a clean source signal, it would be more efficient to directly perform Direction of Arrival (DOA) estimation on the less-noisy frequency bins. Nevertheless, the MWF step could serve as a pre-filtering method before applying DOA estimation algorithms. The filter coefficients at a frequency f can be formulated as

$$\mathbf{W} = \frac{\mathbf{C}_\times^{-1}\mathbf{C} - \mathbb{I}_M}{1 + \mathrm{tr}(\mathbf{C}_\times^{-1}\mathbf{C}) - M} \tag{4.42}$$

where $\mathbb{I}_M \in \mathbb{R}^{M\times M}$ is the identity matrix, tr() denotes the trace operator of a matrix, and M is the number of microphones. In DOA estimation, the scaling of $\mathbf{W}$ is not relevant, thus, the denominator can be omitted. The denoised pressure signals can be obtained as $\tilde{\mathbf{p}}_{\mathrm{MWF}} = \mathbf{W}\mathbf{p}$.
The main drawback of this method is that it requires appropriate scaling of the ego-noise CSM due to the subtraction term in (4.42). In the original method, the ego-noise CSM is estimated in real-time during measurements, from the measurements when the speech source is off, which is impractical in our industrial setting.

Measurement CSM whitening

Rather than filtering the signal, another approach is to whiten the noise components in the measured CSM. Whitening transforms the measured CSM such that the ego-noise becomes spatially white, i.e., its CSM becomes diagonal, making it easier to isolate and process the target signal.
The principle of the Generalized EigenValue Decomposition (GEVD) filtering [26, 89] is to utilize a model of the noise CSM to whiten the ego-noise component in the measurement CSM. Based on the model from Eq. (4.41), the whitening operation of the ego-noise component in the measurement CSM, $\mathbf{C}$, writes

$$\mathbf{C}_G = \mathbf{C}_\times^{-\frac{1}{2}}\mathbf{C}\mathbf{C}_\times^{-\frac{1}{2}} = \mathbf{C}_\times^{-\frac{1}{2}}\mathbf{C}_s\mathbf{C}_\times^{-\frac{1}{2}} + \mathbb{I}_M \tag{4.43}$$

Conventional Beamforming (CBF) and MUltiple SIgnal Classification (MUSIC) algorithms can then

be applied to the filtered CSM, $\mathbf{C}_G$, to produce a beamforming map of the sources other than the ego-noise. Furthermore, assuming the ego-noise CSM can be estimated from an offline ego-noise-only measurement, this method's computational cost is virtually null. Indeed, applying beamforming to the filtered CSM, $\mathbf{C}_G$, is equivalent to applying beamforming using modified the steering vectors $\mathbf{g}_G = \mathbf{g}\mathbf{C}_{\times}^{-\frac{1}{2}}$, which can be computed offline. The topic of the ego-noise CSM estimation will be discussed later in a dedicated section.
Compared to the MWF method, the whitening approach is inherently more robust to mis-scaling of the ego-noise CSM. Indeed, a scaling error in the ego-noise CSM estimation would only scale the whole equation by a term $\kappa \in \mathbb{R}$, thus the term $\mathbb{I}_M$ would become $\kappa\mathbb{I}_M$, which does not degrade the performance of the DOA estimation algorithms.

4.4 | Practical implementation of denoising

It is important to highlight that Eq. (4.43) considers the theoretical CSMs, which can only be estimated in practice. This section investigates the sources of uncertainty in the CSMs estimation that will degrade the performance of the denoising algorithm when used in a realistic context.

4.4.1 | Estimation of the measurement CSM

Due to the real-time constraint, in experiments, the measurement CSM, $\mathbf{C}$, is estimated over a finite number of snapshots L spanning a finite acquisition time T_{acq}. The acquisition time is defined as $T_{\text{acq}} = \frac{1}{\Delta f}[1 + (L-1)(1-\rho)]$, with Δf being the spectral resolution of the discrete Fourier transform, and ρ the overlap ratio of the snapshots (a value of 0 meaning that snapshots do not overlap, and a value of 1 meaning they fully overlap). Based on the model in (4.40), the measurement CSM obtained by Welch's estimator [35] over L snapshots and an acquisition time T_{acq} is given by:

$$\hat{\mathbf{C}}(T_{\text{acq}}, L) = \frac{1}{L}\sum_{i=1}^{L} \mathbf{p}_i\mathbf{p}_i^H = \frac{1}{L}\sum_{i=1}^{L} (\mathbf{p}_{s_i} + \mathbf{p}_{c_i} + \mathbf{p}_{u_i})(\mathbf{p}_{s_i} + \mathbf{p}_{c_i} + \mathbf{p}_{u_i})^H \tag{4.44}$$

where subscript i denotes the snapshot index, and $\mathbf{p}_s$, $\mathbf{p}_c$, $\mathbf{p}_u$ denote the contributions from the SOI, the correlated ego-noise, and the uncorrelated ego-noise, respectively. Developing Eq. (4.44) yields:

$$\begin{aligned}
\hat{\mathbf{C}}(T_{\text{acq}}, L) &= \hat{\mathbf{C}}_s(T_{\text{acq}}) + \hat{\mathbf{C}}_c(T_{\text{acq}}) + \hat{\mathbf{C}}_u(L) + \hat{\mathbf{C}}_R(L) \qquad (4.45)\\
\hat{\mathbf{C}}_s(T_{\text{acq}}) &= \frac{1}{L}\sum_{i=1}^{L} \mathbf{p}_{s_i}\mathbf{p}_{s_i}^H\\
\hat{\mathbf{C}}_c(T_{\text{acq}}) &= \frac{1}{L}\sum_{i=1}^{L} \mathbf{p}_{c_i}\mathbf{p}_{c_i}^H\\
\hat{\mathbf{C}}_u(L) &= \frac{1}{L}\sum_{i=1}^{L} \mathbf{p}_{u_i}\mathbf{p}_{u_i}^H\\
\hat{\mathbf{C}}_R(L) &= \frac{1}{L}\sum_{i=1}^{L} \left(\mathbf{p}_{s_i}\mathbf{p}_{c_i}^H + \mathbf{p}_{c_i}\mathbf{p}_{s_i}^H + \mathbf{p}_{u_i}(\mathbf{p}_{s_i} + \mathbf{p}_{c_i})^H + (\mathbf{p}_{s_i} + \mathbf{p}_{c_i})\mathbf{p}_{u_i}^H\right)
\end{aligned}$$

where $\hat{\mathbf{C}}_s(T_{\text{acq}})$ is the estimated SOI CSM, $\hat{\mathbf{C}}_c(T_{\text{acq}})$ is the estimated correlated ego-noise CSM, $\hat{\mathbf{C}}_u(L)$ is the estimated uncorrelated ego-noise matrix, and $\hat{\mathbf{C}}_R(L)$ is a cross-term matrix that only depends on the number of snapshots and tends to zero as L increases.
The mutual uncorrelation of $\mathbf{p}_s$, $\mathbf{p}_c$, and $\mathbf{p}_u$, along with the self-uncorrelation of $\mathbf{p}_u$, implies that there exists a number of snapshots L_u beyond which the difference between the estimated uncorrelated

ego-noise covariance matrix and its true value becomes arbitrarily small. This can be expressed as:

$$\forall \varepsilon > 0,\ \exists L_u \in \mathbb{L},\ \forall L > L_u, \quad \left| \hat{\mathbf{C}}_u(L) + \hat{\mathbf{C}}_R(L) - \mathbf{C}_u \right| < \varepsilon \tag{4.46}$$

where $\mathbf{C}_u = \mathbb{E}\{\mathbf{p}_u \mathbf{p}_u^H\}$ represents the theoretical uncorrelated ego-noise CSM. In this case, since $\mathbf{p}_u$ is modeled as temporally and spatially white, each snapshot provides an independent realization. As a result, the accuracy of $\hat{\mathbf{C}}_u(L)$ increases as L increases, regardless of the total acquisition time T_{acq}, since the estimation of $\mathbf{C}_u$ depends solely on the number of independent snapshots.
Additionally, under the assumption of stationarity for the correlated ego-noise in experimental measurements, there exists an acquisition time $T_c \in \mathbb{R}_+$ such that, for any desired accuracy $\varepsilon > 0$, the estimated correlated noise covariance matrix $\hat{\mathbf{C}}_c$ converges to its theoretical counterpart $\mathbf{C}_c$ as the acquisition time T_{acq} increases. Formally:

$$\forall \varepsilon > 0,\ \exists T_c \in \mathbb{R}_+,\ \forall T_{\text{acq}} > T_c, \quad \left\| \hat{\mathbf{C}}_c(T_{\text{acq}}) - \mathbf{C}_c \right\| < \varepsilon \tag{4.47}$$

where $\mathbf{C}_c = \mathbb{E}\{\mathbf{p}_c \mathbf{p}_c^H\}$ is the theoretical correlated ego-noise cross-spectral matrix (CSM). In this context, since $\mathbf{p}_c$ is assumed to be temporally coherent, increasing the number of snapshots without extending the acquisition duration does not improve the estimation. The convergence of $\hat{\mathbf{C}}_c$ thus depends primarily on the total acquisition time.
To be more precise, the terms $\hat{\mathbf{C}}_s(T_{\text{acq}})$ and $\hat{\mathbf{C}}_c(T_{\text{acq}})$ also depend on the number of snapshots, but considering the high number of snapshots L_u, that is required so that $\hat{\mathbf{C}}_u(L) + \hat{\mathbf{C}}_R(L)$ converges, it is assumed that the convergence of $\hat{\mathbf{C}}_s(T_{\text{acq}})$ and $\hat{\mathbf{C}}_c(T_{\text{acq}})$ as a function of the number of snapshots is already sufficient.
As explained in Section 4.2 in the analysis of Fig. 4.7, it can safely be assumed that the magnitude of the uncorrelated noise can be expected to be lower than or equal to that of the correlated noise. Under this assumption, and considering that satisfactory uncorrelated noise reduction is achieved when the residual uncorrelated noise magnitude is at least ten times lower than that of the correlated noise, a rough estimation of the number of required snapshots can be obtained using a variance scaling argument. Assuming that the uncorrelated noise vectors $\mathbf{p}_{u_i}$ are realizations of a temporally and spatially white, zero-mean complex Gaussian process, and are uncorrelated with the signal component $\mathbf{p}_c$, both the measurement uncorrelated noise CSM $\hat{\mathbf{C}}_u(L)$ and the cross-term $\hat{\mathbf{C}}_R(L)$ are sample averages of zero-mean, uncorrelated random variables over L realizations. Under these assumptions, the magnitudes of their off-diagonal elements decrease approximately at a rate of $1/\sqrt{L}$. As a result, reduction of the uncorrelated noise by a factor 10 can be achieved by using 100 snapshots. For the rest of this work, the number of $L = 100$ snapshots is thus used to estimate the CSMs. Low that the number of snapshots L_u to mitigate uncorrelated noise is found, we still need to estimate the acquisition time over which the CSM can be considered stationary, T_c.

4.4.2 | Estimation of the ego-noise CSM

Choosing an appropriate estimator for the ego-noise Cross-Spectral Matrix (CSM) is not a trivial task due to the inherent variability in the noise generated by the UAV. Rather than focusing on complex learning-based methods, this work focused primarily on building up our understanding of drone noise recorded from an embedded microphone array. In that scope, the choice was made to study prerecorded acoustic data of the drone in stationary hover conditions. In Eq. (4.47), we hypothesized that despite the fluctuations in the estimated CSM, they oscillate around a well-defined mean value that characterizes the UAV's ego-noise under a given flight configuration. Specifically, this configuration corresponds to a stationary hover with constant payload and stable environmental conditions. This assumption is motivated by the observation that, while the UAV motors are constantly adjusting to

maintain hover, the control signals remain centered around stable average values. Figure 4.6 illustrates this by showing the controller outputs—interpreted as normalized voltages—over time, along with their distribution around the mean for each motor. This suggests that the system, though not perfectly stationary, might exhibit stationary behavior in the statistical sense, allowing us to model the ego-noise CSM as fluctuating around a central, representative value.

Under the assumption of wide-sense stationarity, a natural objective is to define a model of the ego-noise CSM that best represents the underlying process over time. This can be formalized as the problem of minimizing the mean-squared modeling error between the instantaneous measurement CSM $\hat{\mathbf{C}}(t)$ and a time-invariant model $\mathbf{C}$, expressed as:

$$\tilde{\mathbf{C}}_{\times} = \arg\min_{\mathbf{C}} \mathbb{E}\left\{ \left\| \hat{\mathbf{C}}(t) - \mathbf{C} \right\|_F^2 \right\}, \tag{4.48}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Under stationarity, the expectation is minimized when $\mathbf{C}$ is equal to the statistical mean of $\hat{\mathbf{C}}(t)$:

$$\tilde{\mathbf{C}}_{\times} = \mathbb{E}\left\{ \hat{\mathbf{C}}(t) \right\}. \tag{4.49}$$

Since the true expectation is generally not accessible in practice, it is approximated by the time-averaged CSM computed from a sufficiently long recording, which is

$$\tilde{\mathbf{C}}_{\times} = \hat{\mathbf{C}}(T_c, L_u, \mathbf{C}_s = 0) \approx \mathbf{C}_c + \mathbf{C}_u = \mathbf{C}_{\times} \tag{4.50}$$

The experimentally filtered matrix, $\hat{\mathbf{C}}_G(T_{\text{acq}}, L)$, becomes:

$$\hat{\mathbf{C}}_G(T_{\text{acq}}, L) = \tilde{\mathbf{C}}_{\times}^{-\frac{1}{2}} \hat{\mathbf{C}}(T_{\text{acq}}, L) \tilde{\mathbf{C}}_{\times}^{-\frac{1}{2}} \tag{4.51}$$

Naturally, the performances of the GEVD filter can be expected to be as good as the match between $\tilde{\mathbf{C}}_{\times}$ and the noise component of $\hat{\mathbf{C}}(T_{\text{acq}}, L)$, which is $\hat{\mathbf{C}}_c(T_{\text{acq}}) + \hat{\mathbf{C}}_u(L) + \hat{\mathbf{C}}_R(L)$, which directly depends on the acquisition time T_{acq} and the snapshot number L.

From our experience, the sufficient acquisition time T_c for $\mathbf{C}_c$ to converge is practically intractable. As an example, 100 s of the UAV ego-noise only was measured. Two non-overlapping signals of 30 s are extracted from this recording, and the CSMs $\hat{\mathbf{C}}_1$ and $\hat{\mathbf{C}}_2$ are computed from the two signals using 10,000 independent snapshots, over the 2 kHz frequency bin of spectral resolution 10 Hz. Comparing the relative difference between both CSMs using the Frobenius norm yields a 30% difference, i.e., $\frac{2\|\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_2\|_F}{\|\hat{\mathbf{C}}_1\|_F + \|\hat{\mathbf{C}}_2\|_F} \approx 0.3$, questioning the assumption of stationarity of the estimated CSMs.

4.4.3 | Temporal variations of the ego-noise CSM coefficients

This subsection investigates whether temporal trends in the CSM coefficient could be exploited. Since the CSM coefficients are complex, the evolution over time of their magnitude and phase components is studied. Using the previous 100 s recording from the embedded microphone array, 800 CSMs are computed on non-overlapping time windows of $T_{\text{acq}} = 0.125$ s, using L =100 snapshots, with a frequency resolution of Δf =10 Hz. The time variations of the magnitude and phase of a few selected coefficients are plotted in Fig. 4.10.

The coefficient magnitude varies over time and increases by high-amplitude bursts which are coherent on different microphones. The dynamics of these magnitude coefficients is up to 10 times the average value in amplitude. The phase seems to follow a temporal structure with additive noise, which might require additional investigations to fully characterized, for instance by studying the autocorrelation function of the noise overtime. If a temporal structure is observed, Kalman filtering could be a

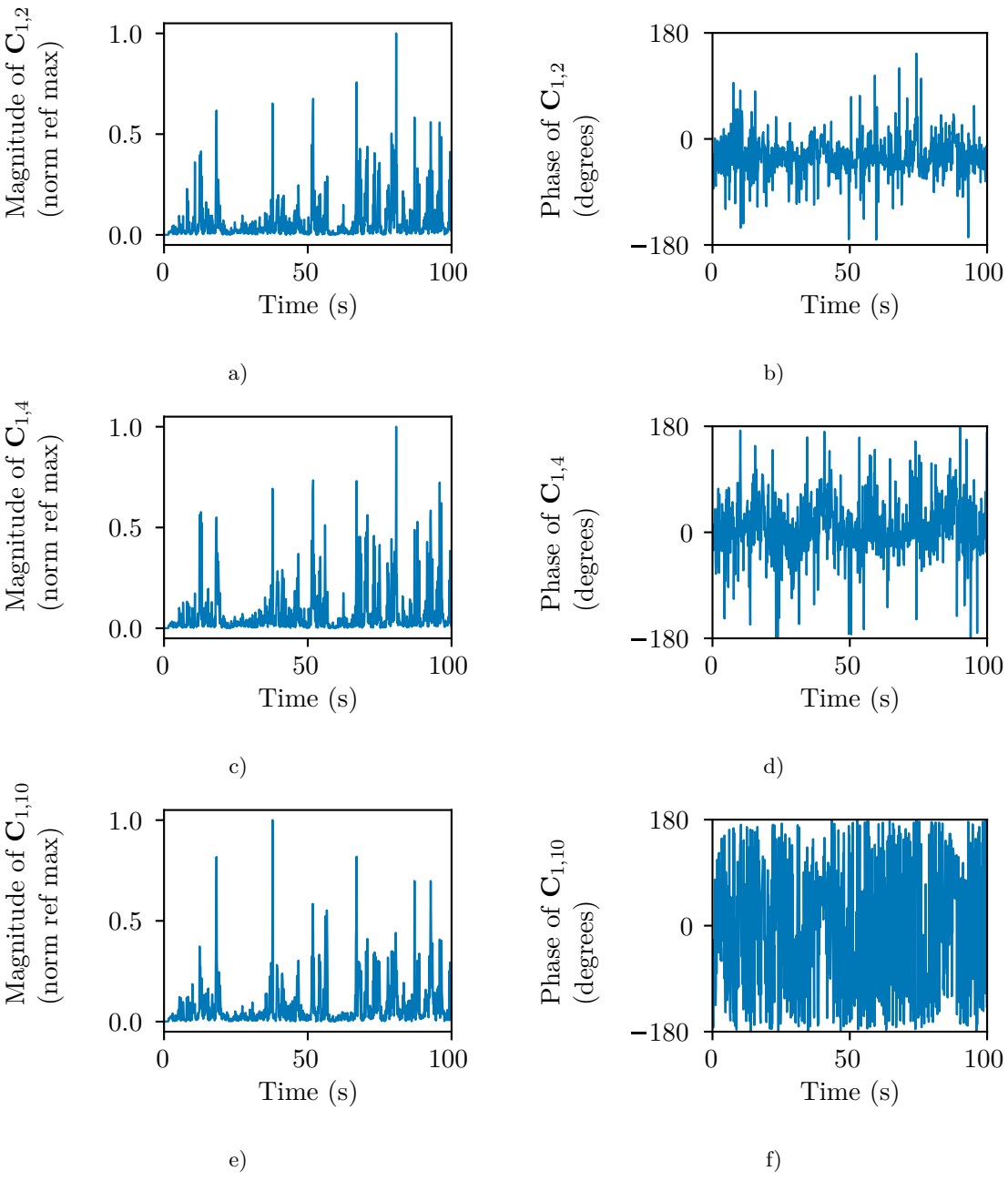


Figure 4.10. Time variations of the magnitude and phase term of the 2 kHz coefficient of the interpectrum for microphone pairs: (a) and (b) 1-2, (c) and (d) 1-4, and (e) and (f) 1-10.

promising candidate to model the CSM during the real-time measurements.
From Fig. 4.10, it is hard to consider that the CSM follows a stationary process over the studied time scales. Despite this, the whitening method using an ego-noise CSM computed using a single offline recording of the UAV obtained promising results, as will be illustrated in Chapter 5. The observations suggest that the ego-noise CSM can not be considered stationary. The denoising method would thus greatly benefit from using a more complex ego-noise CSM model that takes into account the temporal variations of the CSM coefficients.

More complex maneuvers

Based on the ego-noise generation mechanisms described in Section 4.2, the theoretical value of $\mathbf{C}_\times$ is expected to vary over time and depending on the UAV's control state. Specifically, it depends on two parameters: the average rotational speed of each motor, which is directly determined by the maneuver being performed, and fluctuations due to stochastic noise sources, as well as estimation errors. When performing different maneuvers, the average rotational speed of the motors changes, leading to a different equilibrium state of the ego-noise CSM. For instance, during forward motion, the rear motors must produce greater thrust than the front motors, creating a different equilibrium state for the CSM than stationary hover. In practice, using different ego-noise CSMs for different maneuvers does not significantly improve the performance of the denoising algorithm, as will be detailed in the next chapter. It is thus assumed that the impact on denoising of the type of maneuver performed is minimal compared to the stochastic sources and the CSM estimation errors.
Additionally, it was suggested by Zhong *et al.* [75] that a good acoustic design for UAVs should minimize the rotational speed fluctuations, which reduces the ego-noise amplitude. This is directly linked to the maneuvers performed by the UAV, and it calls for stable UAV motor control. In this sense, aggressive maneuvers, such as rapid accelerations or decelerations, is expected to lead to stronger ego-noise and maneuvers with motor control close to an equilibrium state, such as hover or translations in a plane, to lead to a weaker ego-noise.
More complex, time-varying maneuvers would, however, find limitations with a static noise model.

Influence of Maneuvers on Ego-Noise Statistics

As described in Section 4.2, ego-noise arises from both deterministic and stochastic mechanisms linked to UAV propulsion. Consequently, the ego-noise CSM $\mathbf{C}_\times$ is not strictly stationary but varies with the UAV's control inputs and flight dynamics.
Two main factors contribute to this variation: the average rotational speed of each motor, which depends on the maneuver being executed, and fluctuations due to turbulent flow and estimation noise. For instance, during steady forward motion, rear motors typically operate at higher speeds than front motors, creating a distinct control equilibrium compared to that of hovering. Each maneuver is thus expected to correspond to a different equilibrium state in the ego-noise CSM.
Despite this theoretical dependency, results presented in the next chapter show that using maneuver-specific ego-noise models does not significantly improve the denoising performance. The maneuver-induced differences in the CSMs are thus expected to be less impactful than the stochastic variations and estimation errors. Therefore, in this work, we adopt a single reference model for ego-noise $\mathbf{C}_\times$, regardless of the UAV's specific motion state.
Moreover, prior work by Zhong *et al.* [75] emphasizes the importance of minimizing motor speed fluctuations to reduce ego-noise. Since aggressive maneuvers involve rapid acceleration and deceleration, they tend to generate more intense and less predictable acoustic emissions. In contrast, smoother motions such as hovering or planar translations maintain the rotors closer to equilibrium, resulting in lower ego-noise levels.

Future work could explore adaptive or dynamic models for improved performance under highly non-stationary flight conditions. However, the chosen approach throughout this work focuses on smoothing the trajectory and minimize rapid changes in motor speeds, thereby stabilizing the ego-noise characteristics.

More complex models

Note that more complex ego-noise models can be used to approach the noise component of $\hat{\mathbf{C}}(T_{\text{acq}}, L)$ [90, 91], such as machine learning models trained on large experimental datasets, as done by Furakawa *et al.* in [92]. Since these models depend on both the microphone array geometry and the drone model, they require extensive measurement campaigns, which have to be performed every time the measurement setup or the drone model changes. Although suboptimal, the mean ego-noise CSM method has the advantage of requiring only a single ego-noise-only measurement, which can be quickly obtained in practice from flying the UAV in conditions approaching free field. In our case, it is measured in the flight area with the sources turned off, but in practice, in an outdoor localization scenario, we could imagine flying the UAV at a high altitude and far from any reflective surfaces so that free-field conditions can be obtained.

4.5 | Experimental validation

The aim of this section is to assess the performance of the whitening algorithm in different contexts. Each experiment presented here follows the same principle: find a sound source in the presence of an interfering sound source. The angular distance between the two sources, as well as the relative Sound Pressure Level (SPL) ratio, denoted as the Signal-to-Interference Ratio (SIR), and the nature of the noise source, vary. The first part focuses on point-noise-source , and the second part on a stationary hovering UAV as the noise source. For the sake of clarity, the noise source is denoted S_1, and the sound source to be found, S_2.

Experimental setup

First, the experimental setup is introduced. The data used in this section comes from the measurement campaign introduced in Chapter 2, Section 2.3.2, for the validation of the spatial filter. A quick reminder of the experimental setup is provided here. In this setup, two loudspeakers are positioned at a distance of 1.8 m from the microphone array. A reference microphone placed close to the microphone array recorded the scene to measure the SPL in different cases. It is used to compute the Signal-to-Interference Ratio (SIR), which translates how much louder the noise source is, compared to the sound source of interest.

Data processing

The processing of the data is performed as follows: a 1 s sample is extracted from the recordings of S_1 only, S_2 only, and the recording of their mixture $S_1 + S_2$. The direction of S_2 is obtained by applying beamforming to the S_2-only recording. A model of S_1's CSM is obtained from the S_1-only recording. Since these recordings have high SNR (very low background noise), the CSMs typically suffer from conditioning issues. Diagonal loading is employed to ensure the computation of the inverse of S_1's CSM converges. Diagonal loading consists of enforcing linear independence between vectors inside a matrix by adding a scalar term on the diagonal. It is used to artificially improve the conditioning of an inverse problem. Since the diagonal elements of the CSM are relatively irrelevant for the acoustic imaging process, it offers a convenient solution to increase the robustness of the computation of the

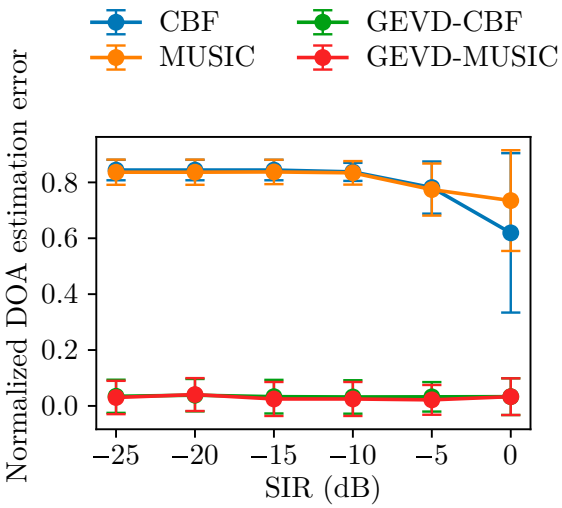


Figure 4.11. The normalized DOA estimation error average over SIR bands.

inverse of S_1's CSM. Tests were also performed without diagonal loading, and the DOA estimation results were significantly degraded, even for the 0 SIR case. The sampling frequency of the microphone array signals is 50 kHz. Since signals are stationary in the case of the loudspeaker noise source, the number of snapshots for the CSM computations is set to 10. All processing is performed on the 2 kHz frequency band of spectral resolution $\Delta f = 10$ Hz. For the $S_1 + S_2$ recordings, the 1 s signal is split into 8 non-overlapping windows of 0.125 s, on which acoustic images using the different algorithms are computed. The DOA estimation error for each image is computed and averaged over time, yielding one DOA estimation error value for each test case. This value is then normalized by the angular distance to enhance the interpretability of the results.

4.5.1 | Case study: noise point source

The first scenario studied is the incoherent noise point source case. In this scenario, S_1 and S_2 emit independent white noise, making the two sources incoherent. Following the methodology introduced in the previous paragraph, each test case was recorded and post-processed. Results are provided in Fig. 4.11. Results are averaged and aggregated in SIR bands, meaning that the results from test cases of similar SIR are averaged, regardless of the angular distance between sources. The 50 ° angular distance case was discarded due to a technical fault during the recording. These results evidence the importance of the whitening operation. Under these controlled experimental conditions, estimations of the CSMs approach their theoretical values, and DOA estimation error remains very low, even for extreme SIR. Additionally, CBF and MUSIC performance are similar, which is an expected result in high SNR conditions. These results allow us to consider that the whitening implementation using diagonal loading is successful, and that for this test case, MUSIC and CBF perform as good when associated with the whitening filter. Additionally, the influence of the angular distance between S_1 and S_2 was studied. Results are presented in Fig. 4.12. Except for an outlier at 50 ° which was not displayed, the DOA estimation error remains very low, down to an angular distance of 30 °, below which the algorithm is not able to separate the two sources. Since the main acoustic sources on the UAV are located at the propellers, the usable detection range is limited to 30 ° away from the propellers. Since we typically look for sources located in the hemisphere below the drone, this will not limit the performance of the algorithm in practice.

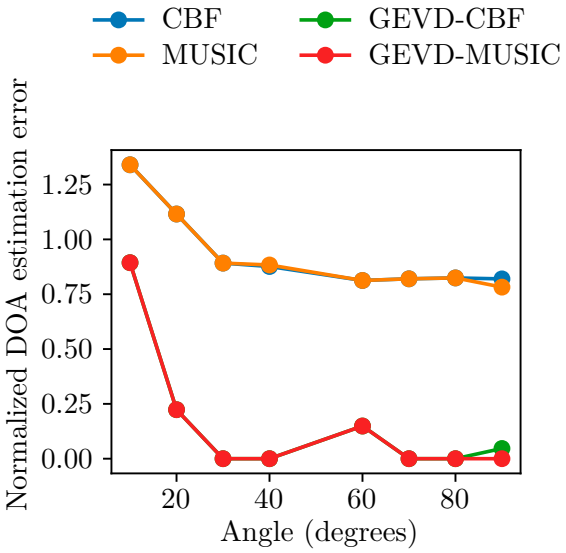


Figure 4.12. The normalized DOA estimation error for different angular distances between sources.

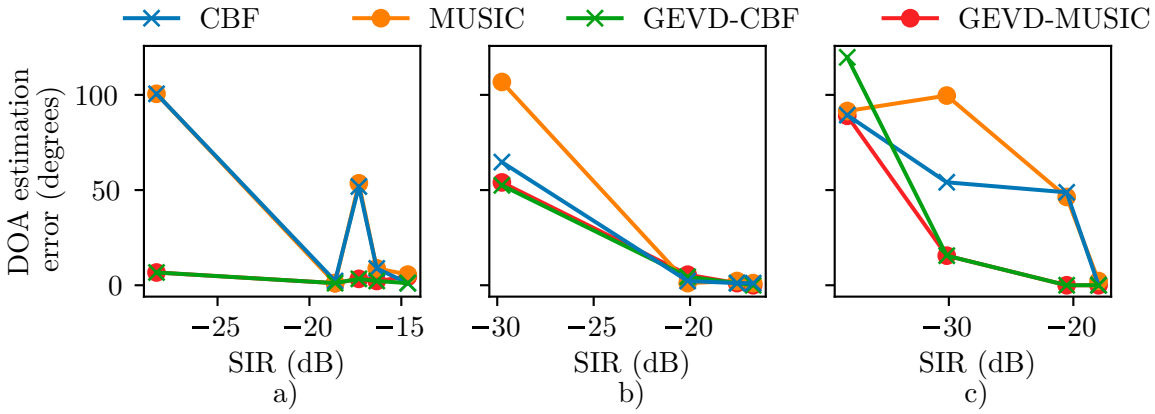


Figure 4.13. DOA estimation error for the different algorithms as a function of SIR for three stationary hover altitudes above the array: (a) low (20 cm), (b) medium (40 cm), and (c) high (60 cm).

4.5.2 | Case study: UAV interference during stationary hover

The second test case consists of emulating that the array is carried by the drone. The aim is to recover the correct direction of the loudspeaker S_2 while S_1 is replaced by a stationary hovering drone. To do so, the UAV presented in Section 4.1.1 performs a stationary hover at fixed altitudes above the microphone array. The data was collected during the flight operation presented in Fig. 4.9d. In this case, no diagonal loading is applied to the interference CSM since the conditioning of the problem is already sufficient from the turbulence noise. For the same reason, the snapshot number used to compute the CSMs is set to 100, and the acquisition time to estimate the UAV noise CSM is set to 25 s. Note that in this case, the source signal used is a 2 kHz sine wave, and the interfering noise is random noise with a non-constant power spectral density, making it hard to estimate. Thus, the SIR is not as relevant as in the previous case, since the power of the noise source is hard to estimate accurately. Additionally, SIR is affected by the uncorrelated noise generated by the UAV, which does not impact DOA estimation performance as much as correlated noise. When comparing two cases, one might have a higher uncorrelated noise level, hence a lower SIR, but a lower correlated noise level, which dictates the DOA estimation performance. Results in terms of SIR should be interpreted carefully. Figure 4.13 presents the results for each algorithm for three hover altitudes. The acoustic maps for each case have been checked manually to confirm that there was no fake positive detection of a source. From these results, it is clear that the DOA estimation performance is systematically enhanced or preserved when using the whitening algorithm. Even under UAV noise, CBF and MUSIC showed similar performances, which further motivates the use of CBF in real-time operation.

Conclusions on experimental whitening

These experiments demonstrated the benefits of using a noise-informed CSM whitening algorithm. This filter was proven to perform extremely well in a controlled environment polluted by a noise point source whose CSM is stationary and fully known. When the noise source was replaced by a stationary hovering drone, the performance degraded but still outperformed the classical CBF and MUSIC algorithms. This loss in efficiency is attributed to the randomness and non-stationarity of the UAV noise CSM. Additionally, both CBF and MUSIC achieved similar results on the cases tested. In conclusion, this study motivates the use of a CSM whitening algorithm to enhance the DOA estimation capabilities of our system.

4.6 | Discussion

In this study, we have explored the impact of UAV-generated noise on acoustic measurements and assessed the effectiveness of various denoising strategies. This section aims to interpret the main findings of the experimental validation and put them into context with the broader objectives of UAV-based acoustic applications.

First, the basic concepts of UAV operation were introduced, highlighting the inherent non-steadiness of the control inputs and hinting at the great complexity of UAV-generated noise. Based on the aeroacoustics literature, a comprehensive summary of the noise components was proposed, as well as the physical mechanisms at their origin. They were distinguished based on their location on the UAV, their dominance on the overall noise, their spatial coherence, and their spectral characteristics.

Several denoising approaches were reviewed critically in the scope of autonomous drone-embedded sound source localization. They were separated into two categories: passive denoising and numerical denoising techniques. The first relies on structural and material modifications, offering a first layer of mitigation by reducing the physical transmission of ego-noise to the microphones. However, they suffer from identified drawbacks, mainly reducing UAV maneuverability.

In contrast, numerical denoising approaches based on the estimation of CSM provided more robust and adaptable solutions. The separation of ego-noise and measurement CSMs is critical to the success of these techniques, and the choice of the ego-noise CSM model plays a major role. While a time-average model obtained from offline noise recording demonstrated convincing results, the high non-steadiness of the ego-noise cross-spectral matrix suggests that results could be greatly improved by including the additional information available (namely smooth control commands and system's dynamic state) into an adaptive ego-noise model, following the examples given in [90, 91, 92]. Research on the potential temporal structure of the noise CSM is ongoing, but has not yielded meaningful results yet. The presence of temporal structure could open the way for modeling the UAV noise dynamically, using, for instance autoregressive process or Kalman filtering.
A set of experiments within a controlled environment has been performed, and concluded that the cross-spectral matrix whitening algorithm was capable of attenuating the UAV interference significantly, both in the presence of a noise point source and under stationary hover conditions. This demonstrated the feasibility of our approach. Only a part of the data from the experimental campaign has been presented in this study, and further work will focus on exploiting it to explore the limits of the denoising algorithm more accurately.
Finally, although the experiments presented in this chapter focused on hovering UAVs, the next chapter considers scenarios involving dynamic flight to demonstrate the proposed method's capabilities with a realistic setup.

4.7 | Conclusion

This work addressed the challenge of ego-noise mitigation for UAV-based acoustic sensing systems by combining passive and numerical denoising strategies. After characterizing the noise sources generated by quadcopters, we proposed and evaluated a framework for estimating and subtracting the ego-noise from acoustic measurements using cross-spectral matrix analysis. Experimental validations confirmed the effectiveness of the approach in both controlled and realistic scenarios. The methods significantly improved signal quality, allowing for the detection of external acoustic events even in the presence of strong UAV interference. These findings contribute to the development of autonomous aerial platforms capable of reliable acoustic monitoring, with potential applications in industrial inspection, but also in various fields, such as wildlife monitoring, search & rescue, or human-robot interaction. Nevertheless, the practical deployment of such systems will require further efforts to enhance robustness in dynamic conditions, to reduce computational costs, and to improve the adaptability of noise modeling techniques.
The next chapter will focus on implementing the current approach to an autonomous UAV, integrating the denoising algorithm presented, and localizing a sound source from acoustic measurements.

Summary tables

The key achievements introduced in this chapter are summarized in Table 4.2, and the main perspectives discussed are listed in Table 4.3.

Aspect	Achievement
Ego-noise theoretical modeling	Summarized ego-noise generation mechanisms and their statistical properties.
Choice of the denoising method	Defined the most suited denoising algorithm available based on theoretical considerations.
Practical denoising concepts	Introduced a model of the online CSM estimation to explain the limitations of denoising.
Validation of the denoising algorithm	Demonstrated the benefits of CSM whitening for UAV ego-noise reduction.

Table 4.2. Summary of the main achievements of this chapter.

Aspect	Perspective
Ego-noise modeling	Identifying temporal structures in the noise CSM or correlation between the available state variables and the noise CSM would enhance the denoising capabilities.
Performance assessment	Finding an appropriate metric to quantify the performance of the denoising would allow generalizing the results introduced in this chapter.
Validation of the denoising algorithm	Demonstrated the benefits of CSM whitening for UAV ego-noise reduction.

Table 4.3. Summary of the main perspectives regarding UAV denoising.

This chapter laid the foundations for more optimized denoising, while validating a realistic strategy, achieving good performances in the presence of a drone hover in the vicinity of the microphone array. The next chapter will focus on implementing the denoising solution onto the UAV platform within the autonomous sound source localization framework presented in Chapter 2.

5. UAV-based sound source localization

Contents

5.1 Platform and validation 103
5.1.1 Experimental setup 103
5.1.2 UAV control 104
5.1.3 Path planning using true DOAs measurements 106
5.2 UAV-based real-time sound source localization 109
5.2.1 Experimental setup 109
5.2.2 Results 110
5.3 Discussions 114
5.4 Conclusion 115

The experimental results of the real-time Autonomous Sound Sources Localization (A-SSL) using an Unmanned Aerial Vehicle (UAV) are introduced in this chapter. The goal is to demonstrate the feasibility of real-time sound source localization from a UAV, using the path-planning algorithm and the triangulation scheme introduced in Chapter 2 and the denoising algorithm presented in Chapter 4. This chapter is structured as follows: first, the experimental setup and the validation of the path-planning algorithm using true DOA data are presented in Section 5.1. Then, the real-time sound source localization from a UAV is introduced in Section 5.2, using acoustically estimated DOAs to control the UAV dynamically. Finally, in Section 5.3, the results introduced in this chapter are discussed and perspectives of this work are presented.

5.1 | Platform and validation

In this first section, the experimental setup and the validation of the path-planning algorithm using true DOA data are introduced. The experimental results obtained allow to validate the implementation of the path-planning algorithm independently of the performance of the DOA estimation algorithm.

5.1.1 | Experimental setup

The experimental setup is similar to the one introduced in Chapter 4. The microphone array is the 32-channel Megamicros system, which is mounted on the Intel Aero Ready-to-Fly drone. The most significant addition to the setup is the communication architecture, which allows the UAV to receive the acoustic data required by the path-planning algorithm from the DOA estimation node deported on a ground station.

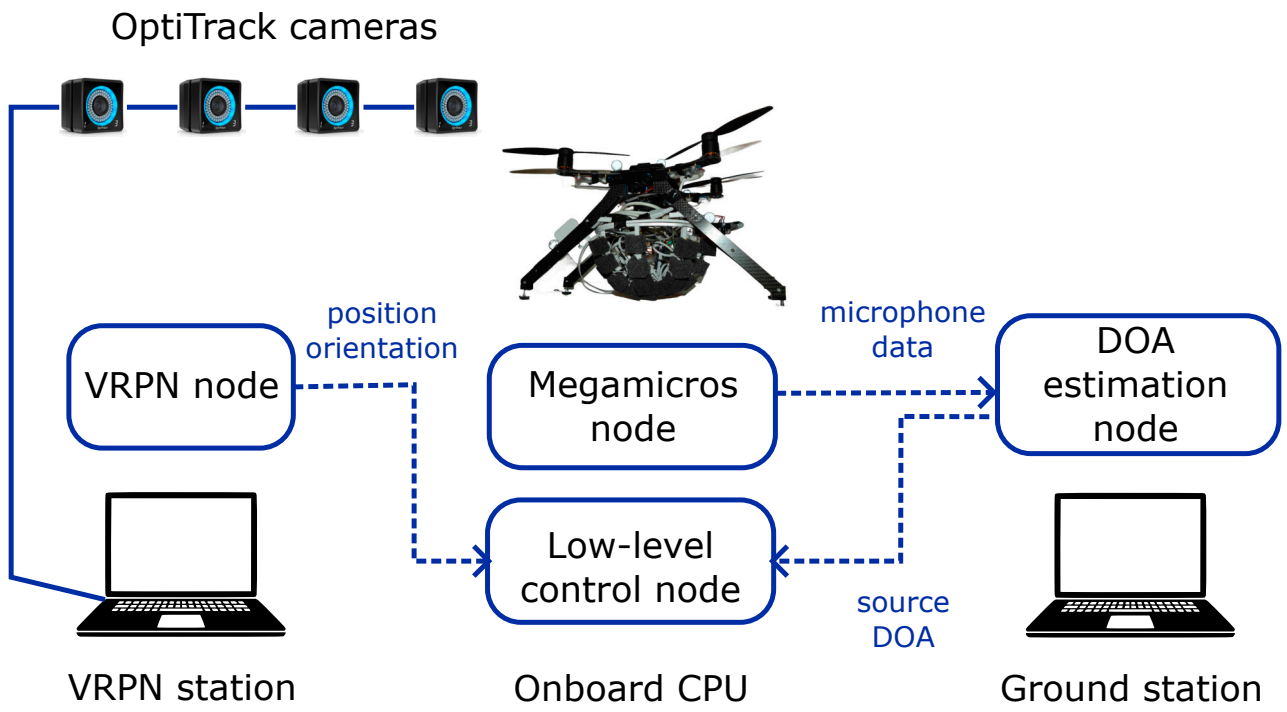


Figure 5.1. Diagram of the experimental setup and the communication architecture. Dashed arrows depict ROS topics broadcasting data via Wi-Fi. The arrow points from the publisher to the subscriber. The nature of the data transiting through the topic appears next to it in light blue. The ROS Master node, hosted on the ground station, is not represented for the sake of clarity.

Communication architecture

Similarly to the experiment on the ground vehicle introduced in Chapter 3, the processing of the acoustic data is performed on a remote platform. The acoustic signal recorded by the Megamicros acquisition system are sent to the ground station, which computes the current DOA and sends this information back to the UAV. The UAV then computes its next position based on the source DOA and its current position measured by the OptiTrack and broadcasted using the VRPN ROS node. The next desired position is obtained using the path-planning algorithm introduced in Chapter 2. The communication architecture follows the same distribution as the ground vehicle setup, except that the on-board data management and light computation are performed by the drone CPU directly. Since most of the setup remains unchanged from the ground vehicle experiment described in Chapter 3, only the major differences are listed here. First, the control node only takes the source DOA as an input from the acoustic imaging node. It uses the position and orientation measured by the OptiTrack to compute its next desired position. This desired position is updated each time a new DOA is obtained. The DOA information is sent directly to the UAV using the Wi-Fi network using a socket communication under the TCP protocol. The denoising of the acoustic recordings is performed as a preprocessing step by the acoustic imaging node. A summary of the setup is presented in Fig. 5.1.

5.1.2 | UAV control

The drone controller follows the cascaded PID architecture introduced in Chapter 4. Given the direction of the sound source, the UAV computes its next position using Algorithm 2, similarly to the ground vehicle. In Chapter 3, we compared the trajectory of the ground vehicle with simulations and observed that the distance at which the vehicle circled the source matched its steering radius, regardless of the step size used in the path-planning algorithm. Here, the term step refers to an effective step size, which depends on the UAV's speed and the acquisition time. It corresponds to the distance between two consecutive measurement positions. Unlike the ground vehicle, the quadcopter

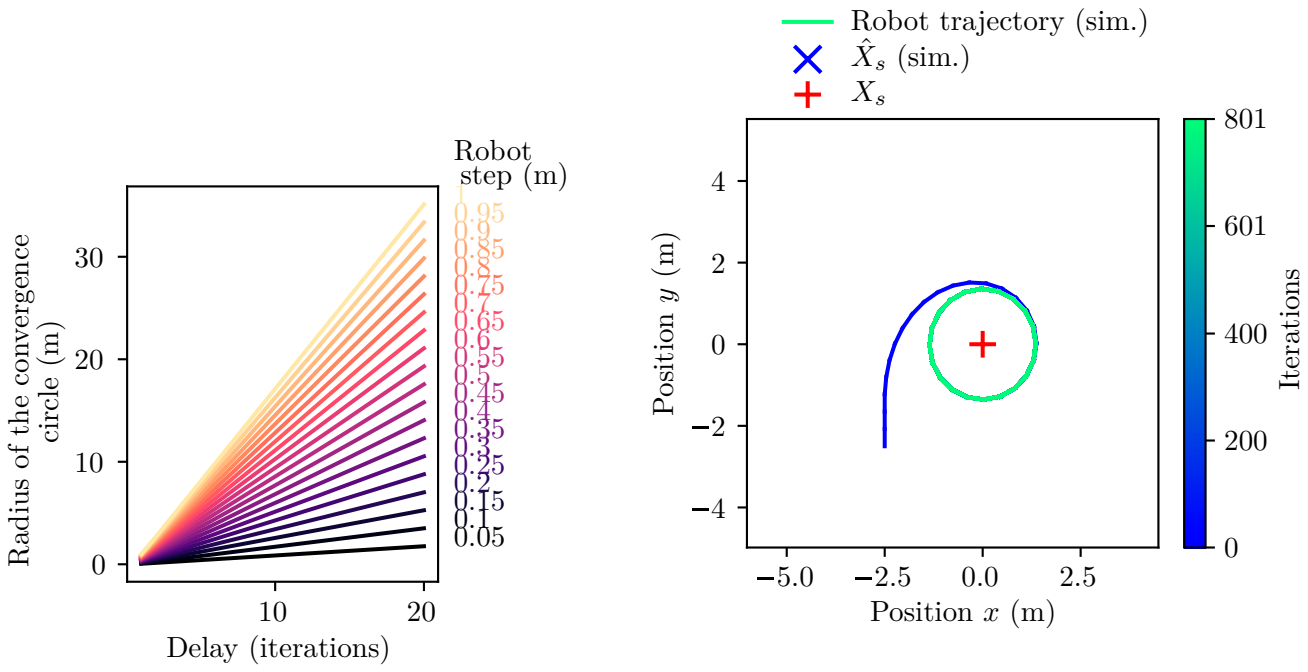


Figure 5.2. (a) Evolution of the convergence radius as a function of the delay introduced on the DOAs, for different values of step. (b) A sample simulation of the vehicle trajectory while introducing a 3-iterations-lag on the DOA estimation and using a step of 0.3 m. These simulations neglect the vehicle's dynamic.

has six degrees of freedom, enabling it to efficiently control its position and orientation. Under perfect control and assuming accurate adherence to the trajectory planned by the path-planning algorithm, the drone is expected to spiral around the source and reach a position on a narrow circle above it. This configuration, however, prevents triangulation if the circle is too narrow, as explained in Appendix B. In practice, the UAV's inertia makes tight circular maneuvers difficult to execute with perfect accuracy. Fortunately, this causes the UAV to 'naturally' converge to a fixed distance from the sound source. Since this effect originates from dynamic constraints, the convergence radius depends on the UAV's speed and, consequently, on the control parameters. This opens the possibility of tuning the convergence radius by adjusting the controller.

Additionally, while exploring the effect of delays in the control loop, we observed that introducing an artificial delay between the DOA estimation and its use in the path-planning algorithm also increases the convergence radius. Deriving analytically the behavior of the robot in these conditions can prove to be difficult, and our attempts remained unsuccessful. However, the behavior of the robot was observed in simulations and can be described efficiently using a simple proportional model illustrated in Fig. 5.2. In these simulations, the path-planning module uses DOA estimations lagged from a given number of iterations earlier rather than the most recent value.

These results confirm that the convergence radius increases proportionally with both the introduced delay and the robot's step size. This supports the idea that the radius can be explicitly increased by introducing lags into the DOA processing pipeline, which does not influence the triangulation performances since triangulation still uses the non-lagged DOAs.

In practice, some delays are inherently present in the system. DOA estimation is performed on a remote computer, which requires processing time and communication latency to send results back to the UAV. Furthermore, since measurements are collected during motion, the DOA estimation for a given frame is associated with the endpoint of the UAV's trajectory over that frame, introducing another source of effective lag.

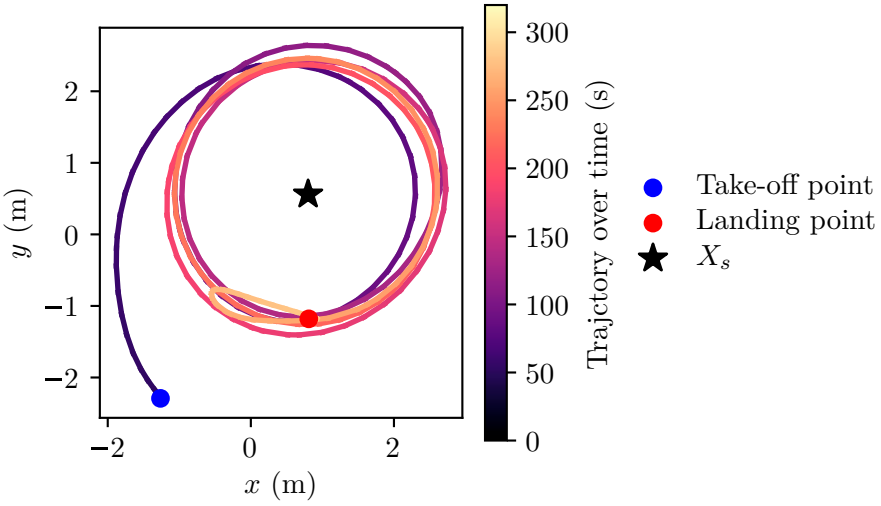


Figure 5.3. UAV trajectory around the sound source during the experiment. The UAV follows the path-planning algorithm, which uses the true DOA of the source to compute its next position. The trajectory is a circle of radius 1.5 m centered on the source.

In conclusion, the UAV's convergence distance around a sound source, which is an important parameter for accurate triangulation, can be controlled either through its dynamic behavior or by introducing intentional lags in the DOA estimation used for navigation. Importantly, neither method degrades the final localization performance, as the non-lagged DOAs can still be used for triangulation independently of control. In our experiments, the naturally observed convergence radius was satisfactory, so no additional adjustments were made. However, if needed, this radius can be modified as follows: to widen the convergence circle, increase the UAV speed or introduce DOA delays; to tighten it, reduce the speed. The UAV's speed could be varied through the PID gains on the x and y axes, the tilt and roll saturation bounds, and the step size used in the path-planning algorithm.

5.1.3 | Path planning using true DOAs measurements

Before testing the path-planning algorithm using acoustically estimated DOAs, we first tested the algorithm using true DOAs information. The true DOAs are obtained from the optical setup, which records the positions and orientation of the UAV and the source. This true DOA is given to the path-planning algorithm, which computes the next position of the UAV. Since the precision of the OptiTrack system is very high, the DOAs are very close to the truth and can be considered as the true DOAs. This experiment allows to validate the implementation of the path-planning algorithm and that the UAV follows the expected trajectory. During this experiment a wireless Jabra loudspeaker emits a 2 kHz sine wave at a constant amplitude. It is considered a point source. The Megamicros system records the acoustical data in order to assess the performances of the denoising and of the acoustic imaging algorithm using the path-planning using the true DOAs. Figure 5.3 shows the trajectory of the UAV around the sound source during this experiment.

Control evaluation

As expected from Section 5.1.2, the UAV follows a circular trajectory around the source, with a radius larger than its theoretical value of $\frac{d}{\sqrt{2}} \approx 0.2$ m for an approximate value of the effective step of $d \approx 0.3$ m. This is due to the inertia of the UAV, which makes it hard to control dynamically, especially when performing narrow circular maneuvers. Moreover, lags might have been introduced by

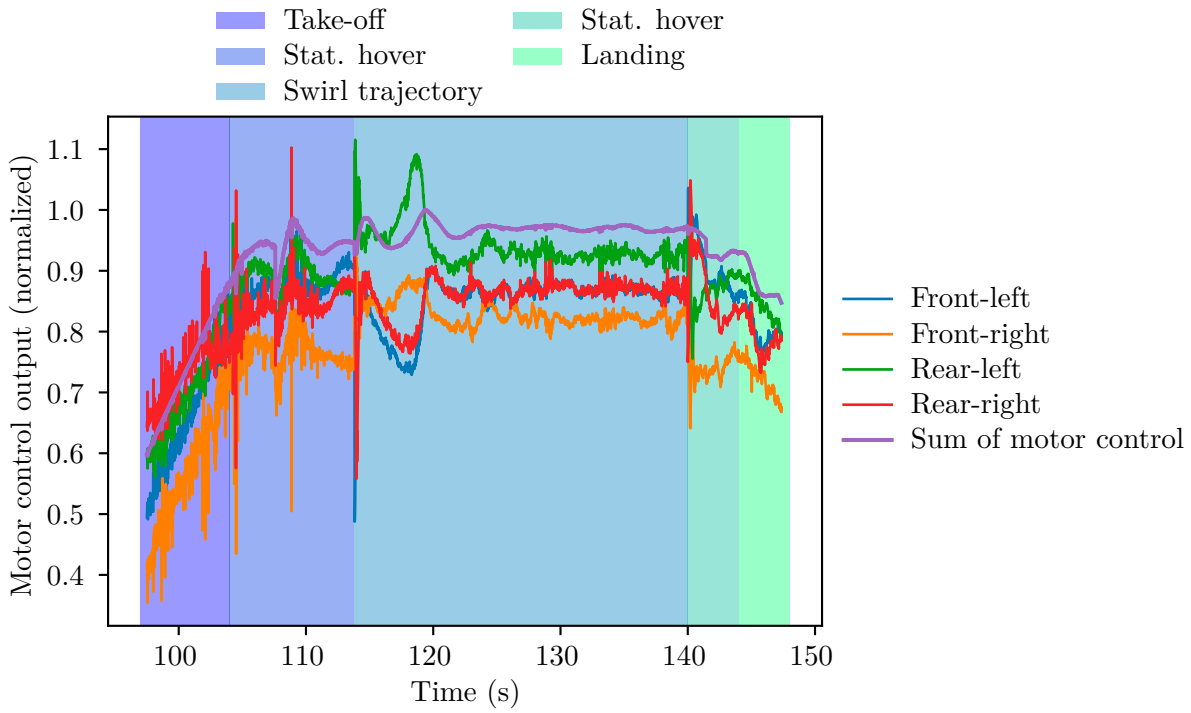


Figure 5.4. Motor controller output during the experiment using the true DOA data.

the communication architecture, which would also increase the radius of convergence. Nevertheless, this experiment demonstrated that the UAV is able to follow a circle trajectory around the source, using the direction of the source obtained from a reference, the OptiTrack system.
Chapter 4 introduced the physical link between the control algorithms and the noise emitted by the UAV. It is thus interesting to observe the output of the motor controllers during this flight. They are shown in Fig. 5.4. The different flight phases are highlighted by a background color.
During take-off, the control sent to the motors is highly unsteady and increases gradually until it starts flying. It is followed by a stationary hover phase where the UAV switches from manual control to autonomous control. Around 115 s, the UAV starts moving toward the source using the DOA obtained from the optical system. During that phase, the motor control output becomes stable around a mean value. This is attributed to the path-planning algorithm, which translate to a steady command of the drone motors. Indeed, by following a circular trajectory and always facing the source, the movement to be performed, seen from the UAV frame, remains constant. Finally, the UAV stabilizes above the landing point and lands.

Acoustic evaluation

From the control perspective, this experiment demonstrated that the UAV is able to perform the desired trajectory around the sound source. However, interesting information can also be extracted from the acoustic data recorded during this experiment. The denoising algorithm presented in Chapter 4 is applied using the 100 s pre-recorded data containing the ego-noise of the UAV. Additionally, Conventional BeamForming (CBF) is applied to the raw data as a reference. The acoustic processing parameters are the ones employed in the ground vehicle experiment, and they are summarized in Table 5.1. This set of parameters satisfies real-time processing.
Figure 5.5 shows the spectra of the acoustic recordings averaged over each microphone, at different times during the experiment. Since the acoustic imaging is performed on a narrow frequency band centered on 2 kHz and of spectral resolution $\Delta f = 10$ Hz, the Signal-to-Interference Ratio (SIR) is observed on that narrow frequency band to denote how challenging the configuration is for acoustic

Parameter	Value	Definition
M	32	Number of microphones
δ_ϕ	1°	Search grid resolution
T_{acq}	125 ms	Acquisition time
f_s	10 kHz	Sampling frequency
n_f	1	Number of frequency bins
Δf	10 Hz	Frequency resolution
L	50	Number of snapshots

Table 5.1. Acoustic processing parameters.

imaging. From Fig. 5.5, we observe that the SIR at the landing position, which is located on the convergence circle, is around -5 dB, indicating unfavorable conditions for acoustic imaging.

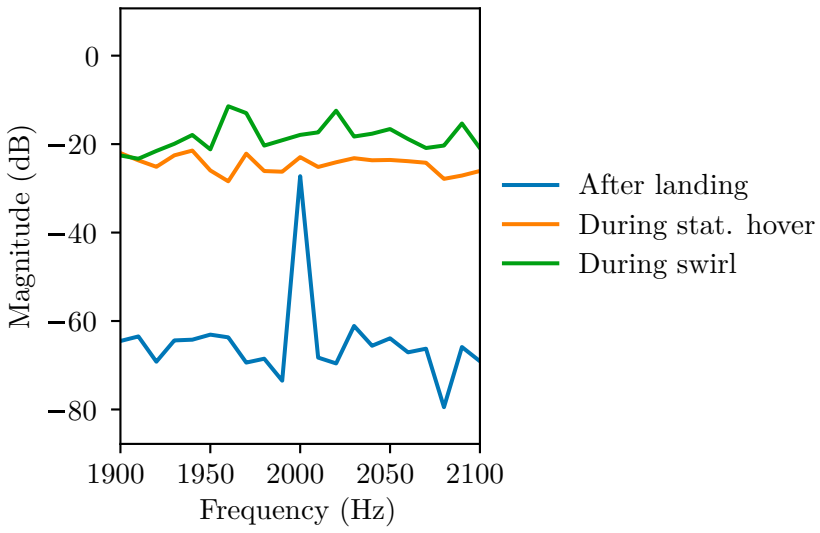


Figure 5.5. Spectra of the acoustic recordings averaged over each microphone, at different times. The frequency resolution matches the one used in the acoustic imaging algorithm, $\Delta f = 10$ Hz.

In order to evaluate the relevance of the denoising algorithm, the DOA of the source is computed and compared to the ground-truth DOA obtained from the optical system. The estimation error between the ground-truth DOA and the estimated DOA is defined as the angle between the two vectors representing the DOAs. It is plotted in Fig. 5.6 for both the denoised (GEVD) and raw acoustic imaging algorithm (CBF).

We observe that this angle error is consistently reduced when using the GEVD-CBF algorithm compared to the CBF algorithm. The standard deviation of the error during the swirl trajectory is 47 ° using the raw CBF algorithm, and 23 ° using the GEVD-CBF algorithm. Having a reliable DOA estimation will be crucial when the command of the UAV will be based on the acoustic imaging results.

The post-process was also performed using a pre-recorded dataset of ego-noise while the UAV performed a swirl maneuver, rather than a stationary hover maneuver. However, the results did not show any significant difference. This suggests that the difference between stationnary hover-produced noise and swirl-produced noise is not significant with regard to the other sources of uncertainty and non-steadiness mentioned in Chapter 4.

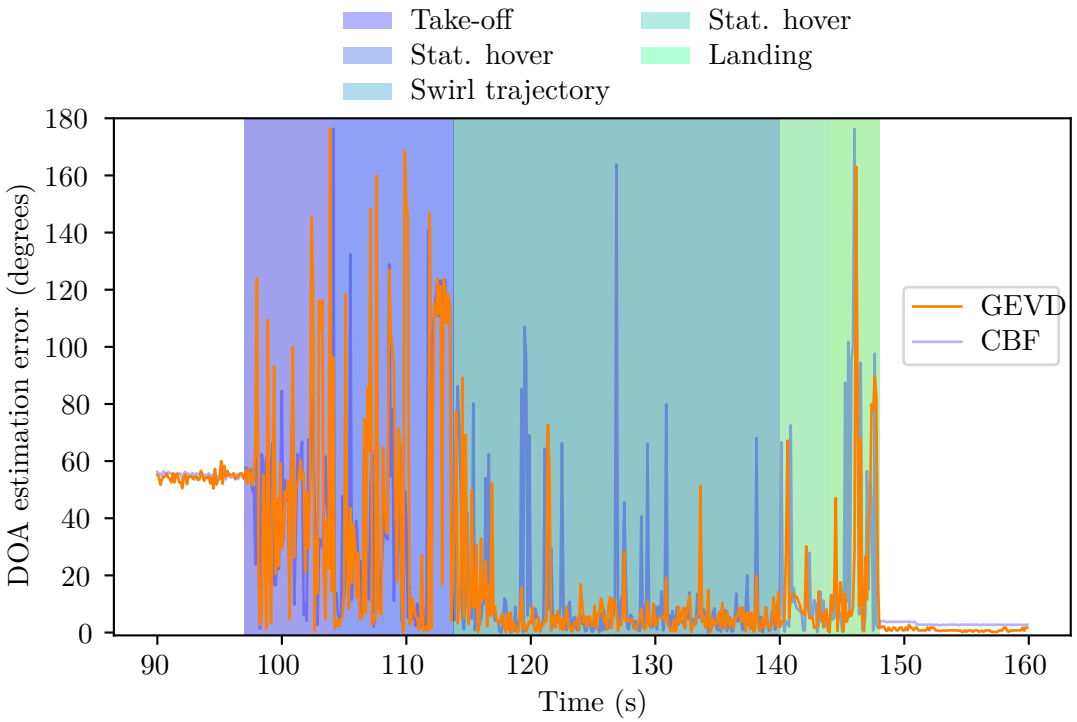


Figure 5.6. Angle error between the ground-truth DOA and the estimated DOA using the GEVD-CBF algorithm and the CBF algorithm. The background color denotes the flight phases. Standard deviation of the angle error during the swirl trajectory phase is 47 ° using the CBF algorithm, and 23 ° using the GEVD-CBF algorithm.

Conclusion

This first experiment allowed to validate the path-planning algorithm for a UAV using true DOA data, and to assess the relevance of the denoising algorithm presented in Chapter 4 for the UAV ego-noise under this smooth trajectory planning conditions. It also highlighted the benefit of the converging trajectory planning.
The next step is to test the path-planning algorithm using acoustically estimated DOAs, which will be used to control the UAV in real-time, the main concern being that a noisier control might lead to stronger and less predictable ego-noise, which in turn could affect the DOA estimation and the path-planning algorithm. Additionally, triangulation will be performed using the DOA estimated from the acoustic data in real time.

5.2 | UAV-based real-time sound source localization

This section introduces the experimental results of the real-time sound source localization from the acoustic-controlled UAV.

5.2.1 | Experimental setup

For this experiment, the UAV is given the task to recover the position of a sound source located on the ground. The same setup as the previous experiment is used, with the Megamicros system recording the acoustic data and the OptiTrack system providing the position and orientation of the UAV and the source. The proportional gains of the PID controller are set to lower values to reduce the speed of the UAV to avoid damaging the equipment. As a result, the convergence radius is expected to be smaller than the previous experiment.

5.2.2 | Results

During this experiment, the UAV applies the path-planning algorithm using the acoustically estimated DOA of the source. It followed the trajectory shown in Fig. 5.7.

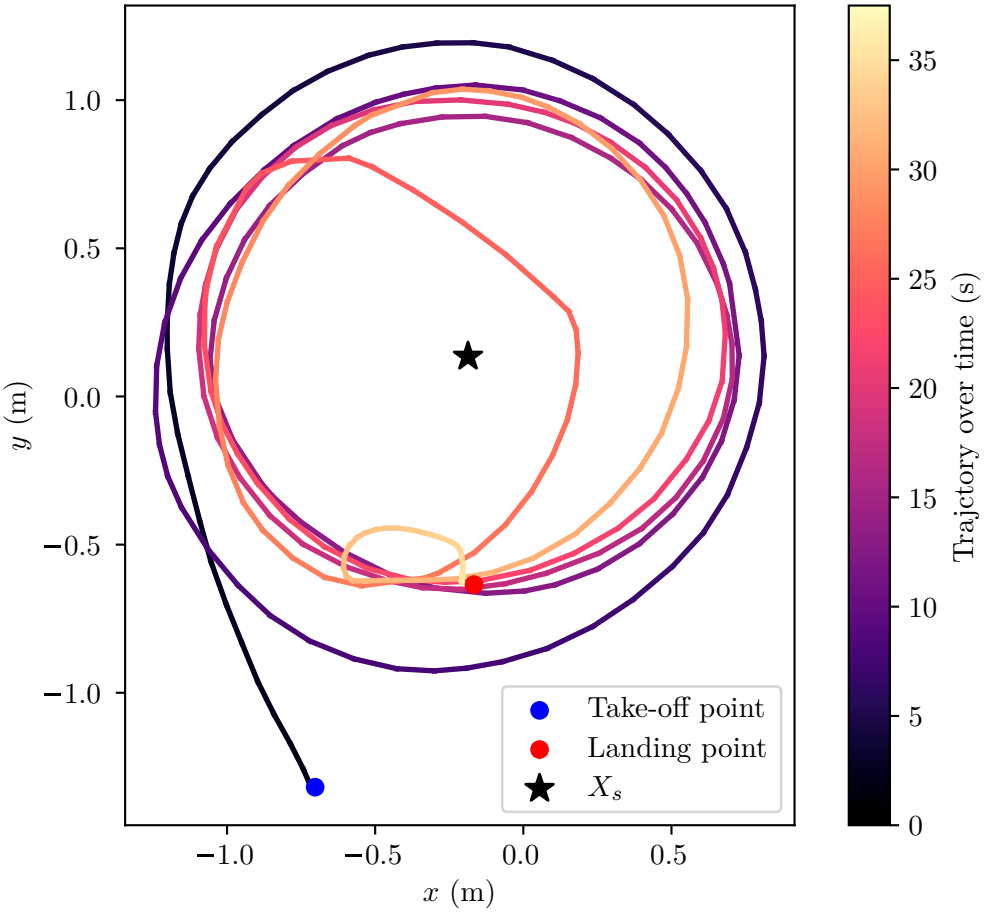


Figure 5.7. Trajectory of the UAV around the sound source during the experiment. The UAV uses the acoustically estimated DOA in its path-planning algorithm. The trajectory is a circle of radius 1 m centered on the source.

Similarly to the previous experiment, the results will be analyzed first from a control perspective, then from an acoustic perspective, although both are tightly linked.

Control evaluation

The trajectory is similar to the one obtained from the true DOA data. The UAV describes a circular trajectory around the source, while facing the source (the heading of the drone is pointed towards the source). Near the end, the UAV gets abnormally closer to the source due to an outlier in the DOA estimations, however, it is able to quickly recover and get back on the expected path. This demonstrates the robustness of the proposed solution.

The DOA estimation using acoustic data is more noisy than the one obtained from the optical system, which is expected to generate a noisier control. From the analysis of the trajectory, it did not seem to affect the UAV movements. However, it might still be observable on the motor controller outputs, which are shown in Fig. 5.8.

As expected, the motor controller output is much more noisy when relying on the acoustically estimated DOA. A highly unsteady rotor speed is known to generate stronger and more unpredictable noise than a steady rotor speed. Thus, the ego-noise is expected to be more challenging to mitigate using acoustic

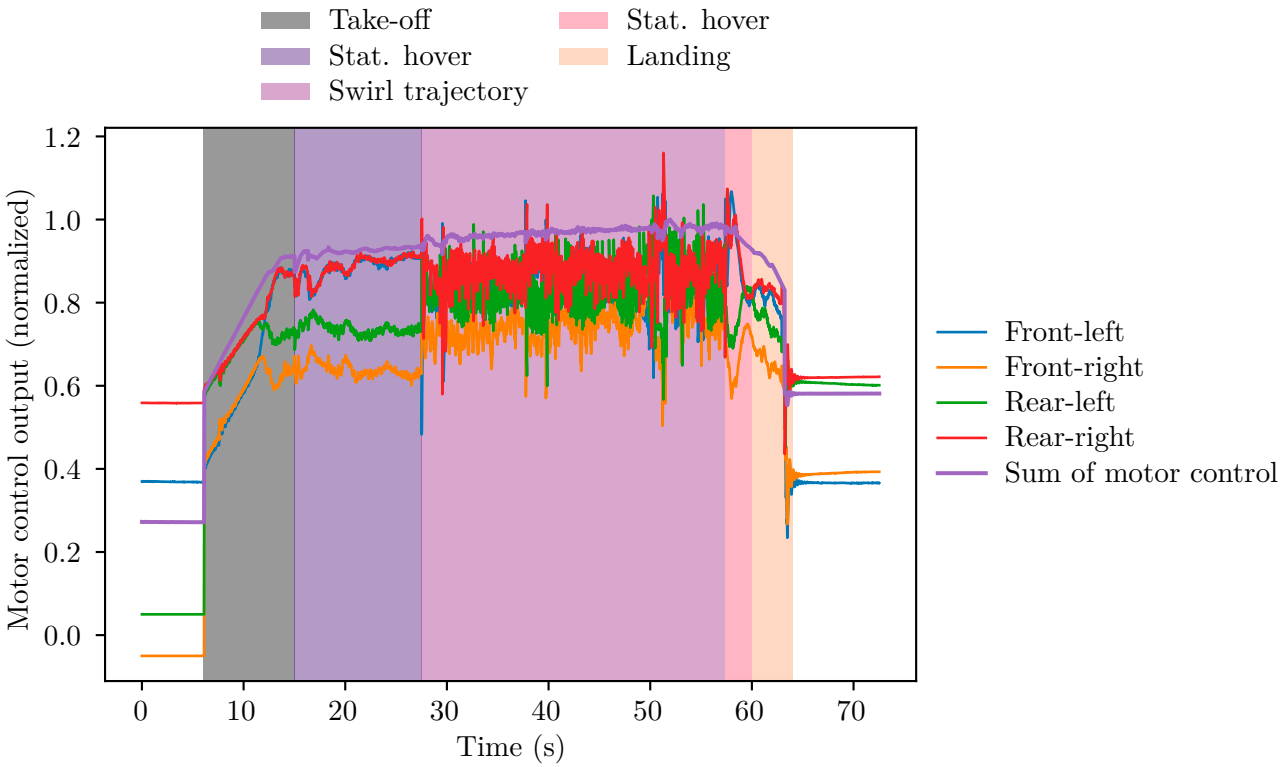


Figure 5.8. Motor controller output during the experiment using the acoustically estimated DOA data.

control than using camera control. However, Fig. 5.8 shows the motor controller output, which is not exactly the rotor speed. Indeed, rotor possesses an inductance, which acts as a low-pass filter on the motor controller output. From the visual analysis of these motor signals, their low frequency component seems to be highly unsteady, suggesting the low-pass filter effect from the inductance might not be sufficient to mitigate the unsteadiness generated by the more erratic acoustic control. This suggests that it would be beneficial to smooth the DOA signals before using them for control, thus making the control smoother as well.

Acoustic evaluation

After considering the control aspect, we now analyze the acoustic data recorded during this experiment. The acoustic processing parameters are the same as in the previous experiment, summarized in Table 5.1. The SIR is calculated from the spectra averaged over all the microphones at different times, shown in Fig. 5.9.

The first spectrum (blue) is recorded at the take-off point, before taking-off, when the UAV is the furthest from the source. The second spectrum (yellow) is recorded after landing, when the UAV is on the circle of convergence around the source. On this circle, the UAV is at a close distance to the source, which makes the SIR higher. The third spectrum (green) is recorded during the stationary hover period, where the UAV maintains its position above the take-off point. Finally, the last spectrum (red) is recorded during the swirl trajectory phase, when the UAV circles around the source. The SIR at the take-off position is around -5 dB. It varies over the course of the experiment, as the UAV moves around the source. The SIR at the take-off point is observed, since it is the most critical point of the experiment, as the UAV is the furthest from the source. However, due to the high non-stationarity of the ego-noise, the SIR does not provide a perfect indication of the instantaneous conditions for localization, but rather an idea of the difficulty of the localization task. With a SIR of -5 dB, it is safe to assume that the source signal is consistently dominated by the ego-noise, and that the localization task will be challenging.

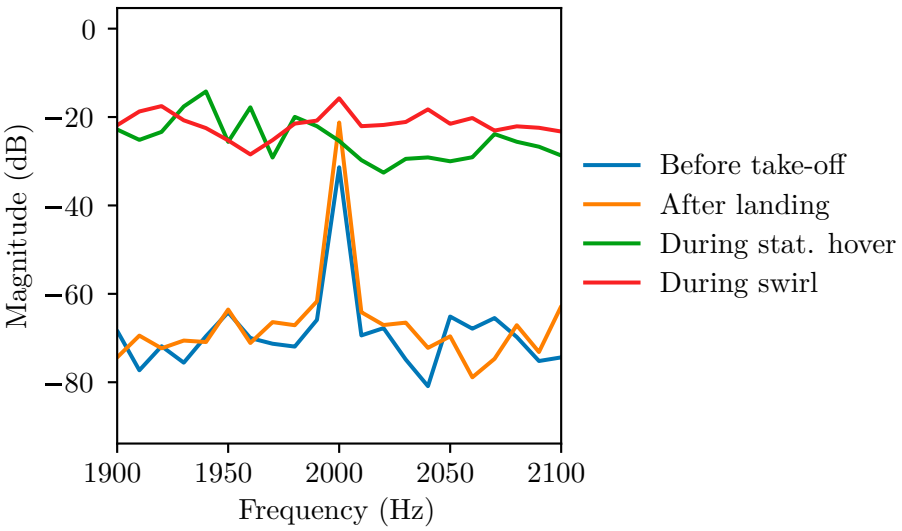


Figure 5.9. Spectra of the acoustic recordings averaged over each microphone, at different times. The frequency resolution matches the one used in the acoustic imaging algorithm, $\Delta f = 10$ Hz.

The DOA of the source is computed using the GEVD-CBF algorithm, and compared to the ground-truth DOA obtained from the optical system. The same offline recording of the ego-noise is used to denoise the data, as in the previous experiment. The DOA estimation error for the GEVD-CBF and raw CBF algorithms are shown in Fig. 5.10.

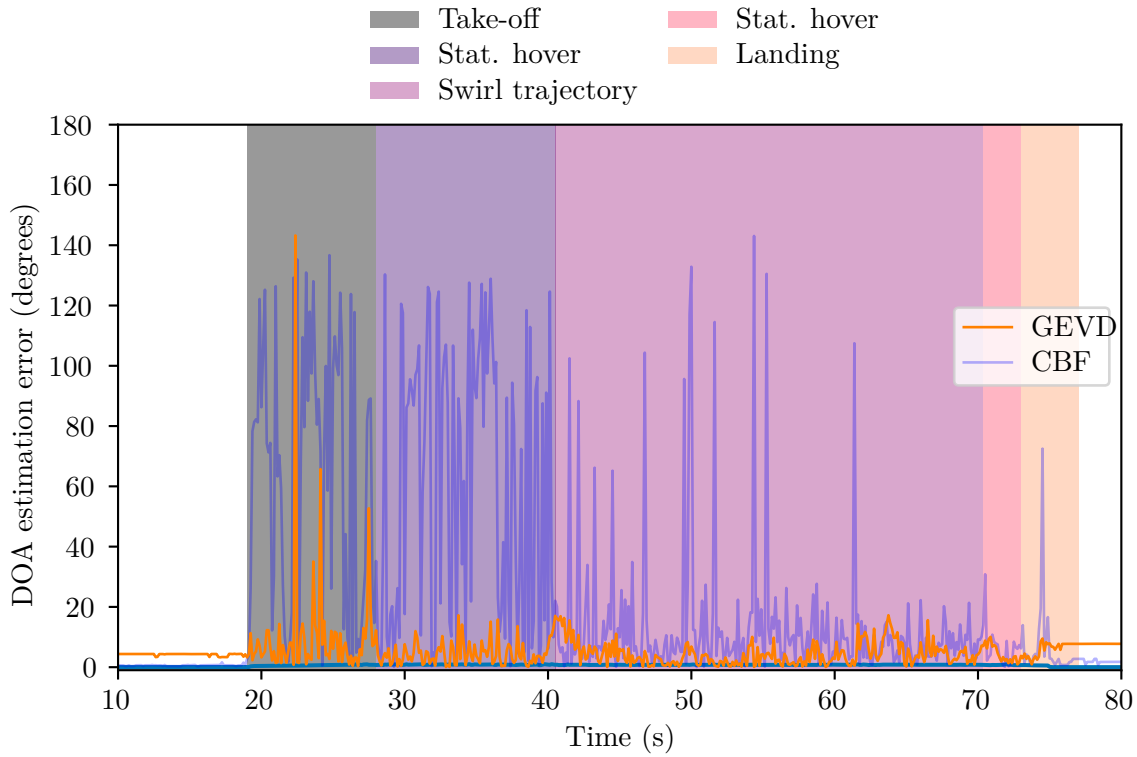


Figure 5.10. Angle error between the ground-truth DOA and the estimated DOA using the GEVD-CBF algorithm and the CBF algorithm. The background color denotes the flight phases. Standard deviation of the angle error during the swirl trajectory phase is 23 ° using the CBF algorithm, and 4 ° using the GEVD-CBF algorithm.

We observe that the angle error is considerably reduced when using the GEVD-CBF algorithm compared to the CBF algorithm, and remains below 10 ° during the swirl trajectory phase. The difference

between these results and the previous experiment is explained by the smaller radius of convergence, which leads to a better SIR at the end of the experiment. Obtaining consistent experimental conditions to assess the performance of the algorithms has proven to be challenging, and will be the focus of further developments. Nevertheless, the results obtained during this experiment are consistent with the ones obtained during the previous experiment, and demonstrate the relevance of the GEVD-CBF algorithm for real-time denoising.

SSL evaluation

Finally, the SSL capabilities of the platform are validated using the acoustic data recorded. The goal of SSL is to recover the position of the sound source from the DOA estimated by the acoustic imaging, and the position and orientation of the UAV measured by the cameras. To do this, the triangulation scheme introduced in Chapter 3, Section 3.3 is used. This consists in solving the triangulation geometrically within the ground plane, using two DOAs and two measurement positions. The two DOAs picked are separated from a time equal to the time the UAV takes to perform a quarter of a circle around the source. This time depends on the speed of the UAV and the radius of the circle of convergence, which were measured afterward. From this time, it is possible to estimate the number of measurements between two orthogonal DOAs, which allow us to pick a set of two orthogonal DOAs. On the other hand, this triangulation requires knowledge of the radius of the convergence circle, it would thus be interesting to investigate further its relationship with the mass and speed of the UAV, which are known during flight, following the first steps from Section 5.1.2.
The instantaneous localization error is defined as the Euclidean distance between the triangulated point at a given measurement point and the real position of the source. It is plotted over the course of the whole experiment in Fig. 5.11.

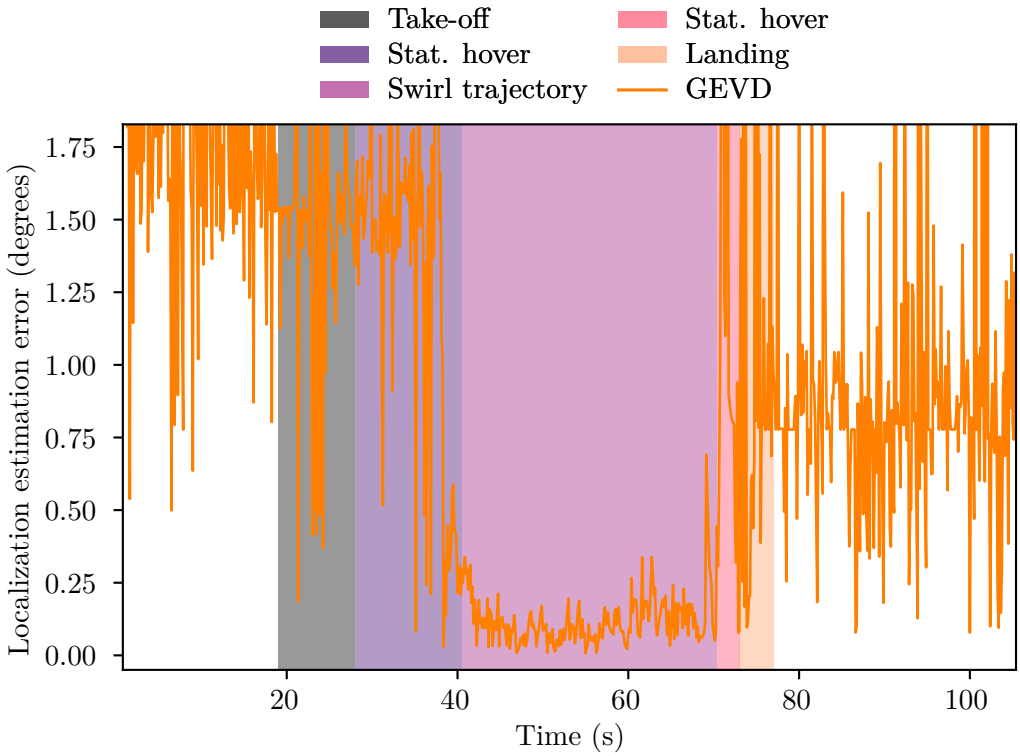


Figure 5.11. Localization error over the course of the experiment. The background color denotes the flight phases.

The instantaneous localization error is observed to be high during the stationary hover phase, which is expected since the UAV is not moving and thus does not provide enough angular diversity to triangulate the source. As soon as the UAV starts moving, the localization error drops below 0.2 m,

which is around the wavelength of the source signal. Comparing Fig. 5.10 and Fig. 5.11, we observe the importance of motion for the triangulation. Even if the DOA estimation is reliable during the stationary hover phases, the triangulation error is high due to the small angle differences between the DOAs. A video of the triangulation process is available at https://youtu.be/gYu6MRTG1zc. These results are satisfactory and suggest that the convergence criterion for the SSL, which is achieved when five successive triangulation estimations are contained within a wavelength-radius-circle around the last estimation, will be easily satisfied in this configuration. Indeed, applying the convergence criterion over this data allows to find the source after only 18 iterations, which corresponds to around 2.25 s of flight time. The final estimation of the source position is obtained by averaging the last five triangulation estimations, which, in this case, yields an error of 0.26 m, which is satisfactory given the geometry of the loudspeaker (about 10 cm wide), and close to the wavelength.

Using more instantaneous triangulation estimations allows to reduce the final localization error, but also increases the time needed to converge to the source position. For instance, using 10 estimations yields a final localization error of 0.03 m, but required 62 measurements, or 7.75 s of flight time to converge. A clear trade-off appears between the final localization error and the time needed to converge to the source position.

5.3 | Discussions

The convergence criterion currently used for the triangulation is based on an arbitrary-defined parameter: the number of triangulation estimations to consider. A perspective for future work would be to evaluate the DOA estimation error of the acoustic imaging algorithm coupled with the filtering algorithm. This would allow to define a convergence criterion based on the Cramer-Rao Bound (CRB) presented in Appendix B, incorporating unexploited knowledge of the minimal source level to be detected, and the radius of the convergence circle. However, the dependency of the filtering algorithm on the ego-noise template used for denoising makes it difficult to perform a reliable error propagation study on the denoising step.

Regarding the convergence radius, it is currently controlled indirectly by the UAV speed and weight. However, it could be interesting to design the speed of the UAV based on the desired convergence radius, possibly introducing artificial lag in the DOAs used for control to enlarge the circle.

The path-planning algorithm utilizes non-smoothed DOA estimations, which leads to a noisier control and, consequently, a stronger ego-noise. Smoothing the DOA estimations before using them for control improves the stability of the motor controller output, which is expected to reduce the ego-noise.

The current implementation of the triangulation algorithm is based on two DOA estimations, which is sufficient for the current experiments, but could be extended to more than two DOAs. The current algorithm has the advantage of simplicity, which allows to derive theoretical bounds, which could be useful in the aforementioned convergence criterion design. An approach using Kalman filtering is envisaged to exploit fully the available DOA information and the knowledge of the vehicle movement.

The experiment presented in this chapter tackled the single-source localization problem, but could be extended to multi-source localization, which was already performed successfully using the ground vehicle in Chapter 3.

Another key assumption throughout this work was that the sound sources are immobile, however, the speed gain of the UAV-based localization enables directly addressing the tracking of a moving source.

Regarding the viability of the algorithm outdoors, it is important to consider the impact of the replacement of the OptiTrack system by a GPS system for the position of the UAV. GPS sensors typically provide a less accurate position than the OptiTrack system, which could affect the triangulation performance. However, the system proved to be quite robust to the noise in the position and orientation

measurements, as shown in Chapter 2, which is promising for further work. On the other hand, Inertial Measurement Units (IMUs) could replace the OptiTrack for orientation estimation, as they are commonly used for that purpose and are highly precise.

5.4 | Conclusion

This chapter concludes the main work in this thesis. Its major contributions are the demonstration of real-time sound source localization using a UAV, and the validation of the path-planning algorithm using acoustically estimated DOAs. It also demonstrates the relevance of the denoising algorithm presented in Chapter 4 for the UAV ego-noise, and the possibility to control the convergence radius of the UAV around a sound source using user-defined parameters. The results obtained during the experiments are consistent with the ones obtained in Chapter 3, and demonstrated that the UAV embedded system represents a faster and more flexible platform for sound source localization than the ground vehicle. The next chapter will conclude this thesis by summarizing the contributions and results, and discussing the perspectives for future work.

Summary tables

This chapter introduced the real-time autonomous sound source localization from a UAV. The key achievements of this chapter are summarized in Table 5.2, and the main perspectives discussed are listed in Table 5.3.

Aspect	Achievement
Experimental setup	The communication architecture allows the UAV to be controlled by the path-planning algorithm using the DOA estimated from the acoustic data.
Path-planning algorithm	The path-planning algorithm is validated using true DOA data, demonstrating that the UAV is able to follow a circular trajectory around the source.
Acoustic evaluation	The GEVD-CBF algorithm is applied to the acoustic data, showing that the DOA estimation error is reduced compared to the CBF algorithm, and that the triangulation error is low during the swirl trajectory phase.
Real-time sound source localization	The UAV is able to triangulate the position of the sound source in real time, with a low localization error of 0.26 m after 18 iterations, which corresponds to around 2.25 s of flight time. This error can be reduced to 0.03 m by using more triangulation estimations, requiring 62 iterations, or 7.75 s of flight time.

Table 5.2. Summary of the main achievements of this chapter.

Aspect	Perspective
Convergence criterion	A convergence criterion based on the time evolution of the triangulation error could be defined, based on the Cramer-Rao Bound of the beamforming algorithm used for triangulation, the minimal sound level of the source to be detected, and the theoretical limits of the 2-DOA triangulation.
Convergence radius	The convergence radius could be controlled directly via the UAV speed, and by introducing artificial lag in the DOAs used for control to enlarge the circle if necessary.
Moving source tracking	The tracking of a moving source could be addressed, using the speed gain of the UAV-based localization.
Outdoor localization	The replacement of the OptiTrack system by a GPS system and inertial measurement unit (IMU) for the position and orientation of the UAV could be investigated, as well as the impact on the triangulation performance.

Table 5.3. Summary of the main perspectives regarding UAV-based autonomous SSL.

General conclusions

This work resulted in the development and validation of two autonomous platforms for sound source localization. The two platforms were able to locate stationary sound sources on the ground without prior knowledge of their position, in an indoor environment. Each platform has its own advantages, and the choice between them depends on the specific application scenario. The ground vehicle is suitable for cases where the sources are low-level and located close to the ground in a navigable environment. The drone-based solution, on the other hand, is more appropriate if the diagnosis needs to be as rapid as possible and the terrain is not accessible to a ground vehicle.
Significant opportunities for future improvement remain, and several research directions are proposed to increase its robustness, challenge its performance under more complex use-cases, or even broaden its use to other fields.

Contributions

The main contributions of this thesis are summarized in Table 5.4. The contributions are organized around five main aspects: acoustic-driven path-planning, numerical evaluation, ground autonomous vehicle for multiple sources localization, drone ego-noise analytical framework, and autonomous drone for sound source localization. Each aspect is briefly described below.

Aspect	**Contribution**
Acoustic-driven path-planning	A path-planning algorithm for autonomous Sound Source Localization (SSL) is proposed. The algorithm is based on the iterative building of a source map, made possible by the autonomous circling of sound sources, which convergence is demonstrated theoretically and observed experimentally.
Simulation setup	The proposed path-planning is validated through numerical simulations, demonstrating its ability to efficiently localize stationary sound sources in a controlled environment, even in the presence of numerous sources.
Autonomous ground platform	An experimental platform for autonomous SSL using a ground vehicle is developed. The platform includes a microphone array, an onboard computer, and a control system for autonomous navigation. Experimental results demonstrate the accuracy of the proposed method in localizing stationary sound sources in an indoor environment.
Drone ego-noise description	An extensive description of the challenges of Unmanned Aerial Vehicle (UAV) embedded microphone measurements is proposed. This description supports the choice of a method for mitigating the ego-noise, enabling the UAV to localize sound sources while flying.
Autonomous aerial platform	An experimental platform for autonomous SSL using a quadcopter UAV is developed. The platform carries a 32-MEMS-microphone-array and is able to locate a sound source on the ground with great precision under minimal time.

Table 5.4. Summary of the main achievements of this thesis.

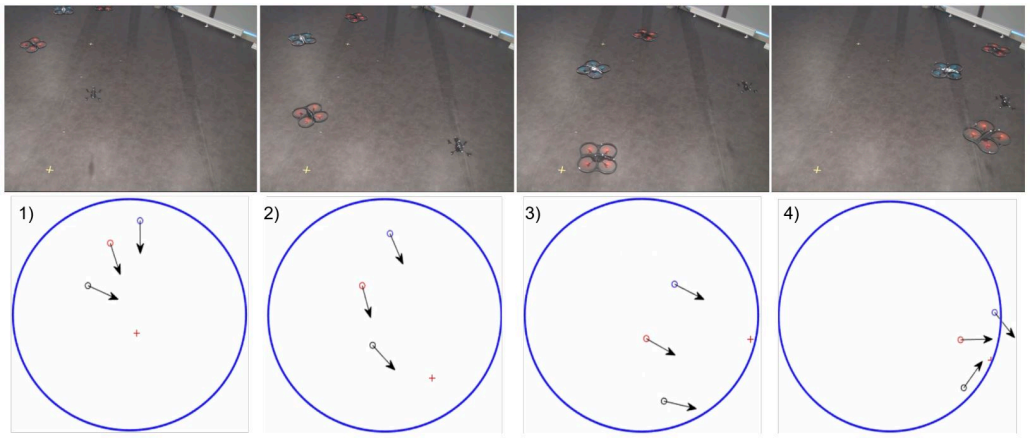


Figure 5.12. Illustration of the multiagent approach for moving source localization by De Souza *et al.* [94]. Two UAVs (blue) are used to 'hunt' a target UAV (red).

Perspectives

The perspectives for this work are numerous and can be categorized into three main areas: technical, methodological, and application-oriented.

Technical perspectives

The technical perspectives include the aforementioned adaptation of the convergence criterion based on theoretical limits of the triangulation and acoustic imaging methods. A robust convergence criterion would be key to ensure the reliability of the localization process in real-life scenarios. Additionally, minimal efforts to enable onboard computing for real-time processing would enable the system to operate autonomously at larger scales. With the addition of onboard sensors such as GPS, or cameras, a demonstration of the method on an outdoor large-scale test area could be possible without further adaptation of the method. Finally, further work on more efficient denoising techniques, such as adaptive denoising methods, could significantly improve the performance of the system in noisy environments. These efforts would align with the current trends in drone audition, and contribute significantly to the field.

Methodological perspectives

The proposed method tackled a relatively simple scenario in a controlled environment. While some limitations of the method could be overcome by technical developments, some require adaptation of the approach itself. For instance, tackling moving sources is a real challenge that is quite different from the immobile scenario considered in this thesis. Quickly moving sources would require a more complex path-planning algorithm that takes into account the motion of the source, as well as the potential changes in the environment. One of the method stemming in the literature to approach this problem is to introduce multiple agents to perform the triangulation synchronously, as proposed by Yen *et al.* [21] through a numerical study. Additionally, recent advances in the field of robust sensor motion control, such as the work of Hu *et al.* [93], where a control law for a set of moving sensors tasked with tracking a target is described. The control law is designed so that the variance of the localization estimator reaches the Cramer-Rao bound. The current approach could serve as a basis for such a multiagent system, where multiple drones would collaborate to localize a moving source. This aligns with the recent advances in the field of autonomous drone coordination such as the work of De Souza *et al.* [94] on the task of autonomous UAV pursuing a target UAV, as illustrated in Fig. 5.12.

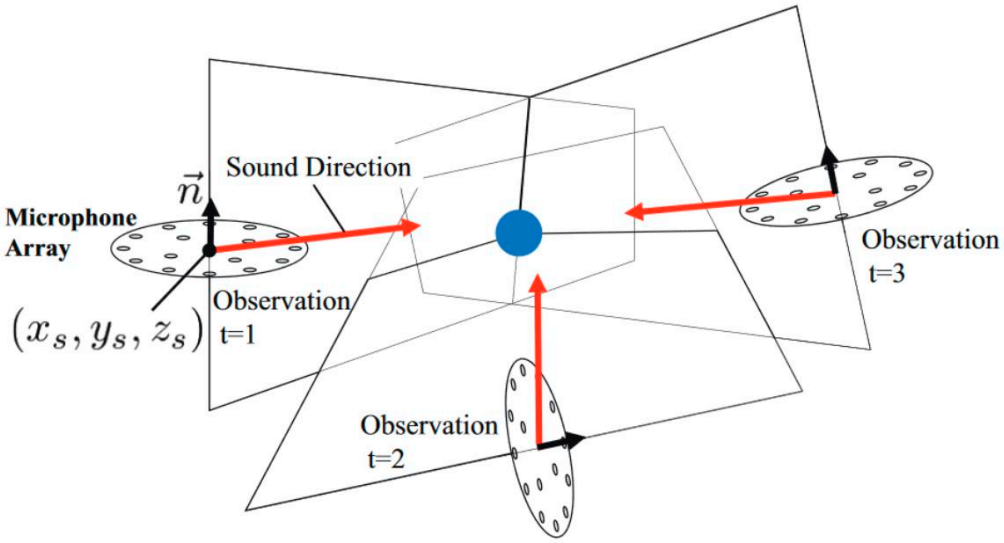


Figure 5.13. Illustration of the 3D triangulation method by Sasaki *et al.* [95].

This multiagent approach could also be extended to intermittent sources, where the source is not stationary, although recent work by Nguyen *et al.* [20] have shown that a Kalman filter for the triangulation, coupled with a voice activity detection algorithm, could be used to estimate the position of an intermittent source.

Another perspective to investigate for practical implementation in complex environments is the 3D triangulation methods. Indeed, in realistic environments, the sound sources are not necessarily on the ground, and the UAV might need to navigate through obstacles to reach the source. Its path would thus be constrained. Taking into account these constraints into the path-planning algorithm would allow reaching another milestone towards real-life applications. The work of Sasaki *et al.* [95] on probabilistic 3D-triangulation of sound sources in complex environments, illustrated in Fig. 5.13, showed promising results and is perfectly fit for an implementation on an UAV.

Finally, the use of AI for source identification could be a valuable addition to the proposed method, especially in the scope of industrial acoustic diagnosis. The work of Tiffeneau *et al.* [96] on AI-based source classification using microphone arrays could be adapted within the proposed framework to enable the UAV to not only localize the source, but also identify its nature. Integrating images from onboard cameras as an additional source of information seems to be a promising direction, as demonstrated by the work of He *et al.* in [97], where SSL was enhanced by the addition of visual cues to detect otherwise invisible sources. The convergence criterion could be modified to include the identification of the source, making the system rotating around the source until it is identified, rather than just localized.

Application-oriented perspectives

Although this approach was initially designed to localize sound sources in an industrial context, it could be adapted to other scenarios. For instance, bioacoustic monitoring is a field that could benefit from this approach. The work of Wang *et al.* [83] on bird call detection from a flying UAV platform using a directive microphone showed promising results and could be extended to the proposed method, enabling autonomous movement of the platform to explore its surroundings. Similarly, the work of autonomous underwater robots for wildlife monitoring by Girdhar *et al.* [98], illustrated in Fig. 5.14 could be adapted to the proposed framework, enabling the underwater robot to autonomously localize, track and monitor marine wildlife.



Figure 5.14. Illustration of the autonomous underwater robot for wildlife monitoring by Girdhar *et al.* [98].



Figure 5.15. Illustration of the building envelope acoustic inspection by Kolsch *et al.* [99].

Additionally, the proposed method could be adapted with minimal changes to search-and-rescue scenarios, where the UAV could autonomously localize persons in danger, for instance buried under rumbles or snow after a natural disaster occurred. Indeed, by setting a threshold on the minimal sound level to detect, it would be possible to determine the altitude and coarseness of a research grid that the UAV would scan, and go off-course using the proposed autonomous path-planning algorithm to localize the person.

Finally, the approach could be used not only to detect gas leaks or partial discharge phenomena in industrial environments, but also to detect air leaks in building envelopes, as proposed by Kölsch *et al.* [99], illustrated in Fig. 5.15. The method could be adapted to autonomously navigate around the building and localize the thermal leaks, enabling a more efficient inspection process, which could support a large scale deployment of the method in the field of building energy efficiency.

In conclusion, the work presented in this thesis opens up new avenues for research and development in the field of autonomous drone audition, and paves the way for future applications in various domains.

A. Convergence of the path-planning algorithm

As stated in Chapter 2, after enough iterations, the proposed path-planning algorithm ensures that the microphone array places itself on a circle $\mathcal{C}$ of radius $\frac{d}{\sqrt{2}}$ around the targeted sound source. A proof of this property is given here.

The movement of the robot 45 ° to the left of the current DOA can be broken down into two movements: one movement of length $\frac{d}{\sqrt{2}}$ directly toward the source, and one movement of length $\frac{d}{\sqrt{2}}$ to the left. Let us introduce $P_{i+\frac{1}{2}}$, the point obtained after the first movement in the direction of the source. Note that this movement never occurs in reality, and is just a useful geometrical construction in the following. Let the sequence U_i be equal to the signed distance between the robot position at iteration i and the circle $\mathcal{C}$ centered on the source X_s, and of radius $\frac{d}{\sqrt{2}}$.

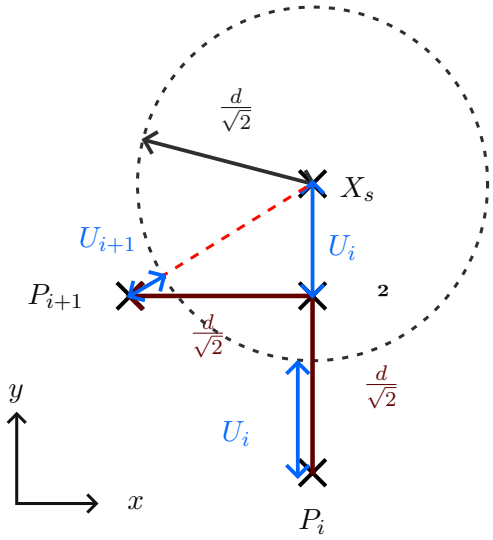


Figure A.1. The robot converging in position to the targeted sound source.

Figure A.1 presents the right triangle $(P_{i+\frac{1}{2}}, S, P_{l+1})$. The Pythagorean theorem in this triangle is

written

$$
\begin{aligned}
(\frac{d}{\sqrt{2}} + U_{i+1})^2 = U_i^2 + \frac{d^2}{2} \\
\Rightarrow |\frac{d}{\sqrt{2}} + U_{i+1}| = \sqrt{U_i^2 + \frac{d^2}{2}} \\
\Rightarrow |U_{i+1}| \leq \sqrt{U_i^2 + \frac{d^2}{2}} - \frac{d}{\sqrt{2}} \\
\Rightarrow |U_{i+1}| \leq \frac{d}{\sqrt{2}}\sqrt{1 + \frac{U_i^2}{\frac{d^2}{2}}} - \frac{d}{\sqrt{2}} \\
\Rightarrow |U_{i+1}| < \frac{d}{\sqrt{2}}(\sqrt{1} + \sqrt{\frac{U_i^2}{\frac{d^2}{2}}}) - \frac{d}{\sqrt{2}} = |U_i| \qquad \text{(A.1)}
\end{aligned}
$$

Thus, the sequence $|U_i|$ is strictly decreasing and bounded below by 0, hence it converges to some limit $L \geq 0$. Thus as $i \to \infty$, $\Rightarrow |U_{i+1}| \leq \sqrt{U_i^2 + \frac{d^2}{2}} - \frac{d}{\sqrt{2}}$ becomes

$$L \leq \sqrt{L^2 + \frac{d^2}{2}} - \frac{d}{\sqrt{2}}$$

Assuming $L \neq 0$, we have

$$L \leq \sqrt{L^2 + \frac{d^2}{2}} - \frac{d}{\sqrt{2}} < \sqrt{L^2} + \sqrt{\frac{d^2}{2}} - \frac{d}{\sqrt{2}} = L$$

leading to $L < L$, which is absurd. Thus $L = 0$ and the sequence U_i admits 0 as a limit. This means that as $i \to \infty$, the positions of the robot converge on the circle $\mathcal{C}$.

B. Cramér–Rao Bound for 2-DOAs Triangulation

The path-planning algorithm introduced in Chapter 2 is based on the principle that the localization error in triangulation can be minimized when the two Directions of Arrival (DOA) measurements are approximately orthogonal. This annex provides a theoretical justification for this principle using tools from estimation theory, specifically the Cramér–Rao Bound (CRB) [100].

The end-goal of sound source localization is to determine the 2D position of a source, denoted $X_s = (x_s, y_s)$, from noisy DOA measurements. The localization performance is often quantified by the Root Mean Square Error (RMSE), which reflects the average distance between the estimated source position, $(\hat{x}_s, \hat{y}_s)$, and its true position, (x_s, y_s):

$$\text{RMSE} = \sqrt{\mathbb{E}[(\hat{x}_s - x_s)^2 + (\hat{y}_s - y_s)^2]}.$$

This can be interpreted as a measure of accuracy, where precise localization translates to a low RMSE.

The CRB provides a theoretical lower bound on this RMSE for any unbiased estimator by setting a minimum possible variance for the estimations of x_s and y_s based on the geometry of the measurements and the noise characteristics. Here, we analyze how this bound depends on the angular separation $\Delta\theta$ between the two DOAs and show that the bound is minimized (i.e., accuracy is maximized) when $\Delta\theta = 90°$.

We consider the localization of a 2D source using two DOAs measurements from sensors located at known positions. Let the sensors be located symmetrically at positions $\mathbf{s}_1$ and $\mathbf{s}_2$, both at distance R from the source, the true DOA angles to the source be θ_1 and θ_2, with angular separation $\Delta\theta = \theta_2 - \theta_1$, and the measurements be corrupted by independent Gaussian noise with variance σ^2_{DOA}. Let $\theta_0 = (\theta_1 + \theta_2)/2$ be the bisector angle.

Each DOA measurement θ_i satisfies:

$$\frac{\partial \theta_i}{\partial X_s} = \frac{1}{R} \begin{bmatrix} -\sin\theta_i \\ \cos\theta_i \end{bmatrix}$$

The CRB can be obtained from the Fisher Information Matrix, which, for this problem, is given by:

$$F = \sum_{i=1}^{2} \frac{1}{\sigma^2_{\text{DOA}}} \left(\frac{\partial \theta_i}{\partial X_s}\right) \left(\frac{\partial \theta_i}{\partial X_s}\right)^T$$

Substituting the expressions, it follows:

$$F = \frac{1}{\sigma^2_{\text{DOA}} R^2} \left[\begin{bmatrix} \sin^2\theta_1 & -\sin\theta_1\cos\theta_1 \\ -\sin\theta_1\cos\theta_1 & \cos^2\theta_1 \end{bmatrix} + \begin{bmatrix} \sin^2\theta_2 & -\sin\theta_2\cos\theta_2 \\ -\sin\theta_2\cos\theta_2 & \cos^2\theta_2 \end{bmatrix} \right]$$

Using trigonometric identities and letting $\theta_1 = \theta_0 - \Delta\theta/2$, $\theta_2 = \theta_0 + \Delta\theta/2$, the FIM simplifies to:

$$F = \frac{1}{\sigma_{\text{DOA}}^2 R^2} \begin{bmatrix} 1 - \cos(\Delta\theta)\cos(2\theta_0) & -\cos(\Delta\theta)\sin(2\theta_0) \\ -\cos(\Delta\theta)\sin(2\theta_0) & 1 + \cos(\Delta\theta)\cos(2\theta_0) \end{bmatrix}$$

For the case where the bisector is aligned with the x-axis ($\theta_0 = 0$), this further simplifies to:

$$F = \frac{1}{\sigma_{\text{DOA}}^2 R^2} \begin{bmatrix} 2\sin^2(\Delta\theta/2) & 0 \\ 0 & 2\cos^2(\Delta\theta/2) \end{bmatrix}$$

The CRB is the inverse of the FIM, which is given by:

$$\text{CRB} = F^{-1} = \frac{\sigma_{\text{DOA}}^2 R^2}{2} \begin{bmatrix} \frac{1}{\sin^2(\Delta\theta/2)} & 0 \\ 0 & \frac{1}{\cos^2(\Delta\theta/2)} \end{bmatrix}$$

The CRB gives the lower bound for estimation variances and covariances of the source position, $(\hat{x}_s, \hat{y}_s)$, as:

$$\text{Cov}(\hat{x}_s, \hat{y}_s) \succeq \text{CRB}$$

with $\text{Cov}(\hat{x}_s, \hat{y}_s)$, the covariance matrix of the triangulation estimator, and where $A \succeq B$ for A and B two given matrices, denotes that A-B is positive semi-definite. More specifically, it follows that:

$$\text{Tr}(\text{Cov}(\hat{x}_s, \hat{y}_s)) \geq \text{Tr}(\text{CRB}) = \frac{\sigma_{\text{DOA}}^2 R^2}{\sin^2(\Delta\theta)}$$

The localization error (RMSE) is therefore:

$$\text{RMSE} = \sqrt{\sigma_x^2 + \sigma_y^2} = \sqrt{\text{Tr}(\text{Cov}(\hat{x}_s, \hat{y}_s))} \geq \frac{\sigma_{\text{DOA}} R}{|\sin \Delta\theta|}$$

with σ_x^2 and σ_y^2, the variance on the triangulated position projected on x and y axes, respectively. It follows that the lower bound for the RMSE is minimal when $\Delta\theta = 90°$, where the bound becomes:

$$\text{RMSE} \geq \sigma_{\text{DOA}} R$$

This configuration minimizes estimation uncertainty and gives a circular uncertainty region. Similarly, we obtain the intuitive result that the lower bound of RMSE diverges towards infinity when DOAs are colinear, with $\Delta\theta = 0°$ or $\Delta\theta = 180°$.

Note that these results are valid for any unbiased 2-DOAs triangulation estimators, such as the one used in Chapter 2.

These results are illustrated in Fig. B.1, obtained by simulating the triangulation process for three different sensors with angular spacing $\Delta\theta$, and a Gaussian error of zero-mean and standard deviation $\sigma_{\text{DOA}} = 2.5°$ on the DOA estimation. The sensors are placed at an equal distance $R = 1$ m from the source. It can be seen that the error ellipses are well aligned with the bisector, as depicted in our analytical model. Moreover, their area increases as $\Delta\theta$ differs from 90°.

Another interesting point to investigate is how the triangulation formula used in Chapter 2 compares to the CRB for different angular separations $\Delta\theta$ and noise variance σ_{DOA}^2. It can be seen from Fig. B.2 that the empirical RMSE values follow the CRB for low σ_{DOA} values; however, they quickly diverge as σ_{DOA} increases. As expected, the configuration where DOAs are spaced by 90° yields the lowest RMSE, and for this configuration, the CRB is achieved up to $\sigma_{\text{DOA}} = 20°$. The RMSE values lower than the CRB demonstrate that the estimator is biased for high DOA-noise values.

These results support that 90°-spaced DOAs are statistically optimal for 2-DOA triangulation under

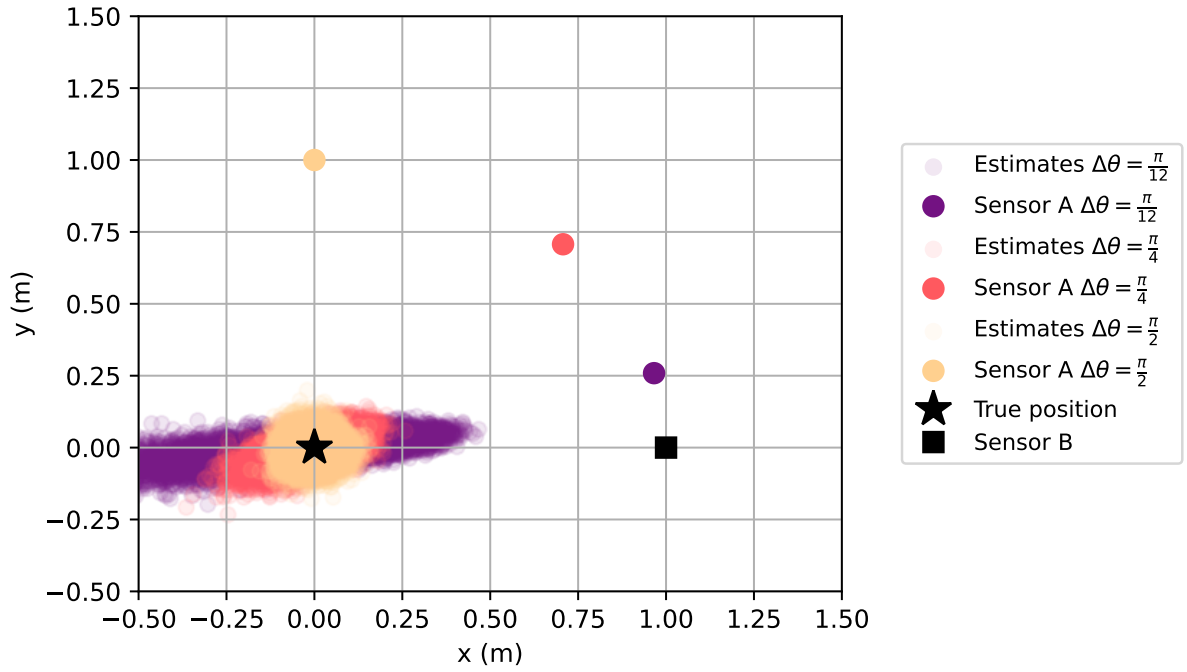


Figure B.1. Example of three sensor positions and the localization error they yield for a given DOA estimation uncertainty.

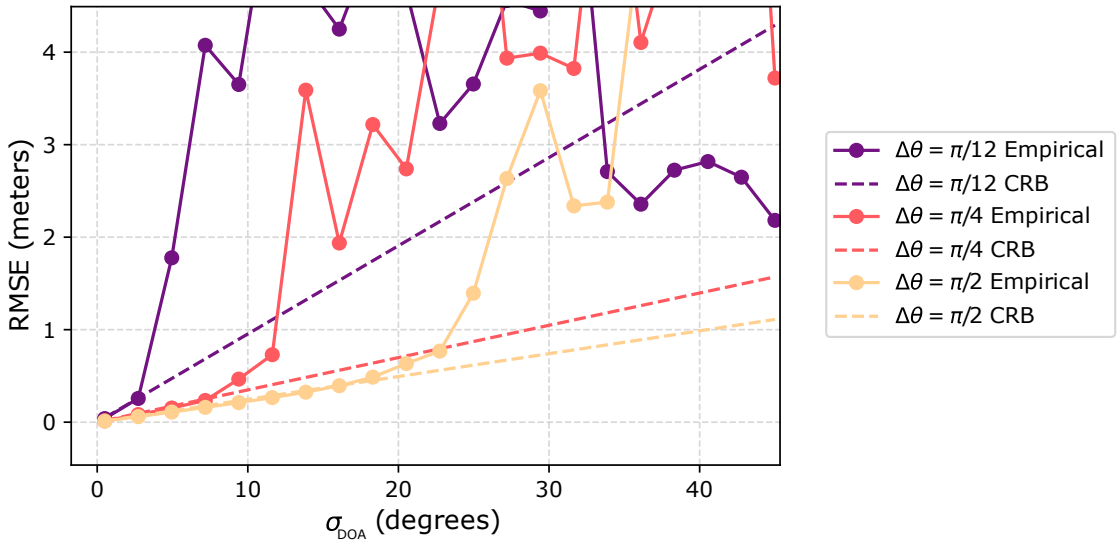


Figure B.2. Comparison between the empirical RMSE and its CRB for different angular separations $\Delta\theta$ and noise variance σ^2_{DOA}.

the Gaussian DOA-noise assumption. Additionally, it was demonstrated that for such a configuration, the RMSE can be expected to be close to $\sigma_{\text{DOA}}R$, with σ_{DOA}, the DOA-noise standard deviation, and R, the sensor-to-source distance.

Bibliography

[1] W. M. Hartmann, "How we localize sound," *Physics Today*, vol. 52, no. 11, p. 24–29, Nov. 1999. http://doi.org/10.1063/1.882727. [Online]. Available: http://dx.doi.org/10.1063/1.882727 (Cited on page 1.)

[2] A. Berta, J. L. Sumich, and K. M. Kovacs, "Chapter 11 - sound production for communication, echolocation, and prey capture," in *Marine Mammals (Third Edition)*, third edition ed., A. Berta, J. L. Sumich, and K. M. Kovacs, Eds. San Diego: Academic Press, 2015, pp. 345–395. http://doi.org/https://doi.org/10.1016/B978-0-12-397002-2.00011-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123970022000119 (Cited on page 1.)

[3] M. Konishi, "C<scp>oding of</scp> a<scp>uditory</scp> s<scp>pace</scp>," *Annual Review of Neuroscience*, vol. 26, no. 1, p. 31–55, Mar. 2003. http://doi.org/10.1146/annurev.neuro.26.041002.131123. [Online]. Available: http://dx.doi.org/10.1146/annurev.neuro.26.041002.131123 (Cited on page 1.)

[4] R. Leiba, "Conception d'un outil de diagnostic de la gene sonore en milieu urbain," Theses, Universite Pierre et Marie Curie - Paris VI, Dec 2017. [Online]. Available: https://hal.sorbonne-universite.fr/tel-01688847 (Cited on page 1.)

[5] W. Fiebig, *Archives of Acoustics*, jul 2023. http://doi.org/10.24425/aoa.2020.132487. [Online]. Available: http://dx.doi.org/10.24425/aoa.2020.132487 (Cited on page 2.)

[6] De Coensel, Bert and Botteldooren, Dick and Van Renterghem, Timothy and Dekoninck, Luc and Spruytte, Vincent and Makovec, Alphonso and Wessels, Peter and van der Eerden, Frits and Basten, Tom, "Robust microphone array beamforming for long-term monitoring of industrial areas," in *Proceedings of the 22nd International Congress on Acoustics*, 2016, p. 10. (Cited on page 2.)

[7] E. Cardenas Cabada, Q. Leclere, J. Antoni, and N. Hamzaoui, "Fault detection in rotating machines with beamforming: Spatial visualization of diagnosis features," *Mechanical Systems and Signal Processing*, vol. 97, p. 33–43, Dec. 2017. http://doi.org/10.1016/j.ymssp.2017.04.018. [Online]. Available: http://dx.doi.org/10.1016/j.ymssp.2017.04.018 (Cited on page 2.)

[8] L. del Val, A. Izquierdo, J. J. Villacorta, M. Herráez, and L. Suárez, "Fault detection methodology for a fan matrix based on svm classification of acoustic images," p. 221–228, 2019. http://doi.org/10.1007/978-3-030-11220-2_23. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-11220-2_23 (Cited on page 2.)

[9] M. Liang, K. Yang, M. Feng, K. Mu, M. Jiao, and L. Li, "Acoustic imaging method for gas leak detection and localization using virtual ultrasonic sensor array," *Sensors*, vol. 24, no. 5, p. 1366, Feb. 2024. http://doi.org/10.3390/s24051366. [Online]. Available: http://dx.doi.org/10.3390/s24051366 (Cited on page 2.)

[10] H. Shakhatreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, p. 48572–48634, 2019. http://doi.org/10.1109/access.2019.2909530. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2909530 (Cited on page 2.)

[11] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," pp. 832–839, 2000. [Online]. Available: http://www.aaai.org/Library/AAAI/2000/aaai00-128.php (Cited on page 3.)

[12] B. J. Zhang and N. T. Fitter, "Nonverbal sound in human-robot interaction: A systematic review," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 4, p. 1–46, Dec. 2023. http://doi.org/10.1145/3583743. [Online]. Available: http://dx.doi.org/10.1145/3583743 (Cited on page 3.)

[13] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, Oct. 2017. http://doi.org/10.1016/j.robot.2017.07.011. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0921889016304742 (Cited on pages 3 and 16.)

[14] Y. Sasaki, S. Kagami, and H. Mizoguchi, "Multiple sound source mapping for a mobile robot by self-motion triangulation," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2006, p. 380–385. http://doi.org/10.1109/iros.2006.281797. [Online]. Available: http://dx.doi.org/10.1109/IROS.2006.281797 (Cited on page 3.)

[15] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot," *Autonomous Robots*, vol. 27, no. 3, p. 221–237, Jun. 2009. http://doi.org/10.1007/s10514-009-9123-1. [Online]. Available: http://dx.doi.org/10.1007/s10514-009-9123-1 (Cited on page 3.)

[16] C. Evers, Y. Dorfan, S. Gannot, and P. A. Naylor, "Source tracking using moving microphone arrays for robot audition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2017, p. 6145–6149. http://doi.org/10.1109/icassp.2017.7953337. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2017.7953337 (Cited on page 3.)

[17] C. Evers and P. A. Naylor, "Acoustic slam," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, p. 1484–1498, Sep. 2018. http://doi.org/10.1109/taslp.2018.2828321. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2018.2828321 (Cited on page 3.)

[18] A. Portello, P. Danes, and S. Argentieri, "Active binaural localization of intermittent moving sources in the presence of false measurements," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2012, p. 3294–3299. http://doi.org/10.1109/iros.2012.6385967. [Online]. Available: http://dx.doi.org/10.1109/IROS.2012.6385967 (Cited on page 3.)

[19] A. Magassouba, N. Bertin, and F. Chaumette, "Aural servo: Sensor-based control from robot audition," *IEEE Transactions on Robotics*, vol. 34, no. 3, p. 572–585, Jun. 2018. http://doi.org/10.1109/tro.2018.2805310. [Online]. Available: http://dx.doi.org/10.1109/TRO.2018.2805310 (Cited on page 3.)

[20] Q. V. Nguyen, F. Colas, E. Vincent, and F. Charpillet, "Motion planning for robot audition," *Autonomous Robots*, vol. 43, no. 8, p. 2293–2317, Jul. 2019. http://doi.org/10.1007/s10514-019-09880-1. [Online]. Available: http://dx.doi.org/10.1007/s10514-019-09880-1 (Cited on pages 3, 32, 54, and 119.)

[21] B. Yen, T. Yamada, K. Itoyama, and K. Nakadai, "A performance assessment on rotor noise-informed active multidrone sound source tracking methods," *Drones*, vol. 8, no. 6, p. 266, Jun. 2024. http://doi.org/10.3390/drones8060266. [Online]. Available: http://dx.doi.org/10.3390/drones8060266 (Cited on pages 4, 31, 54, and 118.)

[22] S. Argentieri, A. Portello, M. Bernard, P. Danès, and B. Gas, *Binaural Systems in Robotics*. Springer Berlin Heidelberg, 2013, p. 225–253. http://doi.org/10.1007/978-3-642-37762-4_9. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37762-4_9 (Cited on page 4.)

[23] K. Nakadai, M. Kumon, Y. Sasaki, K. Hoshiba, and B. Yen, "Swarm active audition system with robots and drones for a search and rescue task," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Dec. 2024, p. 1–6. http://doi.org/10.1109/apsipaasc63619.2025.10848937. [Online]. Available: http://dx.doi.org/10.1109/APSIPAASC63619.2025.10848937 (Cited on page 4.)

[24] K. Nakadai, M. Kumon, H. G. Okuno, K. Hoshiba, M. Wakabayashi, K. Washizaki, T. Ishiki, D. Gabriel, Y. Bando, T. Morito, R. Kojima, and O. Sugiyama, "Development of microphone-array-embedded uav for search and rescue task," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Sep. 2017, p. 5985–5990. http://doi.org/10.1109/iros.2017.8206494. [Online]. Available: http://dx.doi.org/10.1109/IROS.2017.8206494 (Cited on pages 4, 5, and 88.)

[25] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Multiple sound source position estimation by drone audition based on data association between sound source localization and identification," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, p. 782–789, Apr. 2020. http://doi.org/10.1109/lra.2020.2965417. [Online]. Available: http://dx.doi.org/10.1109/LRA.2020.2965417 (Cited on page 4.)

[26] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2009, p. 664–669. http://doi.org/10.1109/iros.2009.5354419. [Online]. Available: http://dx.doi.org/10.1109/IROS.2009.5354419 (Cited on pages 4 and 90.)

[27] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Sep. 2014, p. 1902–1907. http://doi.org/10.1109/iros.2014.6942813. [Online]. Available: http://dx.doi.org/10.1109/IROS.2014.6942813 (Cited on page 4.)

[28] M. Strauss, P. Mordel, V. Miguet, and A. Deleforge, "Dregon: Dataset and methods for uav-embedded sound source localization," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2018, p. 1–8. http://doi.org/10.1109/iros.2018.8593581. [Online]. Available: http://dx.doi.org/10.1109/IROS.2018.8593581 (Cited on pages 4 and 5.)

[29] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP). IEEE, Mar. 2017, p. 496–500. http://doi.org/10.1109/icassp.2017.7952205. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2017.7952205 (Cited on page 4.)

[30] W. Manamperi, T. Abhayapala, P. Samarasinghe, and J. A. Zhang, "Drone audition: Audio signal enhancement from drone embedded microphones," Feb. 2023. http://doi.org/10.36227/techrxiv.20576823.v2. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.20576823.v2 (Cited on pages 4 and 90.)

[31] D. Salvati, C. Drioli, G. Ferrin, and G. L. Foresti, "Acoustic source localization from multirotor uavs," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, p. 8618–8628, Oct. 2020. http://doi.org/10.1109/tie.2019.2949529. [Online]. Available: http://dx.doi.org/10.1109/TIE.2019.2949529 (Cited on pages 4, 5, and 87.)

[32] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2012. http://doi.org/10.1109/iros.2012.6385608. [Online]. Available: http://dx.doi.org/10.1109/IROS.2012.6385608 (Cited on page 5.)

[33] S. F. Chevtchenko, B. J. Rodríguez, R. Vale, A. Soti, Y. Bethi, N. Ibnul, A. Marcireau, M. R. Azghadi, A. Wabnitz, and S. Afshar, "Drone-based sound source localization: A systematic literature review," *IEEE Access*, vol. 13, p. 94256–94274, 2025. http://doi.org/10.1109/access.2025.3572478. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2025.3572478 (Cited on page 5.)

[34] M. Bruneau, *Fundamentals of Acoustics*. Wiley, Jan. 2006. http://doi.org/10.1002/9780470612439. [Online]. Available: http://dx.doi.org/10.1002/9780470612439 (Cited on pages 10 and 11.)

[35] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, p. 70–73, Jun. 1967. http://doi.org/10.1109/tau.1967.1161901. [Online]. Available: http://dx.doi.org/10.1109/TAU.1967.1161901 (Cited on pages 13 and 91.)

[36] P. Stoica and R. Moses, "Spectral analysis of signals," *Prentice Hall*, 01 2005. (Cited on page 14.)

[37] W. Wang and E. W. Gill, "Evaluation of beamforming and direction finding for a phased array hf ocean current radar," *Journal of Atmospheric and Oceanic Technology*, vol. 33, no. 12, p. 2599–2613, Dec. 2016. http://doi.org/10.1175/jtech-d-15-0181.1. [Online]. Available: http://dx.doi.org/10.1175/JTECH-D-15-0181.1 (Cited on page 14.)

[38] R. Merino-Martínez, P. Sijtsma, M. Snellen, T. Ahlefeldt, J. Antoni, C. J. Bahr, D. Blacodon, D. Ernst, A. Finez, S. Funke, T. F. Geyer, S. Haxter, G. Herold, X. Huang, W. M. Humphreys, Q. Leclère, A. Malgoezar, U. Michel, T. Padois, A. Pereira, C. Picard, E. Sarradj, H. Siller, D. G. Simons, and C. Spehr, "A review of acoustic imaging methods using phased microphone arrays: Part of the "aircraft noise generation and assessment" special issue," *CEAS Aeronautical Journal*, vol. 10, no. 1, p. 197–230, Mar. 2019. http://doi.org/10.1007/s13272-019-00383-4. [Online]. Available: http://dx.doi.org/10.1007/s13272-019-00383-4 (Cited on page 16.)

[39] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech &; Language*, vol. 34, no. 1, p. 87–112, Nov. 2015. http://doi.org/10.1016/j.csl.2015.03.003. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2015.03.003 (Cited on page 16.)

[40] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley, Mar. 2002. http://doi.org/10.1002/0471221104. [Online]. Available: http://dx.doi.org/10.1002/0471221104 (Cited on pages 16, 17, and 24.)

[41] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, Mar. 1986. http://doi.org/10.1109/TAP.1986.1143830. [Online]. Available: https://ieeexplore.ieee.org/document/1143830/ (Cited on page 17.)

[42] B. Dong, J. Antoni, A. Pereira, and W. Kellermann, "Blind separation of incoherent and spatially disjoint sound sources," *Journal of Sound and Vibration*, vol. 383, p. 414–445, Nov. 2016. http://doi.org/10.1016/j.jsv.2016.07.018. [Online]. Available: http://dx.doi.org/10.1016/j.jsv.2016.07.018 (Cited on page 17.)

[43] S. Argentieri and P. Danes, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Oct. 2007, p. 2009–2014. http://doi.org/10.1109/iros.2007.4399422. [Online]. Available: http://dx.doi.org/10.1109/IROS.2007.4399422 (Cited on page 17.)

[44] Q. Huang and N. Lu, "Optimized real-time music algorithm with cpu-gpu architecture," *IEEE Access*, vol. 9, p. 54067–54077, 2021. http://doi.org/10.1109/access.2021.3070980. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3070980 (Cited on pages 17 and 45.)

[45] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, p. 320–327, Aug. 1976. http://doi.org/10.1109/tassp.1976.1162830. [Online]. Available: http://dx.doi.org/10.1109/TASSP.1976.1162830 (Cited on page 18.)

[46] I. Rakotoarisoa, J. Fischer, V. Valeau, D. Marx, C. Prax, and L.-E. Brizzi, "Time-domain delay-and-sum beamforming for time-reversal detection of intermittent acoustic sources in flows," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, p. 2675–2686, Nov. 2014. http://doi.org/10.1121/1.4897402. [Online]. Available: http://dx.doi.org/10.1121/1.4897402 (Cited on pages 18 and 19.)

[47] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, p. 107–151, Jul. 2022. http://doi.org/10.1121/10.0011809. [Online]. Available: http://dx.doi.org/10.1121/10.0011809 (Cited on page 18.)

[48] J. Fischer and C. Doolan, "Beamforming in a reverberant environment using numerical and experimental steering vector formulations," *Mechanical Systems and Signal Processing*, vol. 91, p. 10–22, Jul. 2017. http://doi.org/10.1016/j.ymssp.2016.12.025. [Online]. Available: http://dx.doi.org/10.1016/j.ymssp.2016.12.025 (Cited on page 19.)

[49] R. Porteous, Z. Prime, C. Doolan, D. Moreau, and V. Valeau, "Three-dimensional beamforming of dipolar aeroacoustic sources," *Journal of Sound and Vibration*, vol. 355, p. 117–134, Oct. 2015. http://doi.org/10.1016/j.jsv.2015.06.030. [Online]. Available: http://dx.doi.org/10.1016/j.jsv.2015.06.030 (Cited on page 19.)

[50] K. T. Selvan and R. Janaswamy, "Fraunhofer and fresnel distances: Unified derivation for aperture antennas," *IEEE Antennas and Propagation Magazine*, vol. 59, no. 4, p. 12–15, aug 2017. http://doi.org/10.1109/map.2017.2706648. [Online]. Available: http://dx.doi.org/10.1109/MAP.2017.2706648 (Cited on page 19.)

[51] C. Vanwynsberghe, "Reseaux a grand nombre de microphones : applicabilite et mise en oeuvre," Theses, Universite Pierre et Marie Curie - Paris VI, Dec. 2016. [Online]. Available: https://theses.hal.science/tel-01508547 (Cited on page 20.)

[52] I. A. McCowan, "Robust speech recognition using microphone arrays," Ph.D. dissertation, Queensland University of Technology, 2001, presented to the Speech Research Laboratory, School of Electrical and Electronic Systems Engineering, Queensland University of Technology. (Cited on page 24.)

[53] C. Vanwynsberghe, P. Challande, J. Marchal, R. Marchiano, and F. Ollivier, "A robust and passive method for geometric calibration of large arrays," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, p. 1252–1263, Mar. 2016. http://doi.org/10.1121/1.4944566. [Online]. Available: http://dx.doi.org/10.1121/1.4944566 (Cited on page 24.)

[54] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. (Cited on page 24.)

[55] C. Vanwynsberghe, P. Challande, F. Ollivier, J. Marchal, and R. Marchiano, "Geometric calibration of very large microphone arrays in mismatched free field," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, p. 215–227, Jan. 2019. http://doi.org/10.1121/1.5083829. [Online]. Available: http://dx.doi.org/10.1121/1.5083829 (Cited on page 24.)

[56] H. Demontis, F. Ollivier, and J. Marchal, "3d identification of acoustic sources in rooms using a large-scale microphone array," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, Sep. 2018, p. 506–509. http://doi.org/10.1109/iwaenc.2018.8521264. [Online]. Available: http://dx.doi.org/10.1109/IWAENC.2018.8521264 (Cited on page 30.)

[57] E. Sarradj, "Three-dimensional acoustic source mapping with different beamforming steering vector formulations," *Advances in Acoustics and Vibration*, vol. 2012, p. 1–12, Jun. 2012. http://doi.org/10.1155/2012/292695. [Online]. Available: http://dx.doi.org/10.1155/2012/292695 (Cited on page 30.)

[58] H. Demontis, "Identification de sources acoustiques complexes en milieu réverbérant par grands réseaux de microphones," Thèse de doctorat, Sorbonne Université, France, 2019, ⟨NNT : 2019SORUS196⟩. [Online]. Available: https://tel.archives-ouvertes.fr/tel-03053332 (Cited on page 30.)

[59] Y. Zhou, V. Valeau, J. Marchal, F. Ollivier, and R. Marchiano, "Three-dimensional identification of flow-induced noise sources with a tunnel-shaped array of mems microphones," *Journal of Sound and Vibration*, vol. 482, p. 115459, Sep. 2020. http://doi.org/10.1016/j.jsv.2020.115459. [Online]. Available: http://dx.doi.org/10.1016/j.jsv.2020.115459 (Cited on page 30.)

[60] L. Yu, J. Antoni, and Q. Leclere, "Spectral matrix completion by cyclic projection and application to sound source reconstruction from non-synchronous measurements," *Journal of Sound and Vibration*, vol. 372, p. 31–49, Jun.

2016. http://doi.org/10.1016/j.jsv.2016.02.031. [Online]. Available: http://dx.doi.org/10.1016/j.jsv.2016.02.031 (Cited on page 31.)

[61] K. Wächter, Doebler, Herold, Ocker, and Puhle, "Investigations on beamforming in the wind tunnel using multiple microphone array measurements," in *Proceedings on CD of the 7th Berlin Beamforming Conference, March 5-6, 2018*. GFaI, Gesellschaft zu Förderung angewandter Informatik e.V., Berlin, March 2018. [Online]. Available: http://www.bebec.eu/Downloads/BeBeC2018/Papers/BeBeC-2018-S09.pdf (Cited on page 31.)

[62] L. T. Lima Pereira, R. Merino-Martínez, D. Ragni, D. Gómez-Ariza, and M. Snellen, "Combining asynchronous microphone array measurements for enhanced acoustic imaging and volumetric source mapping," *Applied Acoustics*, vol. 182, p. 108247, Nov. 2021. http://doi.org/10.1016/j.apacoust.2021.108247. [Online]. Available: http://dx.doi.org/10.1016/j.apacoust.2021.108247 (Cited on page 31.)

[63] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, p. 717–772, Apr. 2009. http://doi.org/10.1007/s10208-009-9045-5. [Online]. Available: http://dx.doi.org/10.1007/s10208-009-9045-5 (Cited on page 31.)

[64] F. Ning, J. Song, J. Hu, and J. Wei, "Sound source localization of non-synchronous measurements beamforming with block hermitian matrix completion," *Mechanical Systems and Signal Processing*, vol. 147, p. 107118, Jan. 2021. http://doi.org/10.1016/j.ymssp.2020.107118. [Online]. Available: http://dx.doi.org/10.1016/j.ymssp.2020.107118 (Cited on page 31.)

[65] G. Chardon, "Maximum likelihood estimators and cramér–rao bounds for the localization of an acoustical source with asynchronous arrays," *Journal of Sound and Vibration*, vol. 565, p. 117906, Oct. 2023. http://doi.org/10.1016/j.jsv.2023.117906. [Online]. Available: http://dx.doi.org/10.1016/j.jsv.2023.117906 (Cited on page 31.)

[66] É. Bavu, H. Pujol, A. Garcia, C. Langrenne, S. Hengy, O. Rassy, N. Thome, Y. Karmim, S. Schertzer, and A. Matwyschuk, "Deeplomatics: A deep-learning based multimodal approach for aerial drone detection and localization," in *QUIET DRONES 2022 SECOND INTERNATIONAL SYMPOSIUM ON NOISE FROM UASs/UAVs and eVTOLs SYMPOSIUM PROCEEDINGS*, ser. QUIET DRONES 2022 SECOND INTERNATIONAL SYMPOSIUM ON NOISE FROM UASs/UAVs and eVTOLs SYMPOSIUM PROCEEDINGS. Paris, France: INCE/Europe and CidB, Jun. 2022. [Online]. Available: https://hal.science/hal-03707115 (Cited on page 31.)

[67] H. Hou, F. Ning, Q. Zhai, J. Li, K. Yao, W. Li, and J. Wei, "Cross-spectral matrix denoising for beamforming based on schatten-p norm," *Applied Acoustics*, vol. 197, p. 108938, 2022. http://doi.org/https://doi.org/10.1016/j.apacoust.2022.108938. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X22003127 (Cited on page 38.)

[68] R. DeGroat, E. Dowling, and D. Linebarger, "The constrained MUSIC problem," *IEEE Transactions on Signal Processing*, vol. 41, no. 3, pp. 1445–1449, Mar. 1993, conference Name: IEEE Transactions on Signal Processing. http://doi.org/10.1109/78.205753 (Cited on page 39.)

[69] C. Vanwynsberghe, R. Marchiano, F. Ollivier, P. Challande, H. Moingeon, and J. Marchal, "Design and implementation of a multi-octave-band audio camera for realtime diagnosis," *Applied Acoustics*, vol. 89, pp. 281–287, 2015. http://doi.org/https://doi.org/10.1016/j.apacoust.2014.10.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X14002709 (Cited on pages 58 and 62.)

[70] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002. (Cited on pages 61 and 79.)

[71] M. Kumon, H. G. Okuno, and S. Tajima, "Alternating drive-and-glide flight navigation of a kiteplane for sound source position estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Sep. 2021, p. 2114–2120. http://doi.org/10.1109/iros51168.2021.9636136. [Online]. Available: http://dx.doi.org/10.1109/IROS51168.2021.9636136 (Cited on page 74.)

[72] Springer-Verlag, 2005. http://doi.org/10.1007/1-84628-179-2. [Online]. Available: http://dx.doi.org/10.1007/1-84628-179-2 (Cited on page 77.)

[73] Heudiasyc Laboratory et al., "Flair software," 2022. [Online]. Available: https://gitlab.utc.fr/uav-hds/flair/flair-src (Cited on page 81.)

[74] S. Glegg and W. Devenport, "Chapter 16 - open rotor noise," in *Aeroacoustics of Low Mach Number Flows*, S. Glegg and W. Devenport, Eds. Academic Press, 2017, pp. 399–436. http://doi.org/https://doi.org/10.1016/B978-0-12-809651-2.00016-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128096512000163 (Cited on page 81.)

[75] S. Zhong, P. Zhou, R. Fattah, and X. Zhang, "A revisit of the tonal noise of small rotors," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 476, no. 2244, dec 2020.

http://doi.org/10.1098/rspa.2020.0491. [Online]. Available: http://dx.doi.org/10.1098/rspa.2020.0491 (Cited on pages 82 and 95.)

[76] L. Zhao, T. Zhao, L. Xu, Y. Mao, and H. Wu, *Effect of UAV Configuration on Noise: Experimental Investigation and Analysis*. Springer Nature Singapore, 2024, p. 756–767. http://doi.org/10.1007/978-981-97-4010-9_57. [Online]. Available: http://dx.doi.org/10.1007/978-981-97-4010-9_57 (Cited on page 82.)

[77] G. Corcos, "The structure of the turbulent pressure field in boundary-layer flows," *Journal of Fluid Mechanics*, vol. 18, no. 3, pp. 353–378, 1964. (Cited on page 82.)

[78] A. Dinsenmeyer, Q. Leclère, J. Antoni, and E. Julliard, "A bayesian approach for the separation of the acoustic and the correlated aerodynamic wall pressure fluctuations," *The Journal of the Acoustical Society of America*, vol. 149, no. 6, p. 4410–4421, Jun. 2021. http://doi.org/10.1121/10.0005243. [Online]. Available: http://dx.doi.org/10.1121/10.0005243 (Cited on page 82.)

[79] T. Brooks, D. Pope, and M. Marcolini, "Airfoil self-noise and prediction," vol. 1218, 08 1989. (Cited on page 82.)

[80] M. J. Kingan, R. S. McKay, Y. Wu, R. Jung, and S. T. Go, *Unmanned Aerial Vehicle Propeller Noise*. Springer Nature Switzerland, 2025, pp. 103–118. http://doi.org/10.1007/978-3-031-73935-4_6. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-73935-4_6 (Cited on page 82.)

[81] "Acoustics – Determination of sound power levels of noise sources using sound pressure – Engineering method in an essentially free field over a reflecting plane," International Organization for Standardization, Geneva, Switzerland, Standard ISO 3744:2010, 2010. (Cited on page 84.)

[82] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H. Okuno, "Design of uav-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, Nov. 2017. http://doi.org/10.3390/s17112535. [Online]. Available: http://dx.doi.org/10.3390/s17112535 (Cited on page 87.)

[83] L. Wang, M. Clayton, and A. G. Rossberg, "Drone audition for bioacoustic monitoring," *Methods in Ecology and Evolution*, vol. 14, no. 12, p. 3068–3082, oct 2023. http://doi.org/10.1111/2041-210x.14234. [Online]. Available: http://dx.doi.org/10.1111/2041-210X.14234 (Cited on pages 87, 88, and 119.)

[84] H. Lee and D.-J. Lee, "Rotor interactional effects on aerodynamic and noise characteristics of a small multirotor unmanned aerial vehicle," *Physics of Fluids*, vol. 32, no. 4, Apr. 2020. http://doi.org/10.1063/5.0003992. [Online]. Available: http://dx.doi.org/10.1063/5.0003992 (Cited on page 88.)

[85] P. Joseph, P. Chaitanya, S. Elliott, M. Bhardwaj, and T. P. Chong, "Propeller tonal noise reductions through synchrophasing: Mechanisms and performance," *Journal of Sound and Vibration*, vol. 610, p. 119110, Aug. 2025. http://doi.org/10.1016/j.jsv.2025.119110. [Online]. Available: http://dx.doi.org/10.1016/j.jsv.2025.119110 (Cited on page 88.)

[86] M. Polen, C. Fuller, and N. Schiller, "ANC of Radiated Sound from a Small UAS Rotor using Virtual Error Sensors," no. NASA/TM–2020–5004166, Aug. 2020, national Institute of Aerospace and NASA Langley Research Center. [Online]. Available: https://ntrs.nasa.gov/citations/20205004166 (Cited on page 88.)

[87] A. Deleforge, D. Di Carlo, M. Strauss, R. Serizel, and L. Marcenaro, "Audio-based search and rescue with a drone: Highlights from the ieee signal processing cup 2019 student competition [sp competitions]," *IEEE Signal Processing Magazine*, vol. 36, no. 5, p. 138–144, Sep. 2019. http://doi.org/10.1109/msp.2019.2924687. [Online]. Available: http://dx.doi.org/10.1109/MSP.2019.2924687 (Cited on page 88.)

[88] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, p. 508–519, 2022. http://doi.org/10.1109/taslp.2022.3140550. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2022.3140550 (Cited on page 88.)

[89] F. Asano, M. Kimura, D. Shibuya, and Y. Kamitani, "Localization and extraction of brain activity using generalized eigenvalue decomposition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Mar. 2008, p. 565–568. http://doi.org/10.1109/icassp.2008.4517672. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2008.4517672 (Cited on page 90.)

[90] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, May 2011, p. 3517–3522. http://doi.org/10.1109/icra.2011.5979578. [Online]. Available: http://dx.doi.org/10.1109/ICRA.2011.5979578 (Cited on pages 96 and 100.)

[91] P.-O. Lagacé, F. Ferland, and F. Grondin, "Ego-noise reduction of a mobile robot using noise spatial covariance matrix learning and minimum variance distortionless response," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2023, p. 3533–3538.

http://doi.org/10.1109/iros55552.2023.10342193. [Online]. Available: http://dx.doi.org/10.1109/IROS55552.2023.10342193 (Cited on pages 96 and 100.)

[92] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor uav," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Nov. 2013, p. 3943–3948. http://doi.org/10.1109/iros.2013.6696920. [Online]. Available: http://dx.doi.org/10.1109/IROS.2013.6696920 (Cited on pages 96 and 100.)

[93] J. Hu, D. Zachariah, and P. Stoica, "Target tracking using robust sensor motion control," 2025. http://doi.org/10.48550/ARXIV.2502.20986. [Online]. Available: https://arxiv.org/abs/2502.20986 (Cited on page 118.)

[94] C. de Souza, R. Newbury, C. A., P. Castillo Garcia, B. Vidolov, and K. D., "Decentralized Multi-Agent Pursuit using Deep Reinforcement Learning," in *IEEE International Conference on Robotics and Automation (ICRA 2021)*, vol. 6, no. 3, Xi'an, China, May 2021, pp. 4552–4559. http://doi.org/10.1109/LRA.2021.3068952. [Online]. Available: https://cnrs.hal.science/hal-03584945 (Cited on page 118.)

[95] Y. Sasaki, R. Tanabe, and H. Takemura, "Probabilistic 3d sound source mapping using moving microphone array," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2016, p. 1293–1298. http://doi.org/10.1109/iros.2016.7759214. [Online]. Available: http://dx.doi.org/10.1109/IROS.2016.7759214 (Cited on page 119.)

[96] E. Tiffeneau, Q. Leclère, and S. Bouley, "Séparation adaptative de sources acoustiques par clean-stft," in *Congrès Français d'Acoustique*. Paris, France: Société Française d'Acoustique, Apr. 2025, https://hal.science/hal-05064255. (Cited on page 119.)

[97] Y. He, S. Shin, A. Cherian, N. Trigoni, and A. Markham, "Soundloc3d: Invisible 3d sound source localization and classification using a multimodal rgb-d acoustic camera," 2024. http://doi.org/10.48550/ARXIV.2412.16861. [Online]. Available: https://arxiv.org/abs/2412.16861 (Cited on page 119.)

[98] Y. Girdhar, N. McGuire, L. Cai, S. Jamieson, S. McCammon, B. Claus, J. E. S. Soucie, J. E. Todd, and T. A. Mooney, "Curee: A curious underwater robot for ecosystem exploration," 2023. http://doi.org/10.48550/ARXIV.2303.03943. [Online]. Available: https://arxiv.org/abs/2303.03943 (Cited on pages 119 and 120.)

[99] B. Kölsch, I. S. Walker, B. Schiricke, W. W. Delp, and B. Hoffschmidt, "Quantification of air leakage paths: a comparison of airflow and acoustic measurements," *International Journal of Ventilation*, vol. 22, no. 2, p. 101–121, Aug. 2021. http://doi.org/10.1080/14733315.2021.1966576. [Online]. Available: http://dx.doi.org/10.1080/14733315.2021.1966576 (Cited on page 120.)

[100] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993. (Cited on page 123.)